EXPERIMENT 7.1

Taniya Ahmed

21BDS0059

1. **Load the necessary libraries and dataset.**

   CODE:

```
install.packages("corrplot")

library("corrplot")

print("Taniya Ahmed 21BDS0059")


df = read.csv("D:\\Downloads\\titanic.csv")
```

   OUTPUT:

```
> library("corrplot")
Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  display list redraw incomplete
2: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
3: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
> print("Taniya Ahmed 21BDS0059")
[1] "Taniya Ahmed 21BDS0059"
>
> df = read.csv("D:\\Downloads\\titanic.csv")
```

2. **Check the structure and description of the data.**

   CODE:

```
head(df)

str(df)

summary(df)

print("Taniya Ahmed 21BDS0059")
```

OUTPUT:

```
> head(df)
  PassengerId Survived Pclass
1           1        0      3
2           2        1      1
3           3        1      3
4           4        1      1
5           5        0      3
6           6        0      3
                                                 Name    Sex Age SibSp Parch
1                             Braund, Mr. Owen Harris   male  22     1     0
2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
3                              Heikkinen, Miss. Laina female  26     0     0
4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
5                            Allen, Mr. William Henry   male  35     0     0
6                                    Moran, Mr. James   male  NA     0     0
            Ticket    Fare Cabin Embarked
1        A/5 21171  7.2500              S
2         PC 17599 71.2833   C85        C
3 STON/O2. 3101282  7.9250              S
4           113803 53.1000  C123        S
5           373450  8.0500              S
6           330877  8.4583              Q
> str(df)
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence
Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Pee
l)" ...
 $ Sex        : chr  "male" "female" "female" "female" ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : chr  "" "C85" "" "C123" ...
 $ Embarked   : chr  "S" "C" "S" "S" ...
> summary(df)
  PassengerId       Survived          Pclass          Name
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
 Median :446.0   Median :0.0000   Median :3.000   Mode  :character
 Mean   :446.0   Mean   :0.3838   Mean   :2.309
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
 Max.   :891.0   Max.   :1.0000   Max.   :3.000

     Sex                 Age            SibSp           Parch
 Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
 Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
 Mode  :character   Median :28.00   Median :0.000   Median :0.0000
                    Mean   :29.70   Mean   :0.523   Mean   :0.3816
                    3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
                    Max.   :80.00   Max.   :8.000   Max.   :6.0000
                    NA's   :177
    Ticket              Fare           Cabin             Embarked
 Length:891         Min.   :  0.00   Length:891         Length:891
 Class :character   1st Qu.:  7.91   Class :character   Class :character
 Mode  :character   Median : 14.45   Mode  :character   Mode  :character
                    Mean   : 32.20
                    3rd Qu.: 31.00
                    Max.   :512.33

> print("Taniya Ahmed 21BDS0059")
[1] "Taniya Ahmed 21BDS0059"
```

3. **Check for missing values in the dataset.**

CODE:

```
sum(is.na(df))

colSums(is.na(df))

print("Taniya Ahmed 21BDS0059")
```

OUTPUT:

```
> sum(is.na(df))
[1] 177
Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  display list redraw incomplete
2: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
3: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
4: In doTryCatch(return(expr), name, parentenv, handler) :
  display list redraw incomplete
5: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
6: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
> colSums(is.na(df))
PassengerId    Survived      Pclass        Name         Sex         Age
          0           0           0           0           0         177
      SibSp       Parch      Ticket        Fare       Cabin    Embarked
          0           0           0           0           0           0
> print("Taniya Ahmed 21BDS0059")
[1] "Taniya Ahmed 21BDS0059"
```

4. **Doing univariate analysis by plottingthe numerical and categorical variables age and ticket.**

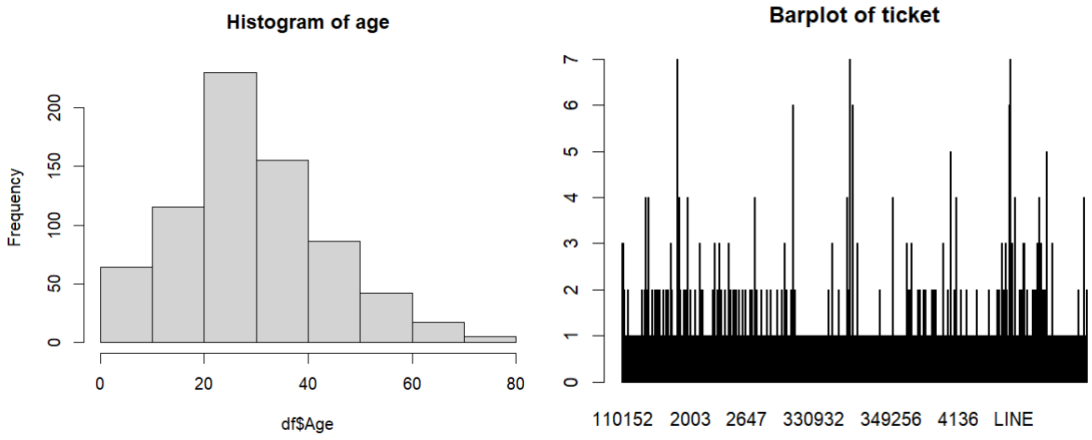
CODE:

```
hist(df$Age, main = "Histogram of age")

barplot(table(df$Ticket), main = "Barplot of ticket")

print("Taniya Ahmed 21BDS0059")
```

OUTPUT:



5. **Doing bivariate analysis by using scatterplot and boxplot.**

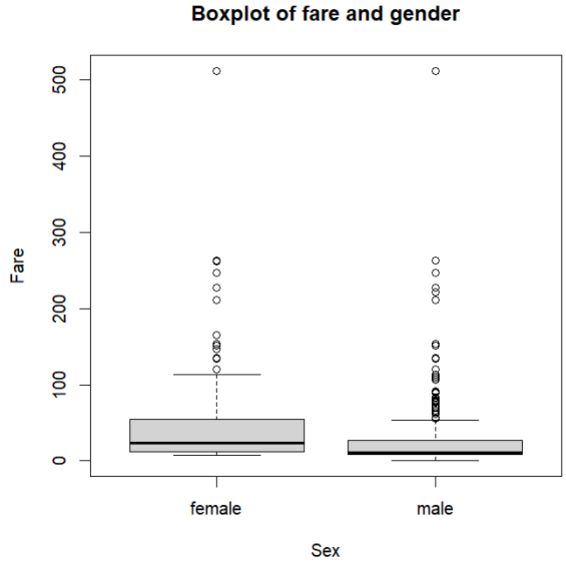
CODE:

```
plot(df$Age, df$Fare, main = "Scatter plot of age vs fare", xlab = "Age", ylab = "Fare")

boxplot(Fare ~ Sex, data = df, main = "Boxplot by category")
```

print("Taniya Ahmed 21BDS0059")

OUTPUT:



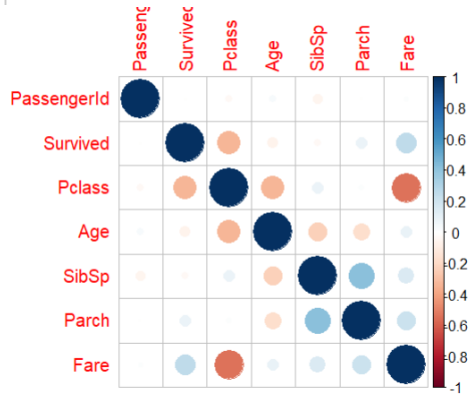
6. **Imputing NA values and Doing multivariate analysis.**

CODE:

```
df$Age[is.na(df$Age)] <- mean(df$Age, na.rm = TRUE)
cor(df[, sapply(df, is.numeric)])
corrplot(cor(df[, sapply(df, is.numeric)]), method = "circle")
print("Taniya Ahmed 21BDS0059")
```

OUTPUT:

```
> df$Age[is.na(df$Age)] <- mean(df$Age, na.rm = TRUE)
> cor(df[, sapply(df, is.numeric)])
             PassengerId     Survived      Pclass         Age       SibSp        Parch        Fare
PassengerId  1.000000000 -0.005006661 -0.03514399  0.03320655 -0.05752683 -0.001652012  0.01265822
Survived    -0.005006661  1.000000000 -0.33848104 -0.06980852 -0.03532250  0.081629407  0.25730652
Pclass      -0.035143994 -0.338481036  1.00000000 -0.33133877  0.08308136  0.018442671 -0.54949962
Age          0.033206546 -0.069808515 -0.33133877  1.00000000 -0.23262459 -0.179190915  0.09156609
SibSp       -0.057526834 -0.035322499  0.08308136 -0.23262459  1.00000000  0.414837699  0.15965104
Parch       -0.001652012  0.081629407  0.01844267 -0.17919092  0.41483770  1.000000000  0.21622494
Fare         0.012658219  0.257306522 -0.54949962  0.09156609  0.15965104  0.216224945  1.00000000
> corrplot(cor(df[, sapply(df, is.numeric)]), method = "circle")
> print("Taniya Ahmed 21BDS0059")
[1] "Taniya Ahmed 21BDS0059"
>
```



7. **Plotting boxplot for outliers and reducing the affect of outliers by applying logarithmic transformation.**

CODE:

```
boxplot(df$Age, main = "Boxplot for Outliers of Age")

df$Age = log(df$Age)

print("Taniya Ahmed 21BDS0059")
```

OUTPUT:

```
> boxplot(df$Age, main = "Boxplot for Outliers of Age")
> df$Age = log(df$Age)
> print("Taniya Ahmed 21BDS0059")
[1] "Taniya Ahmed 21BDS0059"
>
```