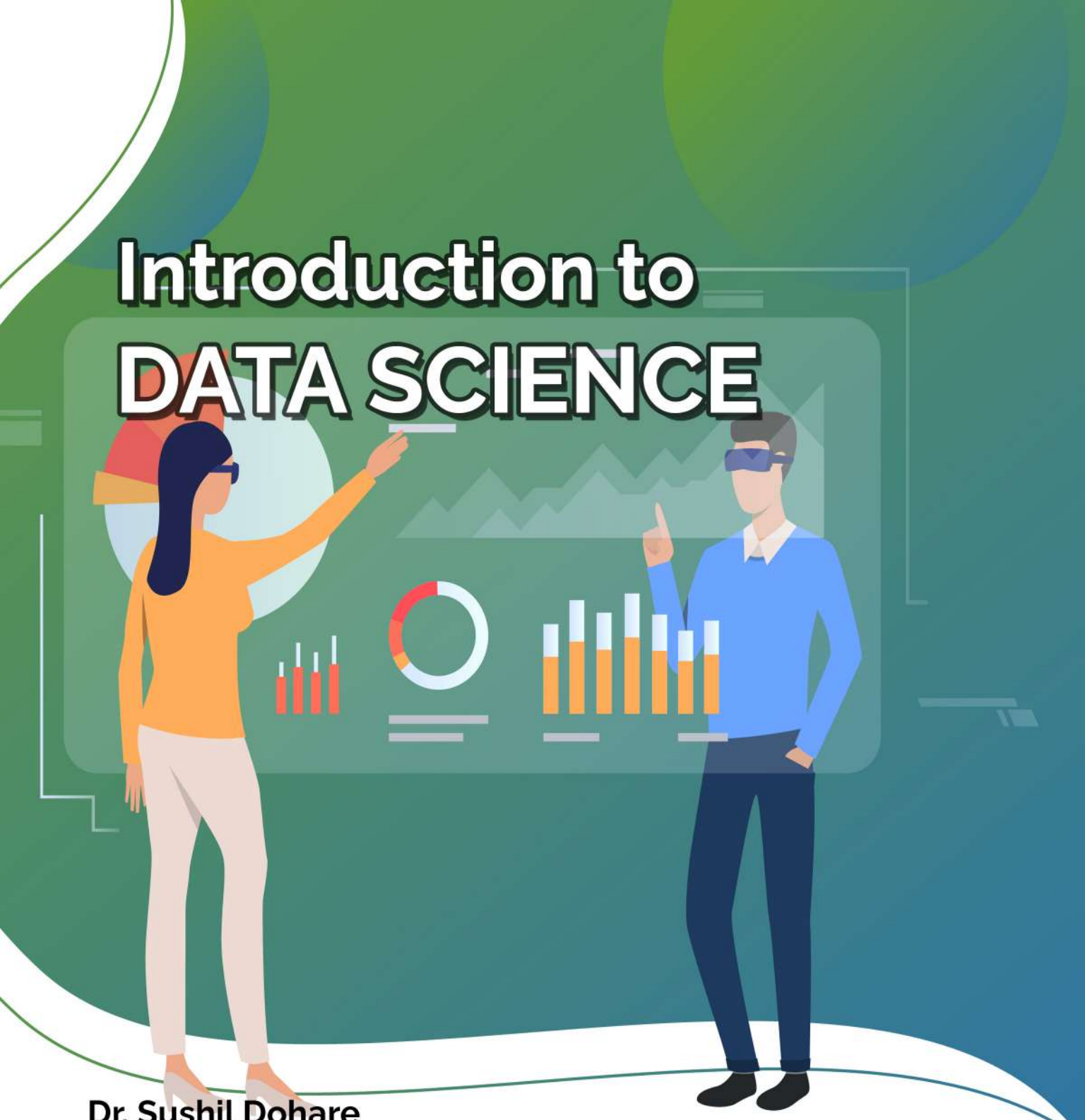
Introduction to DATA SCIENCE



Dr. Sushil Dohare
Dr. V SelvaKumar
Sachin Raval
Dr. Sumegh Shrikant Tharewal

Xoffencer

INTRODUCTION TO DATA SCIENCE

Editors:

- Dr. Sushil Dohare

- Dr. V SelvaKumar

- Sachin Raval

- Dr. Sumegh Shrikant Tharewal

Xoffencer

www.xoffencerpublication.in

ISBN-13: 978-93-94707-67-2 (paperback)

Publication Date: 6 April 2023

MRP: ₹450/-

Published by:

Xoffencer International Publication

Behind shyam vihar vatika, laxmi colony

Dabra, Gwalior, M.P. – 475110

Cover Page Designed by:

Satyam soni

Contact us:

Email: mr.xoffencer@gmail.com

Visit us: www.xofferncerpublication.in

Author Details

Dr. Sushil Dohare

Dr. Sushil Dohare, Experienced Professor of Community Medicine with experience of working with World Health Organization. Skilled in Medical Education, Epidemiology, Personnel Management, Public Health Program Strategic Planning, Public Health Program Implementation and Program Evaluation. Vast international medical education experience as Faculty member in Zawia University Medical College, Zawia, Libya. Presently working as Associate Professor, Department of Epidemiology, College of Public Health and Tropical Medicine, Jazan University, Jazan, Saudi Arabia. Experienced researcher with many original research publications in international journals in areas of non communicable disease Epidemiology, Maternal and Child Health, Application of Nanotechnology in medical sciences. Graduated from MAMC(Maulana Azad Medical College, New Delhi, India); MD from LHMC(Lady Hardinge Medical College, New Delhi, India)

Dr. V SelvaKumar

Dr. V SelvaKumar, Assistant Professor in the Department of Mathematics and Statistics, Bhavan's Vivekananda College of Science, Humanities & Commerce. He did his Ph.D from BITS Pilani, Hyderabad Campus. Dr V Selvakumar has 21 years of experience as an active academician and researcher. He has published 22 papers in different national and international journals, 5 patents, and authored a book to his credit. Also, presented twelve papers at national and international conferences. His areas of interest are Data analytics, Time Series Analysis, Machine Learning and Deep learning.

Sachin Raval

Sachin Raval is a research scholar and am currently pursuing triple Master's degrees in International Finance, Economics, and Law from three prestigious European universities - the University of Macerata in Italy, Nicolaus Copernicus University in Toruń in Poland, and the University of Angers in France. Additionally, he hold a Bachelor's degree in Commerce, a Master's degree in Commerce, a Bachelor's degree in Arts in Shastri-Sanskrit, and an ITI Trade certification course in Computer Operator and Programming Assistant trade. he have gained professional experience by working on several projects at Tata Consultancy Services Limited in India.

Dr. Sumegh Shrikant Tharewal

M.Sc., Ph.D. (Computer Science)

Dr. Sumegh Shrikant Tharewal Currently working as an Assistant Professor at Symbiosis Institute of Computer Studies and Research (SICSR), Symbiosis International (Deemed University), Pune, MH 411016, India, he completed his Ph.D. from Dr. Babasaheb Ambedkar Marathwada University Aurangabad, Maharashtra, India in the Department of Computer Science, and Information Technology. He was Program Head of M.Sc. Blockchain Technology at Dr. Vishwanath Karad MIT World Peace University, Pune, India. He has published more than 42 Research Papers in various national, and international conferences, and International Peer-Reviewed Journals like IEEE, Springer, and Elsevier. he received 214 citations with a 9 h index on Google Scholar for his publication.

Preface

The text has been written in simple language and style in well organized and systematic way and utmost care has been taken to cover the entire prescribed procedures for Science Students.

We express our sincere gratitude to the authors not only for their effort in preparing the procedures for the present volume, but also their patience in waiting to see their work in print. Finally, we are also thankful to our publishers **Xoffencer Publishers, Gwalior, Madhya Pradesh** for taking all the efforts in bringing out this volume in short span time.

Abstract

If you want to manipulate data successfully, you need tools to do it, which means you need computer programming skills and some understanding of algorithms and data structures. If you want to manipulate data efficiently, you need tools to do it. Often there is an additional requirement that must be met. Data science always emphasizes data itself as the main subject. The focus of a project involving data analytics is the process of bringing data from its original state into a form that can be summarized and used through a series of operations. This process starts with the data in its original state. While there is a slight difference between the two, the focus of all work is not what the computer does, but the flow of data and how it changes. It also focuses on why certain data changes were made, what purposes those changes serve, and how those changes help us better understand the data. A data management strategy is just as important as an efficient way to carry out the process involved. Obviously, statistics has a deep and important connection to data science in different ways. In fact, many people think that data science is nothing more than a fancy nickname for statistics that sounds a little cooler and more appealing. This is because the two areas dock close together. Data science sounds more appealing than statistics, but just as data science is only slightly different from computer science, statistics and data science are only slightly different from each other.

Contents

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

The response to this inquiry is not at all easy to comprehend. I'm not sure how simple it is to discover someone who has a complete comprehension of what data science is, but I am certain that it would be challenging to locate two individuals who have fewer than three points of view on the topic. I am not sure how simple it is to locate someone who is well-versed in all aspects of what data science entails. Finding a person who is well-versed in all facets of data science may not be as simple as it initially appears to be. I cannot give you a definite answer. It's safe to say that it's a buzzword, and it seems like every data scientist desires it these days; as a result, having a background in data science is a useful thing to add to a résumé. Because of this, the role of "data scientist" has become increasingly common.

But what exactly does it mean? Because I am unable to provide you with a definition that the vast majority of people will comprehend, I will instead provide you with the definition that I personally employ: The branch of study known as "Data Science" concentrates on the process of deriving information from other types of information that has been gathered. Data This description touches on so many different areas and almost encompasses so much ground that it is almost incomprehensible. It's not a mystery to me at all. Having said that, I believe that the discipline of data science encompasses a huge breadth of subject areas and subfields. There is nothing that makes me feel less ashamed than that. It is possible that the purpose of any scientific endeavor is to gather information from the evidence that has been gathered, and you may be correct if you argue this point.

On the other hand, I would contend that the scientific approach entails more than simply transforming unprocessed data into information that can be understood. This is what I refer to when I make a declaration like this. Data science, in contrast to the conventional emphasis of science, which is to provide solutions to particular inquiries about the

world, is concerned with determining how data can be processed in a way that is both efficient and meaningful. Because of this change in emphasis, data science can now concentrate on determining how to process data in a way that is both effective and significant.

It is not as essential to decide what questions to ask the data as it is to define how we can address any questions that we do end up asking the data. In this respect, it is more preoccupied with mathematics and computer science than it is with natural sciences in general. In this environment, the process of learning the skills necessary to accurately understand data takes a back seat to the more important task of actually working in the natural environment. Experimentation is an essential component of data science, and this applies to both the planning and carrying out stages of the process. If we are able to obtain the essential statistics, we will be able to investigate the subjects that are of interest to us. This can be challenging if the studies that are conducted are not well organized, or if we do not select the data that we acquire with great care.

The planning of studies is not the primary focus of this work; rather, the emphasis is on the methodology that was used in the research, which is one of the most essential components of data science. The analysis of the data that was collected is going to be the primary emphasis of this endeavor. Research in computer science is another fundamental axis of the field; however, the purview of this research is significantly greater than that of mathematics. However, datalogy, which is an older term for data science, has also been suggested for computer science and is the name for computer science in some countries such as Denmark. The use of the name "computer science" places an emphasis on computation, but datalogy has also been proposed for computer science. The term "data science" places a greater emphasis on data than it does on mathematical computations. However, there are many areas of research that intersect with one another.

When you are constructing an algorithm for sorting, do you put the majority of your attention on the computations themselves, or do you give more attention to the data with which you are working? Is it really the case that this is a question that needs to be addressed right off the bat? Because there is a high degree of overlap between computer

science and data science disciplines, the skill sets required of you will also coincide. Therefore, it is essential that you have a solid foundation in both disciplines.

If you want to manipulate data successfully, you need tools to do it, which means you need computer programming abilities and some comprehension of algorithms and data structures. If you want to manipulate data effectively, you need instruments to do it. Often there is an additional prerequisite that must be fulfilled. Data science always prioritizes data itself as the primary subject. The emphasis of a project involving data analytics is the process of bringing data from its original state into a form that can be summarized and used through a sequence of operations. This process begins with the data in its original condition. While there is a subtle difference between the two, the emphasis of all work is not what the computer does, but the flow of data and how it changes. In addition to this, it concentrates on the reasons why certain data changes were made, the functions that those changes serve, and the ways in which those changes help us better comprehend the data.

It is just as essential to have a data management strategy as it is to have an effective method to carry out the procedure that is involved. It should come as no surprise that statistics shares a profound and significant relationship to data science in a variety of different ways. In point of fact, a lot of people believe that data science is nothing more than a flashy alias for statistics that sounds a little bit more hip and interesting. This is due to the proximity of the two regions' docks to one another. Statistics may not have the same allure as data science, but the two fields are actually not that dissimilar from one another. Just as data science is only slightly distinct from computer science, statistics and data science are only slightly distinct from one another.

I wouldn't say that I completely disagree with you, but I also wouldn't say that I concur with you. Perhaps something that, in terms of the magnitude of the disparity, can be compared to general mathematics. A significant portion of the work that must be done in order to conduct statistics involves the creation of mathematical models that are suitable for the data, followed by the application of those models to the data in order to gain insights into the data. Producing statistics requires a significant amount of effort. Our work in the field of data science also includes this other component. I don't see why labelling statistics "data science" would be problematic as long as the emphasis is

placed on the information that is gathered and analyzed. As the emphasis changes from models and mathematics to the computations themselves rather than the data themselves, we have lost interest in data science and have instead shifted our attention to a different subject matter. We are transitioning from the field of data science into the field of computer science as the emphasis moves from data to calculation.

The disciplines of data science, machine learning, and artificial intelligence are all intertwined, and this is another area in which there are numerous similarities between the three subfields. If you specialize in data analytics, you also have roots in data science; This may not come as much of a surprise given that the field has its roots in both computer science and statistics; If you concentrate on data science, you also have roots in machine learning; This may not come as much of a surprise given that the field has its origins in both computer science and data science. To tell you the truth, I've never fully grasped the concept of converting a model in a quantitative model that was merely a statistical model into a model that can be applied to machine learning. This is something that I have a lot of trouble doing. For the purposes of this book, we will continue to use the definition that I have supplied and we will refer to it as Data Science. This is due to the fact that the primary emphasis of our company will be data analysis.

Instead of concentrating predominantly on programming, the first seven chapters of this book present a conversation that is centered on data analysis rather than programming. The viewer will acquire knowledge regarding software architecture, algorithms, data structures, and other related topics throughout these seven chapters. I do not make the assumption that you will have a comprehensive understanding of the principles. In the same vein, I will not presume that you are familiar with the programming language R or have any previous experience using it. However, I will assume that you have some prior understanding in certain fields, such as programming, mathematical modeling, and statistics. Analyzes. If you have experience with other programming approaches, such as scripting or object-oriented languages, you might discover that R programming presents additional challenges.

This is particularly the case if you have prior familiarity with several different programming paradigms. R is a functional programming language, which means that it does not permit the manipulation of data. In addition, while it does have object-

oriented programming systems, it manages this programming paradigm in a very different way than languages like Java and Python, which you have undoubtedly already encountered. R does not permit users to change data because the program does not permit users to amend the data.

R does not permit users to change data because the program does not permit users to amend the data. We will only use R for relatively simple programming tasks throughout the data analysis portion of this book, which includes the first seven chapters; therefore, none of this should present any difficulties. Because we have to create fundamental programs in order to process and summarize data, we have access to various programming constructs such as function calls, if statements, loops, and so on. It is necessary for us to be familiar with the construction of fundamental expressions such as Because we will require it. To be able to cope with all of these different aspects, you need to have a great deal of self-assurance. I won't spend an excessive amount of time talking through each of these constructs one at a time, but I will cover each of them in the book and introduce them to you when it is necessary to do so that you can see how they are expressed in R. I'll proceed in this manner so that you can observe the structure of the R code.

I will build a significant portion of my conclusion on your capacity to comprehend this concept using a variety of illustrations. Keeping this in mind, I won't presume that you already have a solid understanding of how to match data and compare models in R. In point of fact, I seriously doubt that you have any experience similar to that. My working presumption is that you have had sufficient exposure to statistics to become acquainted with fundamental concepts such as parameter estimation, model fitting, explanatory and response variables, and model comparisons. This is my working assumption. This is the hypothesis I'm working with. In any case, even if that's not the case, I believe you're at least capable of grasping what it is that we're discussing. Despite the fact that I don't believe you have a strong foundation in programming and statistics, since this isn't an overnight training to become a data scientist, you should be able to come up with instances on your own so that I don't have to accompanies you every step of the way. You don't seem to have much of a foundation in either programming or statistics, in my opinion.

One of my students completed a data analysis project during one of my first semesters of teaching, and the following portion is a condensed and simplified synopsis of that project. Following the completion of the first seven segments, this undertaking was considered finished. There is a sample of what such an undertaking might look like, but if you want to get started on your own analysis right away, you shouldn't wait until you have finished reading the first seven portions of this book before you get going on it. You have to get started on your research as soon as you possibly can if you want to get anywhere with it. If you want to get the most out of reading this book, you need to make sure that you consistently apply what you've learned into practice. If you want to get the most out of reading this book, you need to do this. You need to put in the effort to get the most out of this work if you want to get the most out of it. Before beginning to read this book, select a dataset that you have an interest in learning more about, and after finishing each chapter, put what you've learned in that chapter into practice by applying it to the dataset. You will have a better chance of remembering the information if you do it this manner.

This book devotes its final seven segments to the discussion of programming as its main subject. You should be familiar with object-oriented programming fundamentals and have some experience in the field before continuing on to the next section. Before we go any further, I just want to make sure that you are well-versed in concepts such as class hierarchies, inheritance, and polymorphic methods. I'll describe how it's done in R, along with the ways in which it contrasts from languages such as Python, Java, and C++. I don't believe you're acquainted with functional programming because I don't think you are (but even if you are, you should still have a lot to learn about R programming from these chapters if you aren't already familiar with it).

In the book, we go over the fundamentals of data processing, which include things like filtering and selecting pertinent data, transforming data into forms that can be readily analyzed, summarizing data, presenting data in an instructive way to examine data, as well as presenting and generating results. your images. Exactly these are some of the subjects that we look into. When conducting research in the field of data science, it is essential to keep these things in mind because they are the most significant factors to take into account. Next, we'll take a look at how to develop new R packages that others

can use in their own projects, as well as how to actually construct those packages. These packages are not only reusable but also simple to substitute, compatible with a wide variety of other packages, and straightforward to use.

These are the fundamental abilities you need to cultivate and demonstrate to the rest of the world how your procedures work. For the purpose of carrying out all of these responsibilities, the programming language known as R (https://www.r-project.org/about.html) is utilized. R has shot up the ranks to become one of the most prominent and widely used computer languages in the field of data analysis in a very short amount of time. In addition, R can be obtained for free by anyone who is interested. Because R is so widely used, it comes with a vast collection of plugins (which are referred to as "packages" in R). These plugins can perform virtually any type of research that you might be interested in. There is no question that brilliance and recognition do not automatically go hand in hand with one another. Because individuals who discover new statistical techniques frequently implement them as R packages, you can navigate R's more sophisticated methods in a very straightforward manner.

This is due to the fact that individuals who develop statistical methods frequently implement them as R packages, which then monitor your utilization of the methods. Because it is highly improbable that you will be the first person to go through any given experience, the majority of the challenges you face can be overcome by conducting a short search on Google. In addition, there are a great deal of R and bespoke package-specific educational resources available online in the form of tutorials. There are also a great number of videos that provide instruction on R and notable R packages; Finally, if you want to learn more, there are a great number of publications that you can purchase. There is access to all of these materials via the internet.

While RStudio is already operating, you are able to enter R statements in the console. The RStudio window has a section on the left side that serves as the location of the interface. When all three stages are finished, R will read an expression that was entered, then it will evaluate that expression, and ultimately it will present the outcome of the evaluation. You will soon be taught how to designate values to variables, and once you do so, those values will be displayed in the environment frame that is located in the top right portion of the screen. You can gain access to this section by selecting the button

labelled "Media" that is located in the top right portion of the screen. If you look in the bottom right corner of the screen, you'll see a shortcut that takes you to the directory where the document was stored.

This directory is used as the default saving location for any and all freshly generated files. You can make a new file by going to the "File" option and choosing "New File" from the drop-down selection that appears there. You have a selection of options available to you in terms of the kinds of files to use. Both the R Script and the R Markdown classes are very essential to us for the work that we do. The first is a file structure for R code that is not accompanied by any documentation text, and the second is a method for producing reports that include both R code and documentation text together. In my opinion, markdown files ought to be utilized for any kind of task that involves conducting data analysis. It is of the utmost importance to keep detailed written documentation of what it is that you are working on at the moment because it is possible that you will need to come back to a project after a few months have passed.

1.2 OBTAINING DOCUMENTS FOR FEATURES

It is simple to neglect the complexities of a function, and more specifically, what each component of a function does when it is called. Due to the fact that these particulars are so easily neglected, it is imperative that the function's specifications be checked on a regular basis. Therefore, it is simple to become complacent and neglect to consult the documentation for your functions. This procedure can be made much easier and completed much quicker by using the free programs R and RStudio. If the documentation for the function was written by the person who created the function, all you have to do is tell R what you want to know about a function, and it will give you the information you need. Take into consideration the duration of the function in the scenario that was presented to you earlier. Simply typing "length" into the R interface will provide you with additional information regarding its operation and function.

When you carry out these steps in RStudio, the documentation will show up in the window to the right, as demonstrated in the image on the right. In addition, there is a construct that is referred to as the next statement, and it is this construct that causes the loop to proceed to the next repetition of the sequence. After going over what loops are

and how they function, I feel obligated to mention that they are not utilized nearly as frequently in the programming language R as they are in a variety of other programming languages. This is something I should point out before moving on to the next topic. Because I believe it to be essential, it is something that I should bring to your attention. The use of loops is fiercely discouraged by a lot of people because of the perception that they make programming slower. The cause is their stellar reputation.

Because it is much simpler to create sluggish code with loops than it is to write code without loops, it follows that loops on their own serve no purpose and should be avoided. Loops are replaced with functions so that function support can be provided for loop functions. In most cases, there is a function for everything that you want to accomplish with a loop, and if there isn't one, you can typically combine three different functions to accomplish what you want to do: matching, filtering, and shortening. In most cases, there is a function for everything that you want to accomplish with a loop. If the goal you want to accomplish with a loop doesn't already have a function associated with it, you can almost certainly find one. Even if there isn't a function that will do what you want, there is almost always a method to do it using a loop. This is true even if there isn't a function that will do what you want. Because this subject goes beyond the purview of the conversation that we are having right now, we will circle back around to it in a later portion of the book.

1.3 DATA FRAMEWORK

No matter what form they take, the vectors that have been discussed so far can all be interpreted as sequences of data. They do not have a structure aside from the sequence in which they are presented, which may or may not be significant when taking into consideration how the data will be interpreted. The sequence in which they were issued is the sole framework that they possess. The data is not displayed in that manner, and we do not intend to investigate this topic further. The majority of the time, we are working with a large number of interconnected variables that are all components of the same observations. You will have a number for each of these variables for every data point that has been recorded. (or evidence of missing data if certain variables were not observed). In point of fact, there is now a table in front of you that contains a row for each measurement and a column for each variable that has been inputted Data.

When dealing with data, R makes use of the frame data type to symbolize the various table kinds that it encounters. A data frame is a collection of vectors, each of which must have the same length and is handled as if it were just a two-dimensional matrix. The length of each vector must be the same. It is required that all of the vectors contained within a data block have the same length. When we conceive of data blocks, we often picture them as having rows that correspond to observations and columns that correspond to the characteristics of those observations. This is because rows and columns are often organized in a table format. Therefore, we interpret them as being arranged in rows and columns. When viewed in this light, data blocks transform into extremely helpful instruments for the statistical modelling and fitting processes.

1.4 PROCESSING WITH MISSING VALUES

"Missing values" are attributes that are not presented or stored inadvertently and ought to be secreted away. The overwhelming majority of datasets have "missing values," also known as parameters that were not inadvertently witnessed or documented. This is because these values were not included in the dataset. Even if the only thing you do to solve the issue is get rid of all the instances that have missing data, you still need to deal with missing data in the research. In point of fact, there is only one method that can resolve the issue, and that is to delete all observations with incomplete data. When conducting an analysis, using data and algorithms to make decisions about how to manage lacking data can be helpful. The specific number NA is used to symbolize any data in R that cannot be found because it does not exist.

This can be accomplished in a number of different methods. (not available). It is essential to keep in mind that R is conscious that NA stands for "missing values," and that he handles NAs in accordance with his own comprehension of what the term "NA" means. It is possible for values of any sort to be lacking, in which case they will be presented as NA. If there is a lack of information, you should never provide a particular number; rather, you should always use the "NA" designation. (-1 or 999 or whatever). R is used for the NA processing technique, but it has no means of understanding that -1 symbolizes anything other than a negative one. This is because R is a symbol-based programming language. Additionally, there is no means to determine whether or not the value -1 symbolizes something other than negative numbers.

The word "NA" alludes to the activities themselves when they are involved in a situation where there are NAs. It is mathematically impossible to perform an operation on missing data and receive an outcome that is anything other than numerous missing values as a consequence of performing the operation on the missing data. This further suggests that the outcome of a comparison between two NAs will also be NA as a consequence of the comparison. NA is not doing well even with himself due to his lack of comprehension, and this is a problem.

Conducting data analysis in this manner does not, in the overwhelming majority of occurrences, result in any kind of difficulty. However, in the vast majority of cases, a lot more steps than these are necessary to be taken. If this is the case, you have two options: either be very inventive when labelling the variables on which you store data, or overwrite the variable names by reassigning them to a variable after the data has been modified. If you choose the former, you will need to be very creative. If your answer is affirmative, the best choice for you is the first one. If that is the case, coming up with creative names for the variables that are used to record the data is going to require a lot of effort on your behalf. Some people find it challenging to remember many variable titles and to change the values of their variables frequently.

It is more likely that you will make a blunder and execute a procedure on the incorrect variable when you have a greater number of variables than when you have a smaller number of variables. You might, for instance, decide to summarize my data variable rather than the clean data field. A warning will be generated for you if you attempt to invoke a procedure using the name of a variable that is not presently being used in the program. On the other hand, it is not feasible to determine in advance whether or not the error will be simple. If you execute a procedure with the incorrect arguments, it's possible that you won't even realize there's a problem. Having said that, there is a very high likelihood that the technique will produce an erroneous outcome. Because of the nature of the error, correcting it won't be a simple task in the years to come. When it happens as part of a reassignment to a variable as part of the operation, the problem is much less severe than when it occurs on its own.

If you are utilizing R in an implementation that allows for user interaction, you will most likely run into issues with it. If you want to go back and change any portion of

the program that you made, you will first need to return to the beginning, which is where the data was received. This is due to the fact that data are received relatively early on in the process. You can't just decide to start using a variable in the middle of a bunch of function calls, especially if it doesn't have the data you were working with when you initially executed the program. This feature is not available in any part of the program. There is no way that can happen. If you always execute your R programs from beginning, you won't have as much of an issue with this particular issue.

However, the most common method to use R is to implement it in an interactive interface or markdown website, both of which have the potential to result in undesirable outcomes. On the other hand, if you continue to execute your R programs from beginning, you will continue to experience this issue. Therefore, one of the options is to avoid running the functions one at a time and instead allocate the results of each transient test to a different variable. A strategy for addressing the issue at hand. In the previous illustration, there were four instructions, one for each function call. However, in this example, you will instead send the outcome of one function call to the subsequent function call. Because of this, there will be no need for any additional notifications to be sent.

Since "big data" can be described in a variety of ways, it may continue to be challenging for the majority of people to grasp such a concept. The term "big data" refers to the aggregation of data sets that are not able to be kept in a single location due to their sheer size. Today, different individuals have various conceptualizations of what "big data" entails and means. Some individuals believe that it is a collection that is bigger than a particular limit, such as more than a terabyte (Driscoll 2010), while others believe that it is data that can be analyzed using popular analysis tools such as Microsoft Excel. Big data, on the other hand, refers to information that demonstrates the characteristics that are the most easily identifiable, such as diversity, speed, and workload. (Laney 2001; McAfee and Brynjolfsson 2012; IBM 2013; Marr 2015). Big data analytics is an innovative strategy that extracts valuable insights from low-value data that for some reason does not fit into regular computer systems.

The strategy makes use of a wide variety of technologies and procedures to accomplish this goal. Utilizing the accessible data to its full potential is the primary objective of

this tactic. In the absence of the moratorium, it appears that there is a description that more accurately describes this occurrence (Dumbill 2013; De Mauro et al. 2015; Korea 2016): inexpensive. read Big data analytics is a cutting-edge illustration that summarizes the many different technologies and procedures that can be used to extract useful insights from data. Even though each of these explanations is accurate in some way, there appears to be one that is more adequate in explaining this occurrence. In recent years, there has been a substantial increase in the number of scientific studies focusing on big data. (Lynch 2008). There are a variety of applications for big data that can be investigated in virtually all areas of scholarly study. The performance advantage that the data-driven business enjoys over its rivals is between 5 and 6%. This is true independent of the particular sphere of endeavor being discussed.

A variety of writers present the following five recommendations for establishing an efficient strategy for large data: putting more of an emphasis on using information to generate value; including behavioral specialists on the team; putting more of an emphasis on learning; and putting more of an emphasis on business problems as opposed to technology-related ones. I emphasize the use of knowledge to add value, I bring behavioral experts onto the team, and I place a strong emphasis on learning. I do this by putting people at the center of big data initiatives; ii by putting a strong emphasis on using knowledge; iii by adding behavioral experts to the team. Home Concentrate on your task rather than the problems caused by technology. Choosing the appropriate data, concentrating on the key performance factor to optimize business operations, and transforming organizational capabilities are the three separate key features that Barton and Court (2012) recognized as being necessary in order to unleash the potential of big data.

To fully realize the possibilities of big data, having all of these capabilities is absolutely necessary. The amassing of data is rapidly developing into a new type of capital, a unique currency, and a primary source of worth. Already acknowledged is the significance of transforming the potential of big data into a workable strategy for the purpose of effectively controlling and expanding business operations. The judgements that need to be made in management, as well as the expansion of the company, are the foci of this strategy. However, it is a known truth that big data applications might not

be useful for all businesses. The primary reason for this is that the organizational components of a company or the company itself might not be suited for such applications. Having a data strategy, on the other hand, is always beneficial, regardless of the amount of data that you possess. In point of fact, having a data strategy can assist you in making effective use of the data that you already possess. Therefore, before you can even begin the process of building a data frame for an organization, you need to first dispel a few common misconceptions, which are as follows:

More data means more precision. Due to the fact that not all data are of a high quality, the presence of poor data within a data collection can eventually have an effect on the standard of the products that are manufactured. It is analogous to receiving a blood transfusion, in which the consequences for the entire body can be catastrophic in the event that the blood used in the procedure is not compatible with the blood type of the beneficiary. In this particular instance, the beneficiary of the transfusion is the organism itself. Second, in spite of the proverb that "if you torture the data enough, nature will always confess" (Cose 2012), there is always the possibility that the data will suit the model very well, which can make it more difficult to acquire new information. This can be a barrier to progress.

It is essential to keep in mind that aiming for perfection in one of the applications that make use of big data is not the most productive approach. When an excessive number of variables are added to a model, the complexity of the model rises, but this does not necessarily make the model more accurate or efficient. The collection of additional data will likely incur additional expenses, but this does not necessarily equate to improved precision. The amount of effort, time, and opportunity expenses that are necessary to acquire data directly correlates to the degree of difficulty associated with making decisions and accurately understanding findings. On top of that, the difficulty of the circumstance is only going to get worse. These are the expenses that are connected. Furthermore, there is an increase in the amount of money spent on maintenance, which applies to both the physical storage and the model preservation.

However, these facts need to be verified and confirmed, and they have the potential to challenge the conventional way of thinking. Obviously, the facts that are used do not have to be implemented in a conventional or typical manner, and this is where the true

advantage resides. In a nutshell, the data strategies that are the most successful begin with an examination of the datasets that are already held by an organization, and then proceed to combine those datasets with data obtained from either public or private sources. This is the most effective method for ensuring that your data approach is as productive as it possibly can be. Don't just store or analyze data for the sake of storing or analyzing data because the amount of data generated each day is getting close to the point where the noise will outweigh the signal. (Silver 2013). The 80/20 rule that was described by Pareto is applicable in this scenario: there is a high possibility that only 20% of the data that is currently accessible can appropriately explain 80% of the events that have been witnessed.

Data equals objectivity. The first stage is to contextualize the data because the "objective" meaning of the data shifts depending on the surroundings in which it is examined. This makes contextualizing the data the most important process. It is possible that this assertion is subjective, despite the fact that it may appear to be controversial. This is the case when statistics represent objective, solely human or societal concepts while recording facts about natural occurrences. While subjective data represent human notions or societal concepts, objective data are recordings of actual occurrences in the natural world. In other words, the data can be factual, in which case it is clearly identical to the minister; it can also be traditional and social, in which case it is more abstract data that warrants the claim of representation from the people's assembly; and finally, the data can be neither factual nor traditional nor social; in this case, it is clearly identical to the minister.

In either scenario, the facts in question may be categorized as either accurate or conventional and societal. Think about things like prices, values, and analogous concepts when deciding what to include in this second category of material, in addition to the various other categories. It is essential to make this differentiation, as the second group is much simpler to manage, while the first group runs the risk of becoming the focus of a self-fulfilling prophesy. As we have seen in the previous section, the interpretation of data is the primary element that plays a role in determining how much value an organization places on that data. In the end, different onlookers may acquire different perspectives on each data category due to the relative character of the problem

frameworks or their ability to analyze the data. This may be the case because of the ability to analyze the data. It's possible that this is due to the fact that different witnesses have varying capacities. (known as the framing effect).

Since data science would lack complete impartiality and reproducibility, it can never be considered a legitimate scientific discipline for this reason. Additionally, not all measurements are capable of having their values determined precisely, so only approximations can be provided for them. We are not conscious that people are susceptible to a variety of behavioral biases, each of which has the potential to make research less objective. All of these biases have the potential to make research less objective. You should bear this in mind moving forward. The most common types of logical fallacies include apophenia, which is the propensity to see patterns where there are none, narrative fallacy, which is the need to put patterns into a set of facts that are unconnected, and confirmation bias, which is the tendency to use only information that confirms one's preexisting beliefs in a given circumstance. context. and selection bias, which is the propensity to only use information that backs up certain principles with the conclusion that the search for evidence will result in the discovery of evidence.

This leads to the conclusion that the search for evidence will result in the discovery of evidence. The desire to find patterns where none exist is an example of the narrative fallacy, which is also referred to as apophenia. Apophenia is defined as the compulsion to find patterns in data that are not connected to one another. (the tendency to always use some type of data, perhaps the most familiar). The widespread acclaim that the well-known performer received in the media contributed to an increase in the stock price of Berkshire Hathaway, which is controlled by Warren Buffett. The purpose of this is to bring to your attention one more fascinating facet of the problem with big data. This occurrence has recently been referred to as the "Hathaway effect," which is a word coined relatively recently. This suggests that there are sometimes connections that have no foundation in reality, are completely meaningless, and either are completely correct or are completely incorrect.

Your data will tell you the whole truth. The questions you need to ask yourself in order to understand the facts are essential before you can make meaning of the facts on their own. Big data has the potential to reveal the solution to everything, including life,

the universe, and everything else, as suggested by Deep Thought in the novel Hitchhiker's Guide to the Galaxy by Douglas Adams, which was first published many years ago. But in order to accomplish that, you need to make sure you're asking the correct question of yourself. Although a computer might be able to provide a quicker response to a specific quantifiable question, human judgement is still required in the end. Although it's still feasible for a human intellect to pose the right question and accurately understand the findings, a computer can answer the question itself much more quickly.

The alternative method, which involves the discovery of data at random and is often referred to as the "let the data speak" method, is extremely unproductive, requires a significant amount of resources, and has the potential to eliminate value. Because we "don't know what we don't know," it makes perfect sense to have a strategy that involves both intelligent data discovery and experimental analysis. For the simple reason that "we don't know what we don't know." (Carter 2011). One of the primary reasons why data mining is frequently ineffective is that it is frequently carried out without justification, which leads to errors such as false positives, overfitting, disregarding false correlations, sample bias, and so on. This also leads to common errors such as reversing the causative link and using false variables. Registration of the model, or choice of it. The fact that data mining is frequently carried out without sufficient substantiation is one of the primary contributors to its frequent inefficiency. (Doornik and Hendry 2015; Harford 2014).

In light of the fact that the observational data only takes into account the second component, a disproportionate amount of attention and research should be dedicated to the question of causality and association. On the other hand, according to Varian (2013), conducting experiments is the single most essential thing to do in order to figure out how to address the situation.

In most cases, the stages involved in data analysis are as follows: first, you capture your data, and then, you save that data to a file, a spreadsheet, or a database. After that, you will be able to move on to the following step of the procedure. After that, an analysis is carried out, which is written in a variety of scripts. You might save some intermediary findings, or you might continue to work on the raw data continuously. After that, you'll

have a few different choices. After developing charts or tables that contain highlights of the pertinent data, continue reporting the findings by using a word processor or text editor. The analysis ought to include a discussion of the ramifications of the conclusions. The standard operating procedure is followed.

This is something that is done more or less by the overwhelming majority of individuals who analyze data. On the other hand, it is a procedure that has a high probability of producing incorrect outcomes due to the nature of its steps. There is a wall that divides the data and the analysis programs, and there is a second wall that divides the actual analysis from the analysis documents. Walls stand between the two walls, separating them from each other. If all of the research is done on unprocessed data, then problem number one is not a significant problem at all. On the other hand, distinct programs are stored for each individual task component by default. As an illustration, a script might record intermediary output, which the following script would then view and process. To be able to reproduce an experiment, you will need to execute all of the scripts in the correct sequence, and each script will illustrate a different stage in the process of analyzing the data.

This proper sequence is only kept in a text file, or even worse, in the memory of the data scientist who is developing the procedure. It takes place quite frequently. The regrettable reality that it won't last very long and will probably vanish before it's required once more is a circumstance that only makes the situation even more precarious. Always make an effort to create your processing routines in such a way that any portion of your process can be replicated completely automatically at any moment. This is something you should try to do as often as possible. Something that you should keep an eye out for at all times. This is the greatest scenario that could possibly occur. The second obstacle presents a problem in that, despite the fact that the process is automatic and simple to restart, there is a tendency for the documentation to become disconnected from the actual processing programs in a relatively brief period of time.

Whether or not the procedure is automatic is something that needs to be considered. Even if you take notes, it is highly unlikely that you will be able to recall to later update the documentation if you make modifications to the programs. However, you do not need to remember to remember to update the documentation for each analytic run. It is

probable that you will not neglect to update outlines, tables, and other related things. Different kinds of modifications include filter choices, function overrides, and similar kinds of overrides. When the documentation departs from the actual analysis to an unacceptable degree, it becomes useless and must be discarded because it is no longer pertinent to the problem at hand.

You can definitely depend on automated routines to accurately represent real data analytics each and every time; this is one of the advantages of using automated analytics processes in the first place; however, there is a high risk that the documentary will be completely made up of fabricated information. One of the many advantages of utilizing automatic analytic techniques is that this is one of them. Let's make the reasonable assumption that you are searching for a method to produce changeable documentation. reports that provide a concise and understandable summary of the research technique for both people and computers. The report is used by computers as part of an automatic procedure, and the investigation can be carried out whenever it is needed. People use it as documentation, and it provides an accurate representation of the analytical methodology that we always use. People use it as documentation.

1.5 WORKFLOW AND DOCUMENTATION INTEGRATION

The practice of literary programming is a method that aims to maintain the most recent version of automatic processes and documentation. Utilizing a programming language is yet another alternative that may be considered. Donald Knuth, a computer scientist at Stanford, is credited with being the first person to conceptualize literary programming, which is a methodology for the creation of software. This method of programming, on the other hand, was never widely adopted, in part because the overwhelming majority of programmers despise the requirement that they create documents. Literary programming is a strategy to computer programming that maintains that the documentation of a program (i.e., documentation of how the program functions, as well as documentation of how the program's algorithms and data structures work) should be written along with the program's source code (i.e. code), which actually implements the program. The term "two-way documentation model" refers to this strategy for approaching documentation in general.

The code itself serves as the primary document when using tools such as Javadoc and Roxygen; the documentation is made up of annotations that are appended to the code. The documentation for literary programming ends up being the primary writing that users consume, and the code itself ends up becoming an essential part of that documentation and is positioned in the locations in which it makes the most sense to have it. Because of this, the documentation becomes the primary writing that people peruse. Each time the application is launched, the website's source code is downloaded and separated in an automated fashion in order to produce the computer program's executable code.

Literacy programs have never enjoyed substantial success in terms of program development; however, the problem lies with data science. It makes sense to consider this document to be the primary outcome of the project because the final product of a data analysis project is typically a report that outlines the models that were used and the findings of the research. As a result, at this point we are concentrating on gathering the necessary paperwork. Text-based programming is permitted so long as the code can be analyzed and included in the documentation report. This is the one and only object that is necessary to successfully accomplish the assignment.

This functionality is supported in a variety of environments by a variety of programming languages. Mathematica is a derivative of the iPython Notebook, and it is possible to create notebooks in Mathematica that contain documents and visualizations intermingled with code that is functional. YR incorporates routines for automatic analysis as well as a number of different approaches to document generation, each of which is capable of acting as a foundation for subsequent reporting. Any one of these functions is suitable for using these materials. Knitr and R Markdown, two frequently used techniques, were used to produce these documents. These are the two most popular techniques. (for analyzing and generating reports).

The label that you provided to the R Markdown file will be given to the freshly produced HTML file when it is written to the disc after it has been created. When the file is committed, this moniker is also appended to the file as an identifier. The file that contains R Markdown markup will have the extension .rmd, while the file that contains HTML markup will have the same beginning but the extension.html. The extension of

both of these folders is the same. Select Knit HTML, and then click the menu symbol that appears next to it, to gain access to additional personalization choices. This opens up additional possibilities for you to consider. In earlier versions of R Studio, this was indicated by a down arrow in the upper-right corner.

You have the choice, when working in RStudio, to present the HTML page in the area to the right of the screen rather than in a distinct window. You have the choice to go in this direction. If you are working on a notebook that does not have a lot of screen space, you might find it more convenient to view the document in a frame instead of a distinct window because it will take up less of your available screen real estate. You have the capability of producing a file or document in Word format as well as an HTML website when you use this utility. RStudio remembers the output format you selected for a file and applies that format when you view the file. This is true even if you saved the file in a different format than the one you originally chose. Alterations are made to the heading information in addition to the conversion of HTML Knit to Knit or Knit Word. When you make adjustments to the description manually, the X-shaped button will automatically update to represent those changes.

If you move the tool symbol to the right one position further, you'll have some additional choices for how you want the document to be displayed once you make that selection. You have the ability to customize the way the document is presented through the use of these choices. Because everything is intermingled, it is highly unlikely that you will even be aware that the process of creating a document involves the use of two distinct instruments as well as three distinct languages. Nevertheless, it is essential because everything is dependent upon one another. There is evidence of the R code having been copied into the text. Following the processing of the R code, the knitr utility in R performs an evaluation of the outcomes of this procedure. After you have provided the program with your preferences, it will continue to process the output, which may include graphical elements as well as data.

Markdown was used as the formatting method for the finished product, which was a document. (registration without R). This template document is then processed by the pandoc tool, and the accountability for creating the result file lies with the pandoc tool. The information that is located within the document's heading is utilized in order to

accomplish this goal. Although this information is recorded in a programming language known as YAML, the actual formatting is done in a programming language known as Markdown. Metadata can be written in YAML, which is a markup language.

Because you have no influence over Pandoc, you should not be concerned about its effectiveness as a document processing service in general. Even if R Markdown is the only markup language you ever use to create documents, RStudio gives you the option to consolidate the pages you create into a variety of different output documents. Even if you only use R Markdown to generate documents, this statement is still accurate. Having said that, you have access to a very powerful document generation tool in the shape of a workflow that goes from R Markdown to knitr to Markdown, and then from Markdown to pandoc and a variety of output formats. You now have access to a very effective document production instrument as a result of this. This book is written in R

Markdown, and each chapter can be viewed as a separate document that can be shared using knitr without compromising the integrity of the other chapters. After that, I get the generated markdown pages by using pandoc with a few different configuration settings, include them, and then generate an output in addition to the epub output. There are many templates that can be used to customize the format with pandoc, and this may vary depending on the result document that is produced using the various templates that are accessible for the various formats. This is feasible due to the fact that Pandoc enables users to utilize numerous documents concurrently. As it is feasible to use different templates for each version within Pandoc, this is not only conceivable but also highly probable.

This work does not have sufficient space to investigate the elements of the instrument that contribute to its high performance or high clarity; therefore, we cannot do so. It is a very effective piece of equipment. Remember that you can use R Markdown to create more complicated documents than is readily feasible in RStudio, and bear in mind that this option is available to you if you're contemplating using it. In the preceding statement, I mentioned that a document composed in R Markdown utilizes not one, not two, but three different languages. Let's look at each one in the sequence that it was presented to us: the first is the heading language, which is called YAML; the second is

the text markdown language called Markdown; and the third is the method of inserting R into a document.

1.6 YAML LANGUAGE

When expressing data with key-value pairs, the language known as YAML is frequently utilized. Because this is a repetitive assertion, the abbreviation YAML stands for "YAML is not a markup language." YAML also stands for "YAML is not a markup language." To be honest, it is possible to argue that when I titled this segment "YAML Language," I shouldn't have included the word "language," but I went ahead and did so anyway. Within YAML, the character "L" stands in place of the word "Language." I believe that this was the finest attainable choice at the time. The word "markup language" typically refers to instructions that are used to markup text, specify formatting, or add structured information to text; however, since YAML does neither of these things, the abbreviation has been altered to reflect this difference.

It was formerly superseded by yet another script Language (YAML), but since the word "markup language" now more commonly applies to the instructions that are used to markup text, the term has been modified to mean "other markup languages," and YAML now stands for "other markup languages." It is possible to make the case that YAML is not really a markup language because its primary function is not to markup text but rather to supply various kinds of information to a computer program that processes a document. This is the major reason why YAML was developed. In point of fact, the primary objective of YAML is not to markup text but rather to provide a variety of formats for input.

There is typically no need to directly modify this description in most circumstances. When you make a change to the settings using the graphical user interface, the label will automatically be adapted to reflect your selections. However, it is outdated and does not include bibliographies, which are going to be discussed in the following segment. If necessary, you can add anything you want to the heading, and doing so through this method, rather than through the interactive user interface, might be simpler. (GUI). On the other hand, you shouldn't be switching the label around too frequently. You will have access to the capacity to generate key-value combinations by

making use of YAML, which will be made accessible to you. You should begin by pressing the: key, and then input the number right away.

Consequently, "My Markdown Document" is the key title, and "Thomas Mailund" is the key creator in the example that was just presented. You should never quote anything, and even if the values don't contain any colons, you should always be able to do so if you want to. There is never an occasion when you have to designate a number. Both the RStudio and Pandoc procedures, when carrying out their individual operations, make use of the YAML heading. RStudio allows the user to specify the output format into which the document will be transformed by clicking the escape button. The Knit taskbar button then represents the choice that was made by clicking the escape button. When adding information to a document that Pandoc has developed or produced, it takes into consideration the document's title, creator, and timestamp in addition to the information that is being added.

1.7 BRAND LANGUAGE

A wordplay on the word "markup," which is precisely what markup languages do, "markup" is essentially a spin on that word. Markup languages mark things up. When it was first developed, its primary goal was to make the process of people producing online sites as straightforward and easy as possible. In addition to its use as a coding language, HTML is also a language that is employed in the production of online sites. Having said that, it is not always simple for individuals to comprehend. Markdown was developed as a solution to this issue by first structuring the text based on some very straightforward formatting principles that are utilized in emails when they are written in HTML, and then developing tools that can transform Markdown to HTML. This allowed for the problem to be resolved. In order to accomplish this, the text was initially formatted according to the formatting principles that were utilized in emails when those emails were composed in HTML.

Markdown is a text-writing language that lets you create text that is understandable by humans by employing markdown expressions. This text can then be transformed into a number of different document formats. Even though it's a language that's used for a lot more than just making webpages these days, Markdown is still a fairly straightforward

and easy-to-understand format. When you write in Markdown, you are essentially writing in normal text but giving it the appearance of being written in Markdown. As a result, the vast majority of the material is composed of simple text and does not include any coding. It is necessary to use an appropriate text editor in order to guarantee that the finished product will consist entirely of text.

Because the file structure of a word processor already includes a substantial amount of formatting information that is difficult to see on a computer screen, you are unable to compose with a word processor. Text processors are something you should already be acquainted with if you have even the remotest interest in learning how to code in the future. If that isn't an option, you can always use RStudio to generate R Markdown files, and once those files are generated, everything will function normally. When you need to construct something more complicated than simple text, you can use markup sentences instead of just plain old text. Because there aren't that many different instructions to remember, learning how to use them doesn't take very long. As you enter, the instructions are organized to help you concentrate on the text rather than the formatting, which means that learning them won't take you very long.

There are instances when a specific component of the research requires a considerable amount of time. This also pertains to the time you spend thinking; however, you do not need to keep going over the same problems in your head. In this particular setting, I'm referring to processing time rather than thinking time; however, the same is true of CPU time and thought time. On the other hand, if you are careless, you might end up having to repeat the same check on your computer more than once. If you are conducting an analysis that includes stages that take a significant amount of time, the process of creating documents will move at an intolerably sluggish pace. When you make a new edition of the document, you must begin the research from the very beginning. Something that you have to do each and every time. You want this feature because it eliminates the possibility that the research will produce content that is not related to documentation. This is the ability you need to work on.

The effectiveness of the procedures, on the other hand, is significantly diminished if the creation of them requires several hours. The outcomes of the extraction can be saved to a memory, and doing so would be one strategy for addressing this problem. If you

want the results of a fragment to be stored in a cache, you need to make sure that the fragment contains the cache=TRUE option. As a result, you are required to include this information in the excerpt heading, just as you would when examining output choices. You are going to need to rename the music before you can make use of it. If you try to recollect the findings by using the name, you won't be able to do so because the number of unidentified sections discovered previously in the document determines whether or not you will be able to retrieve the information. This name is given to components that do not in fact have a name; however, the identification of this name can change contingent on the number of unidentified components that were discovered in previous sections of the document.

Because of this, you ought to come up with a moniker for him. When you generate a document, any named fragments that have been designated for storage are stored in the cache. This occurs regardless of whether or not these fragments have been modified since they were last evaluated. This occurs regardless of whether or not you were the one who originally produced the document. If nothing is altered, it will use the findings of the most recent assessment that was stored. If there has been an update, a fresh collection of results will be generated. Because R is unable to store everything, putting libraries into a cached widget will not cause those libraries to be loaded until after the widget itself has been evaluated. This is because R cannot cache everything. The reason for this is that R cannot store everything.

This demonstrates that there are restrictions on the kinds of things that you are permitted to do. In general, however, it's a fairly useful piece of equipment. It is possible for other fragments to depend on a fragment that has been cached, and it is also possible for a cached fragment to depend on another fragment regardless of whether or not the latter fragment has been cached. Both possibilities can be realized at the same time. Because it won't be reevaluated until you update the code that it includes, if the code fragment is based on something you've altered, it will save findings that are based on data that is no longer accurate. This is because the code won't be updated. This implies that it will now save findings based on inaccurate data if the excerpt is based on something you changed in the past. In this respect, you need to give it your complete and undivided concentration.

On the other hand, you can find a solution to this issue by establishing relationships between the various components. You have the ability to indicate that the outcomes of one part depend on the outcomes of another part by making use of the "dependon=other" component choice. This indicates to the system that the relationship between the two currencies is one of interdependence. In the event that the component upon which the cache is built undergoes any kind of modification, the cache is considered incorrect and a new evaluation is performed. You are the one who is tasked with figuring out which blocks are dependent on which other blocks.

1.8 DATA DISPLAY

When composing a summary on the results of a data analysis, it is only natural to want to include certain discoveries. If you have reached this point, it means that you have already begun the procedure. It is necessary to display the information in some fashion for it to be considered comprehensive. You simply need to add the R expressions that you want to evaluate to the excerpt in order for it to display the outcomes of the evaluation. However, you will frequently find yourself desiring to visualize your data through the use of charts or other graphics. This is of the utmost significance in the event that the report is going to be delivered to individuals who are not acquainted with the R programming language. Tables and pictures saved in the appropriate formats can, thankfully, be transferred with relative ease. If you use the kable() technique that is included in the knitr utility, you will be able to successfully construct a database.

You could attempt including a portion that is analogous to this one in a standard document that is already in existence. Already acknowledged is the significance of transforming the power of big data into an efficient business strategy that is able to both maintain and expand business operations. It has been determined that this is one of the most essential things that must be done. However, reaching an agreement on how and what to implement is challenging, and the suggestion is merely one of many potential solutions to the problem that has been presented. In accordance with the suggestions made by Doornik and Hendry (2015), we have discovered that a streamlined strategy to the resolution of data problems is not only helpful, but above all else efficient. This finding is in step with their recommendations.

In point of fact, time, effort, and resources affiliated with data acquisition and processing are decreased, as are the effects of technological advancements and ex post observations. The fact that it gives people the ability to prevent two situations that couldn't be more different is the most important aspect of the structure. These findings are the aggregate of all of the data or none of the data at all. demonstrates the primary actions that need to be taken in order to accomplish this minimal strategy to big data. These stages include, first and foremost, the identification of the business processes that will be used in the project, and second, the formulation of the analysis framework that will be applied to the work. These two successive stages, in addition to outlining the analytical framework and generating the dataset, include data at rest, which is data that is stored stationary and inactively in a database; data in motion, which is data that is continuously stored in impermanent storage; and data that is being utilized. (which is constantly updated and stored in the database).

The modelling portion is a very essential part of the procedure, and it also contains the evaluation step. When the scaling and measurement components are brought into play, the procedure will have reached its conclusion. A feedback system should prevent internal deadlocks by presenting the findings of the analysis to the business process. This should be done rather than continually constructing the model without receiving any feedback from the business. This is done to ensure that the system does not become corrupted. This is due to the fact that the feedback mechanism needs to prevent an internal delay from taking place. It is necessary to establish common information standards; it is necessary to create a master copy; and ultimately, all information must be verified to determine whether it is accurate and complete. Data should be collected consistently from a variety of information sources and integrated with other systems and platforms.

In addition to conducting an analysis of the necessary characteristics and abilities, one of the other essential steps involves establishing efficient data value networks and putting appropriate procedures into place. These are two of the most important steps. In the end, conduct an assessment of the knowledge and character traits required to evaluate the data. The issue of large stratified intake, in which datasets fail to function as intended due to the fact that different individuals are added to the dataset at different

periods or in different schedules, can be remedied by utilizing a data management system, which provides dependable internal data that is consistent with the golden record. The issue can be resolved in a more efficient manner by making use of these meticulously crafted combinations of available tools.

Even when utilizing a strategy that is guided by data, organizations can still run into a substantial number of roadblocks. Consequently, the development of a structure that can monitor internal progress and obstacles while also providing an outline of the subsequent stages in the analytics journey constitutes an essential first step. In order to fulfil this requirement, a development model that is commonly referred to as the Data Stage of the Development Structure (DS2) was developed particularly. It is a road plan with the purpose of putting into action a data strategy that will generate revenue and have an effect. It is possible to use it to investigate the current status of an organization and gain knowledge regarding the future steps that need to be taken in order to construct big data capabilities within the organization. With this instrument, you can do either one. A matrix that is four rows thick and four columns broad can be seen in Table 2.1. On the top line, the bottom-up phases of development are labelled Primitive, Custom, and Factory.

On the bottom line, the measures that are used to evaluate these stages are labelled Culture, Data, and Technology. o The concluding evaluation is on the very last line, and it discusses the financial effect that the business will experience as a result of having an established data strategy. This line is drawn last because it is the line that contains the final assessment. Therefore, it is drawn last. The first stage, which is also referred to as "awareness," is to realize that data science can have an effect of some kind on the activities of a business. This marks the beginning of the procedure that will be followed. At this stage, none of the previously established administration frameworks or technological advancements have been implemented, and more significantly, there is no agreement anywhere within the organization. However, the outcome of people's excitement for knowledge is always something that can be turned into something that can be acted upon and translated into real action. This result is always something that can be turned into something that can be acted upon and translated into real action.

The currently available features are still in the developmental stages, and the actual data utilization isn't nearly as intricate as it might be. The activities of the organization are not significantly altered as a result of the collection of data because the sole purpose of such collection is to supply management with pertinent information. It is not necessary for the project to be unsuccessful simply because it has reached this level; rather, it merely signifies that the performance and outcomes of the project are extremely variable, unreliable, and not sustainable. The second phase is commonly known as the support phase, despite the fact that this stage is essentially an experimental stage. This is because the word of the stage suggests assistance.

The trial project was successful in demonstrating the value of big data; however, new competencies, technologies, and infrastructure are required. Moreover, a new data administration is required to oversee not only the possibility of data contamination, but also the numerous players who are participating in different phases of the process of data analysis. In point of fact, different players join the process at various points at various periods. The level of managerial dedication is presently extremely low, and as a result, the pool of prospective applicants is restricted to a single particular department or position. Even though they were more sophisticated than the procedures used in the first phase, the processes that were used were still extremely specialized, and they could not be replicated.

This was the situation despite the fact that they had been utilized in the first place. The third phase is responsible for driving a process that is more standardized, streamlined, and reproducible. At this stage, access to the data had been significantly widened, tools had come to the forefront, and a viable recruitment strategy had been developed to combine talent and resources. Because of the high level of dedication displayed by the management team, the initiatives have been able to guarantee continuous financing distributions, which would not have been feasible without the efforts of the management team. The transition into the new work responsibilities is addressed in the fourth stage.

At this stage, each and every function is now data-driven, the transformation is agile-driven (that is, value is delivered progressively, rather than at the end-of-life), and complete management support results in a collection of connected actions. At this

stage, each and every one of these components is accessible. These may include the establishment of a center of excellence (i.e., a project), as well as the optimization of state-of-the-art architectural and technological infrastructure; all of this has a significant influence on the flow of your revenue. Data lakes, also known as repositories that store data in its original format, offer an alternative that is both international and cost-effective in comparison to more expensive preservation choices. Additionally, it supports a wide variety of programming languages, is highly expandable, and is maintained in a central storage location. These characteristics make it possible for the organization to migrate between different platforms at no additional expense. recognizes.

They also ensure a reduced possibility of suffering a loss. However, in order to guarantee the quality of the research and the safety of the data, they need an information management system that can place the data into perspective. In addition, they need to establish stringent standards. The information ought to be retained safely, analyzed utilizing the procedures that are the most applicable, and guarded against access by unauthorized parties. The establishment and upkeep of an information lifecycle is essential, and when doing so, special attention should be paid to the preservation of data in a prompt and effective manner, as well as the maintenance of data retention and the validation of data for the production environment. The "culture" element, which encompasses a number of different phases, ought to also be taken into consideration.

The theory developed by Davenport states that each stage requires a different kind of research. (2015). There are four distinct sub-fields that fall under the umbrella of analytics: descriptive analytics, predictive analytics, normative analytics, and automatic analytics. The field of analytics can be broken down into three distinct subfields: descriptive analytics, predictive analytics, and prescriptive analytics. Analytics can be broken down into three categories: descriptive analytics, predictive analytics, and prescriptive analytics. Descriptive analytics focuses on what is occurring, predictive analytics considers potential future outcomes (sometimes supplemented by diagnostic analyses that look at contributory factors to a specific occurrence), and prescriptive analytics makes recommendations. The final category of analysis is known as automatic analysis, and it entails taking some form of action in response to the

findings of a study. A summary of some of the findings mentioned in the conversation up to this point can be found in Figure 1.1.

The diagram that follows illustrates the connection between managerial support for the analytics function and the degree of complexity and expertise that is necessary to be successful in data-driven organizations. The degree of managerial dedication is represented by the horizontal line (high versus low), and the possibility of the project that is being undertaken is evaluated along the vertical axis. Within the scope of this discussion, the term "feasibility" alludes to the extent to which the necessary abilities to carry out the project are readily accessible, in relation to the degree of difficulty posed by the task at hand.

The matrix is broken up into four portions, each of which corresponds to a unique form of analytics and is arranged in accordance with how well they intersect with big data analytics practicality and management participation. These parts have been given the letters A, B, C, and D, in that order, correspondingly. There are four distinct levels, and each circular symbolizes one of them. (numbered sequentially from "Primitive" to "Scientific"). The size of the effect that each component has on the organization is represented by the circumference of the circular. (i.e. the larger the circle, the higher the return on investment). The second horizontal line is in charge of monitoring the increased quantity in order to ensure that the overall performance does not improve.

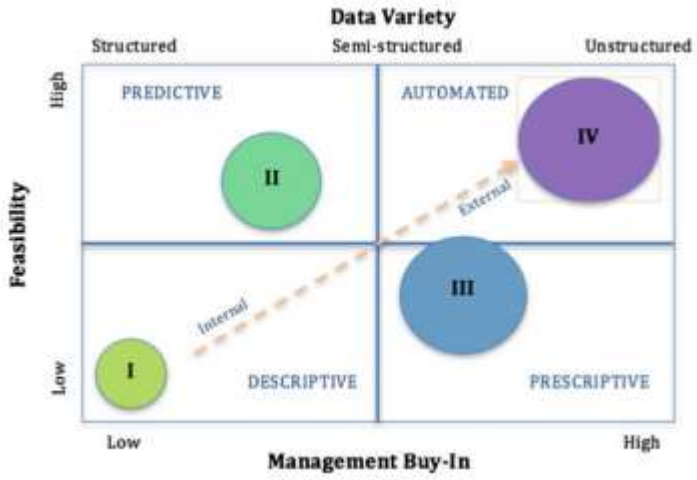


Figure. 1.1 Big Data Maturity Map

Every one of the processes makes use of some form of data, which, depending on the organization of the information, may be structured, semi-structured, or uncontrolled. (e.g. Internet of Things, sensors, etc.). An orange diagonal represents the various kinds of data that are utilized, beginning in the lower left quadrant with closed systems and internal private networks and continuing to the upper right with market, public, and exterior data. Each of these categories of data is represented by the diagonal. After investigating the different choices and indicators (for more information on the framework, see the complete specifics), these can be put to use to establish which stage a business is presently operating in. Please refer to Framework if you are interested in learning more about the framework. or read on for further information.

It is also beneficial to have an understanding of how a business can progress from one stage to the next, and the figure that follows provides particular recommended methods to facilitate that progression. The term "nursing" refers to this particular motion. Experiments should only be carried out by one individual at a time in order to ensure a seamless transformation from the stage of rudimentary development to the stage of tailored development. Their objective is to create proofs of concept or pilot projects in order to answer a very straightforward question utilizing the information that is already available. It is more likely that these proposals will be noticed right away if they will have an effect that is beneficial or favorable on the organization.

You should make an effort to keep the financial costs as low as possible because the project already requires a significant amount of physical effort (through cloud computing, open source software, etc.). As a result, you should attempt to keep the financial costs as low as possible. etc.).). The problem of redundant data needs to be resolved, and the responsibility for doing so lies with the entire business. In order to make the transition from "Bespoke" to "Factory," you will not only need strong backing from project management, but you will also need to implement standard operating procedures and metal records.

During the development of technology, instruments, and infrastructure, it is necessary to engage in experimentation as well as critical thinking in order to guarantee that these things will be useful in the future. As a result, the vision ought to be for the not too far away or not too immediate future. It is essential to make an attempt to boost the

allegiance rate at the topmost level of management, as this will have a significant impact on the organization. At a higher level, this necessitates the completion of data voids, the promotion of new data sources and kinds, and the development of a strategy for platform adaptability.

You are obligated to finish each of these tasks. Specifically, you are tasked with defining the acceptable level of data loss, which is referred to as the recovery point target, as well as the amount of time it takes for unsuccessful discs to recover from their condition, which is also referred to as the recovery point target. re-creation). It is imperative to pay attention to the slightest details in order to progress technologically and become an expert and pioneer in the field of data. Additionally, it is important to optimize models and databases, improve process data collection, improve data quality and accessibility, and find a Blue Ocean strategy that you can follow. Protecting the privacy of users and ensuring the safety of their data is of the uttermost importance, and stockholders have the right to more openness regarding the company's data processing strategy.

Establishing a center of excellence (CoE) and the talent procurement value chain are both essential stages in the process of delegating the responsibility of operating the business to the data science team, which is the eventual objective of this process. However, in order for the Council of Europe to reach its goal of sustainability, it will need to undergo a process of reunification in the not too distant future. despite the fact that the CoE is essential to assisting managers in accomplishing their near-term performance objectives. You can now begin recording and monitoring adjustments, as well as beginning and continuing to register and track returns on investment. (ROI). It is essential to maintain your position as a leader in the data community and gain the respect of other members of the community by actively engaging in the continuous process of learning new things and experimenting at the top.

Figure 1.1 also provides a suggested timeline for each phase, including the following: up to six months to evaluate the current situation, conduct research, and initiate a pilot project; up to one year to leverage a particular project to understand the skills gap, justify higher budget allocations, and plan to expand the team; two to four years to ensure full support of all functions and levels within the organization; and at least five

years to finally reach a futuristic inventory. Figure 1.1. Suggested timelines for each phase. It goes without saying that the demands placed on your time by various organizations will vary, and as a result, you will need to maintain a high degree of adaptability. It is of the utmost importance to elucidate on the organizational foundation of data analysis in a few more sentences. (Pearson and Wegener 2013). Our position is that the Center of Excellence is the most sophisticated enterprise model that is presently accessible, and that it is the model that should be used for integrating and controlling the data activities of an organization.

In point of fact, Microsoft was the company that created the Center of Excellence. The primary responsibility of this organization is to serve as a central point of coordination for the various operations carried out by the various divisions. This includes the upkeep and enhancement of the technological infrastructure, the determination of which data will be collected by which service, assistance with the recruitment of talented individuals, the planning of the knowledge production phase, and the declaration of privacy, compliance, and ethics policies. But there are also other formats, and it is essential to be familiar with these other formats because sometimes they suit the present organizational paradigm better than the format that is being used.

It is essential to have a working knowledge of these documents. It provides a wide variety of distinct data analysis and business model configurations that can be combined in a variety of ways. It can range from business units (BUs) that are completely autonomous to business units that are autonomous but collaborate on certain initiatives to an internal (Enterprise) or external (Center of Excellence) center that coordinates numerous activities.

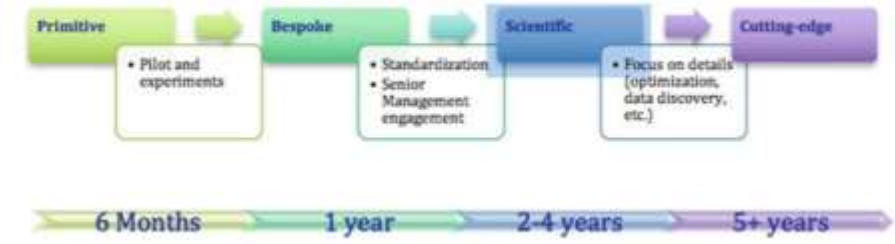


Fig.1.2 Maturity level transitions

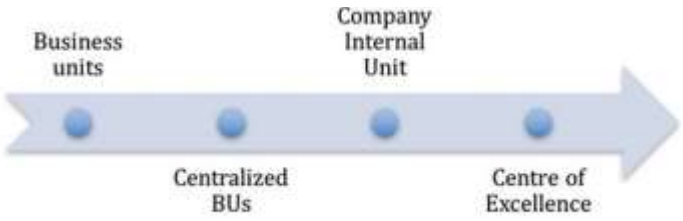


Figure. 1.3 Organizational models for data analysis

However, depending on the characteristics of the organization, each of the evaluations conducted up to this point symbolizes a different set of ideas and provides a unique perspective on the situation. To be more specific, the particular phase of the business lifecycle that the company is operating in at the present time has a substantial influence on the type of strategy that it follows, whereas the degree of data development of the organization is immaterial. (For instance, a business that has only been around for a couple of months might be considered a science-based company, whereas a major investment bank might be considered nothing more than a rudimentary company.)

1.9 VERSION CONTROL AND POOLS

When developing new software, it is strongly recommended that developers make use of a version control system for a number of factors, the most important of which are safety and productivity. Your first order of business should be to record and monitor all of the modifications that are made to the source code. You will then be able to determine when and what changes were made in the future, and if you discover that those changes were incorrect, you can use the register to return to an earlier edition and attempt something else in the event that you discover that the changes were incorrect. In this manner, you will be able to determine at a later time when and what adjustments were made. It keeps a record of how far your software has progressed, allowing you to restart from an earlier point in time if you discover that your choices are leading you in a direction that you do not want to go. In point of fact, it generates a document for the software development you are performing.

The facilitation of communication with other individuals is typically the second goal that a version control system is designed to achieve. Each developer works on their own duplicate of the code as changes are made to the code; when they are finished

making changes, they submit the modified code to the central repository. The idea behind this is that all of the source code and any modifications to the source code are stored in a centralized location referred to as a repository, and that this repository is made accessible to all developers. When utilizing earlier versions of version control systems, it was essential to secure the files in order to make modifications to them before releasing them. Only after making these changes could the files be unlocked. This was done to eliminate any potential issues that could arise from having multiple developers working on the same files at the same time as you. When it comes to changing the same files at the same time, modern version control systems are more forgiving than they were in the past.

In point of fact, it combines the revisions as a matter of course so long as there are no alterations to the entries that intersect. This is the case regardless of whether or not there have been any alterations made to the sections that are in disagreement. (in which case you must resolve the conflicts manually). A large number of programmers who use this kind of version control have the ability to work on multiple portions of code at the same time without having to think about any disagreements between them. If you make an attempt to send modifications to the public repository, it will identify any conflicts that are already present and tell you that you need to address them before you can continue. Due to the fact that it does not even have a singular worldwide repository in the traditional sense, the distributed version control system known as git makes it possible for even more continuous and independent work to be done. At the very least, it lends credence to this hypothesis.

It would be beneficial to have a worldwide repository that contains the version of your software that has been officially approved, and many people already have this. In point of fact, maintaining such a collection is a practical notion. The construction of the system is predicated on the idea of numerous groups, all of which are in continuous communication with one another and are able to share information about the modifications that have been made. This idea is extremely important to the overall structure of the system. When you work on some source code using Git, a private repository of that code will be generated for you automatically as you go along. You can use this repository to develop branches for particular features or versions, as we

will see in the following section; you can also use this repository to document changes, as we will see in the following section.

As you would ordinarily do, you make modifications to your source code, and when you are finished, you have the option of pushing those modifications to your local repository without interacting with the modifications that other people are making to your source code. Because you and they use different local repositories, neither of you can see the changes made by the other, nor can they examine any customizations made in the repository that they use. For your modifications to be preserved in the other repository, you will need to "push" them there from the original repository. If, on the other hand, you wish to retrieve modifications that were made in a different repository, you will need to retrieve it from this location. At this point in the process, it is standard practice to store information in a centralized collection that can be accessed by all members of the organization.

You will submit the changes to your private repository while you are developing a new feature; however, as soon as you finish developing the feature, those changes will be pushed to the public repository. You can ask someone who does have permission to make changes to the public repository to submit the changes on your behalf if you do not have permission to make changes to the repository yourself (perhaps because you copied and modified the code of another user). You also have the option of asking somebody else who does have authorization to make modifications to the public repository if you do not have it yourself. If you do not have permission to make changes to the public repository, you will need to inquire of another individual who does have permission to examine changes made to the repository. This action may also be referred to as a "pull request" in certain settings.

1.10 USING GIT IN RSTUDIO

This is all very speculative, and if it's difficult for me to explain it, then it's probably just as difficult for you to grasp it. Instead, we put Git through its paces by evaluating it in a simulated environment. You are able to set up repositories, send changes to those repositories, and submit changes to other repositories with the help of the fundamental tools for dealing with Git that are supplied in RStudio. All of this is possible without

ever having to leave RStudio. You will need other tools or use the command line version of Git for this, but for everyday version control, this is sufficient for most activities. It does not support all of the things that you can do with Git.

Everything that can be accomplished with Git also necessitates the use of other tools. To accomplish anything with git, you will need to make use of a variety of other programs. You will need to do some searching in order to acquire the knowledge necessary to implement Git in the most effective manner possible on a variety of operating systems. The next stage, after downloading Git on your computer without any problems, is to become familiar with how to make use of it. This needs to be done so that any modifications that are made to your code on your behalf can be attributed to you. You do not need to figure out who made the modifications if you are the only person working on the code at the moment. This is because you are the only person working on the code at the moment. Nevertheless, when multiple people labor together to develop software, it is essential to identify who is accountable for which modifications.

In order to provide Git with information about yourself, type the following instructions into a Terminal1 and execute them: You may be requested for the route to the Git instruction if you have RStudio installed on your computer, and you can provide it if you are. To accomplish this, open the drop-down option in the top right portion of the screen and select Tools > General Settings. A section labelled Git/SVN ought to materialize all of a sudden on the left-hand side of the window. It ought to be easy for you to discover it. You have the choice to inform RStudio in this window where the Git program you want to use is located. RStudio will then use that command. Git is a piece of software that can be installed on your computer and used in the command prompt environment.

You are still able to use RStudio despite the fact that the graphical user interface (GUI) for dealing with Git has some restrictions placed upon its capabilities. You can accomplish a great deal more with Git than you can with RStudio on its own, so it is strongly suggested that you acquire one of the graphical user interface programs. There aren't that many options available for you to pick from among these programs. I

discovered that it was much simpler to use them than it was for me to enter the command lines myself as I grew older and started to forget some of the instructions.

When you use this window for the first time, a yellow question mark will appear next to the condition of every file that you have modified since the object was first generated. This occurs independently of who generated the file, meaning that it could have been you or someone else. (including files created by RStudio during package creation). This indicates that Git is not aware of the existence of these folders at the present time and does not have any information of them. Even though he is conscious of the existence of the files, he has not been given any instructions regarding what he should do with the information that is contained within the files. You are obligated to respond instantly. In the Prepared section, there is a selection designated specifically for each of the folders. When you select one of them, a vibrant green "A" will appear next to the condition of the file you selected, indicating that it has been successfully opened.

This indicates that the file you have prepared is available to be published to the Git repository as soon as you are ready to publish it. When you are ready to publish it, you can publish it. Carry out the procedure described above for every one of the folders. When you carry out this operation in a directory, all of the files contained within that directory will also be added to the collection if you choose to carry out this operation in the directory. So that's the kind of thing we want to watch for. Before you can actually submit the changes, you want to Git, the Git commit procedure requires you to prepare the changes you want to commit. This occurs prior to the modifications being submitted into Git itself. What we have just done is instructed git that these files must be included in the revision the next time you make a modification.

This is the event that took place just a moment ago. If we are determined, the adjustments that are put into action will be constrained to a significant extent by what has been done in the past. As a consequence of this, you are only able to analyze the source code with a portion of the modifications that were made, which can be helpful in some circumstances. You might have made a lot of modifications to a lot of files, but at some point, in the future, you might decide to only submit one bug repair for a new feature that you don't want, rather than a bug fix. I have not yet completed the construction of the product. This can also occur if you have previously made a

significant number of modifications to a large number of folders. You can accomplish this by only releasing the modifications that you want to submit to the system. In any event, we have modified everything, and all that is required of you to validate the changes is to select the "Confirm" option that is located in the taskbar. That is the one and only option available.

This will bring up a new window displaying the modifications you intend to post, as well as giving you the option to compose a message committing to posting those modifications. (top right). This communication has been saved and indexed as a component of the comprehensive change history. Please provide a concise overview of the modifications that were made to this portion, focusing on the most important aspects. You'll need it in the event that you determine at a later time that you need to search for modifications to the register. After that, choose the Confirm option, and when you're finished, dismiss the window. At this stage, the Git window ought to be totally devoid of any content. This is due to the fact that the file has not undergone any further modifications since the previous commit, and the interface only displays files that have been modified between the version of your program that is presently installed and the version that has been stored in Git. Because of this, you are observing this pattern of behavior.

1.11 ADD GO TO AN EXISTING PROJECT

Since you are just starting out with Git, it is likely that you already have a number of projects that do not have Git integrated into them—that is, unless you always select the Git option when establishing new projects. This is because you are only beginning to work with Git at this point. You can install Git in a subfolder that already exists even if you did not initially configure your project to be built with a connected Git repository when you initially set it up. This is the situation regardless of whether or not you constructed your project with a corresponding git repository in mind. When the following dialogue box displays, choose Configure Build Tools from the Build menu, and then choose Git/SVN from the selection menu. After you make the decision to use Git as your version control system, RStudio will automatically configure it for you once you have finished making your selection.

1.12 BARE POOLS AND CLONING POOLS

If you are hosting your repository on a separate website, such as GitHub, the vast majority of the information presented in this portion will be irrelevant to you. There, the procedure of generating a repository as well as the process of communicating with it once it has been established are both carried out via an online interface. It is only necessary for you to be concerned with "cloning" a repository; other than that, you do not need to be concerned at all with the technicalities of the procedure. In this part of the tutorial, we'll physically establish a repository that we'll name "bare," and then we'll investigate how we can use that repository to send changes to other nearby repositories. When R projects are initiated or git init is performed in a directory, the source code that is located in the project directory is saved for the purposes of version control. This saves the code from repositories that are created at the same time. Because of the manner in which they are constructed, working with them is not particularly straightforward for developers.

Even though it is possible, in theory, to combine changes made in one repository with changes made in another repository, in practice, doing so is laborious and not something you should have to deal with very frequently. In order for us to be able to synchronize the modifications that have been made in various repositories, we need a main repository. Because of this, we are able to handle our info in a more effective manner. Because there is no original source code in this repository, you will not be able to make any modifications to the material on your local machine. The presence of this feature prohibits you from making any local modifications, despite the fact that it is not a particularly distinctive one. You are only able to update using modifications that have been brought in from other sources. Because of this, the range of potential changes is reduced.

You have now downloaded the empty repository twice, so you have two versions of it altogether. You will learn how to monitor a clone for updates from a copied repository and how to send changes from a clone to a cloned repository. Both of these tasks will be covered in this lesson. In this class, we will go over both of those activities. Although it has been stated that merely navigating a repository is not the only way to send changes from one repository to another, it is the simplest way to work with Git, and

this is the strategy you should take when utilizing a website such as GitHub. If you follow the instructions on how to complete this task on GitHub that are provided in the following section, the website will automatically install the basic repository for you. Simply create a duplicate of it and save it to a different location on your computer so you can work with it. You are not required to send modifications to the public repository (bare) after each commit; in fact, it is completely conceivable that this is not something that you would prefer to do.

In order to achieve finer-grained version control, you should submit your native code on a regular basis. However, you shouldn't implement these adjustments until after you've finished a feature or at least brought it to a point where it's logical for other people to work on your code. Until then, you shouldn't even consider doing so. It is not a huge problem if you submit code that is completely non-functional for your private repository; however, if you share faulty code to others, it will not be well received. Although this is not an especially significant issue, you should still make every effort to prevent encountering it. The customizations are incorporated into the first copied repository; however, the second repository does not display any indication that they were completed. You are required to download them as a necessary stage in this process.

It initiates a transaction that retrieves the most recent updates and then incorporates those changes into the local repository after the completion of the transaction. After that, finish following the steps in the procedure. This is primarily what will occur if you incorporate the changes you've made with the changes made by others and publish the combined changes to the central repository. Execute the instruction by navigating to the copied repository, within which no modifications have been made. Check to see if the adjustments you made are reflected there. When collaborating with other people on a project, the accepted procedure is to make adjustments first, and only after making those changes should they be saved to the repository. Because you are not yet prepared to share the modifications that you make while using this repository, only you will be able to view them. After that, you can upload the changes to the shared repository when you are ultimately ready to discuss them with you and the other people working on the project.

You have complete discretion over whether or not you want to implement the adjustments suggested by other people. You will receive an error notification if you attempt to send to the public repository after another user has already made revisions but the public repository has not yet verified out the changes. There is absolutely no cause for concern on your part. Simply reverse the adjustments, and then you will be able to submit your own alterations. The process of committing changes to your repository while those changes also include changes that have already been committed but have not yet been committed is referred to as a combine, and in order to finish the process, you will need to create a commit statement. In point of fact, the combine necessitates modifications that have previously been confirmed but have not yet been presented. We strongly suggest that you use the default greeting that has been supplied for this purpose.

Because you have access to two repositories for the purpose of testing, you should experiment with different push and pull configurations, and you should also send the changes to a repository where you've already uploaded the updates. It is possible that the following description will make more sense to you if you have some of your own experiences to draw from. RStudio offers fundamental assistance for the transmission and retrieval of data in a variety of formats. You will be able to attempt starting a new RStudio project that will be placed in your copied repository if you start a new project in RStudio and place it in an existing subfolder first. This will give you access to the choice to do so. When initiating a brand new project in RStudio, you will always have access to this choice. If you follow these instructions, you will observe that the Go panel has been updated with two new buttons labelled Push and Pull. These buttons have been added because the panel was in need of an update. As part of the change, these options were introduced to the page.

1.13 WORKING WITH BRANCHES

The great majority of version control systems have a function known as branching. This function enables users to concurrently work on various versions of code that they have produced, which is a very useful feature. A standard illustration of this would be to divide your team into two distinct branches: one branch would be in charge of creating new features, while the other branch would be in charge of maintaining a stable version

of your product. Your code changes as you attempt to build new features, the implementation of the new feature might be difficult, and the user interface may alter simultaneously from one layout to another. You shouldn't let your consumers use such a version of your product, at least not until you give them a warning that the package they are using is unstable and the user interface may be different. You are aware that it is vital to keep the product code distinct from the common code, right? It wouldn't matter much to you if you merely published new versions of the program on a regular basis and added new features gradually in the time in between new software releases.

Users should not utilize commits that are between published versions; rather, they should use the version that you published as their main source and not use commits between published versions. On the other hand, if you're constructing a build that has an issue that isn't impossible to attain and you want to correct it when it's identified, the world isn't as straightforward as you would believe it to be. Maybe. Troubleshooting is something that you should not put off until the process of integrating any new features that you are presently working on has been completed. This is something that comes highly recommended by us. You made the decision to make some changes to your code just before it was made available to the general public. If there are more defects found in the release code than were anticipated, more bug patches will be added.

In addition to this, you are simultaneously modifying the code that is being used for the creation of the product. It should go without saying that any modifications you make to the development code in order to correct errors in the version that was published will, of course, be included into this code as well. After all, you don't want the next release to have defects that your team has already resolved, even if it is exactly what you want to happen and what you don't want to happen. Here is when the importance of branches becomes apparent. The support that RStudio provides for branches is quite restricted, and the application does not even provide assistance when it comes to the creation of new branches. In order to complete this task, you will need to utilize the command line. In order to create a branch, you will need to run the command git branch name.

If you intended to establish a development branch called development but didn't want to come up with a better name, you should use this command. You will see that the Pull

and Push buttons are disabled and will not display if you move to the development branch. mainly due to the fact that they are no longer considered experts in the relevant functions. You may make modifications to your code and apply those changes while working on a specific branch; but cannot (yet) push or pull during this time. In a moment, we'll go into more detail regarding that topic. After creating and applying certain code changes in the development branch, you will notice that these changes are absent in the master branch when you switch to the master branch. In point of fact, the master branch always contains the most up-to-date version of the code. You can determine what took place by reviewing the Git history as well as the files on their own. (using the Git log or clicking the History button in the Git dashboard).

No matter where the changes are made, they will not be reflected on the development branch as long as they are made in the master branch. This is the comprehensive response to your query. When it comes to transitioning from working on the development version of your product to working on the final version of your product, you are in no way restricted in any manner; all you need to do is change branches. Both of these groups exist in total and utter isolation from one another. You will need to combine the two branches in order to incorporate the modifications that were made in one branch into another branch after those changes were made in the first branch. The process is not really symmetrical; rather, one of the branches must be dependent on the other. You should begin by verifying the branch you want to make changes to, and then use this command to combine the modifications made in one branch with those made in the current branch. In order to do this, you will first need to validate the branch you want to modify.

1.14 TYPICAL WORKFLOW CONTAINS MANY BRANCHES

Git was developed to make it easier for people to work together on projects by providing a number of different branching possibilities. (unlike some version control systems where branching and merging can be quite slow). When working with Git, this idea is mirrored in the fact that many users choose to work with branches. Make a number of separate branches of your code, work on a diagram that shows the differences between them, then merge the branches together as necessary. The structure of a master repository will often include both a development branch and a master

branch. This is a typical practice. On the other hand, developing a new feature always results in the creation of a new branch, which is an extremely frequent approach.

Before beginning work on a new feature, it is standard procedure to first establish a new branch that is distinct from the development branch. After the development of the feature is finished, the branch should be promoted to the development branch. It is also usual practice to have a distinct department dedicated to the troubleshooting of each different kind of issue. When the patching process begins, this branch will often split off from the master branch. When the patch deployment is finished, this branch will be combined back into both the master branch and the development branch. Explore the Git lesson that has been made available by Atlassian. If you create many branches for each new feature or bug repair, you may want to consider removing individual branches after you have completed making changes to the code. In contrast to this, you should typically refrain from pursuing either the Development or Master Branches for the foreseeable future.

In order to sever a branch, you need to use this command. Not only does this result in modifications being deployed, but it also brings attention to the fact that the branch relies on the development branch at the place where the exit point is located. You made a duplicate of another repository and referred to it as the new repository resource while you were in the process of setting up this repository for the first time. 5 It is totally up to you to decide whether or not you want the branch that you are now working on to be included in the public repository for the whole project. If you are presently working on a feature that you intend to share with other people when it is ready, but not before, you should avoid pushing that branch to the public repository until the feature is ready to be shared with others. In general, this applies if you are working on a feature that you want to share with others when it is ready, but not before. The development and domain branches are the two exceptions to this rule; it is vital to have these branches in the main warehouse at all times.

1.15 COLLABORATION ON GITHUB

When you first create a repository on GitHub, it is only visible to you, and only you have the ability to make changes or updates to it. If you clone them, anybody will be

able to see their source code. However, you are the only one who has authority to make changes to the repository. Simply make a copy of them, and anybody may see the source code. The most apparent advantage is that arbitrary individuals are barred from tampering with your code; nevertheless, the disadvantage is that it makes it more difficult for individuals to work together. Sharing the repository's write access with other people is one approach to work together with other people. To begin, you will need to choose the Settings option that is found on the toolbar at the top of the repository site.

After that, you have to pick Collaborators from the list of options on the left. As soon as you are finished, you will be brought to a page on GitHub where you may add collaborators. These contributors are distinguishable from one another based on their user accounts on GitHub. Alterations made by collaborators may be pushed into the repository, which also allows you to publish your own modifications. When a number of people are working on the same piece of code at the same time, it is critical to maintain a level of organization about the process by which changes are merged into the main (and/or development) branch. This will assist in lowering the likelihood of an excessive amount of misunderstanding. Before incorporating the adjustments into the primary project branches, one strategy that is advocated and approved by GitHub is one in which the adaptations are first produced in separate branches and then debated via a procedure known as pull requests.

1.16 DRAWING REQUIREMENTS

If you want to send pull requests in the right manner, you must first deploy any newly developed features, bug fixes, or other modifications to branches that are distinct from either the development or the domain branch. The next step is not to instantly merge two independent branches into a single one, but rather to issue a pull request to the other developer. You may start a pull request by heading to the main repository page, choosing the branch, and then clicking the large green button that is labeled New Pull Request. This will allow you to get the ball rolling on the pull request. Alternately, if you've made recent modifications to your code, you should also notice a green button that's titled "Compare and Capture Request." When you click this button, the process of putting in a pull request will get underway.

When you click the button, you will be sent to a screen where you can offer a description of the modifications you made to the code as well as give your pull request a name. The moment you press the button, you will be brought to the current screen. Additionally, it is up to you to decide on which branch the pull request should be merged. You will notice a drop-down menu just above the title where you have submitted your pull request. From this menu, you will be able to pick two branches: the branch you wish to combine (named Base), and the branch where new modifications will be stored. (Compare). To begin the process of creating the new branch, you must first choose one of the two original branches from which to draw inspiration. After that, you may submit the pull request that you want.

It only offers a web interface that can be used in order to enable the display of modifications that have been made to the website. You will be able to examine recent changes, provide comments on these changes, and provide feedback on the department as a whole by using the website. In the same vein, anybody is permitted to check the branch as well as make their own modifications, provided that they do it in a public manner. Any individual is free to contribute further enhancements to the branch so long as the pull request is still open and the conversation is still taking place. You are free to merge the pull request after you have completed all of the necessary work on your project. (via the large green "Merge Pull Request" button on the pull request discussion webpage).

1.17 FORK POOLS INSTEAD OF CLONING

It is still necessary to have write access to the repository in order to make modifications to the different versions of the repository and then to submit pull requests in order to incorporate those modifications. It is necessary to have this access in order to make modifications to the repository's various versions. Because other people don't want to give you complete write access to their software, this isn't an optimal situation for receiving updates from unknown individuals or making changes to packages developed by other people. This can be problematic when trying to make changes to packages. As a result, none of these pursuits would benefit from its participation. On the other hand, if you have a small group of close buddies, it's a fantastic opportunity to collaborate. On GitHub, it is still possible to work together with other people even if you do not

have write access to the repositories that each of you uses. There is no need for you to be concerned about this. Because of the way that pull requests are managed, it is not necessary to combine different versions together in order for them to be counted as part of the same main repository.

This is because of how pull requests are managed in the system. It is up to you to decide whether or not you want to combine branches from various locations. If you want to make modifications to a repository that you do not have write access to, you can clone the repository instead of making the modifications directly to the original repository. Once the clone has been produced, you can make the modifications to the original repository that you purchased. However, because you do not have access to the repository from which you copied the repository, you are unable to send these modifications to the original repository. Because the modifications you make to your local system are stored on your computer and not on the GitHub server, the other users of GitHub won't be able to view them unless you choose to make them publicly visible and share them.

Create a new repository on GitHub that is a clone of the repository you want to modify but is otherwise vacant before you attempt to push your changes to an existing repository on GitHub. This step is required before you can submit your changes to an existing repository on GitHub. After that, in order to access the repository, you will need to clone it to your local computer. Because this is a centralized repository and you have written access, any edits that you make on your personal computer at home can be posted to the repository on GitHub, where they will be visible to other users of GitHub. It can be found on GitHub. When you make a new repository on GitHub in the same manner as this one, GitHub refers to it as a "clone" of the original repository. Forking a project means creating your own version of it and developing it independently of any previous versions of the project. When a repository is cloned, a new repository that is vacant is produced.

This is the only distinction between cloning and forking. However, from a strictly technological point of view, forking is the same as cloning. The concept of "creating your own version" of a project is referred to as "forking" in the realm of open-source software, which is where the word "forking" first appeared. On the other hand, if you

navigate to the homepage of a repository on GitHub, you will notice a hyperlink in the top right portion of the screen that is designated "Fork." This option can be found to the right of the name of the repository whose branch you are perusing as well as the name of the repository itself that you are browsing. If you have the ability to make changes to the repository you are using, selecting the "Fork" option will cause a duplicate of the repository to be created in a new location for you to use. Even if your repositories are not already forked, you are not permitted to split them. because it should go without saying. Having said that, in the overwhelming majority of cases, you probably don't want to establish your own repositories anyway. I have no idea why this is the case. You can also clone any repository that is maintained by accounts that are controlled by users other than yourself.

When you have a duplicate of the repository in your possession, you can clone it to your computer and modify it in exactly the same way as you would edit any other repository. You have the same level of access to change it as you do with other repositories. The only thing that distinguishes this repository from the one you made yourself is that whenever you make a pull request, GitHub recognizes that you have extracted it from another repository. This is the only distinction between the two. The only thing that distinguishes this repository from the repository that you created is this one single difference.

This one thing is the only thing that sets this storage facility apart from others. You have the ability to select not only the base branch and the comparison branch when you make a pull request, but also the base branch and the primary branch in addition to the base branch and the comparison branch. Both the repository from which you want to impose the settings, known as the root fork, and the repository in which you made the adjustments, known as the root fork, are referred to simply as the repository. When you make a pull request to the project's initial repository, you will not see the basic and major choices if the project has been cloned by another user.

On the other hand, you have the option to turn them on by selecting Analyze across businesses when you are generating the pull request. In this manner, the distinctions between the various segments can be more clearly seen. When you check out modifications from another person's repository, the procedure that you go through is

identical to the one that you go through when you check out your own projects. The only distinction is that you won't be able to integrate the pull request once the conversation about the modification is finished. Those who have been granted the write authorization can carry out this work.

When somebody else wishes to make modifications to your code, a procedure very similar to the one described above is followed. You are free to start a pull request with modifications to your code, but at the conclusion of the conversation regarding payment, you are the only person who has the authority to decide whether or not to submit those modifications to the repository. One of the many benefits of using Git and GitHub is the ability to freely and adaptably co-author the source code with input from anyone, including total acquaintances. This is just one of the many advantages. This is just one of the many elements that are at your disposal.

CHAPTER 2

DATA MANIPULATION

The process of developing data-driven models and performing data modification are both essential components of the field of data science. Rarely do we collect data in a manner that enables us to immediately input it into statistical models or machine learning algorithms, which we intend to use to evaluate the data once we have collected it. Because we want to complete the data analysis as rapidly as feasible, this presents a challenge for us. Figuring out how to import the data into R and then figuring out how to rapidly transform it into a form that can be analyzed are the first steps in any data analysis project.

This step is almost always the first step in any data analysis project. After that, it is decided how the findings of the research should be interpreted in light of the findings. The technique that is outlined below begins with these first stages. Before continuing with the decoding procedure, it is necessary to ensure that the programs magrittr and ggplot2 have been successfully installed. This pertains to the entire segment, in addition to the sections that come after it (in order not to do this explicitly in every example).

2.1 DATA ALREADY IN R

R comes preinstalled with a variety of datasets, but you can also find datasets in other R programs if you search for them. They are helpful in gaining a comprehension of the implementation of novel techniques that can be accomplished with the assistance of its use. It is much simpler to evaluate how well a new strategy is working when one is familiar with a dataset and what that dataset can teach them about how well the strategy is working. This evaluation will be significantly more challenging for you if you do not have prior experience with datasets. One more beneficial application of this information is the comparison of the strategies that you use.

Therefore, it should come as no surprise that their usefulness is significantly reduced when it comes to analyzing current data. Following completion of the installation process, the dataset program will be incorporated into your R distribution

automatically. Using the library() function in R will allow you to successfully install the program. After that, you will be presented with a summary of records that are included along with a concise explanation of each record, as will be demonstrated in the following portion of this article.

The sort of summary that is generated differs depending on the number of categories that are presented. The number of items in each category as well as the TRUE or FALSE values are used to provide a synopsis of the classification data and the Boolean data, respectively, while quarters of the numeric data are used to provide a summary of the information. There is a column in the Iris dataset that is referred to as "Species," and the total for each level serves as a synopsis for that particular column. The classification by category can be found in this section. You can find out what kind of data is stored in each column of the database by using the text() function. The dataset is organized into columns. This provides you with the framework for a data type, but at the moment, it is much more comprehensive than you might want it to be. On the other hand, it gives a synopsis of the many different kinds of sections that can be discovered in a data frame, which is very helpful for the purpose it was designed for.

2.2 READING THE DATA

Data that has been stored in a variety of file formats, such as Excel, JSON, or XML, can be viewed by using a variety of applications. If the data you have is already organized in a specific structure, you should search the internet for step-by-step instructions that explain how to import the data into R. If the data format is Standard, then there is presumably already a package that can help you deal with this, and if there is, you can use it. If there is, however, you cannot use the package, then you will need to create one. On the other hand, data is frequently considered to be a distinct variety of text string. The overwhelming majority of software applications provide users with the option to integrate and export their data.

One of the many built-in functions in R that can be used to read different kinds of data is referred to as the reader. This function can also read other built-in functions. The following syntax can be used to generate a list: read.table The read.table() function is implemented slightly differently in each of these forms; the most notable difference is

in the initial parameters that are utilized by each technique. For instance, read.table() anticipates that the data will be presented in the form of space-separated columns, whereas read.csv() anticipates that the data will be presented in the form of comma-separated values. Nevertheless, data must be delivered in the appropriate manner for either one of these functions to be performed successfully. Accordingly, what differentiates these two functions from one another is the manner in which they analyze the meaning of the column delimiter for the data.

This particular characteristic is what differentiates the aforementioned two functions from one another. The read.table() function accepts a significant number of parameters, which can be sent in as part of the request to the function. These are used to alter it so that it corresponds with the specifics of the text file that you are currently reading. (The only thing that differentiates this procedure from the others is the number that is used by default for each of the parameters that are received.) The choices that I count on the most on a regular basis are as follows:

- header is a boolean value that tells the function whether to treat the first line of the input file as the header. This value is used to specify whether the function should do this. If true, the first row will be used to define the column names of the generated data frame; If set to false, the first row is considered the first row of the data frame. The first row is used to define the column names of the data frame it creates when set to true. If neither true nor false is specified, the value is assumed to be true by default.
- col.names - If the first row is not used to create the header row, this option can be used to name the columns instead of using the first row. This is the case where the first line is not used at all. You have to give it an array vector where a string matches each column of data you get.
- The letter "dec" is used instead of the decimal point seen in numeric expressions. because the spreadsheets I'm dealing with use both. And in terms of decimals, I think this number is pretty important. The number of different nationalities you deal with on a daily basis is probably the most important factor in understanding how important this is to you.

- comment.char: When the function encounters a line that continues after the # character, it assumes that line is the beginning of a comment and ignores any remaining characters in the line. This happens when the function encounters a line followed by the # character. If the # value is actually used elsewhere in your data, you need to make the necessary changes. You can use a different font for the comments, but the meaning remains the same.

- stringsAsFactors - By default, when this argument is presented, the function assumes that columns containing strings should be interpreted as factors. This assumption is made every time it encounters the stringsAsFactors argument. It goes without saying that this is not always the case: there are cases where a string is just an array. By setting this parameter to FALSE, you can instruct the function to read strings in the most appropriate form for those strings. However, there is not a single option that falls into the gray area. If false, none of the columns containing strings are considered factors; On the other hand, if true, all columns containing strings are considered factors.

- colClasses: allows you to define the type each column should have; For example, in this section you can declare that some columns should be factors and other columns should be strings. You are forced to define all columns, which is not only time consuming but also a bit tedious because R is generally pretty good at recognizing the correct types for a column. However, you are responsible for doing so. Even if you are determined to go with this approach, there is a limited amount you can do. It can specify that a column must have a ranking factor, but cannot define levels or other properties on it. Using it speeds up functionality for large datasets, as R doesn't need to find the column types itself. The main reason I use this is to specify which columns should be factors and which should be strings. Its main function is to let me choose which columns to use to store factors and which columns to use to store strings.

If you provide read.table() and its utilities with the appropriate information, they will, for the most part, get you to where you need to be when it comes to accessing data tables. In point of fact, read.table() and the collaborators it works with are intended to

cooperate. If you are having difficulty accessing the data, you should take a careful look at the documentation to determine whether or not the procedures can be altered in any manner to make it possible for the data to be imported. It is feasible most of the time, though obviously not always. I typically give up and create a script in another language when I discover something that isn't exactly what I imagined it would be. This allows the data to be produced in a manner that I can integrate into R.

If I discover something, it won't be what I was expecting to find. was, I can make the necessary changes in R. R is not the most effective program to use when working with simple text. I'm going to be realistic and use the tools that are best suitable for the work at hand rather than trying to force all phases of an analysis to be done in R. It is not always necessary for to be R. This is due to the fact that R is not the most effective instrument for handling simple text. However, before you take extreme measures, such as converting to programming in a different language, you should do a comprehensive investigation to see if anything is being viewed. This will help you determine whether or not you need to make the switch. table() techniques are open for customization. This should be completed before moving on to any other tasks.

After navigating to the Website that will facilitate the data transfer for you, you will be able to save the file to your personal computer. When inputting data from sources other than R code, there are obviously advantages and disadvantages associated with doing so. On the one hand, it is not hard to carry out and gives you the opportunity to examine the data before beginning the analysis process. On the other hand, it adds an additional stage to the procedure of non-automatically repetitive analytics. In spite of the fact that the documentation offers Address information and makes use of a link that does not become outdated with the passage of time, there is still a stage of the process that calls for user participation. Additionally, this is a stage that the vast majority of people get incorrect the vast majority of the time.

Instead, I study the article in its totality straight from the website in order to acquire the knowledge I require. I am unable to provide an assurance that the data will continue to exist on the computer or will remain the same over the course of time because I do not have authority over which server the data is housed on. This level of a pipeline is considered to be one of the most unpredictable sections of the process because of the

fact that it directly results from this truth. There is always the possibility of something going wrong. Even though I have the code in my workflow that enables me to download the data the vast majority of the time, I almost never forget to save a copy of the data to a file as well. When I need it, copying and pasting the code to download the data and transfer it to my local device in a stored markdown fragment only works once. Using the readLines() function, I am able to read the data and obtain it in the form of a line vector. Utilize it in this manner. This makes it possible for me to quickly examine the first or second line of the file to get a general sense of what the remainder of the file includes without having to read the entire file.

2.3 DATA PROCESSING WITH DPLYR

When you want to represent your data in a way that each section can stand for a different numerical measurement that you want to incorporate into your models, using data blocks is an extremely helpful technique. The utilization of data frames lends itself particularly well to the provision of this kind of data representation. The overwhelming majority of R programs, including those that can be used to develop statistical models or machine learning strategies, depend on data frames as the primary data storing mechanism. This is because data frames are the most efficient way to organize tabular data. However, in order to make substantial alterations to a data frame, it is typically necessary to write a significant quantity of code in order to filter, reorganize, and consolidate the data in a variety of different ways.

A few years back, the amount of code necessary to manipulate data frames was significantly higher than the amount of code required to conduct data analysis. The implementation of the dplyr package, which could also be dubbed "d pliers" (where "pliers" is pronounced in the same way as the instrument "pliers"), was a substantial enhancement over its predecessor. This program comes with a number of helpful functions that make it simple to manipulate data chunks in a variety of ways and to connect them together making use of the %>% operator. This product comes with all of these functions already installed.

To the best of my knowledge, this operator was first implemented in this package, and all that is required of you in order to gain access to the operator is for you to import the

dplyr package. However, magrittr supports a variety of enhancements; because of this, I strongly encourage that you always incorporate magrittr in addition to any other programs you may be using. You will receive functions that walk you through the process of developing pipelines to manage data frames when you integrate dplyr.

2.4 SOME USEFUL FEATURES OF DPLYR

In this chapter, I won't be able to discuss everything that dplyr is capable of because there isn't enough time. In any event, it is changed at such a frequent cadence that by the time you read it, there will probably be additional features that were not available when I penned the chapter. This is due to the fact that it is changed on a fairly regular basis. This is due to the fact that it is continually refreshed. Therefore, the first thing you should do is determine whether or not the documentation that was included with the program has been modernized. You will discover explanations of functions that are utilized frequently by me in the following portion.

Since each one of these functions accepts a data frame or something comparable as its first parameter, they are completely consistent with pipelines. When I state that they accept any argument that could be considered a burst argument, what I really mean is that they accept any argument that could be considered a "burst-equivalent." It is not unheard of to come across other burst representations that take precedence over the data structure that the application was developed with. It is sometimes beneficial to use a distinct representation of the integrated data frame when working with massive data sets; Certain alternative data structures are preferred over established data structures because they can work with data that is recorded on storage.

R data frames have to be imported into memory before operations can be conducted on them, but some tasks are more effective when carried out on other structures. Or perhaps the standard of the print job as a whole is very high. When the data frame's name is entered into the R interface, the software displays the contents of the data frame and has the ability to automatically retrieve the data heading from a representation that is provided to it. The dplyr program offers a number of different perspectives on the data. tbl df rendering is something that I come across quite frequently. I do so because the appearance of clichés that come about as a consequence of printing appeals to me.

2.5 TREATMENT OF BREAST CANCER DATA

Let's take a look at how the dplyr program works in practice so that you can get a clearer idea of how it can assist you in data exploration. You should now have a clearer understanding of how beneficial it can be. In a minute, we'll discuss the breast cancer statistics in more detail. To begin, we will investigate the modifications that were made to the original data after it was brought in from the CVS file (and now stored in the Breast Cancer raw variable). The process of formatting this information required editing the factor that corresponds to the class variable. This was a necessary step in the process. It was one of the steps in the process to get to this point. In order to accomplish this goal, numbers were directly assigned to the Class variable making use of the Breast Cancer $class that was configured. Nevertheless, we can execute it directly as a data frame modification by utilizing the modify() function that is made available by dplyr. This provides us with:

Therefore, I cannot state with confidence whether this edition is simpler to comprehend than the preceding one. If you are not used to writing code in pipelines, this may not be an easy task for you to complete. However, as soon as you become accustomed to understanding pipeline code, you will find that it is not nearly as difficult as you might have thought. In any event, this makes the transformation very obvious, and it should go without saying that we produced the structured breast cancer data frame by conducting transformations on the breast cancer data frame without resorting to any other form of modification. This is a fact that cannot be disputed. Let's start by taking a look at some unprocessed statistics now that that's out of the way. This is a fairly straightforward data analysis that we are able to perform, but we do it for the purpose of investigation. It might be fascinating to investigate how different circumstances have an effect on the reaction variable that is referred to as the class variable.

Is there, for instance, a distinction between normal and dangerous tumors in terms of the thickness of their cell membranes? In order to determine whether or not this hypothesis is accurate, we can categorize the data according to the cell characteristic and examine the average thickness of the cell. Once more, information that is helpful to us has come to our attention. When the cell size is large, it appears that there are more dangerous tumors, whereas when the cell size is small, it appears that there are

more innocuous tumors. It would appear that there is an equivalent amount of normal and cancerous lesions in medium-sized cells. It has the potential to serve as a cornerstone and a foundation for our work, even when it comes to the development of statistical models.

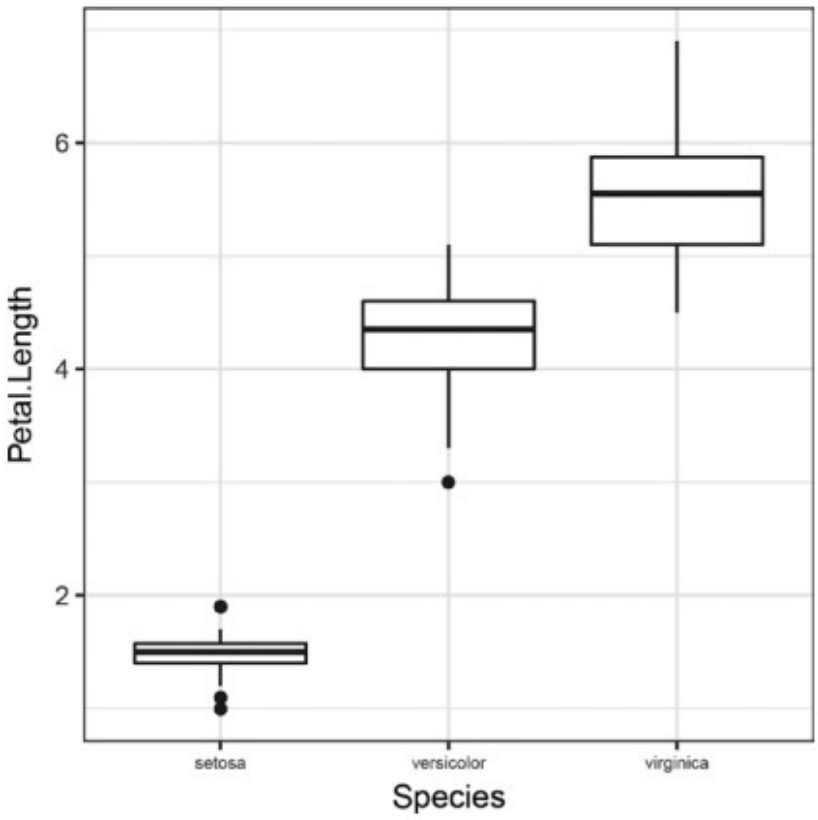


Figure 2.1 Species representation by petal length

The fact that the diameters of the cells can be determined using distinct numbers is the sole reason why this form of classification is helpful. This is the only reason why organizing things in this way makes any kind of logic. In this specific instance, clustering by floating point integer will not produce the desired results. There, graphs might be of greater assistance. It is feasible for us to conduct data analysis by building tables in the manner described above because the cell widths in this particular data collection are represented as integers. This makes it possible for us to do so. One more possibility is that we watch the characteristics being combined as this process goes on.

Since we already know that the nature of a tumor seems to have an effect on both the size of tumor cells and the thickness of tumor cells, we want to investigate how the thickness of tumor cells functions in relation to the cell class and the size of the tumor cells. The behavior of tumor cell thickness in relation to cell class and size can then be analyzed with the assistance of this information.

Because I have one column for the x-axis and another column for the y-axis, this works out perfectly for me. What if, however, I want to plot it on a graph to examine how the various pupil measurements compare to one another? Each parameter has been assigned its very own column in the database, which was just created for it. The titles of their associated properties include, amongst others, elements like Petal.Length and Petal.Width. I recently fell into a small problem as a result of the fact that various measures in my data frame are located in various sections. Placing them on an x-axis and a y-axis is difficult for me to do because there isn't a clear cut method for doing so. The software Tiddr provides a solution to this issue that you can implement. The Collect() function is a useful tool for manipulating the data frame in such a way that certain columns become the factor's titles while other columns become the factor's associated numbers. Utilizing the Collect() function will allow for this modification to be made. In practical application, you would need to change the data frame in order to obtain one column that lists the titles of your initial columns and another column that lists the values that were initially recorded in those columns. As a consequence of this, you now have two columns: the first one lists the titles of your original columns, and the second column lists the values that were initially saved in those columns.

Within the iris collection, we have measurements regarding the length as well as the breadth of the petals. It is not at all difficult for us to carry out an analysis in which we compare species based on the length or breadth of the sepals. If we are to consider both characteristics simultaneously for each category, we are going to find ourselves in a more challenging situation than we were in before. Since the data frame does not have the framework that we require for this purpose, we cannot use it. We can see that these two measurements correspond to different units of measurement if we arrange them against the Sepal.Length and Sepal.Width variables. In order to properly organize the data that we have in our data frame, we need to add not only a column that indicates

whether a measurement is longitude or latitude, but also a column that indicates the actual value of the measurement. This will allow us to properly categorize the data that we have. This is essential in order to structure the data that we have stored in our database in an appropriate manner. Using the Collect() function that is offered by Tiddr, you will be able to accomplish this goal.

This excerpt instructs Collect() to generate two columns: one named Attributes, which lists the column titles of the input data frame, and one named Measure, which lists the values of the key columns. Both columns will be returned in the output of Collect(). Both categories can be found in the data packet that is produced. The data block that was generated by the software contains both entries in their respective positions. You can see that the final data frame contains an attribute column with the names Sepal.Length and Sepal.Width, in addition to an additional column that contains the measurements (you can see this above if you send it through head() ;). only the Sepal.length variable is displayed in the report. In the event that you do not transmit the output from your cranium, you will see this (). As a consequence of carrying out this process, the data are transformed into a structure that makes it possible for us to display characteristics against measurements (see Figure 2.1 for output).

Today, the term "artificial intelligence," more commonly abbreviated as "AI," refers to a change in perspective that is simultaneously driving the expansion of scientific research as well as the development of a broad variety of businesses. As a result of the high level of expertise in a particular field that is required to fully understand the technical aspects of artificial engines, both what artificial intelligence is and what it is capable of accomplishing are frequently misunderstood: the general public is interested in its development and fears scenarios similar to Terminator; investors raise substantial capital; however, they do not have a clear picture of the competitive factors that characterize companies and products; Even though there is a widespread sense of positivity regarding the possibility of advancing artificial intelligence (Muller and Bostrom 2016), I believe that certain concepts need to be developed in order to encourage a faster rate of AI development.

In point of fact, there is a generally positive outlook regarding the development of advancements in AI. The purpose of this study is to investigate additional innovations

brought about by artificial intelligence, both scientifically and in business models; to gain an understanding of where the value lies for investors; and, finally, to stimulate the discussion regarding the potential risks and future developments associated with artificial intelligence. The goal of this article is to investigate additional breakthroughs that artificial intelligence enables, not only in the realm of science but also in the realm of business strategy. Description and explanation of a few central ideas, as well as a brief overview of the development of artificial intelligence (AI) throughout history and advancements in the relevant literature.

2.6 BASIC DEFINITIONS AND CATEGORIZATION

To begin, let's pause for a moment and talk about what we mean when we say "artificial intelligence." According to Bostrom (2014), the concept of artificial intelligence (AI) in the modern world can be understood in three different ways: as something that can answer all of your questions with increasing accuracy ("prophecy"); as something that can do whatever it is told to do ("genie in a bottle"); or as something that can act independently to pursue a long-term goal ("monarch"). On the other hand, artificial intelligence (AI) should not be defined by what it is capable of doing and what it is not capable of doing; rather, a description that is more comprehensive would be more appropriate.

The term "artificial intelligence" refers to a collection of instructions that, when followed, enables computers to generate their own algorithms without needing to be specifically programmed (also called an algorithm). In a nutshell, artificial intelligence (AI) describes a system that is capable of teaching itself new things. We frequently conceive of intelligence as the computational component of our capacity to accomplish certain objectives; however, a description of intelligence that is more accurate would be the capacity to learn new things and find solutions to problems in an environment that is constantly shifting. In point of fact, the relationship between intelligence and flexible behavior is much closer. In a universe devoid of history, there is only one mentality that can guarantee the continuation of life and reproduction: (Lo 2012, 2013; Brennan and Lo 2011, 2012). Whether or not a creature is accountable for bringing the world to the state it wishes to optimize is one of the parameters that can be used to determine whether or not the creature possesses intelligence.

Even uneducated members of our group are able to see that the degree of intelligence presently provided to computers is light years ahead of what is possible for typical humans. This is true regardless of how precisely we characterize the concept. This is the case regardless of whether or not we provide a precise definition of the word. An artificial intelligence is propelled purely by data and does not have any previous understanding of the nature of the relationships between things, in contrast to humans, whose actions are directed by witnessing the physical world and inferring fundamental cause-effect relationships in natural occurrences. somewhere between these two times. On the other hand, individuals have previous awareness of the nature of the relationships that exist between these facts, which stands in opposition to this truth. Therefore, in this specific setting, we refer to it as "manufactured" because it is not derived from the physical principles but rather from the data itself, and not the other way around.

Our conception of the phrase "Artificial Intelligence" (AI) and what it refers to has been modernized as a result of recent occurrences. Nevertheless, as part of this introduction to artificial intelligence, there are two additional ideas to think about: first, how does AI differ from and/or relate to other concepts such as big data and machine learning; and second, what characteristics a system must have in order to be considered intelligent. Both of these ideas can be broken down into subheadings. Within the scope of this introduction to artificial intelligence, it is necessary to conduct analysis on both of these concepts. I believe that artificial intelligence, also known as AI, is an interdisciplinary discipline that necessitates the study of a variety of subfields, including natural language processing, computer vision, the internet of things, and robotics.

Additionally, a broad assortment of other disciplines, such as B., are included in this category as well. automation and natural language translation technology. In this particular setting, the word "artificial intelligence" serves more as a generic expression that can refer to a variety of different facets. It is possible to make comparisons in order to evaluate the degree to which AI is connected to other (sub)fields, and it is also possible to look at AI in a manner that is comparable to a biological creature that is operating at its maximum capacity. . Both types are going to be covered in the following

section. If there are parallels to be drawn between the human body and artificial intelligence, then the AI must also possess a brain. The brain is the component that is responsible for a wide variety of specialized duties and abilities, such as speaking (the processing of natural language), vision (the processing of computer vision), and others.

In the same way that a robot is constructed out of mechanical and electronic parts, the human body is made up of skeleton structures and the muscles that connect them. The process of machine learning can be conceptualized as the development of particular patterns of behavior, thoughts, or actions, which are then perfected through continued interaction with the system. The Internet of Things (IoT) can be compared to the human senses in that both are conduits through which information is conveyed to us about the environment in which we find ourselves living. In a nutshell, Big Data can be compared to the food that we consume, the atmosphere that we breathe, the gasoline that permits us to work, and the various types of information that are brought to us from other countries. the universe that can be perceived through our senses.

Despite the fact that the comparison is quite simplistic, it eloquently and concisely explains how each of the concepts are connected to one another. It is always going to be extremely difficult for AI to determine what characteristics a system must have in order for it to be considered a functional system, despite the fact that many additional similarities can be made and many can be rectified at the same time. In my opinion, the system ought to come equipped with a learning structure, an interface for collaborative communication, and an input processing process that is analogous to sensing procedures. Regrettably, this recommendation lacks scientific precision due to the fact that a variety of ethical, psychological, and philosophical considerations will need to be taken into account.

Rather than spending much more time concentrating on this unprovable notion, I'd rather demonstrate how these characteristics symbolize the various kinds of AI that we're dealing with now and will be dealing with in the future. Those are the varieties of AI that we'll be dealing with in the future. I'd rather just squander my time than do that, if I had a choice. An AI can, in point of fact, be broken down into three distinct categories, which are as follows: There are three distinct kinds of AI. A limited AI, which is nothing more than a particular application or activity that develops over time,

reduces the overall amount of erroneous output that is generated. ; a general AI that encompasses all other types of AI; and a hybrid AI that incorporates general and specialized AI. In this particular instance, "Deep Blue" for chess is the best example; however, more generally speaking, the word "functional technologies" applies to any technology that has a specified purpose. In this particular instance, "Deep Blue" for chess is the best example. Because their architecture calls for them to execute a constrained set of tasks, such systems are typically designed to require only minimal amounts of maintenance effort.

The term "artificial general intelligence" refers to a situation in which a computer program has not been pre-written for a specific task, but rather has the potential to learn from an application and apply the same knowledge to a wide variety of parameters. This type of program is considered to have artificial general intelligence. . Stuart Russell, a computer scientist at Stanford University, is the one who first came up with the phrase "artificial general intelligence" (AGI). This is not an example of technology being provided in the form of a service like the picture that was just presented; rather, this is an example of technology being provided in the form of a commodity.

Even though Google DeepMind is not an artificial general intelligence (AGI), it is the most prominent member of this particular subcategory because it is the finest example and therefore the leader in the field. It is correct that we have not yet arrived at that stage because even DeepMind is not yet capable of performing an intellectual activity on the same level as a person. In order to accomplish this objective, a substantial quantity of additional research into the functioning of the brain structure, methods to enhance brain function, and the development of portable processing capacity is required. It is possible that some individuals are under the impression that artificial general intelligence (AGI) can be created by merely combining numerous specialized forms of artificial intelligence, but in point of fact, this is not at all the case.

It is not important how many different specialized skills a program can support; rather, what matters is how well all of these skills are connected together. However, it does have one major limitation, which is that in its current state, it can only be achieved by passing an infinite amount of data to the engine. This type of intelligence does not need to be fine-tuned or fine-tuned by an expert as limited AI would; however, it does have

one major limitation. Limited AI would. In opposition to this, restricted AI must be operated or modified by a trained professional in order to function properly. His intelligence is far beyond that of human beings, and he is capable of thinking in both a scientific and artistic manner. It is distinguished by extensive general information in addition to social competence and potentially emotional intelligence. In addition to that, it possesses these characteristics.

Instead, the stage that comes after that is referred to as Highly Intelligent Artificial Intelligence (ASI). Although we typically conceive of this intelligence as a single very powerful computer, the reality is that it is more likely to be generated by a network or a collection of different intelligences operating together. On the other hand, the manner in which individuals progress through the various phases is still a contentious topic of discussion, and there are a great deal of diverse schools of thought regarding the topic. It is possible, according to the symbolic method, to symbolize any and all information conceptually. However, due to the limited amount of space that is available for representation, it is recommended that everything be articulated using a rigorous mathematical language.

This method has been used throughout history to study the complexity of the real world; however, it has always suffered not only from computer problems, but also from the inability to even understand the source of the information. Throughout history, this method has been used to study the complexity of the real world. Statistical artificial intelligence, on the other hand, takes a more inductive approach, as opposed to the more deductive method of logical artificial intelligence, and focuses on controlling the uncertainty that is present in the real world. This is in contrast to logical artificial intelligence, which takes a more deductive approach (Domingos et al. 2006).

While statistical artificial intelligence focuses on drawing inferences, logical artificial intelligence uses a more inductive approach to its research. When it comes to the application of unevenly distributed representations, also known as SDRs, for the purpose of information processing, a biological neural network seems to provide an excellent infrastructure for the development of artificial intelligence. This is due to the presence of biological neuronal networks in creatures that are still alive. On the other

hand, the amount of consideration that should be given to the human intellect at this point is not completely obvious.

In spite of all the recent hoopla surrounding AI, its beginnings mostly stretch back to the 1950s. Artificial Intelligence (AI) study is not a novel developmental research field. If we disregard the philosophical line of reasoning that started in ancient Greece and continued all the way up to Hobbes, Leibniz, and Pascal, the official foundation of artificial intelligence as we know it took place in 1956 at Dartmouth College. This is assuming that we ignore the philosophical line of reasoning that started in ancient Greece and continued all the way up to Hobbes, Leibniz, and Pascal. It all starts in the ancient Greek world. Here in this location, a gathering of some of the most well-known and respected specialists in the field was convened to discuss the various approaches that can be taken when carrying out intelligence simulations. It had only been a few years since Asimov had established his three laws of robotics. To be more specific, it had only been a few years since Turing had published his famous paper in which he first proposed the concept of a machine to think about and Turing's favourite machine idea to demonstrate.

Turing's paper was published in 1950. determining whether or not such a device possesses true intelligence. This incident happened only a few short years after Asimov presented his three principles of robotics to the world. It wasn't until a few years after Asimov developed his rule of three robotics that this actually came to pass. Following the publication of the material and ideas developed at that summer conference by the Dartmouth research team, a trickle of government financing was directed to investigate the construction of artificial intelligence. This choice was decided in what felt like a split second. During that time, it seemed as though artificial intelligence was right around the corner. However, as time went on, it became abundantly obvious that this was not the case. Researchers in the late 1960s started coming to the conclusion that the subject of artificial intelligence was indeed a complicated one.

The original spark that had attracted financing started to dip as a direct consequence of this consciousness that began to spread. This phenomenon, which has shaped AI throughout its history, is often referred to as the "AI effect," and it consists of two parts: first, the continued promise of true AI over the next decade; second, updating the

behavior of AI after solving a particular problem and constantly redefining what intelligence means. Throughout its history, AI has been shaped by this phenomenon, which is often referred to as the "AI effect." The term "AI Effect" is frequently used to allude to this impact because it has been moulded by AI throughout its existence. The guarantee that real AI will continue for the next ten years, and the reduction in AI's behavior after it is developed, are the two separate categories that can be used to classify these effects. Initially, it was suggested that the United States Defense Advanced Research Projects Agency (DARPA) provide funding for research into artificial intelligence in order to develop the world's first error-free machine interpreter.

The simultaneous occurrence of two independent events, which marked the beginning of what would come to be known as the first winter of artificial intelligence, however, caused the strategy to fail and was the cause of the disruption. In point of fact, both the "Lighthill Report" (1973) and the report that was produced by the Advisory Committee for Automatic Language Processing (ALPAC) in the United States in 1966 evaluated the applicability of artificial intelligence in light of recent developments and came to the conclusion that it is not possible to create artificial intelligence. A computer that possesses the ability to learn or that is perceived to be knowledgeable. 1966 and 1973, in their individual years, saw the publication of two findings. These two versions, along with the limited data that was available to power the algorithms and the limited computational power of the engines that were available at the time, led to the field collapsing, which paved the way for the development of artificial intelligence. has been mocked relentlessly for the past decade.

However, in the 1980s a new surge of financing was spurred by the development of "expert systems," which were rudimentary instances of circumscribed AI similar to those mentioned above. This funding was stimulated in the UK and Japan. The creation of so-called "expert systems" was a driving force behind this latest round of financing. The development of so-called "expert systems" has been the primary impetus behind all of these financial outlays. However, the mere fact that these computers were capable of imitating the skills of human professionals in certain disciplines was sufficient to spark the emergence of a new fundraising trend. The Japanese government has been the most active participant in recent years, and the hurry to construct the fifth-

generation computer has tangentially prompted both the United States and the United Kingdom to reinvigorate their financing for artificial intelligence research.

The Japanese government has been the most engaged participant over the course of these past few years. Nevertheless, this fortunate period did not last for very long, and a new problem emerged as a result of the failure to meet the financial objectives. The computational capacity of personal computers surpassed that of the Lisp machine in 1987. This marked the culmination of several years' worth of research into artificial intelligence. Because of this event, the beginning of the second AI winter was easily identifiable. This event demonstrated that DARPA was taking a stance against artificial intelligence (AI), and as a result, it was not granted any additional funding. 18 A Brief Overview of Artificial Intelligence. When MIT's Cog project to create a humanoid computer and the Dynamic Analysis and Replanning Tool (DART) returned to the United States in 1993, thankfully, that period came to an end. Since 1950, the state is responsible for all money supplied. Both of these endeavors were carried out as a component of the Dynamic Analysis and Replanning Program (DART).

When DeepBlue defeated Kasparov at chess in 1997, it was evident that artificial intelligence had returned to the forefront of technological advancement. The concept of artificial intelligence as a paradigm change is relatively new; this is despite the fact that scholarly research has made substantial strides in the field over the past two decades. However, there is one particular event that we hold accountable for the pattern that has developed over the course of the past five years. Of course, there are many explanations that will enable us to comprehend why we are investing so much money in artificial intelligence these days. When we take a look at Figure 2.1, we can see that in spite of all the advancements, the word "artificial intelligence" did not achieve widespread acceptance until the late year of 2012.

It comes out that the picture was produced using CB Insights Trends, which is a tool that is both simple to operate and keeps up with current trends. for a wide variety of specific categories or disciplines (in this case, artificial intelligence and machine learning). To be more specific, I believe that this new surge of optimism in artificial intelligence (AI) started on December 4, 2012, the day that I think is the true catalyst for this new wave of optimism to begin. This is the beginning of this new hope. a group

of academicians delivered a demonstration on convolutional neural networks at the Neural Information Processing Systems (NIPS) conference the previous Tuesday of the previous week. These networks previously held the top spots in ImageNet's standings just a few weeks ago.

As a direct result of their efforts, the success rate of the categorization algorithm has increased from 72% to 85%, and the implementation of neural networks has developed into an essential component of artificial intelligence. The categorization was able to attain 96% accuracy in less than two years thanks to advancements in this field, which is slightly higher than the precision achieved by people (about 95%). The image also reveals three significant growth trends in the development of artificial intelligence, which are illustrated by three significant events: the three-year-old DeepMind, which was acquired by Google in January 2014; an open letter signed by more than 8,000 people from the Future of Life Institute and a reinforcement learning study published by Deep mind in February 2015 (Mnih et al. 2015); and finally, the article DeepMi published in Nature in January 2016, to which DeepMind contributed.

The subject of research known as artificial intelligence (AI) is one that must be invested with a significant amount of both time and money over an extended period of time. As a result, AI is highly reliant on financial support. As a consequence of this, there is an increasing concern that we are currently living through the subsequent zenith (Dhar 2016), but that the good times will come to an end in the not too distant future. [More citation is needed] However, I believe that this new era is different for the following three primary reasons: I (Big) Data, because we finally have most of the data needed to power algorithms; (ii) technological advances such as storage capacity, computing power, ever-increasing bandwidth and lower technological costs have allowed the model to be built by absorbing the necessary information; and (iii) democratizing and efficient allocation of resources offered by Uber and Ai. [More citation I (Big) Data as a result of the fact that we now have the majority of the data required to fuel the algorithms. On the other hand, I get the impression that this contemporary period is one of a kind.

Because of the enormous quantities of money that are being invested in AI research at the moment and the growing interest that the general public is showing in the topic, we

are researching AI in a more in-depth manner than ever before. It is obvious that this is because of the potential implications that it may have, the attention that it receives in the media, and the attention that it receives from the general public. Even though this is only true for low-level knowledge, the quick development of machine learning into a commodity is helping to promote a wider democratization of intelligence. In point of fact, it makes information at lower levels easier to obtain. Having said that, this is only applicable to information of a low degree. If, on the one hand, a multitude of services and tools are now available to end-users, and, on the other hand, concentrated in the hands of a few large owners with real power, data, and IT resources, then you can really take 'artificial intelligence' to the next level, and you can say with absolute certainty that the AI industry has reached a milestone.

In point of fact, large taskers possess the data as well as the processing resources necessary to genuinely advance AI to the next level. Putting aside the polarization that exists in technology, the primary challenge that the industry is currently facing can be broken down into two main components: first, the inadequacies of long-term AGI research that has been sacrificed for short-term commercial applications, and (ii) contrary to what people think or think AI is capable of doing, AI does not yet exist. This is the most significant issue facing the sector. Aside from this dichotomy in technological approaches, the primary challenges faced by the industry can be broken down into two distinct categories. The business is currently confronted with its greatest obstacle, which is the fragmentation of technology. Both of these issues are caused by the high degree of technological ability that is necessary in order to comprehend AI. On the other hand, these issues also contribute to an excessive excitement for AI.

The fact that artificial intelligence has already proven to be helpful in streamlining activities that have traditionally been difficult to complete because they require a certain degree of knowledge is evidence that some of the enthusiasm surrounding it is justified. The intimate connection that exists between people and machines, as well as the many different kinds of interaction that can take place between the two, is the second element. We are currently participating in a significant societal transformation that has been going on for a number of years. This is due to the fact that in the beginning, the human being is the controlling entity, and the computer is the protection

system for the disagreeable events that are taking place in the beginning. We are now a member of this transformation because of this reason.

However, in the modern society that we live in, the roles have been switched; machines are typically in charge, while humans only serve to superintend the work at best. The significance of this relationship lies in the fact that it alters who we are as individuals. Some people believe that cross-pollination is a way for humans to become more human, just as people try to do the same thing with computers. While the majority of people believe that machines make humans more like them, other people believe that cross-pollination is a way for humans to become more like machines. . One of the ways in which people can become even more human is through the process of cross-pollination (Floridi 2014). It would appear that the only argument that has garnered widespread support is the idea that in order to accelerate the acceptance of AI, we should learn to stop always relying on our instincts and instead rely on the computer that is already present within us to change the moment. whether it be human or artificial. This is the one and only issue on which there is widespread consensus.

It is the only viewpoint that has garnered approval from a significant number of people. In this setting, everyone wants to know, "Where do humans stand in relation to the machines?" He has every right to pose that question. The reality of the matter is, however, that we are still a considerable distance away from the so-called "singularity," which is the point at which superintelligence will transcend the reasoning capabilities of humans (Vinge 1993). Raymond Kurzweil, a prominent visionary, is credited with being the person who first proposed the concept of the law of accelerated returns in the year 1999. According to this hypothesis, the rate of technological advancement will immediately accelerate at a geometrically increasing rate in the future. Decrease the expense of the processor while simultaneously improving its computational capabilities. According to him, the route of human development follows the curve of an S, with significant benchmarks corresponding to developments in technological innovation along the way.

As a consequence of this, the development of human advancement can be interpreted more as a succession of jumps than as a continuous and unbroken progression. The human brain is capable of performing 1016 operations per second (cps) and can retain

1013 pieces of information at one time. Kurzweil came to the conclusion that, with these capabilities, humankind will achieve artificial general intelligence (AGI) by the year 2030, and the singularity will take place in the year 2045. This was accomplished by making the presumption that Moore's rule would always be accurate. This information comes from Moore's rule, which can be found here. However, I believe that this perspective is excessively optimistic due to the fact that the level of intelligence that machines possess in our present living environment is still quite restricted. They are clueless, they lack common sense, they have no concept of what an item is, they have no recollection of previous unsuccessful attempts, and they do not have any understanding of what an item is. This is the so-called "Chinese Room" argument, which argues that even if a computer does the correct translation from Chinese to English or vice versa, it will still not be able to fully capture the content of the discussion. This is the case even if the computer is perfectly translating between the two languages. Languages.

On the other hand, they tackle issues with a method known as structured reasoning; in addition, they have more storing space, memory that is more dependable, and sheer computational power. On the other hand, people endeavored to become more productive, to pre-select data that might be pertinent (at the risk of missing some essential information), to be creative and innovative, and to become more adept at forecasting the fundamentals. and more quickly than a limited number of cases, and they are also able to translate and apply this information to instances that have never been seen before. In comparison to other species, humans have a much greater capacity to function and generalize when they are exposed to an uncontrolled learning environment. To put it another way, it is simpler for people than it is for machines to carry out duties that are essential but do not place a significant amount of pressure on the practitioner.

While there are some straightforward actions that are almost impossible for a computer to perform (also known as "things people do without thinking"), mathematically demanding tasks are surprisingly simple for both a machine and our brain (also known as "moments of thought").). In a compelling proposition, Moravec's paradox summarizes a portion of this debate by stating that high-order thinking requires

minimal processing and is therefore applicable even to a machine. On the other hand, relatively simple and low-level sensorimotor skills will require a great deal of computational effort on the part of the machine. All of the evaluations that have been done up to this point are not a goal in themselves, but they are helpful in spotlighting essential design problems that should be considered when creating an Intelligence system. This is due to the fact that all of these evaluations have been completed up to this point. In addition, there have been the emergence of a few characteristics that are necessary for the creation of an artificial general intelligence (AIG).

Hybridity, resilience, and safety are a few examples of these. Consequently, hybridization. Russell et al(2015) .'s idea of an artificial intelligence needs to be verified (in the sense that it must behave within official constraints and meet official requirements), authenticated (in the sense that it must not monitor undesirable behavior within the limits set above), secure (in the sense that it must avoid deliberate changes by external or internal third parties), and controlled. A validation of an artificial intelligence takes place when it demonstrates behavior consistent with official limitations and specifications. Evaluation takes place when an AI demonstrates acceptable behavior within the parameters outlined above (humans must have ways to regain control if necessary). Second, if we take Igor Markov's perspective, we have no reason to be concerned for Igor's well-being. Finally, access to energy should be restricted after it was suggested that self-replication of both software and hardware, as well as self-repair and self-development, should be limited. It's inevitable that Intelligence will have serious shortcomings.

Let the computer do the work, and when the circumstances are uncertain, incorporate humans in the decision-making process or let them make the ultimate call. In conclusion, the development of artificial intelligence should center on the utilization of a framework known as composite intelligence. Instruct the computer to do the work, then bring humans in for questionable circumstances or ask them to make the ultimate decision. These are the two possible approaches to taking care of this matter. The other choice is to delegate the responsibilities to the computer. The primary distinction is that the first situation requires the data to be extremely dependable, but it gets things done much more quickly because it allows machines to assume responsibility of decision-

making (while still relying on humans for feedback). That is a significant point of distinction. Depending on what happens, the conclusion of this first chapter can be summed up as follows: the rise of artificial intelligence is on the horizon, but it won't happen as quickly as was anticipated. This spring of AI appears to be different from other stages of the cycle for a number of reasons, and we need to concentrate our resources and energies on developing AI that brings us to an optimistic future in order to make progress.

2.7 DATA STRUCTURES

However, before we start constructing concrete classes and objects, let's first look at some data structures. In this study, we examined a variety of the built-in data structures that are available in R. This section will describe the data structures that are already incorporated into the system. We did not describe how to create something more complicated with such data, but the data categories that are built into R serve as the foundation for creating more sophisticated structures on it. The concept of keeping pertinent data in one location so that it can be handled as a whole is more essential than any object-oriented system that has ever existed or could exist. When we are working on a sizable dataset that is connected in some way, we do not want the data to be dispersed across a variety of variables or perhaps different regions because this makes it extremely improbable that it will spread. You can preserve your consistency. It would be a laborious process to determine what times each variable corresponds to even if the dates were fixed and could not be altered.

This is going to be the situation regardless of whether or not the facts can be changed. As a result of this reason, the data that we examine is typically kept in a framework that is known as a data block. This is an essential strategy for ensuring that the confidentiality of the data is preserved. When we execute functions that use the data frame, we can be confident that all of the data is recovered accurately and continuously because we deal with all of the data in the same data frame. Because we analyze every piece of info in the same block at the same time. Consistent with data frames, at least as far as we are able to guarantee; despite the fact that we are unable to guarantee that the data itself has not been altered in any way, we are able to construct functions that are founded on the presumption that data blocks operate in a particular manner. What

about a customized silhouette that will make your figure look even better? When we apply a model to some data, the outcome is a collection of recorded variables.

These variables provide insight into the manner in which the model is utilized to analyze the data. As we continue to work on the model, we will do everything in our power to ensure that these variables remain grouped together. We have come to this conclusion because we do not wish to commit the gaffe of inadvertently using a combination of variables that corresponds to more than one model. We might want to retain information about what actually matches the model so that it can be used as information about what actually fits the patient in the event that we decide to evaluate it at a later time in the R environment. Alternately, the times that it was predetermined to take place in the first place. Since we don't have any other choice, the only way to bring together multiple pieces of information into one coherent whole is to use a list, which is also the only option we have. Using the R programming language in this manner is the correct way to carry it out.

2.8 EDITING THE LINEAR BAYES MODEL

The final section of the book is devoted to the discussion of the second endeavor, which centers on Bayesian linear models as the primary subject of discussion. We would place the data from a model into a collection in order to convey something like this. Imagine that we have a function that is comparable to the one explained in this article in order to guarantee that the data is presented in the appropriate format. Determine the value of the component by applying the algorithm that is spelled out as part of the model specification. Additionally, it takes into consideration the "resolution" of beta statistics as well as the "historical alpha accuracy." The next step is to figure out the mean as well as the correlation matrix of the model that best matches the data. The findings are shown after everything is done. The mathematical reasoning that underlies the code is broken down in great detail. After that, the person who is using the function will be given a summary that includes the fitted model, the algorithm that was used to fit the model, and the data that was used to fit the model (assuming they are in the mutable box state). This summary incorporates everything that was just discussed. Additionally, the procedure provides statistics that can be used when creating models.

2.9 POLYMORFIC FUNCTIONS

One illustration of a generalized function is the print operation, which prints out information. [Example:] This suggests that the behavior of the function on each call is determined by the class of the first parameter that you passed into the function. R first identifies the class of the object, which in this instance is blm, and then it looks to see if it can locate a function with the name print.blm. When we give R the instruction to print, it will first determine the class of the object, which will turn out to be blm, and then it will determine the class of the object. If it is feasible to do so, this procedure should be invoked with the same parameters as publishing. In the event that this cannot be achieved, this procedure will not be invoked. If that does not work, you can try running print.default directly instead. In the event that the search is fruitless, it will be continued.

There is no additional work required for you to create your own print function that is particular to a class. In order to generate your own instance of a class-specific polymorphic function, the procedure is as easy as described above. The only thing you need to do is take the name of the function, and append.classname to the end of it. If you create a function with this name, it will be called whenever a polymorphic function is called on a member object of that class. If you don't create a function with this name, it won't be called at all. It is not possible to invoke this procedure unless you first construct one with the same name. On the other hand, in terms of the design as a whole, you need to pay attention to the utilitarian interface. The user experience was intended to accommodate this as one of its many features. When I say this, I'm referring to the arguments that the function can take before beginning its task (and command). The minimal number of parameters that must be passed to a polymorphic function is established in advance for each function.

You can learn what they are by reading the documentation that is associated with the function in question. You have the ability to add numerous variables to the function when you create your custom function; however, it should be intended to handle at least the same parameters as the function as a whole. When working on the property description, you have the opportunity to select this alternative. This could lead to issues in the future if someone contacts your function with preconceived notions about what

parameters it requires based on the public interface, and then they come across a bespoke function that operates in a manner that is inconsistent with their expectations. If someone invokes the function while making conclusions about the parameters it requires based on the public interface, then there is a good chance that problems will emerge.

R won't produce any audible output unless you specifically set it up to do so; however, in the future when other people utilize the function, there will be issues. This shouldn't take place at all. It is strongly suggested that the implementation of the function be modified so that it accepts the same input as the global function. Among these are employing the same titles for each of the various process characteristics. It is possible to send named parameters to the public function; however, in order for this to work, the parameters you pass in must have the same titles as the parameters passed into the global function. It's possible that named parameters are being passed to the generic procedure by someone. As a consequence of this, we came to the realization that the number x would be the most appropriate selection for the input component of the print.blm function.

2.10 CLASS HIERARCHY

Object-oriented programming is distinguished by a number of characteristics, including inheritance and generalized functions, both of which are aspects of the former. Larry Alexander is credited with the creation of the object-oriented programming language. When developing more specialized classes as opposed to more general classes, this method serves as the beginning point for the development process. The concept that there can be multiple degrees of specialization within the same subject is the one that sheds the most light on this question. You have more general categories of products, such as furniture; within that larger category, you have more specific categories, such as chairs; within the chair category, you have even more specific kinds of products, such as kitchen chairs, dining room chairs, and eating chairs.

It's possible that a chair could be considered a piece of furniture, and it's also possible that a chair could be considered a piece of furniture. There is no question in my mind that you can achieve the same look with chairs as you can with other pieces of furniture.

For instance, certain pieces of furniture, such as a chair, can be used to ignite a fire; as an illustration, a fire can be started using a chair. However, not all of the things you can do with other types of furnishings are analogous to the things you can do with chairs. This is due to the fact that there are several distinctions between the two. You can't hurl a piano at an unwelcome visitor, but you can certainly pitch a chair in their direction. There are a number of different procedures that can be carried out on generic classes, and this is the mechanism behind how this kind of specialization works. These operations include: Any instance of these classes, including elements of other classes with a higher level of specialization, is able to carry out the aforementioned operations on them. This is true for instances of classes as well, even if those classes are themselves subjects of other classes.

Even though the operations don't accomplish the same thing in precisely the same way (for instance, on blm objects we can modify print, an operation that can be conducted on any object to achieve a different outcome), there are still helpful techniques that execute the can operation. . An operation is said to be class-specific when it is carried out by an instance of a class, even though it is possible for the operation to be carried out by all of the objects that belong to the public class. The distinction lies in the fact that the operation is carried out by the instance of the class in a manner that is more class-specific. On the other hand, specialized sections have the capacity to do more, which indicates that they might be eligible for procedures that are easier to apply to them. This is because they possess a greater variety of abilities. No trouble. If all objects of a specific class can be handled in the same manner as all objects of a more general class, then one can presume that the specialization in question is based on both the interface and the implementation.

2.11 EXPERTISE AS AN INTERFACE

An interface describes the different operations and routines that can be carried out on objects of a particular class. These operations and routines can be conducted on the objects. One way to think of it is as a procedure that regulates the way in which we communicate with objects that are members of the class. If we look at "fit models" as a more general category, we can state that we ought to be able to determine the appropriate parameters for any model and then forecast new values that are satisfactory

for any model. This is because "fit models" is a more general category. If we were developing a general category of "specific templates," then this would be the scenario that would play out. This is the type of thing that would fall under a more general categorization that we could term "modified models." It is a presumption that every model contains these kinds of functions. These functions can be found in R under the titles of coefficients and projections, and they are referred to by their own individual identities.

This demonstrates that I am able to generate code that communicates with a bespoke model by using the public model functions. As long as I adhere to the interface that is displayed to me, I am free to work on any model. If I subsequently determine that I want to transition from a linear regression model to a decision tree regression model, all I will need to do is bring in another appropriate model and speak to the same polymorphic functions. This will allow me to make the switch easily. Because of this, I will always be able to make the transition in the future from using a linear regression model to using a decision tree regression model. This choice is available to me as one of the alternatives to consider in the event that I decide to transfer. If I figure in and wait for generic functions, then the actual functions that will be called will, of course, be different; however, the interface that is used will not change at any point. R is not responsible for implementing such connections on its own initiative and is not responsible for doing so on its behalf.

Classes in C# are not created in the same way that they are in other programming languages such as Java. For instance, in the programming language Java, it is considered a type error to define something as an object that satisfies a particular interface when that object does not actually satisfy that interface. This is because the definition implies that the object satisfies the interface, but the object does not actually satisfy the interface. In R, the word for a class is spelled very differently than in other languages. AR doesn't care. Only in cases where you attempt to invoke a procedure that does not exist; if this is the case, you might run into some issues. On the other hand, it is your responsibility to implement an interface that corresponds to the kind of class or procedure that you believe your class should match. If you don't give your class an implementation of an interface, it won't function correctly.

If you implement functions that are expected of a particular interface (and those functions do something relatively comparable to what the interface expects even though they do not actually have the same name), then you have a specialization in that interface. In point of fact, the functions that it implements behave very similarly to what the interface anticipates those functions will behave like. This is due to the fact that the functions that it implements are exactly those that the interface anticipates those functions to be. It is possible for it to carry out the same duties as other classes that implement the interface; however, the actions that it carries out are obviously tailored to the present class that it belongs to. This is due to the fact that the activities you participate in are also completed by your present class.

It is completely appropriate for your class to have more possibilities than the object class that is more flexible; this is true even if you intend to add even more functionality to your class in the future. This remains the case even if you choose to bolster your character with additional abilities. Because other classes can now implement those operations as well, you now have a wider selection of classes from which to choose when you require something that is capable of performing more specialized operations. This results in the formation of a new category that covers a wider range of topics and has the capacity for further subcategorization of its content. There is a structure of classes, and classes are organized within that hierarchy according to the functions for which each class provides an implementation. Those functions are referred to as "implementations."

2.12 EXPERTISE IN IMPLEMENTATION

When it comes to object-oriented programming, the notion of specialization revolves heavily upon the concept of class hierarchies. In the case of R, this may be accomplished by giving implementations of polymorphic functions. Specialization can be accomplished in a number of different ways, including the provision of more generic or more particular interfaces. As a result, you are free to consider the many kinds of objects to be a single overarching class. This is possible due to the fact that class hierarchies have an additional aspect to take into consideration, and that value is the rate at which code is repeated. You may develop code that functions in a common interface and then reuse it for all of the objects that implement that interface. Of fact,

you already gain the majority of these advantages if you provide interfaces for dealing with objects since you can write code that functions in that interface.

The provision of new interfaces for the manipulation of objects, on the other hand, enables a rise in the value of this advantage. On the other hand, by offering interfaces for the manipulation of things, you may get an even greater amount of value from them. You may achieve a degree of reusability by developing a class hierarchy that progresses from more generic and abstract classes to classes that are, on the other hand, more specific and concrete. You will also be subject to yet another form of fee in addition to this one already mentioned. When you specialize in a class, you produce a new class that, with a few tweaks here and there, implements the same interface as the parent class you choose to specialize in.

This enables you to apply previously developed functionality to a class that is more general, and then utilize that functionality to develop a private version of the class. When you alter a class in this manner, you should avoid implementing new versions of polymorphic methods that are a part of the class's interface. Instead, you should stick to updating the existing versions of such methods. Instead, the current iterations of these procedures should be simply replaced. A sizeable portion of them will exhibit behavior that is identical to the implementation of the most generic class to which they belong, and this similarity will be rather widespread.

This is because we haven't adequately communicated to R that we consider class B to be a more specialized implementation of class A. This is the reason why this is the case. Therefore, we find ourselves in this predicament. We have written the "native code" for the Object() function, which is also referred to as the B function. This is the part of the function that is responsible for the actual creation of the object, and as a result, every B object also includes the data that is there. To do this, the Object() method was renamed to be similar to the B function. Our objective of making it possible for B objects to be manipulated in the same way as A objects using R was never accomplished.

We are able to ensure that using foo on object B will have the exact same outcomes as invoking foo on object A. This is feasible due to the fact that we are able to configure

foo.A to call foo. This will be the approach with the least amount of complexity. A. If there are several polymorphic functions that may be applied to objects of type A, then we need to construct version B of each of those functions. It does not take a lot of effort to develop a single function, but the amount of work that needs to be done rises dramatically when there are several functions to implement. Not only is it taxing on one's physical capabilities, but it also offers little room for error. If only there were some way to convey to R that class B is really just an expanded version of class A! As well as that. It is possible to use a string in lieu of the value of the class attribute of an object, however this is not needed.

It may take the form of a vector containing string values. If we first declare object class B as B, then define it again as A, we get the following result: The system is negatively affected by your activities. When these vectors reach a certain level of complexity, it may become challenging to comprehend how this process works, what functions are called in the order of the class names in the vector, and what code is really being carried out.

There is not a true class hierarchy here in the same sense that there is in programming languages such as Python, C++, or Java. The only method that we are aware of for assigning polymorphic functions is to call them by appending the names of the classes that are present in the class attribute vector. It's the only option we have at this point. It's likely that one of your objectives is to communicate to R the existence of a class hierarchy in which class A is the most generic, class B is a subtype of that class, and class C is the subtype with the greatest level of specialization. On the other hand, it is not what you are conveying to R at this time. Because you can't. You instruct R on how to locate dynamic functions, and it will then utilize the code you provide to locate the functions that you instruct it to locate.

Immediately after the completion of the C functions. Due to the fact that you have not supplied any more information on the operation of B's normal functions, this will need some investigation. It is unclear to her how she would want it to appear in this form. If the idea of providing indexing as a choice strikes you as unusual, it is essential that you understand that everything in the R programming language revolves on invoking functions. This is the single most critical concept to have a grasp on. The process of

indexing an array using a function is identical to the process of calling any other function, and functions may accept named parameters. At this point, there is nothing further that can be included in this discussion. When applied to a list in order to execute a subset operation, the operation will always result in the production of another list.

If the information that has been provided here has caught you off guard, keep in mind that whenever you sign up for a carrier, you also receive a carrier back. This is the case regardless of the circumstances. You don't give it much thought since we have a method that we typically interpret individual values as vectors of a length, and because of this, we are more used to thinking about things in this manner. As a direct consequence of this, you don't give it any thought. When applied on a list in any of these methods, the subset operation will always result in a different list from the original list. Even if you just generate a portion of an object, you won't end up with just that portion of the object; rather, you'll end up with a list that has just that one item. Even if you merely produce a little portion of an article, this statement will still be valid.

2.13 CONTROL STUDY

A program's control structures determine the order in which its instructions are executed during its execution. You can go a long way just by putting together a set of statements or expressions, but at some point, the results of a calculation will have to deviate from what you intended, at which point control structures come into play. . You can go very far by simply putting together a series of phrases or expressions. The two most common types of control structures in R are loops and selection statements, also known as if statements. This is similar to the situation in many other programming languages (for, while or repeat statements).

2.14 A CAUTION ABOUT THE CYCLE

If you keep reading about R, ultimately, you'll come across the claim that the looping operations in R are so painfully sluggish that you won't be able to stand it. You absolutely must get yourself ready for this. Even though the situation is not as dire as some would have us believe it is, the condemnation is partially warranted because it is true. R is a dynamic programming language, which means that the interpreter cannot optimize the code before the program is executed. This is in contrast to other

programming languages. R is a programming language, which is the reason for this. This is due to the fact that the functions and variables of the program are subject to change at any moment while it is being executed (but not unlike other dynamic languages like Python). However, there have not been many attempts made to maximize for loops because, in the majority of instances, there are superior alternatives accessible in R than making use of an explicit loop expression.

As a result, the results of these endeavors have not been very significant. Along with its other capabilities, R is what's known as a functional language, and in general, functional languages don't make substantial use of loops. However, R has this capability, along with a number of others. During the course of the code's processing, the value of any loop variables or Conditional expressions will, of necessity, change. This is because of the way loop structures are constructed. This is the situation regardless of whether or not the code is functional. Although it is not forbidden in functional programming languages, changing the meanings of variables is regarded as "impure" (so clearly R is not a pure functional language).

In place of the more common technique of using loops, recursive functions are utilized here. Even the most fundamental repeating structures are never included in purely functional languages, which means that the overwhelming majority of functional languages are completely devoid of them. Even though R is a very functional language in comparison to others, you will almost always get superior results from your work if you avoid using loops whenever it is feasible to do so. In the following paragraph, we will examine this matter even further.

In the setting of function f, where the g function has been specified but the values of y and z have not yet been described, sets a reference to the variables y and z. But this is exactly the point: our only responsibility is to build the g function, but we don't even give it a name. The next thing we do when calling f is to assign the number y to the variable y, and then we return g to the procedure that called us. The number that is returned by the function is indicated by the symbol h when it is taken out of the context of the function call. If we choose to use h here, it is important to keep in mind that y is specified in f, as well as the value of y after we have returned from f. This occurs if we continue to use h in its current form.

In point of fact, if we do reach this position, it is important to keep in mind that we have recently arrived here from point f. This is the explanation for it. Because az has not yet been given a number to which it can be allocated, getting an error whenever you try to call ah is guaranteed to happen. Since the inner function is constrained and is not specified in any of the neighboring contexts, it would make sense to define z in the global scope at the conclusion of the program; however, z is not defined here. If we go ahead and introduce it there, the issue will be resolved because R will be able to discover the variable in a location that is separate from the location where the function was initially established. If we follow these steps, we will be able to resolve the issue.

Because of the potentially harmful nature of the function I am about to describe, I really shouldn't be providing you with this information because I am about to disclose it. I'm going to show you how you can make characteristics have a greater influence than they do now, despite the fact that traits aren't supposed to have any negative side effects. It is not appropriate for functions to have any kind of secondary consequences. In either case, it is a feature of the language and can be very beneficial if you discover that features have side effects as long as you use them very cautiously. However, if you find that features do not have side effects, then you should avoid using them. However, if you are not cautious, it can pose a significant risk to your safety. On the other hand, if you are not cautious, it has the potential to cause a great deal of harm. Consider the following hypothetical situation in order to get to the meat of the issue:

You want to allocate a variable that is in a different scope than the function that you want to assign it to. When you make an effort to do this, what do you find? You cannot allocate a variable in such a way that will result in the creation of that variable within the context of the present scope. You won't need to allocate a variable because of this. As a consequence of this, it will not be feasible to explicitly designate a number to the property. This section of text includes a getter and setting function, both of which you should study very attentively because they contain some interesting concepts that you might find helpful in the course of your work.

The value of the variable x can be changed using the setting technique, while the getter function provides information about the value of the variable at the present time. On the other hand, declaring x inside the body of the set function causes a local variable to

be defined inside of that function; a step is not allocated to x in the preceding statement. To successfully carry out this function, you will need to make use of the unique assignment operator that was developed specifically for this objective. It does not generate any new local variables; rather, it looks beyond the confines of the program for an existing variable to which it can allocate a value and then searches for that variable. However, if it is possible to reach the absolute boundaries of the global scope, it will generate the variable there if it doesn't already exist if it has already been used. This occurs only if the variable has been used previously. When we employ the assignment operator, we get the behavior that we want, as demonstrated in the previous example.

2.15 BINARY SEARCH

One of the methods that can be utilized to ascertain whether or not an element is present in an orderly collection is the time-honored strategy that is known as binary search. Since we're being completely honest, it's nothing more than a basic iterative function. The most basic example involves a chain that consists of just one piece. This is the circumstance that arises most frequently. You will then have the opportunity to make a straight comparison between the item you are searching for and the item in the collection, giving you the ability to determine whether or not the two items are the same. In other words, it is possible to determine whether or not two things are the same.

If the collection consists of numerous components, the component that should be chosen is the one that is located in the center of the list. If this is the object that you are searching for, then the work that you came here to do has been completed, and you can let us know that the item can be posted. In the event that this is not the document that you were searching for, your work is not yet finished here. If it is less than the item you are looking for, then you know it must be in the penultimate half of the array if it is in the list, and you can search there. If it is greater than the item you are looking for, then you can search anywhere in the array. If it is bigger than the thing you are searching for, you won't know where to start looking for it.

If this is the case, there is a good chance that the component you are searching for is included on the list. If it's bigger than the element it's looking for, it understands it must

be in the first half of the array if it's in the array, so it searches iteratively there. If it's smaller than the element it's looking for, it explores the second half of the array. If it is not more comprehensive than the article you are seeking, you will not be able to determine whether or not it is part of the series. If it is lower than the item you are looking for, then you can be fairly certain that it is not a component of the collection; however, if it is higher than the item you are looking for, then you cannot be certain. Iterative inquiries will need to be carried out while moving through a sub stream in order for you to successfully accomplish this work as outlined. Therefore, it is necessary to reproduce this substring in order to successfully carry out the function call, which results in the implementation being significantly less effective than it should be. You should give it your best effort to complete binary search without using it.

2.16 SELECTING THE SMALLEST WED ELEMENT

If you have n items and you want to get the smallest k items, an easy way is to sort the items and then select the kth item from the list. This will give you the smallest element. While this approach is effective and can be completed in a reasonable amount of time in the vast majority of cases, it is possible to get the job done in less time. Therefore, it is not necessary for us to classify items in any precise order; Instead, we just need to move the smallest sized element to position k in the array. This allows us to complete the task. This problem can be solved by making some changes to the method called Quicksort used in the previous exercise. When we divide an array into less than, equal, and greater than the pivot, we always recursively sort the smallest and largest items. This ensures that the small items are always sorted before the large items.

If our only goal is to find the item that will eventually be sorted at position k in sorted lists, there is no need to sort the non-overlapping array in that index; in this case, we are only concerned with the location of the article. If the array does not overlap this index, there is no need for sorting. If there are mok elements in our collection that have a lower value than the pivot, we don't need to sort them before adding them to the array; Instead, we can put them at the top of the list. We need them there but we don't have to sort them because we only care that the smallest component is placed in the correct index. There is no need to rank the main components that are the same as each other. If we sort them, they all end up with indices greater than k, but we don't really care

where they go because we know that's where they end up. If there are m items of k size less than the pivot and l items equal to the pivot, then the smallest item of size k is equal to the pivot and we can return that value. Apply this algorithm to the situation and see what happens.

CHAPTER 3

VISUALIZING DATA

Nothing tells a compelling story about your data like a good plot. Charts are a much more effective way of visualizing data than summary statistics because they often reveal aspects of data that you cannot distinguish from mere summaries. Summary statistics are a great way to get an overview of your data, but graphs are even more effective. R has very advanced data visualization capabilities. Unfortunately, it actually has more tools than you can use properly. This is not a happy aspect of the product. You can graph data with different frames; However, these frameworks are often not very compatible with each other, making it difficult to combine many methods. In this section, we'll explore the graphics capabilities R offers.

While it's physically impossible for us to review all of the many graphics capabilities available, I'll focus on a few of the different frameworks. First, the main components of the graphics architecture. It's something I don't use often and I wouldn't recommend using it either, but it's important to know that it exists because it's the default for many programs. In the second part of this tutorial, we will introduce you to the ggplot2 framework, which I believe is the data visualization format that gives you the best results. It provides a small domain specific language for generating data and is great for examining data as long as it exists in a data block, which is very important. If you want to explore data, you need to have it in a data frame (and it takes a little more work to create publishable graphs).

3.1 BASIC CARDS

The main drawing system application is included in the graphics package. In most cases it is not necessary to add the package:

This list is not exhaustive, as the basic plot function, plot(), is generic, and many packages write extensions to customize plotting. Therefore, this list does not include all possible plot features. However, if you want to create simple graphs, you can use plot() . Because this function is called a generic function, the behavior it produces

depends entirely on the input passed to it. This feature can be used to create a wide variety of results. You can then give it a number of initial inputs to get graphics of a variety of different objects. A scatter chart is the simplest type of chart to make, showing the dot times x and y values (see Figure 3.1 for an example).

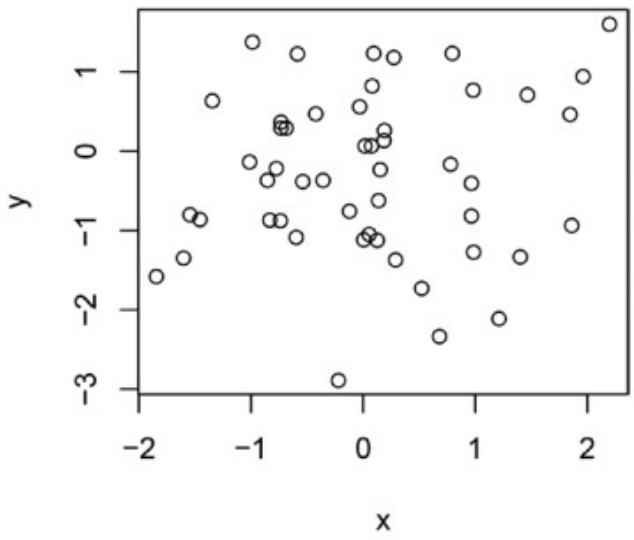


Figure 3.1. A typical point cloud

You can plot data from a data frame using the plot() function, which takes a data argument. However, you can't write code like this to display automatic data inside the dataset package: although plot() comes with the data frame, it doesn't know which variables to plot. x and y parameters. These parameters are x and y. It's important to use the %$% operator to plot() access to variables in a data frame so you can add charts to pipelines. Therefore, adding packages to pipelines requires using the %$% operator. For example, we can display car data as follows:

The data argument of the plot() function is passed when plot variables are specified as a formula. Not used on its own; In contrast, the plot() method's data parameter is used after it has been combined with a formula to perform its intended function. If x and y values are specified in a formula, you can provide the function with a data frame that already contains the variables and then display the data accordingly. If x and y values are not specified, the function automatically generates them. Here's how: By default,

the chart shows data as points; However, you have the option of passing a type parameter (see Figure 3.2) for an example (see also) to display data in various other formats, such as lines or histograms.

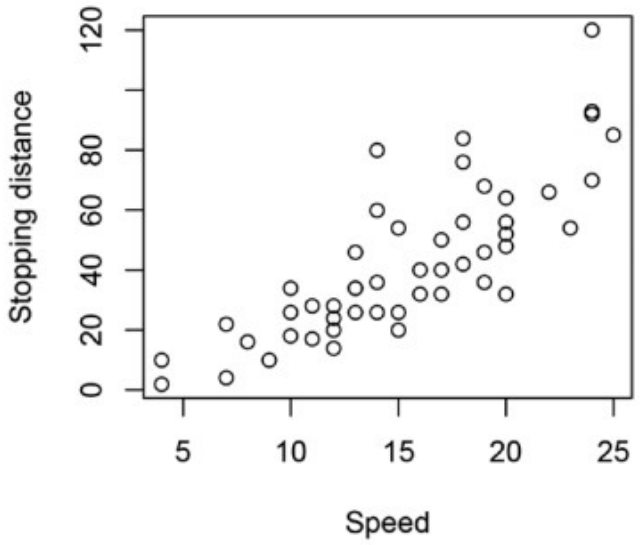


Figure 3.2. Speed and distance point cloud for cars

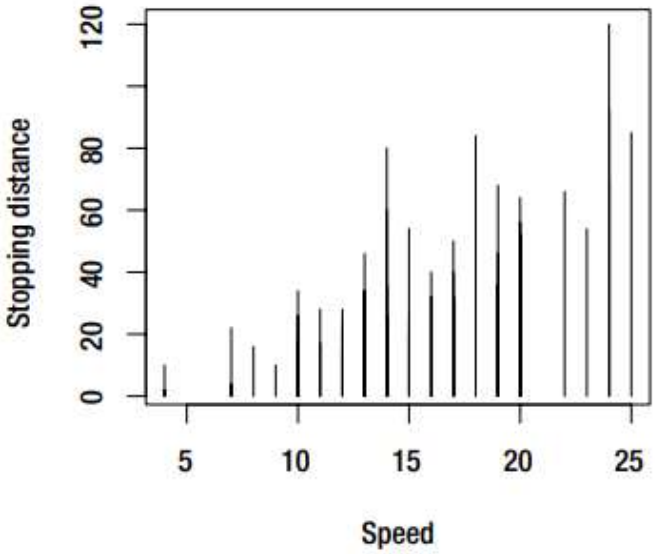


Figure 3.3. Speed and distance histogram for cars

3.2 GRAPHS AND GGPLOT2 PACKAGE

ggplot2 software offers an alternative to standard plotting; This alternative is based on "graph grammar", a term coined by package developers. The rationale behind this is based on the idea that the system provides a narrow domain-specific vocabulary that can be used to create graphs (just as dplyr provides a domain-specific language for manipulating chunks of data). Charts are created using a list of function calls, just as you would create simple charts; However, these function calls do not immediately write to a canvas independently of each other. Instead, diagrams are created with a list of function calls as building blocks. Instead, everyone manipulates a chart in two ways: by scaling the axes, dividing the data into subsets that are displayed in different ways, or adding levels of visualization to the chart. Each of these methods is described in more detail below.

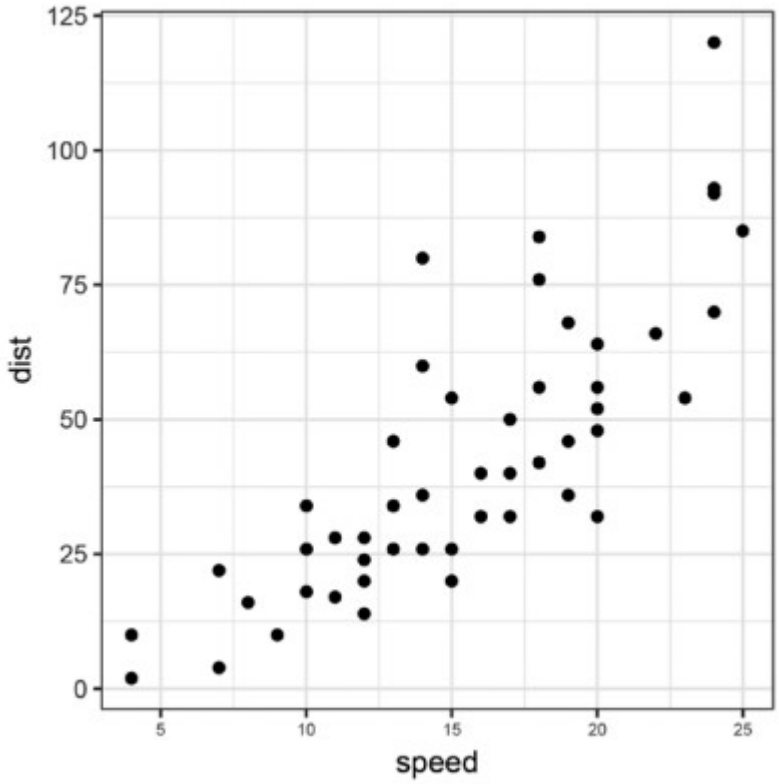


Figure 3.4. Plot the vehicle data with qplot (ggplot2).

You want to broaden your experience in this field. After completing this section, you should be able to create simple diagrams, and if you search the Web for information on creating more complex diagrams, you should be able to find information on how to do it. As a first step in our ggplot2 research, let's start by introducing the qplot() function (which stands for Quickplot). This function performs operations very similar to plot() with a few minor differences; However, it creates the same objects that other ggplot2 functions work with , so it can be used with these functions.

However, the events that unfold are very different. Instead of creating a directed graph, the qplot() function actually creates a new ggplot object. This is done instead of the traditional technique. Printing such objects will cause them to be printed as an undesirable result of the printing process. It might not make much sense, but that's how things really work. Because the function used to print R objects is a generic function, the effect of printing an object depends on the implementation of the print() function offered by the object. In fact, the function used to print R objects is itself a generic function. If the object is already an instance of ggplot, this function plots it for you. However, it works fine with the type of code we wrote, because in the above code, the result of the entire expression is the value returned by qplot(), and when evaluated at the top level in the R prompt, the result is output. Evaluating the above code, the result of the entire expression was actually the return value from qplot(). This indicates that it is compatible with the type of code we have created and works effectively . That's why the ggplot object is plotted. The code above and this one share the same features.

Since the print() function is called automatically when an expression is evaluated, we use it to evaluate an expression at the R prompt instead of the plot() function, which would be the obvious choice in many other contexts. This is because plot() is the obvious choice in many other contexts. Using the print() method allows us to avoid manually printing objects; All we have to do is place the tracking code at the highest possible level of the program. Conversely, when you create a chart within a function, the final product is not created automatically; instead, you are responsible for printing it yourself in accordance with the instructions provided. I'm talking about all these details about objects being created and printed because the typical workflow for using ggplot2 is to create a ggplot object, perform various actions on it to modify it, and

finally print it. The reason I'm mentioning all these details about created and printed objects is because of this typical workflow.

That's why I included all these details. When you use the qplot() function, many changes are made to the drawn object before qplot() gives you control of the object. qplot() does this so you have full control over the object. The term "quick plot" refers to the process where qplot() first makes an educated guess about the type of plot you most likely want, and then applies changes to a plot to produce that type of plot. The term "quick" refers to how fast this process happens, hence the name "quick pull". To achieve our goal of having full control over the final plot, we avoid using qplot() and instead manually make the necessary changes. qplot() is no longer a function I use;

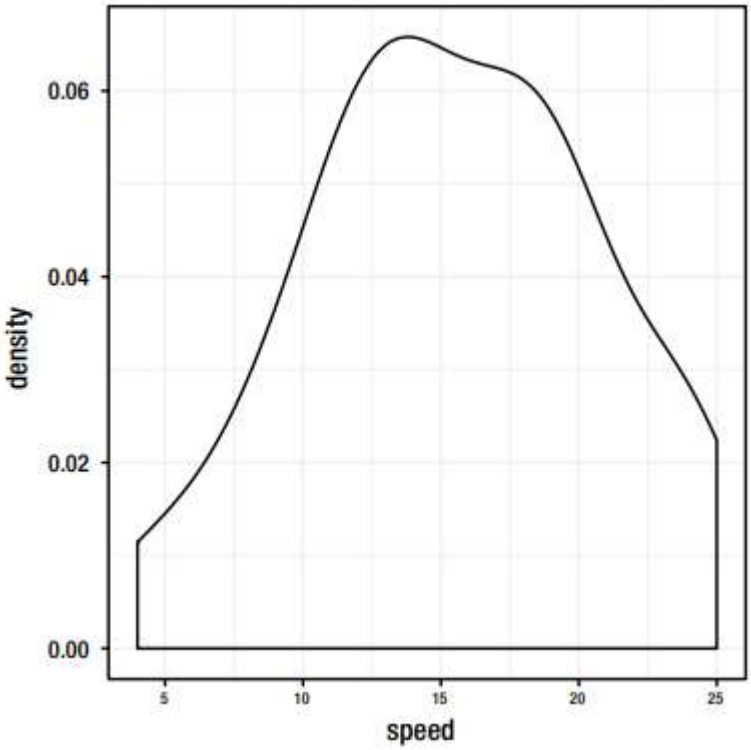


Figure 3.5. Velocity density of the car created with qplot (ggplot2)

However, if you're new to ggplot2 and want to get familiar with it, using qplot() isn't a bad idea and might even be a good idea. Using qplot() it is possible to make the representation of data points dependent on data variables much more directly and

simply than is possible in a graph. This is the opposite of story skills. To color the iris data based on the Types column of the chart, we first needed to do a Mapping(). So we had to make changes to the Types column to preserve the colors. When using the qplot() function, it is sufficient to specify that the colors must be type variable. This can be seen in Figure 3.4.

By defining geometry, you can create line, boxplot, fiddle drawing, and other chart shapes like other chart types. Depending on the geometry, the underlying data should be properly displayed, this is what decides. They may include showing raw data, as they do when we create a scatterplot, or they may include providing summary statistics, as they do when we create a histogram or heat chart; However, they all show how the data should look. When creating a visual with ggplot2, it is often necessary to use more geometry as part of the process of adding geometry to the data. But to know how to achieve this, we first need to exit qplot() and examine how plots previously created using qplot() can be copied using geometries. This allows us to learn how it's done.

3.3 USE OF GEOMETRIES

By chaining several geometry instructions together, it is possible to display the same data in different ways. For example, scatter charts with smooth lines can be displayed together. You can even include data from multiple different sources in a single chart. But before we move on to more complex structures, let's see how qplot() plots can be created by manually calling geometry methods. This is done before more complex structures appear.

An object is created after using the ggplot() function. Automotive information is one of the inputs we provide to you. After the data frame is assigned to this object, subsequent operations access the data. In practice this will be the case. It is possible to change the data frame from which the data we plot comes from, but until something else is provided, we have access to the data we supplied to ggplot() when we trained the first object. However, it is possible to change the data frame from which the data we represent comes from. On the other hand, we have access to all data, unless stated otherwise. The next step is to execute two different procedures within the confines of a single function call. Using the geom point() function, we can tell the program that we

want to plot x and y values as points. Next, we'll map speed to x values and distance to y values using the "aesthetic" function, abbreviated as aes. Finally we publish the result.

Assigning data to images is the responsibility of the aesthetics department. Using the geom point() geometry requires that the graph contain data not only for x but also for y. The function takes the instructions from aesthetics on which data variables to use in each case. The data-graph mapping specification found in the aes() function only takes effect when applied to the geom point() function. Sometimes we don't want to have separate mappings for different geometries, but there are cases where we want to have such mappings. If we want to reuse cosmetic parameters in multiple methods, we can do so with the ggplot() function instead of setting it for each function separately. Then the subsequent methods can access it as well as the data and we don't need to define it continuously in subsequent procedure calls. In fact, they are treated in the same way as data. The ggplot() and geom point() procedures can be combined into a single function using the plus operator. You can update a ggplot object by combining a series of plus (+) expressions. It works very similarly to how we chain a series of data manipulations using the percent sign (%>%).

The only reason these two operators behave differently in this environment is due to historical circumstances; If the %>% operator had been used generically when writing ggplot2, it would certainly have been included. In this case, you must use the plus sign. Because the plus operator behaves slightly differently in ggplot2 and magrittr, you cannot use the function name if the function is not taking an input argument. So you need to make sure the dot geom has the appropriate brackets. Since ggplot expects a data frame as its first argument, a common pattern is to first manipulate the data with a series of %>% operations, then pass it to ggplot, and finally continue with a series of + operations. This is because the first argument expected by ggplot is a data frame. This is possible because the ggplot function requires a data frame as its first input. If we could do this with cars, we would have this basic channel; however, more complex applications often have additional operations involved in both chart composition and pipeline.

In this particular scenario, we also had the option to implement the aesthetic using the ggplot() function instead of the geom point() technique. It was one of our available

options. If you are presenting the color cosmetically, you have the option of making it dependent on another element in the data. On the other hand, if you want a color code or other graphical parameter, you usually just need to move the parameter mapping to a location that isn't inside the aes() method. This allows you to encode the parameter. If the geom point() method takes a color parameter, it uses that color for the points; If no color argument is given, it gets its color from aesthetics. If you do not specify a color parameter, the color defined in the aesthetic is used. See Figure 3.5.

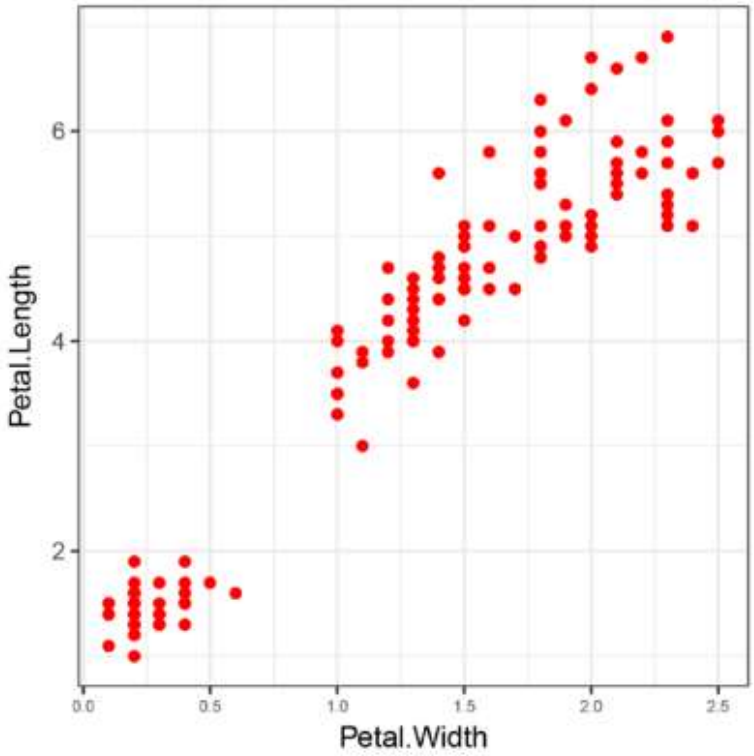


Figure 3.6. Iris data where the spot color is wired

Combining various shapes in different configurations allows you to view your data in different ways. This is not always useful and its relevance depends on the method of summarizing the data; The combination of scatter charts and histograms probably won't do you much good. The significance of this result depends on how the data are summarized. On the other hand, it is also possible to create a graph showing the car's speed both as a histogram and as a density (see Figure 3.6 for an example).

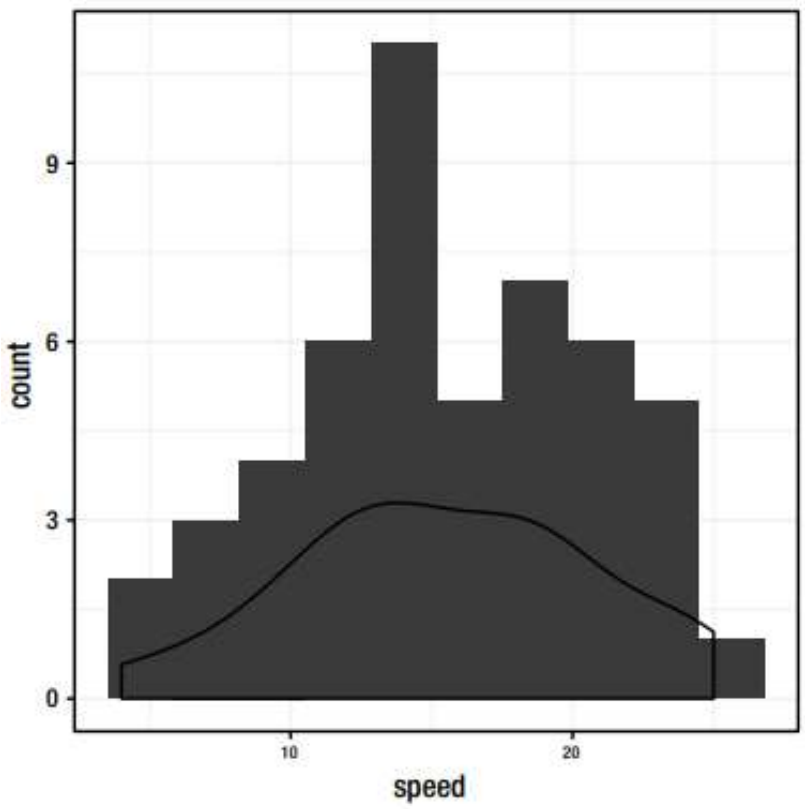


Figure 3.7. Combined histogram and density plot for vehicle data rate

You just need to verify that the geom histom() and geom density() functions are used. We also need to add one more aesthetic option for the y value. By default, histograms report the number of observations on the y-axis that fit in a bin, while intensities fit in a bin. This is because histograms provide the number of observations that fit in a box. When you tell them that y must equal "count", you're telling both geometries to use numbers on the y-axis. y must be equal to "count". You can use the equation y =..density... to get the value of y by replacing y with the density. Additionally, you can get an overview of data statistics by combining features with a point cloud. In the scatterplot created earlier for the vehicle data, we added the result of a linear fit performed on the data. This fit was performed after first performing a linear fit on the data(). As can be seen in Figure 3.7, all it takes to achieve the same result as ggplot2 is to add a call to the geom smooth() method.

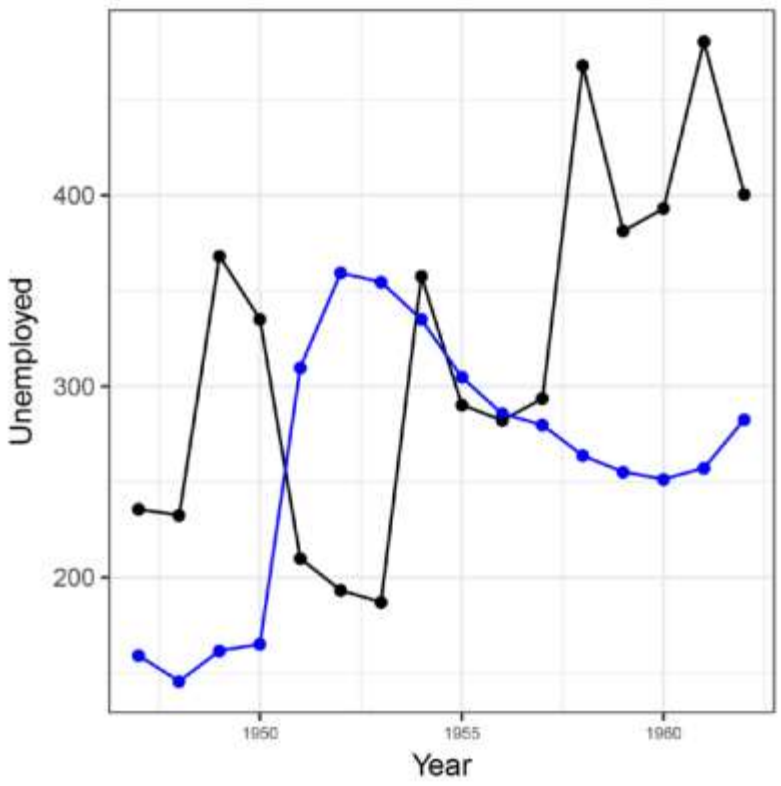


Figure 3.8. Longley data plotted with ggplot2 using points and lines

It's true that graphing two variables with different styles is a strategy that works well for the vast majority of applications, but that doesn't mean it's always the most efficient approach to accomplishing a particular task. The problem is that we present the two indicators of unemployment and military in the same way. These are the two actions in question. The term strengths refers to two separate metrics we collect each year that we can combine into a single visual representation using a custom tracking algorithm. Nothing in the data indicates that it can be used as a basis for our calculations. If we wanted to monitor two metrics separately rather than in the same frame, we would have to generate additional code for the tracking technique.

Refactoring the data frame to have a single column that tells us whether an observation is armed or empty is an approach that will be more efficient than any other. After applying the forces and other usage values, the color was adjusted to match the first

column and the y-axis was adjusted to match the other column. You can do this with the Collect() function, which is part of the Tiddr package, as shown in Figure 3.8. Make sure you have the appropriate permissions to do this.

3.4 SCALE

Scale and geometry are very important factors to consider when deciding how to graph data. Geometries tell ggplot2 how to transform data into visual components such as points or densities, and scales tell ggplot2 how to plot dimensions on the graph. ggplot2 is a drawing tool developed by Google. As usual, the x and y axes, with values mapped to positions on the chart, are the simplest imaginable scales. However, scales can also be applied to observable properties such as color. Using scales simply consists of placing labels on the axes of the scale. If your only concern is to tag things, you can achieve these using methods also available in this scenario. On the other hand, if you look closely at this chart, you will see that the scales are used differently. Point cloud labels with vehicles must be written to identify them.

3.5 THEMES AND OTHER GRAPHIC CONVERSIONS

Working with ggplot2 is mostly about configuring its geometries and scales to exercise control over how data is graphed. However, you can also control the visual presentation of the final graph using functions that only deal with the visual representation of the final result. Much of this is achieved through changes called themes. If you decide to try the examples I have provided in this chapter for yourself, you may find that your results do not match those described in this book. This is because I have assigned a default theme to the book with the following command: The actual appearance of the images on this page is generated by the theme function. This role must be accredited for your job. You can use the plus sign (+) to theme a drawing like with any other ggplot2 setting, or set it as default as I did in this example. In either case, you can use the plus (plus) sign to theme a graphic.

You have access to a wide variety of themes; To use them, see the documentation for functions that start with the theme_ prefix. All this can be changed and gives the player more control over the plot. Beyond themes, there are a number of features that can affect a property's architecture. Themes are one such feature. He's compiled

information to convey everything that could be related to themes and visual transformations in one post, but can he show you an example to give you an idea of what could be saved for a better record? Idea of what it could be. For example, you can change the coordinate system using one of the many coord_ functions available. Both x and y coordinates can be changed with minimal effort using the coordinate. However, there are some situations where reversing the coordinates of a complex diagram may be a more practical alternative than updating the aesthetics in several places.

Of course, this can also be achieved simply by making aesthetic changes in the environment. You may need to make some changes to the iris chart axes we reviewed earlier. If I just reverse the coordinates, suggesting a free y-axis to the faceting method, the axis labels will be wrong on the new x-axis. This is because the labels are relative to the old x-axis. This is because I am using a free y-axis. After translating the coordinates, we can only see the values of one of the y-axes. Because if you had an empty y-axis, you would have unique ranges for y-values, which is exactly what we want. this is the reason. Even so, the remaining numbers appear to be on the same axis. That won't happen. As a direct result of this, the . You can try it and see what results you get.

3.6 MULTI-GRAPHIC FIGURES

While models are useful in many situations where you want to have multiple fields on the same chart, this is not always the case. In other cases it may not be necessary to use a coating. You can use directions to accomplish this task when you want to show different subsets of data in separate panes while keeping essentially the same graph for each subset. Sometimes you want to combine different types of tables or charts created from different datasets and display them as secondary charts on different dashboards. To do this, you need to combine packages that are usually separate from each other. The ggplot2 package does not support merging multiple plots directly, but you can do this with the grid, the plotting system used below the surface. When working with a simple grid you have access to many low-level tools for manipulating the graphics; However, if you want to combine packages, you want additional high-end functionality that you can purchase in the GridExtra package.

As we can see, artificial intelligence is a broad field that encompasses a wide variety of technologies in its different applications. However, in this section, we will try to create a new visual form that can capture all technologies that are considered extremely important related to artificial intelligence. How they can be classified varies greatly; However, how they can be classified is very different. Working with strategic innovation company Axilo, our team set out to create a visual tool that allows users to quickly understand the breadth and depth of this toolkit. We've also laid the groundwork for a map to help users navigate the AI wasteland. If you're looking for a way to organize unstructured information into some kind of ontology, see the table below. The ultimate goal is not to accurately represent all the information currently available in AI; Rather, it's about having a tool that can identify and access some of that information.

The ultimate goal is not to accurately represent all the information currently available in AI. So it's about designing an architecture to access AI information and follow the emerging dynamics, designing a gateway that opens to pre-existing information on the subject and allows you to search for additional information and potentially generate new AI information. In other words, it's about designing an architecture to access AI information and monitor emerging dynamics. The figure below, shown in Figure 4.1, is an attempt to draw a building. Therefore, the usefulness of the final product will help you achieve the following three goals: to have a map to understand what is happening and to trace the path; Understand where AI is used today (compared to where it wasn't used in the past); and understanding what problems need to be reformulated and how many problems need to be reformulated so that AI can solve them (if you know the work of Agrawal et al. 2018, these are direct consequences of the falling cost of predictive technologies).

Now other than that, let's talk about Artificial Intelligence Knowledge Map (AIKM) to Keep You Moving. On the axis you will find two macro groups called AI problem areas and AI paradigms, respectively. Both groups are associated with artificial intelligence. Artificial intelligence or X-axis paradigms are simply alternative approaches to the problems scientists are trying to solve. This idea was developed by Pedro Domingos (2015) and subdivides researchers according to the procedures they used in developing

their work. You can easily map these five tribes using our paradigm classification (ignoring the bodily intelligence group) , i.e. with a logical approach, H. Symbolists (they use logical reasoning based on abstract symbols); Machine learning and connectionists (inspired by mammalian brains); Scalable with search and optimization (inspired by Darwin evolution); Bayesian methods with probabilistic methods (they use probabilistic models); and with probabilistic methods Bayes (they use

The patterns on the boxes do not classify technologies into one category; rather, they divide technologies into two different groups: those with narrow applications and those with broad applications. The language used was chosen to be somewhat misleading; But if you'll tolerate me a little longer, I'll explain what I mean by these words. Anyone new to AI should have a really solid understanding of the differences between weak/narrow AI (ANI), strong/general AI (I), and super AI (ASI). For the sake of clarity, artificial general intelligence (AI) is the ultimate goal and holy grail of science, whereas today we have really limited artificial intelligence, a set of technologies that can't handle anything.

Application scopes (this is the main difference from AGI). The two different types of lines (solid and dashed) used in the chart are to clarify this difference and give you peace of mind that you won't be completely confused by reading more background literature on AI. They do this by emphasizing the difference between the two types of lines on the chart. But at the same time, this distinction highlights technologies that can only solve a specific task (better than humans: narrow applications in general) and other technologies that can solve more tasks and interact with the world (better than many humans - general applications).

These two types of technologies are examples of applications in their respective fields. Other than that, let's start by looking at the graph itself. The world map shows the various sub-studies that fall under the scope of artificial intelligence. I would like to point out that I deliberately avoided naming certain algorithms to assign them to broader topics. Nor will I make an assessment of what it does or does not do, by only listing resources that academics and data scientists can use. Now the problem is to read and understand the map. To help you with this, I would like to give two examples.

Natural language processing is a family of algorithms that use a combination of knowledge-based techniques, machine learning, and probabilistic approaches when trying to solve perceptual problems. These algorithms are part of the natural language processing discipline. If you take a look, you'll see that these approaches are often used together. But if you look at the empty space where logic paradigms and inference problems intersect, you might wonder why there isn't technology there. The lesson to be drawn from the map is certainly not that there are no methods that can fill this space, but that people choose an approach like machine learning when dealing with a thinking problem. This is the message that can be retrieved from the card.

Far from being incompatible, many of these technologies complement each other, and you may choose to use one or more strategies to solve a particular problem. This indicates that you have the opportunity to implement one or more technological solutions. Finally, there is another important categorization (that is, many types of analysis) that I have not included in the table above, but it is important enough to deserve full mention. The classification I am talking about is the classification of many types of analysis. You may encounter not one, but two, but five different types of analytics: descriptive analytics (what happened), diagnostic analytics (why something happened), predictive analytics (what's going to happen), normative analytics (recommended actions), and automated analytics (automatically taking action). performs). You may also want to use this to classify the technologies described above in some way; In reality, however, this classification is more of a functional classification and process classification than a product classification. In other words, any spectral technology is capable of performing these five analytical functions.

After concluding the data analysis discussion in this chapter, we move on to programming and software development, which will be the focus of later chapters in this book. The first chapter provides a pedagogical introduction to R programming; However, this entry ignores a significant number of issues. This section covers most of these details; However, the more complex aspects of R programming, such as functional programming and object-oriented programming, are covered in the following sections.

To begin this new chapter, we'll look at a few sentences. Evaluating expressions is the foundation of everything we do in R. Most expressions are evaluated to perform a calculation and get the result. However, some expressions, such as B. Assignments, have side effects, and we often parse these expressions for side effects.

In each of these cases, x and y can be real numbers, variables that refer to numbers (technically called numeric vectors because R always works with vectors), or other expressions that evaluate to real numbers. Real numbers are the most common value types for x and y. Another possibility is variables that express numbers. When you create expressions with additional operators, you are subject to the same precedence rules you know from math. These policies determine the order in which attachment operators are evaluated. For example, exponentiation comes before multiplication, which also comes before addition. The order of operations is as follows: This suggests using parentheses when you need to evaluate sentences in a different order. Since you already know the instructions, this is unlikely to cause you any difficulties.

This is simply not the case when these sentences are combined with the : operator. It is an addition operator for creating arrays rather than an arithmetic operator, and has a precedence greater than multiplication but less than addition. The precedence of the addition operator is the smallest of the three. We often use the two-digit version in control structures such as never working with vectors, when we need to do boolean arithmetic, or when we want our expressions to work as vectorized expressions. In fact, the two-character version allows our expressions to function as vectorized expressions. On the other hand, if we want our expressions to behave like vectorized expressions, we use the single-character form of the expression. When dealing with multiple values, all arithmetic operators | behaves like. and the & operators. This means that when they work on vectors containing multiple values, they perform vector operations element by element.

R also gives you the ability to work with complex numbers in case you find yourself in a situation where you need this skill. You can create them in two ways: Using the as function directly. adding any number to a complex or imaginary number. An imaginary number is represented by a number preceded by the letter i. It is possible for the

imaginary number to have the value of zero, denoted by the notation 0i. This produces a complex number consisting of a single nonzero real component.

3.7 DATA STRUCTURES

A variety of data structures can be created, starting with the most basic types, combining smaller types to create more complex ones. This can be done in any order, but always start with the simplest types. The main building blocks that make up this structure are vectors, which are all sequences of values of the same type, and lists, which are sequences of values that can contain values of different types.

3.8 VECTORS

Since we talk a lot about vectors in this book, you should be familiar with vectors by now. We did not see a vector with multiple values, as we did not see an expression with more than one integer. As a result, we haven't seen anything that isn't vector. But at this point we will take a more technical approach to our discussion of vectors. The term "atomic arrays" more accurately describes what I have called vectors in the discussion so far. It refers to any major set of categories discussed in the discussion above. These can be created by combining multiple bases using the c function, which is an option.

CHAPTER 4

WORKING WITH LARGE DATASETS

Big data is a term that refers to large amounts of data. These datasets are sets of entities that require data warehouses to store data, often require complex algorithms to manage data, and often require distributed computing to do anything with it. At the very least, we're talking about a significant number of gigabytes of data; However, in the vast majority of cases we are dealing with terabytes or exabytes. Another aspect of data science is dealing with large amounts of data; However, this area of data science is not covered in this book. This section does not deal with datasets that are too large to be analyzed on your desktop computer; Instead, it focuses on large datasets and data management that extends your analysis.

Disregarding the big data issue, the definition of a large dataset largely depends on your purpose for using the data. It is proportional to the difficulty level of the expected difficulty in a one-to-one ratio. Some algorithms are very fast and can scan data in linear time. This indicates that the time taken to analyze the data correlates with the number of data points. On the other hand, some techniques require increasingly exponential time and cannot realistically be used for datasets larger than a few tens or hundreds of data points. The study of what can be done with data in a given amount of time or a fixed amount of storage (RAM or disk space or whatever you need) is called complexity theory and is one of the main areas of study in the field. computer Science. Complexity theory is one of the most important fields of study in computer science.

In practice, however, it often depends on how long you're willing to wait for a test to be done, and that decision is largely a matter of personal preference. In this section, we'll cover a number of situations that have arisen in my business where the amount of data has grown so large that it has prevented me from meeting my goals, forcing me to tackle them in new ways. Various forms. in different ways. These situations are due to the fact that I have accumulated a large amount of data during my studies. Your situation may be different from others, but even so, you can be inspired by the examples given to you.

4.1 COLLECT PARTIAL DATA BEFORE ANALYZING THE ENTIRE DATASET

But the first point I would like to make is this: It is almost never necessary to analyze the entire dataset to get at least a rough idea of how the data is organized and behaves. All you really need to do is look at a sample of the data. Unless you're particularly interested in finding truly unusual events, viewing a few thousand data points will give you the same level of information about the data as viewing a few million. If you see a few thousand data points, you understand several million or so data points. Sometimes you need access to incredibly large datasets to be able to successfully identify the information you are looking for. always z. B. Knowing how to distinguish between genetic variation and dominant enclaves, this is often the case when the ratio between your ridges is very poor and you need a large amount of data to distinguish between random associations and true associations.

In this case, it is necessary to have a large amount of data to be able to distinguish true associations from occasional ones. However, most of the signals of practical importance in the data tend to be found in smaller datasets. This is because smaller datasets contain less information. That's why you should first look at a smaller sample of your data before using the full potential of all your data in one analysis, especially if that analysis turns out to be extremely slow. This is especially important if the scan in question turns out to be extremely slow. This is something that needs to be done, especially as the analysis in question moves quickly. Under these circumstances, choosing a champion is a mandatory step. In addition to the columns that make up a data block, the data itself often has a structure. This structure is likely to have been established at some point during data collection. If your data is sorted chronologically by when they were collected, the previous data points you have access to may differ from the previous data points. Even if it is not explicitly represented in the data, it is always present, even if its structure is hidden.

Data randomization is an approach that can be used to eliminate problems that may arise as a direct result of this situation. Randomization has the potential to eliminate a weak signal, but with the help of statistics we can bring order to the chaos created by random noise. Much more difficult is the challenge of confronting long-standing

prejudices of which we are unaware . If you're dealing with a very large dataset that forces your work to progress more slowly than usual, don't be afraid to pick a random item and analyze only that subset. You can detect signals that are not in the full dataset in the subsample; However, the probability of this happening is much less than you might fear. If you're worried about this possibility, don't worry. When looking for signals in your data, you should always keep in mind the possibility of finding false signals. However, the probability of this does not increase as the dataset size is reduced relative to a larger dataset. When you have a larger dataset to compare results in the future, you're less likely to arrive at wrong conclusions after the study is complete.

This is because larger datasets contain more information. One of the most important considerations when using more traditional hypothesis testing techniques is the risk of generating false data. If you decide to use the 5% p-value threshold to evaluate whether a signal is significant, you can expect to get one in twenty false conclusions. In fact, the 5% threshold is not sensitive enough to detect small signals. If you ignore the possibility of other tests, you will most likely get false results. When you finally apply all the data to your models, it's very unlikely that it will still hold. This is because they are very unlikely to be valid. In any event, the size of the dataset plays a role in the ability to detect statistically significant deviations from an empty model, even if those deviations are irrelevant to the current research. When we do data analysis, we almost always rely on relatively simple blank models, and a simple blank model will never produce a complex dataset. If you have enough information, any problem you study is likely to have significant deviations from the simple null model you are using. If you don't have enough information, your chances are much worse.

Examples that accurately reflect the real world are not from a simplified linear model. There will always be an additional layer of complexity. If you only have a few data points, you cannot observe them. However, if you have enough data points you can rule out the possibility of an empty model. In no way does this mean that the things you are looking at have any relation to the world that actually exists. If there are signals you notice in a smaller subset of your data and those signals persist when you look at the entire dataset, you may be more inclined to believe them.

It goes without saying that to use dplyr for sampling, your data must be in a format that dplyr can process. But just in case, if the data is too large to load into R, it's impossible to even fit it into a data frame to be able to sample it. You will quickly see that dplyr supports using data stored on disk rather than RAM, and this support is available in various backend formats. For example, when users connect a database to dplyr, they have the opportunity to use this method to extract a representative sample from a large dataset.

4.2 LACK OF MEMORY DURING ANALYSIS

When working with R, memory can be used inefficiently. R remembers more than it seems; this means that even if your dataset is small enough to fit in memory and parsing time is not a serious issue, you may run out of memory. . Because R remembers more than is immediately apparent. This is because R remembers more than it seems. Modifying an object in R leads to the development of an entirely new object, because an object created in R cannot be modified after it has been created. Because the project has been handled very carefully, it is necessary to keep different copies of the data only in case of obvious inconsistencies between the two. If you make changes to one of the variables in the data frame, R creates a copy of the data frame with your changes. So you now have double access to the data as it is stored in both variables.

On the other hand, just because two separate variables refer to the same data block does not mean that the data block is displayed twice. But if you reference the data block with only one variable, R won't make a copy of it because it's smart enough to realize it isn't. Pipelines are preferred over manually assigning values to a large number of variables during an analysis for several reasons, one of which is to avoid namespace confusion. Pipelines can be used instead of assigning values to variables. However, reassigning the value of a variable would be fine, and the %>% operator will not give a large increase in the number of copies required. However, despite the use of pipelines, an extremely high level of caution must be used.

There are still many features in R dedicated to copying data. If changes are made to the data, a new version of the data is stored in a local variable of the function. Some of the data may be shared; For example, simply referencing a data frame in a local variable

does not create a copy of the data frame. However, if you split a block of data into training data and test data within a function, you copy the data and the function now displays all the data twice. This can be avoided by not splitting the data frame into training data and test data. You only have to worry about it when you're dangerously close to consuming all the RAM your computer can store, as that memory is removed when the function finishes its calculations.

Conversely, if the copied data is retained in the output of the function instead of being released when it is processed by the function, the allocated resource continues to be used. For example, it's not uncommon for models-fit functions to keep in the object all the fitting data returned by the function. This is not uncommon for model fitting methods. The lm() function used to perform linear regression records not only the input data frame, but also the response variable with any explanatory factors. It does this by making copies of the data so that each piece of data is stored not just once, but twice (and not reusing memory).

If you want to prevent this from happening, you have to give it very specific instructions not to use the pattern parameter and the x, y, and qr parameters. If you're doing data analysis in R and you're running out of memory, usually the problem isn't that you can't visualize your data to begin with; Instead the problem is that you have many copies of your data. You can avoid this problem to some extent if you don't store temporary data frames in variables and implicitly store data frames in the output of your functions. However, you can avoid this problem completely by using the method of intentionally deleting the saved data, thereby freeing up disk space.

4.3 TOO LARGE FOR THE LAND

When dealing with large datasets, it's not the first time I run into problems when my memory is full; more when trying to visualize data. Box plots and histograms are useful and rarely difficult to summarize data. This is especially true when creating scatter charts. There are two potential problems when creating point clouds with large amounts of data. The first thing to keep in mind is that when you create files from point clouds, you will get a graph containing each point cloud. This is the point cloud output generated from the files. This can be a very large file. Worse still, the planning process

will take forever as the audience has to weigh every point. You can get around this problem by creating raster graphics instead of PDFs, but that brings us to the next puzzle to solve. If a scatterplot contains too many points, it loses its ability to provide meaningful information. Because of the overlap of the points, it is difficult to determine how many individual data points are shown in the graph. This becomes an issue in the vast majority of cases, long before the time required for computation becomes an issue.

In other words, the final result will not be particularly informative; For a picture, see Figure 5-1. When the dots overlap, it's impossible to tell if the huge black dot cloud has different densities because the dots are stacked on top of each other.

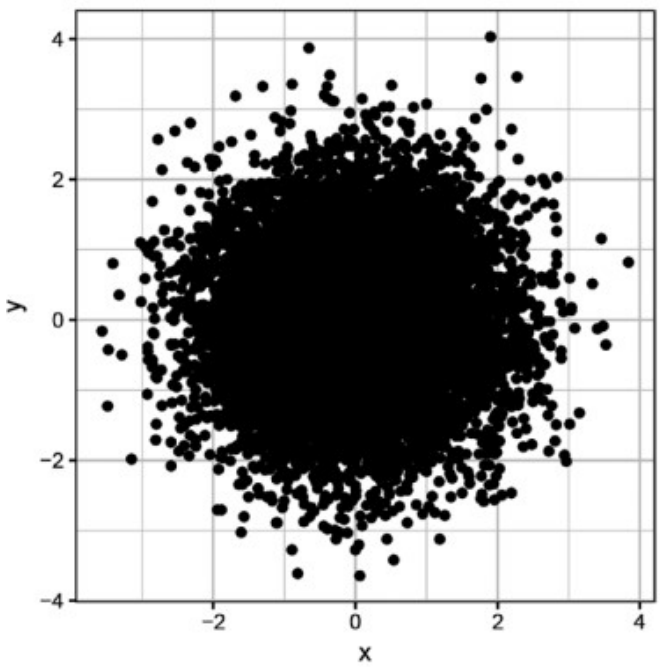


Figure 4.1. A scatter chart with too many points

The problem can be resolved if the points are rendered so that they remain visible even when surrounded by many other overlapping points. If the points actually overlap because they have the same x or y coordinates, you can arrange the points as in the previous section. Another way to solve a similar problem is to plot the points with alpha levels so that each point is only partially visible. However, as shown in Figure 5-2, you

get a graph with a significant number of points, although the density of the dots is obvious because they are somewhat transparent. It can be seen this way.

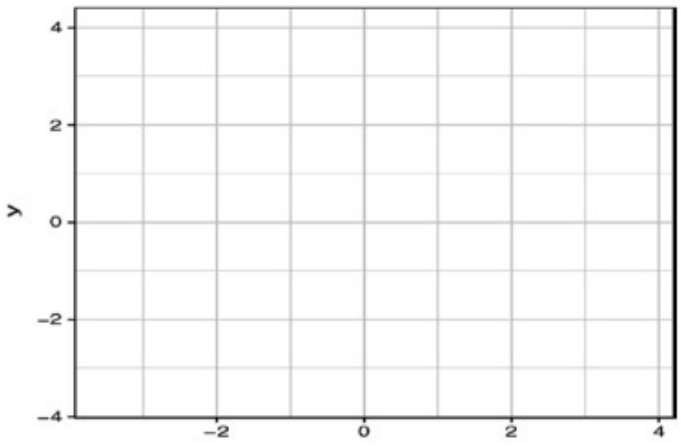


Figure 4.2. A scatter chart with alpha values

This doesn't fix the problem of files drawing all the dots, causing printing problems and file size issues. However, a transparent scatterplot is another way to represent 2D density, and you can visualize this more simply using the geometric density function, as shown in Figure 5-3. It can be done similarly.

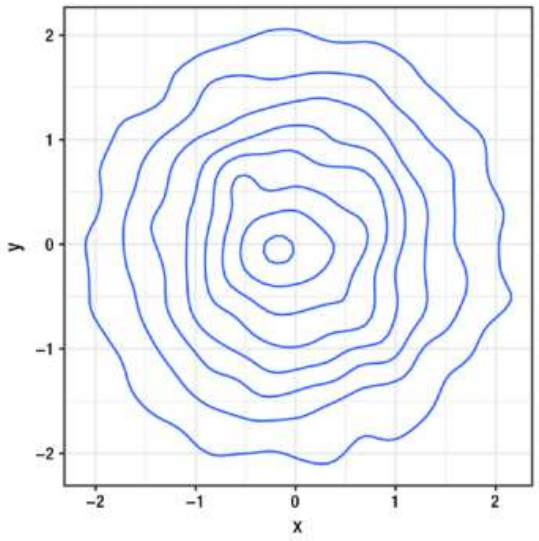


Figure 4.3. 2D density graph

4.4 TOO SLOW TO ANALYSIS

When plotting data, the problem almost never occurs with chart types other than scatter charts. If not, don't worry about too many stitches or too large a print file. Even when drawing a large number of points, the real problem only arises when you extract a PDF drawing, place it in the viewer and send it to the printer or send it to print. Only then will the problem be resolved. However, most scans will be slower if there are enough data points, and this can be a problem. This could be a problem. Again, the simplest approach is to take a subsample of your data and work with it. It will shed light on the important signals embedded in your data, without slowing down the pace of your research in any way. If this is not the solution for you, you need to choose the most effective forms of analysis. We almost often refer to algorithms that do their job in a linear fashion.

Unfortunately, most standard algorithms do not have linear time, and even if they do, the application may not make it easy to fit data in batches to allow batch updating of model parameters. This is the case even if the algorithm has linear time. While the algorithm itself works in linear time, this statement is still true. You will often be asked to find software specifically designed for the job at hand or write your own. You can use the biglm() function instead of the lm() function for linear regression, and the bigglm() function instead of the glm() function for advanced linear regression. Both functions are explained in more detail below. Each of these responsibilities is discussed in more detail in the following paragraphs. If you are using a data chunk format that stores data on disk and is supported by biglm (see the next section for more information), the package splits the data into chunks that you can load into memory and parse. If you are using a data frame format that does not store data on disk, the packet will not split the data.

Unless you're using a burst format that stores data on disk, the program won't split the data for you. If you do not have access to a software tool that can do this task automatically, you can manually fragment the data using any method you see fit. To illustrate, let's look at the car's dataset and try to use that data to calibrate a linear model of braking distance as a function of speed. But we must do this step in batches of ten data points at a time. Of course, we can easily accept such a small data set without the

need to collect it; In fact, we don't even need to use the biglm() technique for this; but this is sufficient for the purposes of the example.

However, Bayesian models based on previous conjugates are really suitable for such applications. Bayesian model fitting techniques are, in part, notorious for being slow; However, this type of implementation works better with legacy conjugate-based Bayesian models. If you have a conjugate priority, the posterior distribution you obtain as a result of analyzing a dataset can be used as a predecessor for the dataset you are examining later using it as the previous in the previous dataset. The concept of "with pre-conjugate" refers to this particular meaning of the word. With this technique, data can be segmented, the first segment fitted using a real precedent, and subsequent segments fitted using the result of model fits performed on the segments.

The Bayesian linear regression model implemented in Project 2 is a good example of such models. There we implement a function that creates a new model by fitting it to a dataset using an already fitted old model. This method basically uses an existing model for fitting. By applying this to automotive data and splitting it into bits of size 10, the results will be pretty close to those obtained in the large sample. Better still, the models allow different parts of the dataset to be examined independently, and their results can then be combined to create a model that applies to the dataset. They can not only be processed in batches, but also in parallel using multiple cores or many compute nodes. Batch processing is just one of the many processing options available.

It is possible to calculate the gradients separately for each section used in gradient descent optimization techniques and then combine these gradients to form a step in the optimization process. Other optimization techniques include iterative optimization and iterative gradient descent optimization. However, there are no universal solutions to the problem of handling datasets that are too large to be analyzed effectively. This is a long standing problem. You need to think about the algorithms you use, and in many cases a specific implementation of these algorithms is also required. Unless you're lucky enough to find a package that can handle batch data. You need to think about the algorithms you use, and in many cases a specific implementation of these algorithms is also required.

4.5 TOO BIG TO LOAD

Rather than storing the data you're currently working on, R prefers to keep it in the memory you're currently working on. If your computer doesn't have enough RAM, you're out of luck and you won't be able to recover it. Even when working with the most basic data representations like data frames you are dealing with standard data representations. Because R often prefers 32-bit integers for indexes and uses both positive and negative numbers for indexes, the maximum number of data points R can index is approximately 2,000 million. This is because R uses positive and negative numbers for indices. While your memory is capable of storing more, there are several alternatives you can choose to deal with this situation. An example of this would be the ff package. It is compatible with the type of desk we have used so far; But instead of using traditional files to represent data, it loads chunks of data into memory as needed.

This is a significant improvement over the previous method. Of course, this translation is not necessary if you can already represent a data frame in memory; no need in this case. If you can already do it, you don't have to. But besides these forms, ff also has methods for creating ffdf objects from files. For example, if you have a very large comma-separated values() file, you can use the read.csv.ffdf command. Using ff gives you access to a variety of techniques for efficiently generating summary statistics from memory-mapped flat files. These methods are accessible from the program's command line interface. Since they are implemented as generic functions (the topic of generic functions is covered in the title, it includes a description of the generic functions.) However, there are scenarios where the functions are (implicitly) in an ffdf because not all functions support it. It works like a normal data frame object. This is because not all functions support it. This can often cause this method to fail because there is too much data to add. Because there is not enough disk space to store all the information.

If you want to work with it, you need to do data analysis in batches, which means you need special functions to analyze data, and most of the time this means you develop analysis techniques entirely on your own, generic functions for models Linear models and generalized linear models can be found in the old biglm package . Linear and generalized linear models are included in these models. This indicates that the actual code being executed varies depending on the format in which the input data is provided.

This can be recognized by running the code differently. If you give them an ffdf object, they treat it like a data frame object and don't take advantage of the fact that data can be grouped in chunks. But if you give them a data frame object they will use this function. This problem is solved by the ffbase package, which includes the implementation of a custom bigglm() function that can be called on ffdf objects. This method can be used to calculate Bigglm values. Although designed for generalized linear models, you can still use it for linear regression, since linear models are only special cases of generalized linear models.

The user has the ability to change the rate at which data points are stored in memory using a parameter known as the block size accepted by the function. There is a more reasonable default value of 10 than we've used in the past, but in general you can use bigglm() on ffdf objects, just as you would with lm() or glm(). to use. frame objects. There is a default value of 10 that is more reasonable than the one we used earlier. More specifically, this means you have to call the lm() or glm() method. The inability to use ffdf objects with dplyr is the main disadvantage of using ff as a data representation mechanism. However, this feature is available in a development version of the software called ffbase2 available here. See the GitHub repository at for more details.

It is true that the dplyr package supports a wide variety of backends, including relational databases. When it can be processed as a flat file, there is no need to store the data in a database; However, large data sets are usually stored in databases that can be searched using SQL (Structured Query Language). The study of this language is not covered in this book because it is not the subject of the book. However, it is a language that should be mastered because it is useful. In any case, dplyr is a tool that can be used to access such databases. This indicates that it is possible to design dplyr pipelines consisting of function calls to manipulate data. The results of these queries are translated into SQL statements and made available to the database management system for processing. Using dplyr gives you access to popular database management systems like MySQL and PostgreSQL. However, in order to use these systems, you must first set up a data server. Therefore, if your data is not stored in a database, using LiteSQL would be the easiest option.

If it is already possible to process the data frame in working memory, in most cases it is not necessary to transfer the data to the database (RAM). Therefore, scanning while navigating a database system only serves to slow down the process rather than having the ability to store data in memory. Note that you can use dplyr to access database data even if you create the database with a program other than R. This is an important point to keep in mind. Since this method includes a temporary option, the table you fill in the database remains unchanged even if you change session. This is because the temporary option is part of the method. If you don't temporarily set it to FALSE, it will only stay as long as the database is open; After the database is closed, it will be deleted. If you do not temporarily set it to FALSE, it will remain permanently. If you set it to TRUE, it will still exist even if you delete it. This is useful for a variety of operations, but not what we need for this particular task.

After successfully establishing a database connection, you can use artificial intelligence to fetch a table from the database. The field of artificial intelligence is probably one of the most complex and advanced fields of study today. It is also developing rapidly. The level of difficulty in capturing and processing is not what is meant by complex in this context (although obviously very high); Rather, complexity refers to the degree of interrelationship with other seemingly separate areas. Connectivity start from the idea that it is necessary to draw inspiration from the neural networks of the human brain, while symbolists prefer to turn to knowledge banks and established rules of how the universe works to formulate their theories. Essentially, there are two schools of thought on how to properly build an AI. These schools of thought are called connectionism and symbolism. Both schools of thought are fundamentally right in their assumptions.

With these two pillars, they believe it is possible to design a system that can think and understand what it sees. You can solve a problem in two ways: You can use a simpler method (using the iteration technique) that improves the accuracy of the problem over time, or you can break the problem down into smaller and smaller parts. Also, there is a clear distinction inherent in the problem-solving approach: You can approach a problem by breaking it down into increasingly manageable parts. Instead, there is a strategy that involves combining the two approaches (Parallel Sequence Decomposition Approach). There is still no definitive answer to the question of which

approach or school of thought is most successful today. As a result, I find it acceptable to quickly review important findings in both pure machine learning methods and neuroscience from an agnostic perspective.

The main difference between supervised and unsupervised machine learning methods is whether the data from which lessons are drawn is labeled (as in supervised learning) or not (as in unsupervised methods). This distinction can be seen as a rough (unsupervised) classification. During a conversation about artificial intelligence, the concept of reinforcement learning, which is the third category (RL), may come up. RL is a machine learning strategy based on the simple idea of reward feedback that under certain conditions the machine works with the aim of maximizing the probability of a future reward. This concept states that the machine learns from its experience (cumulative reward). In other words, it is a trial-and-error learning method that falls between supervised and unsupervised learning: data labels are provided only after the action (i.e., they are rare and delayed), not for each training sample. punctual).

RL often poses two main problems, the credit allocation problem and the exploit discovery dilemma, as well as a number of technical problems such as the curse of dimensionality, non-stationary environments, or the partial observability problem. Credit allocation problem and exploitation dilemma are two of the most common problems with RL. At the heart of both issues is the distribution of credits among players and whether a particular region can be explored. First, incentives are lagging by definition, and you may need to take a certain number of actions in a given time to get what you want to do. command. After that, the challenge is to determine which of the above actions is actually responsible (and rewarded at the time) for the final result, and if so, what contribution that responsibility has made to the end product that existed before. has been responsible.

The last problem, on the other hand, is known as the optimal search problem and requires the software to create as accurate a map of the region as possible to determine the structure of your reward. To do this, the software needs to create an environment map that is as accurate as possible. There is a challenge that some call the optimal stopping question, and it can be seen as a unique type of satisfaction. This question begs the question: To what extent should the agent continue to scan the room for better

strategies or start using what they already know (and know how to work with)? In other words, should the broker keep looking for better strategies?

In addition to the existing classifications, it is also possible to classify machine learning algorithms according to the results they provide. Regression, clustering, density estimation and dimensionality reduction methods are all included in these classifications. The new wave of AI has spurred the production of new and innovative techniques and at the same time revived a somewhat outdated concept, especially the use of artificial neural networks. These innovations inspired the creation of Unique and Innovative Methods (ANN). Artificial neural networks have been likened to the human brain in that they are based on biological principles and allow computer programs to learn from experience using observational data. This comparison is possible because it allows computer programs to learn from experience. McCulloch and Pitts (1943) are believed to have invented the first artificial neural network (ANN), later called the Threshold Logic Unit (TLU).

But forty years later, Rumelhart et al. (1986) made significant progress in this area by developing the back propagation training technique for feedforward multilayer perceptron's. It was a step in the right direction (MLP). Any artificial neural network system (ANN) typically has an architecture consisting of several nodes arranged in an input layer, an output layer, and various customizable hidden layers. This is called traditional architecture (characterizing the depth of the web). Below are the results of multiplying the inputs by a predetermined link weight for each level. The total is then compared to a threshold value. The signal generated after addition is subject to the application of a transfer function. Thus, the transfer function produces an output signal that is used as an input by the next level.

Training takes place during the procedure of many rounds of this and is quantitatively measured for a set of specific training data by choosing the weighting variable that results in the least amount of error between inputs and inputs in the map. exit. Learning takes place in practice at many stages of this process. Because implementing ANNs requires no prior knowledge, it also means they are vulnerable to fraud due to lack of prior knowledge. In fact , implementing ANN requires no prior knowledge. They are also often referred to as deep learning (DL), especially in the context of multilevel

systems that perform computational tasks. DL stands for "Deep Learning". The best-known types of artificial neural networks (ANNs) include recurrent neural networks (RNNs), convolutional neural networks (CNNs), and biological neural networks;

However, there are many other types of ANNs (BNNs) that are increasingly being used. Because RNNs use sequential information, they can provide accurate predictions. In traditional ANNs, each entry is treated as if it exists in a separate and independent world. RNNs, on the other hand, are responsible for performing a unique function for each element in the array. This allows them to keep some sort of memory of past calculations. On the other hand, convolutional neural networks (CNNs) try to mimic the structure of the human visual brain, and each of its layers acts as a detection filter to recognize certain patterns found in the original data. These patterns reside in the data (and are therefore really suitable for object recognition).

Finally, biological neural networks, also known as BNNs, are more of a subset of artificial neural networks, also known as ANNs, rather than an application. In my opinion, Hawkins and George of Numenta, Inc. The Hierarchical Temporal Memory Model (HTM) designed by A.S. is the most influential example in this category. HTM is a technology that captures the structural and algorithmic aspects of the neocortex, and this model is what I consider to be the best in its class. HTM stands for "High Performance Modeling". Despite the great excitement surrounding deep learning opportunities, all that glitters isn't gold. DL is certainly a monumental step towards building an AGI; however, it has some limitations.

The main disadvantage is the large amount of data that must be entered into the system for it to work properly. This is the most critical limitation as it represents the biggest obstacle to large-scale implementation of the approach. DL debugging is incredibly difficult, and the vast majority of timing issues are resolved by constantly feeding more data into the network. This leads to an increased reliance on big data. Also, DL is very useful in shedding light on previously hidden connections and correlations, but by no means provides insight into the factors (causes of things) underlying observed phenomena. The network training process takes a long time due to the amount of data to be provided . Networks are often run in parallel to reduce the overall time required for this process.

This can be achieved in two ways: by splitting the model across multiple machines using multiple GPU cards (model parallelism) or by reading different (random) datasets through the same model while working on multiple machines to adjust the settings. . . Both methods are explained in more detail below (data parallelism). Because of the limitations we discussed in the previous section, different types of tools have been developed throughout human history. Particle swarm optimization, commonly known as PSO, is a technique that uses a computer to iteratively develop valid solutions to optimize a particular problem. Google is the company that developed the PSO (Kennedy and Eberhart 1995). The original candidate population, also called a particle, drifts in the search space. This population also includes individual particles that can optimize their location both locally and to the global search space, ultimately leading to the formation of an optimized swarm.

The agent-based computing economy, abbreviated as ACE, is a complementary tool that allows agents to interact in simulated environments according to predetermined rules (Arthur 1994). After the modeler imposes an initial condition, the dynamic system evolves over time as a direct result of the many actors involved in various forms of interaction with each other (and learning from previous interactions). Alternatively, evolutionary algorithms (often abbreviated as EA) are a more general category of problem-solving methods that use insights into the natural evolutionary process to arrive at optimal solutions to optimization challenges. The principles of selection, mutation, inheritance and crossover are included in this category. Another type is genetic algorithms, which is a form of algorithm, while evolutionary algorithms are a subset of genetic algorithms.

An excellent example of an evolutionary algorithm (EA) is the genetic algorithm (GA), an adaptive heuristic that attempts to mimic the process of natural selection. It is a search optimization strategy for evolutionary computing that starts with a base population of candidate solutions and then develops these solutions based on the idea of survival of the fittest. It is a scalable approach to optimizing computational search. Genetic programming, commonly known as GP, is an extension of GA that applies a GA to a population of computer programs. GP is sometimes called "genetic programming". He was the first person to describe Koza in 1992. It begins with the

generation of chromosomes, also known as the initial program population, consisting of a predetermined set of functions and a set of terminals, and then assembles these components into a random tree structure. Chromosomes are also called the starting population of programs.

In this context, the above terminology takes on a slightly different connotation: replication refers to the act of copying another computer model of a pre-existing population; Crossover refers to the process of randomly recombining selected parts of two different computer programs; and mutation refers to the process of randomizing the replacement of a functional or terminal node. Reproduction, crossover and mutation have slightly different meanings in this context. EPRs, also known as evolutionary polynomial regressions, are a type of hybrid regression that uses genetic algorithms (GA) to choose exponents from polynomial and numerical regressions (aka least squares regression) to calculate true coefficients.

EPRs are also known as evolutionary setbacks. AG is used to calculate the exponents of the polynomial (Giustolisi and Savic 2006). Recent advances by Sentient Technologies, LLC have led to the creation of a model that can be called Evolutionary Intelligence (EI) or Evolutionary Computing (EC). Each of these templates has its own unique features that make it interesting. Then the random evolution process of billions of potential solutions, called genes, begins. All of these genes are prone to malfunction due to the nature of the process. Your exercise data is then used for evaluation, and the algorithm uses a fitness score to decide in what order to present the best answers (and discard the worst). The process is then repeated until convergence is achieved; meanwhile, components of emerging candidates are introduced into the development of new populations.

There are two other considerations to keep in mind to wrap up this section. Shannon (1948) was the first to propose generative models (GMs), but OpenAI, a San Francisco-based not-for-profit artificial intelligence research institute, has recently moved them to the forefront of the discussion (Salimans et al., 2016; Chen et al.) . get. 2016). Intuitively, this category of models is known as models where we can generate random data provided there are some hidden parameters. After collecting the data, the system proceeds to generate a common probability distribution and set of labels based on the

collected information . Second, Cao and Yang (2015) proposed an innovative method that essentially changes the learning algorithm instead of jumping directly from one training data point to another. This strategy is described in the article. Their paper described the new approach they had developed. Machine learning opt-out, commonly known as MU, is the term for a technology that gives computers the ability to "forget" unwanted information.

They actually contain an aggregation level as an intermediate step between the algorithm and the training data points. This step ensures that the algorithm produces accurate results. As a result, the training algorithm and data points no longer depend on each other, but only on the sums; This not only makes the training process much faster, but also allows it to be updated incrementally without having to train the model from scratch. is time consuming and expensive if it has to be run multiple times. Both authors proposed the concept of using the word "line" to refer to the entire data transmission network. Therefore, if some data and its descendants need to be deleted, the system no longer needs to recalculate the entire array; Instead, you can recalculate a limited number of totals. In fact, the result of the above sentence is that some data and its descendants will be deleted. This is because the two authors themselves coined the word "line."

4.6 PROGRESS IN NERVOUS SCIENCE

Alongside the numerous advances in pure machine learning research, we've made great strides towards better understanding how the brain works. While much is still unknown, we now have a relatively clearer understanding of the processes occurring in the brain that could contribute to the development of artificial general intelligence (AGI). It should come as no surprise that a strategy aimed at exactly replicating the workings of the human brain is both possible and not the best course of action. Orienting yourself to brain activity, on the other hand, is a whole different game. Neuroscience work not only has the ability to inspire the creation of innovative architectures and algorithms, but also has the potential to legitimize the use of existing machine learning research to create artificial general intelligence. [Example:] (AGI). Numenta researchers believe the human neocortex should serve as a model for artificial intelligence to gain even more detail.

While a general theoretical framework for the cortex has not been widely adopted by the scientific community, Numenta believes a cortical theory should be able to explain the following. How can different layers of neurons learn sequences; Features of II DSPs; III, an unsupervised learning mechanism with the transmission of temporary data streams; IV connection between different neuron layers; How different parts of the brain shape the outside world and produce behavior; and I saw the hierarchy between the different parts of the brain. These six guiding principles make up biological or artificial intelligence and should be found in any system that claims to be intelligent. Whether an intelligence is biological or artificial, these six guiding principles make it intelligent.

Because the neocortex learns from sensory information and then builds a sensorimotor model of the environment, it appears to be a logical model, at least from an intuitive point of view. Neocortical learning occurs when sensory input is combined with motor output. Any AI to be developed needs to be both flexible and robust. In fact, we still don't have a perfect understanding of how the neocortex works, making it imperative that any AI be built accordingly. In more recent research, Hawkins and Ahmad (2016) drew their attention to a neuroscientific challenge considered fundamental to the development of artificial general intelligence (AGI). They set out to explain how neurons integrate input from hundreds of synapses and the large-scale network activity resulting from that integration, and they were successful in both efforts. Since it is not clear why neurons have active dendrites, almost all ANNs developed so far do not use artificial dendrites. This is due to the lack of information.

This leads us to believe that something is missing in our handcrafted buildings, which would be a reasonable conclusion. His idea explains how neural networks communicate with each other, given that the human brain has thousands of synapses. They proposed a set of memory models common to all neocortical tissues; If true, this model will have a huge impact on how we build and use artificial brains. The excitatory neurons studied formed the basis of the model. Rocki (2016) has pointed out some features that are particularly relevant for creating a biologically inspired artificial intelligence. Specifically, the components required to develop a general purpose learning algorithm were highlighted in this discussion. It is believed that humans do not learn in a

supervised manner, but (unsupervised) learn to interpret input from the environment and filter out as much data as possible while maintaining information integrity. Important information for them.

This belief is based on the idea that people learn to interpret and interpret environmental inputs (Schmidhuber 2015). The human brain works according to a principle similar to the Pareto law or the rule of minimum descriptive length, depending on the circumstances. This principle allows the human brain to only hold and store information that can explain much of what is going on. According to Rocki, unsupervised learning is responsible for organizing and compressing information, turning our brains into data concentrators (Bengio et al. 2012; Hinton and Sejnowski 1999). According to Rocki, the design of a general learning algorithm should be compositional, sparse and distributed, aimless and scalable.

In addition to the criterion of not following the algorithm, another criterion must be met. The human brain is hardware-wired to learn in a hierarchical way, starting with basic patterns and breaking down more complex topics into the smallest building blocks it grasps. The human brain is designed to learn hierarchically. Deep learning is a proven method to capture this compositional and hierarchically structured learning style. Ahmad and Hawkins (2015) have shown that sparse representations are required and are much more resistant to noise effects than their dense counterparts. This discovery was made possible by the fact that sparse representations have higher information density. But there are many other quirks that add to the appeal of SDRs. For example, although the brain lacks region-specific algorithms, cortical columns act as independent feature detectors.

In response to a particular stimulus, each column activates while laterally inhibiting the activity of other columns in the immediate vicinity. This leads to the formation of rather weak business models. Due to the rarity of these signals, Candès et al. (2006) states that decoding a given external signal is much easier to extract information from it. Alternatively, the feature of staining is useful for understanding what causes pattern variations. [Example:] [Example:] Using SDR not only makes it easier to remove unwanted information, as mentioned above, but also makes it easier to remove unwanted information. It is a representation of the lowest entropy codes developed by

Barlow et al in 1989. These codes offer a generalized learning process with low time dependence.

its predecessors support the idea that this is the only way to capture and generate transferable ideas. While it's debatable why the learning process shouldn't have a clear goal, Rocki argues it's the only way to get what needs to be done. Additionally, Rocki emphasizes the importance of scalability as an important part of an overall learning architecture. Since each brain region is responsible for both processing and storing information, the brain can actually be thought of as a parallel system (which is why GPUs are much more efficient at deep learning than CPUs). . This means that an AI will have a hierarchical structure that separates higher-level connections (synaptic updates) from lower-level connections (local learning), and a self-computing memory to generate the information necessary to do so. reduce data transfers. Synaptic updates are an example of higher order connectivity.

Local learning is an example of lower-level connectivity. Ultimately, Rocki concludes that developing an AI requires the inclusion of complementary parts that are more functional than structural. Compression, prediction, comprehension, sensorimotor, spatio-temporal invariance, context updating and model completion are some of the components that make up this system. We have already discussed the importance of compression and sensorimotor functions, and we can think of AGI as a universal compressor that produces stable representations of abstract concepts. We have previously discussed the relevance of compression and sensorimotor function.

That's how we might think about AGI, but this particular issue is moot because of the no free food theorem (Wolpert and Macready 1997), which implies that this algorithm cannot exist. If we have this perspective, we can argue that understanding and learning to predict are the same thing. Prediction can also be viewed as a relatively weak mode of space-time coherence in the universe. If we have this perspective, we can argue that understanding and learning to predict are the same thing. In summary, we need to create a feedback loop of continuous bottom-up prediction and top-down contextualization into our learning process. With the help of this contextual spatio-temporal concept, it can also clarify the situation in which several (contradictory) predictions are made.

4.7 HARDWARE AND Chips

One of the factors contributing to the current explosion in artificial intelligence (AI) research and the rapid rise of the industry to the top of the scientific food chain is the exponential rate of technological progress we have faced in recent years. . This has been one of the contributing factors to the current explosion in AI research. However, it is important to note that AI is already having a significant impact and will play a key role in determining the future of technology. This is a point to consider. First of all, graphics processing units, better known as GPUs, have evolved from their first use in traditional GUI applications to their current use in alternative parallel processing processes.

This transition occurred as a result of its use in alternative computing processes. NVIDIA is at the forefront of this trend, dominating the industry with its CUDA platform and the recently launched Tesla P100 platform. Both platforms were recently introduced, the first GPU designed for large data center applications. Together with the P100, they created the first complete server appliance platform and called it the DGX-1. This platform was developed by them. The level of deep learning that can be achieved through this platform is taken to a whole new and higher level. Also, they recently launched the Titan X, the most powerful graphics processing unit (GPU) ever produced (3584 CUDA cores). Overall, the most important advances we've seen are in chips, specifically neuromorphic processing units (NPUs) designed to mimic human brain function. The most significant improvements we've seen are especially in chips.

Major officials have developed a variety of artificial intelligence chips, including but not limited to: IBM first introduced the TrueNorth chip in 2016, and the company claims it is remarkably mammal-like. Brain. The device has 5.4 billion transistors and can simulate up to 1 million neurons and 256 million connections between these neurons. It can also simulate 256 million connections between these neurons. It has 4,000 nuclei, each with 256 input lanes (axons) and the same number of output lanes as input lanes (neurons). The electrical charges in each core must reach a certain threshold before signals can be sent. This configuration is somewhat similar to the Neurogrid developed by Stanford, but the academic version consists of sixteen independent chips as opposed to the single chip proposed by the software giant.

On the other hand, Google commented on the design of an application-specific integrated circuit (ASIC) specifically designed and adapted for neural networks. This announcement was made by Google. Its application-specific integrated circuit is known as a tensor processing unit (TPU). DeepMind and RankBrain, also known as Google Search, are powered by TPU, which aims to maximize performance while minimizing power consumption associated with machine learning challenges (e.g. AlphaGO). Knights Landing is the name of the latest iteration of a series of Intel processors that are functionally identical to those developed by Xeon Phi. The codename of these processors was Knights Landing (KNL). There is no longer a need to send machine learning tasks to individual coprocessors as the KNL has the potential to act as the main processor rather than the GPU.

It has a maximum capacity of up to seventy-two cores. Qualcomm also spent a lot of time and energy producing the Snapdragon 820; which will eventually lead to the development of a deep learning software development kit (SDK) for Snapdragon's neural processing engine and engine. . Intelligence platform. The total cost of all these chips is extremely high (billions of dollars for R&D and hundreds of thousands of dollars to sell, respectively) and currently their only intended use is for commercial applications in retail customers who can't afford it. to them. The most notable exception to this general trend is the early 2016 release of a massively commercial artificial intelligence chip called Eyeriss by a group of MIT academics. This chip is the most notable outlier of this overall trend.

Consisting of 168 processors, this chip is built on the balance of smartphone performance and is therefore highly energy efficient. But since a smartphone is designed with the power budget in mind, it has computational limitations. Despite being a very expensive game, many startups and small businesses make significant contributions to the industry. An open-source version of Numenta's NuPIC, an intelligent computing platform that analyzes streaming data, serves to explain this concept. Chips known as memristors, a type of device that can change the resistance of its internal circuits in response to electrical impulses applied to them, are now available (and used as volatile memory) from Knowm, Inc. Developed by KnuEdge (with one of its subsidiaries, KnuPath), LambaFabric is based on a completely new design that

differs not only from the traditional GPU architecture, but also from the TPU architecture.

KnuPath was a subsidiary of KnuEdge. A new Application Specific Integrated Circuit (ASIC) called High-Bandwidth Memory has recently become available for purchase. This ASIC was designed and developed by Nervana Systems. Horizon Robotics is another company that makes significant contributions to this sector. Another company actively involved in this field is krtkl. They developed a new low-cost dual-core ARM processor called the Snickerdoodle. This processor also has FPGA, Wi-Fi and Bluetooth. Movidius deserves an extra mark for being the first company to invent a completely original concept, an all-in-one USB stick for deep learning. This should be considered an argument in favor of the company. It has a processor called Myriad 2 and its internal codename is Fathom Neural Compute Stick. This chip was developed in collaboration with Google to solve all the complex image recognition problems (but has also been used to power drones and robots of all kinds).

CHAPTER 5

UNSUPERVISED LEARNING

When we enter supervised learning, we are presented with a set of variables and one or more goals that we want to predict using those variables. This is in contrast to unsupervised learning, where we do not receive this information. However, building predictive models is only part of analyzing data. There are times when we are only concerned with better understanding the structure that actually exists in the data we analyze. There are many different hypotheses that could be true about this. There are times when previously unknown structures can reveal previously unknown pieces of data. There may be times when we want to make an active effort not to approach an unfamiliar building. For example, if we have datasets that need to be comparable, the last thing we want to know going forward is if there are any systematic inconsistencies between them. Whatever the purpose, the primary purpose of unsupervised learning is to recognize previously unknown patterns or structures in data.

5.1 SIZE REDUCTION

When you already have high-dimensional data, you can use a set of strategies known as size reduction to map high-dimensional data to fewer dimensions. These methods are used as their names suggest. The purpose of this activity is usually to create a visual representation of the data to find hidden patterns in the charts. Analytics often add nothing new to knowledge; all it really does is render the data a bit different. Some information may be lost during the process; However, analysis may be easier if you reduce the number of dimensions. This step is important if your data contains multiple columns with the data type you are referring to. There may not be many observations, but each observation may have many variables and relatively little information in a column. This is possible even when there are not many observations. An example of this can be seen in genetic data, where it is common to observe hundreds of thousands, if not millions, of genomic loci in any given individual.

For each of these gene sites, we have a number that can range from 0 to 2, which tells us how many instances of a particular genetic variation are present in these markers.

This number corresponds to a genetic location. While the information contained in each bookmark is quite small, it has the potential to tell a lot about a person when all bookmarks are viewed together. Principal component analysis, the first example we'll explore in this section, is commonly used to map hundreds of genetic markers to a few dimensions that are most useful for determining correlations between different people. In this section, we will consider an example of principal component analysis. I will not use very high dimensional data; Instead, I will use smaller datasets to illustrate techniques where they may still be useful. I will not use very high dimensional data in this presentation.

5.2 ANALYSIS OF MAIN COMPONENTS

When your data is assigned using principal component analysis (PCA), it moves from one vector field to another vector field that has the same number of dimensions as the original vector field. Therefore, the total number of dimensions does not decrease. However, choose the coordinate system for the new area so that the first coordinate has the most information, the second coordinate has the second, and so on until all coordinates have the same amount of information. It is possible to simplify it down to the ground state, which takes the form of a linear transformation. Change the basis of your vector space so that the direction in which the first basis vector moves has the greatest data variance, and subsequent basis vectors have progressively lower data variance. In effect, it changes the basis so that the direction in which the first basis vector is moving has the greatest variance in the data.

Components are the basic building blocks of the new vector space; Evaluating their behavior is central to the term "primary component", which refers to the methodology of focusing on the first components that are most important. The first part of the transformation may contain changes to normalize the data, but the final stage of the transformation is always a linear mapping. This is done to normalize the data. Therefore, after the conversion, your data still contains the same amount of information; The difference is that the information is displayed in different sizes.

While PCA simply modifies your data, it requires your data to be represented as digital vectors from the start. Data editing is the first step in working with categorical data.

One technique is to express the elements of each level as a binary vector; this is similar to the way pattern matrices are stored in supervised learning. This is possible for any level. It's a method. However, if a large number of factors in the data need to be taken into account, PCA may not be the most appropriate strategy to use in this situation. It is beyond the scope of this book to discuss principal component analysis (PCA) theory in detail; however, it has been covered extensively in many other guides; So we'll just look at how it's used in R.

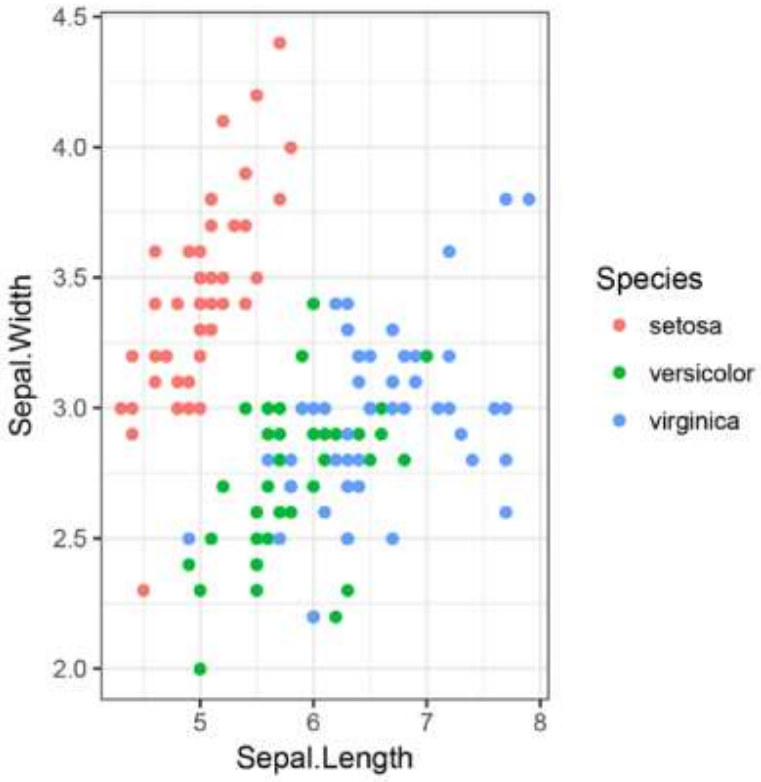


Figure 5.1 Diagram of iris sepal length versus sepal width

What comes out contains a lot of information about the final result. Using standard deviations, we can determine how much variation there is in each component, and by shifting the data we can determine what the linear transform is. When we view the PCA object, as shown in Figure 5.1, we can see how much variance is in the data corresponding to each component.

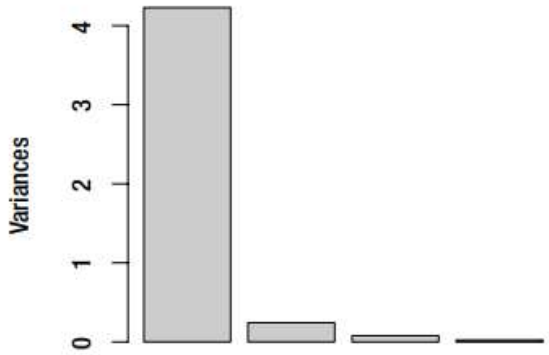


Figure 5.2. Plot of variance of each principal component for iris dataset

The first thing to do after completing the transformation is to examine how the scattering is distributed among the components. If the first components don't explain most of the variance, the transformation didn't help you. When the time comes, plotting the first few components can tell you about the data. You can do this by looking at the chart.

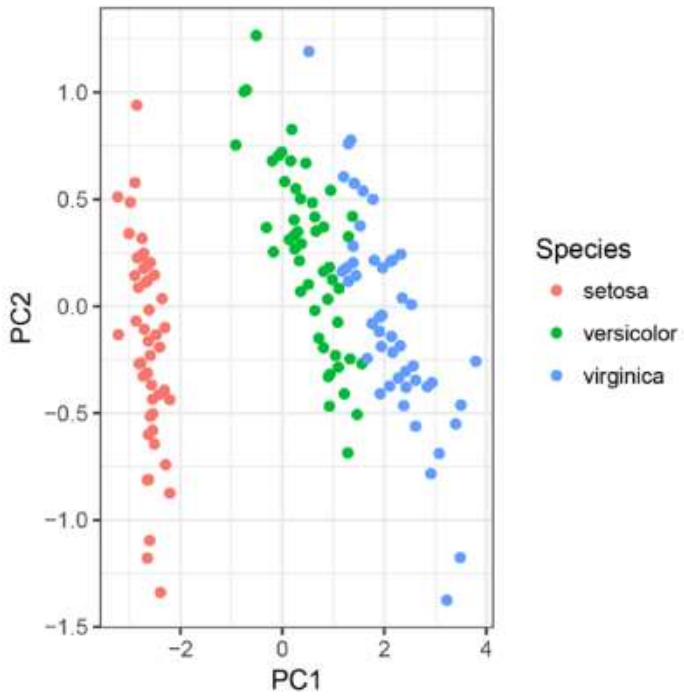


Figure 5.3. Representation of the first two main components of the iris dataset

This can also be used with new data that was not originally part of the PCA object creation process. At this point we only need to provide you with the old data we use. It is not necessary to remove the type variable because the program will decide which one to use based on the column names. Therefore, it is not necessary to remove the Type variable. Now that we have the two initial components, we can represent them relative to each other as shown in Figure 5.2.

The reserved Iris object returned by the prediction function is an array, not a data frame, as expected. In fact, a data block cannot contain more than one row. So we need to use as data frame to convert it back to data frame because ggplot() function is not working properly in its current state and we need it to work. Since we want to color the table according to the types, we need to add the following information again; However, note that the pca object does not consider the data for this component; We use the cbind() command for this. So we sit down and formulate our strategies. We did not learn much from this meeting. The modified data provides approximately the same amount of information as the original columns with approximately the same level of specificity in their respective formats. But now that we've seen PCA in action, let's try it with a slightly more interesting example.

This data collection takes into account the votes cast by Republicans and Democrats on a total of sixteen different proposals. The recommendations are divided into their respective categories below. There are three types of votes: affirmative (yes), negative (no), and unknown (none). The lack of data in this case suggests that someone made a conscious decision not to vote, as the probability of votes being accidentally lost is very low; Therefore, it is not an absence in the traditional sense.

This is because entries are less likely to be accidentally lost. There is a possibility that some information may also be stored somewhere. Whether there are differences in voting preferences between Republicans and Democrats is an interesting possibility we can explore. This is something we expect, but the real question is whether you can actually see this in the data. Because each column is binary (or triple if you think the missing data really matters), it contains very little data, and there is no apparent difference between the two groups. We can try to run a PCA on the data.

5.3 MULTI-DIMENSIONAL SCALE

Instead of using numerical vectors to describe the distance between two objects being compared, it can be helpful to have a metric for the distance between the two. Take, for example, the case of strings. You can convert them to numbers depending on the encoding, but the range of possible strings is pretty wide, even infinite if you didn't limit the length of the strings, which means this method isn't particularly useful in practice. However, there are several approaches to determining the degree of difference between the two sequences. Converting text values to numeric values is more difficult than generating a distance measure, at least when dealing with strings. Defining a distance measurement is a simpler task. If we have a distance measure, we can define the data we have as the distance matrix, which is a matrix containing all the distances in pairs.

If we have a distance measurement, we can define the data we have as a distance matrix. However, if there are less than a few thousand data points, this is not so important. The number of matches increases with the square of the number of data points. With so many data points, it makes sense that this strategy cannot be successfully applied. After using the matrix containing all binary distances as a starting point, multidimensional scaling translates each data point into linear space while trying to keep the binary distances as precise as possible.

We use cmdscale to create a representation of these distances in two-dimensional space. The dimensionality at which the points will be placed can be specified with a parameter named k that accepts. We wouldn't be able to see it even if we gave it a large enough number of k, but if we did, it would fit all binary distances exactly. We would benefit most from the low dimensionality and decided to use two variables to represent the data. We also benefit more from having low dimensionality. The final result is a matrix with one row for each of the initial data points and one column for each of the desired dimensions; In this particular case there are two dimensions.

In this formula, the axes corresponding to x and y are represented by the names V1 and V2, respectively. This takes advantage of the column names in a data frame where we provide no names are named Vn where n is an increasing number. In particular, it

allows us to use this function. It is no accident that the plot is practically similar to the PCA used in the past; The only difference is that it displays differently . We can do the same with voting data; more precisely, we can use the own data available for the quantified variables; The result of this procedure is shown in Figure 5.4.

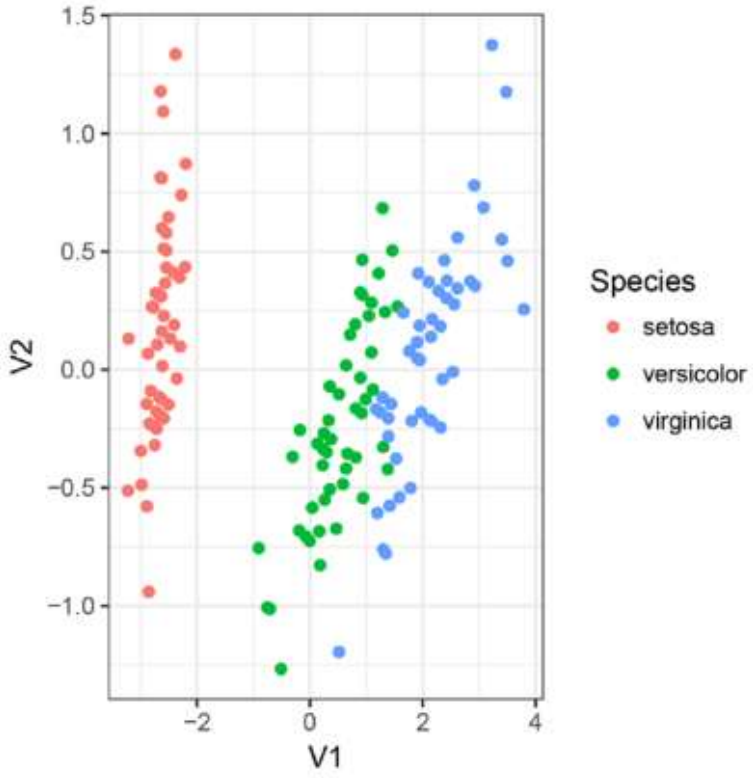


Figure 5.4. Multidimensional scale chart for iris data

5.4 GROUPING

The purpose of clustering algorithms is to first detect patterns of similarity between data points and then divide the data points into clusters based on these patterns. When the structure of such clusters is hierarchical, each data point is associated with multiple clusters, ranked from the most specific to the most general; however, in the case of a non-hierarchical structure, only one group is usually assigned to each data point. In the following sections, a chapter is devoted to each type of grouping, followed by a review of the two most commonly used grouping types.

5.5 K-AVERAGE GROUP

When using k-means clustering, your goal is to organize the data into k distinct groups, and the exact number of groups is at your discretion. In other words, you control the number of groups. Typically, data must be delivered as digital carriers most of the time. From a mathematical point of view, if you have a method for calculating the mean of a series of data points and the distance between two series of data points, you can use this strategy successfully. In other words, both elements are needed for this strategy to be effective. Numerical data are needed to use the R function known as kmeans, which performs K-means clustering. In short, the algorithm starts by making reasonable guesses about the "centers" of the proposed clusters. After this step, each data point is assigned to the nearest center and a data pool is created as a result. Each center is then moved to the position that is considered the average of its associated groups. It happens eventually. This process should be repeated several times to maintain balance. Multiple calls to the function do not always produce the same result, as the initial centers are randomly selected. This is because the starting centers are chosen randomly. At the very least, you should be prepared for different calls to generate different set tags.

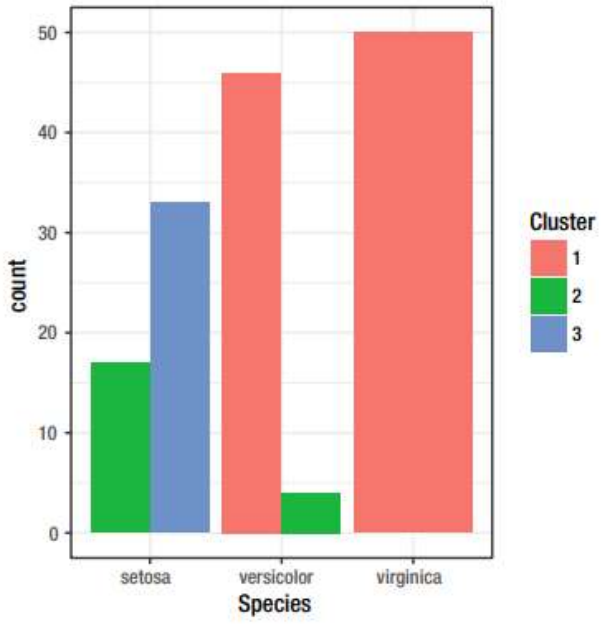


Figure 5.5. Group tasks for the three types of iris

We'll start by integrating the iris dataset with the cluster association we get from the clusters, and then move on to the next step, bar chart creation. Because the value of the position parameter is dodge, cluster assignments appear side by side instead of stacked. Setosa looks markedly different from the other two species that seem to share common characteristics based on at least four measurements we have. Setosa looks distinctly different from the other two species. Having already seen the data graphically, this finding didn't surprise us, nor were we surprised that Setosa appeared to be seemingly unique between the other two species. There is also a luck factor to consider in this scenario. The outcome that can be achieved is affected by having another starting point where the initial centers will be placed. For example, if you placed two sets of Setosa data points in the cloud, you would separate those points into two sets and combine the Versicolor and Virginia points into a single cluster.

This would happen if you put two clusters in the cloud. Had the starting point been different from it, the end result might have been different. Never skip the step of visually assessing the relationship between the clustering result and the actual locations of the data points . This is an important step that should never be skipped. This can be accomplished by plotting each data point and observing how the data points are grouped and ordered. Since we know how to plot the data, we can try to plot the data for the first two main components. However, we can also try to plot the data for any other feature pair we choose. Alternatively, we can show the points for any available combination of features. You may remember that the Predict() technique gives us the ability to map data points from key components to key features. This applies not only to the original data used in the construction of principal component analysis (PCA), but also to the centers obtained by combining k-means.

If, after looking at Figure 5.6 again, you get the impression that some square points are closer to the center of the "group of triangles" than to the center of the "group of squares", or vice versa, you are correct in your assumption. This is because the center of the "triangular group" is closer to the center of the "square group". It's important not to get too discouraged because there are actually two lies going around here. Since the axes are not on the same scale, it can be concluded that the distances are further along the x-axis than the y-axis. One of the problems is this. The distances used to group the

data points are in the four-dimensional space of the original elements, but the graph is a two-dimensional plane projection of the first two principal components.

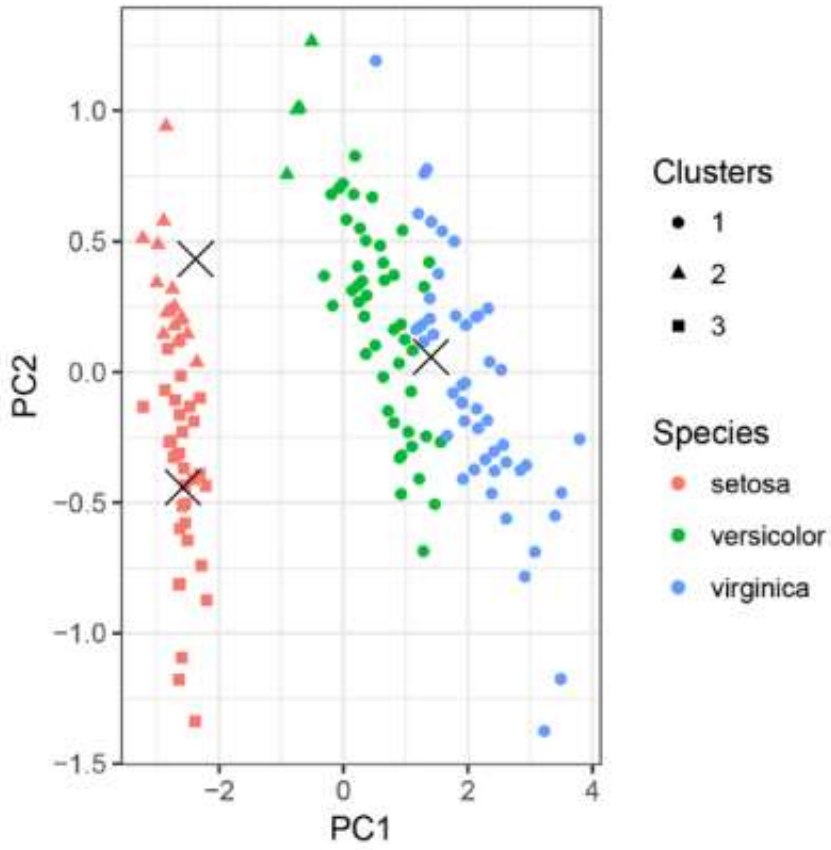


Figure 5.6. Clusters and types of irises for sparse grouping

There is a second distinction between the two. However, there are reasons for concern with different distances. However, if you have one axis in centimeters and another axis in meters, the distance along one axis is numerically 100 times greater than the distance along the other axis. In fact, the focus depends on the distance between the cluster centers and the data points. Trying to represent all properties with a single unit is not the only solution to this problem. There are other options. First of all, this cannot be guaranteed at all hours of the day or night. It is not possible to meaningfully convert units of time or weight into distances. [Example:] Even so, the object to be measured

is still important to the entity being considered. [Example:] It makes sense to express certain quantities in meters, such as a person's height; However, it is not recommended to use certain sizes, e.g. B. the volume of a cell expressed in meters. Principal component analysis also encounters some difficulties in dealing with this problem.

A method that tries to produce a basic vector space based on the variance of the data is clearly affected by the units used in the data fed into the method. In fact, the variation in the data can be taken as a measure of the available headroom. Most of the time, the problem can be solved by scaling all the input features to be zero-centered and have a variance. This will solve the problem. At the end of this step, the average of the features from each data point is subtracted and the resulting value divided by the standard deviation. Seen through the prism of standard deviations, this indicates that all dimensions have the same total variance.

The prcomp() function requires passing parameters to perform the scaling. Using the parameter center option with a default value of TRUE, data points are converted to zero mean and parameter scale. With the default value of FALSE, this option changes the data points to have a variance across all dimensions (note the trailing dot). Even if there is a complete one-to-one correspondence between clusters and species, the confusion matrix will only be diagonal if clusters and species occur in the same order. This is because the set has no type information. It allows us to classify species accurately by assigning any species to the group to which the majority of members of that species belong.

Although this is not an ideal solution, this way two different species can be put in the same group and still not be able to construct a confusion matrix, this will help us in the case discussed here. In short, we should keep this in mind. We can calculate the number of observations of each observed cluster of each species by counting the following: Since k is a parameter that must be specified, the question now is how to choose it. Since we know three different species in this region, we decided to use these three for k. First, since we don't know how many clusters are in the data, how do we choose which value of k to use, if any? Unfortunately, a comprehensive answer to your problem cannot be given. Some solid standards need to be met, but there is no single solution that is 100% reliable and viable in all circumstances.

5.6 HIERARCHIC GROUPS

After you create a distance matrix from your data, you can use a technique known as hierarchical clustering to organize your data in a more logical way. The general goal here is to create a tree structure of nested clusters. This is achieved by iteratively combining groups. First you need to place each data point in the appropriate singleton cluster. It then goes through an iterative process to find two physically close clusters and then merge them into a new cluster. Continue in this manner until each data point falls into the same large group. There are several alternative approaches, and their main difference is how they choose which clusters to combine and how they estimate the distance between multiple clusters. The iris dataset may allow us to see the hclust() method in action. This function is responsible for implementing various algorithms in R; The parameter method decides which algorithm to use. First you need to create a distance matrix. When I get to this step, I will start scaling the data.

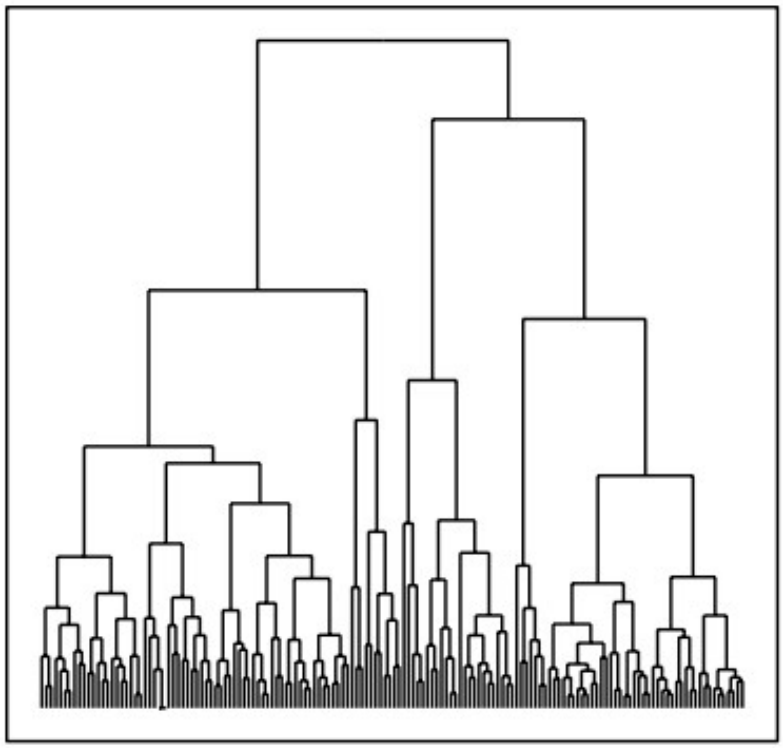


Figure 5.7. Hierarchical grouping of iris data drawn with ggdendro

Using ggdendro gives you access to the raw parts of the graph and gives you control over a significant portion of the graphical representation of the tree. As it is often not sufficient to just look at the clustering, we need to be able to extract the clusters that actually occur to use the result. This is necessary because viewing the cluster alone is often not sufficient. This is possible thanks to a technique called cuttree(), which allows you to do this, even though it only has a t in its name. You can give it a number of k to cut the tree to exactly the level of k clusters, or give it a parameter h to divide the tree into clusters by dividing the tree by the h height dimension.

In all cases, the groups are clipped at exactly the tree level where there are k groups. Whether one of these two options is chosen or not, the tree is cut into bundles. So if we want to base clustering for any classification, we likewise need to construct a confusion matrix. On the other hand, hierarchical clustering contributes much less to classification than k-means clustering. With k-means grouping, it is easy to determine which cluster center a new data point is most similar to by taking that point and comparing it to the others. If you were to use hierarchical clustering , you would have to rebuild the entire tree to determine where each node belongs in the tree hierarchy.

5.7 ASSOCIATION REGULATIONS

The final unsupervised learning method we'll look at is designed to work with categorical data, regardless of the order in which the categories appear in the data. To use association rules, you must convert numeric data into factors. Similarly, in order to use methods such as principal component analysis, factors must be converted into numerical data. This isn't usually a problem, and if you want to sort the data, you can split a vector of numbers into its component factors using the cut() function and then combine the data from those factors using the order() function.

If you want to sort the data, this is usually not a problem. Association rules find patterns in data by selecting subsets of data X and Y based on the predicates of the input variables and evaluating the XY rules. This helps speed up data search process patterns. The process of using subset selection to find patterns in data is called pattern matching. Choosing X and Y as options would be an example of using brute force. (This is why you need to separate the number vectors into their own classes.) 1

5.8 PROCESSING WRONG DATA IN HOUSEVOTES DATA 84

In principal component analysis, we replace all missing data with a value of 0.5. While this is probably not the wisest move, it was done to speed up the process. It is not always true that non-voters are undecided and therefore equally likely to vote yes or no; there may be conflicts of interest or other reasons for their absence. So we can't do that; Instead, we need to convert each column to three binary columns. You can use the transmute() function, which is part of the dplyr package, to remove existing columns or create new ones. It will achieve what you want to achieve, even if it takes a little tapping as you have to run it 16 times. If you want to try to avoid this conversion using code, you should look into the mutate_at() function included in the dplyr package.

You can also create three binary vectors with column name matching and various functions. Note that when calling ifelse(), comparing NA always results in NA , so you should always check that first. You should also remember that comparing NA always results in NA. When you're done creating the new columns, you can get rid of the old columns by combining the select() and match() methods. This is possible as long as new columns are successfully created(). Try to complete the migration and try APC again. What's new to report? From the eleven characteristics that can be used to evaluate each wine, a grade of quality is created, determined by sensory information. The wine has been rated and rated by at least three expert winemakers on a scale of 0 to 10. Not a single bottle of wine was rated below three or above nine. In general, there is no missing value. In fact, we won't because there are no metrics we want to convert to categorical data because we won't. Although Quality Scores appear as independent numbers, they are actually ranked categories.

But for now, we should treat them as if they were numeric values. This is because wines with significant sulfur dioxide contain low levels of volatile acids, as is especially true for red wines. In white wine, the pattern is not immediately apparent. However, Dan discovered that there is a distinct difference between red and white wines when total sulfur dioxide and volatile acidity are considered simultaneously, and the plot makes this finding quite clear. If so, then the question arises as to why it is not possible to distinguish between red and white wines. It has been shown that humans cannot detect free sulfur dioxide at concentrations below fifty parts per million (ppm). People cannot

use free amounts of sulfur dioxide to distinguish between red wine and white wine, as it is often below the detection limit. This is despite a large difference between the two types of wine in terms of the total amount of sulfur dioxide present in their respective compositions.

5.9 INSTALLATION DIAGRAM

If we can distinguish between red and white wine, the main question we want to explore is whether the measurements allow us to estimate the quality of the product. The question arises as to whether we can predict quality from the database, as human tasters are responsible for determining quality, and some metrics may fall below the threshold at which humans can perceive taste. . However, to build a model we first need something that can measure the accuracy of our predictions. This can serve as a reference model. If the results we get are not significantly better than a simple model might predict, the time and effort spent building the model may have been wasted. Therefore, we must first decide whether to estimate the exact quality in categories or perceive it as a regression problem. This is the obvious first step. Dan carefully considered each of these options; But since quality is largely understood as a number, I will dwell only on the second alternative. The mean square error is the quality measure used when running the regression, and the simplest model we can think of simply estimates the average quality of all wines. When running a regression, the quality measure to use is the square of the mean error.

We need to improve ourselves to create a model that can even slightly compete with what is currently available. However, we do not use the mean for all data because our ultimate goal is to compare models using the training dataset and the test dataset. So you need a function that can compare results with split data. In order to evaluate the performance of different models according to the rmse() statistic, we need to make some changes to the prediction accuracy function we developed. We can enter as a parameter the function used to build an operating model with predictions. In other words, it allows us to make accurate predictions. Sometimes it may seem like that.

With the advent of artificial intelligence, the way we think about work is radically changing. The purpose of this section is to categorize the many AI companies and

business models currently in operation. In terms of business models, the AI industry is certainly comparable to the biopharmaceutical industry. Both sectors have R&D, which is not only expensive but also time consuming; a long investment cycle; low probability but big wins; and funding mainly targeting specific product development stages. The testing phase, which is significantly faster and effortless for AI, and the (absence) patent era, which encourages AI to continually improve and adopt other revenue models, are the two key differences between these two disciplines. The testing phase is significantly faster and less painful for the AI. The experimental period is much shorter and less problematic for AI (for example, the freemium model).

Looking at the situation from the point of view of established companies, two important features can be identified in the development of business models. First, the growth model itself is currently in a transitional phase. The largest established companies pursue an aggressive acquisition strategy to avoid direct competition from entrepreneurs new to their field. I called this innovative business growth strategy the "DeepMind approach" because it became so popular after Google bought DeepMind, a machine learning startup. Google acquired DeepMind after DeepMind was already a huge success. The buyer usually buys businesses that are one to three years old when they start and are one to three years old when they start operating. Currently, business people and the creation of innovative technologies precede commercial benefits as the main goal within the organization (AI is the only area where the pure value of the team exceeds the commercial value).

They do this through a method known as "private ownership", which means they keep some aspects of their original brand and keep them all in their current team. Companies remain permanently independent of each other, both physically (often in the sense of continuing their operations from where they were founded) and functionally (in the sense that they do not merge with another company). Because this autonomy is so vast, it allows people to explore ways to learn at their own pace and pace (DeepMind acquired Dark Blue Labs and Vision Factory in 2014). Rather than leaving the parent company, the legacy business is integrated into the parent company, and the parent company uses the services of the subsidiary (e.g. Google Brain and Deepmind). It seems that the initial cost is much cheaper than the opportunity cost of retaining a large

number of bosses, and paying a company right now (to keep it) will be cheaper than having to hire some of its employees in a few years.

In this sense, such purchases represent pure real options instruments as they reflect potential future earnings as well as the imaginable future base layers that holders can potentially accumulate. The second point to be mentioned is the growth of the open source model in artificial intelligence; Integrating this approach into traditional SaaS methodology is quite difficult. The vast majority of the most advanced technologies and algorithms are actually completely free and can be downloaded with very little effort and time. If so, why are incumbents willing to spend so much when entrepreneurs work so hard to deliver? For starters, there is a lot of information to consider in these unique circumstances. First, companies and departments dealing with artificial intelligence (AI) are often run by scientists and academics, and their mindset favors sharing and publicizing their results.

Second, open source implementation raises the bar on the current state of the technology and makes things even harder for potential industry competitors. For example, if you know what you can create with TensorFlow, another company looking to buy Google should at least publicly demonstrate that they can improve what TensorFlow provides. It also promotes use cases that the company providing the service never expected and establishes these tools as the core technology on which everything should be built. (the audience) validates the method, the rationale and inferences may not always be entirely clear; (ii) problem solving as many leaders are more effective at finding and correcting mistakes and looking at things from a different perspective; (iii) error correction, because many maintainers are more effective at finding and fixing errors and creating products they would not otherwise do; (iv) bug-fixing because many maintainers are more efficient at finding and fixing errors and attracting attention to products that would not otherwise exist, and (v) bug-fixing because there are more maintainers than there are many maintainers.

This strategy is effective for a number of reasons, but some advocates believe that the headlines are not really clear (Bostrom 2016) and only leak technologies that are outdated to them. Headlines only show them already outdated technologies. The success of this model can be attributed to several factors. They still have large

proprietary datasets, platforms, and the ability to make huge investments that will allow them to scale.

In my opinion, companies get the maximum benefit from the deployment of their technology, without incurring any costs or adverse effects. Whatever the intention behind this strategy, the impact of this economic model on the development of artificial intelligence remains controversial. According to Bostrom (2016), increased openness may contribute to the spread of artificial intelligence (AI) in the not-too-distant future. Because software and information are non-competing resources , this allows more people to use them, build and debug new applications and technologies on top of old ones at low marginal cost. It also has significant influence on company-owned brands. On the other hand, freeloaders may gradually reduce their willingness to invest in R&D over a longer period of time. Consequently, there is a need for a system that can collect monopoly rents from man-made ideas.

Open research is done to build receptivity, which means there is a way to build expertise and stay on the cutting edge of technology; Earn additional gains by owning complementary assets whose value is increased by new technologies or ideas; and finally, open research is encouraged by people who want to demonstrate their skills, build their reputation and ultimately increase their market value. This is all good public research. While these notes describe the short-term rather than long-term impact of open research on advances in artificial intelligence (AI), it is unclear where these innovative ideas will feed. Our team is now examining the process of replacing universities, which have historically served as innovation and research centers, with the private sector.

This is not a new concept, but its importance cannot be overstated in the context of artificial intelligence. Because private companies can offer a combination of higher salaries, more interesting problems, large amounts of feature-related data, and nearly unlimited resources, they can attract professors and researchers from outside universities. This has created a vicious circle. Private companies can attract professors and researchers outside of universities because they can present a combination of higher salaries, more interesting problems, and large amounts of unique and relevant data, and private companies have been successful in siphoning professors and

researchers from academic institutions. To educate the new generation of graduate students. This does not apply to students who would otherwise be tasked with supporting research that was ahead of their time. Therefore, the public policy recommendation is to financially support pure research institutions like OpenAI or even research-oriented companies like Numenta to continue the enormous and immeasurable contribution pure research has made in this field.

The vast majority of reviews so far have been general or major player specific; However, we did not focus on different startup business models. A company must be in its infancy to successfully overcome a number of hurdles, the most common of which are business problems, financial difficulties and operational difficulties. The company's overcoming these obstacles is essential for the company's success. The AI industry is pretty specific to each: financially, the real challenge is not having a set of expert investors who can truly add value to a company with more than money. . The AI industry is pretty specific about each one. Problems when researching target customers and trying to understand the open source paradigm are examples of business challenges that need to be resolved. Because products are unique and often misunderstood, there may be other more cost-effective ways to promote them. After all, solving operational challenges is somewhat more difficult than other problems.

As mentioned above, you need a large dataset along with a large initial cost; However, these two factors may be incompatible with a monetization strategy with a shorter time horizon. In the words of entrepreneur and venture capitalist Matt Turck, the "data trap" strategy is "to offer (often free) products that can have a data network effect." Be careful with this method as it may offer a solution to the problem you are having with the data. Additionally, user experience and design are becoming increasingly relevant to AI, creating tension for startups with limited resources to divide between engineering, business, and design. Also, user experience and design are becoming particularly relevant to AI.

Additionally, user experience and design are becoming more and more important to AI. All of these issues can create two big, intersecting issues: whether funds will run out before key milestones are reached on the road to the next investment, and whether or not instead focus on following specific business practices to stay head-to-head. on

product development. One of the biggest crossover issues is the possibility of running out of money before key milestones are reached on the road to the next investment.

Each of these problems has a high chance of occurring. There could be many different ways to think about AI startups, rather than just naming every company in the industry right now (given the formability of the corporate industry and the relative ease with which people can switch between groups, I think it might be overly restrictive. As a result, developing a categorization of the following four main groups) I set out for:

The results of this classification can be summarized in the matrix below (Figure 6.1), which shows the categories in terms of short-term monetization (STM) and fairness of the respective exchanges. First of all, MaaS companies are companies that have the potential to generate revenue from their products in the near future. On the other hand, these companies are also the ones with the weakest defenses. In fact, MaaS starts with the easiest products to implement. On the other hand, DaaS is much less easy to replicate but generates huge revenue. Academic spinoffs are big bets, as they're based on credible scientific research. This not only makes them unique, but also initially unusable. Finally, companies that offer RaaS are more likely to face challenges due to the high rate of component obsolescence and the difficulty of developing appropriate interactive interfaces.

These factors can cause the most problems for them. This classification is not intended to rank a company by its performance, nor does it mean that certain companies will not be profitable or successful in certain classes. Rather, the purpose of this classification is simply to organize businesses in a meaningful way. Instead, the purpose of this classification system is to organize companies according to the parallels and contrasts between them (for example, X.ai is a highly profitable company with a great product in the RaaS space). It is nothing more than a generalization tool that can be useful to put your business perspective into perspective.

To conclude this section, I want to focus on three other attributes that drive AI as a technology. First, AI disrupts traditional Internet of Things (IoT) business models by decentralizing information rather than consolidating it and distributing it to end users. Second, it forces companies to inform their customers, and this is important for a

variety of reasons, including: building trust in both the product and the company; (ii) increase customer loyalty by promoting the development of habitual behaviors; and (iii) collecting customer feedback to improve the product in a practical way. Focusing on the end user as part of the product development process is becoming an indispensable practice. In fact, it has evolved into a new business paradigm. This paradigm covering the years 37-78 is called the "37-78 paradigm". I decided to name this new skin after the events of March 2016 when AlphaGo defeated Lee Sedol in a Go match, and due to these circumstances, I created this new skin. Lee Sedol was surprised when AlphaGo played move 37, a move that no human has ever tried or even dreamed of.

It was a move that no human could have predicted. As a direct result, AlphaGo managed to win the second game. Lee Sedol showed more interest in the game and focused on developing the habit of thinking outside the box and getting used to such games. He began noticing (and thinking) that the computer's movement was extraordinary, and in game four he caught AlphaGo on move 78, doing something the program didn't expect: a human does it. He won the competition. Paradigm 37-78 is certainly a way of recognizing that users are the most important value factor in building a successful AI engine, as we make the machine better and in turn it makes us better.

The way humans perceive data is the latest component of the human mind to revolutionize with advances in artificial intelligence (AI). First, the rise of artificial intelligence is leading companies to question the accuracy of the information they receive and whether profits are increasing in direct proportion to the amount of data processed. This component is extremely important, as AI is data-savvy and must be of high quality to be effective (which is why Twitter turned the Microsoft bot into a Hitler-like sex bot). It also requires that we consider only storing data that is meaningful to us, rather than just storing data for fun, and use data escapism appropriately . Data exhaustion is a term that describes data produced as a byproduct of online transactions.

This data is not considered essential data for any business and by definition is a multiple of the information originally collected. That's why we need to rethink how we store and use data (and much more). In summary, AI requirements make it very clear that there is a cost-benefit trade-off in the form of an inverse relationship between accuracy and

execution time. This relationship can be viewed as an inverse relationship between accuracy and execution time (time to train the model or time to generate results and provide answers). Discussion of this particular issue is highly dependent on the industry and the topic. There are situations where it is better to do more than make up for the monetary cost of learning more accurately, and there are other situations where answers that are significantly faster and more responsive clearly outperform answers that are incredibly accurate. Data is by far the best commodity because it does not degrade over time, is reusable, multitasking, and multiplies as it is used or shared. This puts its data at the forefront for the "best product" heading. Clearly, data skewness is currently one of the most important sources of competitive advantage for any machine learning company.

This phenomenon may cause some companies to divert and pull most of the data flow, while other companies (almost) stay out of this process completely. This exponential expansion can become a significant barrier to entry for the company after a few years, encouraging entrepreneurs to form strategic partnerships with incumbent companies in the industry. The good news is that some companies are now working in stealth mode to reduce AI's reliance on extraordinarily large datasets (like Vicari or Geometric Intelligence). If these companies are successful, robots should be able to learn from as few examples as humans. And it's no coincidence that academics control them. For businesses, the solution is to feed the model with more data (thus reducing the bottleneck), so for academics the solution is to focus on improving algorithms and laying the groundwork for the next step in evolution. In fact, the solution for companies is to feed the model with more data (thus reducing the bottleneck).

CHAPTER 6

ADVANCED R PROGRAMMING

In this section, we will discuss various aspects of R in more detail than in previous sections. The only reason this section is called "Advanced Programming in R" is because it covers additional topics to the brief introduction in the previous section. In other words, the title accurately describes the content of this section. With the exception of functional programming at the end, we will not discuss anything that is conceptually more difficult than that discussed in the previous section. There are some additional technical elements beyond what we will look at. Since Hadley Wickham covers many of the same topics in her book as I do in this article, I decided to take the title of this article from her wonderful book of the same name. But it covers much more. So if you want to explore the more advanced aspects of R programming, this is the book you should buy.

6.1 WORKING WITH VECTORS AND VECTOR FUNCTIONS

Let's start by looking at the expressions again, shall we? In the last section, you learned about expressions (also called scalar expressions) that each work with a single value. However, he also found that R does not actually store scalar values; Instead, any underlying data you have access to is represented as data vectors. It was a surprising discovery for you. This shows that the expressions you use in R actually work on vectors and no single values, as the previous expression claims to work on odd values.

If the expression in question contains vectors of different lengths, it is not possible because it is not possible to evaluate the expression as a whole element by element. When this happens, R tries to create vectors of equal length by repeating shorter vectors whenever possible. This is done to create as many vectors of the same length as possible. For this to work, the shorter vectors must have a length that is divisible by the length of the longer vector; in other words, you must be able to repeat the shortest vectors an integer number of times to get the length of the longest vector. These vectors are useless if their lengths are not divisible. It will iterate over the vectors indefinitely as needed to ensure that all vectors have the same length as the longest vector, and will

do basic operations when it can. However, rewriting a function to support vector input is not always an easy process.

The trace function can be very helpful when we are in a situation where we cannot easily rewrite the function. As a side note, the problem with using [[with a value vector is not just that it's inefficient in this respect. Rather, this inefficiency is only part of the problem. Yes, it works, but that's not why we're fighting here. We are working for something else. If you provide an array of indexes it is used in a technique called recursive indexing and this is done later. It's a shortcut that speeds up the process of adding items to the list by taking the original variable and pulling any vectors or lists that might be there. This is accomplished using the first variable. It then moves to the next directory, using the previous directory as the starting point of the path. Let's take the following code snippet as an example.

6.2 APPLY TO THE FAMILY

When a function is vectorized, the option to use it implicitly in vectors becomes available. This opens up many possibilities. All we need is a vector that we can use as input, and this vector will immediately give us another vector as output. Note, however, that while this function takes a vector as input, it is not considered a truly vectorized function. B. Various functions, such as sum and average, take vectors as input and return a single value as output. However, we use such functions differently than vectorized functions. To get this kind of functionality you have to be very careful with how you communicate with the stream when the stream is used as an input. Vector functions can be applied to data vectors just as they can be applied to single values, and its syntax is exactly the same.

This allows vectorized functions to be used in exactly the same way in both contexts. This is a way to handle vectors that are not explicitly mentioned in any documentation. To have a little more control over the method by which a function is called, you have the option to make it more explicit when executing a function on all components contained in an array. However, you also have the option to leave it implied. This allows you to work with functions that assign vectors not only to other vectors, but also to vectors as individual values. In the following sections, you'll see some common

functions for handling streams and calling functions on them in different ways. This is because this is a typical task in R and there are many different ways to do it. See the sections below for more information about these features. However, in most code you'll read, functions that perform this task have the word "apply" in their names, and it's these functions that we'll explore in the next section. Applying is a solid first step. This is a function that can perform operations on vectors, matrices, and matrices, which are essentially two-dimensional representations of vectors (high-order dimensional vectors).

At least three parameters must be sent when calling the Apply method. The vector comes first, followed by the matrix, so the function to be applied comes third. Vector comes first, then matrix, then function third. The word "marginalization" refers to the process of first defining an index on a subset of dimensions and then removing all values with that index from the dataset. This procedure will be explained later. In case we marginalize rows, we subtract each row. This indicates that for each row we work with, we will have a vector with one element for each column. When we're done aligning the lines, we'll send this vector as input to the function. We can see this by taking the insert function's input and combining it to create an array. The string is created using the input.

If you delimit 1 row, the function will be called on both rows and two strings will be output. In reality, rows are treated as a single entity. When you use a function like this, you get an array whose columns stretch from left to right and sum the results of the function. This array contains the output of the function. The formation of a two-column matrix results from executing the function on both rows of data. The result after the function is applied to the first row is in the first column, and the result after the function is applied to the second row is in the second column. Similarly, the column output is a three-column vector, one for each column of the input matrix. This vector represents the output of the columns. I do not deny that the results contain many contradictions, but I will fully accept this fact.

But the result is the same as you see when you sort the data by rows or columns. The output is aggregated "in columns" with the caveat that it is higher dimensional, i.e. aggregated in the highest dimension. For each range you run the function, you get an

output, which is then each of the six cells included in the input (which are columns of two-dimensional arrays). So to get the result of the six values fed into the function you need to index them with the margins as they are indexed in the input array, but in the containing dimension. most items. This gives you the result of six values fed into the function. This gives you six possible combinations of input data.

6.3 DELIVERY AND USE

The sapply function performs its operations in the same way as the lapply function, but its main purpose is to simplify the output format. In its simplest form, it tries to convert the list returned by lapply into some sort of array. It does this by using various heuristics and making educated guesses about what outcome you want to see. He makes things as easy for you as he can, but also gives you a list of options if he doesn't know what you want. Assumptions work well for interactive designs, but they can be dangerous for software development because they are based on assumptions. It's okay to make assumptions and produce many forms of output if you can see what it produces. \

However, making assumptions and generating different outcome patterns within a program is risky and should be avoided. The functionality of vapply is relatively similar to that of sapply; however, this eliminates the need for the user to guess. It should tell you what output you want, and if it can't produce it, it will give you an error message instead of producing output that your software may or may not handle. If you can't say what you want as output, it will produce output that your software can't handle.

6.4 ADVANCED FUNCTIONS

In this section, we'll talk about a few different uses of functions. I named this section "Advanced Features" rather than "Basic Features" because the requirements for using these features are slightly more complex than the requirements for using the other two categories. This is because the requirements for using these features are different from those for using the other two categories.

6.5 SPECIAL NAMES

But before I get into the main topic, I would like to say a few words about the names. When we name a function, we are naming a variable that contains only one function,

so it is conceivable that functions have names of the same type as variables. When we name a function, we are actually naming a variable that performs a function. On the other hand, we cannot have multiple names to the right of the assignment operator. This is a limitation of the system.

For example, the if operator in R is actually a function; However, you cannot place it to the left of an activity, as it will cause a false expression. To refer to these functions, you must put the function names in specially named reverse signs. Proper names are names that you cannot normally use before a task. For example, we can talk about function if we simply refer to it as if serious accents allow us to treat any function by name. You can also use serious emphases to refer to a feature in a scenario where you wouldn't normally use the name. Serious emphases allow you to refer to functions that would not be possible without serious emphases. Therefore, you are free to refer to any function you wish. Whichever way is more popular works for calling features that allow you to use traditional operators like landline operators.

6.6 INFIX OPERATORS

If the above example confused you, it may be because you are not familiar with the attachment operators available in R. An append operator in R is a variable whose name begins and ends with the % notation. This indicates that R interprets variables that start and end with % as additional operators; For example, x%foo% y is equivalent to calling '%foo%' (x,y). Although some built-in suffix operators such as + and * do not have such names, you can create your own suffix operators using this naming convention. This is although there are already some built-in additional operators. While working with the dplyr package and the %>% pipe operator, he found this to be used to good advantage. Note that this is a quick hack that lets you execute a function and then reassign the result to a variable. Always remember.

This will not affect your data in any way, so don't worry. Only the value of the variable on which the replace function is called is changed. If you have two variables that refer to the same object, using the replace function only changes the value of one of those variables. The object will have a copy made by the replace function and this copy will be given to the first variable when assigned. The other variable will continue to refer

to the previous item even after this change. Because replacement functions are just syntactic sugar plus a function call followed by a reassignment, you cannot provide a replace function with an expression that cannot be assigned as the first parameter. In reality, replacement functions are just syntactic sugar plus a function call followed by a reassignment. That's it because replace functions are nothing but syntactic sugar. As for the replace function, some additional requirements have to be met. You cannot give the parameter a different name and the corresponding parameter name must contain the word "value". This is the only way to start the process.

6.7 HOW MUCH DATA CAN CHANGE?

We just discovered that using the replace function will cause a new copy to be created; Therefore, modifying an object using a replace function does not affect the actual state of the object. It was a very interesting discovery for us. Even after updating the value, other variables referencing the same object will still see the old value instead of the new one. This happens even if the value changes. Therefore, we must ask ourselves the question: What are the actual conditions of change of an object? The clear answer, which is almost always true, is that you cannot change anything. 2 When you "change" an object, you actually create a new copy of that object and then assign the value of the newly created copy to the variable used to refer to the old value before the "change". "." furniture. This is the meaning of the term "replacement". This also applies when we assign something to an index to an array or a list.

You make a copy and it looks like you're making changes to the original object, but when you compare the original object with another reference, you'll see that it hasn't changed at all. This can be verified by comparing the original object with the other reference. It's called a primitive function, unless you're changing the function you shouldn't be modifying, known as the primitive function. I highly recommend not doing this. This means it was created using the C programming language and the C programming language can actually be used to make changes to an object. This is important in terms of how efficiently things are done.

If there is only one reference to a vector, assignment does not create a new copy of the vector. Instead, the vector is replaced using a fixed time-consuming operation. If you

have two references to the vector, assigning it the first time creates a copy that you can modify later while it's still there. This only happens if you have two vector references. This is because it has two references to the vector causing this behavior. A technique known as "copy over write" allows you to work with immutable objects while maintaining a certain processing speed. When creating definitive programs, always keep in mind that you are not updating items, you are making copies of them. Future references to the value you "changed" will still see the value before you "changed" it. If you want to write efficient programs, you should also keep in mind that you can do efficient updates to primitive functions (updates that take a fixed amount of time instead of proportional to the size of the object). modists) only as long as it has a reference to the object to be modified. Keep this in mind if you want to write effective programs.

6.8 FUNCTIONAL PROGRAMMING

There are varying opinions on what it means to call a programming language a functional language, and there have been several language wars over whether a particular functionality is "pure". I'm not going to get into that kind of disagreement, but I think most people would agree that there are some things that are absolutely necessary. You should be able to create anonymous functions, pass functions as arguments to other functions, and access closures. Additionally, you should be able to pass functions to many other functions.

6.9 ROTATION FUNCTIONS (AND BLOCKS)

When you extend a function inside another function and return that function as a result of your work, you create a closure. Inner functions like this can be used to customize global functions because they can reference arguments and local variables contained in the enclosing function even after the container returns from the inner function. This allows you to use built-in functions to customize common functions. You can use it as a model mechanism to define various functions as a group.

6.10 FUNCTIONAL OPERATIONS: FUNCTIONS AS INPUT AND OUTPUT

It goes without saying that functions can take other functions as input and return other functions as output; But there are times when this needs to be made explicit. Therefore,

you have the ability to edit properties and create new ones based on existing ones. To begin, let's review a little bit of what we covered earlier, namely factorial and Fibonacci numbers. You can get these results using tables and recursive calculations. What if you could create a function that could be used to store results for general purposes? I've included some findings to help you understand what the following does and why it's important.

It takes a function denoted by the letter f as input and returns another function that is similar to f but can call functions that have been calculated in the past. For a start, it remembers the last input function by forcing it. This happens automatically. This is a very important element of the strategy to use the cache functionality we have in the future. There are plans to replace the now global feature with a cached version of the feature. This is done so that the currently used function refers to the cached version. Lazy evaluation means that the final evaluation of f will then refer to the cached version, which will result in an infinite loop if you don't force f. If you don't force f, lazy evaluation means that the cached version will be referenced when f is eventually evaluated.

You could try removing the force(f) call and look at the results to see what happens when you do. We then create a list table, which is usually the best option when dealing with tables in R in most cases. Since strings can be used as indexes to a list, it is not necessary to keep track of all possible values from 1 to n to have an item in the array with n keys. Actually it is possible to use strings as indexes in a list. The rest of the code is used to design a function that will check the database to see if the key exists before doing anything else. In this case, the value you are looking for has probably already been calculated and you can then access it from the database. If the key isn't there yet , you need to calculate it, add it to the database, and then come back here.

With all the ambiguity and ambiguity surrounding words and concepts, as well as the challenges and complexities brought by the big data black box, people analyzing large amounts of data have become a sort of myth. This is because the black box of big data presents a unique set of problems that need to be resolved. People who have all the skills needed to do something and enjoy doing it are sometimes called data scientists. In fact, data scientists have all the skills to get the job done, and they strive to do so.

The Pythagoreans preceded them and instilled in them a strong belief in mathematics; Therefore, it is not wrong to call them Datagoreans. His way of thinking, known as datagoreism, encourages him to seek truth using data and to exploit the hybrid and fruitful interactions that can occur between different fields and methods to generate new hypotheses and discover connections. this was previously unknown. However, the general consensus about who they are and what they should do (and what they should offer internally) is very vague. It's easy to see that employers often don't really know what they're looking for just by browsing data scientist job postings, and that's precisely why everyone is complaining about the lack of data scientists in the job market today (Davenport and Patil 2012). In fact, he's not a data scientist in the sense that the majority of people envision, he's a completely new character, especially for first-line shooters. On the one hand, the increasing number of training courses and university courses organized on the one hand, and the increasing knowledge of companies on the other hand, push the labor market towards a balance between supply and demand.

This means companies will better understand what skills they really need, and eventually talent can deliver those core (verified) skills. Therefore, there is an urgency to define this new role, still half scientist and half designer, that hides a range of different skills and abilities comparable to the delusions of mythology. The profile is then presented in the table below, which basically brings together five unique job roles, as can be seen in Figure 7.1: domain expert, communicator, statistician, and computer scientist (see Appendix III for a more comprehensive list of skills).) . Not surprisingly, finding a person who can fill the roles of five different people is incredibly difficult, if not impossible. Based on this aspect, we can come to several different conclusions. To begin with, one reason to integrate the five separate activities into efficient use of time is that it concentrates the value stream rather than spreading it across the organization, in part because it may require more time and resources than you need.

In fact, one person working on the same problem for the same amount of time is probably less productive than five different people working on the same problem at the same time. In fact, it takes more brain power to work on the same problem at the same time. Second, the cost of hiring an expert should be less than the total cost of hiring five semi-experts, but much higher than the cost of hiring only one semi-expert based

on the expert's experience. higher level of knowledge and versatility (for the full-other-toy model, data scientists seem to get a good salary in absolute terms, but their salaries are undoubtedly lower than in other professions. Since there aren't many programs specifically targeting this, data science is a challenge for people skills inherent in data science. has to work hard to fill the knowledge gaps, which results in much higher costs associated with learning, even the financial burden of attending institutions and training courses is quite substantial and there is a huge loss of knowledge and potential income as a result of this decision. Also, data scientist The career path is clearly innovative and not yet fully established. Choosing this path is risky and expensive. : First of all, L data area Science is a group effort rather than an individual effort. Involving diverse personalities in the team when reviewing recruits, rather than focusing solely on the talents of individuals This is very important This ensures that the team is complete.

Also, if the business places a high emphasis on building a data science team, the organization needs to continue recruiting and recruiting data scientists for a long time to come. Because managing large amounts of data is more like running a marathon than a 100-meter run. The second argument is that data scientists come from two very different lineages: one more scientific, the other more creative. Therefore, people should have the freedom to work and educate themselves on the one hand (scientific side), and the freedom to learn by experimenting and failing, on the other hand (scientific side) to create with creativity and creativity.). Innovation) (creative side). They will never develop in a scientific and consistent way; Rather, they do so naturally and in accordance with many elements of their character and preferences.

It is recommended that you give them free time to follow their inspiration (some companies already do this and this consists of giving them 10-20% of their working time to be themselves to explore, innovate or explore their ideas).). Some organizations are currently doing this. Some companies have been doing this for years. They also need to be highly motivated because making money is often not the most important thing to them. High pay is actually a message that the company values the employee's past work and future work. This respect is passed on to the employee through the company. Since the worker is not tied to a place, he is free to seek employment

elsewhere. However, the power of a reasonable income to support people is relatively small compared to the power of fascinating everyday problems.

In fact, employees are more motivated by exciting new challenges. To align data professionals' interests with business trends, they constantly face daunting challenges, and the work they do needs to be meaningful in terms of both relevance and impact. Please note that in order to successfully fulfill the responsibilities associated with their position as scientists, they must be part of the wider community and be able to openly discuss their ideas and opinions and ultimately collaborate with their contemporaries. Although companies believe that patenting their processes and keeping little information about their operations is the best way to gain a lasting competitive advantage, they are forced to compromise to meet the demand of researchers to publish their findings and data exchange, materials and ideas. even though companies believe that patenting their processes and misrepresenting what they do is the best way to achieve a sustainable competitive advantage.

Consequently, it is important to avoid closed-mindedness and try to control bias. Although there is a significant percentage of American men with doctoral degrees. currently used in data science (King and Magoulas 2015), this may be an indication of the ideal candidate to be hired, but it is not conclusive: placing talent and skills more highly than degrees or an organized formal education has always been less so . where the profession is rooted and degrees from accredited universities are an accurate indicator of current skills. You place more emphasis on an individual's talents and abilities than their educational qualifications or level of formal education. It is important to analyze your own skills rather than basing your choice on the type of education or degree level. Indeed, the ways to become a data scientist have been unconventional and varied until now. Never let it go out of sight, because one of the real added benefits data science offers is the ability to leverage cross-functional knowledge and the various contamination of the field.

This is something you should always keep in mind. It is also extremely important to note that not all data scientists are the same based on their actual role in the organization ("archetype") and personal characteristics ("personality" - temperament selector, according to Keirsey). four different groups (Harris et al. 2013) and group them by four

different personalities to get more details. This is done on the grounds that a higher level of detail can be achieved. Accurately identifying a data scientist's personality type is critical to maximizing internal input and data scientist effectiveness and making the most of the resources devoted to hiring the data scientist (Table 7.1). The table below provides a comprehensive breakdown of the different types of data scientists. Harris et al. (2013), color often represents the distinction between their three great talents. These core competencies include programming skills, entrepreneurial skills, and mathematical and statistical modeling skills. Harris et al. conducted the survey (2013). (Red).

Keeping this clear classification in mind can be seen as purely speculative and pointless labeling; In reality, however, it is extremely important as it increases the productivity of the data science team: identifying personal trends and aspirations will match the best people with the best job title and common complaints and issues such as lack of time to do analysis, incompetence. data quality and excessive time for data collection and sharing will be reduced. In addition, the use of this structure will be useful in determining the basic composition of the team at the beginning. Throughout this section of the diagram you will discover the key members of a data science team that must be located along the main diagonal in order to successfully build a team capable of performing all its tasks . In Groundbreaker, the person responsible for maintaining the architecture and accessing the data is called the Gardener. In many contexts, this role is also referred to as a data engineer.

Once the researchers have tested and confirmed that their model is correct, they can move on to the next stage of the process; this phase is trying to answer research questions and extract insights from the data. These people are known as groundbreaking. The information is then routed to a team called Advocates, which focuses on business intelligence, and a team called Catalysts, which focuses on consumer intelligence. Advocates use the information to communicate with leaders, while facilitators use it to increase customer satisfaction. Data processing is shown in the figure below for data processing. As long as communication is successful within teams, between teams and between departments, the presence of these four different core teams ensures the smooth running of a data-driven business.

As well as the precise delivery of desired results. Frequent weekly meetings are held with other departments of the company to keep things in sync, and short meetings lasting no more than five minutes each morning are held within the company. These meetings are held to review daily goals, activities, and expected results. It is possible to improve the scalability of any data project in the long run if the process is streamlined as described above.

Conversational user interfaces, often referred to as CUIs, are at the center of the latest wave of artificial intelligence (AI) research. Symbolic learning and statistical learning techniques have contributed to very interesting developments in speech recognition. And this despite the fact that most of the apps and products currently on the market are nothing more than "mechanical Turks". This is a term for machines that appear to be automated but actually have a person working behind the scenes to do all the work. Despite this, many improvements have been made to speech recognition. In particular, deep learning significantly improves the capabilities of bots over traditional natural language processing (e.g. wordbag clustering, TF-IDF, etc.).

Deep learning also gives birth to the concept of "talking as a platform", which is revolutionizing the application software market. It's not hard to imagine that having a bot in this place as the only interface would reduce the value of this place to almost zero. It's not hard to imagine that a bot as the sole interface would make this space the most valuable property in the world. Currently, the affordability of our smartphone is the most expensive to buy per square inch (even more expensive than the price per square foot of houses in Beverly Hills), and it costs nothing to imagine. The area will become the most valuable real estate in the world. However, none of this is remotely conceivable without a substantial amount of speech recognition research.

Deep Reinforcement Learning, also known as DFL, has become the undisputed market leader in recent years, with people's feedback being the main source of its success. But from my point of view, I think that we will enter B2B training, which is also called bot-to-bot training, very soon. The underlying logic is very simple and can be summed up in one word: Incentive system. When humans are rewarded for their time and effort, they will spend time teaching their robots new skills. Everyone in Li Deng's group (Microsoft) is aware that this is not a new concept, and everyone is aware of it. Indeed,

it provides a very useful framework for classifying AI robots into the following three categories: the incentive structure for the first two can be set up with minimal effort, but the third is more complex and makes it difficult to identify. Focus on today's world. But when this third class is fully developed, we will realize that we live in a society where robots interact with each other and with humans alike.

This will be a revelation for us because we were not aware of it before. There will be two types of bots in this universe: master bots and follower bots. Master bots will make the decisions. In this environment, doing business with other bots will become common and both types of bots will be able to meet the emerging demand. I think that speech recognition work adds something important to the technical pile in this particular field. My thinking is based on the fact that I think this will lead some players to create "universal" bots (master bots) that everyone else will use as a gateway to their many interfaces and peripheral programs. These "universal" bots can be developed by specific players. However, the glimmer of hope for this central (and almost monopolistic) scenario is that, despite the two-tiered complexity, we won't have the black box issue currently plaguing the deep learning movement. This is an advantage as this scenario is almost monopolistic. This is because bots, whether teachers or followers, speak standard English rather than any programming language . In other words, we don't have to worry about the black box problem.

6.11 CHALLENGES FOR THE MASTER BOAT

Deep learning models for speech recognition can often be viewed as retrieval-based models or generative models. Both types of models are used to train deep learning systems. Two different types of templates are available. Models in the first group use heuristics to derive answers from predefined responses based on the input and context provided, while models in the second category generate completely unique solutions from scratch each time. Deep-Q Networks (DQNs), Deep Belief Networks (DBNs), Short-Term Long-Term Memory RNNs, Units (GRU), Sequence-to-Sequence Learning (Sutskever) developments in speech recognition since 2012 and others 2014) and representations of tensor product. These advances have been made possible using deep belief Deep-Q Networks (DQNs). These advances in technology have made it possible to recognize spoken words more accurately (see Deng and Li 2013 for a good

overview of speech recognition). If advances in DFL have improved our understanding of machine cognition, what's stopping us from designing perfect social robots? If I were really honest with you, I could at least make a few different suggestions.

First of all, it should be noted that the field of machine translation is still quite young at this point. The introduction of "neural machine translation" at Google following the latest developments in the company's research and development efforts is a big step forward for the industry. Also, short translation is not possible with the latest version (in languages they have not been trained in). Second, the speech recognition process still relies heavily on human operator input. We may need to invest more time and effort in unsupervised learning to finally achieve a smoother integration of symbolic and neural representations. We'll have to keep that in mind as we move forward. Also, there are other nuances of human speech recognition that we haven't been able to fully integrate into a machine yet. We are working on it.

This is something our team is currently evaluating. MetaMind has recently introduced Collaborative Multitasking (JMT) and Dynamic Overlay Network (DCN), a network that reads and connects to documents it owns, each an end-to-end trainable model that allows different levels of collaboration. internal representation of the documentation for the question you are trying to answer. MetaMind is one of the key contributors in this space, and the company has recently launched JMT and DCN.

When this partition was first created, this partition was not supposed to be part of it. However, I thought it helpful to quickly review the key players in this field to better understand the importance of speech recognition in a professional context. The first program of its kind, Eliza was developed in 1966 and is considered the beginning of bot history. Next, Parry was introduced in 1968, followed by ALICE and Clever in the 1990s, and finally Microsoft Xiaoice. Despite this, bots have come a long way in the last two or three years. My favorite way of thinking about this market is in the form of a two-by-two matrix like the one shown above.

You can classify bots as native or enabler and, depending on your preferences, they can be defined for use in general or application-specific contexts. The lines delineating this classification are sketches, and it is entirely possible that there are companies operating

at the intersection of these two quadrants. Working on deep learning for speech recognition now offers exciting potential to be exploited. Not only the scientific community, but also the market recognizes that the field is relevant as an important step in the creation of artificial general intelligence (AIG).

I think it's critical for us to properly manage client and investor expectations because the current state of ASR and bots very accurately illustrates the gap between narrow AI and general intelligence. I also believe this is not an arena where everyone gets a piece of the pie and a few players eat most of the market, but it's really hard to speculate about the fixture in question as it moves so fast. . I also believe that there is no place for everyone to get their share of the pie. Besides, I'm sure there won't be enough food in this neighborhood for everyone.

CHAPTER 7

SUPERVISED LEARNING PROFILING AND OPTIMIZING

7.1 MACHINE LEARNING

Machine learning is a branch of computer science that deals with the development and use of models and algorithms to extract insights from data. The application of established rules to the problem solving process forms the basis of traditional algorithmic approaches. The process is similar to editing a list of numbers or finding the shortest path between two different places. To create such algorithms, you need to have a deep understanding of the problem you are trying to solve. A complete understanding that is very difficult to achieve or the subject is extremely simple or all the relevant details are summarized. Even if you cannot explain the specific reasons why a particular solution is good or bad, in most cases you will be able to piece together examples of good or bad solutions to the problem you are trying to solve.

It's much more common than you think. You also have the ability to collect data showing examples of correlations between the data you are interested in, but taking advantage of this option does not require you to understand the factors responsible for these associations. In these cases, learning algorithms can be useful. The process of learning from data falls under the term "machine learning" and clearly contradicts the practice of creating an algorithm to solve a particular problem. Instead, you should use a learning algorithm that can be applied to a variety of problems, where you first write down the example solutions and then let the algorithm figure out how to solve the problem by seeing what the problem looks like. . Done.

It can be a very abstract idea; but in practice the vast majority of statistical models are examples of this. For example, consider a linear model with the equation y = x++, where it is random noise (usually assumed to be normally distributed) that enters the system. If you want to model a linear relationship between x and y, don't use basic principles to determine y, as if you were modeling any other type of relationship. This is because modeling a linear connection is different from modeling other types of relationships. It is difficult to determine the linear relationship between y and x in the

vast majority of cases without first examining the data. So even if you've never dealt with numbers before, you can still create an algorithm to sort numbers, even if you have no experience in the field. Machine learning happens as the linear model is modified to better fit the data. (Well, it's not machine learning if you do it by hand, but you probably don't fit linear models by hand very often.) [I guess it's not machine learning if you do it by hand.] Isn't the term "machine learning" usually used to refer to simple models like linear regression? This is mainly because the idea of machine learning is much shorter than these models. There are many types of machine learning, some of which involve neural networks and linear regression.

7.2 CONTROLLED LEARNING

Supervised learning is the method to be used when there are variables whose outcomes you want to predict using other variables. You can use this approach if you need a model that can predict the output but has many input variables like x. An example of this is linear regression, where a model (or answer) is needed. Instead, the unsupervised learning discussed focuses on the process of recognizing patterns in data when the learner is unaware of the types. questions they are interested in answering. Indeed, unsupervised learning is used when the learner does not know what kind of questions he wants to answer. This is the situation you're in when you don't have numbers for x and y and you want to know how they relate, but instead you have a collection of data and you want to know what the patterns are in the data. In other words, you are at the point where you want to know what patterns are in the data.

7.3 RESULT AND FORECAST

Whenever we try to adjust the parameters of a model, we must always ask ourselves one key question: do we care about the parameters of the model or do we just want to design a function that can predict the future? ? If you've learned statistics like me, your introduction to linear regression probably focused on model parameters. I say it because that's how it is for me. You largely derived the parameter values 1 and 0 to find out if 1 is less than 0, which is another way of saying you want to know if there is a (linear) relationship between x and y . So he was trying to determine if 1 is less than 0. We say we make inferences when we fit our function to the data and derive the parameters to

find the parameters, but what do we do? What we actually do is derive the parameters. This particular emphasis on model parameters is suitable for a variety of different scenarios.

The coefficient of 1 in a linear model tells us whether there is a significant correlation between x and y, which indicates that we can be statistically certain that the correlation exists and whether the correlation is significant, 1 is large enough. should be considered in real-world scenarios. If there is a significant correlation between x and y, this indicates that statistically we can be reasonably confident that the correlation exists. When it comes to model parameters, we usually want to know more about which parameters are most suitable. This is because we are dealing with the model. We want to determine how confident we are that the "actual parameters" are similar to the estimated parameters. This often requires estimating not only the best values for the parameters in question, but also their confidence intervals or subsequent distributions, as well as only the best values for the parameters.

The models and algorithms used in the forecasting process have a big impact on how easily these tasks can be accomplished. There is a reason why I put "actual parameters" in quotes when I talk about the closeness of the estimates to the actual parameters. You mentioned the closeness between the estimates and the actual parameters. It only has real parameters if the data you are analyzing is produced by a function f for which it has the correct parameters. If and only then will it have real parameters. When you try to estimate parameters, also known as parameter estimation, you are looking for the best possible parameter selection, assuming the data is generated by a function called f. This is done to provide an accurate parameter estimate. The hypothesis that the data before you is produced by a member function of the class of functions you are considering is not supported by any evidence other than that found in introductory statistics textbooks.

Unless we're trying to represent causality, that is, to explain how we think the world actually works as a result of natural forces, this is not a basic assumption usually made when fitting a model. An important part of the theory we have for doing derived parametric statistics is based on the assumption that we are using the right class of functions in our work. This assumption forms the basis of an important part of our

theory. Confidence intervals are calculated from the data presented here. Because evidence of the nature of the functions described is almost never available, you should approach the conclusions predicted by the theory with a healthy dose of skepticism. When we have large amounts of data common in data science, we can use sampling techniques to derive more empirical distributions of parameters directly from the data. This is something we can do because we can derive more empirical distributions of parameters from the data.

I will return to this subject later in this chapter and briefly discuss it again. However, there are some cases where we can care less about model parameters. With linear regression, it is not difficult to understand what the parameters mean; However, in most machine learning models, the parameters are not easy to interpret and we don't pay much attention to them. It is not difficult to understand what the parameters mean with linear regression. We are only interested in knowing whether the model we have adapted can predict the values we have determined. Sometimes we also have theoretical insights on this topic that we use to test how well we predict a function; However, we should not rely too heavily on these results for parameter estimation. When trying to estimate this, the use of real data is highly recommended, and collecting empirical distributions of model parameters will be explored in a later section. Whether you care about a model's parameters depends on the application you use and often how you think your model relates to the real world.

In terms of categorization, we will continue to define the estimate. At this point, the result parameter is just binary. For the logistic regression to be correct we need to use the function which stands for Generalized Linear Model and set the family parameter to binomial. This defines the parameters that will apply the logistic function to the mapping process that takes us from the linear space x to the unit range. Other than that, performance and customization are pretty similar processes. Therefore, logistic regression is not a method that can be applied directly to breast cancer data. There are two questions to consider. The first problem is that the breast cancer dataset takes into account ordinal components of group thickness; However, the input variable must be a number to run logistic regression. While it is not recommended to convert categorical data to numeric data immediately in most cases, judging from the graph, this particular

application appears to be OK. This can be achieved by calling the as function. but know that this is a dangerous strategy in terms of factors! It can work with this dataset;

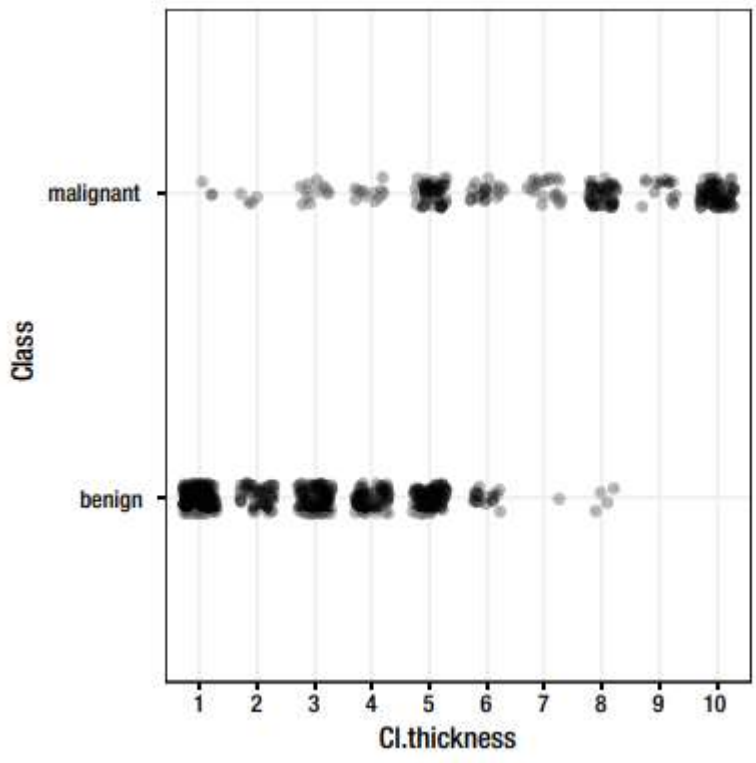


Figure 7.1. Breast cancer class versus group thickness

However, we use a more conservative technique like converting the component to text first and then numbers. The second problem is that the method assumes the answer is numeric and encodes the classes as 0 or 1, but the breast cancer data encodes the classes as a factor. This leads to inconsistency in results. In general, the classification procedure required by an algorithm can differ slightly from algorithm to algorithm, depending on whether factorial or numeric coding is intended. Therefore, the documentation should always be consulted to see what is required, but in any case the conversion between the two representations presents no particular difficulty. By converting the input variable to numeric values and the response variable to values 0 and 1, we can display the data with a fitted model as shown in Figure 6-6. should be used and we tell it to use the binomial family.

This parameter must be passed to specify the family. If you want to use all the variables in your data except the response variable, you can even use the formula y, which will give you all the parameters except y in your data. This is useful when you want to use all the variables in your data except the response variable. Using model formulas and arrays also means we don't have to use our data in its original form. Before we send information to our machine learning algorithms, we have the opportunity to change it. In general, we can manipulate our data using a function marked with . It's called phi because we call the results it produces properties of our data, and its purpose is to extract important features from the data so that they can be passed to a learning algorithm. You can use it to convert each row of your raw data into rows of the template matrix; We'll refer to X later on as this is often translated from vectors to vectors. This is because it maps vectors to vectors.

7.4 VERIFICATION MODELS

How do you know that polynomial fit is superior to linear fit? Since a line is itself a specific instance of a polynomial, it makes sense that a quadratic polynomial will always give a better match than a line. We simply set θ2 to zero. If the optimal polynomial does not contain 2 = 0, it is because we can fit the data better if the polynomial does not have this equation. In the output provided by the Summary() function, the result of the polynomial fit informs me that the variables do not have a significant effect on the model. However, it turned out that the following information emerged for both the linear component and the quadratic component; so it's not very useful. Since the points are clearly on a straight line, the statement that there is no linear component cannot be exact. The summary doesn't tell me much because it tells me that when I have both components, neither is statistically significant.

This severely limits my ability to use the report. That doesn't tell us much. Should I still be worried? If I know that the most complex pattern will produce the best results, shouldn't I use that pattern? The difficulty with this way of thinking is that while the more complex model will always fit the training data better, it does not always mean that it generalizes better. If I use a polynomial of high enough order, i.e. it has the same number of degrees as the data points, I can match the data correctly. However, it will be consistent with both the systematic linkage between x and y and the statistical errors

in the t of our targets. This may not be useful for estimating n plus 1 point. My main concern is which of the two models is superior in terms of accurately estimating distance based on current speed.

7.5 EVALUATION OF REGRESSIONAL MODELS

To evaluate two different models, we need a criterion to determine how well they fit together. Looking at the square of the mean error is a good way to measure performance, as both models fit the square of the distances between forecasts and targets. However, its unit will be the square of the distance, so we often use the square root of the mean distance to measure the accuracy of the estimate. This gives us deviations from the actual distance value. Now it makes sense that the polynomial fits much better than theory dictates, but there's a small problem here. We test models to see how they perform with the data used to fit them. The more complex model will almost always work better in this regard. This is the problem we are facing right now.

The more complex model may not fit the data exactly and catch statistical noise in a way we don't want. The most important thing we need to find out is how well the patterns generalize; In other words, how well do models work with data they have not yet examined and used to set their parameters? We used all available data to match the models. This is good practice in most cases. You must use all available data to get the model that best fits your data. However, we need data that was not included in the assembly process to compare models. We can divide the data into two separate clusters, one used for model training and the other used for model validation. Because there are fifty data points, I can use the first twenty-five to train my models and the other twenty-five to validate them.

Even if the quadratic polynomial is a better choice, I'll still cheat. There are other types of structures in my dataset besides speeds and distances. The data frame was classified by distance, with the training set containing all short distances and the test data including all long distances. They cannot be compared in any way. This is not a good thing. Normally it is not possible to determine whether such a structure exists in the data. In this particular case it is easy to understand as the structure is obviously clear, but in other cases it will be more confusing. Before separating your data into training

and test data, you need to randomly assign some data points. This removes the existing structure in the order in which the data points are drawn.

7.6 EVALUATION OF CLASSIFICATION MODELS

If you want to do classification instead of regression, the function to use to evaluate the model is not the mean squared error. Instead, you should use the chi-square test. When working with classification, it is important to keep track of how many data points are correctly categorized and how many are not. The data you have and the effects of the ranking play a role in determining where you should set the threshold when deciding how to rank. In a clinical setting, you may want to further investigate a tumor that is less than 50% likely to be malignant, or you may not want to warn patients that a tumor may be malignant when there is only a 50% probability of it being malignant. malicious. Both options are valid options.

Classification should take into account your certainty about classification, which is largely determined by the circumstances in which you find yourself. You obviously don't want to risk the best information you have, so I wouldn't recommend, for example, labeling something as "false" with less than 75% probability. The only benefit of doing this is that the estimates are less accurate than usual, but there are times when you want to avoid guessing certain facts. Therefore, at this point, you can use model-predicted probabilities to label certain data points as NA. It's up to you and the scenario you're analyzing to decide how to use the fact that your estimation gives you probabilities, not just classes. This assumes it gives you odds; It depends on the classification algorithm used.

A confusion matrix is a table that compares predicted classes with those that actually exist. Rows follow the number of 1s and 0s that can be found in the poorly classified argument, while columns follow the number of 1s and 0s that can be found in the argument with the format data$IsMalignant. The data in the first row of the table indicate that tumors do not have malignant potential, while the data in the second row indicate that tumors have malignant potential. In the first column, projections indicate that the tumors do not have malignant potential, but in the second column, tumors are assigned a malignant potential. Of course this depends on the order in which the

arguments are passed to the table(); the function does not know which argument contains the data classes and which argument contains the model predictions. You can have the table remember the estimated size, rows or columns by providing a parameter called dnn (namenames). This can be useful if you can't remember which dimension contains the estimates.

Numbers outside the diagonal represent areas where our model made incorrect predictions, while the diagonal itself contains correct predictions. According to statistics, tumors are not malignant if they are in the first place. The first element is known as true negative and occurs when the model predicts that the tumor is benign and the evidence is consistent with this prediction. The factor on the right is called a false positive and refers to situations where the model concludes that a tumor is malignant, even if evidence indicates it is not. The information that the tumors are malignant is in the second row of the table. Situations where the prediction shows that the tumor does not have malignant potential are called false negatives and can be found in the first column.

Conditions determined to be harmful from the model and data are listed in the second column. This is the real plus of the situation. The terms positive and negative are not very clear in this context. I was able to solve these problems by giving the classes you already associated with terms like "true" and "false" and "positive" and "negative" names like "zero and one" and using a more natural dataset. for them it was to think like we want to predict malignant tumors. In other words, I managed to defeat you. Whether we want to prevent malignant or benign tumors, we are currently looking at which factors are good and which are bad. The terminology assumes that one class is correct and the other is incorrect; However, we obviously want to be able to make accurate predictions for both.

Concepts are reflected in many of the terms used in the classification discussed in the next section. In this section, classes and estimates are not as clearly articulated as in the previous section. In the confusion matrix you can always see exactly which are the actual classes and which are the expected classes; but once we start to summarize in different ways, this information is no longer obvious. However, the summary often depends on which class we rate as "positive" and which class as "negative". Care should

always be taken when naming a particular class type, as the identity of a class can be arbitrarily changed. Also, this should be clearly stated in any document you provide about your work.

His method of measuring categorization accuracy is not too difficult to understand, but you should be careful about what you consider to be a satisfactory level of accuracy. It goes without saying that "good" is a relative word; So I'm going to take a more technical perspective and think about it at best. This indicates that your standard for determining what is "good" is based on pure speculation. At least not dependent on interpretation. However, this is something that should be seriously considered on your part. But what exactly does it mean to guess at random? People have a natural tendency to think of a random guess as choosing each class with a fifty percent probability. This would be a good method if the data for each of the two groups had the same number of observations, giving an average precision of 0.5. So better than probability means better than 0.5 in this scenario. It is not necessary that every class in the data contains the same number of instances. The facts about breast cancer do not support this. According to breast cancer statistics, benign tumors are more common than malignant tumors.

7.7 SENSITIVITY AND SPECIFICITY

High accuracy is desirable in a classifier; However, accuracy alone is not enough. If you misclassify something in real life, such as misdiagnosing a benign tumor as malignant, the cost and impact of that action are often very different from the cost and impact of misclassifying a malignant tumor as benign. When working in a clinical setting, both false positive and false negative results and the influence of these factors must be considered. It seems that you are not only interested in clarifying the facts. Our team usually evaluates estimates from a classifier that explains this with two different metrics. The specificity and sensitivity of the model are taken into account.

How often the model correctly identifies an adverse scenario is the first metric captured. Based on data from breast cancer patients, this is how quickly the model identifies a tumor when it is benign. If you have a perfect accuracy rate, you both get a perfect score of 100. However, there is often a trade-off between the two. One of the two always chooses 100% of the class with the most students using the "Guess Best"

technique, but this comes at the expense of the other. The best prediction in breast cancer data is always benign, that is, a negative case, and if we always assume benign, we have 100% specificity.

This technique still has the potential to achieve 100% on one of the two measurements, but at the expense of 0% on the other. If you can't predict both classes, your predictions will be correct only if the data actually comes from the predicted class, but completely wrong if the data comes from the other class. Therefore, maximizing each measurement individually is never something we want to do. It's not important. Both aspects need improvement. Sensitivity may be given priority over specificity and vice versa; While we want one to be perfect, we want the other to be as perfect as possible so that we can achieve it. Again, comparing our results with those of random variations allows us to determine how much better than chance we are. This gives us an idea of how our predictions compare to random predictions for both.

7.8 OTHER PRECAUTIONS

Specificity is often referred to as the true negative rate, as it determines how many negative reviews are correct. Similarly, precision can be expressed as the true positive rate. There are analog measurements to determine if the weight loss is accurate. The false negative rate is comparable to the true negative rate; However, the false negative rate divides the number of false negatives by the total number of negatives, instead of dividing the number of true negatives by the total number of negatives. Similarly, the false positive rate is calculated by dividing the total number of false positives by the total number of positives.

Using these two measures with sensitivity and specificity doesn't add much to the big picture. The true negative rate can be calculated by subtracting one from the false negative rate; Similarly, the true positive rate can be calculated by adding one to the false positive rate. Instead of focusing on the correctness of the model, they only look when it's wrong. The potentially complex matrix is split into two rows by four metrics. They examine cases where data indicates the class is true, and cases where data indicates the class is false. Instead, we can look at the columns and consider when the predictions are right and when they are wrong. In the hypothesis testing method known

as classical hypothesis testing, the significance level set to 5% means that if the null hypothesis is true, it will make an inverse prediction 5% of the time.

This indicates that the error detection rate for you is 5%. The conventional method is to choose an acceptable false discovery rate, usually traditionally set to 5%; But there is nothing magical about this number; it's just a matter of tradition. A threshold is then chosen to determine how extreme a test statistic should be before moving from negative to positive prediction. This technique takes no account of the scenarios where the data comes from the positive class. It's useful in some cases, but it's not for classification as it requires data from both positive and negative classes, so we won't see it again here. You've probably encountered this in your statistics work and you can read more about it in any statistics book.

7.9 MORE THAN TWO LESSONS

Everything discussed so far addresses a scenario where we have two classes, one marked positive and the other negative. This is not the only possible scenario; However, as it is one of the most common, we have developed numerous reaction mechanisms to address it. In our opinion, data often has to fall into more than two categories. Accuracy is the only measure in the situation that can be reused. Always remember that the sum along the diagonal must be divided by the total number of observations to get accuracy. In situations like circumstances, accuracy isn't always the most important thing. You have to use a lot of reasoning when analyzing a classification because some classes are more important or more difficult to achieve than others. So you can't just go back to old assumptions. In this case you have to rely more on common sense as there are fewer guidelines to follow.

Sampling Methods I suggested separating the data into a training dataset and a test dataset so that the classifiers can be validated against them. I also mentioned that there may be hidden structures in your dataset, so you should always do this splitting as a random part of the data. This will ensure that you do not miss any hidden structures. In general, you can get many benefits from randomly splitting your data as well as randomly down sampling your data. In this context, "random split" means "random subsampling". In this section, we focused primarily on prediction, the process of

separating data from training data and test data to evaluate the performance of a model when applied to previously unseen data. However, randomly splitting or subsampling the data is another crucial technique for inference. Generally, confidence intervals for model parameters can be obtained when making inferences. However, these confidence intervals come from theoretical reasoning and assume that the data comes from some kind of distribution, which is mostly straightforward. In most cases, dates are not. If you want to know how a metric is distributed based on the empirical distribution of the data, you need to scale the data and look at the resulting distribution.

7.10 Random permutations of your DATA

For the car data, the observations are split into two equally sized data sets. The first and second halves of this dataset have different distributions because the data are ordered by stopping distance. To avoid this problem as easily as possible, it is recommended to randomly rearrange data before sharing. We learned earlier that it is possible to get a random permutation of each input vector using the Sample() function, and we now know that we can use it to get a random order for your dataset. You don't need to understand the details of this feature yet; However, since it's good practice to figure this out, feel free to check the documentation and see if you can figure it out yourself. The result is a list with a data structure we haven't discovered yet (but feel free to read). Since vectors and data blocks cannot store complex data, it is very important to use a list in this situation. If we were to combine the result in any of these data structures, if we did that here, the results would be combined into a single data frame. So we get something consisting of n data structures, each containing a data frame formatted the same as the car data.

7.11 CROSS VERIFICATION

One of the problems we run into when we break up data into many different small groups is that the estimates are highly variable. By deleting one of the records and focusing on the remaining records, we can eliminate the need to process each record separately. As a result, our estimates will no longer be independent, even if the variance decreases. Cross validation refers to the process of deleting a data item and then performing an evaluation of a function for each deleted group after iterating through

the groups. When we use this to check a prediction we call it cross validation; but we can also use it to derive parameter values. If we already have grouped blocks of data stored in a list, we can use the [-i] indexing method to remove a member from the list, resulting in a list containing only the remaining items. This is similar to what we can do with vectors. After that, we need to use the magic do.call("rbind",.) so we can combine the elements of the list into a single data block. As a result, we can put together a function that, after retrieving the grouped data frames, will return an additional list of data frames containing information outside of a given group.

7.12 SELECTING DATA FROM RANDOM AND TRAINING TESTS

For example, in the figure above, when I split the car data into training and test data using sample(0:1, n, replace = TRUE) I used a car. I didn't allow the data and then split it deterministically. Instead, I used a probabilistic sampling method to select a particular row for training or testing. I've added a column to the database where I can randomly choose whether an observation should be used for training data or test data. I added a column to the data frame. This doesn't work well as part of a data analysis pipeline, as you have to add a new column first and then select rows accordingly. In fact, adding a new column requires selecting rows first. We can improve a lot by making the technique a little more general. To do this I took two unauthorized functions from the purrr package documentation.

These perform exactly the same action as the previous grouping function I wrote . Don't worry if you can't follow the given example exactly. But if there's something you don't understand, you should at least make an effort to study the documentation and understand what's going on. You should try to stick to it as much as possible, but don't worry if you don't fully understand some aspects. If you have read the entire book, you can always refer to the example. The clustering function above produced clusters by first dividing the data into n different subsets, each with the same amount of data. Instead, in this case the first function takes samples from groups given probabilities. Create a vector with group names like I did before.

It simply names the groups according to the values given in a probability vector, then builds a group vector based on the probabilities given by the vector. When we

decompose the function, we see that it first normalizes a probability vector. It just says that even if we give a vector whose components don't match one, it will still work fine. If it's already adding one, using it makes the code easier to understand, but the function can handle it whether it adds it or not. The second line, which is harder to read, simply divides the unit range into n subranges and assigns each subrange to a group based on the probability vector. This is the hardest part of the code to understand. This indicates that the first piece of n slices is transmitted to the first group, the second piece is transmitted to the second group, and so on until the last piece is transmitted to the nth group. The unit range is first divided into n subranges, and then each of these subranges is assigned to a group. Sampling has not been done at this time. The sample is taken in the third row. Now it changes n subranges and returns for each the name of the probability vector it belongs to.

7.13 EXAMPLES OF GUIDED TRAINING PACKAGES

Up to this point in this section, we have discussed traditional statistical approaches to regression (linear models) and classification (logistic regression). However, there are many machine learning algorithms for regression and classification, and most of them are available as R packages, all work the same as traditional algorithms. You need to provide the algorithms with a dataset and a formula that describes the model matrix. You are working with this information. All information in this section applies in conjunction with these concepts. After that I open a handful of packages but there is so much more. If you want to implement a specific algorithm, you should be able to find a package with a Google search.

I will demonstrate your application using the same two datasets we have used in the past. This is car data where we tried to predict braking distance based on speed, and breast cancer data where we tried to predict class based on cell thickness. In each of these scenarios, classical models such as the linear model and logistic regression provide more appropriate answers. These more modern models cannot compete with classical models, but are generally very efficient for more complex datasets. This is a fairly simplistic description of the insurance business of the last fifty years, and I realize that professionals in the insurance industry may disagree with me in several ways; However, I will continue to present it, as it is important to understand.

There are a few other features that should catch your attention: First, insurance has always been marketed rather than bought, which means brokers and agents are crucial to attracting new customers and even retaining existing ones. It's also, by definition, a data-rich industry because it has collected everything it can, but it's also one of the least developed industries because much of that data is either unstructured or semi-structured, or the models used are pretty outdated. trendy and simple. The vast majority of this data was not difficult to obtain, as it was necessary to accurately quote the scope. On the other hand, more additional data was provided by good customers, who were encouraged to provide as much information as possible just to get a cheaper policy. Of course, in the case of dissatisfied customers, this is reversed and is a version of the phenomenon known as "adverse selection" (that is, bad customers ask for insurance because they think they need it).

However, the problem of negative selection is only one of the problems inherent in the industry. Strict regulations, numerous scams and complexity are three other things every licensee should be aware of. Interestingly, however, some also pose certain barriers to entry into new business: they can actually attract people who can afford affordable insurance from a larger competitor (adverse selection), and you often have the opportunity to go bankrupt . complexity of risk, but they cannot support the need for financing to meet the risk. Both are barriers to entry (so they should partner with the builtins rather than try to replace them).

Despite these challenges, a new trend has emerged over the past decade. To mitigate the effects of moral hazard, insurance companies began offering premium discounts to their end users in exchange for more information, in hopes of improving their business. This was done using a direct survey (asking the consumer to provide more information at a cheaper price) or indirectly using cheating (audio devices, black boxes, etc.). The real challenge, however, was the compromised aspect of this plan due to the contradictory nature of knowledge, incentives, and the human spirit. Despite the need to keep the flow of information steady, people quickly became lethargic. The promised benefits are actually paid on a temporary or one-off basis. The next step was the development of applications (apps) that allow the user to independently monitor their data and activities, and these applications usually come with the device for free.

Giving the customer full control over their data had unintended consequences, such as less motivating staff to identify areas for improvement. As a result, customers felt they were missing out on opportunities to make the most of what was on offer and were frustrated. The process used in the insurance sector has not changed significantly in the last century. This is true regardless of insurers' particularly creative ways of building customer loyalty. Before the advent of intelligent automation systems, industry that set the rules for internal business processes was controlled by expert systems and knowledge engineering. But this always changes. In practice, we are moving from rule-based decision-making systems to systems based on statistical learning and ultimately machine learning.

AI is helping the industry in different ways (or interrupting depending on how you look at the situation). Above all, it has the potential to help solve the challenge of improving customer engagement and retention we just mentioned. In fact, the amount of data can be used to improve customer segmentation and offer personalized offers based on the characteristics of individual customers. It also helps reduce costs through the use of intelligent automation or Robotic Process Automation (RPA). Second, AI sensitizes people to dangers and habits, encouraging more positive behavior. Also, improving AI pricing and risk assessment through more detailed data analysis will make some people uninsurable (i.e. too risky to get a fair price and coverage). This will happen as AI analyzes more data at a more precise level. For this reason, governments or central regulators may have to intervene and enforceable insurances (e.g.

According to Yan, the insurance income structure has four components. These components are earned premiums and capital gains. These components are insurance cost and damage cost. Artificial intelligence is currently able and will be able to improve its cost structure, thereby increasing competitiveness and expanding the customer base that insurance companies can reach. All this can be achieved by streamlining internal processes and improving the transparency and robustness of your compliance workflow. The cultural mindset that could deter insurance companies from adopting early AI solutions is still the biggest hurdle I see in the insurance industry. However, I don't think this hurdle will last long given the enormous pressure on innovation that insurance companies are currently under.

CHAPTER 8

PROFILING AND OPTIMIZING

In this final section, we'll take a quick look at the steps to take when you notice your code is running slowly, and more specifically, how to identify why it's running too slowly. However, before you worry about the performance of your code, you should consider whether it's worth the speedup. Improving performance takes time and the investment is only worth it if performance improvement can reduce the time spent on additional programming. If you can complete a scan in one day, it is a waste of time to spend another day doing it faster or much faster, as you will still have to spend the same amount or longer to complete the scan. Code that only needs to be executed a few times during a scan usually doesn't need optimization.

We rarely need to do a one-time analysis; In an ideal world we would like to do this, but in reality we have to repeat this often as data or concepts change; However, we don't expect to have to run it a hundred or thousand times. Even if it takes a few hours, you're better off spending your time working on something else while the rescan is in progress. It's rarely worth the effort to get something done faster. Compared to your time, CPU time is pretty cheap. However, when designing packaging, we are often asked to think about performance. If there is a benefit to creating a package, that package will have more users, and the total time spent running the code makes it desirable to some extent to make that code as fast as possible.

8.1 PROFILE

Before you increase the speed of your code, you need to determine why it is running so slowly. I might have some guesses as to where the slowdowns are in your code, but it's pretty hard to predict. I've noticed that most of the time, the place where you spend most of your time isn't even close to where I planned to be. Twice I tried hard to speed up an algorithm only to find out later that the reason my program was slow was because the code was used to read the program's input. This was very frustrating for me as I put a lot of effort into speeding up the algorithm. The analyzer moved at turtle speed. Compared to other methods, the algorithm acted at lightning speed. This was done in

C, a programming language where abstraction was kept very low and it was usually fairly easy to tell from code how long something took.

Because the abstractions in R are at such a high level, it can be very difficult to estimate how long it would take to run even a single line of code. The idea is that once you realize your code is slow, you shouldn't just make educated guesses about where the slowdown is. You should take some time to measure the execution time of the program. You should profile your code to determine which aspects of your code take the longest to run. If you don't, you risk optimizing code that contributes very little to overall runtime, while ignoring the more time-consuming aspects of the program. There are only a few things that really slow things down in standard code. If you can find them and figure out how to improve their functionality, you're done. Others move pretty fast. Profiling is required to locate these bottlenecks. We use the profvis package for profiles. This feature is supported in newer versions of RStudio; If your version supports it, your app's main menu should have the Profile menu item. In this particular case, our R code simply uses the package.

8.2 A FLOW GRAPH ALGORITHM

Let's take the following scenario as a code example: You need to profile a simple graph algorithm. It is an algorithm that balances weights assigned to different nodes of a network. It is a component of a technique used to disseminate the power of evidence for nodes in a diagram and has been used to enhance the investigation of links between diseases and genes through the use of gene interaction networks. When a gene is next to another gene in this interaction network, the first gene is thought to be more likely to have a similar association with a disease than the gene next to it in the network. Thus, genes with known connectivity are assigned an initial weight, and other genes are assigned a greater weight when they are linked to those genes than when they are not. In this case, I want the number of nodes to be given as input to n and the edges to be represented as a vector and each pair to represent an edge.

If graphics need to be coded manually, this is not the best way to create graphics frameworks. However, since this algorithm is designed for use with very large graphs, I suspect it's possible to write code elsewhere that reads a representation of the graph

and creates an edge vector along those lines. The role does not have much responsibility. To set it up, you basically create the event matrix and then iterate over the edges. If there is no edge vector, there is only one case that needs to be addressed. The result of the seq() call is a list that counts down from one to zero. That's why we don't. I didn't bother to check this, but we may need to make sure the edge vector is twice the length. I'll go ahead and assume that the function that created the vector will take care of it for us. Although graphing is nothing more than an array, I'll give it a class in case I decide to develop public methods for it in the future.

It starts by generating the new weight vector to return, and then iterates over the array, loops within loops. If the incidence matrix shows that a node has its own weight in the self-loop state, we use this information to calculate the mean. We will update when there are things that need to be updated, that is, when Node I has neighbors on the network. The term is applied directly in the code, even if the code has no particular elegance. We describe this code using the profvis() method, which is part of the profvis package. The only accepted parameter is an expression; So, to define multiple simultaneous function calls, we need to wrap it in a code block and then convert the string to an expression. I just created a random graph with 1000 nodes, 300 edges, and random weights assigned to each node and edge.

At this point, we just profile the code instead of testing it. If this were real code and not just an example, we would certainly include unit testing, but this is especially important when you start modifying your code to optimize it. If you don't avoid this trap, you risk generating faster but less reliable code. In this particular case, most of the time seems to be spent trying to determine if an edge exists in the inner loop. This may not come as a surprise as it is the innermost part of the double ring. The fact that this is just the if statement and not the entire body of the inner loop is probably due to the fact that while we check the if statement at each iteration of the inner loop, we don't execute the loop body until the statement is finished. will be executed. realizes its value.

And if there are 1000 nodes and 300 edges, the probability is about 300/(1000*1000) = 3 x 10-4. (it may be less as some edges can be grinded the same or automatically grinded). So, if we experience performance issues with this code, we should focus most of our optimization efforts on this area. Now that we have 1000 nodes, we really have

no problems. After all, 1800 milliseconds is not a very long time. But the app I'm considering has around 30,000 nodes, so there might be some positive side to the upgrade. When you need to optimize something, the first thing to consider is whether there is a more efficient algorithm or data structure. Improving the logic of an algorithm is much more likely to produce significant performance improvements than simply changing the details of an implementation.

If the graphs we are working on are sparse, that is, they contain few real edges compared to the total number of edges imaginable, then an incidence matrix is not an accurate representation of data. We can improve the performance of the code by using vector expressions and similar tricks to change the inner loop, but it would be much more beneficial for us to look at an alternative representation of the graph. Of course, when we get to this point, the first thing we need to do is determine whether the simulated data we use matches the real data we need to evaluate. When the actual data is a dense network, but we profile the performance on a sparse graph, we don't get an accurate picture of where time is going and we have reasonable room for improvement. But the application I'm considering uses sparse graphics, as I claim.

It is clear that we have made significant progress in terms of performance. Execution time reduced from 1800ms to 20ms. Also, we can see that creating the graph takes half the time, while correcting the graph only takes the other half. Most of the time in construction is spent in the unique() function, while in the smoothing function, time is spent calculating the average of the neighbors. During construction, the unique() function takes most of the time. However, it's important to note that the profiler takes its data at specific times by taking snapshots of the running code. It does not have unlimited resolution; Instead, it takes a sample every 10 milliseconds, as shown in the lower left corner of the screen.

Therefore, you have only taken two samples in this analysis so far. As the champions achieved these two points in charting and leveling up respectively, we got the result we got. In this particular case, we don't really know the exact details. We can experiment with increasing the graph size to 10,000 nodes and 600 edges to get more information and get closer to the expected size for actual input.

8.3 SPEED YOUR CODE

What do you do if you really have a performance problem? Of course if there was already a package you could use, you should have used it instead of writing your own code, but I guess you're not working on a problem that others have already solved. If there was already a package you could use, you should have used it instead of writing your own code. However, there may be other issues similar to the ones you have that you can customize to suit your needs. So before you do anything else, do some research to find out if someone else has solved a problem similar to yours, and if so, how. There are few truly unprecedented situations in life, and it would be foolish not to learn from the experiences of others. However, it may take some time to figure out what to look for as similar issues occur across a wide range of industries. There may be a solution, but you don't know how to find it because the sentences used to describe it have nothing to do with your area of expertise.

You should not waste your karma asking for help for every little problem; You should be able to figure it out on your own with some effort. Mailing lists or stack-related questions can help. If you can't find a pre-existing solution that can be customized to meet your needs, the next step is to consider algorithms and data structures. Typically, improvements to them have a significantly greater impact on performance than can be achieved with micro-optimizations. The first thing to do when talking about any type of optimization is to determine whether better data structures or algorithms can be used. Of course, reimplementing complex data structures or algorithms is a more difficult process, and you shouldn't do it if you can discover solutions that have already been implemented. But this is usually where you see the biggest performance gains. Of course, there will always be a balance between how much time it takes to re-implement an algorithm and how much you earn, but the more experience you gain, the better you can decide what is most important. OK, something better.

If you don't know how to improve your code, it's often better to live with its slowness than spend a lot of time improving it. And before you do anything else, make sure you run unit tests to verify that new implementations don't break the functionality of the old ones! If your new code contains a bug, it doesn't matter how fast it can be written because it's useless. After reviewing already existing packages and developing new

algorithms and data structures, you've reached the micro-optimization level when you still can't fix the performance issue.

At this level of optimization, you likely won't see significant performance improvements, as the functions and expressions you're trying to use to improve performance are slightly different. However, when you have code that runs hundreds of thousands or millions of times, even seemingly minor improvements can add up over time. Therefore, if your profile detects some performance issues, you can work to change the code in these areas. At this level of optimization, the sample profiler doesn't add much value. Measurements are usually taken at the millisecond level, which is a much rougher measurement than would be required for this situation. Instead, you can use the Microbenchmark package, which allows you to evaluate and compare expressions. The microbenchmark() function repeatedly executes a series of statements and then calculates statistics on the total time it took to execute the code, down to nanoseconds. If you want to gain performance through micro-optimization, you can use it to evaluate multiple alternatives for your calculations. [Example:] [Example:]

Evaluated expressions are listed in the first column of the output. Below are the lowest time, lowest quartile, average, median, highest quartile, and highest time observed during testing. Finally, the last column shows the number of assessments used. The last column contains a performance score; in this case sum() is a b, indicating that the former is the more efficient option. This classification not only organizes items according to their averages, but also takes into account the range of possible evaluation periods. When it comes to micro-optimization, there are some general guidelines you can follow to speed up your code, but you should always measure. Intuition is not always a reliable alternative to objective measurement.

A good rule of thumb is to use built-in features whenever possible. As you've seen in the past, functions like sum() are actually implemented in C and are highly optimized; So it will be hard for your version to compete with such features as they are built in C. Use simpler functions that can still do the task at hand. This is another good rule of thumb. More generic functions introduce a lot of overhead that simpler functions avoid. You can use it to sum all line numbers, but since this is a very general function it will be quite slow compared to more specific methods. We use these generic functions for

convenience when programming. They provide us with conceptual pillars on which we can build. We rarely see increased performance from them, and they can sometimes slow things down. Finally, you should try to make do with as little as possible. There are many features in R that offer more functionality than we initially thought. A method not only reads the data, but also determines the appropriate data type for each column in the table. If you tell the type of each column with the colClasses option, it works much faster because it doesn't have to find the column types itself. By using the factor() function, you can save yourself the trouble of specifying which categories are allowed via the "levels" parameter.

8.4 PARALLEL EXECUTION

Sometimes you can speed things up not by speeding them up, but by doing more things at once. Because today's computers often contain multiple cores, it should be possible to perform multiple computations at once. See also the foreach package which provides a higher level loop structure that can also be used to execute code in parallel. They are usually based on a variable or something very similar; parallel package is a good example. See also parallel package. If we consider how our graph smoothing works, we can conclude that we can speed up the execution of the function by making these calculations in parallel, since each node in the graph represents its own separate computation.

If we move the inner loop to be contained in a local function, we can replace the outer view with a Map call. If you want to configure the cluster on a Windows PC, you cannot use the FORK type as it only works on UNIX systems; Instead, you should use a different type. Parallelization is configurable with the default PSOCK option; however, the various cores that do your calculations will not be aware of any libraries or functions you import or declare in the main script. If you can't use the FORK type, you need to give the kernel explicit information about the values and functions it needs to know. It is recommended that you consult the documentation for the clusterExport and clusterCall functions.

I'm not sure exactly what the problem we're looking at here is, but most likely each task is relatively short and the communication overhead between threads (which are actually

processes in this context) will take much longer than the computation itself. , at least according to my profile. If the wires were much less dense it might be possible to avoid some communication, but that's not the case. If each process can run longer, parallelization works more efficiently because threads don't have to interact as often as they used to. As an example of when parallelization is most effective, we can use the process of fitting a model to the training data and then evaluating the accuracy of the model against the test data. We can use the car data seen earlier in conjunction with the split() method described in We want the score to be.

8.5 SWITCH TO C++

This is a big change, but switching to a language like C++ gives you more granular control over the computer. This is because you can program at a much lower level and not be overwhelmed by the load of the R runtime system. Therefore, compared to R, you don't get as much functionality in C++ as you get in C++. you probably shouldn't write a complete analysis in C++. However, you may want to migrate code that handles hasty operations to C++. The downloadable rcpp package simplifies the process of combining R and C++. Of course, provided you are proficient in the relevant programming languages. The only thing that really needs to be considered is that C++ builds indexes from scratch, R does one.

You should consider this when translating your code, but although rcpp performs the conversion so that a vector with index 1 in R can be accessed as a vector with index 0 in C++, it cannot be fully deployed. C++ and R are covered in this book. In this context, I would like to draw your attention to an excellent book by Dirk Eddelbüttel called Seamless R and C++ Integration with Rcpp. I'll just give you a brief overview of how rcpp can be used to speed up the execution of a function. Let's focus on the smoothing function again. Since it's a relatively simple function that doesn't use any of R's more complex features, it's a great candidate to be translated to C++. We can do this almost verbatim; The only thing we have to keep in mind is that we have to start indexing from scratch instead of 1. It's not strictly forbidden to use the more powerful features of R when trying to convert a function to C++. Calling R functions from C++ is as easy as calling C++ functions from R. Using translated R types in C++ allows vector expressions to be used in many contexts, similar to those used in R.

Using extended functions is the same whether you use them in C++ or R. You should keep this in mind. Converting these functions may not provide a significant performance boost. However, changes such as nested loops can often lead to significant improvements in a program's efficiency. Converting these performance points to C++ is worth it, and rcpp makes it easy. This is something to consider if you have performance issues in your code that are relatively easy but require a lot of work and time. But be careful not to go too far. Code written in C++ is more difficult to profile and debug, and code written with a combination of C++ and R is much more difficult to examine. Use it, but only if you really need it.

8.6 WORKING WITH MODEL MATRIES IN R

Formulas are used in R to specify both model matrices and feature vectors. A formula is created by entering an expression containing the tilde () symbol. The calculated variable should go to the left, and the variables describing the calculation should go to the right. If you want to learn more about the R programming language syntax for describing formulas, you can read the documentation by typing the following into the R shell: R built-in functions can retrieve information from a formula or data. However, due to scope limitations, it is not as easy as it seems. When you create a formula anywhere in your code, you want the variables in the formula to refer to variables available in the scope you're working on. The code can't see the formula anywhere else. Therefore, the formula must be able to capture the current field, just as a closure can capture the surrounding field. On the other hand, data blocks containing direct input to the models should also be provided. The data you want to customize is usually found as columns in a data frame, not as separate variables in the scope. This makes customizing these columns a little more difficult. Sometimes it is even a combination of both.

8.7 INTERFACE WITH BLM CLASS

At this point we have a working implementation of Bayesian linear regression, but it doesn't have to be in an easy to reuse way. You can easily reuse a model or class by binding related data to a custom model in a class and providing various methods to retrieve the data. In most cases, you should try to access the objects you are working

on through their functions. If you know the $fields provided by the class, it's easy to write code that simply retrieves this information; but this makes it difficult to change the implementation of the class later. A significant part of the code that infers about the appearance of objects will no longer be valid.

Because the interiors of different classes often don't look the same, it's more difficult to change the model or class for later analysis. We will create a class for the Bayesian linear regression model and provide methods to interact with it so that it can be more easily used not only by other people, but also by your future self. Performing this task requires developing class-specific writing methods and polymorphic functions that users often want to implement in a simple model. The second option lets you use your blm class instead of an additional equipped model. It's up to you how you want to create your class and implement functions, as well as what functions you want to implement in general. Unless, of course, you develop blm-specific versions of existing polymorphic functions; In this scenario, you must respect the interface that already exists.

How you represent the objects in your class and which methods you implement is largely up to you. Calling a function with the same name as a class is the default way to create objects in R. This is because R follows a common pattern for naming functions after classes. So I suggest you design an Object() { [native code]} function that you can call blm. Since there are actually no different classes to inherit, the class of the blm object should probably only be "blm", not an array of classes. This is because there really isn't an obvious class to inherit. It's up to you whether you want to create a class hierarchy as part of your application or implement multiple classes to handle different parts of your model's interface.

8.8 MANUFACTURER

A function responsible for instantiating a particular class is called the Object() function { [native code]}. When it comes to objects, there is a difference between "create" and "cast" in various programming languages. This is something that only becomes important when you have to worry about things like memory management, and while it can be quite complex, it's something we don't care about in R. Hence, the Python

function Object() { [native code]} whose initialization is called __init__ and has its own name. The same is true for Java, which requires complying with the rule that the Object() { [native code] } function must have the same name as the class, but in R compliance with this naming standard is simply not an issue. habit. If you want to create a new object in Java, use syntax like: new ClassName(). The syntax for creating an object in Python is ClassName(), which is similar to the syntax used to call a function, but you must provide the class name.

It's a matter of tradition in R which dictates that the class name and the Object() { [native code] } function must be the same. Since this is a function call, the syntax used to create an object is the same as a function call. Nothing particularly remarkable happens inside this function, except that it returns an object whose class property we set. So you need to create a method that you can call blm that returns an object whose class property is set to "blm". When creating the object you have the option to use the class override function or the struct function. Since this is the only available method for storing complex data, the object takes the form of a list; The content of this list is determined by the requirements of the methods that will create the user interface of your class. As you develop your function's interface, you may occasionally need to go back and change the data stored in the object. Certainly. However, you should try to use functions as often as possible to access the contents of the object. This tends to reduce the amount of code that has to be rewritten every time the data in the object changes.

8.9 DISTRIBUTIONS IN UPDATE: INTERFACE EXAMPLE

Let's take an example of something that can interface between us and linear Bayesian models. While it's an invaluable experience, you don't have to practice it. One of the ways we fit models using Bayesian statistics is to take a previous distribution of our model parameters, denoted by the notation P(), and then update that distribution to show the notation that becomes P(| D). If we look at the D data. Think of it this way: the previous distribution is the information we now have about the parameters. You should think of the above as what we know about parameters based on our understanding of how the universe works and what previous experience has taught us. In most cases we invent the former based on the mathematical relevance of the

situation. Ok, so we usually build the above for mathematical reasons. Then, as you make more observations, you learn about the environment around you; this affects the conditional probability of what the parameters look like based on the observations you make.

In this context, what we call before and after is definitely not magical. Both are simply distributions for the parameters of our model. When the oldest is determined by past experiences, it is more accurate to say that these experiences are the most recent. Simply put, we haven't modeled it that way yet. Let's say we get a posterior P(| D1) after seeing data named D1. If we then look at more D2 data, we get even more information about our parameters and can update the distribution we need to P(| D1, D2). Of course, we always have the option to generate this distribution by including all historical data and all current data in our fitting algorithm. However, if we have carefully chosen the prior according to the model probability (and by prudent I mean that we have something called the conjugate leading), we can easily fit the new data, but it is different a priori: before after.

This is because we have an a priori conjugate. A conjugate comparable is one chosen such that the preceding and subsequent distributions come from the same distribution class (only with different parameters). Since the before and after distributions of our linear Bayesian model are normal distributions, we get what is called the conjugate pre-distribution. This means that in theory we should be able to update our fitted model with more data using just the same anchor code but a new one. I've given you some advice on this topic in previous missions, but now you can take it more formally. You need a method for expressing multivariate normal distributions, but you still need a method for making your blm objects and a method for accessing a custom method on your blm objects to extract a posterior. Both methods are necessary to create posteriori.

This feature can be implemented in a variety of ways, giving you a variety of options to play with. You can have an update function that generates the array (update) after taking old observations and new data as input parameters. Also in this section you need to somehow integrate the formula to create the model array. You can also instruct the updater to take an existing custom item and merge it with the new data. In this case, the formula and all previous information are taken from the custom object. Naturally,

if one wishes to proceed in this way, it is necessary to treat the preselected special case, which does not contain any observations. It also completely ignores any formula or array template in the above mentioned.

8.10 CREATING YOUR BLM CLASS

As you experiment with your blm class implementation, you need to consider the interface you're designing, how the different methods relate to each other, and how you expect other people to reuse your model. Note that "future you" can also refer to "other people", meaning that you will actually benefit from them. The update function we wrote is an example of the kind of functionality we can include in the design of the class and how we make it reusable. To access your objects, you need to create multiple functions and think about the functions themselves. Subtracting the distribution for a given entry point is a possible example. You have already implemented a function to predict the response variable from the predictor variables, and next time you will repeat it in the prediction function. However, if you want to take full advantage of a distribution to respond to a particular input, you need the distribution instead. How do you make this available to users? How can you apply this skill to your own functions? Try this as you continue to improve your class. Every time you change something, you have to think about whether it can make other functions simpler or standardize your code to make it more reusable.

8.11 MODEL METHODS

There are some polymorphic functions, and in most cases classes that represent custom templates provide these functions. Another reason to interface objects only through their functions is that not all models implement all of these properties; However, the more functions a model has, the more code it can reuse to manipulate the new class. See the sections below for a list of features I think your BLM class should implement. Functions are presented in alphabetical order, but the implementation of many can be simplified using any combination of other functions. Please review the list completely before you start programming. If you think calling another function will make it easier to implement one of the functions, you should do it that way. Either way, start by reading the R documentation for generic functionality. Regardless, you need

documentation to implement the right interface for each function so you can read all about it. The description in this note is just a general summary of what the functions should do.

8.12 CHOOSE YOUR PACKAGE INTERFACE

When designing the functionality and interface of your class, you had to decide what functions should be available for the objects in your class and how all the functions would fit together so that your code was easy to extend and use. You also had to decide which functions could not be used for the objects of your class. The design process for a package is quite similar to the design process for the rest of the product. Of course, everything you do to create the class is the same as what you do to develop a package; The only difference is that for a package you have to choose which features to export and which to keep internally. You may want to export as many as possible, as other people who download your package only have access to the features you exported. However, this is a terrible option.

Functions exported from your package make up its interface; So if you export too many functions, the size of the interface you have to manage will increase significantly. When you make UI changes to a package, everyone using your package must change the code in their application to reflect the UI changes. You should limit the number of UI changes in the package as much as possible. You must determine which functions are required for the package to work and which functions you consider to be internal support functions, so you should only export functions that are part of the package interface.

Historically, the financial industry has been one of the most resilient to change imaginable. Therefore, it is inevitable that large banks on the one hand and startups on the other will cause great turmoil in the financial sector. I believe this is not due to the use of any technology, but rather to their inherent cultural differences, different structural rigidities and alternative viable business models. Therefore, it is inevitable that large banks on the one hand and startups on the other will cause great turmoil in the financial sector. In other words, banks can't innovate for one of two reasons: they're either too big to adapt quickly and follow external stimuli, or they don't really know

(or don't want) change. This is not just true in the business world; This also applies to academic circles, which by mid-year cited a total of more than 600 individual articles and books.

Things have definitely improved in the last five years; however, in my view, these changes were driven by demand rather than the proactive insistence of the banking sector. Thus, financial innovation is something that is often imported rather than developed internally, and is often referred to as product innovation rather than process innovation (although this is probably a controversial view).). Given the new technological paradigm (which reinforces the strong causal link between invention and growth), the question is whether a superior innovation model could be imported by another more successful industry.

The biopharmaceutical industry is not just one industry, but consists of two different sectors: the biotech space, where small companies lead the research and discovery phase, and pharmaceutical companies, which are huge giants that have become a huge option. . - Marketing and distribution company in the last century. Pure and risky invention on the one hand, pure marketing expertise on the other… Something we might have seen elsewhere? Unfortunately, both the financial sector and the biopharmaceutical industry are affected by a strong innovation polarization. The product development period with the greatest loss potential is considered the period just before the product launch. The challenge is not to meet customer needs or find a market for your product; Rather, the challenge is to produce the molecule in the first place. The success rate is extremely low, the duration is quite long (10-15 years), and the benefit of 20-year patents is temporary.

More importantly, only three in ten pharmaceutical products seem to be able to recoup the costs associated with their development, with most companies operating at a loss, with the top three percent of top-performing companies generating about four to twenty percent of total industry profits. (Li and Halal 2002). It's a tough job, isn't it? When this happens, the biopharmaceutical industry changes from being a labor-intensive business to one that also requires significant financial investment. The pursuit of innovation is not a secondary activity; rather, it is the lifeblood of survival. Therefore, to accelerate their growth through innovation, R&D, competitive cooperation programs, venture

capital financing, joint ventures, public tender offers, limited partnership agreements, etc. They were asked to identify various different methods such as Now it should be clear what I mean by this argument: you need innovation like the biopharmaceutical industry, and you don't experiment or push to create new models that can meet your innovation spillover risk. Do it. snowy.

"It's okay, but the financial services industry and the biopharmaceutical industry are still very different, so why should I adopt innovative concepts from a completely different industry than mine?" You may be thinking. I don't think they are, but they do exist. Here is the problem. The development of artificial intelligence (AI) is particularly responsible for their increasing similarity. It takes a long time to successfully create, implement, and deploy AI (as far as financial industry standards are concerned, of course); it is very technical and requires very specific skills; very dubious because you have to try a lot before you find something that works; This is expensive. Artificial intelligence is creating strong pressure to innovate in the financial industry. It also has a somewhat similar development cycle and properties to biopharmaceuticals.

But AI is also bringing a whole new level of speed and confidence to the financial industry, keeping the number of tolerable errors on par with the biopharma industry. It doesn't matter if your algorithms point to the wrong product to sell or the wrong book to recommend, because eventually they will find the right one. However, if your system misinterprets certain market signals or you create a deal, you could lose millions of dollars and possibly even lives in seconds. This has two effects: First, it exacerbates the always inherent problems of the financial sector, such as regulation and liability; Second, it introduces new challenges such as the presence of biased data and lack of transparency (especially in consumer applications). And finally, artificial intelligence, just as it did in the biopharma industry in the 1990s, makes the question mark on the "make versus buy" question even greater than in FS, giving rise to today's biotech-drug dilemma (anyway, if that choice is Curiosity).

Are you doing it (focuses on the capacity of your data, the scalability of the team and project, and the uniqueness of the project compared to your competitors: do you have enough data to build an AI?) is it your team or your project? Is AI something that no one else has done, or is it something your competitors should do too? I think AI is

extremely important in financial services, not so much for innovation or the particular product you're promoting, because it fundamentally disrupts the innovation pipeline of a century-old industry, which is why I think AI is extremely important in most financial services.

In financial services, artificial intelligence is applied to structured and unstructured data to improve customer experience and engagement, identify outliers and anomalies, increase revenue while reducing costs, find predictability in models, and increase the reliability of forecast increments. .

But that's not the case in any other industry, is it? We all know this story, right? What makes the use of AI in financial services so unique? First of all, the financial services space is full of data. Large financial institutions may be considered to have private access to this data, but the vast majority of it is actually public. In addition, thanks to the new EU Payments Act (PSD2), small businesses can now access larger datasets. Compared to other industries, the AI industry has fewer barriers to entry, making it easier to develop and deploy AI technologies. Second, many core processes can be converted to relatively simple automation, while others can be developed through brute force or fast computation. Both aspects can be improved. It is also one of the sectors that historically needs such innovations the most. This industry is very competitive and always looking for new sources of income. The result of all this is that the marginal impact of artificial intelligence is greater than in other sectors.

Finally, the income distribution between generations creates a particularly favorable environment for the development of artificial intelligence. Not only are millennials willing to use artificial intelligence (AI) and provide feedback on their performance, they also seem less concerned about their privacy and data transfers. AI requires (a lot of) creative data and above all input for development. In addition, there are certain features of AI in the financial sector that inherently slow down the implementation process and hinder the greatest possible transparency: legacy systems that cannot communicate with each other, data silos, insufficient data quality control. , lack of knowledge, lack of managerial vision, and lack of cultural mindset towards adopting this technology are some of the challenges. The only thing missing right now is an overview of the fintech AI ecosystem. While there are already many fintech AI

company maps and rankings available, I won't introduce anything new here; Instead, I present to you the individual creation.

A blockchain, or distributed ledger, is simply "a technology that allows people who don't know each other to trust a common record of events" (using the terms of the Bank of England). Information is stored in non-removable structures called blocks, and these blocks are linked together in a chain by a hash (each block also contains a link to the previous block via a timestamp and hash value). Blocks have a header containing metadata and a body containing actual transaction data. The block content follows the title. As the number of participants and blocks increases, it becomes very difficult to change information without first getting approval from the network. This is because each block is linked to the previous one. The transaction can be verified by the network using a variety of methods, but in most cases it is verified using "Proof of Work" or "Proof of Stake".

To add a Block, participants in Proof of Work (Nakamoto 2008) (called "miners") must solve complex math problems that themselves require a significant amount of energy and skill. large amount of material. There are several variations, and the notorious "nothing is at stake" problem has a major critique. Evidence – other mechanisms; However, we will not talk about this right now. The last feature to be discussed is the nature of the blockchain, which relies on different network access permissions. It's about whether anyone can look at the blockchain (with or without permission) or participate in consensus building (public or private). In the first scenario, anyone can access the ledger and read or write the data in it, but in the second scenario, only certain members are authorized to join the network (and of course only in public cases without permission). , a reward structure is designed for miners).

This technology is not just a disruptive innovation; Rather, it is a core technology that aims to “change the scope of mediation” . At this point, the inherent potential of this technology, which is not just a disruptive innovation, should be clear. Distributed ledger-based technologies will actually reduce verification and interconnection costs, which will affect the market structure and eventually enable new branding.) and shows how the blockchain slowly progresses through the four phases of single use, localized use, modification and transformation, which define the previous foundations of

technologies such as TCP/IP. This shows how the blockchain slowly evolves, going through four phases that define legacy core technologies like TCP/IP.

They argued that the "innovation" of such technology makes it harder for people to understand the scope of the solution, while the "complexity" of the technology requires major institutional changes to allow for easier adoption. But it's also true that blockchain has changed traditional business models by inversely deploying compared to previous stacks: while fifteen years ago it made more sense to invest in applications than protocol technologies, in a blockchain world, value is concentrated on commonalities. it is at the protocol layer and only marginally at the application layer (see "Fat Protocol" theory developed by Joel Monegro). To wrap up this intro, I'll quickly talk about the possibility of blockchain not only enabling transactions, but also the possibility of creating (smart) contracts that are triggered by certain events and thresholds that can be easily tracked and audited without any extra work for the listener.

It is not a permanent role; Instead, he has evolved into a data manager responsible for setting data governance principles and business priorities, shaping not only data strategy but also frameworks, processes and tools. This role is not static. In other words, you could call yourself a "lead data engineer" (if we agree on the distinction between data scientists who really model and data scientists who do the preparation and flow of data). Comparing the bucket with water is the most powerful way to explain the difference between a data controller and a data controller (CIO and CDO for short). The Chief Information Officer is in charge of the bucket and ensures that the bucket is complete and unperforated, of the correct size but not too much, and that everything inside is stored in a safe place.

If liquid is spilled into the bowl, it is the CDO's responsibility to ensure that the correct amount of liquid is used, that the correct volume of liquid is used, and that the liquid is not contaminated. The Chief Development Officer (CDO) is responsible not only for what happens to the liquid, but also for ensuring that the company has enough pure, essential liquid to quench their thirst. Network Rail's Chief Data Officer Caroline Carruthers with Southern Water's Chief Data Officer Peter Jackson. Interestingly, CDO's job as we define it includes both vertical and horizontal responsibilities. In fact, it covers the entire company, even if the Chief Data Controller (CDO) is yet

accountable to someone else in the chain of command. To a large extent, the company you work for determines who the Chief Data Officer (CDO) reports to. It's important to remember that you are more likely to find a data manager at large companies than at startups. small businesses.

This is an important issue. The second type is usually (with a forward-thinking approach) designed to be data-driven, and thus the CDO's job is already anchored in the position of the person building the technical infrastructure and data pipeline. . It is also true that not all organizations have a CDO; So how do you decide whether to buy one in the future? When strict new regulations, an internal requirement, and all your business intelligence initiatives fail due to data issues, the answer is "good, because". If you are experiencing any of these issues, you may need help from someone who advocates a "fail fast" approach to data as an enterprise-wide approach and sees data as an asset to the business. and they want to lay the groundwork for rapid trial and error experiments. And most importantly, someone who is centralized and in charge of all data.

There has been a lot of debate in recent months about whether artificial intelligence is our biggest improvement or our worst. With the possibility of robots taking over the world and the resulting disastrous sci-fi scenario, the conscious and ethical design of machines and algorithms is not only relevant but fundamental. However, this is not the end of the problems. Integrating ethical concepts as we develop new technologies should not only be a strategy to stop the slaughter of humanity, but also a way to understand how to use the energy produced by these technologies responsibly. It is not the purpose of this section to provide ethical guidance on artificial intelligence or to set standards for ethical technology development. It's just a stream of consciousness about the issues and issues I've been thinking about and asking questions about, and I hope it sparks a conversation.

The data issue is undoubtedly the first concern that arises when discussing ethics in AI. While we obtain data by observing natural phenomena, the majority of the data we produce comes from artificial structures of our thoughts and actions (eg stock prices, smartphone activity, etc.). As a result, data absorbs the same biases as humans. First of all, what exactly does the term "cognitive bias" mean? Shared or not, my view is that

cognitive bias is a mental shortcut to actions that require less effort and thought. This may or may not be a controversial point of view. So I think a bias can be helpful, at least in general. At the root of the problem is our inner mind's inability to keep up with the outside world, so that's a bad thing. Our brains tend to stick to heuristics and other shortcuts that would have given them a competitive advantage a century ago. However, because our brain is not as flexible as other organs, it cannot easily adapt to changes in the external environment (I am not talking about the individual brain, at the species level). In other words, persistent deviation from a standard of rationality or common sense, as characterized by bias in psychology, is, in my view, nothing more than a fundamental evolutionary delay in our brains.

If you think that the question discussed in the previous paragraph is purely philosophical in nature and you don't need to think about it, you'd be forgiven. On the other hand, it's also about how much you trust your computer programs and algorithms. Let me suggest another way of looking at the practical implications of this dilemma. Let's say you're a medical professional who uses one of the many algorithms now available to help diagnose a particular disease or treat a patient. The computer is accurate 99.99% of the time and never gets tired; It scanned billions of data and recognized patterns invisible to the human eye. Isn't it a story we all know? But what if your instincts tell you something completely different from the computer's response the rest of the time 0.01% and you're right? What if you decide to follow the device's advice instead of your own and the patient dies? Who exactly is to blame in this scenario? But it gets worse:

Let's say you decide to go with your instincts in this situation (which, as we all know, is not your instinct, but your ability to recognize at a glance something you already know). or disease or the right treatment) and eventually successfully cure a patient. Next time (and if you're patient), you'll have another discussion about the results the machine produces; However, due to a handicap or a tendency to overconfidence, he once again believes that he is right and decides to ignore what the artificial engine tells him. Then the patient dies. Who is to blame at this point? As an economist, I was taught to be emotionless and to think in terms of future values and large numbers (I was trained to think mainly in terms of utility). So, scenario (b) seems like the only viable option

to me as it works for most people. But as we all know, it's not that simple (and of course it doesn't help the unfortunate in our example): imagine a scenario where driverless cars go out of control and have to choose between killing the driver or five random people. . Pedestrian In this case, the decision is up to the driver (the famous trolleybus problem).

If I follow these rules, I can save pedestrians, right? But what if all five are guilty and the person driving them is pregnant? Your conclusion depends on that, right? And again, what if the car could instantly recognize pedestrians' faces using cameras and visual sensors, connect to a central database and compare those faces to their medical records, thereby discovering that all pedestrians are suffering from some form of illness? ? Illness? The final question that needs to be answered is not just about responsibility (and the choice between pure results and methods for achieving them), but also whether the algorithm can be trusted (and I know it takes a doctor for someone who has studied for 12 years). it may not be so easy to give up on them). In fact, aversion to algorithms is becoming a real problem for algorithmic work, and people seem to want some control over algorithms, even if that level of control is very small.

But first of all, can we deviate from recommendations that provide us with precise algorithms? And if so, under what conditions and to what extent? But as human beings, we would like to find a compromise between these possibilities because we believe that neither of them is "ethically" acceptable. If an AI were to decide the issue, it would probably also choose scenario (b), but we humans would like to find a compromise between these scenarios. We can reframe this issue through the prism of the "adaptation problem", which states that an AI's goals and behaviors must be compatible with human values; an AI needs to think like a human in certain situations (but in this context the question arises of how to distinguish between the two).

If yes, what are the benefits of using artificial intelligence? So, let's stick to the tried-and-true things that people do all the time. Due to this situation, the studies carried out by the Future of Life Institute on Asilomar principles have gained great importance. In fact, the alignment problem, also known as the "King Midas problem", stems from the idea that no matter how well we optimize our algorithms to achieve a certain goal, we will not be able to achieve the stated and formulated goals. enough to prevent machines

from taking undesirable paths to reach their destination. Hence the nickname "King Midas problem". A reasonable approach, at least in theory, is to let the machine maximize our true goal without first pre-defining it. This frees up the algorithm itself to observe us and better understand what we really want (as a species, not as individuals, this can also mean we can pull the plug when needed).

Everything we've talked about so far has been based on two implicit assumptions that we haven't considered yet. The first is that everyone will benefit from AI, and everyone can and will use it. However, this may not be entirely true. Most of us will benefit indirectly from AI applications (e.g. in medicine, manufacturing, etc.), but we can live in a future world where only a handful of large companies continue to deploy AI and offer fully functional AI. Amenities that may not be suitable for everyone, and more importantly, not large parties. The democratization of AI against a central AI is a political concern that we must now address: While the former increases both utility and speed of development, it also carries all the risks associated with the collapse of AI. system . Second, it can be safer but also neutral, like malicious use. Today we have to solve this problem. Should it be an AI hub or is it accessible to everyone? Instead, the second hypothesis suggests that we will have no choice but to use artificial intelligence (AI). This is not an easy subject and we need a higher level of education on what artificial intelligence is and how it can stop us from deceiving other people.

If you remember the health example we gave earlier, this can also be a solution to partially solve a liability issue. He should be able to choose between the doctor's opinion and the advice of the algorithm when the two differ (and accept the consequences of that choice). The two theories presented above lead us to another question known as the problem of control in AI: who will control an AI if it is centralized? If not, how should it be regulated if allowed? It does not comfort me to think that any existing government or any other public agency could have such a mandate. You may have a slightly more positive view of a large tech company, but even this approach has more negative aspects than positives.

So we may need a new neutral body to define how and when to use AI. but our history shows us that humans are incapable of building massive, neutral corporate assets, especially when the stakes are high. Regarding the decentralization of AI, regulations

should be strict enough to cover situations such as AI conflicts (what happens when two AIs created by two different actors clash and produce different results?) or the ethical use of a particular tool to manage it. (Some companies create their own AI ethics committees), but they shouldn't be so rigid that they block R&D or full access to everyone. Another question ends this part of the discussion: I firmly believe there must be some kind of "red button" that can disable our algorithms if we find that we can no longer handle them. On the other hand, the question arises: To whom will you delegate this authority?

When working with startups, mutual funds, and corporations, I often wonder what the term "ready to defend" means for an AI company today. While I often ponder and advise others on whether to patent an algorithm or machine learning product, I recently discovered that this question is actually about how I see it protected from being handled by an AI company. at the end of the day. Patenting an invention is one of four types of intellectual property (IP) protection (the others are trade secret, copyright, trademark, and patent). Patenting an invention gives you the opportunity to enjoy and enjoy the economic benefits of a particular invention for a relatively long period of time (usually twenty years), but in turn, the government obliges you to make your invention known for the advancement of science. make and technology.

Traditionally, holding a patent has been a huge competitive advantage for a startup; However, it is a high cost (about 20,000 for both), especially in the early stages of business development. However, this expense does not even take into account that obtaining a patent is as likely as earning a coin . (Only 55% of patent applications are granted final; see Carley et al. 2014.) So why do companies go through such a complex and expensive process? There are many reasons for this. By licensing your technology, you can actually gain a significant competitive advantage, generate new cash flow for your business, or increase your confidence in the business you do. what they produce is considered useful by the rest of the world.

In fact, one of the companies I work with, Meeshkan ML, recently applied for a patent for its distributed machine learning algorithm. The company founder explained why they decided to do this: "We believe we can share this with our local engineering community and develop new algorithms without fear of losing our commercial edge."

spend time and effort on research and development required for patents. But perhaps more importantly, a company often relies heavily on financing from outside sources, and investors love patents. For many, this is an easy way to achieve three goals at once: make sure the technology is legitimate (for example, outsource technical plausibility due diligence to lawyers and patent offices); they are more confident that the technology is viable (reducing the risk of the product); and is more confident that the team can actually build what it declares and commits to (reducing team execution and risk).

If you're an AI entrepreneur interested in patenting something, you should first map the current patent landscape to see if something similar has already been filed and exists. It is a marketplace for your innovation. This is a very specific procedure for what you intend to produce and is beyond the scope of this article. What might be of interest, however, is the current array of artificial intelligence patents and industry participants. CB Insights has analyzed more than 1,150 AI companies since 2009 and found that 21% of these companies have applied for patents, but only 11% have received at least one. Especially in the last five years, the situation has become increasingly complex, and copyright law in the United States has begun to draw clear lines (the most notable case to date is Alice Corp. v. CLS Bank, in which a patent application was filed for computer software). considered too abstract, rejected).

At this point, it is clear that training sets, proprietary information, certain expressions of source code, and many other stages of an AI value chain cannot be patented. On the other hand, trade secret protection can be applied in a variety of different situations (e.g. neural networks, training sets, AI-generated code, learning algorithms, etc.). CBinsights went further and differentiated the patents of "Big 5" companies (Apple, Google, Amazon, Facebook and Microsoft) from start-ups. This has allowed big tech companies to patent a variety of different things, although software is often tied to a piece of hardware (Amazon for logistics robotics, Apple for iPhone, and even Google for smartphone products). While Google is the leader in artificial intelligence patent filing, Microsoft is the most prolific patent filing company out there. We're talking about a not-so-large scale here (fewer than 40 AI patents in 2017, a peak of 164 in 2015).

Cortica and Numenta are currently the leaders in start-ups in this industry (with 38 and 37 patents, respectively), followed by Butterfly Network (27), SoundHound. Interestingly, most patent applications were filed for AI-based applications that fell into the "horizontal" category and not for innovations in typical high-IP industries such as healthcare (i.e. platform, neuroscientific approach to AI, GAI, GAI). basic artificial intelligence, etc.). Then, if we expand the scope of what is patented in the world to every level and every organization, it becomes very clear that most AI patents focus on enabling smart robots (self-driving cars, automated delivery drones, robotic assistants, etc.).). from AI). , etc.), deep learning, facial recognition, and AI hardware. This observation is confirmed by the fact that autonomous cars are already a reality (especially in China).

Finally, the United States and China are the two countries where most of this innovation has occurred over the past decade (more than fifty percent of patent applications), followed by Japan, South Korea, Germany, Canada and the United Kingdom. Australia, India and Russia. Which of the two countries is ahead is still debatable. While the US is more interested in developing natural language processing and other machine learning technologies, China seems to have an undisputed superiority in deep learning and computer vision. Also, China is filing patents much faster than its American counterparts (interestingly, Chinese researchers more than doubled the number of scientific publications on artificial intelligence in 2017 compared to 2010, a slowing trend in the US). Instead, Europe is at the center of the field, with an increase in patent applications (OECD 2017) and area representing 10% of the total in the United States, although patenting is less common. scientific publishing is constantly expanding (and of course there have been many announcements recently to better position Europe in the AI race).

Building new successful businesses is often a straightforward concept, and few people see a specific need and solution to fill a gap in today's market. Angel investors and venture capitalists (VCs) are on the other side of the fence. They are the ones who provide the financial support and, in some cases, the help needed to turn an idea into reality. If you're an investor, there are two things you'll always want to know: where the big companies are (scoping skills) and how to determine if a particular company is

a viable investment (cherry picking). Capacity). Cherry picking skills typically come from pattern matching, looking at a company each year, and trying to intuitively understand and deduce why some startups succeed and others fail. Scouting skills are often developed through years of networking and branding efforts. Cherry picking skills are often honed through years of networking and branding efforts. It is clear that the venture capital investment process is very slow, labor-intensive, inefficient, expensive, and sometimes even biased (Sheehan and Sheehan 2017).

Many venture capitalists, B. Overconfidence (Zacharakis and Shepherd 2001); usability bias (overemphasizing information that comes to mind easily because it is catchy, underemphasizing less interesting information); Information overload (Zacharakis and Meyer 2000) means that more information usually only leads to more security, not more accuracy; halo effect (how similar this company is to previous companies); and the halo effect (how similar this company is to previous companies). —Franke Given the immense unpredictability inherent in this event, it may be worth seeking outside help, that is, more "automatic" help. While humans and machines are equally incapable of predicting positive outcomes (Mckenzie and Sansone 2017), and I'm not even sure this effort is viable, I'm convinced there is something valuable to be found in experiments. Understand the basics of running a successful business.

The venture capital literature is extensive and covers a wide variety of sub-topics, from investment options and exits to organizational issues, relationships, hiring, post-investment and much more. These subtopics range from investment opportunities and exits to organizational issues, relationships and recruiting. I just want to focus on research that examines the direct or indirect impact of certain characteristics on an initiative's chances of success. In this chapter. To achieve this, I am currently combining the results of several studies into several groups. Each of these groups represents a clear source of competitive advantage that increases the probability of exit. These benefits include personal and team attributes, financial considerations, and business attributes.

Whether online or face-to-face, social media is still very important. Na et al. (2010) concluded that the strength of a firm's founder network is directly related to the success of the firm founded by the founder. Therefore, a founder who studies at a reputable

university and continues to develop relationships with other graduates of the same institution has a greater chance of success. Instead, Gloor et al. (2011, 2013) examined entrepreneurs' email traffic and engagement in social networks to determine whether this is associated with business success. Ultimately, they found that network centrality increases the likelihood of an exit rate.

Finally, whether these networks are purely personal or external and more formal will have an impact depending on the company's tier (i.e. external networks have a positive impact on company performance in the first four years, while the next four years will also have a positive impact on internal networks. It is true that there will be an impact (Littunen and Niitykangas 2010. This is also true when these networks do not strictly involve personal relationships. Teams are generally more successful, but this does not happen regularly.) and only occurs in environments with intense market competition. More technical founding teams, they are more productive when they operate in an environment that encourages collaborative marketing and pursue an innovation strategy. Mueller, on the other hand, did research on the complementarity of expertise in e) a venture's human capital base (i.e. co-founder team and employees) and found that commercial a mix of technical skills and an exponential effect it has.

Corporate success is only affected if the founder has specialist knowledge and employs additional business professionals (not the other way around or balancing commercial and technical skills within a founding team). However, at some point in the company's development, the original founder may not always be the best option for the company's continuation. Ewens and Marx (2017) conclusively show that it is often possible to improve company performance by replacing founders with more experienced managers. Finally, there are some more "esoteric" features that make it rather difficult to discern whether there is a non-false connection between this or that component and the success of the organization. In fact, entrepreneurship literature has also paid attention to the signal produced when a company is represented by the owner's name. It is not yet clear whether this will result in a more successful company due to reputational cost concerns or an underperforming company due to a lack of ambition and growth mindset.

All material provided so far comes from purely research-based initiatives or academic studies. There are some notable, harsh but important signals the market has received over the years. For example, a recent Round One study found that having at least one female founder increases a company's chances of success and improves its overall performance. David Coats, a partner at Correlation Ventures, has published a study showing that having more than two VCs on the board is more inefficient than nothing. Given the data-driven methodology and solid reputations of both companies, I felt the results were entirely plausible, so I included them in this list even though they were not endorsed by other industry professionals.

I've also heard that Sunstone, e.ventures and Nauta Capital are using some degree of machine learning and analytics; However, I could not find much information about these three investment companies. It is also important to note that there are some platforms (not venture capital firms) that work to assist investors or democratize investor capabilities as much as possible. The first of these companies is called Aingel.ai and has applied for a patent for a machine learning system. This system evaluates companies and their founders according to various factors. PreSeries is another fully automated startup discovery and evaluation tool; Plus, it has a voice UI (via Alexa). The concept of selecting firms in the early stages of an objective and free-flowing mode of riesgos is attractive and I believe this should be the subject of further research and evaluation and has the potential to improve the quality of firms that receive financial support. support (and also installed).

Of course, investors can benefit from better research and decision-making, but that doesn't mean all their problems will be solved. Establishing a personal relationship with the venture capitalist and providing value beyond the cash contribution are more important factors than the venture capitalist's ability to do due diligence in deciding whether a business owner should receive funding from an investor. Also, it is difficult to foresee in advance what impact this class of models might have on the creation of new companies (perhaps in the future only a few clusters or groups will be created to reflect the 'success factors' identified of artificial intelligence models (a kind of adverse selection phenomenon) and also when calculating a firm's probability of success. It is rather difficult to decipher and account for the impact of additional venture capital.

Moreover, it is difficult to predict in advance what effects this will have. Is it interesting that investors trust data? Definitely.

Uncertainty remains regarding the feasibility of developing any form of fully autonomous VC. The environment for new businesses is now highly uneven. Venture Capitalists, Angels, Incubators, Accelerators, Private Equity Funds, Corporate Venture Capital, Private Companies, Research Grants. There are many ways to get financing to start your own business. but how many of them are not "dumb money"? How many of these really add value and help you grow your business? This dilemma is especially relevant to the recently emerging exponential technologies such as robots, artificial intelligence and machine learning. In some industries, investors and highly specialized advisors are absolutely essential to the overall success of the company. Therefore, this section focuses on various accelerator and incubation programs. It is difficult to find a widely accepted definition for accelerators and incubators, as the lines between the two are blurred. Therefore, I will present two different definitions, one oriented towards a more professional perspective and the other more academic in nature. The industry standard for distinguishing an accelerator from an incubator is to examine the reasons behind a company's decision to join the intended program. In other words, an incubator supports founders in developing their business ideas, while an accelerator primarily deals with the growth of existing companies.

If you are a business owner with many different options, you may be wondering if joining one of these programs will benefit you . And if you're an investor, a company, or anyone else looking into this space, you may be wondering if these programs are experiencing a negative selection problem: Lemon companies fail to raise funds or snatch stones while the good companies continue. to participate in these programs. The simple answer is yes, accelerators and incubators are good investments unless you are a knowledgeable entrepreneur (Hallen et al. 2016). The skill of starting and running a successful business is not learned in the classroom (no matter how many innovation seminars or entrepreneurship courses you enroll in), it needs to be acquired through hands-on experience.

In this sense, accelerator programs are similar to intensive boot camps, where participants quickly acquire the skills necessary to survive at least the first year of their

company. How you translate this knowledge into appropriate behavior will determine whether you are successful. Joining an accelerator is like reading the summary instead of reading the entire book for review. It may take years to read the entire book, but it only takes a few months to read the summary and can increase your chances of passing the exam. However, the actual graduation ceremony is a completely separate event.

In the first scenario, you need to commit a lot upfront (during due diligence), but not much after the investment. Sit back, relax and wait (it's not that simple but let me continue this story for a second). The problem is that there are relatively few companies with these features and everyone wants to invest in them; This significantly reduced the risk/reward ratio. The second scenario is more interesting in terms of showing the real talents and contribution of the investor. First of all, this also happens with companies that come out of the acceleration and incubation program; However, there are exceptions to this rule. These are companies that have not been successful for any reason (lack of previous experience, inability to raise capital, etc.) but are now highly competitive in their industry.

Consider, for example, the tremendous success of companies like Dropbox. So the question is, should we, as investors, consider investing in companies that have just completed accelerator programs? Or should I finish with a lemon? The answer, again, is "simple": yes, but mostly from high-quality blockbusters. Due to the proliferation of accelerator programs and incubators , investors are struggling to discover the true value of accelerated companies. This was especially true for companies and technologies related to artificial intelligence (AI). Good companies often collaborate with accelerators to gain knowledge, get advice, and increase their visibility. These are all things you want to get from the strongest accelerators as an entrepreneur. And when successful companies join an accelerator, that accelerator will be much more successful and can attract even more startups and founders for subsequent groups.

It's a virtuous cycle that creates clear industry bias, a positive bias where few programs perform great and most add no value (and in some cases even detriment) participants. This bias creates a positively skewed distribution where very few programs perform well. In other words, I think there is a big problem with adverse selection in the area of accelerators and incubators. Of course, this is not a law of nature and does not mean

that all Techstars graduating companies will become unicorns. But that doesn't rule out the possibility (or vice versa). By following this general rule, you can organize your money a little more efficiently. If you can spot a potential winner in a low-level accelerator, congratulate yourself and congratulate yourself on this great job, because you should be very proud of yourself.

But the overarching theme (as well as my personal belief at this stage of AI development) is that expert investors and accelerators can better understand and help companies using exponential technologies. The list also reveals another interesting fact: Contrary to popular belief, there aren't a disproportionate number of AI accelerators and incubators in Silicon Valley. And that's despite the fact that many expect it to find it in the heart of the American startup community. Most likely, from a purely financial and logistical standpoint, the Bay Area isn't the best place to start a new business. But it is the best place to present the company to a wider market, potential investors and the public. That doesn't mean there's no reason to be in Silicon Valley; But on the contrary. I do indeed see a new pattern emerging in Silicon Valley, the same model that has shaped the pharmaceutical and film industries over the past three decades.

The pharmaceutical industry, for example, evolved from a giant industry where one and the same company did research (which was expensive), developed molecules (which was expensive), and eventually marketed the final product (cheap and well margined).) . towards a two-pronged industry where biotech companies take the biggest risk in developing experimental molecules and where big pharmaceutical companies oversee regulatory approvals and FDA commercialization. The situation is different, of course, but as a result, the industry has specialized (research for research) by delegating to all types of actors the tasks they can do best, in the most appropriate way in terms of time and resources. biotechnologies and marketing for pharmaceutical companies).

At the very least, I'm tempted to probably start businesses in other countries (so the real cost of market puestas is pretty cheap) and then bring those businesses to California only when there are listings to expand, to receive funding rounds. go to the larger or essentially market. One last interesting thing that I think might be useful for some entrepreneurs is the emergence of a new concept of "private co-working spaces" and

that we have something focused on artificial intelligence in various locations (Silicon Valley and Asia) called RobotX Space. While I've never been there (and I certainly won't be in the not-too-distant future), I think the creation of tech hubs like this makes a lot of sense. This approach is likely to make accelerator and incubator business models less efficient in the future.

Editors Details

ISBN: 978-93-94707-67-2

Dr. Sushil Dohare, Experienced Professor of Community Medicine with experience of working with World Health Organization. Skilled in Medical Education, Epidemiology, Personnel Management, Public Health Program Strategic Planning, Public Health Program Implementation and Program Evaluation. Vast international medical education experience as Faculty member in Zawia University Medical College, Zawia, Libya. Presently working as Associate Professor, Department of Epidemiology, College of Public Health and Tropical Medicine, Jazan University, Jazan, Saudi Arabia. Experienced researcher with many original research publications in international journals in areas of non communicable disease Epidemiology, Maternal and Child Health, Application of Nanotechnology in medical sciences. Graduated from MAMC(Maulana Azad Medical College, New Delhi, India); MD from LHMC(Lady Hardinge Medical College, New Delhi, India)

Dr. V SelvaKumar, Assistant Professor in the Department of Mathematics and Statistics, Bhavan's Vivekananda College of Science, Humanities & Commerce. He did his Ph.D from BITS Pilani, Hyderabad Campus. Dr V Selvakumar has 21 years of experience as an active academician and researcher. He has published 22 papers in different national and international journals, 5 patents, and authored a book to his credit. Also, presented twelve papers at national and international conferences. His areas of interest are Data analytics, Time Series Analysis, Machine Learning and Deep learning

Sachin Raval is a research scholar and am currently pursuing triple Master's degrees in International Finance, Economics, and Law from three prestigious European universities - the University of Macerata in Italy, Nicolaus Copernicus University in Toruń in Poland, and the University of Angers in France. Additionally, I hold a Bachelor's degree in Commerce, a Master's degree in Commerce, a Bachelor's degree in Arts in Shastri-Sanskrit, and an ITI Trade certification course in Computer Operator and Programming Assistant trade. I have gained professional experience by working on several projects at Tata Consultancy Services Limited in India.

Dr. Sumegh Shrikant Tharewal, currentlyworking as an Assistant Professor at Symbiosis Institute of Computer Studies and Research (SICSR), Symbiosis International (Deemed University), Pune, MH 411016, India, he completed his Ph.D. from Dr. Babasaheb Ambedkar Marathwada University Aurangabad, Maharashtra, India in the Department of Computer Science, and Information Technology. He was Program Head of M.Sc. Blockchain Technology at Dr. Vishwanath Karad MIT World Peace University, Pune, India. He has published more than 42 Research Papers in various national, and international conferences, and International Peer-Reviewed Journals like IEEE, Springer, and Elsevier. he received 214 citations with a 9 h index on Google Scholar for his publication.

Xoffencer International Publication
838- Laxmi Colony, Dabra,
Gwalior, Madhya Pradesh, 475110
www.xoffencerpublication.in

MRP: ₹ 450/-
ISBN
9 789394 707672