AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Faculty of Science and Technology

MID TERM PROJECT REPORT

Assignment Title:	Modified Titanic Dataset Project Report			
Assignment No:	1		Date of Submission:	18 July 2023
Course Title:	Introduction to Data Science			
Course Code:	01153		Section:	C
Semester:	Summer	2022-23	Course Teacher:	Abdus Salam

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*
** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.:

No	Name	ID	Program	Signature
1	FAISAL, MD. OMAR FARUK	20-43669-2	BSc [CSE]	
2			Choose an item.	
3			Choose an item.	
4			Choose an item.	
5			Choose an item.	
6			Choose an item.	
7			Choose an item.	
8			Choose an item.	
9			Choose an item.	
10			Choose an item.	

Faculty use only

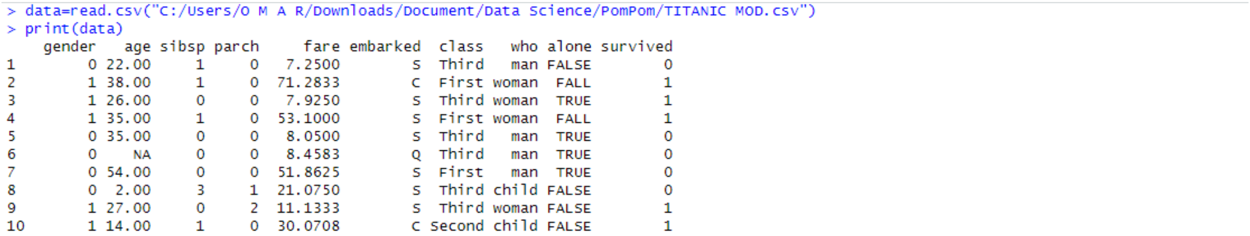
FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

Overview: Real-world data is in most cases incomplete, noisy, and inconsistent. The probability of gathering anomalous or incorrect data is relatively high because nowadays, data generation is growing rapidly and an increasing number of heterogeneous data sources. So, it's most important to process data for the best possible quality. Transforming raw data into a useful, understandable format by data preprocessing. To perform data analysis on this dataset need to preprocess data because the Project dataset contains noisy data, missing values, and errors or outliers. To prepare a cleaned dataset to need to perform the following tasks of data pre-processing using R language:

1. Data cleaning : a. Smooth Noisy Data b. Handling Missing Data c. Data Wrangling or Munging 2. Data Integration 3. Data Transformation 4. Data Reduction 5. Data Discretization

Tool Used: RStudio, MS Excel

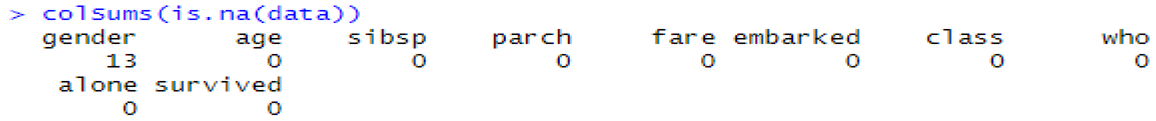
Insertion of Datasheet: TITANIC MOD.csv



```
> data=read.csv("C:/Users/O M A R/Downloads/Document/Data Science/PomPom/TITANIC MOD.csv")
> print(data)
    gender   age sibsp parch     fare embarked  class   who alone survived
1        0 22.00     1     0   7.2500        S  Third   man FALSE        0
2        1 38.00     1     0  71.2833        C  First woman  FALL        1
3        1 26.00     0     0   7.9250        S  Third woman  TRUE        1
4        1 35.00     1     0  53.1000        S  First woman  FALL        1
5        0 35.00     0     0   8.0500        S  Third   man  TRUE        0
6        0    NA     0     0   8.4583        Q  Third   man  TRUE        0
7        0 54.00     0     0  51.8625        S  First   man  TRUE        0
8        0  2.00     3     1  21.0750        S  Third child FALSE        0
9        1 27.00     0     2  11.1333        S  Third woman FALSE        1
10       1 14.00     1     0  30.0708        C Second child FALSE        1
```

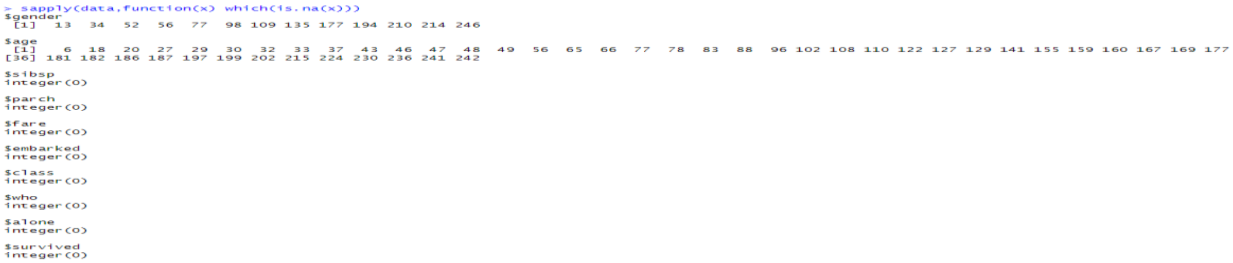
Data Cleaning: Handling Missing Data: This dataset contains missing values in the assault variable. In R programming the missing value will be undefined and with undefiled, any arithmetic operation will produce a NAN. So we have to replace these missing values with the mean values of the respective variables

Counting number of Null values in each column



```
> colSums(is.na(data))
   gender       age     sibsp     parch      fare  embarked     class       who 
       13         0         0         0         0         0         0         0 
    alone  survived 
        0         0 
```

Specific position of Null value

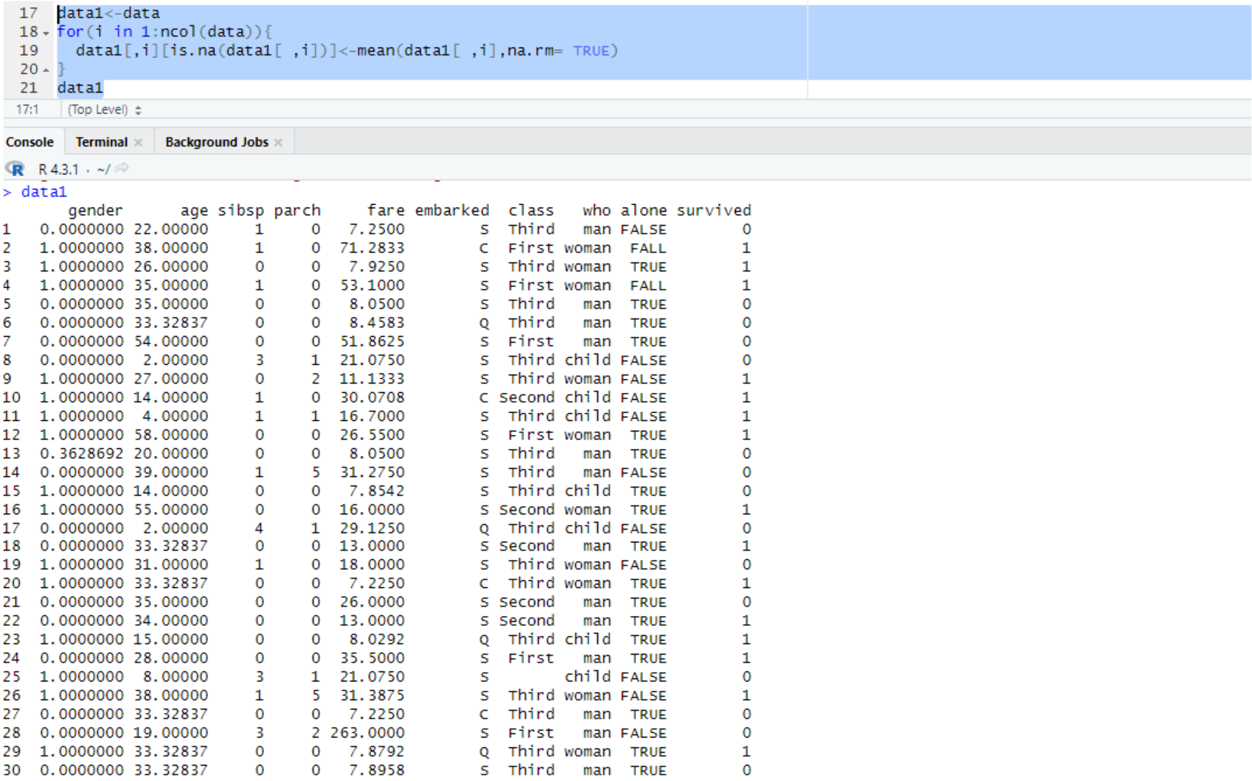


```
> sapply(data,function(x) which(is.na(x)))
$gender
 [1]  13  34  52  56  77  98 109 135 177 194 210 214 246

$age
 [1]   6  18  20  27  29  30  32  33  37  43  46  47  48  49  56  65  66  77  78  83  88  96 102 108 110 122 127 129 141 155 159 160 167 169 177
[36] 181 182 186 187 197 199 202 215 224 230 236 241 242

$sibsp
integer(0)

$parch
integer(0)

$fare
integer(0)

$embarked
integer(0)

$class
integer(0)

$who
integer(0)

$alone
integer(0)

$survived
integer(0)
```

Remove all null value :remove<-na.omit(data)

Then we replcae the missing value value with MEAN value

```
data$age[is.na(data$age)]<-mean(data$age,na.rm= TRUE)
print(data)

data1<-data
for(i in 1:ncol(data)){
  data1[,i][is.na(data1[ ,i])]<-mean(data1[ ,i],na.rm= TRUE)
}
data1
```

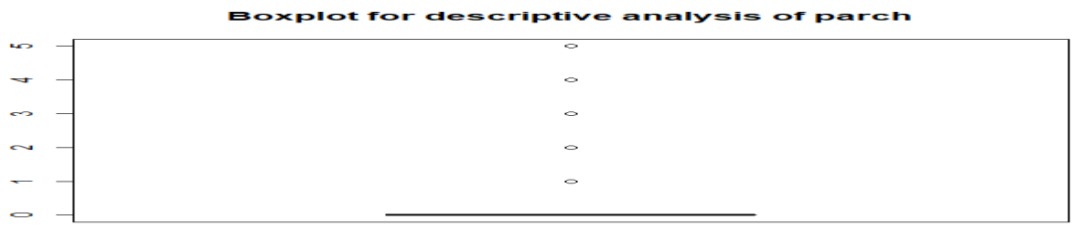


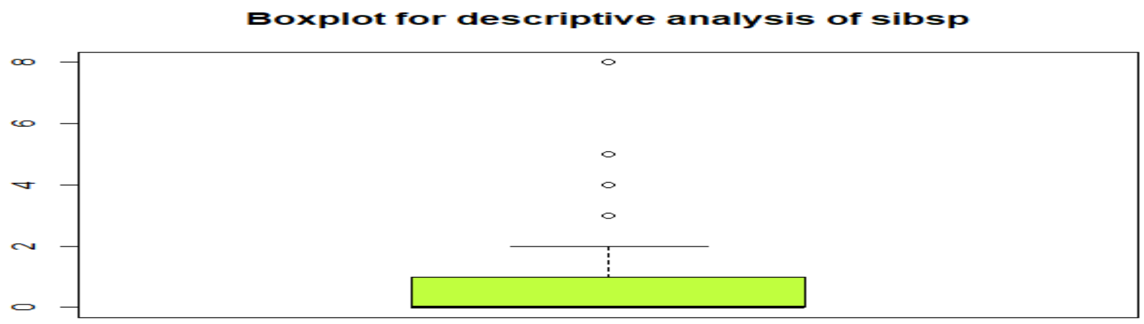
```
data1<-data
for(i in 1:ncol(data)){
  data1[,i][is.na(data1[ ,i])]<-mean(data1[ ,i],na.rm= TRUE)
}
data1
```

```
> data1
      gender      age sibsp parch     fare embarked  class   who alone survived
1  0.0000000 22.00000     1     0   7.2500        S  Third   man FALSE        0
2  1.0000000 38.00000     1     0  71.2833        C  First woman  FALL        1
3  1.0000000 26.00000     0     0   7.9250        S  Third woman  TRUE        1
4  1.0000000 35.00000     1     0  53.1000        S  First woman  FALL        1
5  0.0000000 35.00000     0     0   8.0500        S  Third   man  TRUE        0
6  0.0000000 33.32837     0     0   8.4583        Q  Third   man  TRUE        0
7  0.0000000 54.00000     0     0  51.8625        S  First   man  TRUE        0
8  0.0000000  2.00000     3     1  21.0750        S  Third child FALSE        0
9  1.0000000 27.00000     0     2  11.1333        S  Third woman FALSE        1
10 1.0000000 14.00000     1     0  30.0708        C Second child FALSE        1
11 1.0000000  4.00000     1     1  16.7000        S  Third child FALSE        1
12 1.0000000 58.00000     0     0  26.5500        S  First woman  TRUE        1
13 0.3628692 20.00000     0     0   8.0500        S  Third   man  TRUE        0
14 0.0000000 39.00000     1     5  31.2750        S  Third   man FALSE        0
15 1.0000000 14.00000     0     0   7.8542        S  Third child  TRUE        0
16 1.0000000 55.00000     0     0  16.0000        S Second woman  TRUE        1
17 0.0000000  2.00000     4     1  29.1250        Q  Third child FALSE        0
18 0.0000000 33.32837     0     0  13.0000        S Second   man  TRUE        1
19 1.0000000 31.00000     1     0  18.0000        S  Third woman FALSE        0
20 1.0000000 33.32837     0     0   7.2250        C  Third woman  TRUE        1
21 0.0000000 35.00000     0     0  26.0000        S Second   man  TRUE        0
22 0.0000000 34.00000     0     0  13.0000        S Second   man  TRUE        1
23 1.0000000 15.00000     0     0   8.0292        Q  Third child  TRUE        1
24 0.0000000 28.00000     0     0  35.5000        S  First   man  TRUE        1
25 1.0000000  8.00000     3     1  21.0750        S        child FALSE        0
26 1.0000000 38.00000     1     5  31.3875        S  Third woman FALSE        1
27 0.0000000 33.32837     0     0   7.2250        C  Third   man  TRUE        0
28 0.0000000 19.00000     3     2 263.0000        S  First   man FALSE        0
29 1.0000000 33.32837     0     0   7.8792        Q  Third woman  TRUE        1
30 0.0000000 33.32837     0     0   7.8958        S  Third   man  TRUE        0
```

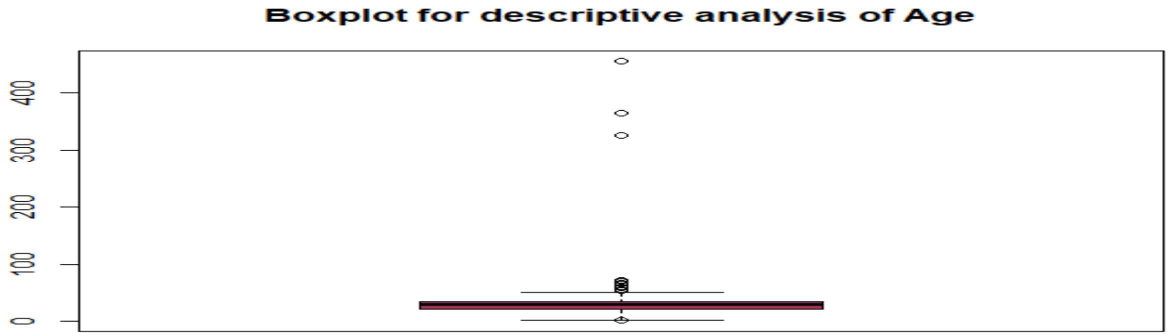
Smooth Noisy Data: Data smoothing refers to a statistical approach of eliminating outliers from datasets to make the patterns more noticeable. By using the Boxplot method, we can detect the outliers.

```
boxplot(data$age,col="maroon",main="Boxplot for descriptive analysis of Age")
boxplot(data$gender,col="blue",main="Boxplot for descriptive analysis of gender")
boxplot(data$sibsp,col="olivedrab1",main="Boxplot for descriptive analysis of sibsp")
boxplot(data$parch,col="red4",main="Boxplot for descriptive analysis of parch")
boxplot(data$fare,col="green",main="Boxplot for descriptive analysis of fare")
boxplot(data$survived,col="skyblue",main="Boxplot for descriptive analysis of survived")
```

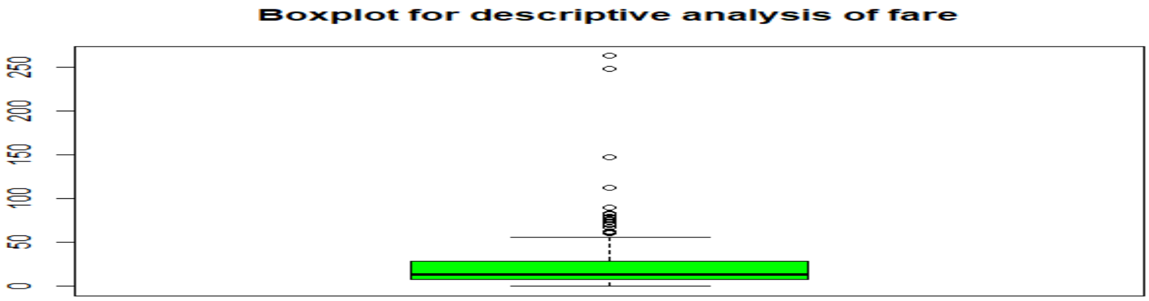




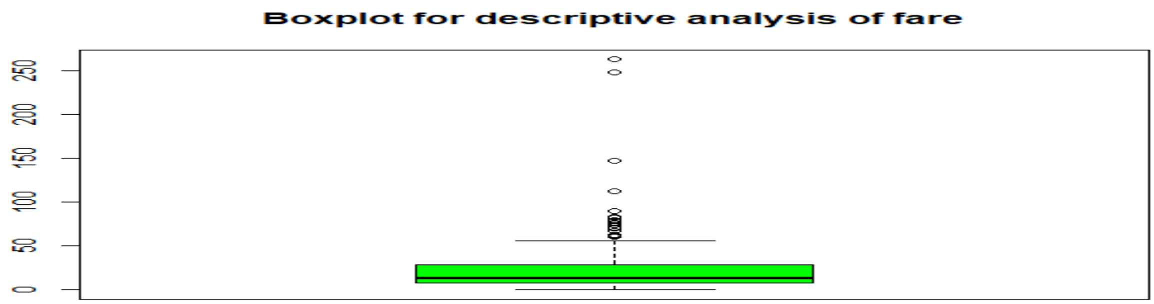
Boxplot for descriptive analysis of sibsp

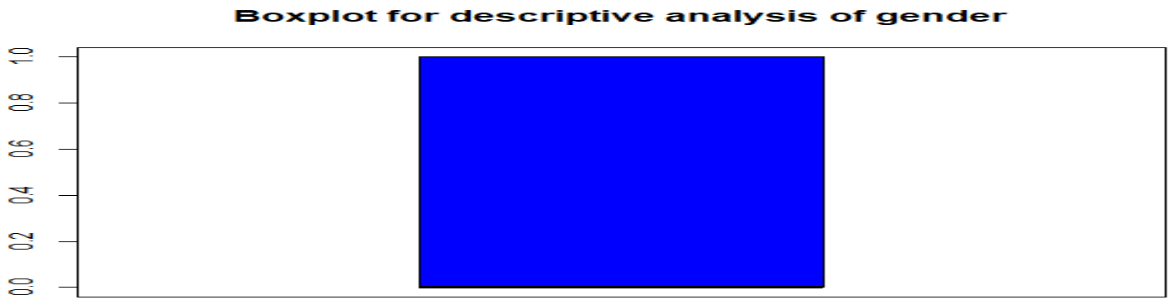


Boxplot for descriptive analysis of Age



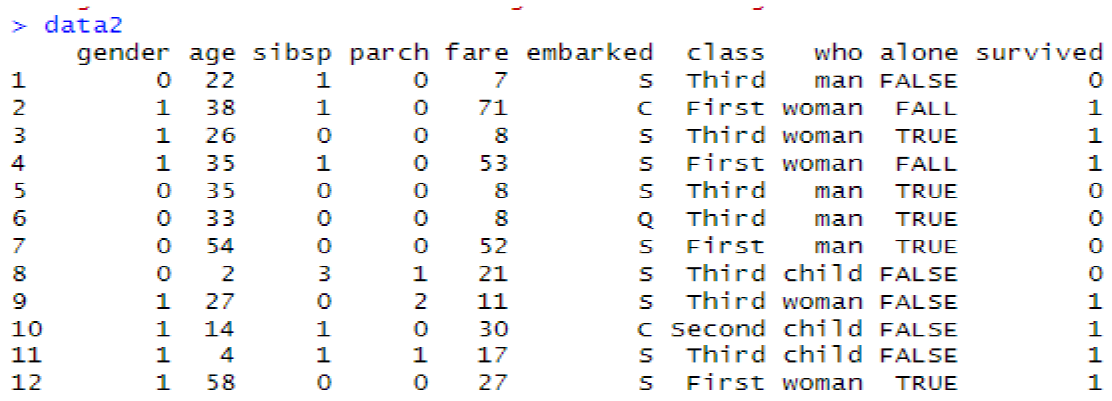
Boxplot for descriptive analysis of fare





Data Integration: No need for data integration. Because there is no other dataset

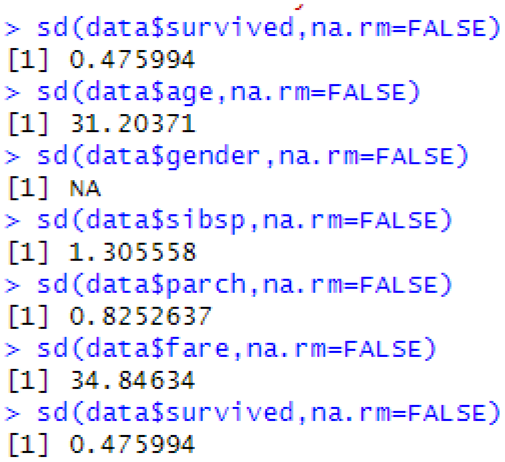
Data Transformation: Converting the age and fare values to integers.



```
> data2
   gender age sibsp parch fare embarked  class   who alone survived
1       0  22     1     0    7        S  Third   man FALSE        0
2       1  38     1     0   71        C  First woman  FALL        1
3       1  26     0     0    8        S  Third woman  TRUE        1
4       1  35     1     0   53        S  First woman  FALL        1
5       0  35     0     0    8        S  Third   man  TRUE        0
6       0  33     0     0    8        Q  Third   man  TRUE        0
7       0  54     0     0   52        S  First   man  TRUE        0
8       0   2     3     1   21        S  Third child FALSE        0
9       1  27     0     2   11        S  Third woman FALSE        1
10      1  14     1     0   30        C Second child FALSE        1
11      1   4     1     1   17        S  Third child FALSE        1
12      1  58     0     0   27        S  First woman  TRUE        1
```

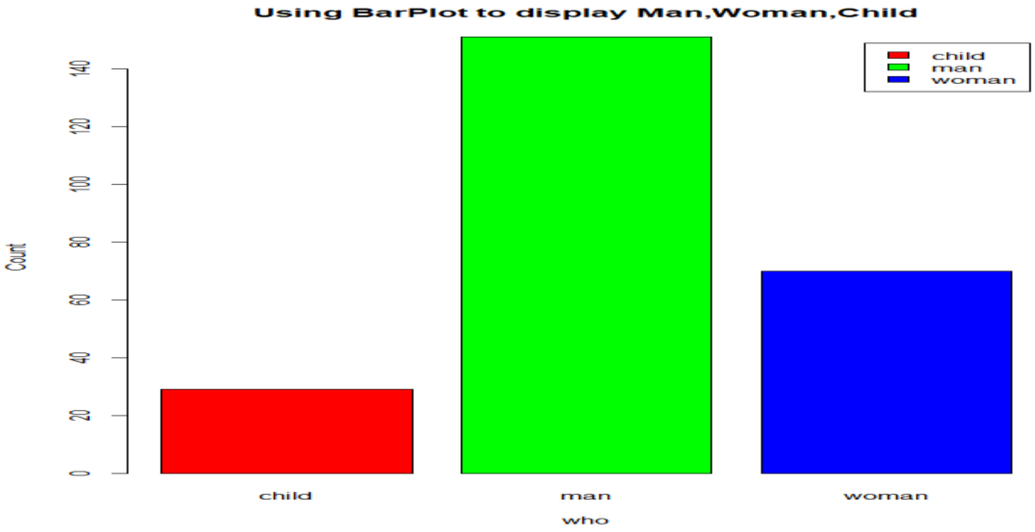
Data Reduction: Large dimensions datasets lead to large computation/training time and some algorithms do not perform well when we have large dimensions datasets. As their data size is very small so no need for data reduction.

Standard Deviation:

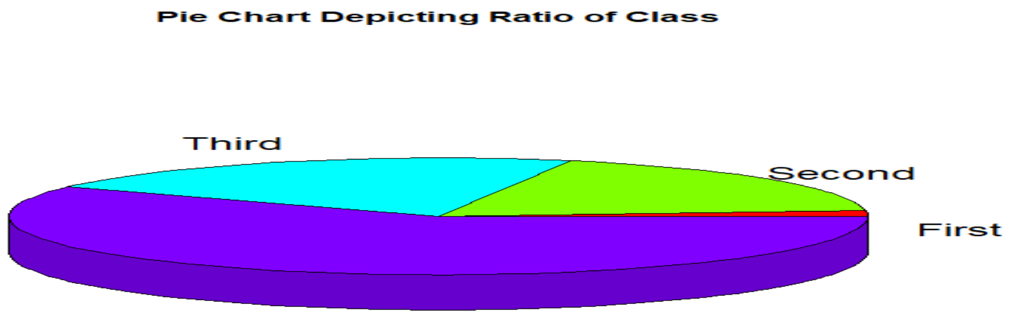


```
> sd(data$survived,na.rm=FALSE)
[1] 0.475994
> sd(data$age,na.rm=FALSE)
[1] 31.20371
> sd(data$gender,na.rm=FALSE)
[1] NA
> sd(data$sibsp,na.rm=FALSE)
[1] 1.305558
> sd(data$parch,na.rm=FALSE)
[1] 0.8252637
> sd(data$fare,na.rm=FALSE)
[1] 34.84634
> sd(data$survived,na.rm=FALSE)
[1] 0.475994
```

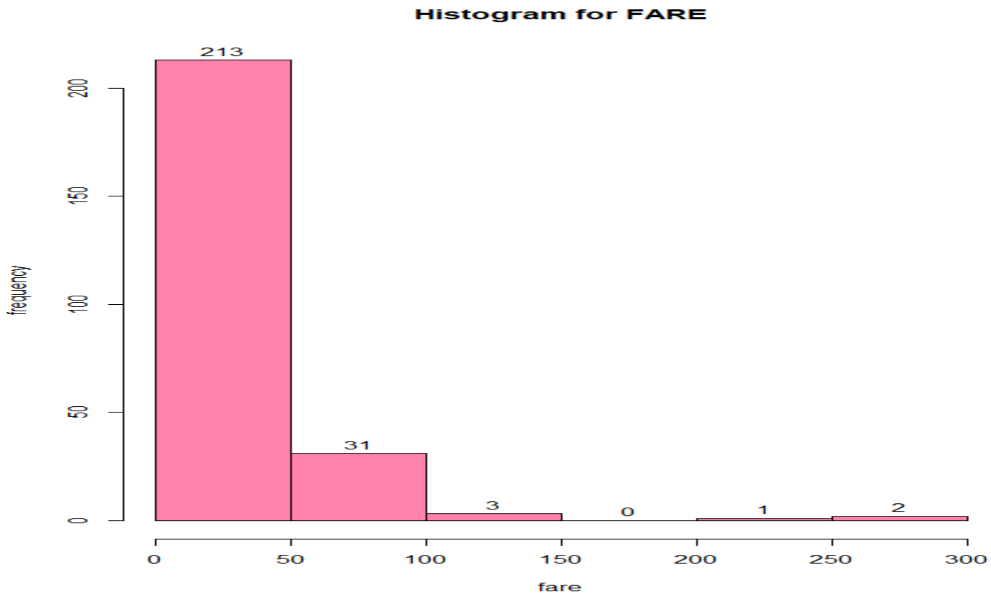
Use of Barplot: Used for ”Who” attribute



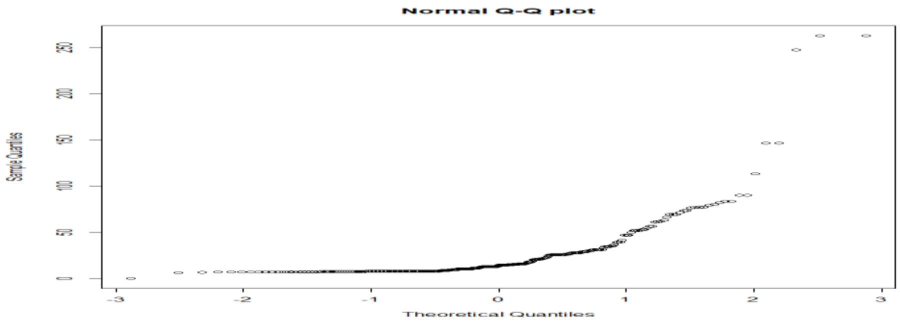
Use of Pie chart:



Use of Histogram:

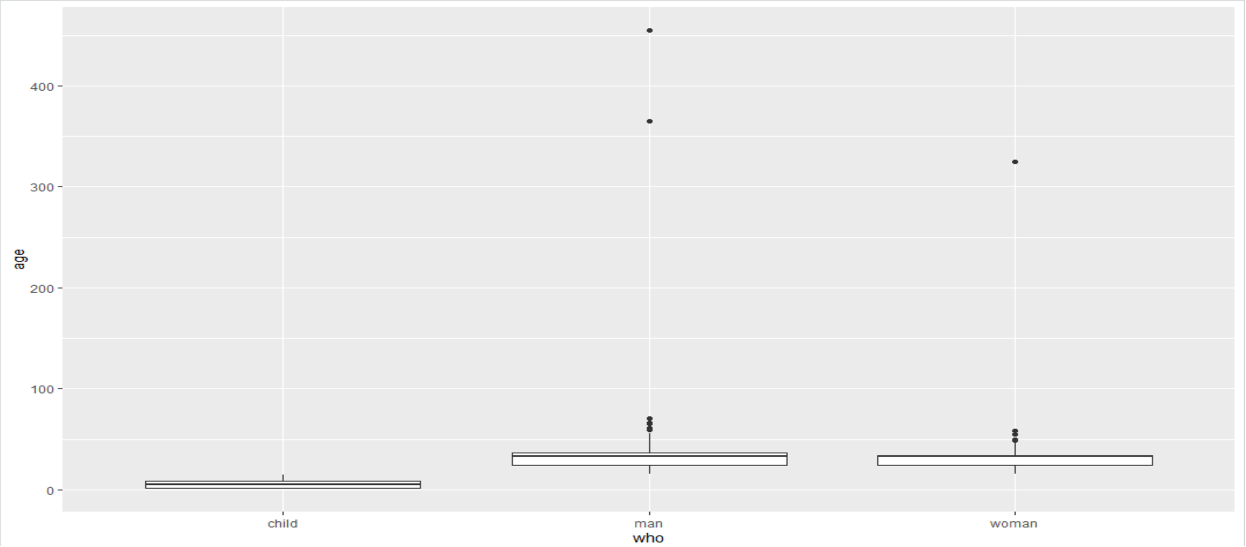


Use of Q-Q Plot:



```
library(ggplot2)
```

ggplot(data,aes(x=who,y=age))+geom_boxplot()



Discussion: Data preprocessing is the most important to process data for the best possible quality. Transforming raw data into a useful, understandable format by data preprocessing. This project is a very good opportunity to learn how to preprocess data by using R Language.

Conclusion: By preprocessing data using the R language generated a cleaned dataset. Now the data is ready for the analysis phase.