K. J. Somaiya College of Engineering,
Mumbai-77
(A Constituent College of Somaiya Vidyavihar University)

Batch: HADS - 3 Roll No.: 16010122151

Experiment No. 4

Title: Exploratory Data Analysis

Aim: Use R libraries to implement exploratory data analysis on chosen datasets.

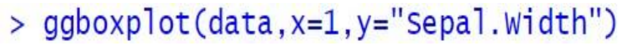
Expected Outcome of Experiment:
CO3: Explain the significance of exploratory data analysis (EDA) in data science
CO5: Apply basic tools to carry out EDA for the Data Science process. **Books/ Journals/ Websites referred:**

1. Data Mining Concepts and Techniques Jiawei Han, Michelin Kamber, Jian Pie, 3rd edition

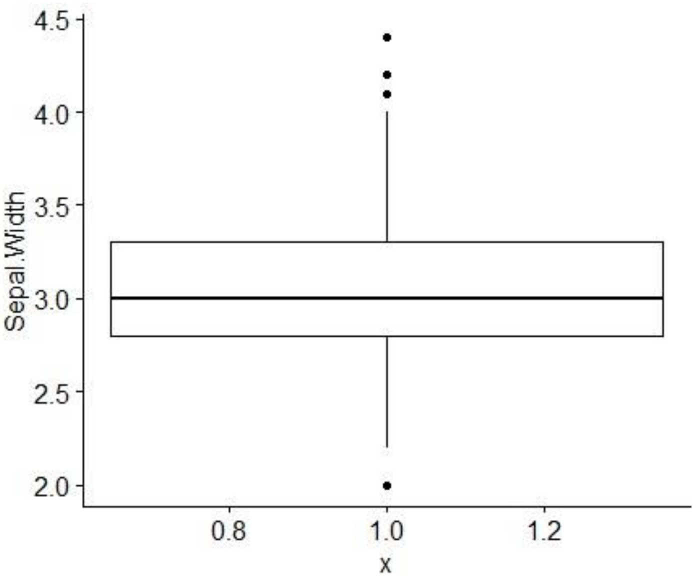
---

How to Remove Outliers from Data Including Multi-Variables in R

In our case, we select the quantitative variables from iris data. We have four variables – sepal length, sepal width, petal length and petal width and 150 observations. The outliers can be observed using the boxplot() function.



```
> ggboxplot(data,x=1,y="Sepal.Width")
```



There exist four outliers seen in the sepal width variable.

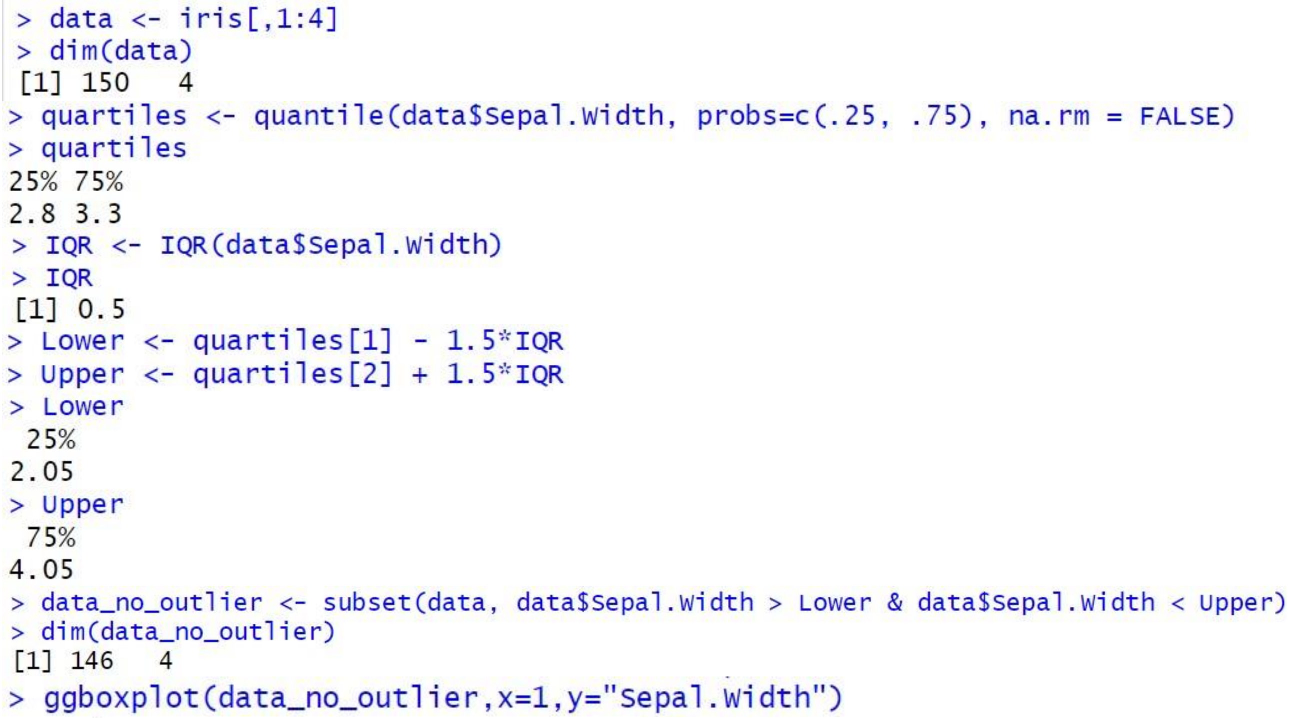
Outliers are either 1.5×IQR or more above the third quartile or 1.5×IQR or more below the first quartile.

We find first (Q1) and third (Q3) quartiles by using the quantile() function.

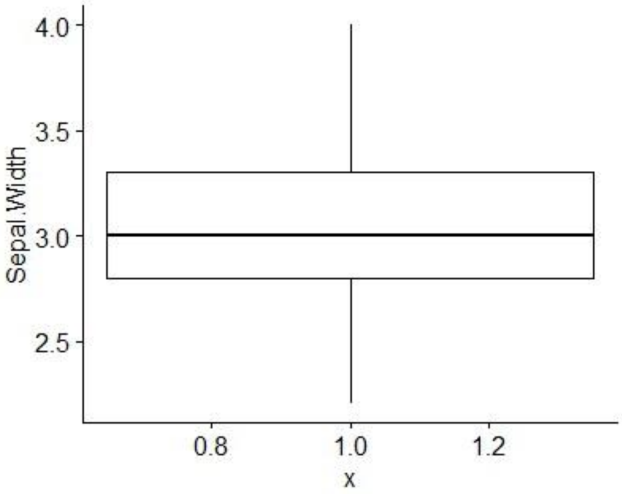
Then, the interquartile range (IQR) is found by the IQR() function.

Then, we calculate Q1 – 1.5*IQR to find the lower limit for outliers.

After that, we calculate Q3 + 1.5*IQR to find the upper limit for outliers. Then, we use the subset() function to eliminate outliers.



```
> data <- iris[,1:4]
> dim(data)
[1] 150   4
> quartiles <- quantile(data$Sepal.Width, probs=c(.25, .75), na.rm = FALSE)
> quartiles
25% 75%
2.8 3.3
> IQR <- IQR(data$Sepal.Width)
> IQR
[1] 0.5
> Lower <- quartiles[1] - 1.5*IQR
> Upper <- quartiles[2] + 1.5*IQR
> Lower
 25%
2.05
> Upper
 75%
4.05
> data_no_outlier <- subset(data, data$Sepal.Width > Lower & data$Sepal.Width < Upper)
> dim(data_no_outlier)
[1] 146   4
> ggboxplot(data_no_outlier,x=1,y="Sepal.Width")
```



Handling missing values

1. Multiple NA or NAN values can exist in a vector.
2. To deal with NA type of missing values in a vector we can use is.na() function by passing the vector as an argument.
3. To deal with the NAN type of missing values in a vector we can use is.nan() function by passing the vector as an argument.

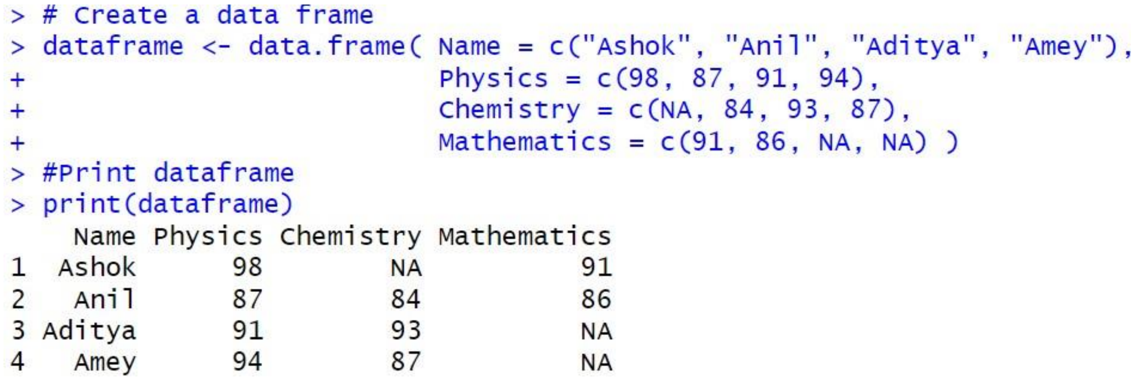
4. Generally, NAN values can be included in the NA type but the vice-versa is not true.

Removing Missing Data/ Values

1. na.omit − It simply rules out any rows that contain any missing value and forgets those rows forever.
2. na.exclude − This ignores rows having at least one missing value.
3. na.pass − Take no action.
4. na.fail − It terminates the execution if any of the missing values are found.

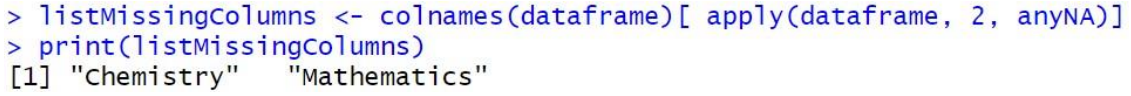
Filling Missing Values with Mean or Median

Consider a dataframe that contains NA values for example:



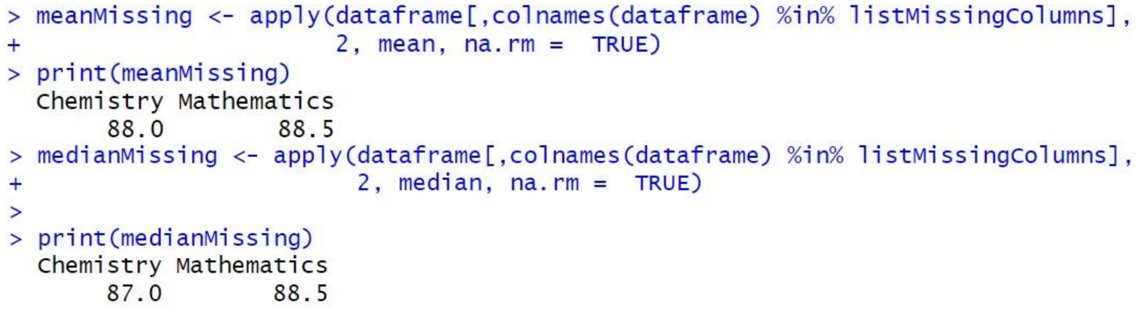
```
> # Create a data frame
> dataframe <- data.frame( Name = c("Ashok", "Anil", "Aditya", "Amey"),
+                          Physics = c(98, 87, 91, 94),
+                          Chemistry = c(NA, 84, 93, 87),
+                          Mathematics = c(91, 86, NA, NA) )
> #Print dataframe
> print(dataframe)
    Name Physics Chemistry Mathematics
1  Ashok      98        NA          91
2   Anil      87        84          86
3 Aditya      91        93          NA
4   Amey      94        87          NA
```

Create a list of columns having at least one NA value.



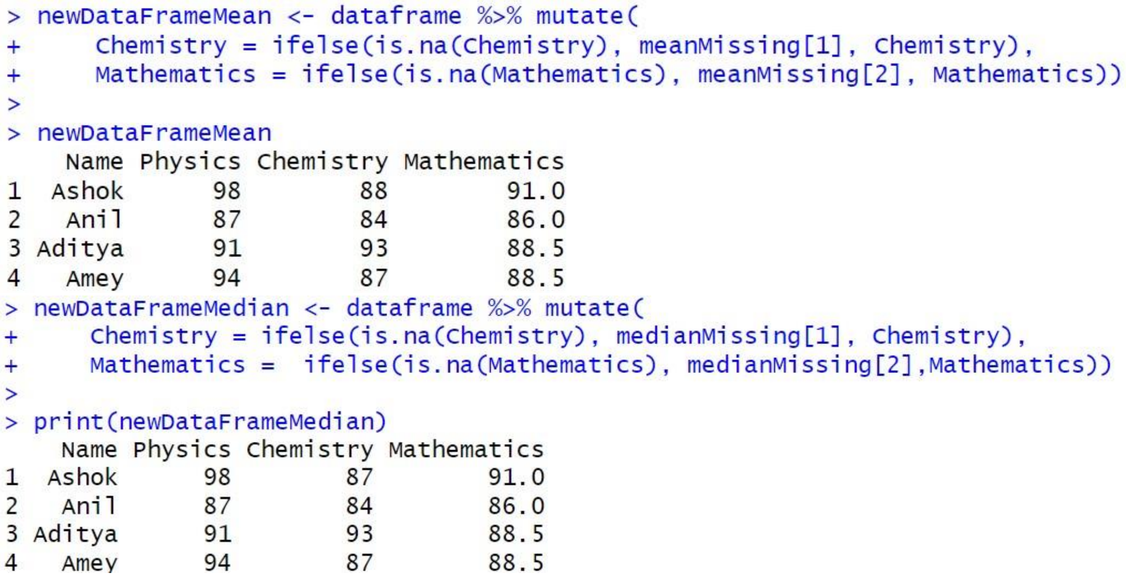
```
> listMissingColumns <- colnames(dataframe)[ apply(dataframe, 2, anyNA)]
> print(listMissingColumns)
[1] "Chemistry"   "Mathematics"
```

Compute the mean and median of the corresponding columns. Since we need to omit NA values in the missing columns, therefore, we can pass the "na.rm = True" argument to the apply() function.



```
> meanMissing <- apply(dataframe[,colnames(dataframe) %in% listMissingColumns],
+                      2, mean, na.rm =  TRUE)
> print(meanMissing)
  Chemistry Mathematics
       88.0        88.5
> medianMissing <- apply(dataframe[,colnames(dataframe) %in% listMissingColumns],
+                        2, median, na.rm =  TRUE)
>
> print(medianMissing)
  Chemistry Mathematics
       87.0        88.5
```

Now our mean and median values of corresponding columns are ready. In this step, we will replace NA values with mean and median using the mutate() function which is defined under the “dplyr” package.



```
> newDataFrameMean <- dataframe %>% mutate(
+     Chemistry = ifelse(is.na(Chemistry), meanMissing[1], Chemistry),
+     Mathematics = ifelse(is.na(Mathematics), meanMissing[2], Mathematics))
>
> newDataFrameMean
    Name Physics Chemistry Mathematics
1  Ashok      98        88        91.0
2   Anil      87        84        86.0
3 Aditya      91        93        88.5
4   Amey      94        87        88.5
> newDataFrameMedian <- dataframe %>% mutate(
+     Chemistry = ifelse(is.na(Chemistry), medianMissing[1], Chemistry),
+     Mathematics =  ifelse(is.na(Mathematics), medianMissing[2],Mathematics))
>
> print(newDataFrameMedian)
    Name Physics Chemistry Mathematics
1  Ashok      98        87        91.0
2   Anil      87        84        86.0
3 Aditya      91        93        88.5
4   Amey      94        87        88.5
```

What is Exploratory Data Analysis (EDA) ?

Exploratory data analysis (EDA) is the process of analysing data to uncover their key features. Exploratory data analysis is used to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed

Types of EDA

- Univariate non-graphical
- Multivariate non-graphical
- Univariate graphical ● Multivariate graphical

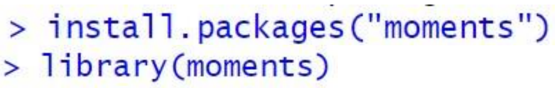
Univariate non-graphical EDA

A simple tabulation of the frequency of each category is the best univariate nongraphical EDA for categorical data.

For numerical data, find the measure of central tendency and spread including skewness and kurtosis

Estimate Skewness and Kurtosis

Load the moments library

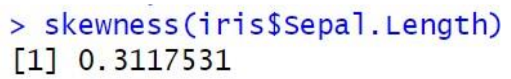


```
> install.packages("moments")
> library(moments)
```

Calculate skewness. Skewness is a measure of symmetry.

Negative skewness: mean of the data < median and the data distribution is left-skewed.

Positive skewness: mean of the data > median and the data distribution is right-skewed.



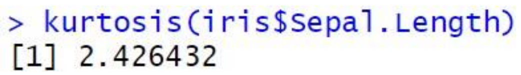
```
> skewness(iris$Sepal.Length)
[1] 0.3117531
```

Distribution is skewed towards the right.

The normal distribution has zero kurtosis and thus the standard tail shape. It is said to be mesokurtic.

Negative kurtosis would indicate a thin-tailed data distribution, and is said to be platykurtic.

Positive kurtosis would indicate a fat-tailed distribution, and is said to be leptokurtic.



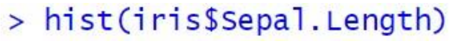
```
> kurtosis(iris$Sepal.Length)
[1] 2.426432
```

It is leptokurtic.

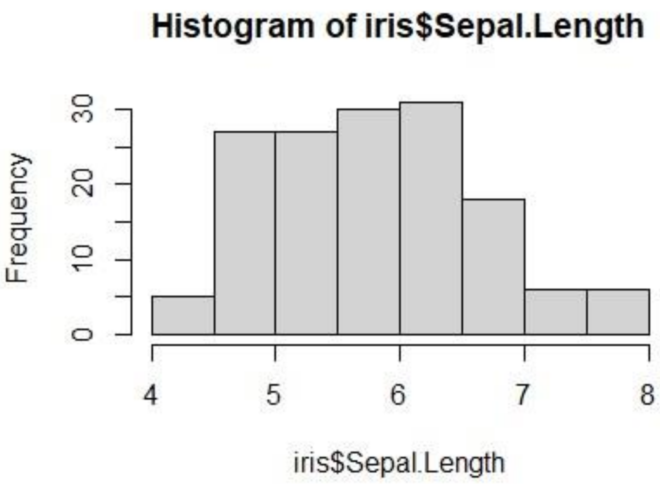
Uni-variate graphical EDA

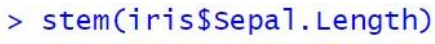
Common sorts of univariate graphics are:

1. ***Histogram***: The foremost basic graph is a histogram, which may be a barplot during which each bar represents the frequency (count) or proportion (count/total count) of cases for a variety of values. Histograms are one of the simplest ways to quickly learn a lot about your data, including central tendency, spread, modality, shape and outliers.
2. ***Stem-and-leaf plots***: An easy substitute for a histogram may be stem-and-leaf plots. It shows all data values and therefore the shape of the distribution.
3. ***Box Plots***: Another very useful univariate graphical technique is the boxplot. Boxplots are excellent at presenting information about central tendency and show robust measures of location and spread also as providing information about symmetry and outliers, although they will be misleading about aspects like multimodality. One among the simplest uses of boxplots is within the sort of side-by-side boxplots.
4. ***Quantile-normal plots***: The ultimate univariate graphical EDA technique is the most intricate. It's called the quantile-normal or QN plot or more generally the quantile-quantile or QQ plot. it's wont to see how well a specific sample follows a specific theoretical distribution. It allows detection of non-normality and diagnosis of skewness and kurtosis

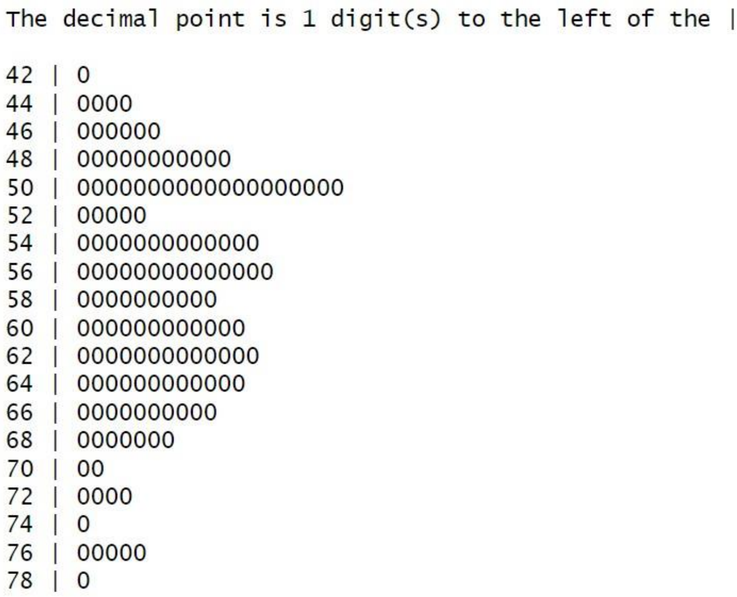


```
> hist(iris$Sepal.Length)
```

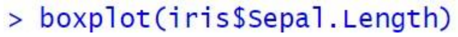
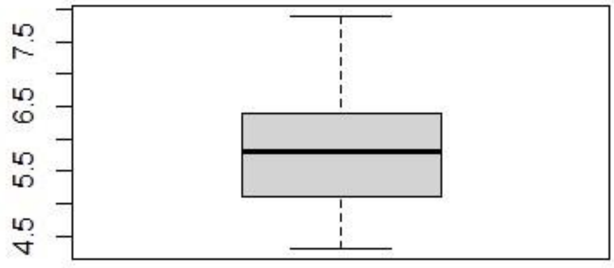
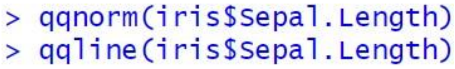
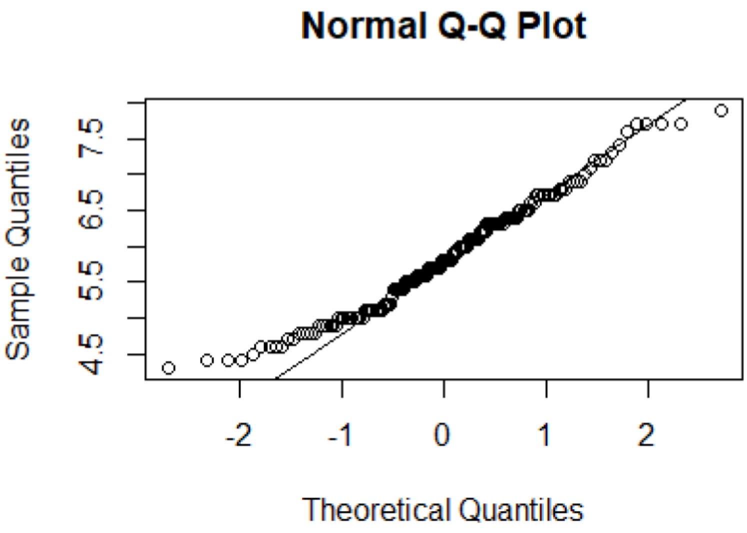




```
> stem(iris$Sepal.Length)
```

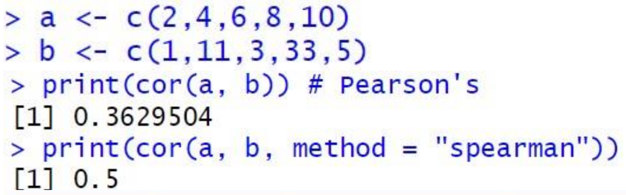


```
The decimal point is 1 digit(s) to the left of the |

42 | 0
44 | 0000
46 | 000000
48 | 00000000000
50 | 0000000000000000000
52 | 00000
54 | 0000000000000
56 | 00000000000000
58 | 0000000000
60 | 000000000000
62 | 0000000000000
64 | 000000000000
66 | 0000000000
68 | 0000000
70 | 00
72 | 0000
74 | 0
76 | 00000
78 | 0
```

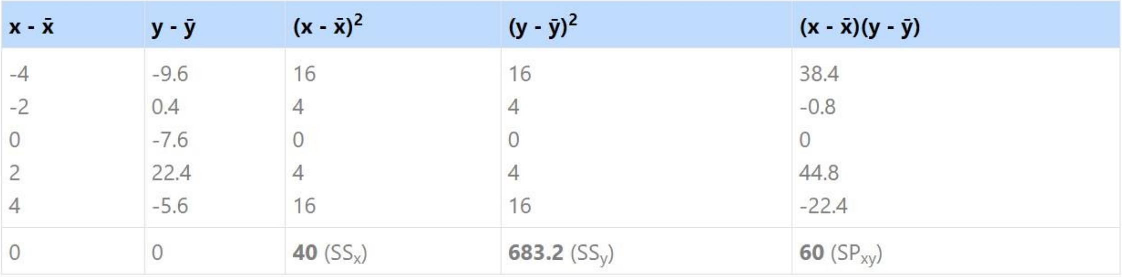
 `> boxplot(iris$Sepal.Length)`	
 `> qqnorm(iris$Sepal.Length)` `> qqline(iris$Sepal.Length)`	

Multi-variate non-graphical EDA

Multivariate non-graphical EDA techniques generally show the relationship between two or more variables in the form of either cross-tabulation (please refer to experiment 2) or by calculating the correlation coefficient.

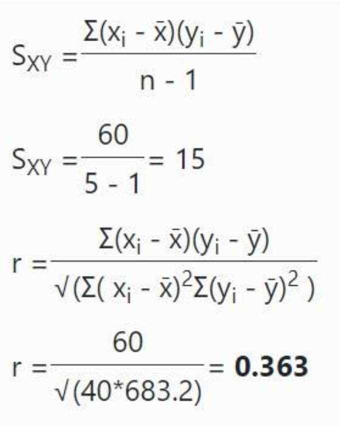


```
> a <- c(2,4,6,8,10)
> b <- c(1,11,3,33,5)
> print(cor(a, b)) # Pearson's
[1] 0.3629504
> print(cor(a, b, method = "spearman"))
[1] 0.5
```



x - x̄	y - ȳ	$(x - \bar{x})^2$	$(y - \bar{y})^2$	(x - x̄)(y - ȳ)
-4 -2 0 2 4	-9.6 0.4 -7.6 22.4 -5.6	16 4 0 4 16	16 4 0 4 16	38.4 -0.8 0 44.8 -22.4
0	0	**40** (SS_x)	**683.2** (SS_y)	**60** (SP_{xy})

Pearson’s correlation coefficient

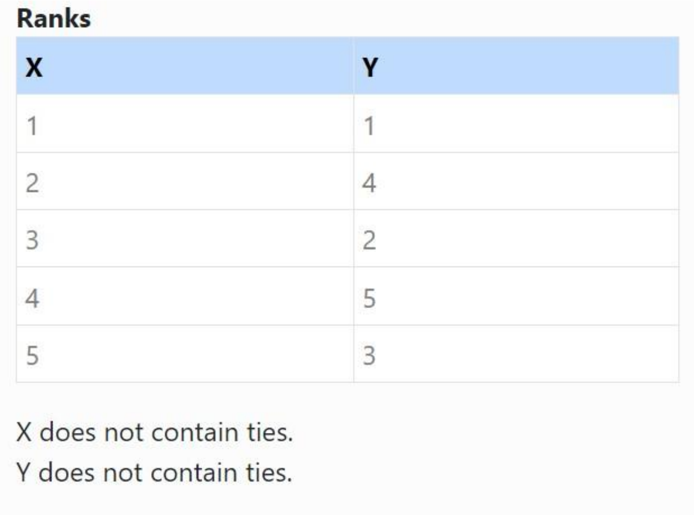


$$S_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$S_{XY} = \frac{60}{5 - 1} = 15$$

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2)}}$$

$$r = \frac{60}{\sqrt{(40*683.2)}} = \mathbf{0.363}$$

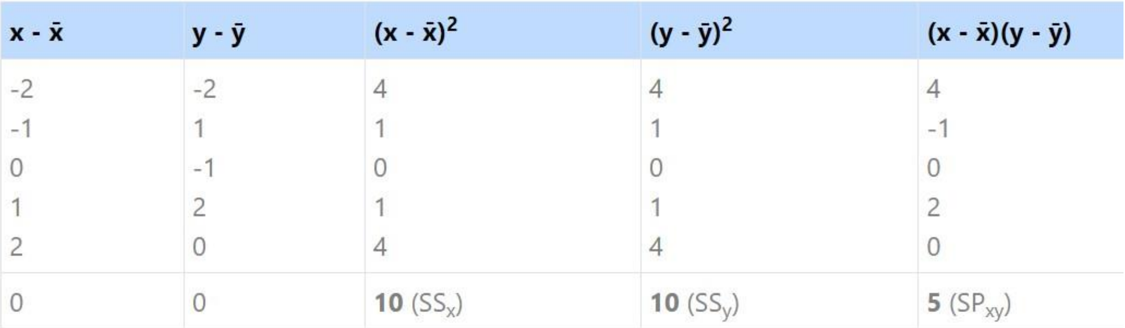


Ranks

X	Y
1	1
2	4
3	2
4	5
5	3

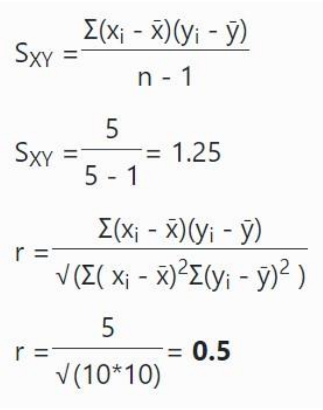
X does not contain ties.

Y does not contain ties.



x - x̄	y - ȳ	$(x - \bar{x})^2$	$(y - \bar{y})^2$	(x - x̄)(y - ȳ)
-2 -1 0 1 2	-2 1 -1 2 0	4 1 0 1 4	4 1 0 1 4	4 -1 0 2 0
0	0	**10** (SS_x)	**10** (SS_y)	**5** (SP_{xy})

Spearman’s rank correlation coefficient



$$S_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$S_{XY} = \frac{5}{5 - 1} = 1.25$$

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2)}}$$

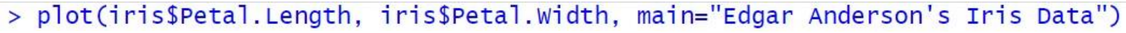
$$r = \frac{5}{\sqrt{(10*10)}} = \mathbf{0.5}$$

You can also perform **chi-squared tests** for multivariate graphical exploratory data analysis (EDA). For further reading on chi-squared tests in R, you may refer to the article Chi-Squared Test | R-bloggers.

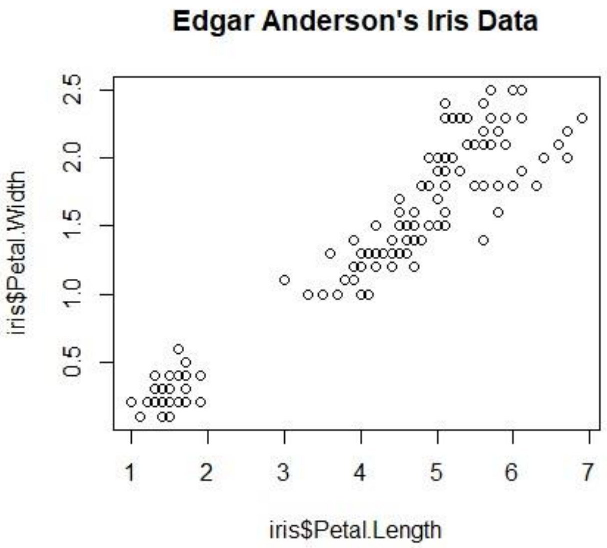
Multi-variate graphical EDA

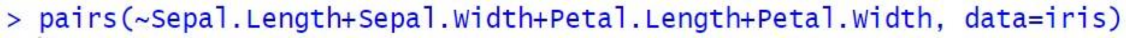
Side-by-side boxplots are the best graphical EDA technique for examining the relationship between a categorical variable and a quantitative variable, as well as the distribution of the quantitative variable at each level of the categorical variable. **Stripcharts** can also be used. Please refer to experiment 2 for the code and examples.

For two quantitative variables, the basic graphical EDA technique is the **scatterplot** which has one variable on the x-axis, one on the y-axis and a point for each case in your dataset. If one variable is explanatory and the other is outcome, it is a convention to put the outcome on the y (vertical) axis.
Scatter plots can be extended to *n* attributes, resulting in a ***scatter-plot matrix***.

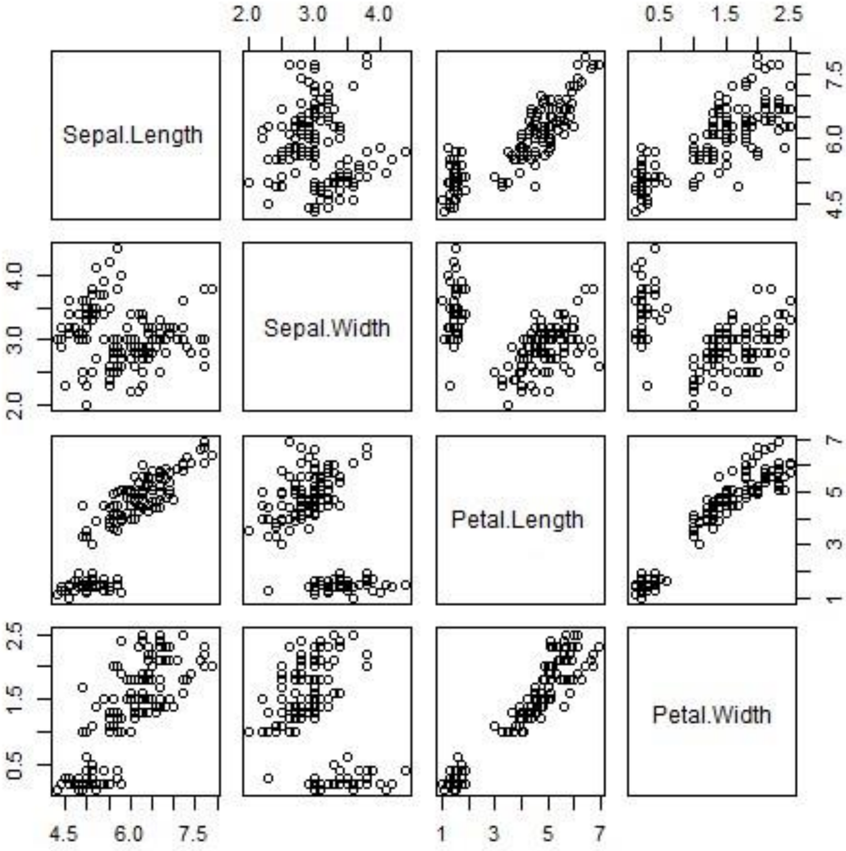


```
> plot(iris$Petal.Length, iris$Petal.Width, main="Edgar Anderson's Iris Data")
```





```
> pairs(~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width, data=iris)
```


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
2.0
3.0
4.0
0.5
1.5
2.5
7.5
6.0
4.5
4.0
3.0
2.0
7
5
3
1
2.5
1.5
0.5
4.5
6.0
7.5
1
3
5
7

Procedure for Implementation in lab :

1. Identify a Dataset for Exploratory Data Analysis
2. Handle the missing values appropriately 3. Detect and remove outliers 4. Perform the following:
    a. univariate non-graphical EDA
    b. univariate graphical EDA
    c. multivariate non-graphical EDA
    d. multivariate graphical EDA

5. What is your understanding of the data after implementing steps of EDA identified above?

Understanding of the Data after Implementing EDA Steps:

- After performing the steps above, you should have insights into the distribution of individual variables, relationships between variables, and potential patterns or trends in the data.
- You will have a clearer understanding of how missing values and outliers impact the dataset.
- EDA helps in making informed decisions about feature engineering, model selection, and any necessary data preprocessing steps before building predictive models.

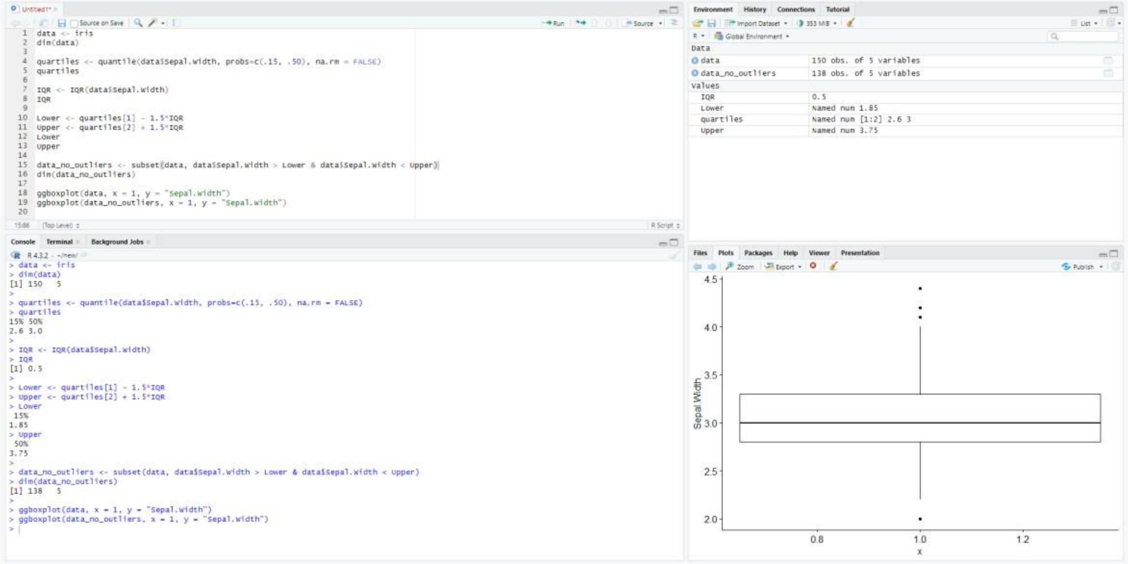
Students should add their R code and screenshots of output. Also students should provide the following details of the dataset:

Data set used: iris

Title: iris Source: R
Studio Number of
instances:

Number of attributes: Attribute
information **:**

1]

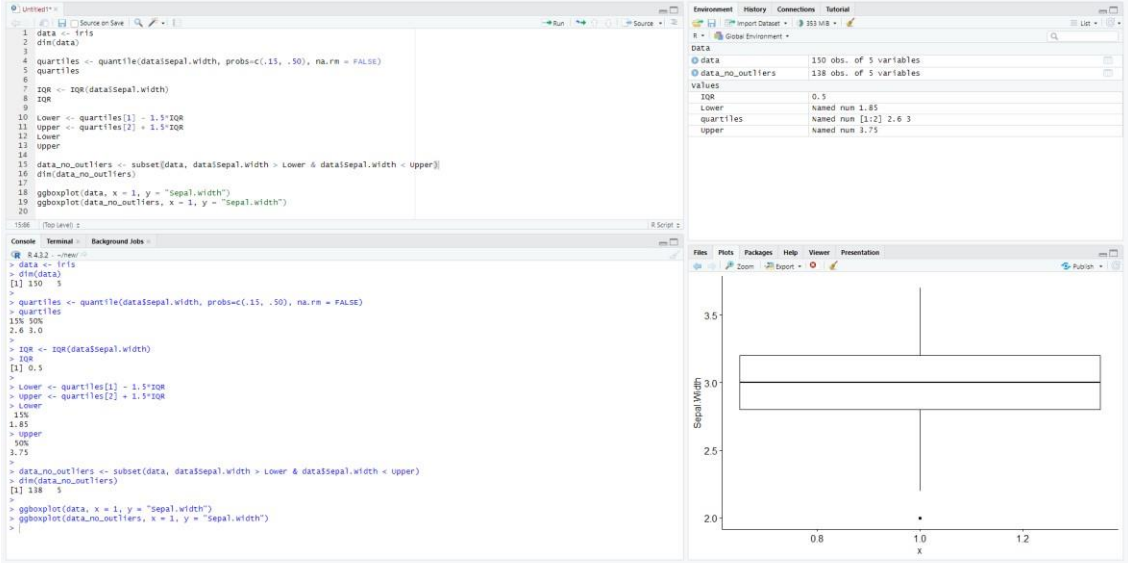


```
data <- iris
dim(data)

quartiles <- quantile(data$Sepal.width, probs=c(.15, .50), na.rm = FALSE)
quartiles

IQR <- IQR(data$Sepal.width)
IQR

Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR
Lower
Upper

data_no_outliers <- subset(data, data$Sepal.width > Lower & data$Sepal.width < Upper)
dim(data_no_outliers)

ggboxplot(data, x = 1, y = "Sepal.width")
ggboxplot(data_no_outliers, x = 1, y = "Sepal.width")
```

```
> data <- iris
> dim(data)
[1] 150   5
>
> quartiles <- quantile(data$Sepal.width, probs=c(.15, .50), na.rm = FALSE)
> quartiles
15% 50%
2.6 3.0
>
> IQR <- IQR(data$Sepal.width)
> IQR
[1] 0.5
>
> Lower <- quartiles[1] - 1.5*IQR
> Upper <- quartiles[2] + 1.5*IQR
> Lower
 15%
1.85
> Upper
 50%
3.75
>
> data_no_outliers <- subset(data, data$Sepal.width > Lower & data$Sepal.width < Upper)
> dim(data_no_outliers)
[1] 138   5
>
> ggboxplot(data, x = 1, y = "Sepal.width")
> ggboxplot(data_no_outliers, x = 1, y = "Sepal.width")
>
```

```
Data
data                150 obs. of 5 variables
data_no_outliers    138 obs. of 5 variables
Values
  IQR               0.5
  Lower             Named num 1.85
  quartiles         Named num [1:2] 2.6 3
  Upper             Named num 3.75
```

2]

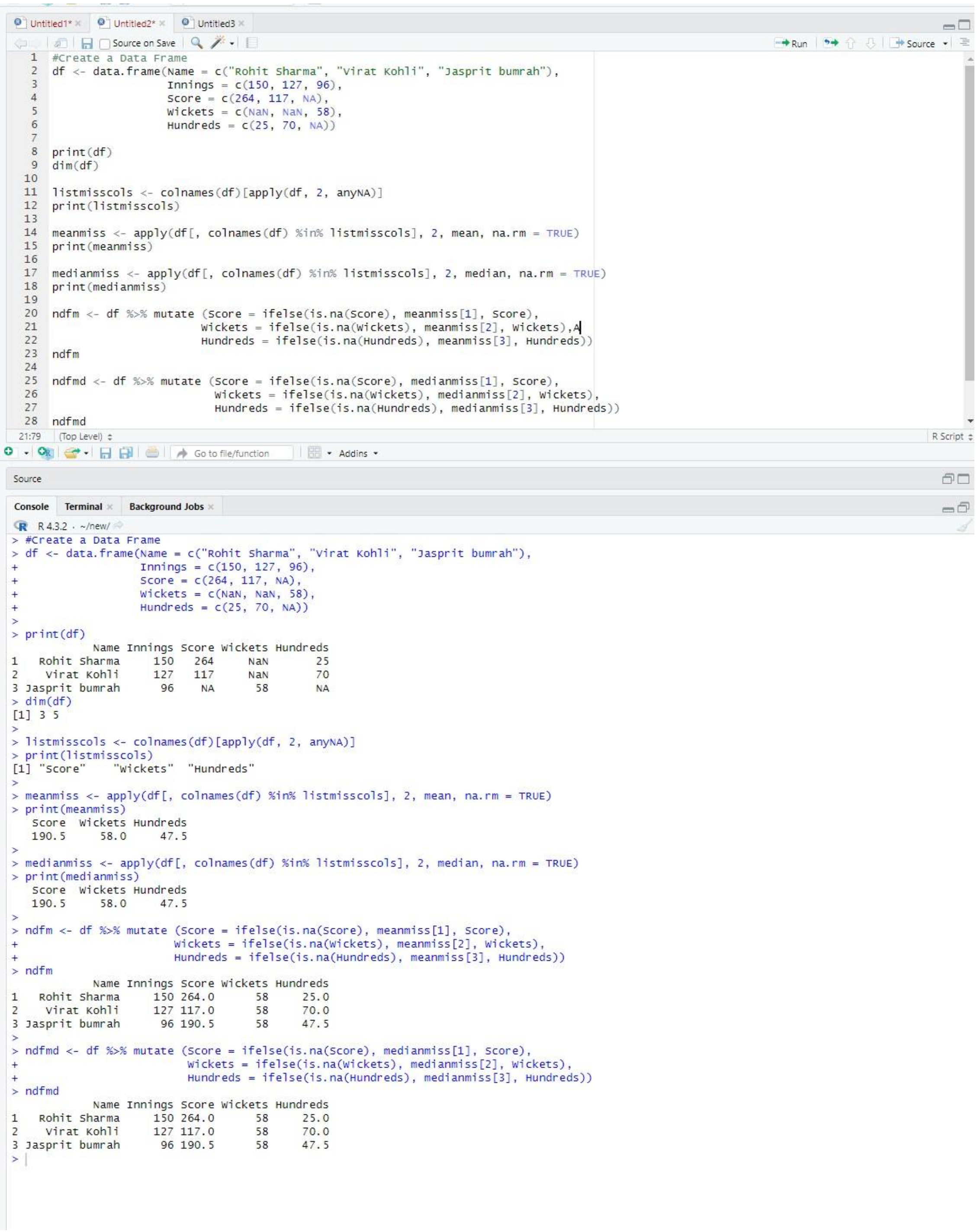


```
data <- iris
dim(data)

quartiles <- quantile(data$Sepal.width, probs=c(.15, .50), na.rm = FALSE)
quartiles

IQR <- IQR(data$Sepal.width)
IQR

Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR
Lower
Upper

data_no_outliers <- subset(data, data$Sepal.width > Lower & data$Sepal.width < Upper)
dim(data_no_outliers)

ggboxplot(data, x = 1, y = "Sepal.width")
ggboxplot(data_no_outliers, x = 1, y = "Sepal.width")
```

```
> data <- iris
> dim(data)
[1] 150   5
>
> quartiles <- quantile(data$Sepal.width, probs=c(.15, .50), na.rm = FALSE)
> quartiles
15% 50%
2.6 3.0
>
> IQR <- IQR(data$Sepal.width)
> IQR
[1] 0.5
>
> Lower <- quartiles[1] - 1.5*IQR
> Upper <- quartiles[2] + 1.5*IQR
> Lower
 15%
1.85
> Upper
 50%
3.75
>
> data_no_outliers <- subset(data, data$Sepal.width > Lower & data$Sepal.width < Upper)
> dim(data_no_outliers)
[1] 138   5
>
> ggboxplot(data, x = 1, y = "Sepal.width")
> ggboxplot(data_no_outliers, x = 1, y = "Sepal.width")
>
```

```
Data
data                150 obs. of 5 variables
data_no_outliers    138 obs. of 5 variables
Values
  IQR               0.5
  Lower             Named num 1.85
  quartiles         Named num [1:2] 2.6 3
  Upper             Named num 3.75
```

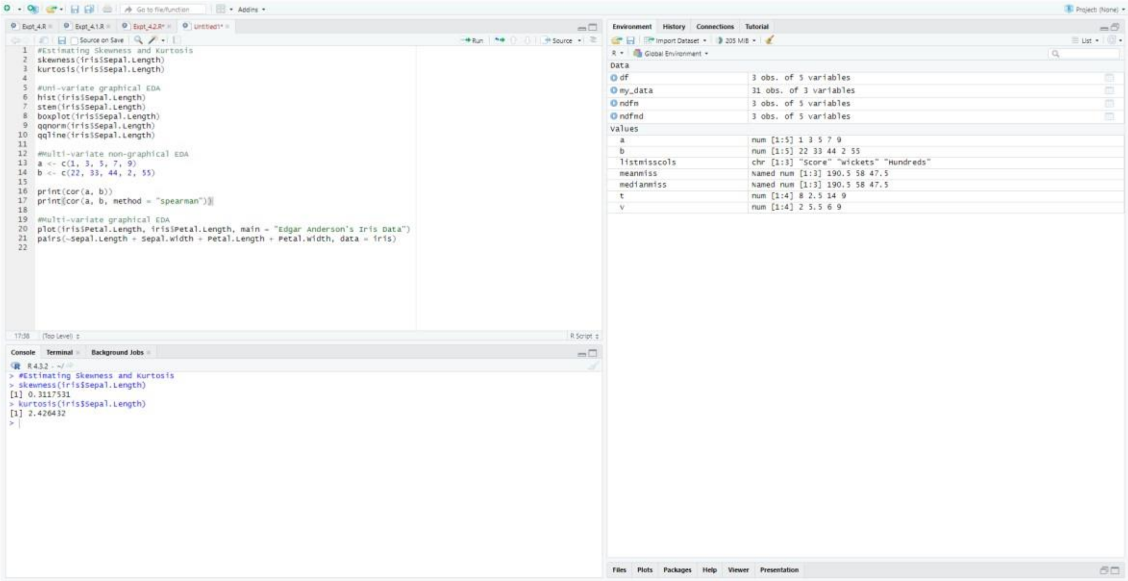
3]



```r
#Create a Data Frame
df <- data.frame(Name = c("Rohit Sharma", "Virat Kohli", "Jasprit bumrah"),
                 Innings = c(150, 127, 96),
                 Score = c(264, 117, NA),
                 Wickets = c(NaN, NaN, 58),
                 Hundreds = c(25, 70, NA))

print(df)
dim(df)

listmisscols <- colnames(df)[apply(df, 2, anyNA)]
print(listmisscols)

meanmiss <- apply(df[, colnames(df) %in% listmisscols], 2, mean, na.rm = TRUE)
print(meanmiss)

medianmiss <- apply(df[, colnames(df) %in% listmisscols], 2, median, na.rm = TRUE)
print(medianmiss)

ndfm <- df %>% mutate (Score = ifelse(is.na(Score), meanmiss[1], Score),
                      Wickets = ifelse(is.na(Wickets), meanmiss[2], Wickets),A
                      Hundreds = ifelse(is.na(Hundreds), meanmiss[3], Hundreds))
ndfm

ndfmd <- df %>% mutate (Score = ifelse(is.na(Score), medianmiss[1], Score),
                        Wickets = ifelse(is.na(Wickets), medianmiss[2], Wickets),
                        Hundreds = ifelse(is.na(Hundreds), medianmiss[3], Hundreds))
ndfmd
```

```
> #Create a Data Frame
> df <- data.frame(Name = c("Rohit Sharma", "Virat Kohli", "Jasprit bumrah"),
+                  Innings = c(150, 127, 96),
+                  Score = c(264, 117, NA),
+                  Wickets = c(NaN, NaN, 58),
+                  Hundreds = c(25, 70, NA))
> 
> print(df)
            Name Innings Score Wickets Hundreds
1   Rohit Sharma     150   264     NaN       25
2    Virat Kohli     127   117     NaN       70
3 Jasprit bumrah      96    NA      58       NA
> dim(df)
[1] 3 5
> 
> listmisscols <- colnames(df)[apply(df, 2, anyNA)]
> print(listmisscols)
[1] "Score"    "Wickets"  "Hundreds"
> 
> meanmiss <- apply(df[, colnames(df) %in% listmisscols], 2, mean, na.rm = TRUE)
> print(meanmiss)
   Score  Wickets Hundreds 
   190.5     58.0     47.5 
> 
> medianmiss <- apply(df[, colnames(df) %in% listmisscols], 2, median, na.rm = TRUE)
> print(medianmiss)
   Score  Wickets Hundreds 
   190.5     58.0     47.5 
> 
> ndfm <- df %>% mutate (Score = ifelse(is.na(Score), meanmiss[1], Score),
+                       Wickets = ifelse(is.na(Wickets), meanmiss[2], Wickets),
+                       Hundreds = ifelse(is.na(Hundreds), meanmiss[3], Hundreds))
> ndfm
            Name Innings Score Wickets Hundreds
1   Rohit Sharma     150 264.0      58     25.0
2    Virat Kohli     127 117.0      58     70.0
3 Jasprit bumrah      96 190.5      58     47.5
> 
> ndfmd <- df %>% mutate (Score = ifelse(is.na(Score), medianmiss[1], Score),
+                         Wickets = ifelse(is.na(Wickets), medianmiss[2], Wickets),
+                         Hundreds = ifelse(is.na(Hundreds), medianmiss[3], Hundreds))
> ndfmd
            Name Innings Score Wickets Hundreds
1   Rohit Sharma     150 264.0      58     25.0
2    Virat Kohli     127 117.0      58     70.0
3 Jasprit bumrah      96 190.5      58     47.5
> 
```

4]



```r
#Estimating Skewness and Kurtosis
skewness(iris$Sepal.Length)
kurtosis(iris$Sepal.Length)

#Uni-variate graphical EDA
hist(iris$Sepal.Length)
stem(iris$Sepal.Length)
boxplot(iris$Sepal.Length)
qqnorm(iris$Sepal.Length)
qqline(iris$Sepal.Length)

#Multi-variate non-graphical EDA
a <- c(1, 3, 5, 7, 9)
b <- c(22, 33, 44, 2, 55)

print(cor(a, b))
print(cor(a, b, method = "spearman"))

#Multi-variate graphical EDA
plot(iris$Petal.Length, iris$Petal.Length, main = "Edgar Anderson's Iris Data")
pairs(~Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, data = iris)
```

```
R 4.3.2 · ~/
> #Estimating Skewness and Kurtosis
> skewness(iris$Sepal.Length)
[1] 0.3117531
> kurtosis(iris$Sepal.Length)
[1] 2.426432
> 
```

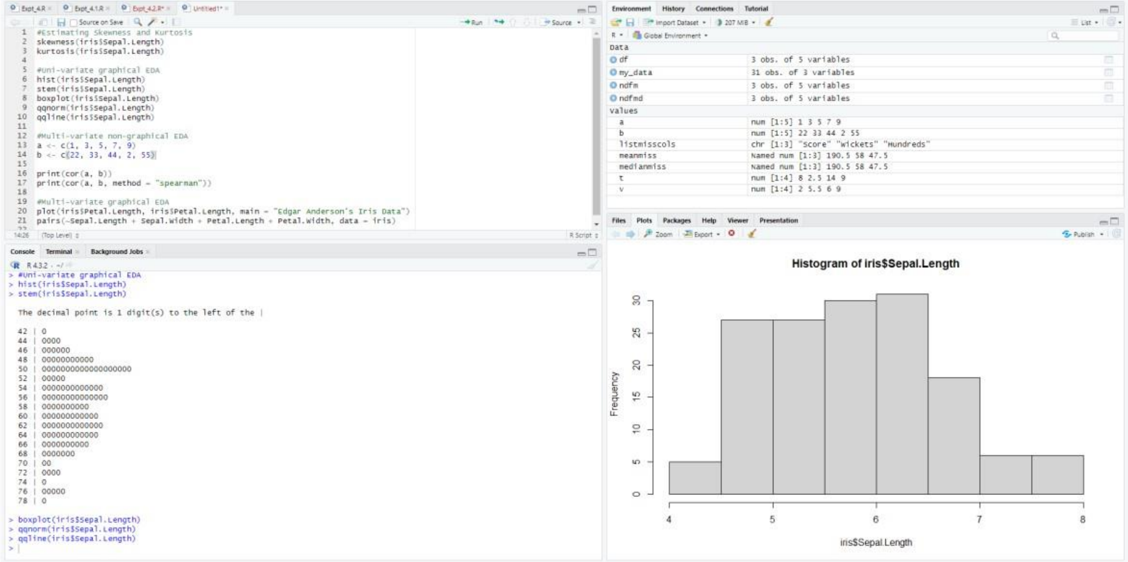
Data

df	3 obs. of 5 variables
my_data	31 obs. of 3 variables
ndfm	3 obs. of 5 variables
ndfmd	3 obs. of 5 variables

Values

a	num [1:5] 1 3 5 7 9
b	num [1:5] 22 33 44 2 55
listmisscols	chr [1:3] "Score" "Wickets" "Hundreds"
meanmiss	Named num [1:3] 190.5 58 47.5
medianmiss	Named num [1:3] 190.5 58 47.5
t	num [1:4] 8 2.5 14 9
v	num [1:4] 2 5.5 6 9

5]



```r
#Estimating Skewness and Kurtosis
skewness(iris$Sepal.Length)
kurtosis(iris$Sepal.Length)

#Uni-variate graphical EDA
hist(iris$Sepal.Length)
stem(iris$Sepal.Length)
boxplot(iris$Sepal.Length)
qqnorm(iris$Sepal.Length)
qqline(iris$Sepal.Length)

#Multi-variate non-graphical EDA
a <- c(1, 3, 5, 7, 9)
b <- c(22, 33, 44, 2, 55)

print(cor(a, b))
print(cor(a, b, method = "spearman"))

#Multi-variate graphical EDA
plot(iris$Petal.Length, iris$Petal.Length, main = "Edgar Anderson's Iris Data")
pairs(~Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, data = iris)
```

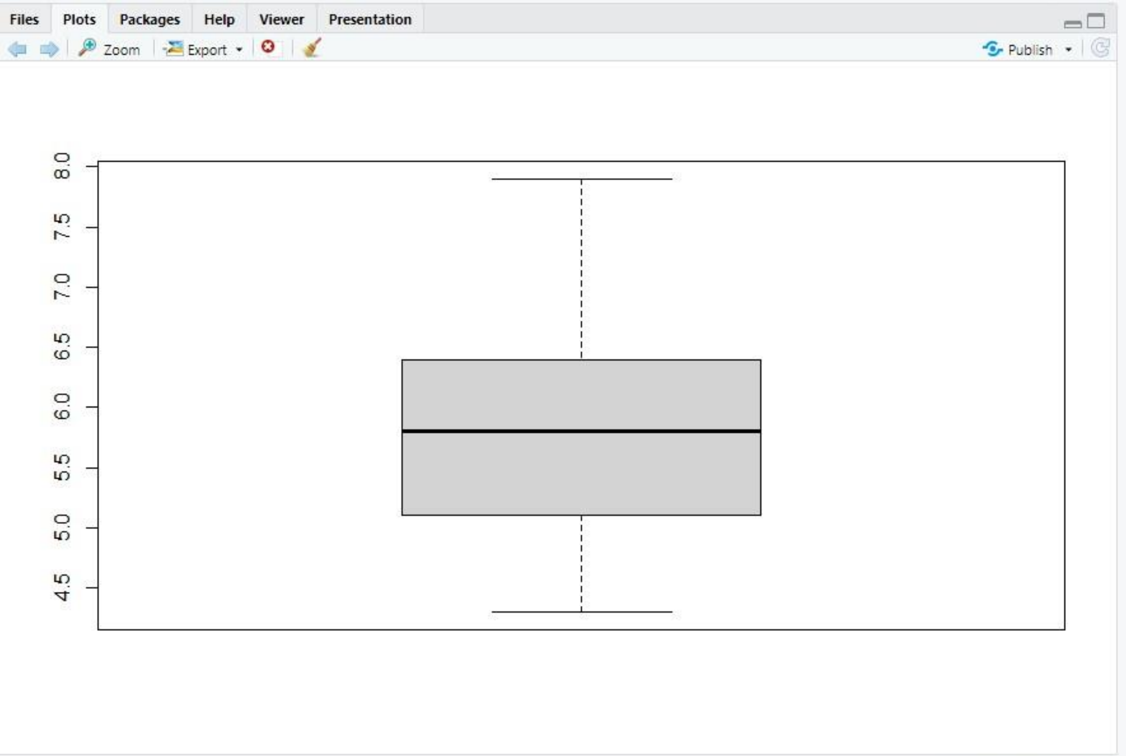
```
R 4.3.2 · ~/
> #Uni-variate graphical EDA
> hist(iris$Sepal.Length)
> stem(iris$Sepal.Length)

  The decimal point is 1 digit(s) to the left of the |

  42 | 0
  44 | 0000
  46 | 000000
  48 | 00000000000
  50 | 0000000000000000000
  52 | 00000
  54 | 0000000000000
  56 | 00000000000000
  58 | 0000000000
  60 | 000000000000
  62 | 0000000000000
  64 | 000000000000
  66 | 0000000000
  68 | 0000000
  70 | 00
  72 | 0000
  74 | 0
  76 | 00000
  78 | 0

> boxplot(iris$Sepal.Length)
> qqnorm(iris$Sepal.Length)
> qqline(iris$Sepal.Length)
> 
```

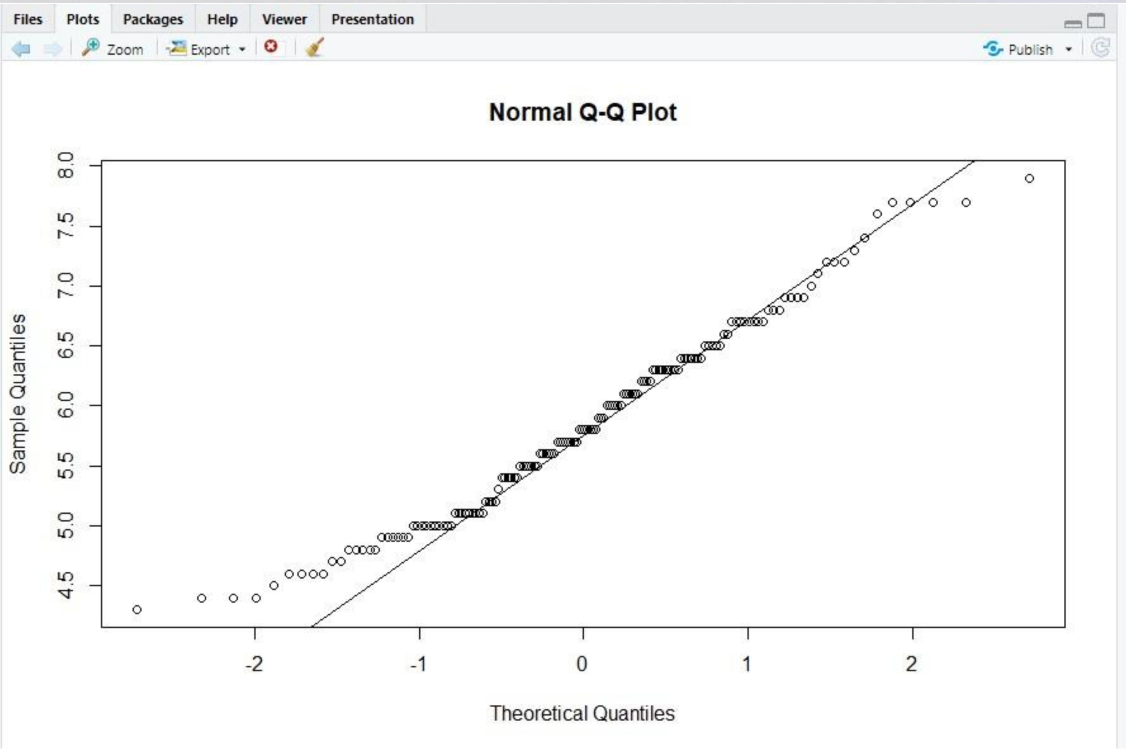
Data

df	3 obs. of 5 variables
my_data	31 obs. of 3 variables
ndfm	3 obs. of 5 variables
ndfmd	3 obs. of 5 variables

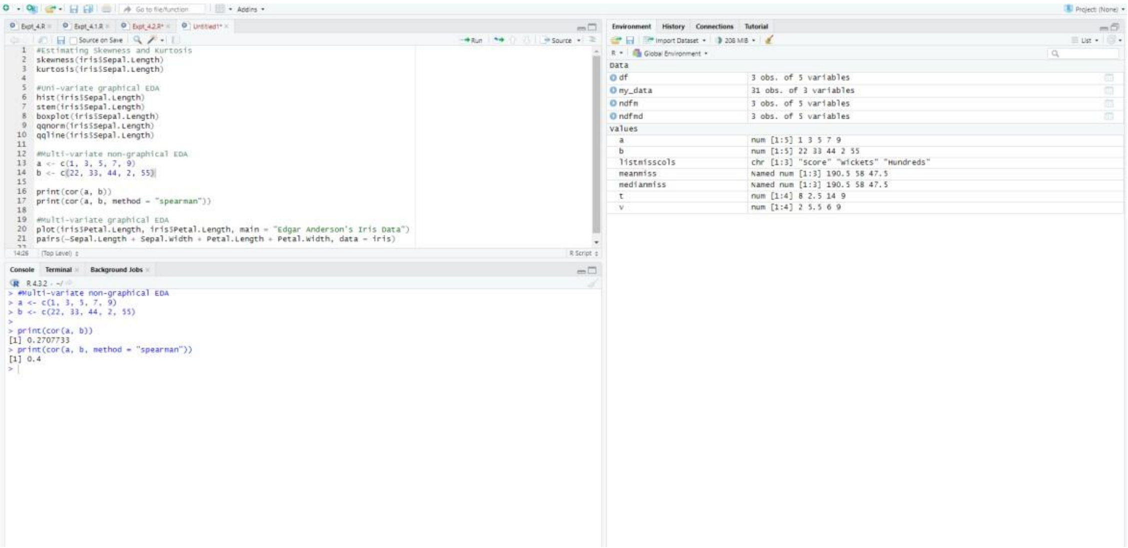
Values

a	num [1:5] 1 3 5 7 9
b	num [1:5] 22 33 44 2 55
listmisscols	chr [1:3] "Score" "Wickets" "Hundreds"
meanmiss	Named num [1:3] 190.5 58 47.5
medianmiss	Named num [1:3] 190.5 58 47.5
t	num [1:4] 8 2.5 14 9
v	num [1:4] 2 5.5 6 9


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Normal Q-Q Plot
Sample Quantiles
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
-2
-1
0
1
2
Theoretical Quantiles

6]

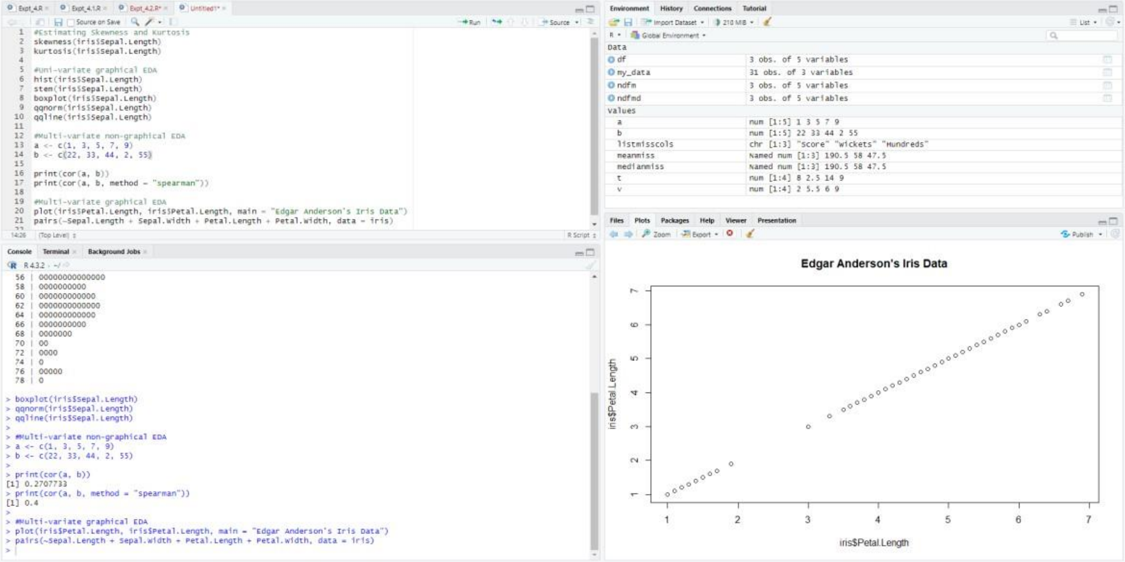


```
#Estimating Skewness and Kurtosis
skewness(iris$Sepal.Length)
kurtosis(iris$Sepal.Length)

#Uni-variate graphical EDA
hist(iris$Sepal.Length)
stem(iris$Sepal.Length)
boxplot(iris$Sepal.Length)
qqnorm(iris$Sepal.Length)
qqline(iris$Sepal.Length)

#Multi-variate non-graphical EDA
a <- c(1, 3, 5, 7, 9)
b <- c(22, 33, 44, 2, 55)

print(cor(a, b))
print(cor(a, b, method = "spearman"))

#Multi-variate graphical EDA
plot(iris$Petal.Length, iris$Petal.Length, main = "Edgar Anderson's Iris Data")
pairs(~Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, data = iris)
```

```
> #Multi-variate non-graphical EDA
> a <- c(1, 3, 5, 7, 9)
> b <- c(22, 33, 44, 2, 55)
>
> print(cor(a, b))
[1] 0.2707733
> print(cor(a, b, method = "spearman"))
[1] 0.4
>
```

```
Data
df            3 obs. of 5 variables
my_data       31 obs. of 3 variables
ndfm          3 obs. of 5 variables
ndfmd         3 obs. of 5 variables
Values
a             num [1:5] 1 3 5 7 9
b             num [1:5] 22 33 44 2 55
listmisscols  chr [1:3] "Score" "Wickets" "Hundreds"
meanmiss      Named num [1:3] 190.5 58 47.5
medianmiss    Named num [1:3] 190.5 58 47.5
t             num [1:4] 8 2.5 14 9
v             num [1:4] 2 5.5 6 9
```

7]

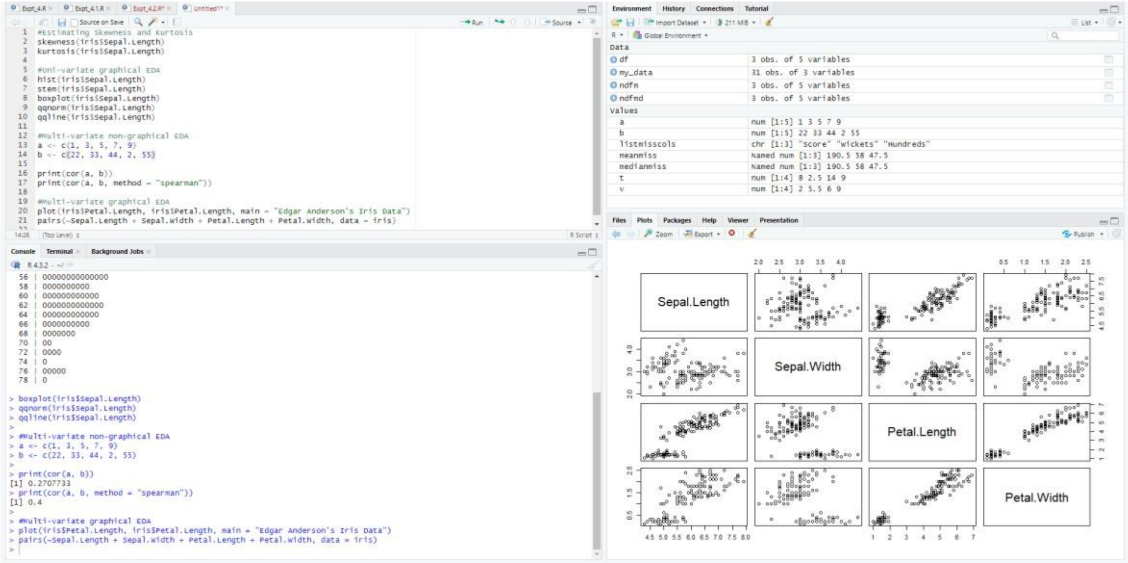


```
#Estimating Skewness and Kurtosis
skewness(iris$Sepal.Length)
kurtosis(iris$Sepal.Length)

#Uni-variate graphical EDA
hist(iris$Sepal.Length)
stem(iris$Sepal.Length)
boxplot(iris$Sepal.Length)
qqnorm(iris$Sepal.Length)
qqline(iris$Sepal.Length)

#Multi-variate non-graphical EDA
a <- c(1, 3, 5, 7, 9)
b <- c(22, 33, 44, 2, 55)

print(cor(a, b))
print(cor(a, b, method = "spearman"))

#Multi-variate graphical EDA
plot(iris$Petal.Length, iris$Petal.Length, main = "Edgar Anderson's Iris Data")
pairs(~Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, data = iris)
```

```
  56 | 00000000000000
  58 | 0000000000
  60 | 000000000000
  62 | 0000000000000
  64 | 000000000000
  66 | 0000000000
  68 | 0000000
  70 | 00
  72 | 0000
  74 | 0
  76 | 00000
  78 | 0

> boxplot(iris$Sepal.Length)
> qqnorm(iris$Sepal.Length)
> qqline(iris$Sepal.Length)
>
> #Multi-variate non-graphical EDA
> a <- c(1, 3, 5, 7, 9)
> b <- c(22, 33, 44, 2, 55)
>
> print(cor(a, b))
[1] 0.2707733
> print(cor(a, b, method = "spearman"))
[1] 0.4
>
> #Multi-variate graphical EDA
> plot(iris$Petal.Length, iris$Petal.Length, main = "Edgar Anderson's Iris Data")
> pairs(~Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, data = iris)
>
```

```
Data
df            3 obs. of 5 variables
my_data       31 obs. of 3 variables
ndfm          3 obs. of 5 variables
ndfmd         3 obs. of 5 variables
Values
a             num [1:5] 1 3 5 7 9
b             num [1:5] 22 33 44 2 55
listmisscols  chr [1:3] "Score" "Wickets" "Hundreds"
meanmiss      Named num [1:3] 190.5 58 47.5
medianmiss    Named num [1:3] 190.5 58 47.5
t             num [1:4] 8 2.5 14 9
v             num [1:4] 2 5.5 6 9
```

8]



```
#Estimating Skewness and Kurtosis
skewness(iris$Sepal.Length)
kurtosis(iris$Sepal.Length)

#Uni-variate graphical EDA
hist(iris$Sepal.Length)
stem(iris$Sepal.Length)
boxplot(iris$Sepal.Length)
qqnorm(iris$Sepal.Length)
qqline(iris$Sepal.Length)

#Multi-variate non-graphical EDA
a <- c(1, 3, 5, 7, 9)
b <- c(22, 33, 44, 2, 55)

print(cor(a, b))
print(cor(a, b, method = "spearman"))

#Multi-variate graphical EDA
plot(iris$Petal.Length, iris$Petal.Length, main = "Edgar Anderson's Iris Data")
pairs(~Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, data = iris)
```

```
R 4.3.2 · ~/
  56 | 00000000000000
  58 | 0000000000
  60 | 000000000000
  62 | 0000000000000
  64 | 000000000000
  66 | 0000000000
  68 | 0000000
  70 | 00
  72 | 0000
  74 | 0
  76 | 00000
  78 | 0

> boxplot(iris$Sepal.Length)
> qqnorm(iris$Sepal.Length)
> qqline(iris$Sepal.Length)
>
> #Multi-variate non-graphical EDA
> a <- c(1, 3, 5, 7, 9)
> b <- c(22, 33, 44, 2, 55)
>
> print(cor(a, b))
[1] 0.2707733
> print(cor(a, b, method = "spearman"))
[1] 0.4
>
> #Multi-variate graphical EDA
> plot(iris$Petal.Length, iris$Petal.Length, main = "Edgar Anderson's Iris Data")
> pairs(~Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, data = iris)
> |
```

Data	
df	3 obs. of 5 variables
my_data	31 obs. of 3 variables
ndfm	3 obs. of 5 variables
ndfmd	3 obs. of 5 variables
Values	
a	num [1:5] 1 3 5 7 9
b	num [1:5] 22 33 44 2 55
listmisscols	chr [1:3] "score" "wickets" "Hundreds"
meanmiss	Named num [1:3] 190.5 58 47.5
medianmiss	Named num [1:3] 190.5 58 47.5
t	num [1:4] 8 2.5 14 9
v	num [1:4] 2 5.5 6 9

Post lab Questions:

1. What is an appropriate way to visualize a list of the eye colors of 120 people?
    I. Boxplot
    II. Pie-chart
    III. Histogram
    IV. Scatterplot

    Boxplot
2. You want to investigate whether households in California tend to have a higher income than households in Massachusetts. Which summary measure would you use to compare the two states?
    I. median household income
    II. mean household income III. 3rd quartile of household income
    IV. IQR

Median household income

3. Suppose all household incomes in California increase by 5%. How does that change the median household income?
    I. median household income goes up by 5%
    II. the median household income doesn't change
    III. cannot be determined from the information given

Median household income doesn't change (II)

4. Suppose all household incomes in California increase by $5,000. How does that change the interquartile range of the household incomes?
    I. cannot be determined from the information given
    II. the interquartile range of the household incomes doesn't change
    III. the interquartile range of the household incomes goes up by $5,000

II.The interquartile range of the household incomes doesn't change.

5. The median sales price for houses in a certain county during the last year was $342,000. What can we say about the percentage of sales represented by the houses that sold for more than $342,000?
   I. the houses that sold for more than $342,000 represent more than 50% of all sales
   II. the houses that sold for more than $342,000 represent exactly 50% of all sales
   III. the houses that sold for more than $342,000 represent less than 50% of all sales

The houses that sold for more than $342,000 represent less than 50% of all sales.

6. Suppose all household incomes in California increase by $5,000. How does that change the standard deviation of the household incomes? I. the standard deviation of the household incomes doesn't change
   II. cannot be determined from the information given
   III. the standard deviation of the household incomes goes up by $5,000

The standard deviation of the household incomes goes up by $5,000.

7. Which of the following graphical displays can be used to understand the distribution of data? I. Box Plot
   II. Quantile-Normal plot
   III. Histogram
   IV. Scatter Plot

Histogram

8. Correlation between two variables X&Y is 0.85. Now, after adding the value 2 to all the values of X, the correlation coefficient will be
   a. 0.85
   b. 0.87
   c. 0.65
   d. 0.82

a. 0.85