Data Analysis and Visualizations

Introduction:

R programming was created for statistics and is used in academic and research fields. R programming has evolved over time and many packages have been created to do data mining, text mining, and data visualizations tasks. R is very mature inthe statistics field, so it is ideal to use R for the data exploration, data understanding, or modeling stages of the CRISP DM model.

R programming:

R is widely used in data science by statisticians and data miners for dataanalysis and the development of statistical software. R is one of the most comprehensive statistical programming languages available, capable of handling everything from data manipulation and visualization to statistical analysis.

R is an implementation of the S programming language, which Was created by Ross Ihahka and Robert Gentlemen at the University of Auckland. R and its librariesare made up of statistical and graphical Techniques, including descriptive statistics, inferential statistics, and Regression analysis. Another strength of R is that it is able to produce Publishable quality graphs and charts, and can use packages like ggplot for Advanced graphs.

R is a software environment and statistical programming language built for statistical computing and data visualization. R's numerous abilities tend to fall into threebroad categories:

- Manipulating data
- Statistical analysis
- Visualizing data

HIGH LEVEL AND LOW LEVEL LANGUAGES:

HIGH LEVEL LANGUAGES:

A high-level programming language (HLL) is designed to be used by a human and is closer to the human language. Its programming style is easier to comprehendand implement than a lower-level programming language (LLL). A high-level programming language needs to be converted to machine language before being executed, so a high-level programming language can be slower.

LOW LEVEL LANGUAGES:

A low-level programming language, on the other hand, is a lot closer to themachine and computer language. A low-level programming language can be executed directly on computer without the need to convert between languages before execution.Thus, a low-level programming language can be faster than a high-level programminglanguage. Low-level programming languages like the assembly language are more inclined towards machine language that deals with bits 0 and 1.

R is a HLL because it shares many similarities to human languages. For Example, in R programming code,

- Var1 <- 1;
- Var2 <- 2;
- >
- Result <- var1 + var2;
- Print(result)
- [1] 3
- >

The R programming code is more like human language. A low-level programming language like the assembly language is more towards the machine language, like 00110110:

0x52ac87: movl7303445 (%ebx), %eax

0x52ac78: call 0x6bfb03

What Is Statistics?

Statistics is a collection of mathematics to deal with the organization, analysis, and interpretation of data. Three main statistical methods are used in the data analysis:descriptive statistics, inferential statistics, and Regressions analysis.

Descriptive statistics summarizes the data and usually focuses on the distribution, the central tendency, and the dispersion of data. The distribution can be normal distribution or binomial distribution, and the Central tendency is to describe thedata with respect to the central of the Data. The central tendency can be the mean, median, and mode of the data.

The dispersion describes the spread of the data, and dispersion can be thevariance, standard deviation, and interquartile range.

Inferential statistics tests the relationship between two data sets or two samples, and a hypothesis is usually set for the statistical relationships between them. The hypothesis can be a null hypothesis or alterative Hypothesis, and rejecting the nullhypothesis is done using tests like the T Test, Chi Square Test, and ANOVA. The Chi Square Test is more for categorical variables, and the T Test is more for continuous variables. The ANOVA test is for more complex applications.

Regression analysis is used to identify the relationships between two variables. Regressions can be linear regressions or non-linear regressions.

The regression can also be a simple linear regression or multiple linearregressions for identifying relationships for more variables.

Data visualization is the technique used to communicate or present data usinggraphs, charts, and dashboards. Data visualizations can help us understand the data more easily.

What Is Data Science?

Data science is a multidisciplinary field that includes statistics, computer Science, machine learning, and domain expertise to get knowledge and insights fromdata. Data science usually ends up developing a data product. A data product is the changing of the data of a company into a product to solve a problem.

For example, a data product can be the product recommendation system usedin Amazon and Lazada. These companies have a lot of data based on shoppers'purchases. Using this data, Amazon and Lazada can identify the shopping patterns ofshoppers and create a recommendation system or data product to recommend other products whenever a shoper buys a product.

The term "data science" has become a buzzword and is now used to representmany areas like data analytics, data mining, text mining, data visualizations, prediction modeling, and so on.

The history of data science started in November 1997, when C. F. Jeff Wu characterized statistical work as data collection, analysis, and decision making, and presented his lecture called "Statistics = Data Science?" In 2001, William S. Clevelandintroduced data science as a field that comprised statistics and some computing in hisarticle called "Data Science: An Action Plan for Expanding the Technical Area of the Field of Statistics."

Statistics is important in data science because it can help analysts or Data scientists analyze and understand data. Descriptive statistics assists in Summarizing the data, inferential statistics tests the relationship between two data sets or samples,and regression analysis explores the relationships between multiple variables. Data visualizations can explore the data with charts, graphs, and dashboards. Regressionsand machine learning Algorithms can be used in predictive analytics to train a model and predict a variable.

What Is Data Mining?

Data mining is closely related to data science. Data mining is the process ofidentifying the patterns from data using statistics, machine learning, and data warehouses or databases.

Extraction of patterns from data is not very new, and early methods include theuse of the Nayes theorem and regressions. The growth of technologies increases the ability in data collection. The growth of technologies also allows the use of statistical learning and machine learning algorithms like neural networks, fuzzy logic, decision trees, Generic algorithms, and support vector machines to uncover the hidden patternsof data. Data mining combines statistics and machine learning, and usually results in the creation of models for making predictions based on historical data.

The cross-industry standard process of data mining, also known as CRISP-DM,is a process used by data mining experts and it is one of the most popular data mining models.

The CRISP-DM model was created in 1996 and involves SPSS, Teradata, Daimler AG, NCR Corporation, and OHRA. The first version was depicted at the fourth CRISP-DM SIG Workshop in Brussels in 1999. Many practitioners use the CRISP-DM model, but IBM is the company that focuses on the CRISP-DM model and includes it inSPSS Modeler.However, the CRISP-DM model is actually application neutral.

What Is Text Mining?

Text mining, text data mining (TDM) or text analytics is the process of deriving high-quality information from text. It involves "the discovery by computer ofnew, previously unknown information, by automatically extracting information fromdifferent written resources.

Text mining, also known as text data mining, is the process of transformingunstructured text into a structured format to identify meaningful patterns and new insights.

Text mining usually involves the process of structuring the input text (usuallyparsing, along with the addition of some derived linguistic features and the removal ofothers, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output.

While data mining is usually used to mine out patterns from numerical data,text mining is used to mine out patterns from textual data like Twitter Tweets, blog postings, and feedback. Text mining, also known as text data Mining, is the process of deriving high quality semantics and knowledge from textual data.

Text mining tasks may consist of text classification, text clustering, and entityextraction; text analytics may include sentiments analysis, TF-IDF, Part-of-speech tagging, name entity recognizing, and text link analysis. Text mining uses the same process as the data mining CRISP-DM Model, with slight difference.

Natural Language Processing

Natural language processing (NLP) is an interdisciplinary subfield of computer science and information retrieval. It is primarily concerned with giving computers the ability to support and manipulate human language. It involves processingnatural language datasets, such as text corpora or speech corpora, using either rule-based or probabilistic (i.e. statistical and, most recently, neural network-based) machinelearning approaches. The goal is a computer capable of "understanding" the contents ofdocuments, including the contextual nuances of the language within them. To this end, natural language processing often borrows ideas from theoretical linguistics. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Natural language processing (NLP) is an area of machine learning **a**nd computer science used to assist the computer in processing and **u**nderstanding natural language. NLP can include part-of-speech tagging parsing, porter stemming, name entity recognition, optical character recognition, sentiment analysis, speech recognition,and more.

NLP works hand in hand with text analytics and text mining. The history of NLPstarted in the 1950s when Alan Turing published an article called "Computing Machinery and Intelligence." Some notable Natural language processing software was developed in the 1960s, such as ELIZA, which provided human-like interactions. In the 1970s, software was developed to write ontologies. In the 1980s, developers introduced Markov Models and initiated research on statistical models to do POS tagging.

Three Types of Analytics

Selecting the type of analytics can be difficult and challenging; luckily, analytics can be categorized into descriptive analytics, predictive analytics, and prescriptive analytics. Noanalytic type is better than the others, but they can be combined with each other.

- **Descriptive Analytics:** Uses data analytics to know what happened.
- **Predictive Analytics:** Uses statistical learning and machine learning topredict the future.
- **Prescriptive Analytics:** Uses simulation algorithms to know what should be done.

Descriptive Analytics

Descriptive analytics uses statistics to summarize the data using descriptivestatistics, inferential statistics to test the two data sets and samples, and regression analysis to study the relationships between multiple variables .

Predictive Analytics

Predictive analytics predicts a variable by implementing machine learning and statistical learning algorithms. In statistics, regressions can be used to predict a variable. For example, y = mx + c. You can determine m and c by training a linear regression model using historical data. Y is the variable to predict, x is the input variable. If you put in x value, you can predict the y.

Prescriptive Analytics

This is a field that allows a user to find the number of inputs to get a certain outcome. In simple form, this kind of analytics is used to provide advice. For example, y = mx + c. You have the m and c values. You want a Y outcome, so what value should you put into x? To get the x value, what kind of things does your company need to do orwhat kind of advice do you need to give to the company? If you have multiple linear regressions, there are many x variables, so you need some simulation or evolutionary search algorithm to get the x values.

Big Data

Big data is data sets that are very big and complex for a computer to **p**rocess.Big data has challenges that may include capturing data, data **st**orage, data analysis, and data visualizations. There are three properties or characteristics of big data.

Volume:

People are now more connected, so there are many more data sources, and as a consequence, the amount of data increased exponentially. The increase of data requires more computing power to process and analyze it. Traditional computing poweris not able to process and analyze this data.

Velocity:

The speed of data is increasing and the speed of data coming in is so fast**t**hat it is very difficult to process and analyze the data. Tradition computing **m**ethodscan't process and analyze at the speed of data coming in.

Variety:

More sources means more data in different formats and types, such as Images, videos, voice, speech, textual data, and numerical data, both **u**nstructured and structured. Various data formats require different **m**ethods to extract the data from them.This means that the data is difficult to process and analyze, and traditional computing methods can't process **s**uch data.

Data grows very quickly, due to IoT devices like mobile devices, wireless sensor networks, and RFID readers. Based on an IDC report, Global data will increasefrom 4.4 zettabytes to 44 zettabytes from 2013 to 2020.

Relational databases and desktop statistics and data science software have challenges to process and analyze big data. Hence, big data requires parallel and distributed systems like Hadoop and Apache Spark to process and analyze the data.

Why R?

When learning data science, many people struggle with choosing which programming languages and data sciences to learn. There are many programming languages available for data science, like R, Python, SAS, Java, and more. There are many data science software packages to learn, such as SPSS Statistics, SPSS Modeler,SAS Enterprise Miner, Tableau, RapidMiner, Weka, GATE, and more.

I recommend learning R for statistics because it was developed for Statistics in the first place. Python is a real programming language, so you can developreal applications and software via Python programming.

Hence, if you want to develop a data product or data application, Python can be a better choice. R programming is very strong in statistics, so it is ideal for dataexploration or data understanding using descriptive Statistics, inferential statistics, regression analysis, and data visualizations.

R is also ideal for modeling because you can use statistical learning like Regressions for predictive analytics. R also has some packages for data Mining, text mining, and machine

learning like Rattle, CARET, and TM. R Programming can also interface with big data systems like Apache Spark Using Sparklyr. SAS programming is commercial, and Java has direct Interfaces with GATE, Stanford NLP, and Weka. SPSS Statistics, SPSS Modeler, SAS Enterprise Miner, and Tableau are data science software Packages with GUIs and are commercial. RapidMiner, Weka, and GATE are Open source software packages for data science.

R is also heavily used in many of the companies that hire data Scientists. Google and Facebook have data scientists who use R. R is also used in companies likeBank of America, and so on.

R is also heavily used in academia, and R is very popular among academic researchers, who can use R graphics for publications. Scripts written in R can be used on different operating systems, including Linux, Apple, and Windows, as long as the Rinterpreter is installed. This is not possible with languages like C#.

What Is R?

R programming is for statistical computing and is supported by the R Foundation for Statistical Computing. R programming is used by many academics and researchers for data and statistical analysis, and the popularity of R has risen over time.

R is a GNU package and is available under the GNU General Public License, which can be assumed to be free to a certain extent and is open source. R is available in a command line application.

R programming is an implementation of the S programming language,its libraries consist of statistical and data visualization techniques, and it can conduct descriptive statistics, inferential statistics, and regressions analysis. You will explore the differences between the R programming Command line application and the RStudio IDE, as well as the basics of the descriptive statistics features and the data visualizationfeatures.

The Integrated Development Environment

An integrated development environment (IDE) is a software application that provides comprehensive facilities for software development. An IDE normally consists ofat least a source-code editor, build automation tools, and a debugger.

An IDE is a software application that helps programmers develop software more easily and more productively. An IDE is made up of a code Editor, compiler, anddebugger tools. Code editors usually offer syntax highlighting and intelligent code completion.

Some IDEs, like NetBeans, also have an interpreter and others, like SharDevelop, don't. Some IDEs have a version control system and tools like a graphicaluser interface (GUI) builder, and many IDEs have class and object browsers. IDEs are developed to increase the productivity of the developer by combining features like a code editor, compiler, debugger, and Interpreter. This is different from a programming code text editor like VI and NotePad++, which offer syntax highlighting but usually don't communicate with the debugger and compier.

The beginning of IDEs can be traced back to when punched cards were submitted to the compiler in early systems. Dartmouth BASIC was the First programming language to be created with an IDE. Maestro I was later created by Softlab Munich and can be considered the first full IDE between 1970s and 1980s.

RStudio: The IDE for R

In R programming, RStudio is the most popular IDE. RStudio has a code Editor that consists of syntax highlighting and intelligent code completion functions. RStudio also has a workspace showing all the variables and history. You may double-click the variables to view them using tables and other options.

The R console is in RStudio so you can view the results of the R scripts afterrunning the scripts; you can also type into the R console with R code to do some simplecomputing. The Plots and Others portion is available in RStudio to let you view the charts and graphs plotted from R scripts. The Plots and Others portion allows you to easily save the graphs and charts.

Installation of R and RStudio

1. In order to code R scripts, you must install the R programming command line application. You can download the R programmingcommand line application from www.r-project.org/.
2. In this book, you will download R for Windows. You can also downloadfor Linux and Mac OS.
3. To install the software, double-click the download setup file and follow the instructions of the installer to install the R programming command Line application.
4. After the R programming command line application is installed, you canstart it.
5. You can create your own Hello World application by using the print() Function. The Hello World application is the standard first application tobe developed when learning a programming language. Type the following Code into the RGui:Print("Hello World");The print() function is

used to print some text on the console screen.You may print any textother than the “Hello World”.

6. RStudio is the most popular IDE for the R programming language. RStudio helps you write R programming code more easily and more productively. To download and install RStudio, visit www.rstudio.com/.
7. Download the latest version. For this book, you will download the 64-bitWindows version. After downloading the RStudio installer or setupfile, Double-click the file to install the RStudio IDE.
8. After installing the RStudio IDE, you can run the RStudio IDE software.
9. Before running the script, you need to select the R programming command line application version to use. Click Tools ➤ Global Options.
10. Click the Change button to select the R version.
11. For the beginner, choose the R version . If you Want to change the Rversion in the future, you can use this method to do so
12. After clicking OK and choosing the R version, you must restart the RStudio IDE.
13. After restarting RStudio, the Console tab should show the selected R version.

Writing Scripts in R and RStudio

1. You can read a comma-separated values (CSV) file using the read.csv() Function in the R programming language.

```
myData <- read.csv(file=”D:/data.csv”,

header=TRUE, sep=”,”); myData;
```

2. In the R programming language, you can use the summary() functionto get the basic descriptive statistics for all the variables. Summary(myData)
3. In the R programming language, you can plot a scatterplot using thePlot() function.
   Plot(myData$x, myData$x2);
4. RStudio is an IDE that provides a GUI for the R programming Command line application. RStudio provide word suggestions and syntax highlighting for the R programming language. The RStudio IDEfor the R Programming language .
5. With RStudio, you can write all the code into the code editor and runthe script.
   myData <- read.csv(file=”D:/data.csv”, header=TRUE, sep=”,”);myData;
   summary(myData);
   plot(myData$x, myData$x2);
   As you type the code, RStudio shows the intelligent code completion.
6. You must select all the R code in the code editor and click Run or Ctrl +Enter to run the script.
7. Or you can click Code ➤ Run Region ➤ Run All.
8. The RStudio IDE offers syntax highlighting features in the code editor. When you run the R script, you can view the results in the Console taband see the scatterplot in the Plots tab. By double-clicking myData in the Global

Environment tab, you can view the data loaded from the .csv file.

Basic Syntax

You will use R for applied statistics, which can be used in the data understanding and modeling stages of the CRISP-DM data mining model. R programming is a programming language with object-oriented Programming features. Rprogramming was created for statistics and is used in the academic and research fields.However, before you go into Statistics, you need to learn to program R scripts.In this chapter, you will explore the syntax of R programming. I will discuss the R console and code editor in RStudio, as well as R objects and the data structure of R programming, from variables and data types to lists, vectors, matrices, and data frames. I will also discuss conditional statements, loops, and functions. Then you will create a simple calculator after learning the basics.

Writing in R Console

The R console offers a fast and easy way to do Statistical calculations and some data visualizations. The R console is also like a calculator, so you can always usethe R console to calculate some Math equations.To do math calculations, you can just type in some math equations like. 1+1

- 1 + 1

[1] 2

1 – 3

- 1 – 3

[1] -2

1 * 5

- 1 * 5

[1] 5

1 / 6

- 1 / 6

[1] 0.1666667

Tan(2)

- Tan(2)

[1] -2.18504

To do some simple statistical calculations, you can so the following:
Standard deviation

Sd(c(1, 2, 3, 4, 5, 6))

>sd(c(1, 2, 3, 4, 5, 6))

[1] 1.870829

Mean

Mean(c(1, 2, 3, 4, 5, 6))

- Mean(c(1, 2, 3, 4, 5, 6))

[1] 3.5

Min

Min(c(1, 2, 3, 4, 5,6))

- Min(c(1, 2, 3, 4, 5, 6))

[1]1

Plot(c(1, 2, 3, 4, 5, 6), c(2, 3, 4, 5, 6, 7))

- Plot(c(1, 2, 3, 4, 5, 6), c(2, 3, 4, 5, 6, 7))

To sum up, the R console, despite being basic, offers the following

Advantages:

- High performance
- Fast prototyping and testing of your ideas and logic
- Before proceeding further, such as when developing windows Form applications
- Personally, I use the R console application to test algorithms and other codefragments when in the process of developing complex R scripts.

Using the Code Editor

The RStudio IDE offers features like a code editor, debugger, and compiler that communicate with the R command line application or R console. The R code editoroffers features like intelligent code completion and syntax .

- To create a new script in RStudio, click File ➤ New ➤ R Script
- You can then code your R Script. For now, type in the following code,A <-
  1;

  B <- 2;

  A/B;

  A * B;

  A + B;

  A – B;

  A^2;

  B^2;
- To run the R script, highlight the code in the code editor and click Run.
- To view the results of the R script, look in the R console of RStudio.
- You can also see that in the Environment tab, there are two variables.

Adding Comments to the Code

You can add comments to the code. Comments are text that will not be run by the R console. You can add in a comment by putting # in front of the text. The comment is foryou to describe your code to let anyone read it more easily.

```
#Create variable A with value 1A

<- 1;

#Create variable B with value 2B

<- 2;

#Calculate A divide BA/B;

#Calculate A times B

A * B;

#Calculate A plus BA

+ B;

#Calculate A subtract B
```

A – B;

#Calculate A to power of 2

A^2;

#Calculate B to power of 2

B^2;

You can rerun the code and you should get the result

Variables

Let's look into the code and scripts you used previously. You actually created two variables, A and B, and assigned some values to the two variables.

A <- 1

B <- 2

In this code, A is a variable, and B is a variable also. <- means assign. A <- 1 means variable A is assigned a value of 1. 1 is a numeric type. B <- 2 means variable B is assigned a value of 2. 2 is a numeric type.If you want to assign text or character values,you add quotations, like A <- "Hello World".

Variable A is assigned a text value of "Hello World". Character and numeric are datatypes.

Data Types

Data types are the types or kind of information or data a variable is holding. Adata type can be numeric and character.

For example,A

<- “abc”

B <- 1.2

In R, data types are automatically determined. Because of the quotations surroundingthe values, variable A is of the character data type, While variable B is of the numeric data type.R is also capable of storing other data types.

Vectors

A vector is a basic data structure or R object for storing a set of values of the same datatype. A vector is the most basic and common data structure in R. A vector is used when you want to store and modify a set of values. The Data types can be logical, integer, double, and character. The integer data type is used to store number values without a decimal, and the double data type is used to store number values with a decimal.

- Vectors can be created using the c() function.
- You can check the data type of the vector using typeof() and class():
- You can check the number of elements or values in a vector using the Length()function:
- You can also use the operator : to create a vector:
- To retrieve the second element or value of a vector, use the [] brackets and putin the element number to retrieve:
- You can also retrieve the elements in the vector using another vector
- You can also retrieve elements of a vector using a logical vector:
- You can also use more than or less than signs to retrieve element
- You can modify a vector as follows using assign, <-:

Lists

A list is like a vector. It is an R object that can store a set of values or elements, but a listcan store values of different data types. A list is also another common data structure in

R. You use a list when you want to modify and store a set of values of different data types. A vector can only store values of the same data type. The syntax to create a listis as follows:

Variable = list(..., ..., ...)

Matrix

A matrix is like a vector, but it has two dimensions. You usually use a matrix to modify and store values from a data set because a matrix has two dimensions. A matrix is goodwhen you plan to do linear algebra types or mathematical operations. For a data set with different types, you need to use a data frame.

Data Frame

A data frame is a special list or R object that is multidimensional and is usually used tostore data read from an Excel or .csv file. A matrix can only store values of the same type, but a data frame can store values of different types. To declare a data frame, usethe following syntax:

Variable = data.frame(colName1 = c(..., ..., ...),

colName2 = c(..., ..., ...), ...)

Logical Statements

If...else statements are usually the logical fragments of your code in R. They give yourprogram some intelligence and decision making by specifying the if rules:

```
    If (Boolean expression) {
 #Codes to execute if Boolean expression is true
}else {
#code to execute if Boolean expression is false }
```

Boolean Operator

== Equal to

>= Greater than or equal to

<=Lesser than or equal to

>Greater than

< Lesser than

!= Not equal to

Loops

Loops are used to repeat certain fragments of code. For example, if you want print the “This is R.” Message 100 times, it will be very tiresome to type print(“This is R. “); 100 times. You can use loops to print the message 100 times more easily. Loops canusually be used to go through a set of vectors, lists, or data frames. In R, there are several loop options:

For Loop

```
For (value in vector) {
#statements
}
```

While Loop

You can also use while loop to loop until you meet a specific Boolean

Expression:

While (Boolean Expression) {

#Code to run or repeat until Bolean Expression is fals

}

Break and Next Keywords

In loop statements, you can use the break keyword and the next keyword. Thebreak keyword is to stop the iterations of the loop. The next keyword is to skip the current iteration of a loop.

Repeat Loop

The repeat loop repeats the code many times, and there is no BooleanExpression to meet. To stop the loop, you must use the break keyword.

Repeat {

#code to repeat

}

Functions

Functions help you organize your code and allow you to reuse code fragments whenever you need. To create a function, use the following

Syntax:

Function_name<- function(arg1, arg2, ...) {#

Codes fragments

Function_name = #value to return }

Descriptive Statistics

Descriptive statistics is a set of math used to summarize data. Descriptive statistics can be distribution, central tendency, and dispersion of data. The distribution can be a normal distribution or binomial distribution. The central tendency can be mean,median, and mode. The dispersion or spreadness can be the range, interquartile range,variance, and standard deviation.

What Is Descriptive Statistics?

Descriptive statistics summarizes the data and usually focuses on the distribution, the central tendency, and dispersion of the data. The distributions can be normal distribution, binomial distribution, and other distributions like Bernoulli distribution. Binomial distribution and normal distribution are the more popular and important distributions, especially normal distribution. When exploring data and many statistical tests, you will usually look for the normality of the data, which is how normal the data is or how likely it is that the data is normally distributed .The Central Limit Theorem states that the mean of a sample or subset of a distribution will be equal to the normal distribution mean when the sample size increases, regardless whether the sample is from a normal distribution. The central tendency, not the central limit theorem,is used to describe the data with respect to the center of the data. Central tendency canbe the mean, median, and mode of the data. The dispersion describes the spread of thedata, and dispersion can be the variance, standard deviation, and interquartile range. Descriptive statistics summarizes the data set, lets us have a feel and understanding of the data and variables, and allows us to decide Or determine whether we should use inferential statistics to identify the relationship between data sets or use regression analysis to identify the relationships between variables.

Reading Data Files

R programming allow you to import a data set, which can be comma separated values (CSV) file, Excel file, tab-separated file, JSON file, or others. Readingdata into the R console or R is important, since you must have some data before you can do statistical computing and understand the data. Before you look into importing data into the R console, you must determine your workplace or work directory first. Youshould always set the current workspace directory to tell R the location of your current project folder. This allows for easier references to data files and scripts.

To print the current work directory, you use the getwd() function:# get

the current workspace location

Print(getwd());

Print(getwd());

[1] “C:/Users/gohmi/Documents”

You can set the work directory using the setwd() function:#set the

current workspace location

Setwd(“D:/R”); #input your own file directory, for here we use “D:/R”

Setwd(“D:/R”);

To get the new work directory location, you can use the getwd()

Function:

#get the new

workspace

Print(getwd());

Print(getwd());

[1] “D:/R”

You can put the data.csv data set into D:/R folder.

Basic Data Processing :

Data processing, manipulation of data by a computer. It includes the conversion of raw data to machine-readable form, flow of data through the CPU and memory to output devices, and formatting or transformation of output. Any use of computers to perform defined operations on data can be included under data processing. In the commercial world, data processing refers to the processing of data required to run organizations and businesses.

After importing the data, you may need to do some simple data processing like selecting data, sorting data, filtering data, getting unique values, and removing missingvalues.

The four main stages of data processing cycle are:

- Data collection.
- Data input.
- Data processing.
- Data output.

Data collection

The first stage of data collection involves gathering raw data from various sources, suchas sensors, databases, or customer surveys. It is essential to ensure the collected data is accurate, complete, and relevant to the analysis or processing goals.

Data input

The next stage is data input. In this stage, the clean and prepped data is fed into a processing system, which could be software or an algorithm designed for specific data types or analysis goals. Various methods, such as manual entry, data import from external sources, or automatic data capture, can be used to input data into the processing system.

Data processing

In the data processing stage, the input data is transformed, analyzed, and organized to produce relevant information. Several data processing techniques, like filtering, sorting,

aggregation, or classification, may be employed to process the data. The choice of methods depends on the desired outcome or insights from the data.

Data output

The data output and interpretation stage deals with presenting the processed data in aneasily digestible format. This could involve generating reports, graphs, or visualizations that simplify complex data patterns and help with decision-making. Furthermore, the output data should be interpreted and analyzed to extract valuable insights and knowledge.