AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Faculty of Science and Technology

Project Report Cover Page

Assignment Title:	Mid Term Project: Titanic Dataset Preparation	
Assignment No:	01	Date of Submission: July 17,2023
Course Title:	INTRODUCTION TO DATA SCIENCE[B]	
Course Code:	02399	Section: B
Semester: Summer		Course Teacher: TOHEDUL ISLAM

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	NAHIUN,AHNAF HASNAIN	20-42359-1	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	

Project Overview :-

The Titanic Dataset contains 301 instances with 10 attributes. This dataset provides information on various factor such as gender,age,class and survival status.Each instance represents a passenger details and attributes describe different factor those are related with titanic tragedy. All this information making it excellent resource for analyzing patterns and drawing.

The ten attributes are given below:

- Gender: the gender of passengers(integer)
- age: the age of the passengers(numeric)
- sibsp: sibling of passengers (integer)
- parch: Parents or children aboard with passenger (integer)
- fare: Each passenger fare (numeric)
- embarked: Port of embarkation (Char)
- calss: ticket class (Char)
- who: categories to passengers (char)
- alone: passenger was alone in ship or no (logi)
- survived: passenger survived or not (integer)

Data Preparation :

Data Exploration :

Load the Data

```
df <- read.csv("E:/R Mid Project/Dataset_midterm_Section(B).csv",
               header = TRUE, sep = ",", na.string = c(""))
head(df)
```

```
> df <- read.csv("E:/R Mid Project/Dataset_midterm_Section(B).csv",
+                header = TRUE, sep = ",", na.string = c(""))
> head(df)
  Gender age sibsp parch    fare embarked class     who alone survived
1      0  24     0     0  7.7958        S Third   mannn  TRUE        0
2      0  17     0     0  8.6625        S Third     man  TRUE        0
3      1  21     0     0  7.7500        Q Third  womann  TRUE        0
4      1  NA     0     0  7.6292        Q Third   woman  TRUE        0
5      1  37     0     0  9.5875        S Third womannn  TRUE        0
6     NA  16     0     0 86.5000        S First   woman  TRUE        1
>
```

View column Name

```
names(df)
```

```
> names(df)
 [1] "Gender"   "age"      "sibsp"    "parch"    "fare"     "embarked" "class"    "who"
 [9] "alone"    "survived"
```

Summary of the data

```
str(df)
```

```
> str(df)
'data.frame':   301 obs. of  10 variables:
 $ Gender  : int  0 0 1 1 1 NA 0 1 0 0 ...
 $ age     : num  24 17 21 NA 37 16 18 33 NA 28 ...
 $ sibsp   : int  0 0 0 0 0 0 1 0 0 0 ...
 $ parch   : int  0 0 0 0 0 0 0 2 0 0 ...
 $ fare    : num  7.8 8.66 7.75 7.63 9.59 ...
 $ embarked: chr  "S" "S" "Q" "Q" ...
 $ class   : chr  "Third" "Third" "Third" "Third" ...
 $ who     : chr  "mannn" "man" "womann" "woman" ...
 $ alone   : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
 $ survived: int  0 0 0 0 0 1 0 1 1 0 ...
>
```

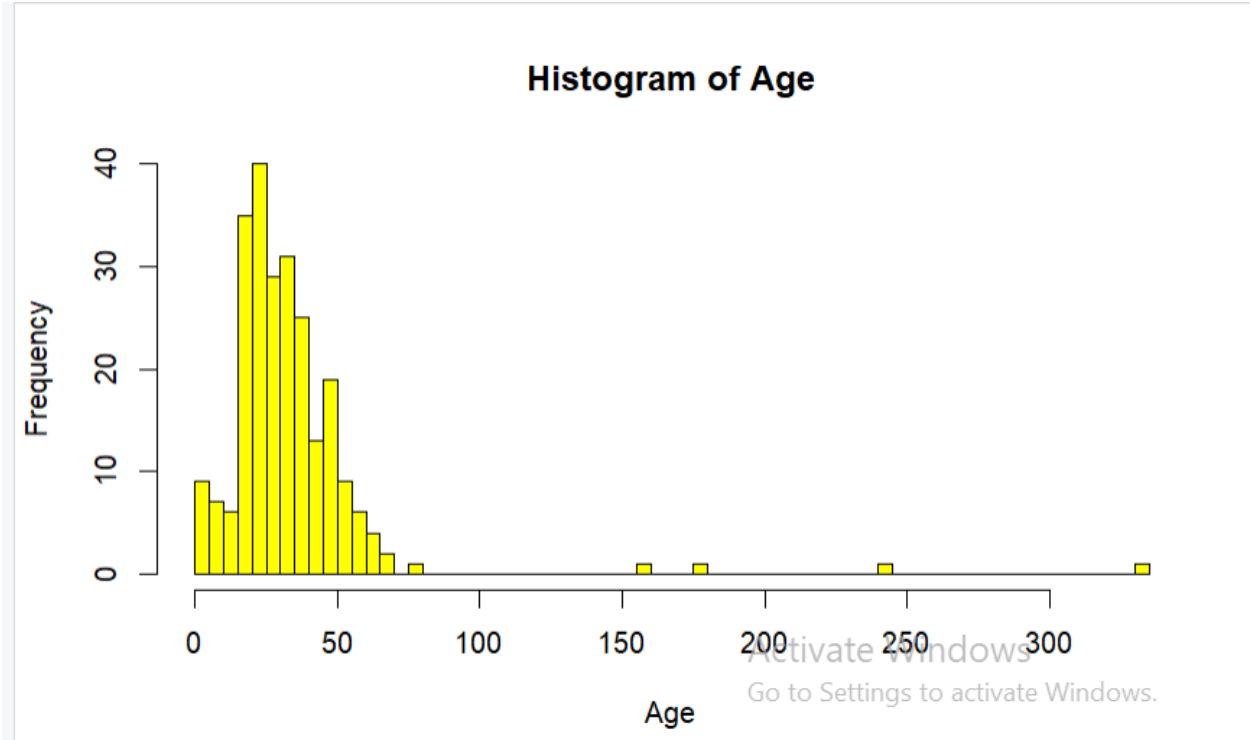
Descriptive Statistics

```
summary(df)
```

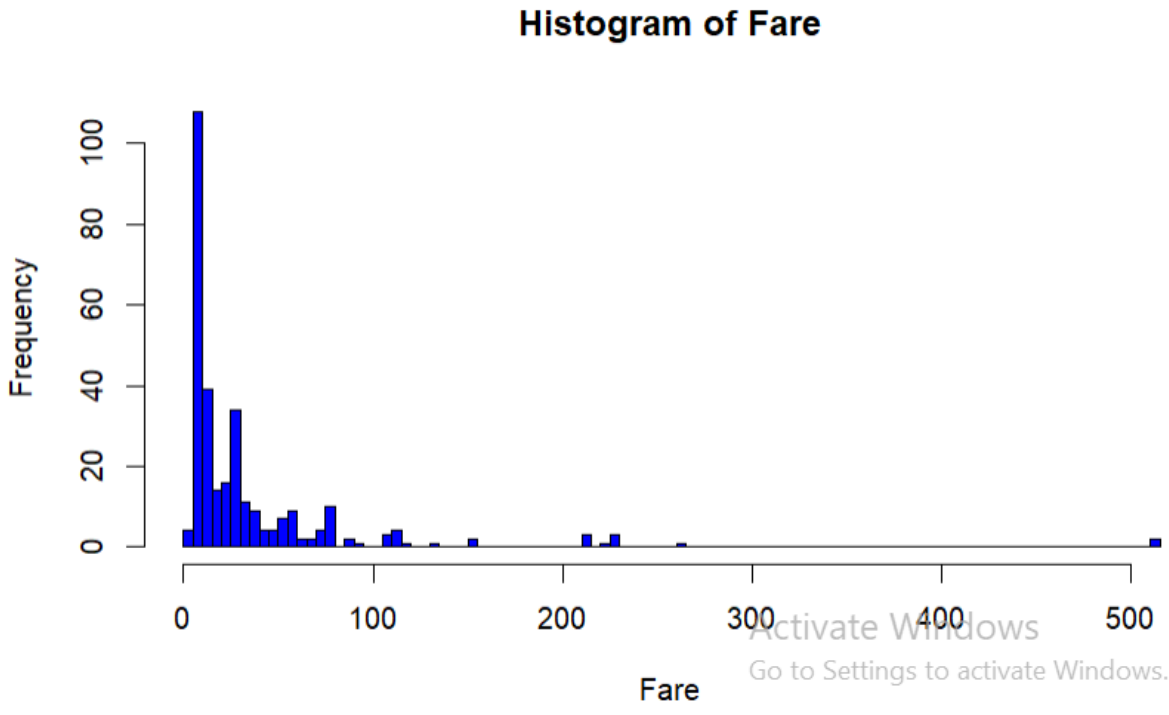
```
> summary(df)
     Gender            age             sibsp            parch             fare
 Min.   :0.0000   Min.   :  0.67   Min.   :0.0000   Min.   :0.0000   Min.   :  0.000
 1st Qu.:0.0000   1st Qu.: 21.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  7.896
 Median :0.0000   Median : 30.00   Median :0.0000   Median :0.0000   Median : 15.000
 Mean   :0.3199   Mean   : 34.04   Mean   :0.4252   Mean   :0.3621   Mean   : 35.041
 3rd Qu.:1.0000   3rd Qu.: 40.00   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 34.375
 Max.   :1.0000   Max.   :331.00   Max.   :8.0000   Max.   :6.0000   Max.   :512.329
 NA's   :4        NA's   :61
   embarked            class               who              alone            survived
 Length:301         Length:301         Length:301         Mode :logical   Min.   :0.0000
 Class :character   Class :character   Class :character   FALSE:109       1st Qu.:0.0000
 Mode  :character   Mode  :character   Mode  :character   TRUE :192       Median :0.0000
                                                                          Mean   :0.3821
                                                                          3rd Qu.:1.0000
                                                                          Max.   :1.0000

>
```

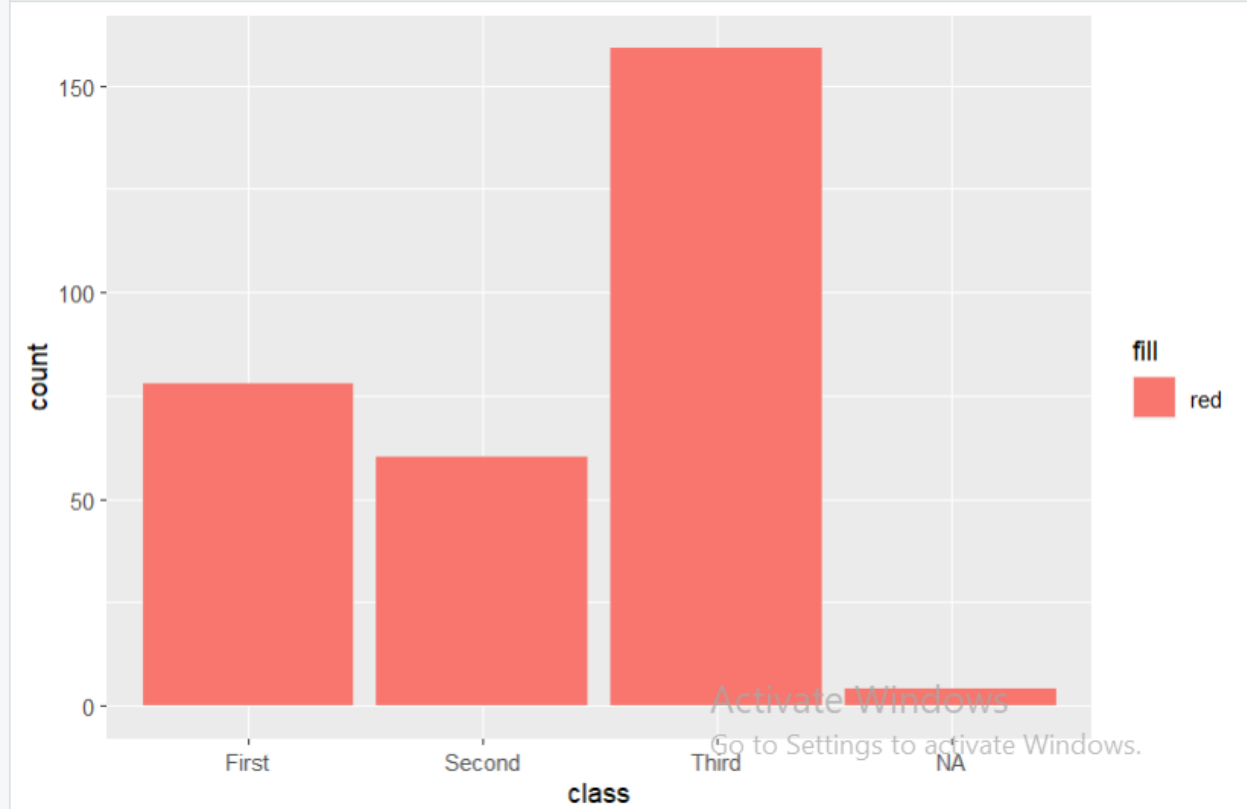
```
hist(df$age,breaks = 100,col = "yellow",main = "Histogram of Age",xlab = "Age")
```




```
hist(df$fare,breaks = 100,col = "blue",main = "Histogram of Fare",xlab = "Fare")
```



```
ggplot(data =df,aes(x=class, fill ="red")) + geom_bar()
```



Missing and Duplicate Values

Counting number of NULL values in each column

```
colSums(is.na(df))
```

```
> colSums(is.na(df))
  Gender      age    sibsp    parch     fare embarked    class      who    alone survived
       4       61        0        0        0        0        4        0        0        0
> |
```

Null Value replace by Mean value

```
> mean(df$age,na.rm = TRUE)
[1] 34.03508
> df$age[is.na(df$age)] <-mean(df$age,na.rm=TRUE)
> |
```

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

Null Value replace by Mode value

```
Mode <- function(x){
  ux <- na.omit(unique(x))
  tab <- tabulate(match(x,ux)); ux[tab == max(tab)]
}
Mode(df$class)
df$class[is.na(df$class)] <- Mode(df$class)
```

```
> Mode <- function(x){
+   ux <- na.omit(unique(x))
+   tab <- tabulate(match(x,ux)); ux[tab == max(tab)]
+ }
> Mode(df$class)
[1] "Third"
> df$class[is.na(df$class)] <- Mode(df$class)
> |
```

Removing Null and Missing Values

```
df <- remove <- na.omit(df)
colSums(is.na(df))
```

```
> df <- remove <- na.omit(df)
> colSums(is.na(df))
  Gender      age    sibsp    parch     fare embarked    class      who    alone survived
       0        0        0        0        0        0        0        0        0        0
> |
```

Cheacking Duplicates value

```
duplicated(df)
```

```
> duplicated(df)
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [16] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
 [31] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [46] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [61]  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [76] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
```

Removing Duplicates values

```
distinct(df)
str(df)
```

```
> duplicated(df)
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [16] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [31] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [46] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [76] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

Data types and Conversion :

Annoting Class column First =1, Second =2, Third =3

```
df$class <-(factor(df$class,
                   levels = c('First', 'Second', 'Third'),
                   labels = c(1,2,3)))
str(df)
```

```
> df$class <-(factor(df$class,
+                   levels = c('First', 'Second', 'Third'),
+                   labels = c(1,2,3)))
> str(df)
'data.frame':   277 obs. of  10 variables:
 $ Gender  : int  0 0 1 1 1 0 1 0 0 0 ...
 $ age     : num  24 17 21 34 37 ...
 $ sibsp   : int  0 0 0 0 0 1 0 0 0 0 ...
 $ parch   : int  0 0 0 0 0 0 2 0 0 0 ...
 $ fare    : num  7.8 8.66 7.75 7.63 9.59 ...
 $ embarked: chr  "S" "S" "Q" "Q" ...
 $ class   : Factor w/ 3 levels "1","2","3": 3 3 3 3 3 1 2 1 3 3 ...
 $ who     : chr  "mannn" "man" "womann" "woman" ...
 $ alone   : logi  TRUE TRUE TRUE TRUE TRUE FALSE ...
 $ survived: int  0 0 0 0 0 0 1 1 0 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:4] 6 23 32 43
  ..- attr(*, "names")= chr [1:4] "6" "23" "32" "43"
>
```

Converting Char to Integer Data type

```
df$class <- as.integer(df$class)
str(df)
```

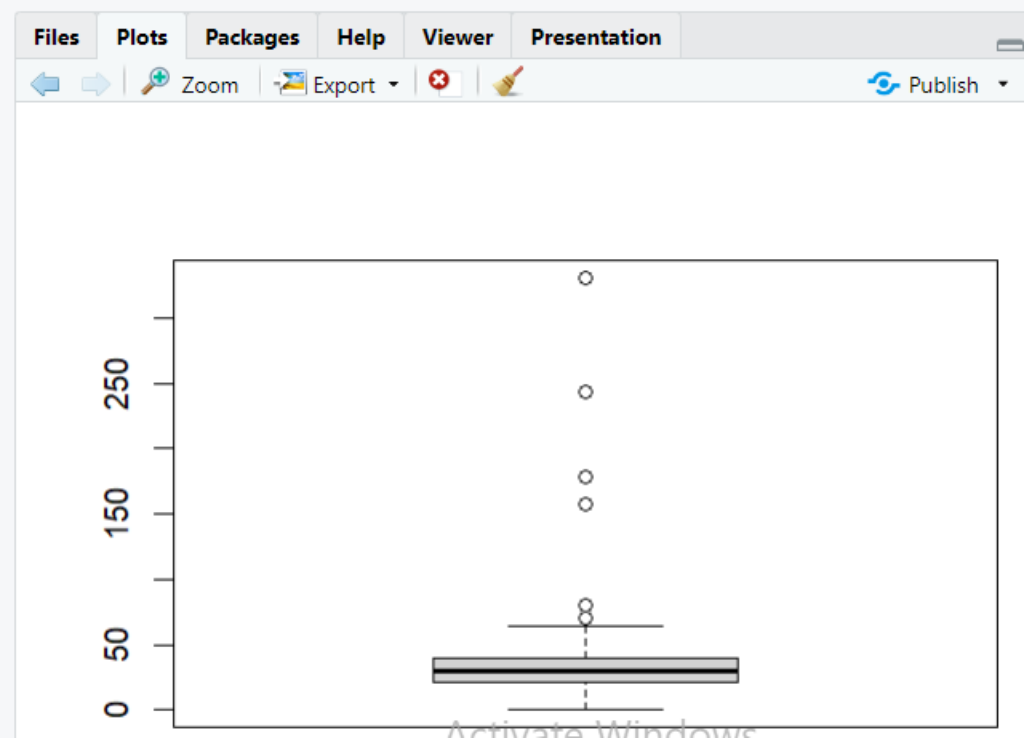
```
> df$class <- as.integer(df$class)
> str(df)
'data.frame':   277 obs. of  10 variables:
 $ Gender  : int  0 0 1 1 1 0 1 0 0 0 ...
 $ age     : num  24 17 21 34 37 ...
 $ sibsp   : int  0 0 0 0 0 1 0 0 0 0 ...
 $ parch   : int  0 0 0 0 0 0 2 0 0 0 ...
 $ fare    : num  7.8 8.66 7.75 7.63 9.59 ...
 $ embarked: chr  "S" "S" "Q" "Q" ...
 $ class   : int  3 3 3 3 3 1 2 1 3 3 ...
 $ who     : chr  "mannn" "man" "womann" "woman" ...
 $ alone   : logi  TRUE TRUE TRUE TRUE TRUE FALSE ...
 $ survived: int  0 0 0 0 0 0 1 1 0 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:4] 6 23 32 43
  ..- attr(*, "names")= chr [1:4] "6" "23" "32" "43"
>
```

#Outlier :

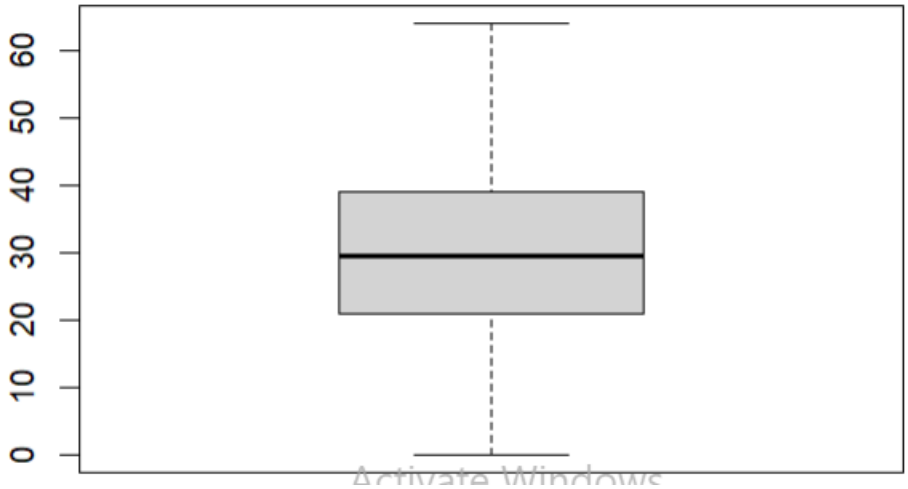
Detecting Outlier for age attribute

```
summary(df1$age)
IQR_age <- 40 - 21
upper_age <- 40 + 1.5*IQR_age
upper_age
lower_age <- 21 -1.5*IQR_age
lower_age

boxplot(df1$age)
data <- df1[!(df1$age<lower_age | df1$age>upper_age),]
boxplot(data$age)
summary(data$age)
```



Removing Outliers

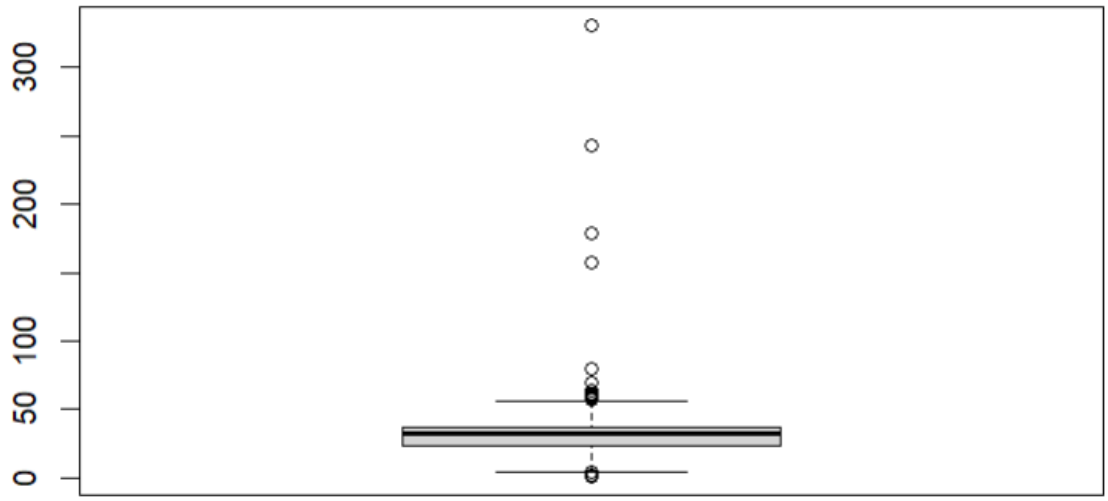



```
> summary(df1$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.67   24.00   33.00   34.45   37.00  331.00
> IQR_age <- 40 - 21
> upper_age <- 40 + 1.5*IQR_age
> upper_age
[1] 68.5
> lower_age <- 21 -1.5*IQR_age
> lower_age
[1] -7.5
> boxplot(df1$age)
> data <- df1[!(df1$age<lower_age | df1$age>upper_age),]
> boxplot(data$age)
> summary(data$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.67   23.00   32.50   31.16   36.00   64.00
```

Detecting Outlier for fare attribute

```
summary(df1$fare)
boxplot(df1$fare)
IQR_fare <- 39 - 7.925
upper_fare <- 39 + 1.5*IQR_fare
upper_fare
lower_fare <- 7.925 - 1.5*IQR_fare
lower_fare

boxplot(df1$fare)
grid()
data <- df1[!(df1$fare<lower_fare | df1$fare>upper_fare),]
boxplot(data$fare)
grid()
```



```
> summary(df1$fare)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   7.925  16.100  36.762  39.000 512.329
> boxplot(df1$fare)
> IQR_fare <- 39 - 7.925
> upper_fare <- 39 + 1.5*IQR_fare
> upper_fare
[1] 85.6125
> lower_fare <- 7.925 - 1.5*IQR_fare
> lower_fare
[1] -38.6875
> boxplot(df1$fare)
> data <- df1[!(df1$fare<lower_fare | df1$fare>upper_fare),]
> boxplot(data$fare)
> summary(data$fare)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   7.896  14.479  23.036  29.781  79.650
```

Standard deviation of ‘age’ and ‘fare’ attributes

```
print(sd(data$fare)
print(sd(data$age))
```

```
> print(sd(data$fare))
[1] 19.72141
> print(sd(data$age))
[1] 29.04188
```

 # <u>Univariate Exploration(mean,median,mode, Standard Deviation)</u> :

```
data %>% summarize_if(is.numeric, mean)

data %>% summarize_if(is.numeric, median)

data %>% summarize_if(is.numeric, mode)

data %>% summarize_if(is.numeric, sd)
```

```
> data %>% summarize_if(is.numeric, mean)
     Gender      age     sibsp     parch     fare    class  survived
1 0.3228346 34.51547 0.4448819 0.3661417 23.03648 2.354331 0.3740157
> data %>% summarize_if(is.numeric, median)
  Gender age sibsp parch     fare class survived
1      0  33     0     0 14.47915     3        0
> data %>% summarize_if(is.numeric, mode)
   Gender     age   sibsp   parch    fare   class survived
1 numeric numeric numeric numeric numeric numeric  numeric
> data %>% summarize_if(is.numeric, sd)
     Gender      age     sibsp     parch     fare     class survived
1 0.4684832 29.04188 0.9083024 0.8641952 19.72141 0.8102853 0.484823
>
```

<u>Feature Selection :</u>

Feature selection

```
data <- subset(data, select = -c(sibsp, parch, embarked, who))
head(data)
View(data)

titanic_target <- data[6]
head(titanic_target)

titanic_feature <- data[1:5]
head(titanic_feature)
view(data)
```

```
> titanic_target <- data[6]
> head(titanic_target)
  survived
1        0
2        0
3        0
4        0
5        0
8        1
> titanic_feature <- data[1:5]
> head(titanic_feature)
  Gender      age    fare class alone
1      0 24.00000  7.7958     3  TRUE
2      0 17.00000  8.6625     3  TRUE
3      1 21.00000  7.7500     3  TRUE
4      1 34.03508  7.6292     3  TRUE
5      1 37.00000  9.5875     3  TRUE
8      1 33.00000 26.0000     2 FALSE
>
```