AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

408/1, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh

Assignment Title: Mid Term Project Report	
Assignment No: 01	Date of Submission: 18/07/2023
Course Title: INTRODUCTION TO DATA SCIENCE	
Course Code: 01153	Section: C
Semester: Summer 2022 - 23	Course Teacher: DR. 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	MD QUANET UL AHKAM ROKON	20-43582-1	BSc [CSE]	Quanetul
2				
3				
4				
5				
6				
7				
8				
9				
10				

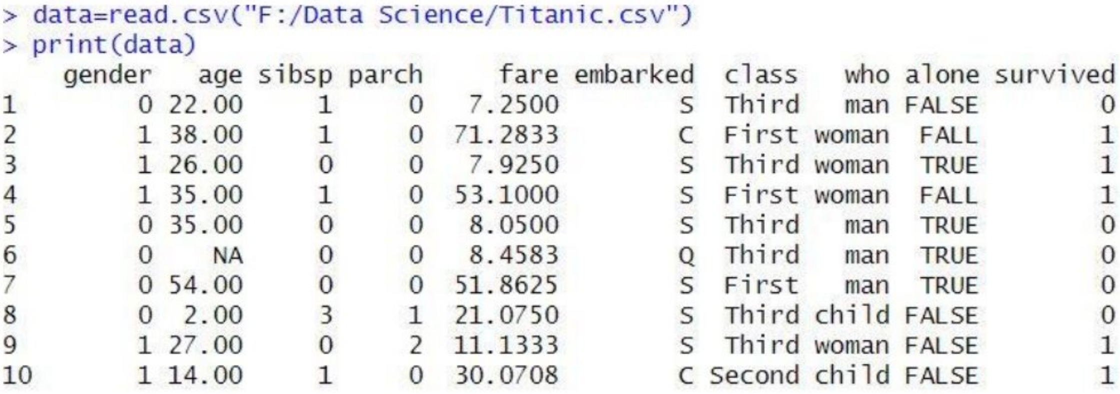
Faculty use only		
FACULTYCOMMENTS	Marks Obtained	
	Total Marks	

Overview: In the real world, data is frequently inconsistent, incomplete, and riddled with inaccuracies. There is an increased chance of running into aberrant or erroneous data due to the quick expansion of data creation and the rise in varied data sources. Processing the data is essential to achieving the highest quality data possible. Data preparation is required to convert the raw data into a format that is practical and clear. Due to the existence of noisy data, missing values, mistakes, or outliers, preprocessing of the data is required before doing data analysis on the provided dataset. The following data pretreatment operations must be carried out using the R programming language in order to produce a clean dataset.

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

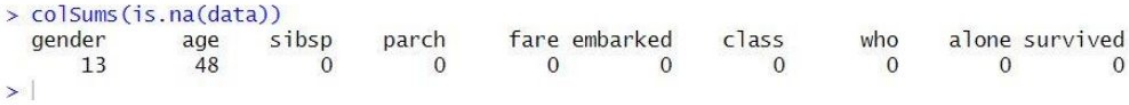
Insertion of Datasheet: Titanic.csv



```
> data=read.csv("F:/Data Science/Titanic.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
```

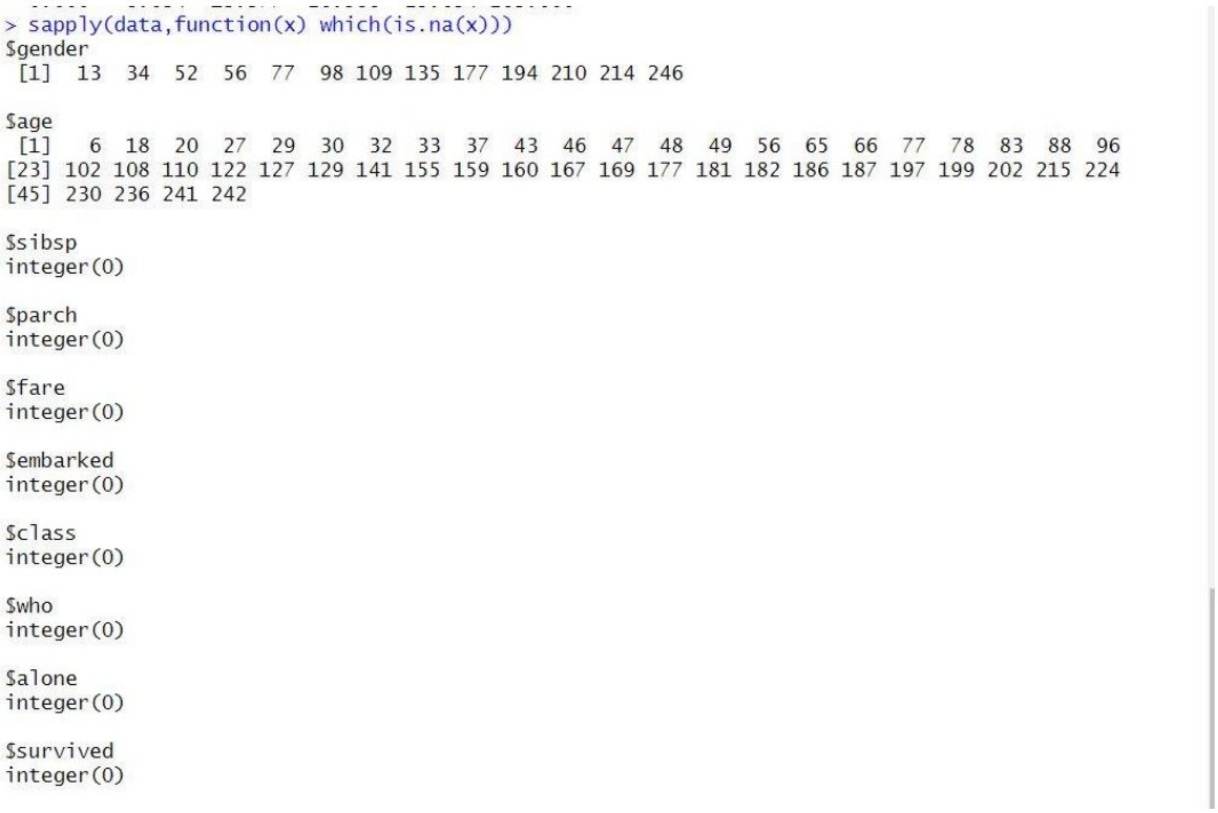
Data Cleaning: Handling Missing Data: The assault variable in this dataset has missing values. These missing values are denoted as "undefined" or "NA" in R programming, and any arithmetic operation with them will result in a "NAN" result. As a result, the mean values of the relevant variables must be used to replace these missing values.

Counting number of Null values in each column



```
> colSums(is.na(data))
  gender      age    sibsp    parch     fare embarked    class      who    alone survived 
      13       48        0        0        0        0        0        0        0        0 
> |
```

Specific position of Null value

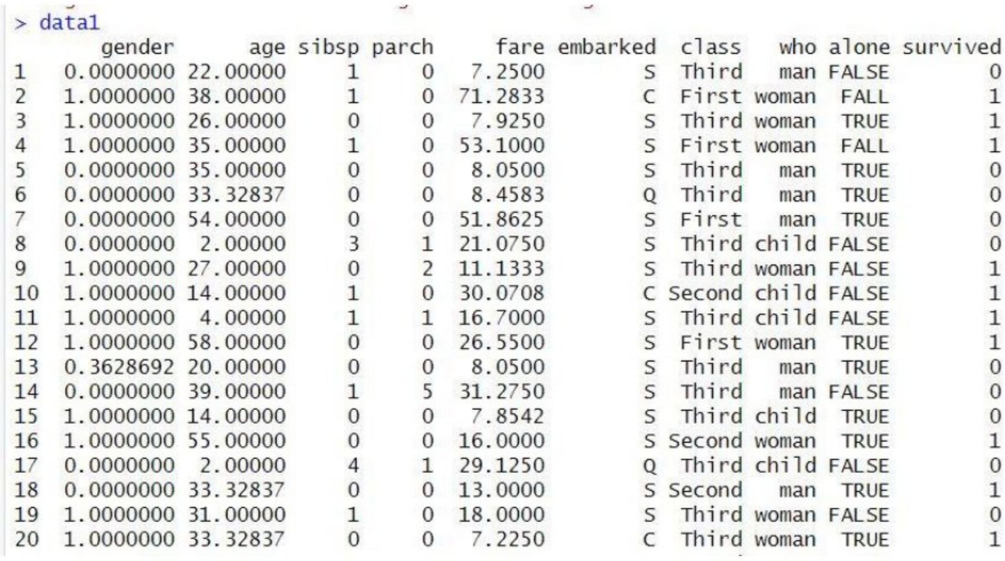


```
> sapply(data,function(x) which(is.na(x)))
$gender
 [1]  13  34  52  56  77  98 109 135 177 194 210 214 246

$age
 [1]   6  18  20  27  29  30  32  33  37  43  46  47  48  49  56  65  66  77  78  83  88  96
[23] 102 108 110 122 127 129 141 155 159 160 167 169 177 181 182 186 187 197 199 202 215 224
[45] 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 replace the missing value with MEAN value

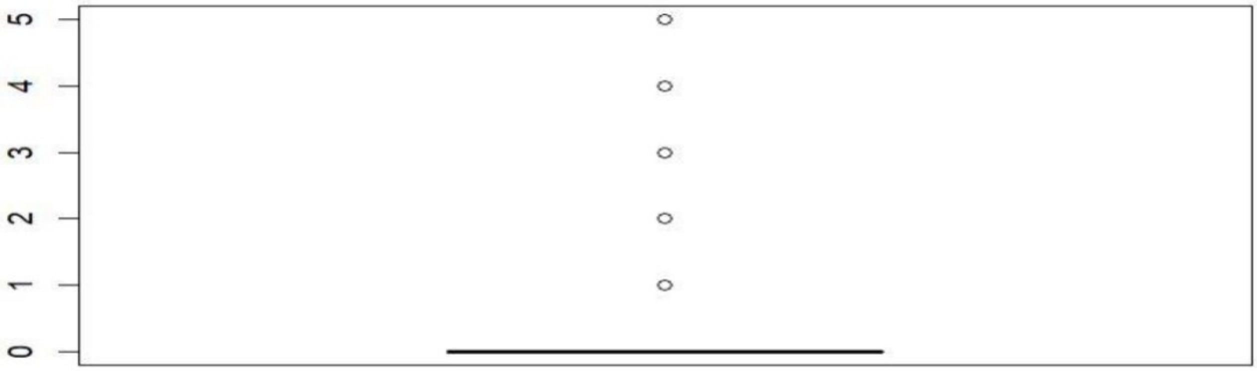
```
data$age[is.na(data$age)]<-mean(data$age,na.rm= TRUE)

print(data)

data1<-data

for(i in 1:ncol(data)){

data1[,i][is.na(data1[ ,i])]<-mean(data1[ ,i],na.rm= TRUE)

}

data1
```



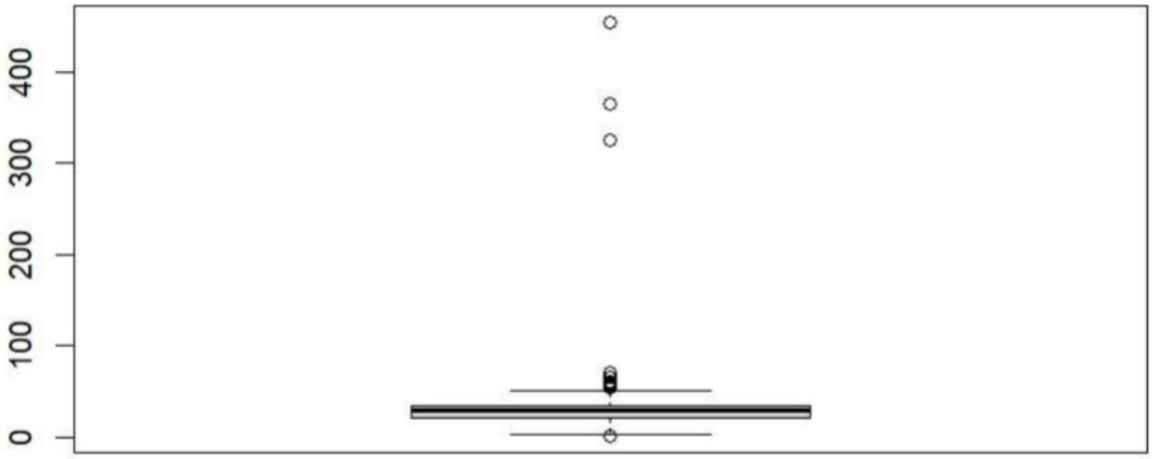
```
> data1
      gender      age sibsp parch    fare embarked  class   who alone survived
1  0.0000000 22.00000     1     0  7.2500        S  Third   man FALSE        0
2  1.0000000 38.00000     1     0 71.2833        C  First woman  FALL        1
3  1.0000000 26.00000     0     0  7.9250        S  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
```

Smooth Noisy Data: Data smoothing is the process of using statistical techniques to remove outliers from datasets so that the underlying patterns may be more easily seen. The Boxplot approach is one that is frequently used to find outliers.

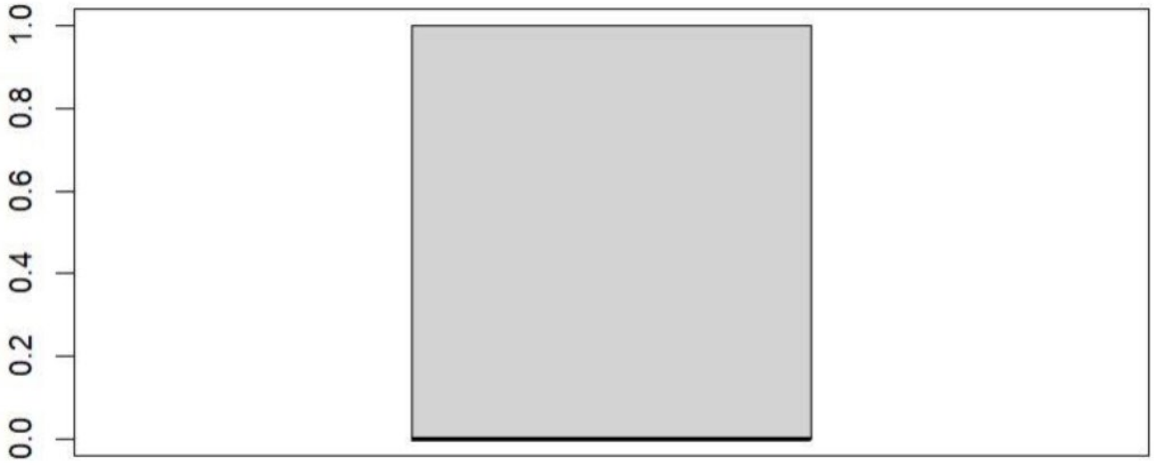
```
boxplot(data$age)

boxplot(data$gender)

boxplot(data$sibsp)

boxplot(data$parch)

boxplot(data$fare)

boxplot(data$survived)
```



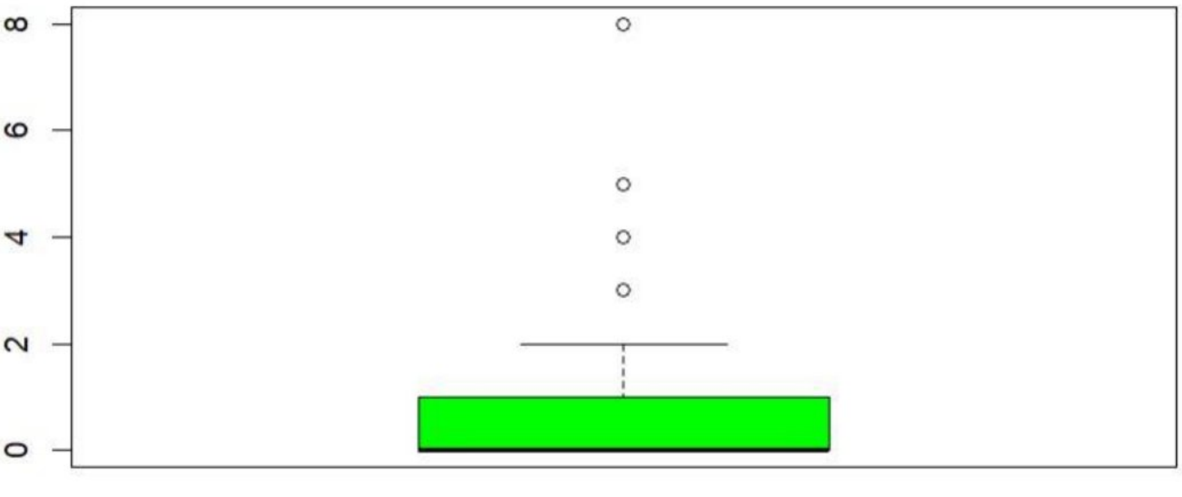
Analysis of parch



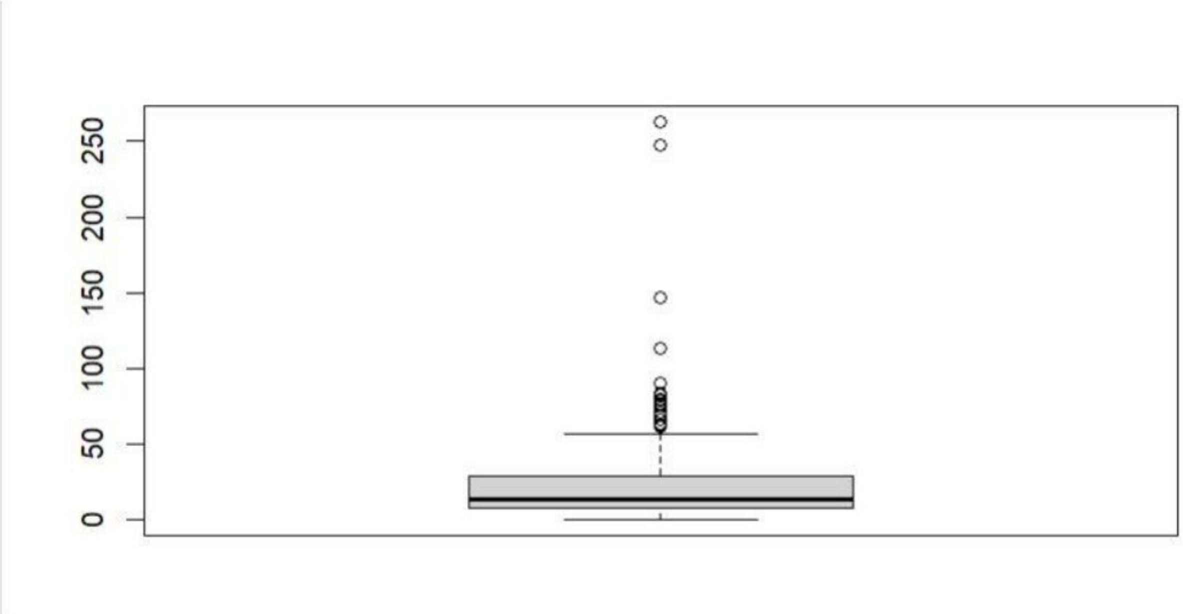
Analysis of Age



Analysis of Gender



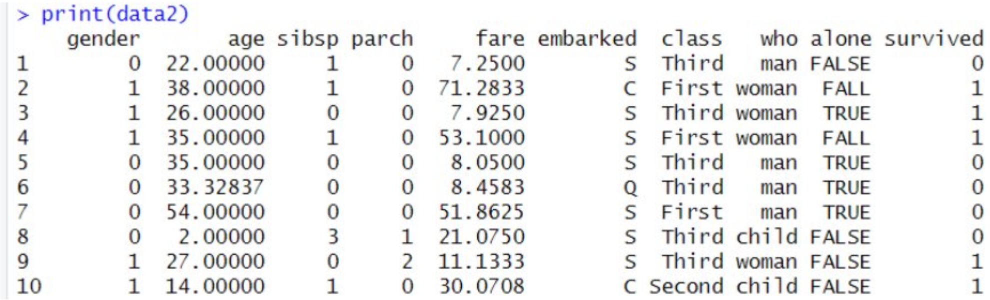
Analysis of sibsp



Analysis of fare

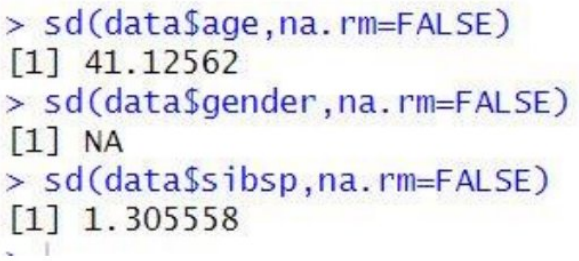
Data Integration: No need for data integration. Because there is no other dataset

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



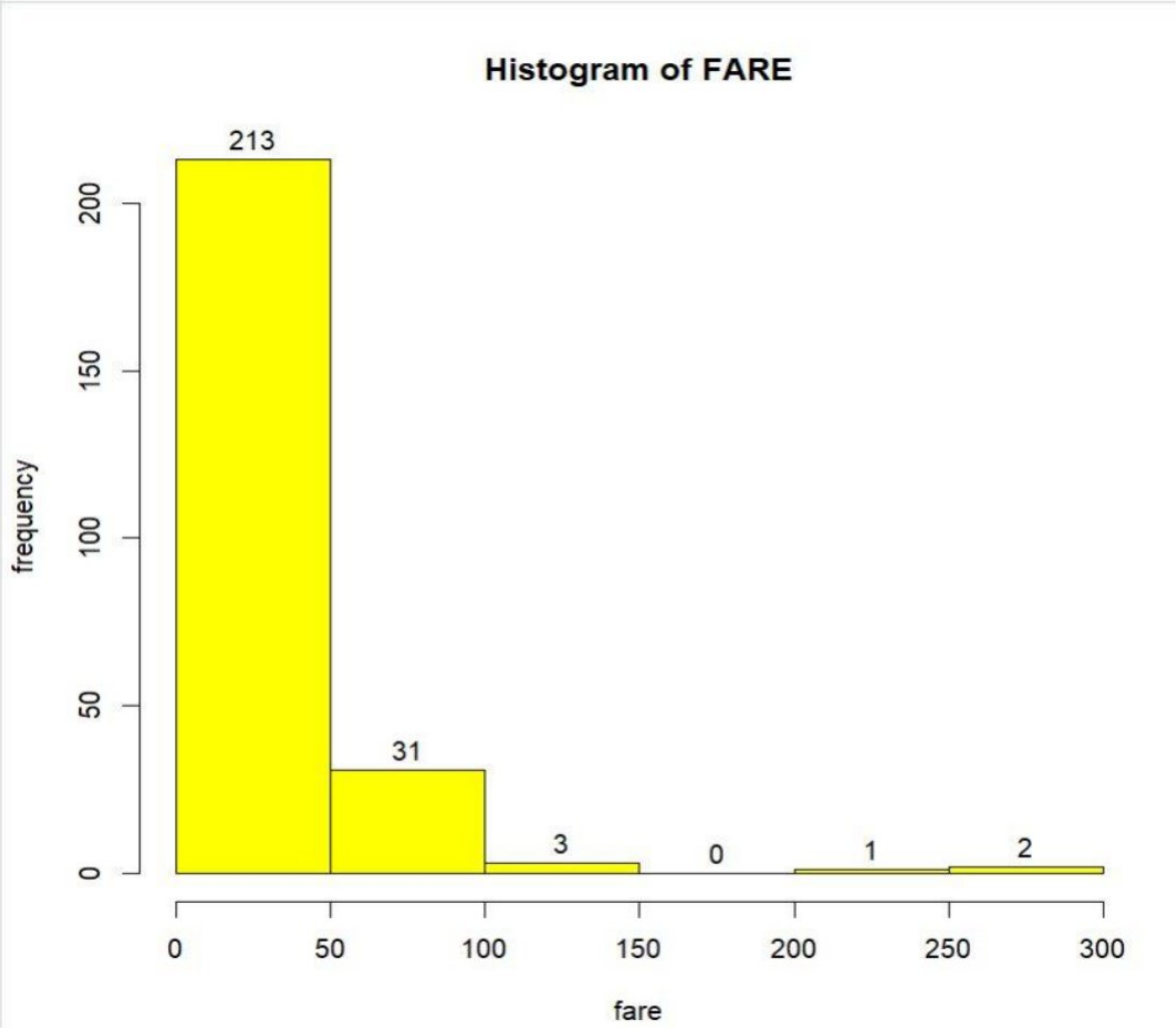
```
> print(data2)
   gender       age sibsp parch     fare embarked  class   who alone survived
1       0  22.00000     1     0   7.2500        S  Third   man FALSE        0
2       1  38.00000     1     0  71.2833        C  First woman  FALL        1
3       1  26.00000     0     0   7.9250        S  Third woman  TRUE        1
4       1  35.00000     1     0  53.1000        S  First woman  FALL        1
5       0  35.00000     0     0   8.0500        S  Third   man  TRUE        0
6       0  33.32837     0     0   8.4583        Q  Third   man  TRUE        0
7       0  54.00000     0     0  51.8625        S  First   man  TRUE        0
8       0   2.00000     3     1  21.0750        S  Third child FALSE        0
9       1  27.00000     0     2  11.1333        S  Third woman FALSE        1
10      1  14.00000     1     0  30.0708        C Second child FALSE        1
```

Data Reduction: When working with high-dimensional datasets, it might take a long time to compute and train, and some techniques might not function well. Data reduction approaches, however, might not be required in the case of tiny datasets because of the small amount of the data.Data smoothing is the process of using statistical techniques to remove outliers from datasets so that the underlying patterns may be more easily seen. The Boxplot approach is one that is frequently used to find outliers.

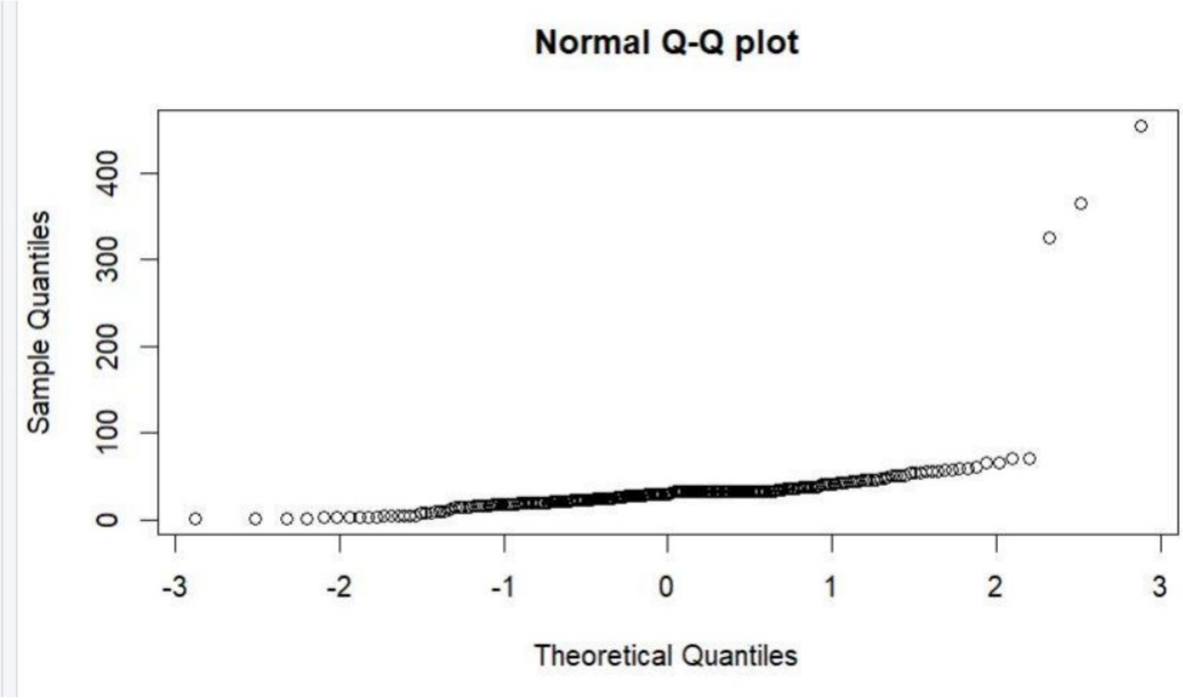


```
> sd(data$age,na.rm=FALSE)
[1] 41.12562
> sd(data$gender,na.rm=FALSE)
[1] NA
> sd(data$sibsp,na.rm=FALSE)
[1] 1.305558
>
```

Use of Histogram:



Use of Q-Q Plot:



Discussion: The analysis covered the value of data preparation in dealing with incomplete, noisy, and inconsistent real-world data. It stressed how crucial it is to fill in any missing values in the assault variable with the means of the relevant variables. In order to smooth noisy data, the Boxplot approach for outlier detection was also discussed.

Preprocessing data is essential because it guarantees data quality and gets the dataset ready for precise analysis. The data's integrity is preserved by addressing missing values and outliers, and any subsequent analysis will be built on a solid basis. Additionally, data smoothing techniques make underlying patterns and trends more obvious, facilitating interpretation and the extraction of valuable insights from the dataset.

Conclusion: In conclusion, each project involving data analysis must include data pretreatment. This project emphasizes how handling missing values and outliers calls for data cleansing and modification. The dataset is made more complete and analytically ready by substituting relevant measures, such mean values, for any missing values.

The research also highlights the significance of Boxplot outlier identification, which aids in finding and managing noisy data. The data can be cleaned up and outliers removed to improve the ensuing analysis's precision and focus.

Overall, the data preparation methods covered in this project contribute significantly to improving data quality and enabling more accurate and insightful analysis. Researchers and analysts can get more precise and worthwhile insights from the information by maintaining data cleanliness and minimizing the influence of noisy or inconsistent values.