EDA

Rahul Sinha

2022-09-22

Exploratory Data Analysis

Exploratory Data Analysis (EDA) – Exploratory Data Analytics refers to the approach used for the analysis of data using several visualization techniques. As we know that not everyone is familiar with the technical terms of the data and is not able to analyze the data accordingly, thus EDA is used so that it becomes easier for people to understand the data through visualization.  
With the use of graphical representation and statistical summaries, EDA is used to identify trends and patterns in the data. The process for conducting EDA includes many different steps those are: -

• Data Collection

• Data Description

• Data Relations

• Data Preprocessing

• Data Plots

Data Collection

It’s essential to first define data before discussing collection. The short answer is that data is a variety of types of information formatted in a specific way. As a result, data collecting is the act of gathering, gauging, and analyzing precise data from a range of particular essentials in order to overcome the problems, provide answers, assess results, and predict trends and possibilities. Additionally, the data can be divided into two types those are qualitative and quantitative types. Where the qualitative type of data is used to cover the description of the data through color, size, quality, and appearance. On the other hand, quantitative data is used to deal with numbers, such as statistics, poll numbers, percentages, etc. The data collection breaksdown into two types namely Primary Data and Secondary Data. In this case the data has been collected from a secondary source, the data that has been used in this assignmnet is ‘iris’ which is a default dataset that has been saved in the directory of R-Studio.

# Data History

The Iris flower data set, often known as Fisher’s Iris data set, is a multivariate data set that was first used as an illustration of linear discriminant analysis by British statistician and biologist Ronald Fisher in his 1936 paper The use of numerous measurements in taxonomic issues. Because Edgar Anderson gathered the data to measure the morphologic variation of Iris blossoms from three related species, it is also referred to as Anderson’s Iris data set. In the Gaspé Peninsula, two of the three species were taken “all from the same pasture, chosen on the same day, and measured at the same time by the same person using the same apparatus.”

The data set includes 50 samples from each of the three Iris species.(Iris versicolor, Iris setosa, and Iris virginica). From each sample, the length and width of the sepals and petals, both in centimetres, were measured. Fisher created a linear discriminant model to separate the species based on the combination of these four features.

# Data Set

The data set contains the record of 5 variables and 150 records

library(ggplot2)  
library(corrplot)

## corrplot 0.92 loaded

# LOADING DATASET  
data<-iris  
data

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa  
## 4 4.6 3.1 1.5 0.2 setosa  
## 5 5.0 3.6 1.4 0.2 setosa  
## 6 5.4 3.9 1.7 0.4 setosa  
## 7 4.6 3.4 1.4 0.3 setosa  
## 8 5.0 3.4 1.5 0.2 setosa  
## 9 4.4 2.9 1.4 0.2 setosa  
## 10 4.9 3.1 1.5 0.1 setosa  
## 11 5.4 3.7 1.5 0.2 setosa  
## 12 4.8 3.4 1.6 0.2 setosa  
## 13 4.8 3.0 1.4 0.1 setosa  
## 14 4.3 3.0 1.1 0.1 setosa  
## 15 5.8 4.0 1.2 0.2 setosa  
## 16 5.7 4.4 1.5 0.4 setosa  
## 17 5.4 3.9 1.3 0.4 setosa  
## 18 5.1 3.5 1.4 0.3 setosa  
## 19 5.7 3.8 1.7 0.3 setosa  
## 20 5.1 3.8 1.5 0.3 setosa  
## 21 5.4 3.4 1.7 0.2 setosa  
## 22 5.1 3.7 1.5 0.4 setosa  
## 23 4.6 3.6 1.0 0.2 setosa  
## 24 5.1 3.3 1.7 0.5 setosa  
## 25 4.8 3.4 1.9 0.2 setosa  
## 26 5.0 3.0 1.6 0.2 setosa  
## 27 5.0 3.4 1.6 0.4 setosa  
## 28 5.2 3.5 1.5 0.2 setosa  
## 29 5.2 3.4 1.4 0.2 setosa  
## 30 4.7 3.2 1.6 0.2 setosa  
## 31 4.8 3.1 1.6 0.2 setosa  
## 32 5.4 3.4 1.5 0.4 setosa  
## 33 5.2 4.1 1.5 0.1 setosa  
## 34 5.5 4.2 1.4 0.2 setosa  
## 35 4.9 3.1 1.5 0.2 setosa  
## 36 5.0 3.2 1.2 0.2 setosa  
## 37 5.5 3.5 1.3 0.2 setosa  
## 38 4.9 3.6 1.4 0.1 setosa  
## 39 4.4 3.0 1.3 0.2 setosa  
## 40 5.1 3.4 1.5 0.2 setosa  
## 41 5.0 3.5 1.3 0.3 setosa  
## 42 4.5 2.3 1.3 0.3 setosa  
## 43 4.4 3.2 1.3 0.2 setosa  
## 44 5.0 3.5 1.6 0.6 setosa  
## 45 5.1 3.8 1.9 0.4 setosa  
## 46 4.8 3.0 1.4 0.3 setosa  
## 47 5.1 3.8 1.6 0.2 setosa  
## 48 4.6 3.2 1.4 0.2 setosa  
## 49 5.3 3.7 1.5 0.2 setosa  
## 50 5.0 3.3 1.4 0.2 setosa  
## 51 7.0 3.2 4.7 1.4 versicolor  
## 52 6.4 3.2 4.5 1.5 versicolor  
## 53 6.9 3.1 4.9 1.5 versicolor  
## 54 5.5 2.3 4.0 1.3 versicolor  
## 55 6.5 2.8 4.6 1.5 versicolor  
## 56 5.7 2.8 4.5 1.3 versicolor  
## 57 6.3 3.3 4.7 1.6 versicolor  
## 58 4.9 2.4 3.3 1.0 versicolor  
## 59 6.6 2.9 4.6 1.3 versicolor  
## 60 5.2 2.7 3.9 1.4 versicolor  
## 61 5.0 2.0 3.5 1.0 versicolor  
## 62 5.9 3.0 4.2 1.5 versicolor  
## 63 6.0 2.2 4.0 1.0 versicolor  
## 64 6.1 2.9 4.7 1.4 versicolor  
## 65 5.6 2.9 3.6 1.3 versicolor  
## 66 6.7 3.1 4.4 1.4 versicolor  
## 67 5.6 3.0 4.5 1.5 versicolor  
## 68 5.8 2.7 4.1 1.0 versicolor  
## 69 6.2 2.2 4.5 1.5 versicolor  
## 70 5.6 2.5 3.9 1.1 versicolor  
## 71 5.9 3.2 4.8 1.8 versicolor  
## 72 6.1 2.8 4.0 1.3 versicolor  
## 73 6.3 2.5 4.9 1.5 versicolor  
## 74 6.1 2.8 4.7 1.2 versicolor  
## 75 6.4 2.9 4.3 1.3 versicolor  
## 76 6.6 3.0 4.4 1.4 versicolor  
## 77 6.8 2.8 4.8 1.4 versicolor  
## 78 6.7 3.0 5.0 1.7 versicolor  
## 79 6.0 2.9 4.5 1.5 versicolor  
## 80 5.7 2.6 3.5 1.0 versicolor  
## 81 5.5 2.4 3.8 1.1 versicolor  
## 82 5.5 2.4 3.7 1.0 versicolor  
## 83 5.8 2.7 3.9 1.2 versicolor  
## 84 6.0 2.7 5.1 1.6 versicolor  
## 85 5.4 3.0 4.5 1.5 versicolor  
## 86 6.0 3.4 4.5 1.6 versicolor  
## 87 6.7 3.1 4.7 1.5 versicolor  
## 88 6.3 2.3 4.4 1.3 versicolor  
## 89 5.6 3.0 4.1 1.3 versicolor  
## 90 5.5 2.5 4.0 1.3 versicolor  
## 91 5.5 2.6 4.4 1.2 versicolor  
## 92 6.1 3.0 4.6 1.4 versicolor  
## 93 5.8 2.6 4.0 1.2 versicolor  
## 94 5.0 2.3 3.3 1.0 versicolor  
## 95 5.6 2.7 4.2 1.3 versicolor  
## 96 5.7 3.0 4.2 1.2 versicolor  
## 97 5.7 2.9 4.2 1.3 versicolor  
## 98 6.2 2.9 4.3 1.3 versicolor  
## 99 5.1 2.5 3.0 1.1 versicolor  
## 100 5.7 2.8 4.1 1.3 versicolor  
## 101 6.3 3.3 6.0 2.5 virginica  
## 102 5.8 2.7 5.1 1.9 virginica  
## 103 7.1 3.0 5.9 2.1 virginica  
## 104 6.3 2.9 5.6 1.8 virginica  
## 105 6.5 3.0 5.8 2.2 virginica  
## 106 7.6 3.0 6.6 2.1 virginica  
## 107 4.9 2.5 4.5 1.7 virginica  
## 108 7.3 2.9 6.3 1.8 virginica  
## 109 6.7 2.5 5.8 1.8 virginica  
## 110 7.2 3.6 6.1 2.5 virginica  
## 111 6.5 3.2 5.1 2.0 virginica  
## 112 6.4 2.7 5.3 1.9 virginica  
## 113 6.8 3.0 5.5 2.1 virginica  
## 114 5.7 2.5 5.0 2.0 virginica  
## 115 5.8 2.8 5.1 2.4 virginica  
## 116 6.4 3.2 5.3 2.3 virginica  
## 117 6.5 3.0 5.5 1.8 virginica  
## 118 7.7 3.8 6.7 2.2 virginica  
## 119 7.7 2.6 6.9 2.3 virginica  
## 120 6.0 2.2 5.0 1.5 virginica  
## 121 6.9 3.2 5.7 2.3 virginica  
## 122 5.6 2.8 4.9 2.0 virginica  
## 123 7.7 2.8 6.7 2.0 virginica  
## 124 6.3 2.7 4.9 1.8 virginica  
## 125 6.7 3.3 5.7 2.1 virginica  
## 126 7.2 3.2 6.0 1.8 virginica  
## 127 6.2 2.8 4.8 1.8 virginica  
## 128 6.1 3.0 4.9 1.8 virginica  
## 129 6.4 2.8 5.6 2.1 virginica  
## 130 7.2 3.0 5.8 1.6 virginica  
## 131 7.4 2.8 6.1 1.9 virginica  
## 132 7.9 3.8 6.4 2.0 virginica  
## 133 6.4 2.8 5.6 2.2 virginica  
## 134 6.3 2.8 5.1 1.5 virginica  
## 135 6.1 2.6 5.6 1.4 virginica  
## 136 7.7 3.0 6.1 2.3 virginica  
## 137 6.3 3.4 5.6 2.4 virginica  
## 138 6.4 3.1 5.5 1.8 virginica  
## 139 6.0 3.0 4.8 1.8 virginica  
## 140 6.9 3.1 5.4 2.1 virginica  
## 141 6.7 3.1 5.6 2.4 virginica  
## 142 6.9 3.1 5.1 2.3 virginica  
## 143 5.8 2.7 5.1 1.9 virginica  
## 144 6.8 3.2 5.9 2.3 virginica  
## 145 6.7 3.3 5.7 2.5 virginica  
## 146 6.7 3.0 5.2 2.3 virginica  
## 147 6.3 2.5 5.0 1.9 virginica  
## 148 6.5 3.0 5.2 2.0 virginica  
## 149 6.2 3.4 5.4 2.3 virginica  
## 150 5.9 3.0 5.1 1.8 virginica

# GETTING TO KNOW THE DATASET BETTER   
names(data)

## [1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"

dim(data)

## [1] 150 5

str(data)

## 'data.frame': 150 obs. of 5 variables:  
## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...  
## $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...  
## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...  
## $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...  
## $ Species : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...

class(data)

## [1] "data.frame"

head(data, n = 5)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa  
## 4 4.6 3.1 1.5 0.2 setosa  
## 5 5.0 3.6 1.4 0.2 setosa

tail(data, n = 5)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 146 6.7 3.0 5.2 2.3 virginica  
## 147 6.3 2.5 5.0 1.9 virginica  
## 148 6.5 3.0 5.2 2.0 virginica  
## 149 6.2 3.4 5.4 2.3 virginica  
## 150 5.9 3.0 5.1 1.8 virginica

# Data Summary

# SUMMARY OF THE DATA  
## applying the necessary library for conducting the summary function  
library(pastecs)  
library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following object is masked from 'package:psych':  
##   
## describe

## The following objects are masked from 'package:base':  
##   
## format.pval, units

##conducting the summary functions of the different library  
summary(data)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100   
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300   
## Median :5.800 Median :3.000 Median :4.350 Median :1.300   
## Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199   
## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800   
## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500   
## Species   
## setosa :50   
## versicolor:50   
## virginica :50   
##   
##   
##

describe(data)

## data   
##   
## 5 Variables 150 Observations  
## --------------------------------------------------------------------------------  
## Sepal.Length   
## n missing distinct Info Mean Gmd .05 .10   
## 150 0 35 0.998 5.843 0.9462 4.600 4.800   
## .25 .50 .75 .90 .95   
## 5.100 5.800 6.400 6.900 7.255   
##   
## lowest : 4.3 4.4 4.5 4.6 4.7, highest: 7.3 7.4 7.6 7.7 7.9  
## --------------------------------------------------------------------------------  
## Sepal.Width   
## n missing distinct Info Mean Gmd .05 .10   
## 150 0 23 0.992 3.057 0.4872 2.345 2.500   
## .25 .50 .75 .90 .95   
## 2.800 3.000 3.300 3.610 3.800   
##   
## lowest : 2.0 2.2 2.3 2.4 2.5, highest: 3.9 4.0 4.1 4.2 4.4  
## --------------------------------------------------------------------------------  
## Petal.Length   
## n missing distinct Info Mean Gmd .05 .10   
## 150 0 43 0.998 3.758 1.979 1.30 1.40   
## .25 .50 .75 .90 .95   
## 1.60 4.35 5.10 5.80 6.10   
##   
## lowest : 1.0 1.1 1.2 1.3 1.4, highest: 6.3 6.4 6.6 6.7 6.9  
## --------------------------------------------------------------------------------  
## Petal.Width   
## n missing distinct Info Mean Gmd .05 .10   
## 150 0 22 0.99 1.199 0.8676 0.2 0.2   
## .25 .50 .75 .90 .95   
## 0.3 1.3 1.8 2.2 2.3   
##   
## lowest : 0.1 0.2 0.3 0.4 0.5, highest: 2.1 2.2 2.3 2.4 2.5  
## --------------------------------------------------------------------------------  
## Species   
## n missing distinct   
## 150 0 3   
##   
## Value setosa versicolor virginica  
## Frequency 50 50 50  
## Proportion 0.333 0.333 0.333  
## --------------------------------------------------------------------------------

summary.abund(data)

##   
## Sorting of descriptors according to abundance for:   
##   
## Coefficient f:   
## 0 variables sorted  
##   
## Number of individuals (% of most abundant in log):  
## NULL  
##   
## Percent of non-zero values:  
## NULL

# Data Description

stat.desc(data)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## nbr.val 150.00000000 150.00000000 150.0000000 150.00000000 NA  
## nbr.null 0.00000000 0.00000000 0.0000000 0.00000000 NA  
## nbr.na 0.00000000 0.00000000 0.0000000 0.00000000 NA  
## min 4.30000000 2.00000000 1.0000000 0.10000000 NA  
## max 7.90000000 4.40000000 6.9000000 2.50000000 NA  
## range 3.60000000 2.40000000 5.9000000 2.40000000 NA  
## sum 876.50000000 458.60000000 563.7000000 179.90000000 NA  
## median 5.80000000 3.00000000 4.3500000 1.30000000 NA  
## mean 5.84333333 3.05733333 3.7580000 1.19933333 NA  
## SE.mean 0.06761132 0.03558833 0.1441360 0.06223645 NA  
## CI.mean.0.95 0.13360085 0.07032302 0.2848146 0.12298004 NA  
## var 0.68569351 0.18997942 3.1162779 0.58100626 NA  
## std.dev 0.82806613 0.43586628 1.7652982 0.76223767 NA  
## coef.var 0.14171126 0.14256420 0.4697441 0.63555114 NA

#CHECKING NULL VALUES  
is.na(data)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## [1,] FALSE FALSE FALSE FALSE FALSE  
## [2,] FALSE FALSE FALSE FALSE FALSE  
## [3,] FALSE FALSE FALSE FALSE FALSE  
## [4,] FALSE FALSE FALSE FALSE FALSE  
## [5,] FALSE FALSE FALSE FALSE FALSE  
## [6,] FALSE FALSE FALSE FALSE FALSE  
## [7,] FALSE FALSE FALSE FALSE FALSE  
## [8,] FALSE FALSE FALSE FALSE FALSE  
## [9,] FALSE FALSE FALSE FALSE FALSE  
## [10,] FALSE FALSE FALSE FALSE FALSE  
## [11,] FALSE FALSE FALSE FALSE FALSE  
## [12,] FALSE FALSE FALSE FALSE FALSE  
## [13,] FALSE FALSE FALSE FALSE FALSE  
## [14,] FALSE FALSE FALSE FALSE FALSE  
## [15,] FALSE FALSE FALSE FALSE FALSE  
## [16,] FALSE FALSE FALSE FALSE FALSE  
## [17,] FALSE FALSE FALSE FALSE FALSE  
## [18,] FALSE FALSE FALSE FALSE FALSE  
## [19,] FALSE FALSE FALSE FALSE FALSE  
## [20,] FALSE FALSE FALSE FALSE FALSE  
## [21,] FALSE FALSE FALSE FALSE FALSE  
## [22,] FALSE FALSE FALSE FALSE FALSE  
## [23,] FALSE FALSE FALSE FALSE FALSE  
## [24,] FALSE FALSE FALSE FALSE FALSE  
## [25,] FALSE FALSE FALSE FALSE FALSE  
## [26,] FALSE FALSE FALSE FALSE FALSE  
## [27,] FALSE FALSE FALSE FALSE FALSE  
## [28,] FALSE FALSE FALSE FALSE FALSE  
## [29,] FALSE FALSE FALSE FALSE FALSE  
## [30,] FALSE FALSE FALSE FALSE FALSE  
## [31,] FALSE FALSE FALSE FALSE FALSE  
## [32,] FALSE FALSE FALSE FALSE FALSE  
## [33,] FALSE FALSE FALSE FALSE FALSE  
## [34,] FALSE FALSE FALSE FALSE FALSE  
## [35,] FALSE FALSE FALSE FALSE FALSE  
## [36,] FALSE FALSE FALSE FALSE FALSE  
## [37,] FALSE FALSE FALSE FALSE FALSE  
## [38,] FALSE FALSE FALSE FALSE FALSE  
## [39,] FALSE FALSE FALSE FALSE FALSE  
## [40,] FALSE FALSE FALSE FALSE FALSE  
## [41,] FALSE FALSE FALSE FALSE FALSE  
## [42,] FALSE FALSE FALSE FALSE FALSE  
## [43,] FALSE FALSE FALSE FALSE FALSE  
## [44,] FALSE FALSE FALSE FALSE FALSE  
## [45,] FALSE FALSE FALSE FALSE FALSE  
## [46,] FALSE FALSE FALSE FALSE FALSE  
## [47,] FALSE FALSE FALSE FALSE FALSE  
## [48,] FALSE FALSE FALSE FALSE FALSE  
## [49,] FALSE FALSE FALSE FALSE FALSE  
## [50,] FALSE FALSE FALSE FALSE FALSE  
## [51,] FALSE FALSE FALSE FALSE FALSE  
## [52,] FALSE FALSE FALSE FALSE FALSE  
## [53,] FALSE FALSE FALSE FALSE FALSE  
## [54,] FALSE FALSE FALSE FALSE FALSE  
## [55,] FALSE FALSE FALSE FALSE FALSE  
## [56,] FALSE FALSE FALSE FALSE FALSE  
## [57,] FALSE FALSE FALSE FALSE FALSE  
## [58,] FALSE FALSE FALSE FALSE FALSE  
## [59,] FALSE FALSE FALSE FALSE FALSE  
## [60,] FALSE FALSE FALSE FALSE FALSE  
## [61,] FALSE FALSE FALSE FALSE FALSE  
## [62,] FALSE FALSE FALSE FALSE FALSE  
## [63,] FALSE FALSE FALSE FALSE FALSE  
## [64,] FALSE FALSE FALSE FALSE FALSE  
## [65,] FALSE FALSE FALSE FALSE FALSE  
## [66,] FALSE FALSE FALSE FALSE FALSE  
## [67,] FALSE FALSE FALSE FALSE FALSE  
## [68,] FALSE FALSE FALSE FALSE FALSE  
## [69,] FALSE FALSE FALSE FALSE FALSE  
## [70,] FALSE FALSE FALSE FALSE FALSE  
## [71,] FALSE FALSE FALSE FALSE FALSE  
## [72,] FALSE FALSE FALSE FALSE FALSE  
## [73,] FALSE FALSE FALSE FALSE FALSE  
## [74,] FALSE FALSE FALSE FALSE FALSE  
## [75,] FALSE FALSE FALSE FALSE FALSE  
## [76,] FALSE FALSE FALSE FALSE FALSE  
## [77,] FALSE FALSE FALSE FALSE FALSE  
## [78,] FALSE FALSE FALSE FALSE FALSE  
## [79,] FALSE FALSE FALSE FALSE FALSE  
## [80,] FALSE FALSE FALSE FALSE FALSE  
## [81,] FALSE FALSE FALSE FALSE FALSE  
## [82,] FALSE FALSE FALSE FALSE FALSE  
## [83,] FALSE FALSE FALSE FALSE FALSE  
## [84,] FALSE FALSE FALSE FALSE FALSE  
## [85,] FALSE FALSE FALSE FALSE FALSE  
## [86,] FALSE FALSE FALSE FALSE FALSE  
## [87,] FALSE FALSE FALSE FALSE FALSE  
## [88,] FALSE FALSE FALSE FALSE FALSE  
## [89,] FALSE FALSE FALSE FALSE FALSE  
## [90,] FALSE FALSE FALSE FALSE FALSE  
## [91,] FALSE FALSE FALSE FALSE FALSE  
## [92,] FALSE FALSE FALSE FALSE FALSE  
## [93,] FALSE FALSE FALSE FALSE FALSE  
## [94,] FALSE FALSE FALSE FALSE FALSE  
## [95,] FALSE FALSE FALSE FALSE FALSE  
## [96,] FALSE FALSE FALSE FALSE FALSE  
## [97,] FALSE FALSE FALSE FALSE FALSE  
## [98,] FALSE FALSE FALSE FALSE FALSE  
## [99,] FALSE FALSE FALSE FALSE FALSE  
## [100,] FALSE FALSE FALSE FALSE FALSE  
## [101,] FALSE FALSE FALSE FALSE FALSE  
## [102,] FALSE FALSE FALSE FALSE FALSE  
## [103,] FALSE FALSE FALSE FALSE FALSE  
## [104,] FALSE FALSE FALSE FALSE FALSE  
## [105,] FALSE FALSE FALSE FALSE FALSE  
## [106,] FALSE FALSE FALSE FALSE FALSE  
## [107,] FALSE FALSE FALSE FALSE FALSE  
## [108,] FALSE FALSE FALSE FALSE FALSE  
## [109,] FALSE FALSE FALSE FALSE FALSE  
## [110,] FALSE FALSE FALSE FALSE FALSE  
## [111,] FALSE FALSE FALSE FALSE FALSE  
## [112,] FALSE FALSE FALSE FALSE FALSE  
## [113,] FALSE FALSE FALSE FALSE FALSE  
## [114,] FALSE FALSE FALSE FALSE FALSE  
## [115,] FALSE FALSE FALSE FALSE FALSE  
## [116,] FALSE FALSE FALSE FALSE FALSE  
## [117,] FALSE FALSE FALSE FALSE FALSE  
## [118,] FALSE FALSE FALSE FALSE FALSE  
## [119,] FALSE FALSE FALSE FALSE FALSE  
## [120,] FALSE FALSE FALSE FALSE FALSE  
## [121,] FALSE FALSE FALSE FALSE FALSE  
## [122,] FALSE FALSE FALSE FALSE FALSE  
## [123,] FALSE FALSE FALSE FALSE FALSE  
## [124,] FALSE FALSE FALSE FALSE FALSE  
## [125,] FALSE FALSE FALSE FALSE FALSE  
## [126,] FALSE FALSE FALSE FALSE FALSE  
## [127,] FALSE FALSE FALSE FALSE FALSE  
## [128,] FALSE FALSE FALSE FALSE FALSE  
## [129,] FALSE FALSE FALSE FALSE FALSE  
## [130,] FALSE FALSE FALSE FALSE FALSE  
## [131,] FALSE FALSE FALSE FALSE FALSE  
## [132,] FALSE FALSE FALSE FALSE FALSE  
## [133,] FALSE FALSE FALSE FALSE FALSE  
## [134,] FALSE FALSE FALSE FALSE FALSE  
## [135,] FALSE FALSE FALSE FALSE FALSE  
## [136,] FALSE FALSE FALSE FALSE FALSE  
## [137,] FALSE FALSE FALSE FALSE FALSE  
## [138,] FALSE FALSE FALSE FALSE FALSE  
## [139,] FALSE FALSE FALSE FALSE FALSE  
## [140,] FALSE FALSE FALSE FALSE FALSE  
## [141,] FALSE FALSE FALSE FALSE FALSE  
## [142,] FALSE FALSE FALSE FALSE FALSE  
## [143,] FALSE FALSE FALSE FALSE FALSE  
## [144,] FALSE FALSE FALSE FALSE FALSE  
## [145,] FALSE FALSE FALSE FALSE FALSE  
## [146,] FALSE FALSE FALSE FALSE FALSE  
## [147,] FALSE FALSE FALSE FALSE FALSE  
## [148,] FALSE FALSE FALSE FALSE FALSE  
## [149,] FALSE FALSE FALSE FALSE FALSE  
## [150,] FALSE FALSE FALSE FALSE FALSE

after checking the dataset through is.na() we can conclude that the dataset does’nt contain any NULL value. That means we don’t have to alter any of the values in it.

# Data Relations

There are many methods that are used for the purpose of Data Relations and those methods are of various types, mainly Data Visualization is used for checking out the relations between different variables in a data. Now, we would learn about the Data Visualization,

## Data Visualization

The depiction of data through the use of typical graphics, such as infographics, charts, and even animations, is known as data visualization. These informational visual representations make complex data relationships and data-driven insights simple to comprehend. The different Data Visualization techniques are:

### Correlation Matrix:

Simply said, a correlation matrix is a table that shows the correlation coefficients for various variables. The correlation between all potential pairs of values in a table is shown in the matrix. It is an effective tool for compiling a sizable dataset and for locating and displaying data patterns. The correlation heatmap is considered to be arguably the best when it comes to Data Visualizations.

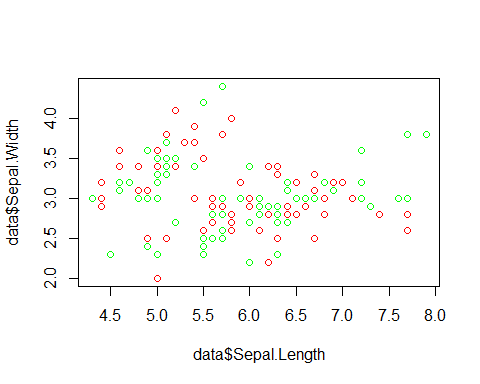
#DATA VISUALIZATION  
## correlation of the data  
pairs(data)



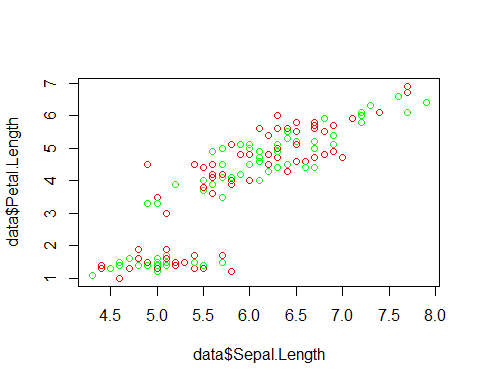
### Scatterplot:

When it comes to data visualizations, correlation heatmaps are relatively difficult to beat, but scatterplots are perhaps one of the most helpful representations. A scatterplot is a form of a graph in which two variables, such as age and height, are “plotted” along two axes. There are numerous uses for scatterplots. Similar to correlation matrices, it aids in the rapid comprehension of the relationship between two variables, the detection of outliers, and the application of polynomial multiple regression models.

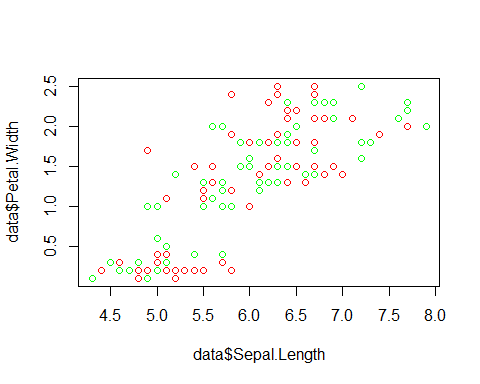
### setting colors to the scatter plot for better understanding  
colors<- c("red", "green")  
plot(data$Sepal.Length, data$Sepal.Width, col = colors)



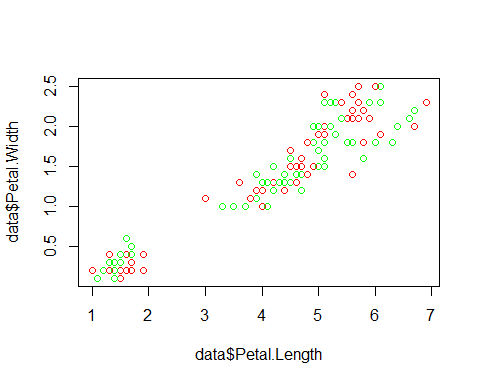
plot(data$Sepal.Length, data$Petal.Length, col = colors)



plot(data$Sepal.Length, data$Petal.Width, col = colors)



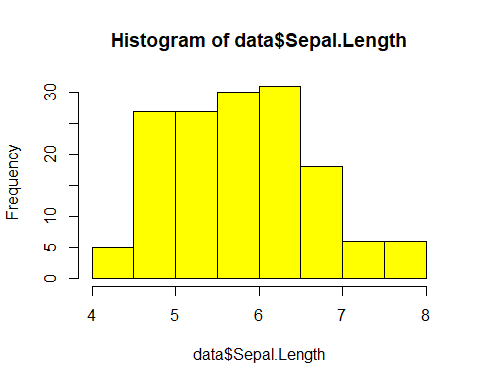
plot(data$Petal.Length, data$Petal.Width, col = colors)



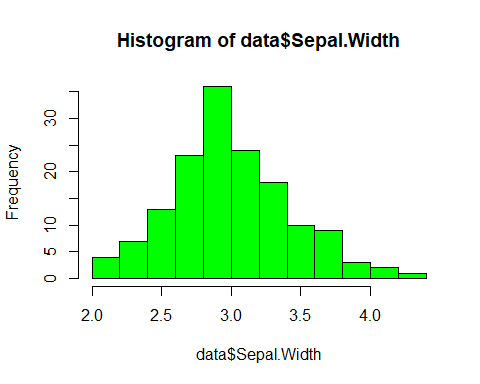
### Histogram:

For examining the relationship between two variables, correlation matrices and scatterplots are helpful. But what if you simply wanted to look at one particular variable? Histograms are useful in this situation. Although histograms resemble bar graphs, they display the distribution of a variable’s range of values.

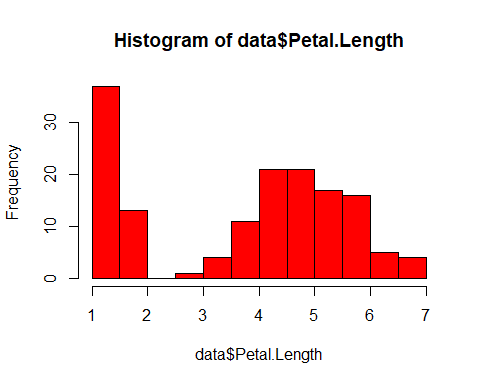
## HISTOGRAM  
hist(data$Sepal.Length, col = 'yellow')



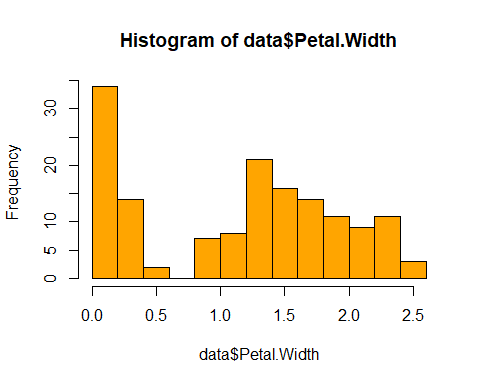
hist(data$Sepal.Width, col = 'green')



hist(data$Petal.Length, col = 'red')



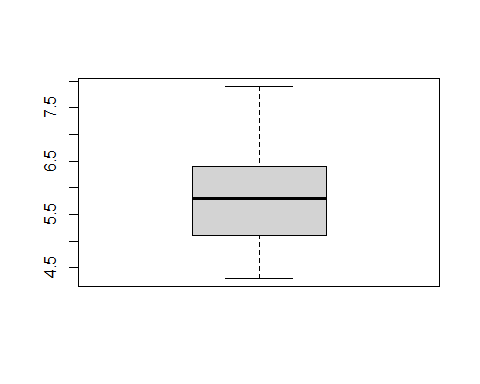
hist(data$Petal.Width, col = 'orange')



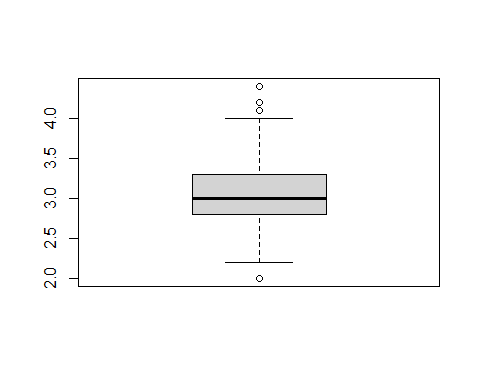
### Boxplot:

You may discover that you require more information than just the measures of central tendency for specific distributions or datasets (median, mean, and mode). You must have information regarding the data’s variability or dispersion. A boxplot is a graph that effectively conveys the distribution of the values in the data. Boxplots have the advantage of taking up less area than a histogram or Scatter plots, which is helpful for comparing distributions across numerous groups or datasets. Despite the fact that they may appear simple in comparison.

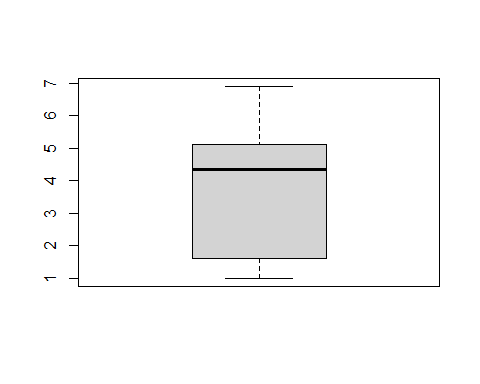
## BOXPLOT  
boxplot(data$Sepal.Length)



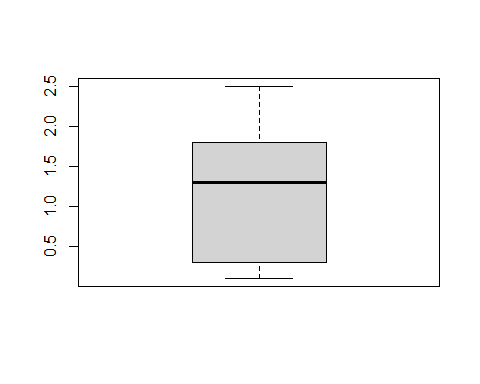
boxplot(data$Sepal.Width)



boxplot(data$Petal.Length)



boxplot(data$Petal.Width)

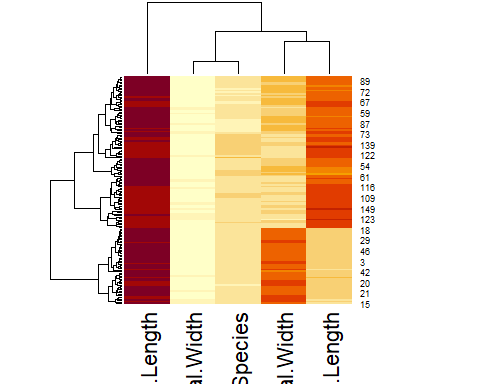


here we can see that sepal width has some outliers in it, as it is exceeding the maximum limit.

### Heatmap:

A data visualisation approach known as a heat map (or heatmap) uses colour to depict a phenomenon’s size in two dimensions. The reader will receive clear visual clues about how the occurrence is clustered or fluctuates over space from the fluctuation in colour, which may be by hue or intensity.

## HEATMAP  
### for heatmap to work we needed numeric vector, so we converted our dataframe into a numeric matrix, as  
a = data.matrix(data)  
heatmap(a)



# Bibliography:

<https://en.wikipedia.org/wiki/Iris_flower_data_set>