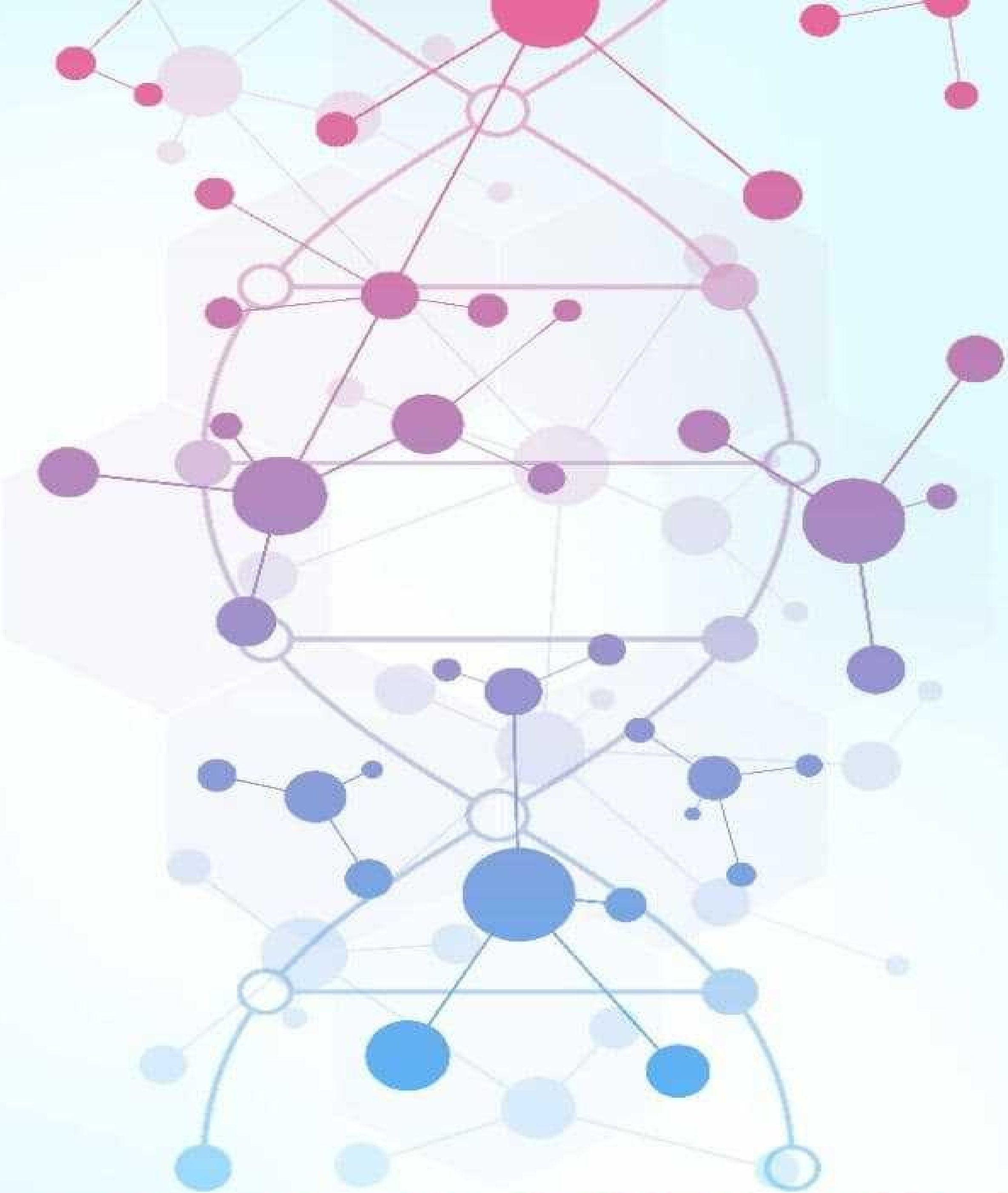
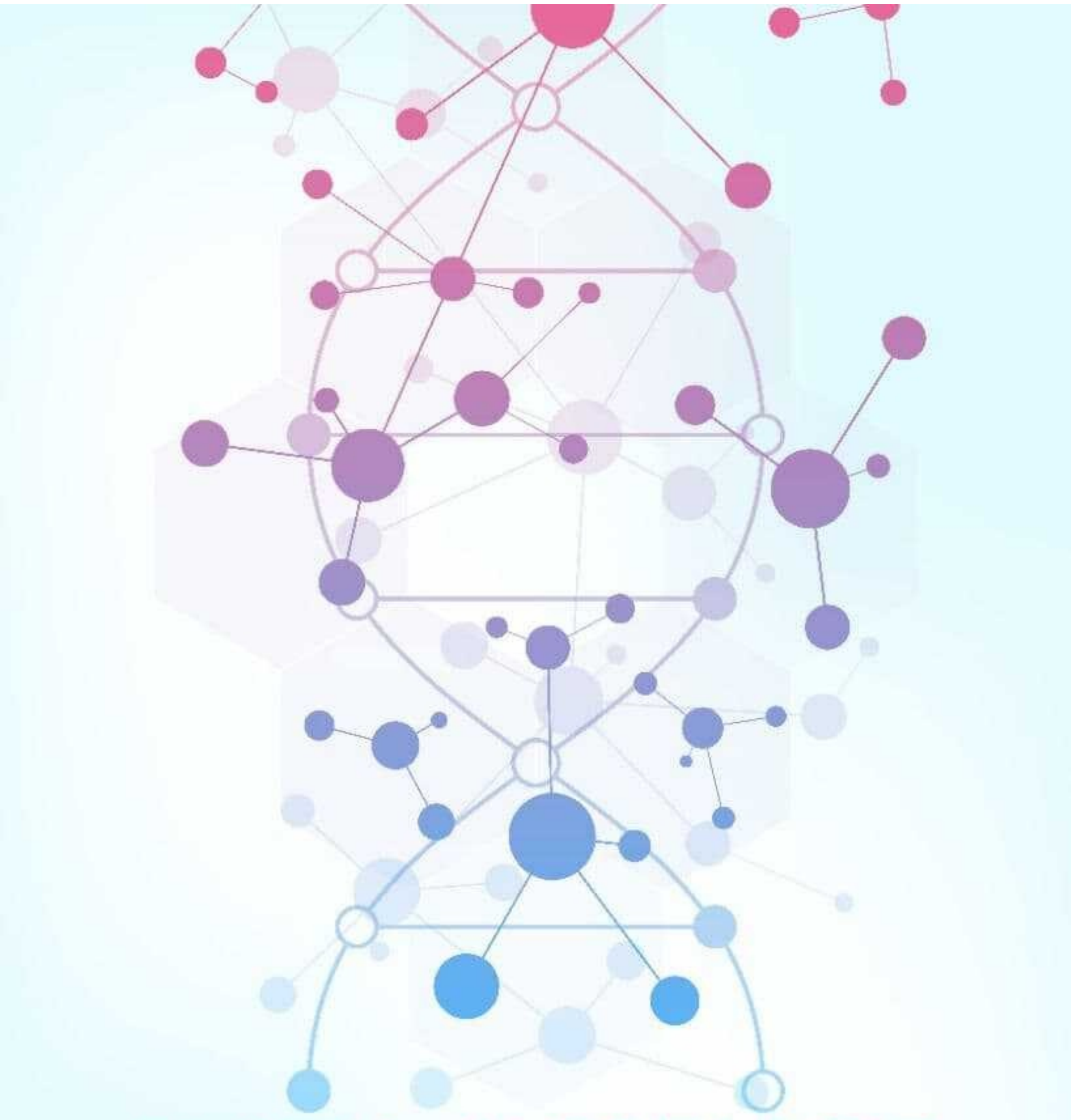


DATA SCIENCE

FOR BEGINNERS

Comprehensive Guide to
Most Important Basics in
Data Science

Alex Campbell



DATA SCIENCE

FOR BEGINNERS

Comprehensive Guide to
Most Important Basics in
Data Science

Alex Campbell

Book Description

Do you wonder what the fascination is around data these days? How do we obtain insights from this data? Do you know what a data scientist does? What is artificial intelligence and machine learning? Are these the same as data science? What does it take to become a data scientist? If you have ever wondered about these questions, you have come to the right place!

There are many resources and courses online that you can use to learn more about data science, but with so much information available, it can become overwhelming. One of the best ways to learn about data science is to understand different machine learning concepts, statistics, and artificial intelligence to help you design models to perform an analysis.

This book has all the information you need to learn what data science is, and what the prerequisites are to become a data scientist. If you're a beginner or if you already have experience in data science, this book will have something for you.

In this book, you will:

- Learn what data science is about.
- Discover the difference between data science and business intelligence.
- Explore the tools required for data science.
- Find out the technical and non-technical skills every data scientist must have.
- Figure out how to create a visualization of the data set with clear and easy examples.
- Get advice on developing a Predictive Model Using R.
- Uncover detailed applications of data science.
- And much more!

The book has been structured with easy-to-understand sections to help you learn everything you need to know about data science. In this book you will learn about the prerequisites of data science and the skills you need to become a data scientist. So, what are you waiting for? Grab your copy of this comprehensive guide now!

Data Science for Beginners

Comprehensive Guide to Most Important Basics in Data Science

Under no circumstances will any legal responsibility or blame be held against the publisher for any reparation, damages, or monetary loss due to the information herein, either directly or indirectly.

Legal Notice:

Disclaimer Notice:

Please note the information contained within this document is for educational and entertainment purposes only. Every attempt has been made to provide accurate, up to date and reliable complete information. No warranties of any kind are expressed or implied. Readers acknowledge that the author is not engaging in the rendering of legal, financial, medical or professional advice. The content of this book has been derived from various sources. Please consult a licensed professional before attempting any techniques outlined in this book.

By reading this document, the reader agrees that under no circumstances is the author responsible for any losses, direct or indirect, which are incurred as a result of the use of information contained within this document, including, but not limited to, —errors, omissions, or inaccuracies.

Table of Contents

References

Introduction

In today's world, organizations and people focus more on big data and artificial intelligence. It may seem surprising that over 2.5 exabytes are created and collected by people and organizations each day. This means the volume of data has risen significantly over the years. Most companies have changed their business model and centered it more on data. Some organizations have also added new departments in the firm to perform data analysis. Statisticians would need to analyze the data quantitatively in the past, but this is not enough since the results of the analysis could only talk about the present. When strong computing processes, cloud technology and analytical tools came into existence; people began to use them to perform analysis. They began to develop models to analyze data.

Let us first understand what data science is before we delve into various aspects of data science. In simple terms, data science is a branch of mathematics and statistics to obtain useful and meaningful insights about the data set and trends from the raw data or information. You can process and manage the data set using programming, business and analytical skills. This sounds tough, does it not? Most people do not know how to work with data science or understand how to develop skills effectively.

This book deals with all the information needed to learn and work with data science techniques and tools. Since the world is turning to make decisions every day, it is important to learn more about what data science is and how it helps make predictions. This book aims to give you a brief of what data science is and how you approach a problem.

The field of data science goes back to its roots in statistics. Having said that, this field is a combination of programming, business acumen and statistics. It is important to learn more about each topic, so you have an idea of how you approach the learning process. The art of finding any hidden insights and trends from the data set goes way back. Ancient Egyptians analyzed census data to help them collect tax efficiently. They also used data analysis to forecast when there could be floods in the Nile. It is important to learn from past data to identify a trend or insight in the data set. This helps the business make informed decisions.

It is no longer a secret that data scientists are in demand, and if you enjoy working with data, you should pick this field. If you learn data science, you need to grab an opportunity to work in this field. Employees skilled in data science make it easier for the company to make informed decisions. The number of data science jobs has increased over the last few decades.

A data scientist is expected to develop the necessary technical and non-technical skills. If you want to excel in this field, you need to develop these skills. The book sheds some light on these skills, which will help you identify any patterns in the data set. Using these skills, you can identify hidden patterns and insights in the data set. You can also understand the different changes they may need to make in the existing products and services to improve the business's revenue.

With the information in this book, you will learn everything there is to know about data science. You will learn more about the various subjects and frameworks used in data science. So, let's get started!

Chapter One: Introduction to Data Science

Data has become the new oil, and every company, regardless of the industry, is looking for ways to manage and store large volumes of data. This was a challenge for most companies until 2010. Every company's objective was to define a framework or solution that allowed them to store large volumes of data. The introduction of Hadoop and other platforms has given organizations an easier way to store large volumes of data because of which they now focus on methods and solutions to process information. This can only be done using data science. It is important to note that data science is the future of technology. It is important to know what data science it, especially if you want to add some value to your business.

Introduction to Data Science

Data science is a mix of numerous algorithms, tools, principles, and languages to identify the hidden patterns within the variables in the data set. This may lead you to wonder how this is different from what has been done on data for years. The answer is that earlier we could only use tools and algorithms to explain the variables in the data set, but using data science, it becomes easier to predict the outcomes.

A data analyst uses the data only to explain what is happening in the present using historical data set. On the other hand, a data scientist only looks at the data to obtain any insights from the data set. He also uses some advanced algorithms to identify the probability of the occurrence of an event. He looks at the data from various angles and aspects.

Data science is used to make informed decisions based on predictions made using the existing data set. You can apply numerous analytics to the data set to obtain this information. We will discuss these in brief in the subsequent sections.

Predictive Casual Analytics

If you want to develop a model predict the possibilities or outcomes of a futuristic event, you need to use predictive causal analytics. Let us assume you work for a credit company, and you loan people money based on their credit. You are going to be concerned with your customers' ability to repay the amount you have loaned to them. You can develop models to perform a predictive analysis on the data using the payment history. This can help you determine if the customer will pay you on time or not.

Prescriptive Analytics

You may need to use a model that can take the required decisions and modify the parameters based on the data set or question. To do this, you need to use prescriptive analytics. This form of analytics is more about providing the right information so that you can make an informed decision. You can also use this type of analytics to predict a range of associated outcomes and prescribed actions. An example of this type of analytics is a self-driving car. We have looked at this earlier as well. You can run numerous algorithms on the data collected from the cars and use the results to make the car more intelligent. This makes it easier for the car to take the right decisions to turn, slow down, speed up or identify the direction to take.

Machine Learning

Make Predictions

Numerous machine-learning algorithms allow you to make predictions using unstructured, semi-structured and structured data sets. Let us assume you work for a finance company and you have the transactional data available. You need to develop a model to determine the trend of future transactions. To perform this analysis, you need to use a supervised machine-learning algorithm. Such algorithms are used to train the machine with an existing data set. You can also use supervised machine learning algorithms to develop and train a model to detect future frauds based on historical information.

Pattern Discovery

Not every data set has variables you can use to make the necessary predictions. This is not true. There is a hidden pattern in every data set, and you need to find those patterns, so you make the required predictions. To do this, you need to use an unsupervised model since you do not have any pre-defined labels in the data set using which you can group the variables. One of the most common algorithms used to identify patterns is clustering. Let us assume you work for a phone company, and you are tasked with identifying where to set up towers in an area to establish a network. You can then use the clustering algorithm to identify where you can set up towers to ensure every user in the area receives the optimum signal strength.

Based on the examples above, it is important to understand how data science and data analytics approaches are different. The latter includes the use of predictions and descriptive analytics only to an extent. On the other hand, data science is more about the use of machine learning and predictive casual analytics.

Now you have an idea of what data science is, let us understand why organizations need to use it in the first place.

Why Use Data Science?

Before organizations collected data from every device used, they worked with small volumes of data. It was easy to analyze and understand the data and relationships within the data set using some business intelligence tools. Most traditional business intelligence tools only worked on structured data sets, but most of the data collected today is either semi-structured or structured. It is important to understand that most data collected now are semi-structured or unstructured.

Simple business intelligence tools cannot process this type of data, especially since large volumes of data are collected from different instruments. It is for this reason we need to develop advanced and complex analytical algorithms and tools to process, analyze and draw some insights from the data.

It is not only for this reason why data science has gained popularity. Let us look at how data science is used in different domains.

Customer Service

How great would it be if you could know exactly what your customers want? Do you think you can use existing data to learn more about your customers, such as purchase history, browsing history, income, and age. You may have had this data with you in the past, as well. Since you use different mathematical and statistical models, you can effectively work with large volumes of data and identify the right products to recommend to your customers. This is a great way to bring more business to your firm.

Self-Driven Cars

How would you feel if your car could drive you home? Numerous companies are trying to develop and improve the workings of a self-driven car. The cars collect live information from various sensors, such as lasers, radars, and cameras. to create a map of the surrounding environment. The algorithm in the car uses this data to decide to speed up, slow down, park, stop, overtake, etc. These algorithms are often machine learning algorithms.

Predictions

Let us now consider how you can use data science in predictive analytics. Consider weather forecasting. The algorithms used take data from aircraft, satellites, radars, ships and other parts to collect and analyze data. This helps you build the required models. You can use these models to predict the occurrence of any natural calamities. Using this information, you can take the necessary measures to save lives.

Who is a Data Scientist?

If you look for the words data scientist on the Internet, you may come across numerous definitions. A data scientist uses data science to answer some business questions and concerns. The term data scientist was coined when people learned that a data scientist uses data, various mathematical or statistical functions and operations, and other scientific fields and applications to make sense of the data in the database.

Functions Performed by Data Scientists

A data scientist is one who cracks various data problems using their expertise in specific scientific disciplines. He works with different mathematical, statistical, and computer science elements. He does not necessarily have to be an expert in these fields. That said, he would use some technologies and solutions to develop the right solutions and reach conclusions crucial for the organization's development and growth. A data scientist finds a way to present the data in a useful form when compared to the data available in the data set. They work with both structured and unstructured data.

Let us now look at what business intelligence is, and how it is different from data science. You may have heard about business intelligence, and most people confuse data science with business intelligence. We will look at some differences between the two, so you understand better.

Differences Between Data Science and Business Intelligence

Before we look at the differences between data science and business intelligence, let us understand these terms better.

Using business intelligence (BI), an organization can find insight and hindsight in the existing data set to describe various trends in the data set. Through BI, businesses can take data from both internal and external sources, prepare that data, and run queries on the data set to obtain the required information. They can then create the required dashboards to answer different questions or identify solutions to various business problems. BI can also help businesses evaluate certain futuristic events.

Data science, on the other hand, is a different approach of looking at data. You can take a forward-looking approach and explain any information or insight in the data set. Using data science, you can analyze the current or past data that helps you predict the outcomes. This is one way most organizations do their best to make informed decisions. They can answer various open-ended questions. The following are some features that differentiate data science from business intelligence:

Features	Business Intelligence (BI)	Data Science
Data Sources	Structured (Usually SQL, often Data Warehouse)	Both Structured and Unstructured (logs, cloud data, SQL, NoSQL, text)
Approach	Statistics and Visualization	Statistics, Machine Learning, Graph Analysis, Neuro- linguistic Programming (NLP)
Focus	Past and Present	Present and Future
Tools	Pentaho, Microsoft BI, QlikView, R	RapidMiner, BigML, Weka, R

Now you have an idea of what data science is, let us look at the lifecycle of the data science. Most people rush into using the models they develop on the data sets without understanding the basics of data science. You need to understand these basics and assess the business requirements before you rush into using the model. Make sure to follow the phases of the data science lifecycle to ensure your results are accurate.

Lifecycle

This section gives you a brief overview of the phases in the data science lifecycle.

Phase One: Discovery

Before you work on the project, you need to understand the following:

- Business requirements
- Specifications
- Required or approved budget
- Priorities

If you want to pursue a career in data science, you need to possess the ability to ask important questions. You need to assess if you have the right resources, people, technology, data and time to support the work done on the project. This is the phase where you frame the problem and identify the initial hypothesis you want to test.

Phase Two: Data Preparation

When you identify the required resources needed to work on the analysis, you need to develop or identify an analytical sandbox where you can perform the testing and analysis of the data. You need to process, explore and condition the data before you model it. You also need to perform the following operations to move the data into the sandbox environment:

- Extract
- Transform
- Load
- Transform

Most data scientists use R or Python to clean transform and visualize the data used in the analysis. These programming languages help you to identify the outliers in the data. You can also use the information to develop or identify a relationship between variables. Once the data is cleaned and prepared, you can perform different types of analysis on the data. Let us look at how you can do this.

Phase Three: Plan the Model

During this phase, you need to identify the techniques and methods to help you draw the relationship between the different variables in the data set. These relationships will help you determine the algorithms you can use in the next phase of the lifecycle. To do this, you need to apply exploratory data analytics methods and tools using various formulae and visualization methods. Let us look at some tools used for this below:

- **R** : This programming language has various modeling capabilities. It is also a good platform to use to develop the right models if you are a beginner.
- **SQL** : SQL provides a set of methods used to perform analysis within the database using different predictive models and mining functions.
- **ACCESS or SAS** : These tools can be used to access data from various storage platforms, like Hadoop, and use that data to create a reusable and repeatable model.

The market has numerous tools you can use to develop modeling techniques, but R is commonly used. At the end of this phase, you will have the required insights in your data that will help you determine the algorithm to use. The next phase is where you apply this algorithm and develop the model.

Phase Four: Build the Model

Now that you have decided which algorithm to use, you need to split the data set into training and testing data sets. In this phase, you need to consider the existing tools and determine if they are sufficient for the purpose of building a model. Make sure you identify a robust environment to run the models. You need to analyze different techniques, such as clustering, classification, and association, to develop the model. You can use different tools to build the model.

Phase Five: Operate the Model

In this phase, you run the data through the model and deliver the reports and necessary technical documents. Additionally, you may also need to run the model in the production environment to test if it works the way it needs. This gives you an idea of how the model performs on real-time data. You can also determine any constraints in the model.

Phase Six: Communicate the Results

It is important to evaluate if the model has given you the results you needed. You can do this by analyzing your hypotheses. This is the last phase of the data science lifecycle and is where you identify the key findings and communicate the same to the organization. You can determine the results of the model based on the criteria you identified in the first phase.

Chapter Two: Pros and Cons of Data Science

Data science is an upcoming field, and there are multiple opportunities. Having said that, this field has its share of pros and cons. This chapter looks at some pros and cons of data science, to help you take the necessary course of action.

Pros

There are many advantages to data science, and this section lists them.

Fastest Growing Field

Data science is an upcoming field and is quite in demand. If you want to become a data scientist, now is the time to do it!

Abundance of Roles

Only some people have the skills necessary to become a data scientist. You need to learn different skills and constantly learn if you want to survive in the field. This makes the field less saturated when compared to other machine learning and big data jobs. You have a lot of opportunities if you want to switch to the field of data science. There is a very low supply of data scientists.

A Versatile Field

Data science can be applied in numerous fields, but is often used in healthcare, consultancy services, e-commerce industries and banking. Data science is versatile, and you have the chance to work in different fields.

Makes Data Easier to Use

Every company needs skilled employees to collect, process, analyze and visualize the data they collect. These employees are data scientists and they not only analyze the data but also improve the quality of that data. A data scientist knows what to do to improve and enrich the data that helps the company make informed decisions.

A Prestigious Career

A data scientist allows a company to make the right decisions. Many companies have hired data scientists to provide the right information they can use to make informed decisions. This gives a data scientist an important position in the organization. Since most companies are looking for data scientists, you can earn a lot of money. According to Glassdoor, you may earn close to $160,000 per annum.

Remove Redundancy

Data science is used in different industries, and most algorithms used in data science help to reduce the number of redundant tasks performed by people. Most companies collect historical data, and they can use this data to train machines to perform redundant tasks, which helps to simplify some jobs performed by people.

Products Become Smarter

Data science is a field that involves the use of machine learning. There are different types of algorithms used in machine learning, known as supervised, unsupervised and reinforcement learning, which look at data sets to determine customer's behavior. For instance, most e-commerce websites use recommendation systems to provide insights to the customers based on their purchase history. This has made it easier for computers to understand how humans behave.

Save Lives

The healthcare sector uses data science to improve diagnoses and predictions made about patients. Through the use of machine learning algorithms, the healthcare sector has found a way to detect tumors and cancer at an early stage. There are many other benefits the healthcare industry gains from using data science.

Aid in Personal Growth

Data science is not only a great career, but it also helps you grow in your career and personally. If you choose to become a data scientist, you will develop the right attitude and thought process to solve problems. Since the field of data science is a mixture of management and IT, you learn from both areas of business.

Cons

Data science is a hot career, and many people choose to become a data scientist because it pays well. That said, there are some disadvantages to the field. If you want to better understand data science, you also need to look at the disadvantages of data science.

Data Science is an Ambiguous Term

There is no definite definition or meaning of data science. It has become a term people use often to talk about analysis, so it is hard to know what data science actually is and what a data scientist can do. The role of a data scientist is dependent on what the company does.

Impossible to Master Data Science

As mentioned earlier, data science is a mix of numerous fields, such as computer science, mathematics and statistics. It is impossible to master the fields used in data science, which means you cannot become an expert in those fields. While most online courses have been doing their best to fill the gap that people in the data science industry are facing, it is impossible to do this. People with a background in statistics may not necessarily have the necessary information in computer science. This field changes frequently, and you need to keep learning different areas of data science if you want to stay abreast.

Requires a Lot of Domain Knowledge

From what you have read in the previous chapter, you know you need a lot of domain knowledge. If you have enough knowledge about computer science, statistics, and mathematics, you may not easily solve a data science problem if you do not have background information. The same can be said for the reverse, as well. Let us assume you work for a health-care company where you need to analyze genomic sequences. To do this, you need to have some information about molecular biology and genetics. This is the only way you can make calculated decisions that help the company. If you do not have this background, it will be difficult for you to work on analyzing genomic syndromes.

Unexpected Results

Data scientists analyze the information in the data set and use the patterns and variables in the data set to make informed decisions. This helps you make informed decisions. There are times when the data provided is arbitrary and you may not obtain expected results. The results may also vary due to the poor utilization of resources and weak management of data.

Lack of Data Privacy

Data is the new oil for numerous industries, and most companies hire data scientists to use the data they collect to make informed decisions. Having said that, the data used in these processes may lead to a privacy breach. The personal data of most clients is stored by parent companies, and some companies do not have enough security to prevent any data leaks. Many countries now have issued regulations and guidelines to prevent such leaks and secure the data.

Chapter Three: Statistics for Data Science

This chapter sheds some light on the statistical concepts used in data science. It is important to remember that data science is not a new concept, and most statisticians can work as data scientists. Data science uses many concepts from statistics, since this is the best tool to help you process and interpret information in the data set. There is a lot of information you can gather from the data set using statistical tools. If you want to master data science and become an expert in the field, you need to learn everything you can about statistics. While there are many topics in statistics a data scientist should know about, the following are the most important aspects to consider:

- Descriptive Statistics
- Inferential Statistics

Descriptive Statistics

Descriptive statistics is the process of expressing or looking at the data to help you read it easily. This method helps you deal with quantitative summarization of the data through numerical representation or graphs. Some topics you need to learn about are listed below:

- Normal Distribution
- Central Tendency
- Variability
- Kurtosis

Normal Distribution

A normal distribution, also known as a Gaussian distribution, is a continuous distribution often used in statistics. Any data set following a normal distribution is spread across a graph, also a bell-shaped curve. In normal distributions, the data points in the set peak at the center of the bell-shaped curve, which represents the center of the data set. When the data moves away from the mean, it will fall to the end of the curve. You need to ensure the data you look at is distributed normally if you want to make inferences from the data set.

Central Tendency

Measures of central tendency help you determine the center values of the data set. There are three measures commonly used – mean, median and mode. The mean or the arithmetic mean of any distribution is found at the center of the data set. You can use the following formula to calculate the mean or average of the data set: (sum of all points in the data set) / (number of data points)

The median is another measure that is the mid value of the data set when the points have been sorted in an ascending order. If you have an odd set of values, you can easily find the mid value, but if you have an even number of data points, you take the average of the two data points in the middle of the data set. The last measure is the mode, and this value is the data point, which occurs the maximum number of times in the data set.

Variability

Variability is a factor which helps you determine the distance between the data points and the average or mean of the data points in the data set. This value also shows the difference between the data points selected. You can view and assess the variability based on the measures of central measure, such as range, variation and standard deviation. The range is a value, which is the difference between the smallest and largest value in the data set.

Kurtosis and Skewness

The skewness of the data set helps you determine how symmetrical the data set is. If the data set is distributed uniformly, it will take the shape of a bell curve. If the curve is shaped evenly, the data is not skewed. If the curve moves either to the right or left side of the data points, it means the data is negatively or positively skewed, respectively. This means the data is either dominant on the left or right side of the measures of central tendency. Kurtosis is a measure that helps you determine the tails of the distribution. When you plot the data points on a graph, you can determine if the data is light or heavy-tailed. You can make this assumption based on the middle section of the distribution.

Inferential Statistics

Descriptive statistics provide information about the data, but inferential statistics are about obtaining insights about the data set. Inferential statistics is about concluding about the large population based on a small data set or sample. Let us assume you need to count the number of people in Africa who have received the polio vaccine. You can perform this analysis in two different ways:

- Ask every individual in Africa if they have been given the polio vaccine

- Take a sample of people from the continent, make sure it is from different parts of the continent, and extrapolate the findings across the entire continent

The first process is difficult to complete, and it is impossible to do it. You cannot go around the entire country and ask people to tell you if they have received the vaccine. The second method is the better way to do this since you can draw insights or conclusions from the sample selected and extrapolate the findings to the large population. The following are some tools used in inferential statistics.

Central Limit Theorem

According to the central limit theorem, "the average of the sample is same as that of the entire population." This shows the properties and measures of the data's central tendency, such as standard deviation, are the same for both the sample and population. This means you can increase the number of data points selected, resulting in a normal curve.

You need to understand the concept of confidence intervals if you want to use the central limit theorem. This goes to show the approximate value of the mean of the population. The process of creating an interval in the population has things like the sum of an error margin. You can calculate this error by "multiplying the standard error of mean with the z-score of the percentage of confidence level."

Hypothesis Testing

Hypothesis testing is the extent to which you can test any assumption you make about the data set. In this process of testing, you can collect the results of your analysis of the hypothesis on a smaller population. The hypothesis you need to test is called the null hypothesis, and we need to check the validity of this hypothesis against the alternative hypothesis. The null hypothesis is the case you need to test. Consider the following example – you are conducting a survey to determine who smokes and who does not, and how people who smoke have cancer.

When you conduct this survey, you begin with the assumption that the number of subjects with cancer who smoke is the same as the number of subjects with cancer do not smoke. This is your null hypothesis, and you need to test it to reject the hypothesis. The alternate hypothesis will be that the number of subjects with cancer who smoke is greater than the number of subjects with cancer who do not smoke.

Based on the given data and evidence, you can test the hypotheses and analyze the data to determine if the null hypothesis is correct or not.

ANOVA

ANOVA is another statistical concept that is used to test hypotheses across different groups of data. This concept helps you determine if the groups you are checking have similar averages and variances. ANOVA allows you to carry out this form of analysis with limited error rates. You can use the F-ratio to calculate ANOVA. The F-ratio is used to calculate the ratio of the mean square error between groups to the mean square error in specific groups. The following are the methods to calculate ANOVA:

1. Write the hypotheses and understand the need for those. Every analysis should have a null and alternative hypothesis
2. In case of null hypothesis, you need to assume the average of the groups is identical
3. The alternative hypothesis will have a different average

Chapter Four: Skills Required for Data Science

If you want to work in the field of data science, or become a data scientist, you need to have some information about different topics mentioned in this chapter. This chapter lists some technical and non-technical skills you should develop if you want to improve in this field.

Technical Skills

As a data scientist, you need to have some technical skills that help you with statistical analysis. You need to learn how to leverage and work with different frameworks and software to mine, collect, process, collate, analyze, interpret and visualize large volumes of data. You need to develop programming skills to perform such activities. An easy way to do this is to ensure you have the necessary academic background. Most data scientists have a Ph.D. or master's degree in engineering, statistics, and computer science. This is the only way to determine if they have the foundation to help them connect with different technical points that are the foundation of the practice. Many schools offer such programs to help people pursue data science.

If you do not want to go through hardcore courses, you can look at options, such as:

- Boot camps
- Massive Open Online courses (MOOC)

These programs help you develop a basic understanding of core data science subjects. These courses also provide some information that is outside of textbook learning. You will be given real-time scenarios and asked to develop models to assess and predict futuristic events. The following are some skills you need to develop.

Understanding Data

Data science is about working with and understanding different types of data. You need to understand and love working with data. The following are some questions you can answer to help you understand whether you love data or not:

- Do you know where you need to collect data from?
- Do you understand and know what data is?
- Have you determined what information you should look at?
- How frequently do you work with data?

- Have you learned how to work with structured and unstructured data?

The most important question you need to answer is whether you love working with data or not. If yes, you need to obtain certifications, so you develop the skills to become a data scientist.

Algorithms

Algorithms are sets of instructions that you write. You can use algorithms to instruct a machine to perform specific functions and tasks. Let us try to write an algorithm using that we will instruct a computer to add two numbers.

1. Identify two variables and declare it to the machine
2. Initialize the variables
3. Ask the user to assign a value to each of these variables
4. Declare another variable to hold the sum of the first two variables
5. Calculate the sum of the first two variables and assign that value to the third variable

You can also use algorithms when you solve puzzles on paper. As a data scientist, you need to understand what algorithms are and how a machine understands them since you work with algorithms to help you analyze data. As a data scientist, you also need to learn how to design different algorithms that perform the necessary functions to help you analyze data. Let us assume you need to key in 10 numbers into the system. You may enter any ten numbers and leave it to the machine to identify the set's third largest number. To do this, you should write an algorithm to help the machine identify the number. As a data scientist, you need to write the necessary logic and develop an algorithm which helps you find the third largest number.

Programming

It is important to learn different languages, such as Java, C++, R, Python, Perl, SQL and other languages. The languages used commonly are R and Python. You can collect, clean, process, organize and analyze the information in the data set which helps you work with unstructured data.

Examples

Example 1

This takes care of only variables and constants – Program to build Hello World

```
> # We can use the print() function
> print("Hello World!")
[1] "Hello World!"
> # Quotes can be suppressed in the output
> print("Hello World!", quote = FALSE)
[1] Hello World!
> # If there is more than 1 item, we can concatenate using paste()
> print(paste("How","are","you?"))
[1] "How are you?"
```

In the program above, we have used print(), a built-in function to print the required string Hello World! The quotes that you see are printed by default. To avoid that, we can add an argument called quote = FALSE. Also, if there is more than a single item, you can use paste() or cat() functions to concatenate the strings together.

Example 2

We can add the elements of the vector by the function sum()

```
> sum(2,7,5)
[1] 14
> x
[1]  2 NA  3  1  4
> sum(x)    # if any element is NA or NaN, result is NA or NaN
[1] NA
> sum(x, na.rm=TRUE)    # this way we can ignore NA and NaN values
[1] 10
> mean(x, na.rm=TRUE)
```

```
[1] 2.5

> prod(x, na.rm=TRUE)

[1] 24
```

When a vector has NA (not applicable), or NaN (not a number), the functions that are used here such as sum(), mean(), prod(), etc make NA or NaN, respectively.

Example 3

This example will deal with an interactive screen, i.e. take inputs from the user.

```
my.name <- readline (prompt="Enter name: ")

my.age <- readline (prompt="Enter age: ")

# convert character into integer

my.age <- as.integer(my.age)

print (paste ("Hi,", my.name, "next year you will be", my.age+1, "years old."))
```

Output:

```
Enter name: Mary

Enter age: 17

[1] "Hi, Mary next year you will be 18 years old."
```

As you can see, we have used the function readline() to get input from the user.

Here, you can see that you can display an appropriate message for the user with the prompt argument.

In the above example, you convert the input age, which is a character vector into integer by the function as.integer().

This is necessary for doing further calculations

Example 4

In this example, we will find if a year is leap year or not by taking inputs from the user.

```
# Program to check if the input year is a leap year or not
year = as.integer(readline(prompt="Enter a year: "))
if((year %% 4) == 0) {
if((year %% 100) == 0) {
if((year %% 400) == 0) {
print (paste (year,"is a leap year"))
} else {
Print (paste (year,"is not a leap year"))
}
} else {
Print (paste (year,"is a leap year"))
}
} else {
Print (paste (year,"is not a leap year"))
}
```

Output 1:

```
Enter a year: 1900
[1] "1900 is not a leap year"
```

Output 2:

```
Enter a year: 2000
[1] "2000 is a leap year"
```

Here we have used the logic that a leap year is exactly divisible by 4 except for the years ending with 00. The century year is a leap year only if it is perfectly divisible by 400.

Nested if else is used to implement the logic in the above program.

Example 5

In this example, we will find the HCF of two numbers

```
# Program to find the H.C.F of two input number
# define a function
hcf <- function (x, y) {
# choose the smaller number
if(x > y) {
smaller = y
} else {
smaller = x
}
for(i in 1:smaller) {
if((x %% i == 0) && (y %% i == 0)) {
hcf = i
}
}
return(hcf)
}
# take input from the user
num1 = as.integer (readline(prompt = "Enter first number: "))
num2 = as.integer (readline(prompt = "Enter second number: "))
```

```
print (paste ("The H.C.F. of", num1,"and", num2,"is", hcf(num1, num2)))
```

Output:

```
Enter first number: 72

Enter second number: 120

[1] "The H.C.F. of 72 and 120 is 24"
```

This program asks the user to input two integers and the pass them to a function which give the output as the H.C.F.

The function first determines the smaller of the two numbers given as an input since the H.C.F can only be less than or equal to the smallest number among the two.

We then use a 'for' loop to go from 1 to that smaller number.

In each loop we check if our number perfectly divides both the input numbers.

If yes, then we store the number as H.C.F. At the completion of the loop we will end up with the largest number that perfectly divides both the numbers.

Example 6

In this example, we will show you how to develop a calculator of your own

```
# Program makes a simple calculator that can add, subtract, multiply and divide using functions

add <- function(x, y) {

return(x + y)

}

subtract <- function(x, y) {

return(x - y)

}

multiply <- function(x, y) {
```

```
return(x * y)
}
divide <- function(x, y) {
return(x / y)
}
# take input from the user
print ("Select operation.")
print ("1.Add")
print ("2.Subtract")
print ("3.Multiply")
print ("4.Divide")
choice = as.integer (readline (prompt="Enter choice[1/2/3/4]: "))
num1 = as.integer (readline (prompt="Enter first number: "))
num2 = as.integer (readline (prompt="Enter second number: "))
operator <- switch(choice,"+","-","*","/")
result <- switch (choice, add(num1, num2), subtract(num1, num2), multiply(num1, num2), divide(num1, num2))
print (paste (num1, operator, num2, "=", result))
```

Output:

```
[1] "Select operation."
[1] "1.Add"
[1] "2.Subtract"
[1] "3.Multiply"
[1] "4.Divide"
```

```
Enter choice[1/2/3/4]: 4

Enter first number: 20

Enter second number: 4

[1] "20 / 4 = 5"
```

In this code, we will first ask the user what operation he wants to carry out. Then, the user is asked to type in two numbers, and we use 'switch' branching is used to carry out a particular function.

The functions such as add(), subtract(), multiply() and divide() areal user-defined functions.

Analytical Tools

Every data scientist needs to learn how to use different analytical tools and understand the use of Software as a Service (SaaS). This understanding will help you obtain some information from the data you have cleaned and processed using different programming languages and tools, such as Hadoop, R, SAS, Spark, Hive and Pig. Different platforms allow you to improve your skills, and you can obtain certifications that will help you move ahead in your career.

Using Unstructured Data

Data scientists need to collect, process, store, manage, clean, understand and analyze the data collected from various sources. Most of these data are not structured, and they can be collected in different forms, such as images, videos, text, emails and other forms. For instance, if you work with a credit company, you need to learn how to use historical data and patterns to assist the company in identifying the customers they can trust. You also need to learn how to collect, process and analyze the data.

Non-Technical Skills

The following are some non-technical skills you need to develop if you want to be a data scientist. These are some personal skills a user needs to develop, and these are very different from qualifications and certifications.

Strong Business Acumen

As a data scientist, you need to have strong business acumen. You need to understand various elements important to the business model. Otherwise, you cannot use the necessary tools and skills to analyze the data set. It is only when you learn and develop these skills that you can identify the various problems and solutions to develop if you want to help the business make informed decisions so that it can sustain and grow. It becomes extremely hard for you to help an organization explore various opportunities if you do not know how to code.

Communication Skills

Data scientists can understand data better than anybody else in the organization. If you want to succeed as a data scientist, you need to learn to communicate your decisions and analysis to the stakeholders. This is the only way you can make it easier for the organization to make informed decisions. It is important to learn to communicate your findings to people in the organization, especially if they are non-technical. You need to develop the necessary communication skills to communicate your findings.

Intuition

This is an important non-technical skill you need to develop as a data scientist. It is true you may already have an insight about the data when you look at it, but you may not have the ability to perceive any hidden patterns in the data set. You need to know where to look and what to look at, so you add more value to the analysis.

These are the skills that make you efficient in your work. Do not worry about cramming too much information in a short time. You learn and improve as you gain experience.

Chapter Five: Tools Required for Data Science

We have looked at different aspects of data science and know why you need to use data science. Let us now look at the different tools you need to work with to become a data scientist. The following are only some tools you need to use if you want to work in data science.

R Programming Language

We have referred to R earlier in the book. This is a programming language widely used by data scientists, and many organizations use this language to perform data analysis. Organizations use R widely since it is an object-oriented programming language, and it is used for the analysis of reporting, visual graphic representations and statistics.

Python Programming Language

Python, like R, is another programming language that has features similar to other object-oriented programming languages. It is used widely in data science and to develop models. Python can be used on different operating systems, and you can develop numerous applications using this language. This programming language is used to develop machine learning models and applications that use artificial intelligence.

Hadoop

As mentioned earlier, Hadoop came into picture when it became difficult for companies to store large volumes of data. This is an open-source application which allows you to store and process big data. You can use a distributed computing framework to analyze the data. It is best to use this application if you want to store data without having to process it. You can store as much data as needed before you decide which variable you need to use.

Structured Query Language (SQL)

SQL or Structured Query Language is used for data storage. You can create new tables and databases in SQL, and the only thing you need to do is update the records in the database. You can retrieve, modify and delete the records in SQL databases and tables. You can use this language to manage the data stored in a database.

SAS

SAS or Software as Service is an important tool to understand if you wish to work as a data scientist. This tool is designed to perform various statistical and mathematical operations. Large organizations use SAS to identify or draw hidden patterns within the data set. SAS uses SAS coding language that is often used by statisticians to develop models.

The language was developed by statisticians and professionals and is based on Reliable Commercial Software (RCS). This programming language has many statistical tools and libraries that data scientists can use to perform modeling. They can also use this language to keep the records in order. SAS is not an open-source programming language. This language is costly and can only be used by large organizations. There are some packages and libraries in SAS that cannot be accessed by organizations in the base pack, and you need to spend money to update the language.

Apache Spark

Apache Spark is a popular analytics tool used by data scientists, and it is designed for stream processing and batch processing. There are numerous APIs developed by data scientists across the globe to develop a frequent path for data to flow through for any machine learning algorithm. This application is a win when compared to Hadoop and other data processing and storing applications since it works faster. It comes with numerous APIs for machine learning and deep learning which enables data scientists to obtain hidden insights in the data set. The application can stream large volumes of data and is more efficient when compared to other big data platforms. This tool can work with real-time data, and allows you to code in Java, R and Python. You can use this application when you work with clusters of data. This tool is more efficient when compared to platforms, such as Hadoop and MapReduce.

D3.js

This is a Javascript library, and you can use this to develop visualization on the Internet. There are numerous APIs of this library, and you can use various functions to make visualizations more dynamic. You can also analyze the data set collected on the Internet. One of the advantages of using this tool is it allows you to use animated transitions to improve data visualizations. When you use this tool, you can create dynamic documents and use options to change the data used in the visualization.

You can also use this tool along with CSS to create transitory and eminent visualizations that would show you the required graphs on the Internet. This is an important tool to use in data science, and you can work with different IoT devices. The tool allows you to work with client-side interventions for data visualization and processing.

Chapter Six: Application of Data Science

Now that we know the importance of data science and the prerequisites and skills needed for data science, it is important to know how data science can be applied in the real world.

Risk and Fraud Detection

Data science was first used in the financial and banking sector. Many financial institutions were sick of bad debts and their losses at the end of each year. Since they had collected large volumes of data when the initial paperwork is carried out, they had a wealth of information they could use to decide who they could give the credit to. They hired a data scientist to prevent future losses caused by bad debts. Financial institutions now have learned to collect, process, divide the data and assess it. They divide the data based on customer profiling. It also helped in the sales of banking products as they could do targeted marketing based on how much a customer earned every year.

Healthcare

Many healthcare companies use medical image analysis, content-based medical image indexing, wavelet analysis and machine learning models to detect tumors, organ failure and other procedures and frameworks. These processes make it easier to analyze and find the optimal parameters or points to improve diagnoses.

Genetics and Genomics

You can use data science to personalize treatment for patients, improving the treatment administered to patients for various illnesses based on research in genomics and genetics. This process aims to determine what the consequences of DNA are on our health and determine if there are any connections between genes and diseases.

Drug Development

It takes 12 years to develop a drug and a few months before it is tested and becomes an official drug. The process of identifying the drug and assessing if it will work on people is complicated and involves the use of many disciplines. Billions of tests often restrict the greatest ideas. You can use machine learning and data science to simplify the process and reduce the time taken to initially screen the drug compounds used to develop the drug. Algorithms and data science can also predict how the body will react to certain compounds in the drug using various statistical and mathematical simulations and models. This is faster to when compared to traditional lab experiments. The models can predict future outcomes with better accuracy, thereby simplifying the process.

Internet Search

When you think of data science, the first thing that comes to your mind is the search algorithm used on the Internet. Every search engine uses various data signs to provide the user with the best results of any query you key into the search tab. Google would not be the search engine it is today if the developers did not use data science algorithms to improve search results.

Website Recommendations

Most websites, such as Netflix, Amazon, IMDB and Twitter, use historical data to send suggestions to the users or customers to improve their experience. For example, if you want to make a purchase of earrings or clothes online, you may look at one website. Google will use that historical search to send you more suggestions. Instagram uses a similar profile as well. The recommendations made by these platforms are based on previous searches you may have performed.

Every marketing company makes use of data science algorithms to send suggestions to customers. Every advertisement you see at the airport or theatre on a digital billboard or screen on the website uses data science algorithms. Digital advertisements have been doing as well as they have since they can get a higher click-through rate when compared to traditional advertisements. You can use data science algorithms to use historical data based on the past behavior.

Advanced Image Recognition

You may have uploaded a picture on Instagram, Facebook or other social media platforms. These platforms may send you a suggestion to tag your friends in the picture based on your previous pictures. This feature uses the face recognition algorithm to suggest whom you can tag. Google also allows you to search for some images by uploading specific images. It provides related search results by using image recognition.

Virtual Assistance

Many organizations have developed a mobile application that helps patients visit doctors virtually to determine what is wrong with them. These applications are powered with artificial intelligence and have a chatbot that allows them to speak to healthcare professionals to understand their symptoms. There are many applications you can use to help remind you to take medication on time. Another example is the use of chatbots on different websites. You can speak to the chatbot and ask them any question you may have.

Speech Recognition

Siri, Cortana, Google voice and other voice assistant products use speech recognition to improve your experience with different gadgets. Your life may stop if you are not in a position to type. These tools convert the words you say to text, but it is important to note that this method of conversion does not perform accurately.

Planning Routes for Airplanes

Airline companies have begun to use data science to identify different routes to optimize the airline routes. To reduce the heavy losses, airline companies bear due to an increase in fuel prices and heavy discounts to customers due to high competition, companies can use data science for the following:

- Predict delays in flights
- Decide the type of aircraft to use or buy
- Determine which flight you should suggest to a customer without any halt or stop
- Developing customer loyalty programs

Gaming

Most developers have started using machine-learning algorithms in games. These algorithms learn from how the user plays the game and uses that information to help the player upgrade or move to higher levels. Even in motion gaming, the computer uses previous or historical data and analyzes it to provide the best experience. EA Sports, Sony, Zynga and Nintendo have taken the experience of gaming to a whole new level by using various concepts of data science in their code.

Augmented Reality

This is an interesting and exciting application of data science, and it is something many companies are trying to incorporate in their products. If you do not know what I am talking about, think about virtual reality headsets and glasses. These products contain computing algorithms, logic and knowledge to improve your viewing experience.

Data science and virtual reality have always had a strong relationship. Virtual reality products allow you to experience things you could not have in the past. For example, in the show' Bones,' Angela Montenegro, an employee of the Jeffersonian institution, sets up a virtual reality monitor that allows a child to take a trip in the Louvre. There are similar activities in Disney Land, United States, where you can take a parasailing trip without actually being in air.

You may relate to the use of data science in virtual reality when it is easier to price the value.

Chapter Seven: Algorithms and Statistics

Do you need to become a Ph.D. in mathematics or statistics to become a data scientist? No, you do not. The extent to which you need to know depends on what role you want to take up in the organization. This chapter gives you an idea of all the concepts you need to know if you want to become a data scientist. As a data scientist, you first need to understand the basic concepts of probability theory and statistics. If you want to master different statistics and mathematics algorithms, you need to practice them regularly. You need to follow a top-down approach if you want to learn how to code. You need to learn how to use a data stack, familiarize yourself with real-world projects and use libraries and documentation. When you notice your lack of theoretical background, you can see the bigger picture. It is important to learn how these algorithms work. Start off with the mathematical aspects before you move onto those areas, which are more complex. This chapter has the list of statistical and mathematical concepts you need to learn.

Naïve Bayes

A Naïve Bayes classifier is an algorithm which uses the principles of probability. One of the most important principles it uses is that the value of one feature in the data set is independent from the value of other features in the data set. According to this theorem, you can predict the probability that an event will occur based on the conditions of the occurrence of that event. It is important to learn how these classifiers work, and to do this, you need to study the basics and principles of conditional probability and probability. One of the easiest ways to understand this theorem is to study the concepts of probability. You can then check on Bayes' theorem and how to code this theorem.

Linear Regression

Linear regression is a basic form of regression, and it allows you to determine the connection between two variables in the data set. One of the variables is termed as the predictor variable and the other is known as the response variable. These variables are often continuous in nature. A simple linear regression model uses various data point sets and then extrapolates the trend of the values and how they will progress in the future. Linear regression is a common algorithm used in parametric machine learning, and in this form of learning, you can train the machine to develop a mathematical function using existing data sets. You can use this function to make predictions or forecasts about future events. These functions are termed as models in machine learning terminology. The best way to learn more about regression is to develop an understanding of elementary statistics. If you want to learn more about regression and the various concepts, you need to take an advanced course of statistics.

Logistic Regression

Logistic regression is a process sued in different cases where you have a binary dependent variable. This form of regression focuses on the estimation of probability of the occurrence of an event. Logistic regression is also an algorithm used in parametric machine learning. Like linear regression, this algorithm also results in a mathematical function. The difference between these forms of regression is that logistic regression leads to developing a mathematical function that estimates the values based on their logarithms or other statistical functions. Another difference between linear and logistic regression is that the former uses real numbers and gives those as the output. The latter uses a model which provides the logistic function. A logistics function produced is also termed as a sigmoid function that takes care of the values to give you an output between 0 and 1. Let us understand why a sigmoid function returns a probability value. It is because of the algebraic concepts used in the algorithms that a negative exponent is taken.

Neural Networks

A neural network is a form of machine learning algorithm. These forms of networks are based on the structure of the neurons in the brain. The neural network uses a series of activation points or weights to help you make the necessary predictions. The neurons take the input, apply the transformation function and give you an output that helps you make the necessary decisions. When it comes to capturing any nonlinear relationships within the data set, it is best to use a neural network best since it helps in tasks, such as image and audio processing. The fundamental concept of neural network is to transform the input and processing it to generate the output. If you want to understand the math better in a neural network, you need to take up courses to help you learn more about linear algebra and geometry. This is a better way to start preparing. If you want to delve deeper, you need to learn more about graph theory, matrix theory and real analysis.

K-Means Clustering

The k-means clustering algorithm is a type of unsupervised machine learning algorithm. You can use this algorithm to categorize any unlabeled data. This type of data does not have any defined categories or groups, but it works to identify or define any hidden categories in the data set. The number of variables identified by the algorithm are 'k' in number. The algorithm iterates over the various data points in the data set, so every point is assigned to one of many groups. This algorithm depends on the idea that the distance between the data points and the center of the cluster is the same across all data points. You can use any function that shows the distance between the elements in the data set. Suppose you want to work with this algorithm. In that case, you need to understand the basics of algebra, including addition and subtraction to identify the distance between the data points easily. If you want to delve deeper, you need to learn about Euclidean and non-Euclidean geometry.

Decision Tree

Decision trees are an easy way to determine the outcome of any decision you take using a flowchart. The flowchart is in the form of a tree structure that uses a branching method. Every node in the decision tree is based on a specific variable and every branch is an image of the outcome. A decision tree is based on information theory, since this is how they are constructed. An important concept of a decision tree is known as entropy. Entropy is a measure of uncertainty in any variable. When you use entropy, you can construct a decision tree easily. When there is higher entropy, there is more uncertainty in the data. this means you need to split the tree in a way to decrease any uncertainty.

Information gain helps you determine how much information you can gain from someone. When it comes to a decision tree, you can calculate the information gain on every column based on previous columns based on the information present in the columns in the data set.

Here is a small tip – you need to learn basic algebra and probability to know what decision trees are and how they work. If you want to learn more about decision trees, you need to learn about logarithms and probability.

Some Final Thoughts

Mathematics and statistics are difficult subjects, and they can feel daunting and dry at times. You will, however, feel equipped when you compare your skills against the skills of your peers. It is also important to learn how you can apply these concepts on various data science problems.

The topics highlighted in this book will need to be read well and understood. This is the only way you can develop and build algorithms. You can rely on different machine learning libraries and tools to complete this work for you. It is useful for a data scientist to understand statistics and math, which helps you determine what is happening in the tools or algorithms. This allows you to choose the algorithm that works best for your data set and allow you to make the necessary predictions.

Dive deep and work hard to develop the skills necessary to become a data scientist.

Applying Math to Data Science Models

If you want to work as a data scientist, you need to have some mathematical skills. This is the only way you can understand data and the significance of the patterns in the data set. These are important data science skills since you use them to build and develop models and perform predictive analysis and hypothesis testing. Data scientists use math to develop decision models and they can use different mathematical algorithms to predict the future. This book explains some concepts you need to learn and the skills you need to develop if you want to become a data scientist.

Deriving Insights from Statistical Methods

If you are a data scientist, you need to understand the basics of statistics. You should understand the significance of data, develop hypotheses and validate them by simulating certain scenarios. This makes it easier for you to predict futuristic events. It is very difficult to develop advanced statistical skills, but if you are keen about pursuing a career in data science, you need to understand some basics of logistic regression, linear regression, time series analysis, Bayes classification, etc.

Some Essentials for Data Science

Understanding programming languages and coding in various languages, especially R and Python, are important for data science. You need to learn how to write code to instruct the machine on how it should process, manipulate, analyze and visualize the data. It is important to understand how to write code in R and Python. There are different libraries and functions in these languages that allow you to manipulate, analyze and visualize the analysis of the data set.

You can also use SQL to run queries on the data set that allows you to extract and modify data in the database. People also use JavaScript library to develop interactive, dynamic and custom visualizations on the Internet. To do this, you need to learn to write code in R or python. It is important to learn to code if you want to become a data scientist, and these are easy languages to learn.

Chapter Eight: Data Science in a Sector

Most statisticians have been stating that data science is exactly what they have been doing for the last few decades. While this may be true to an extent, you cannot agree with them. Data science is completely different and separate from traditional statistical and mathematical approaches. Data science is different from statistics in many ways. Data scientists use programming languages to improve the efficiency of their analysis. If you are a statistician, you need to know the subject in its entirety. As a data scientist, you need only to understand the basics of mathematics and statistics.

Statisticians have very limited expertise about anything apart from statistics. They need to reach out to subject matter experts if they want any help or information about other subjects. They also need some help from experts to help them determine and understand their findings. This is the only way they can learn how they can move forward.

On the other hand, a data scientist has enough information about different subjects that helps him understand and analyze the findings with ease. A data scientist generates deep insights and uses his subject matter expertise to understand what the data is trying to tell him.

The following are some ways to use data science to improve the performance of the business in specific sectors.

- Some architectural firms have begun to use machine learning algorithms to optimize the consumption of energy in modern designs
- Marketing firms and businesses use logistic regression and other statistical algorithms to predict how customers may behave in the future
- The healthcare industry uses data science to determine illnesses and personalize treatment plants. This allows doctors to predict and check for any future health problems
- A data journalist takes information from the website to identify any patterns and connections within the data set
- Law enforcement agencies and organizations use data science to prevent and predict any criminal activities

Communicate the Insights Captured Through Data

If you want to become a good data scientist, you need to have sharp written and oral communication skills. As a data scientist, you need to communicate the information you have inferred from the data. If you cannot do this, all the information in the data set means nothing to you or the business. A data scientist must find a way to explain the insights gathered from the data set so that he can develop meaningful and clear visualizations. It is important to remember that people are visual, so they learn better when they can visualize anything. As a data scientist, you need to be pragmatic and creative about the data collected and find the right way to communicate them clearly.

Leverage Cloud-Based Solutions

If you are not used to working with large volumes of data, you can use cloud-based solutions to learn more about the patterns in the data set. Most organizations and data analysts have understood the use of data analytics and big data and their importance. It is a new domain for some people and organizations, but they are working as hard as possible to prepare for the flood of big data. many cloud applications, both private and new, are dedicated to improving the way companies handle their big data.

You can also use cloud services, such as Tableau, to clean and process data sets. Most cloud-based applications are open-source and code-free. You can use these solutions to model and visualize data using various statistical tools and techniques. You need basic knowledge of mathematics and statistics before you apply any data science techniques on the data set. If you do not have knowledge about programming, you can use applications that do not require you to write any scripts or code.

Regardless of whether you use cloud-based solutions to perform data analysis or not, you need to train your employees and gather more information about what data science is. You also need to learn about the algorithms discussed earlier. It is important for you to learn more about how to process data, design a model, build the model and analyze it if you want to obtain insights about the data. You cannot expect any cloud-based platform to make it easier for anybody to work with the data set if they do not have basic knowledge of mathematics and statistics.

Chapter Nine: Data Cleaning Using R

Data is the new fuel for all organizations, and they collect data from every source possible. This means they work with unstructured, semi-structured and structured data. It is for this reason we need to clean and process the data, so you transform raw data into structured data that is easier to analyze. Cleaning and processing of data influences the analysis you perform. Most data scientists use R to perform an analysis on data, and this language offers a comprehensive and effective method to analyze data. Let us look at the process of cleaning the data.

Step One: Exploratory Analysis of Data

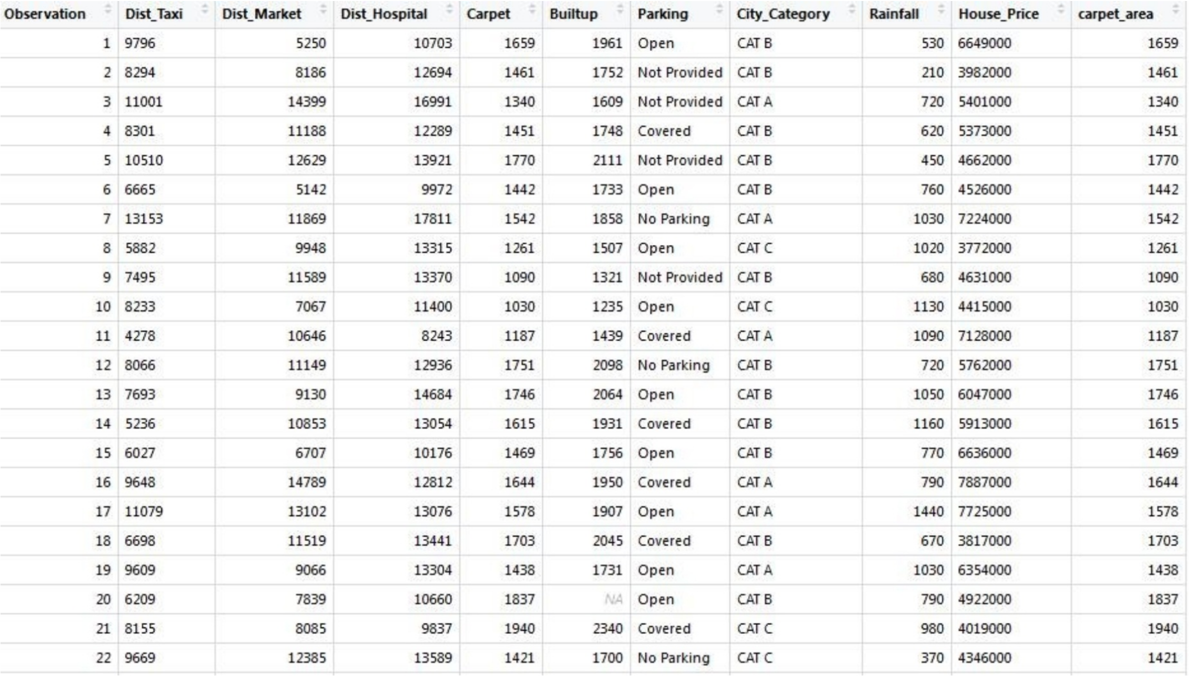
When it comes to processing information in the data set, you need to explore the data. You can import the data into R and use different operations to explore the data set. Before we perform the analysis, it is important for you to understand how to import the data into R.

```
setwd("C:/Users/USER/Desktop/House Pricing ")

dir()

data<-read.csv(“Regression-Analysis-House Pricing.csv“,na.strings = “”)

View(data)
```



Observation	Dist_Taxi	Dist_Market	Dist_Hospital	Carpet	Builtup	Parking	City_Category	Rainfall	House_Price	carpet_area
1	9796	5250	10703	1659	1961	Open	CAT B	530	6649000	1659
2	8294	8186	12694	1461	1752	Not Provided	CAT B	210	3982000	1461
3	11001	14399	16991	1340	1609	Not Provided	CAT A	720	5401000	1340
4	8301	11188	12289	1451	1748	Covered	CAT B	620	5373000	1451
5	10510	12629	13921	1770	2111	Not Provided	CAT B	450	4662000	1770
6	6665	5142	9972	1442	1733	Open	CAT B	760	4526000	1442
7	13153	11869	17811	1542	1858	No Parking	CAT A	1030	7224000	1542
8	5882	9948	13315	1261	1507	Open	CAT C	1020	3772000	1261
9	7495	11589	13370	1090	1321	Not Provided	CAT B	680	4631000	1090
10	8233	7067	11400	1030	1235	Open	CAT C	1130	4415000	1030
11	4278	10646	8243	1187	1439	Covered	CAT A	1090	7128000	1187
12	8066	11149	12936	1751	2098	No Parking	CAT B	720	5762000	1751
13	7693	9130	14684	1746	2064	Open	CAT B	1050	6047000	1746
14	5236	10853	13054	1615	1931	Covered	CAT B	1160	5913000	1615
15	6027	6707	10176	1469	1756	Open	CAT B	770	6636000	1469
16	9648	14789	12812	1644	1950	Covered	CAT A	790	7887000	1644
17	11079	13102	13076	1578	1907	Open	CAT A	1440	7725000	1578
18	6698	11519	13441	1703	2045	Covered	CAT B	670	3817000	1703
19	9609	9066	13304	1438	1731	Open	CAT A	1030	6354000	1438
20	6209	7839	10660	1837	NA	Open	CAT B	790	4922000	1837
21	8155	8085	9837	1940	2340	Covered	CAT C	980	4019000	1940
22	9669	12385	13589	1421	1700	No Parking	CAT C	370	4346000	1421

When you enter the data into R, you need to check the class of the frame in your data.

class(data)

When you do this, you will obtain the output shown below. In this output, you can see that the dataset is stored in R in the form of a data frame.

[1] “data.frame”

We need to perform the next step to check the number of columns or rows present in the data frame. The following is the code to use.

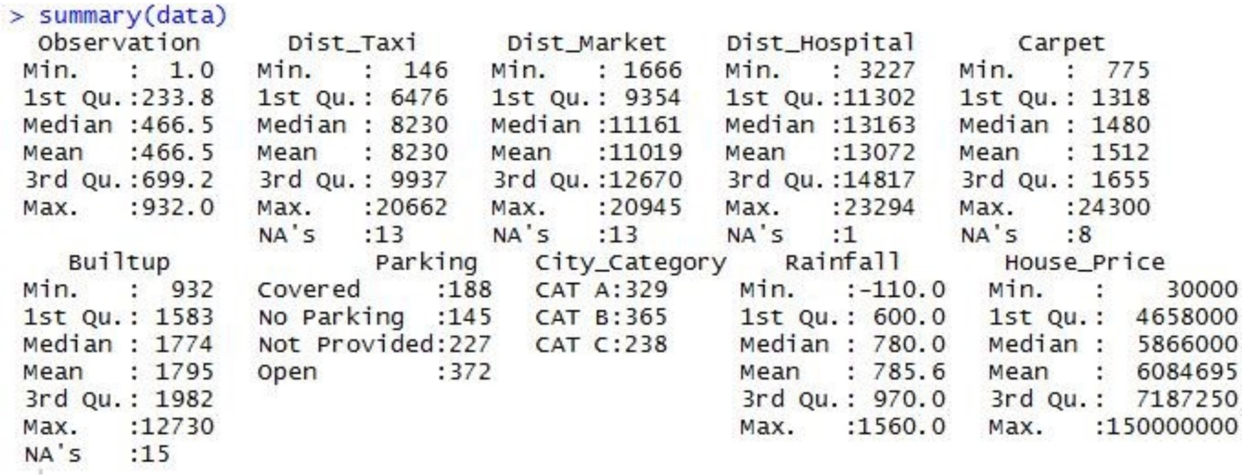
dim(data)

The output is: [1] 932 10

From the above output, we see that the data set has 10 columns and 923 rows. You can use the following function to view the columns and rows' details in the data set.

summary(data)

This leaves you with the following output:



```
> summary(data)
  Observation      Dist_Taxi      Dist_Market     Dist_Hospital      Carpet     
 Min.   :  1.0   Min.   :  146   Min.   : 1666   Min.   : 3227   Min.   :  775  
 1st Qu.:233.8   1st Qu.: 6476   1st Qu.: 9354   1st Qu.:11302   1st Qu.: 1318  
 Median :466.5   Median : 8230   Median :11161   Median :13163   Median : 1480  
 Mean   :466.5   Mean   : 8230   Mean   :11019   Mean   :13072   Mean   : 1512  
 3rd Qu.:699.2   3rd Qu.: 9937   3rd Qu.:12670   3rd Qu.:14817   3rd Qu.: 1655  
 Max.   :932.0   Max.   :20662   Max.   :20945   Max.   :23294   Max.   :24300  
                 NA's   :13      NA's   :13      NA's   :1       NA's   :8      
    Builtup                Parking    City_Category    Rainfall         House_Price       
 Min.   :  932   Covered      :188   CAT A:329     Min.   :-110.0   Min.   :    30000  
 1st Qu.: 1583   No Parking   :145   CAT B:365     1st Qu.: 600.0   1st Qu.:  4658000  
 Median : 1774   Not Provided:227   CAT C:238     Median : 780.0   Median :  5866000  
 Mean   : 1795   Open         :372                 Mean   : 785.6   Mean   :  6084695  
 3rd Qu.: 1982                                     3rd Qu.: 970.0   3rd Qu.:  7187250  
 Max.   :12730                                     Max.   :1560.0   Max.   :150000000  
 NA's   :15                                                                             
```

Step Two: Visual Exploratory Analysis

After you import the data into R, you can use visual analysis to learn more about the hidden parameters. You can use the following plots to look at the data visually:

- Histogram
- Box Plot

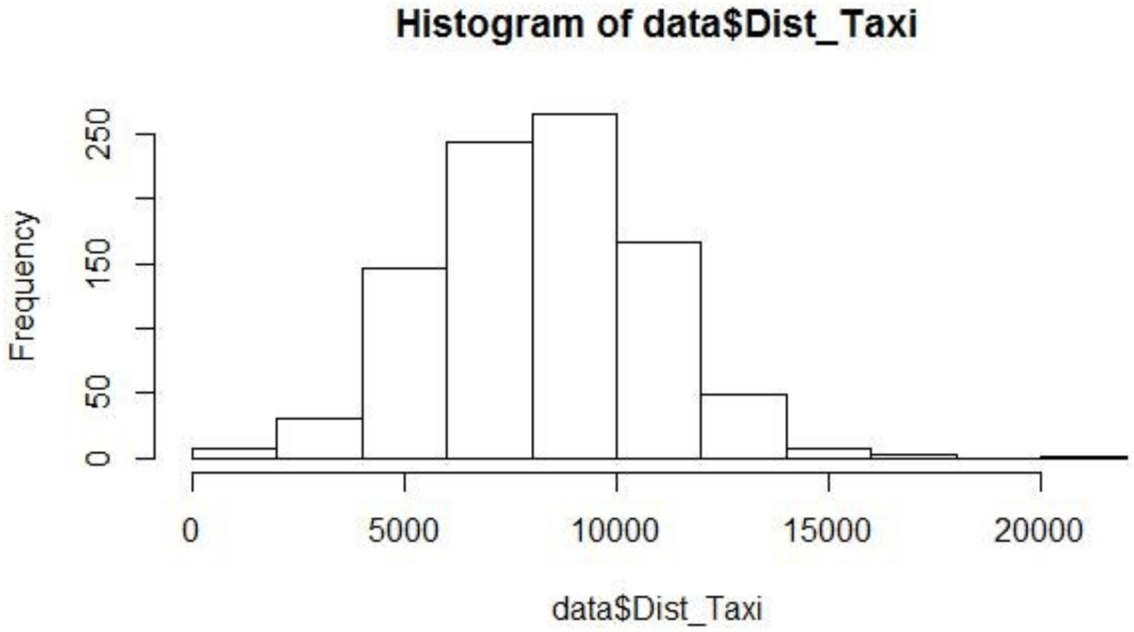
Histogram

A histogram is a visualization tool used to look at the data set on the basis on its numerals. Using a histogram, it becomes easy to determine if the data is distributed using a unimodal and bi-modal distribution. Most data sets are normally distributed, so if your data is not distributed normally, you can convert the data into a normal distribution. A histogram can also help you determine if there are any outliers in the data set or the column under study. Use the code given below to plot the data as a histogram.

```
install.packages(“plyr”)
```

```
library(plyr)
```

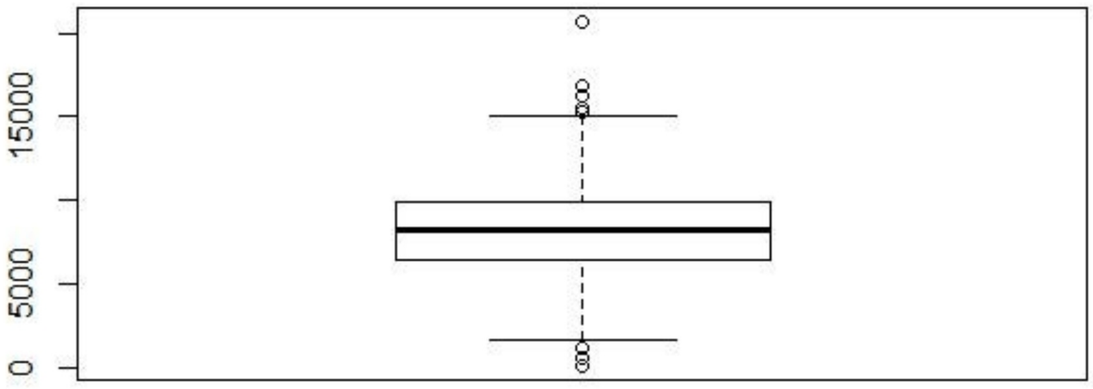
```
hist(data$Dist_Taxi)
```



Box Plot

A box plot is extremely useful since it shows you how skewed the data is. A box plot shows you the median and quartiles in the data set. You can also use this to spot any outliers in the data set. Use the code below if you want to visualize the data in the form of a box plot:

```
boxplot(data$Dist_Taxi)
```



Step Three: Correct Errors

In this step, you should identify the errors in the data set, and determine methods and models you can use to correct the errors in the code. You can also change names of the data frames to differentiate between the accurate frames used. The syntax to change the name of a data frame is:

```
data$carpet_area<-data$Carpet
```

Using the above syntax, you can rename the column 'Carpet' to carpet_area. There may be times when the data set imported in the language are stored as numeric values. In these cases, you need to change the column type using the following code below:

```
data$Dist_Taxi<-as.character(data$Dist_Taxi)

class(data$Dist_Taxi)

"character"
```

You can perform numerous array conversion operations in R. Some of these operations are listed below:

```
as.character()

as.numeric()

as.integer()

as.logical()

as.factor()
```

Since most data collected by organizations is unstructured or semi-structured, you need to find ways to transform data as needed. String manipulation comes in handy since most elements are text based. You can change the text from lowercase to uppercase and vice versa in a specific column. To do this, you can perform the operation below:

```
#Making all uppercase

data$Parking<-toupper(data$Parking)

#Making all Lowercase

data$Parking<-toupper(data$Parking)
```

You can also remove any whitespaces or additional spaces within the data using the syntax below. You can do this to specific cells or columns, as well.

```
#Installing and loading the required packages

install.packages("stringr")

library(stringr)

#Trimming all whitespaces

data$Dist_Taxi<-str_trim(data$Dist_Taxi)
```

If you want to replace a specific letter or word in a specific column, use the code below:

```
#Replacing "Not Provided" with "Not Available"

data$Parking<-str_replace(data$Parking,”Not Provided”,”Not Available”)
```

If you want to replace any outliers in the data set using some summary statistics, like the median or mean, you can use the code below:

```
#Replacing the outliers of a particular column with median

vec1<-boxplot.stats(data$Dist_Taxi)$out;

data$Dist_Taxi[data$Dist_Taxi %in% vec1]<-median(data$Dist_Taxi)
```

The next steps will help you determine how you can identify and remove any missing values in the data set:

```
#Checking for missing values in the entire dataframe
any(is.na(data))
#Checking for the total number of missing values in the entire dataframe
sum(is.na(data))
#Checking for the total number of missing values in a particular column
sum(is.na(data$Dist_Taxi))
#Eliminating missing values completely from the entire dataframe
na.omit(data)
#Eliminating missing values completely from a particular column
na.omit(data$Dist_Taxi)
#Replacing the NA's in the entire dataframe with '0’s
data[is.na(data)]<- 0
#Replacing the NA's in a particular column with '0’s
data$Dist_Taxi[is.na(data$Dist_Taxi)]<-0
#Replacing the NA’s in a particular column with a summary statistics like median
data$Dist_Taxi[is.na(data$Dist_Taxi)]<-median(data$Dist_Taxi)
```

Use the tools below if you want to combine two or more columns in the data set:

```
#Installing and loading the required package
install.packages(“tidyr”)
library(tidyr)
data1<-unite(data = data,col = city_category_with_parking,City_Category,Parking)
View(data1)
```

The function ‘unite’ uses the following arguments:

- Data frame
- New column name
- First column
- Second column

The last two variables in the code are the names of the columns you want to combine.

You can also separate columns if needed:

```
data2<-separate(data = data1,city_category_with_parking,c(“City_Category”,”Parking”), sep = “-“)
```

```
View(data2)
```

The separate function has four arguments:

- Data frame
- Columns to separate
- Name of the new column
- Indicator used to separate the data in the column

When you follow the steps given above, you will end up with a clean data set.

Chapter Ten: Develop a Predictive Model Using R

If you want to understand what predictive modeling is and the strategic areas, let us break the process down into the essential components. You can divide the process into four parts. Every component needs a certain amount of time to execute the process. Let us look at each step and consider the time it will take:

1. Performing descriptive analysis on the data – this takes 50% time
2. The next step is to treat the data and identify any outliers and missing values – 40% time
3. Data modeling – 4% time
4. The estimation of performance – 6% time

These are only assumed times since the time taken by each programmer will vary depending on how well you know the concepts and language. Using the above data, you can cut the time down. Now, let us put the steps discussed above into action. The code used below is not the entire code, but it gives you an idea of what type of code to write. This is a skeleton of the entire algorithm in R.

The first thing to do is to split the entire data set into training and testing data set. Once you do this, you need to import the training data set into R. Use the code below to do this:

```
setwd("C:\\Users\\Tavish\\Desktop\\Kagg\\AV")
```

```
complete <- read.csv("complete_data.csv", stringsAsFactors = TRUE)
```

The next step is to read the names of the columns and the summary in the data set.

```
colnames(complete )
```

```
[1] "ID" "Gender" "City"  "Monthly_Income" "Disbursed" "train"
```

When you extract the summary, you should look for the following data:

- Variables
- Numeric variables
- Factor variables or predictors
- Target variables

If you notice any values missing in the data set, you need to create flags. Use the example below:

```
missing_val_var <- function(data,variable,new_var_name) {

data$new_var_name <- ifelse(is.na(variable),1,0))

return(data$new_var_name)}
```

Now, impute the values in the data set:

```
numeric_impute <- function(data,variable) {

mean1 <- mean(data$variable)

data$variable <- ifelse(is.na(data$variable),mean1,data$variable)

return(new_var_name)

}
```

You can also input other categorical variables in the data set to ensure every missing value in the code says 'Null.' The next thing to do is to pass any imputed variables in the process.

```
create_model <- function(trainData,target) {

set.seed(120)

myglm <- glm(target ~ . , data=trainData, family = "binomial")

return(myglm) }
```

The final step is to make the necessary predictions.

```
score <- predict(myglm, newdata = testData, type = "response")

score_train <- predict(myglm, newdata = complete, type = "response")
```

Once you develop this model, you can test the model using the testing data set.

```
auc(complete$Disbursed,score_train)
```

Chapter Eleven: Create Data Visualization to Communicate the Insight

Once you complete your analysis of the data, you need to use visualization techniques to explain to the stakeholders what you found in the data set. Visualization patterns show people the insights and patterns in the data set, which cannot be absorbed or shown in any other way. It is important to note that data is just bits and pieces of information, but you need to visualize it if you want to make sense of the information available.

In this chapter, we look at what data visualization is in a broad aspect. For instance, some people may look at loading data into excel or any other visualization platform as a form of data visualization. You can then convert the data imported into the tool into a picture of the screen. Having said that, you cannot learn everything about the data set using a table because tables do not help you identify any patterns, hidden or specific, in the data set immediately. A common example of this is the geographical pattern that is studied on a map only after you visualize it.

Bear in mind that every data visualization tool or software allows you to identify a hidden story or pattern in the data set. When you try to visualize the data using different tools, you learn about new patterns in the data set. You may have identified some patterns and relationships in the data set when you had initially downloaded it. There may be other inferences that you may not have identified. Some insights may be due to an error in the data set, while others may result in the development of new insights. You can only identify these patterns when you visualize the data.

You can obtain a new direction and perspective of the data points and variables in the data set through data visualization. You can visualize data in numerous ways. If you have very small data points in the data set, you can easily identify the patterns using a table. A table has the labels, and the values in the data set organized in a structured manner. You can improve the analysis and understanding of the data set by filtering and sorting the data. Tables, however, have limitations. You can use tables to identify and understand the data set, and it’s easy to use tables to compare and understand the difference in the data points. If you want to compare multi-dimensional data sets, do not use tables.

You can also use charts to map the various data points and variables in the data set. These charts allow you to plot the data points in geometric shapes. You can also use scatterplots and other graphs to plot the data. Scatterplots can be used to map two-dimensional data sets. You can also use colors and thermal imaging to learn more about the data points in the set.

Most data collected by organizations are related to people. For example,

- Businesses may collect information about customers to understand and analyze their purchase history and preferences
- They may also collect historical data to determine a customer's credit history
- Businesses can collect information about financial transactions to determine when a customer may spend more or less money
- Governments and other legislative bodies may collect information about the crimes in certain areas and determine which region may have more crimes when compared to others. They may also collect information to determine how some areas in the city or country are doing when compared to others.

One of the most common forms of visualization is graphs. A graph allows you to identify the interconnections in the data sets. You can calculate each data point's position by using graph layout algorithms, and these algorithms make it easier for you to identify the structure in the data set easily. You need to find a way to model the network, and this is an important aspect to consider when it comes to data visualization. Bear in mind that every data set will have some obvious and hidden relationships.

Once you visualize the data using different visualization techniques, you need to understand and find any information from the data set. When you use visualization techniques, consider the following questions:

- What can you see in the image?
- Are there some patterns you have missed?
- Is this visualization exactly what you expected?
- What would happen to the data set if you included more data points?

You may end up with visualizations that do not reveal the information you are looking for. All you should do is try a new visualization. This way, you learn something new about the data.

Chapter Twelve: Document Your Insights from the Data

In most situations, you work with new data. You may develop some expectations or assumptions about the data set before you look at the data. It is best to make a note of these assumptions based on what you think about the data set. This will help you determine if you are biased towards specific trends in the data set. It also reduces the probability of misinterpreting the data set's variables or data points because of the original bias. This is an important tip to keep in mind when it comes to analyzing data.

It is for this reason documentation is extremely important in the process of data analysis. This is the step most people often skip. In the process of documentation, you need to plot the data points and understand them better. You may have created numerous charts as part of the visualization, but these can be confusing to understand. It is for this reason you need to take some time to document the following information:

- Why was this chart created?
- Has the data been updated to fit it into the chart?
- What information can I gather from the chart?

The information you document will help the stakeholders understand what you are trying to convey through the visualization.

Let us now look at which tool you can use to visualize the data in your analysis. When you look at the tools separately, you identify what each tool is good at. It makes sense for you to consider tools that allow you to take data wrangling and visualization into account. You can choose to use different tools if you want to perform the processes separately. When you separate tasks, you need to import and export large volumes or chunks of data. The following are some tools:

- Non-programming visualization tools like Tableau
- Visualization libraries, for example, d3.js, Flare
- Spreadsheets like Excel or Google Docs.
- A programming framework like R, Python, or Pandas
- Data Wrangling software like Google refine or Datawrangler

Conclusion

Data science is a new field of science that is used by organizations to analyze and interpret large volumes of data. This analysis helps businesses make informed decisions about the business. If you are new to data science, this book has all the information you need about data science.

The book introduces the concept of data science and the various processes needed to perform data analysis. You will learn about the various skills, both technical and non-technical, you need to develop to become a data scientist. The book also sheds some light on what different algorithms you can use to become a data scientist.

This book has some information about the different applications of data science. You can learn to implement different data science tools and techniques you learn in this book in different areas of your life. It is important to note that the skills mentioned in the book are essential to develop if you want to become a data scientist. What is most important to understand is that data science is a new field; you need to develop the skills if you want to identify any hidden patterns in the data set.

Bear in mind that you cannot learn everything at the same time. You need to give yourself some time to understand the basics. Once you do this, you can work on improving those skills to help you become an indispensable part of an organization. You can help businesses make informed decisions.

Thank you for purchasing the book. I hope you got the information you are looking for.

References

Ahuja, R. (n.d.). Introduction to Data Science. Coursera website: https://www.coursera.org/specializations/introduction-data-science

Castrounis, A. (n.d.). What Is Data Science, and What Does a Data Scientist Do? InnoArchiTech website: https://www.innoarchitech.com/blog/what-is-data-science-does-data-scientist-do

Data Science Tutorial for Beginners: What is, Basics & Process. (n.d.). www.guru99.com

Holtz, D. (2019, March 27). What is Data Science? 8 Skills That Will Get You Hired in Data | Udacity. Udacity website: https://blog.udacity.com/2014/11/data-science-job-skills.html

Martin, R. (2019, June 10). Linear Regression for Predictive Modeling in R –.Dataquest website: https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/

Perfect way to build a Predictive Model in less than 10 minutes. (2019, June 11). Analytics Vidhya website: https://www.analyticsvidhya.com/blog/2015/09/perfect-build-predictive-model-10-minutes/

Pros and Cons of Data Science - Why Choose Data Science for Your Career - DataFlair. (2019, March 27). DataFlair website: https://data-flair.training/blogs/pros-and-cons-of-data-science/

Sharma, H. (2017, January 5). What Is Data Science? A Beginner's Guide To Data Science. Edureka website: https://www.edureka.co/blog/what-is-data-science/

Sharma, M. (2018, May 2). DATA CLEANING USING R. Data Analytics Edge website: http://dataanalyticsedge.com/2018/05/02/data-cleaning-using-r