Project 2
Name of group members: 1. Ajay Danda (AXD180068) 2. Satyam Bhikadiya (SXB180124)
Contribution of each group member:
Ajay Danda:

- Equal contribution in solving Q1
- Equal contribution in solving Q2
- Equal contribution in Documenting the Report.

Satyam Bhikadiya:

- Equal contribution in solving Q1
- Equal contribution in solving Q2
- Equal contribution in Documenting the Report.

Question 1:

Consider the dataset roadrace.csv posted on eLearning. It contains observations on 5875 runners who finished the 2010 Beach to Beacon 10K Road Race in Cape Elizabeth, Maine. You can read the dataset in R using read.csv function.

(a) Create a bar graph of the variable Maine, which identities whether a runner is from Maine or from somewhere else (stated using Maine and Away). You can use barplot function for this. What can we conclude from the plot? Back up your conclusions with relevant summary statistics.

Answer:
We plotted the bar graph using the barplot function. The read.csv() function was used to read the main data.
The conclusion that we obtained after plotting the bargraph is that there are more than 4000 people from Maine that participated in the Road Race. We also concluded that the population of people that are not from Maine and participated in the Road Race is far less than the population of people from Maine. The people that are not from Maine are termed as "Away". The Away population is between 1000 to 2000 people.
These conclusions were further backed up when we computed the summary statistics using the summary() method. The summary method gave us the frequency of the each of the two type of population that participated in the Road Race.The summary statistics showed that there were 4458 people from Maine and 1417 Away people that participated in that particular Road Race.

(b) Create two histograms the runners' times (given in minutes) | one for the Maine group and the second for the Away group. Make sure that the histograms on the same scale. What can we conclude about the two distributions? Back up your conclusions with relevant summary statistics, including mean, standard deviation, range, median, and interquartile range.
Answer:
In order to plot histograms we used the hist() method. To obtain two histograms of data from the our main roadrace.csv file, we divided the entire data into two separate data sets. The two data sets are, one data set has only "Maine" in the Maine attribute column. The other data set has only "Away" values in the Maine attribute column. Therefore our distinguishing condition of creating the data sets was whether the value of Maine attribute was "Maine" or "Away" now for each of the two data sets, we plotted histograms for "Time…minutes..". Plotting histogram for the Maine group we can conclude that more than 1500 people were able to run in the Road Race for 50 to 60 minutes. The histogram also shows that there are almost same number of people on either side of the tallest bar, which means that the mean of Maine group running in the Road Race will be between 50 to 60 minutes. The histogram shows that there were very few who completed the Road Race in 160 minutes and everybody from the Maine group ran at least for 30 minutes. This was proved when we computed the summary statistics of the same. The result obtained was as follows:

```
    summary(dataMaine$Time..minutes.)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.57   50.00   57.03   58.20   64.24  152.17
```

Now, plotting histogram for the Away group we can conclude that approximately 480 to 500 people were able to run in the Road Race for 50 to 60 minutes. Here this is the maximum number of people from the Away group that ran for 50 to 60 minutes. This bar is the tallest and has approximately same number of people on either side of the bar. Therefore, we can conclude that the mean for the Away group also lies between 50 to 60 minutes.There were few swh o completed the Road Race from the Away group in approximately 140 minutes and everybody ran at least 20 minutes from the Away group. This was closely proved when we computed the statistics of the same.The results obtained were as follows:

```
summary(dataAway$Time..minutes.)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
27.78   49.15   56.92   57.82   64.83  133.71
```

From the above two statistics and the histogram distributions we can say that the Time…minutes for Maine and Away group exhibits a Normal Distribution. This is backed up when we see that the Mean values and Median values for both the groups is very close to each other in a particular group. Both the distributions are right-skewed. We can also conclude that even if there were more than 4000 people from Maine and fewer from Away, the Mean values for both of them is very close to each other. This shows that the Away group had more good runners than the Maine group.

(c) Repeat (b) but with side-by-side boxplots.
Answer:
To plot the boxplot for the above cases where Maine and Away groups have to be plotted separately. There we get side by side boxplots for the same. We use the method boxplot() to plot the boxplot. The syntax given below is used to plot the side by side boxplots by grouping the Maine attribute column by its values and then plots them with Time…minutes… these boxplots give us the five point summary for the Maine and Away group. These boxplots show us the outliers for each of the groups. We can see that there are many outliers for both the groups, however, Maine group has way more outliers than Away group. The dotted lines on either side of the box gives us the Q3+1.5*IQR and Q1-1.5IQR range. If any data values that lie beyond this range are known as outliers. The boxplots also show the spread of data for each of the group along with identifying outliers.

(d) Create side-by-side boxplots for the runners' ages (given in years) for male and female runners. What can we conclude about the two distributions? Back up your conclusions with relevant summary statistics, including mean, standard deviation, range, median, and interquartile range.
Answer:
To plot the boxplot for runners age for male and female runners, we use the same method as above where the boxplot() function will group the M and F values of the Sex attribute column and plot the boxplots accordingly. From the obtained boxplots, we can say that lowest value for Female runners will be approximately 5-6 years and lowest for Male runners will be 10 years. The median values for age of Female and Male runners will be approximately 35 years and 41 years respectively. The maximum values will be approximately 73 years for Females and 83 years for Male. Note that is maximum value is different from the one that is observed in histograms. The box plot also shows that there are more outliers for the Female group of runners while there is only one outlier for the Male group. This also proves that the spread of data for Male groups is better than that of the Female groups. This shows that there are way more older Males that ran in the Road Race than Female runners. From the boxplots, we can also conclude that the Ages of Female runners exhibit Normal distribution since the length of Q3+1.5IQR is almost same as length of Q1-1.5IQR, however because of the outliers the distribution is Right-skewed. All this was proved when we computed the summary statistics of the same.

```
summary(dataMale$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   9.00   30.00   41.00   40.45   51.00   83.00 
IQR(dataMale$Age)
[1] 21
range(dataMale$Age)
[1]  9 83
summary(dataFemale$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   7.00   28.00   36.00   37.24   46.00   86.00 
range(dataFemale$Age)
[1]  7 86
IQR(dataFemale$Age)
[1] 18
```

<u>Question 2:</u>

Consider the dataset motorcycle.csv posted on eLearning. It contains the number of fatal motorcycle accidents that occurred in each county of South Carolina during 2009. Create a boxplot of data and provide relevant summary statistics. Discuss the features of the data distribution. Identify which counties may be considered outliers. Why might these counties have the highest numbers of motorcycle fatalities in South Carolina?

<u>Answer:</u>

To plot the boxplot for Fatal Motorcycle Accidents in different South Carolina county, we use the boxplot() function. From the obtained boxplot, we can say that lowest value for accidents will be 0. The median values for accidents will be approximately 14. The maximum value will be approximately 45 accidents.

The summary statistics that we obtained shows the following:

```
summary(data$Fatal.Motorcycle.Accidents)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    6.00   13.50   17.02   23.00   60.00
```

The summary statistics shows that that the maximum value is 60 accidents. However, the boxplot detects these values as outliers. We obtain 2 outliers after plotting the boxplot. We can say this because the IQR for the data set is:

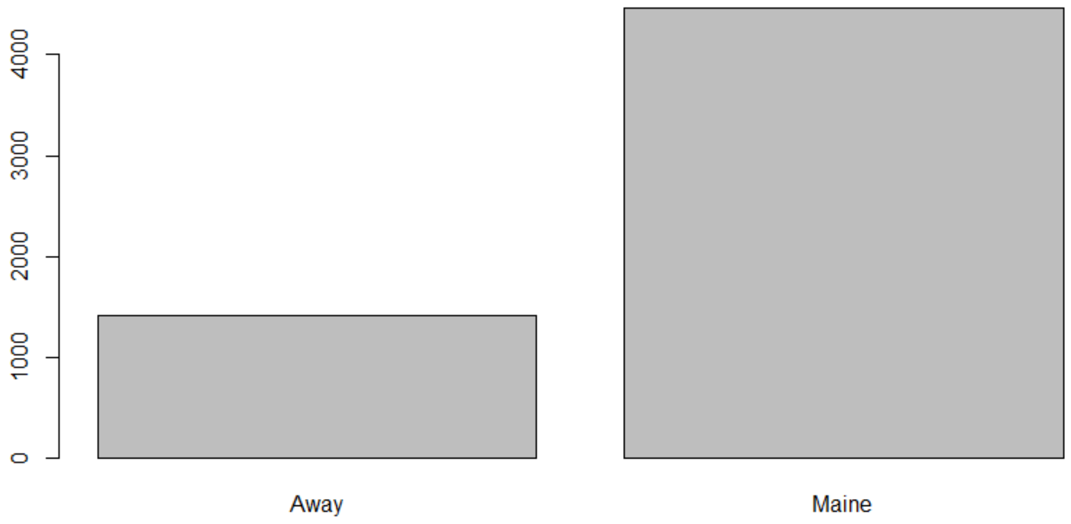
```
IQR(data$Fatal.Motorcycle.Accidents)
[1] 17
```

Therefore all the data points that lie outside Q3+1.5*IQR are detected as outliers by the boxplot. The value of Q3+1.5 *IQR is 48.5
All the data points having value greater than 48.5 are outliers. We see that there are two outliers for the boxplot.
The values of these outliers are 60 and approximately 51. We see from the data set that 60 accidents occur in HORRY county and 51 accidents occur in GREENVILLE county.

To understand the distribution of Fatal accidents, we plot a histogram using the hist() function. After plotting the histogram, we observe that the distribution is a Normal Distribution which is right-skewed.
We can see from the boxplot that the HORRY and GREENVILLE county have maximum number of fatal accidents which proves to show that there are many rash motorcyclists in these counties.

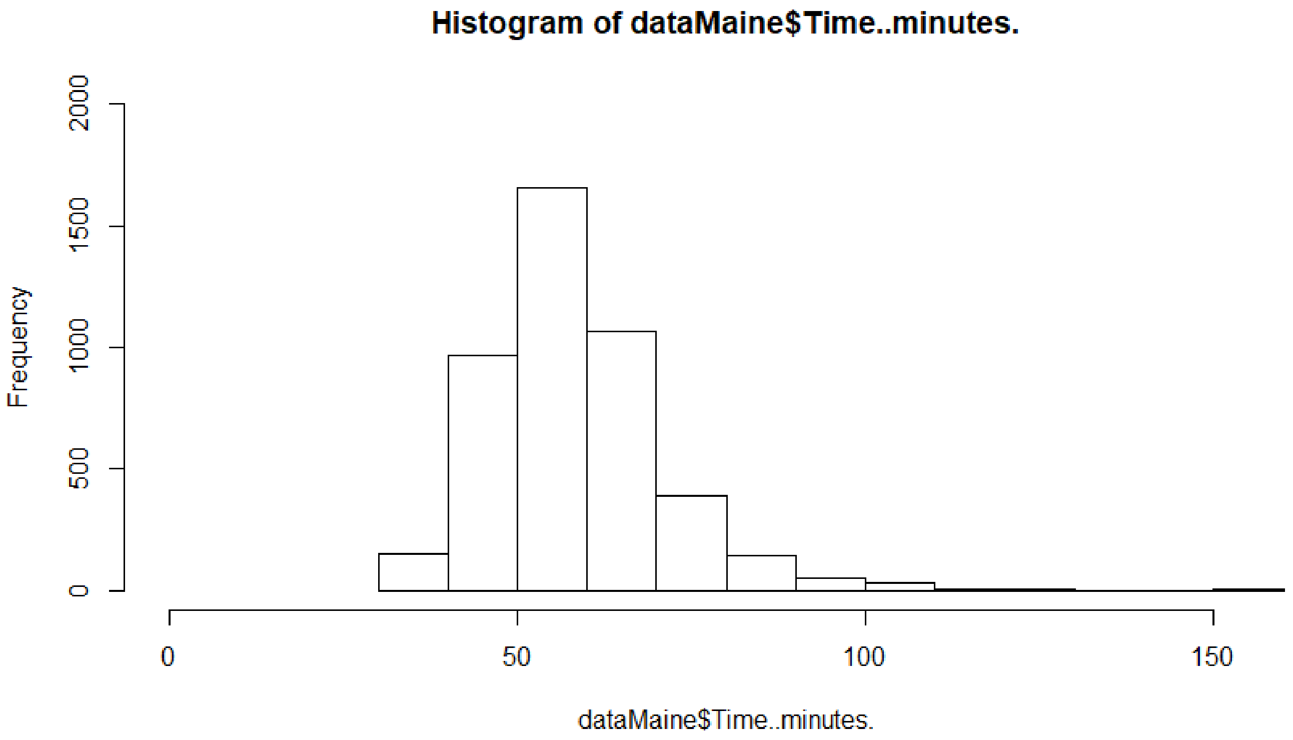
R Code for Q1:

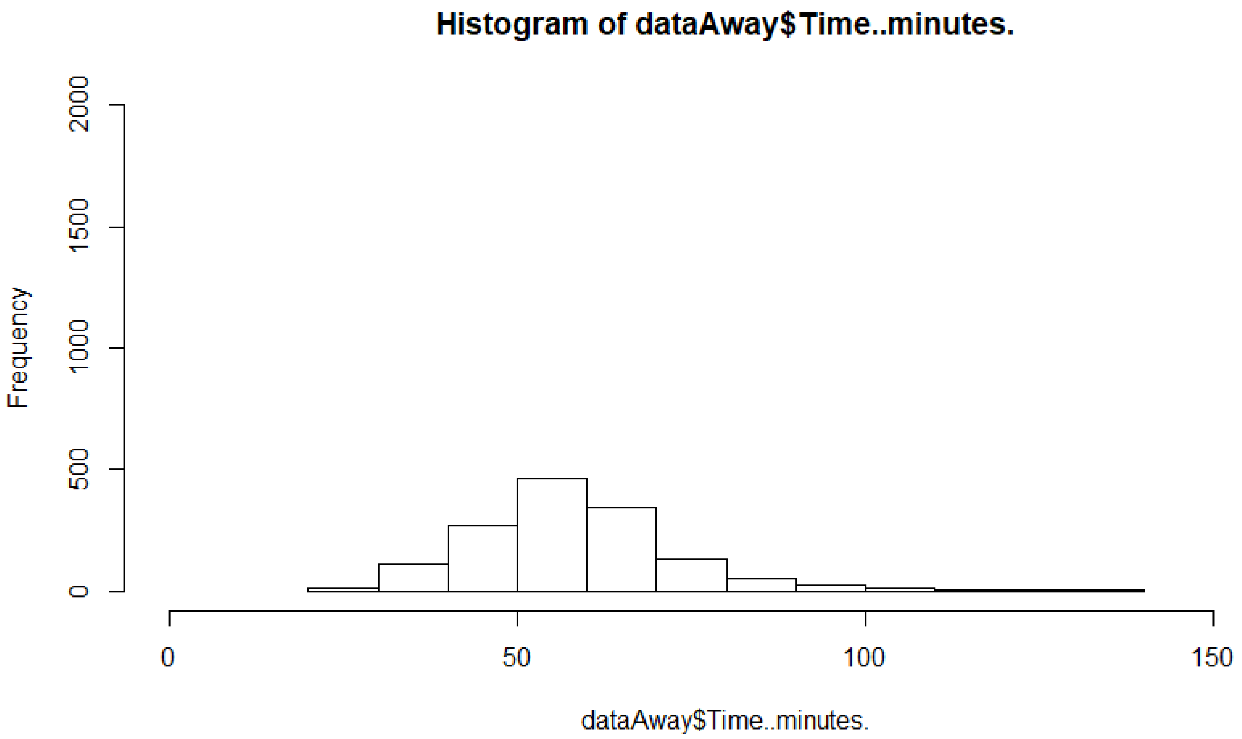
```
A] data = read.csv("roadrace.csv",sep=",")
 Maine = table(data$Maine)
 barplot(Maine)
 summary(data$Maine) Away Maine
                      1417  4458
```



B]

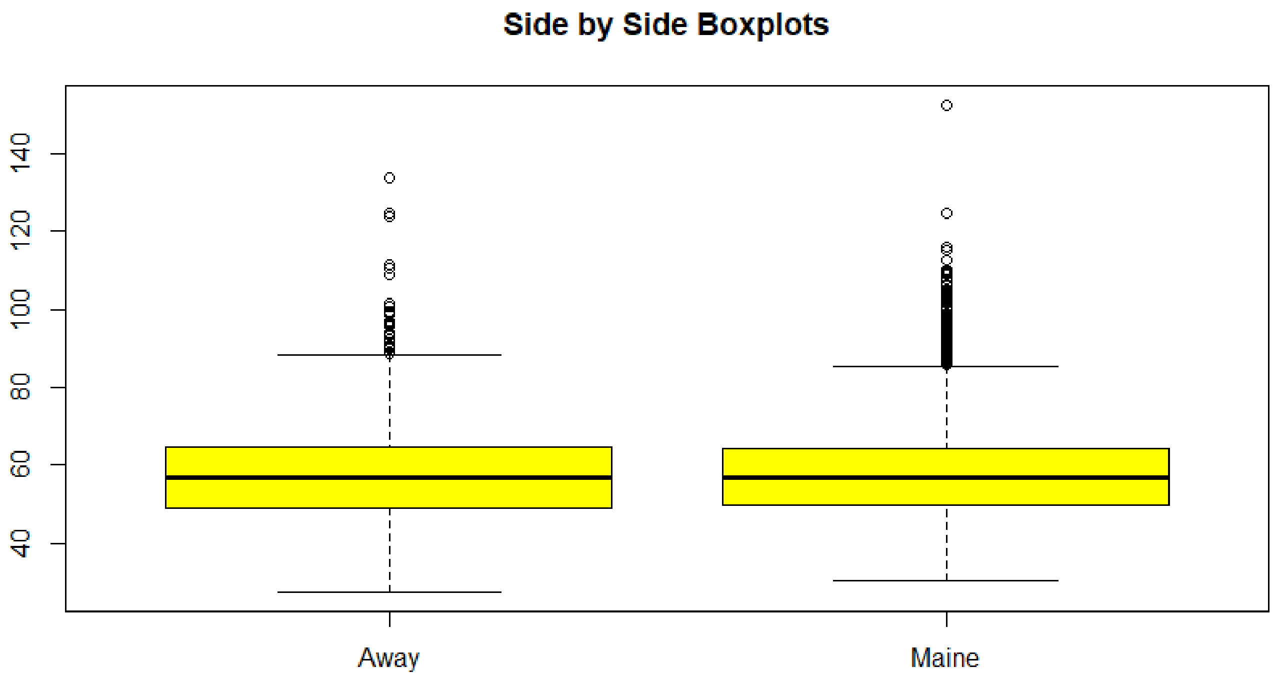
```
data = read.csv("roadrace.csv",sep=",")
dataMaine<-data[which(data$Maine=='Maine'),]
hist(dataMaine$Time..minutes.,ylim=c(0,5000))
hist(dataMaine$Time..minutes.,ylim=c(0,2000))
View(dataMaine)
dataAway<-data[which(data$Maine=='Away'),]
hist(dataAway$Time..minutes.,ylim=c(0,2000))
hist(dataMaine$Time..minutes.,xlim=c(0,160),ylim=c(0,5000))
hist(dataMaine$Time..minutes.,xlim=c(0,160),ylim=c(0,2000))
hist(dataMaine$Time..minutes.,xlim=c(0,160),ylim=c(0,2000))
hist(dataAway$Time..minutes.,xlim=c(0,160),ylim=c(0,2000))
summary(dataMaine$Time..minutes.)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  30.57   50.00   57.03   58.20   64.24  152.17 
summary(dataAway$Time..minutes.)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  27.78   49.15   56.92   57.82   64.83  133.71 
IQR(dataMaine$Time..minutes.)
[1] 14.24775
IQR(dataAway$Time..minutes.)
[1] 15.674
range(dataMaine$Time..minutes.)
[1]  30.567 152.167
range(dataAway$Time..minutes.)
[1]  27.782 133.710
```





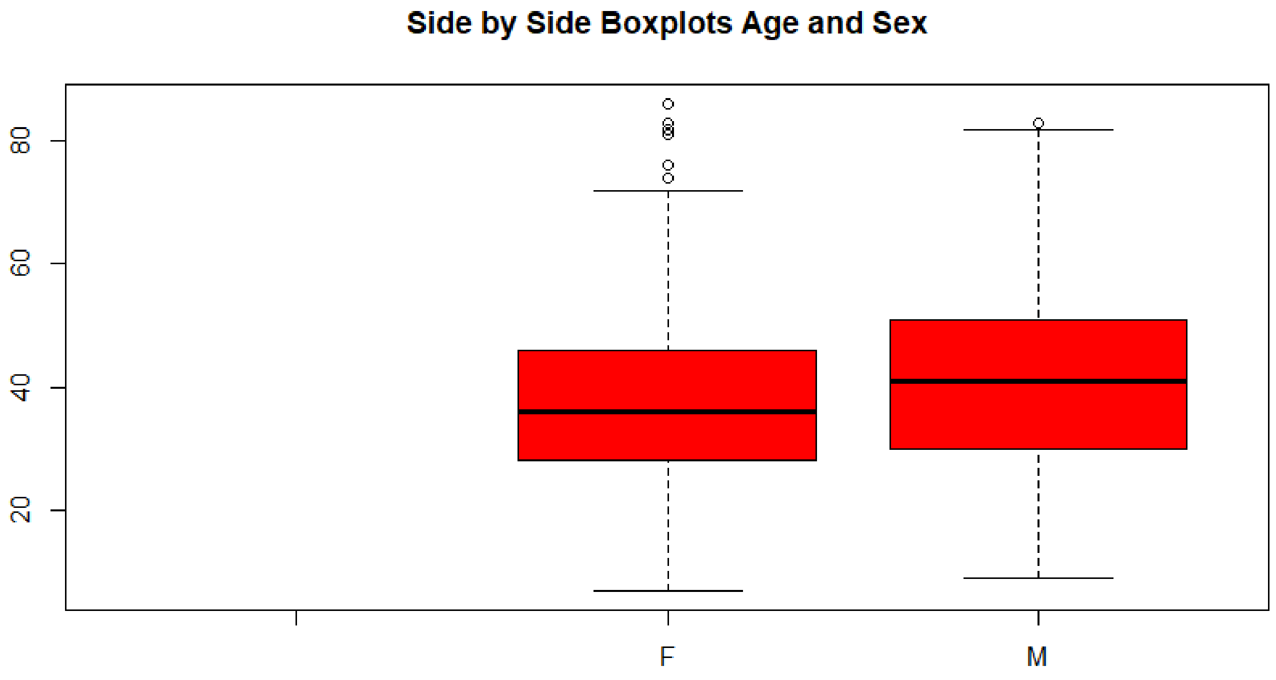
C]

```
setwd("C:/Users/Satyum/Desktop/Statistics for Data Science/Project 2")
data=read.csv("roadrace.csv",sep=",")
boxplot(data$Time..minutes.~ data$Maine, main="Side by Side Boxplots", col="yellow")
```



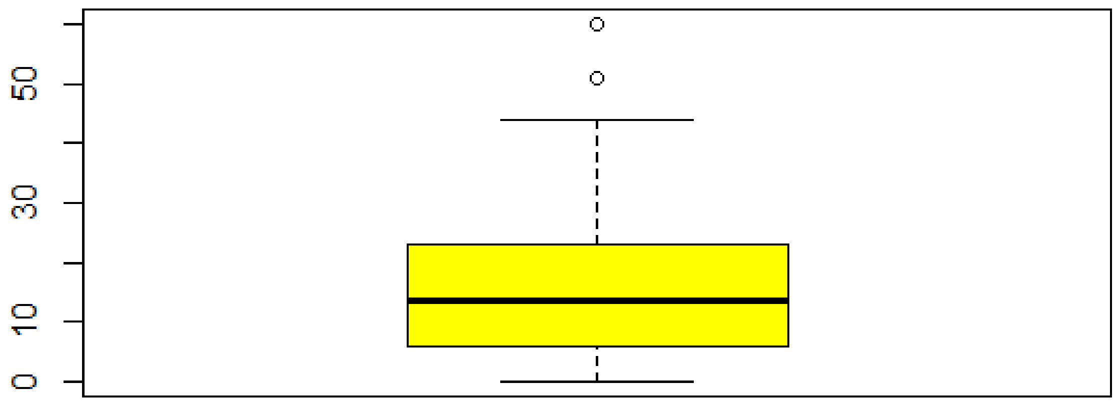
D]

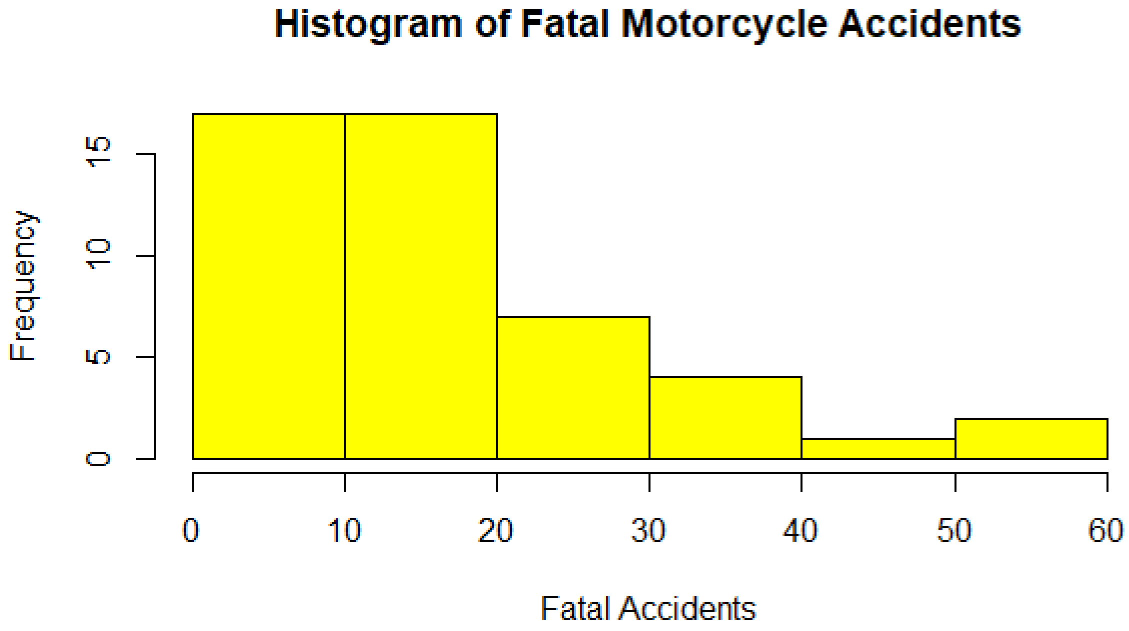
```
setwd("C:/Users/Satyum/Desktop/Statistics for Data Science/Project 2")
data=read.csv("roadrace.csv",sep=",")
boxplot(data$Age~ data$Sex, main="Side by Side Boxplots Age and Sex", col="red")

summary(dataMale$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   30.00   41.00   40.45   51.00   83.00
IQR(dataMale$Age)
[1] 21
range(dataMale$Age)
[1]  9 83
summary(dataFemale$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   28.00   36.00   37.24   46.00   86.00
range(dataFemale$Age)
[1]  7 86
IQR(dataFemale$Age)
[1] 18
```



R Code Q2:

```
setwd("D:/Ajay/Ajay MSCS Semester 1/SMDS/Project 2")
> data=read.csv("motorcycle.csv", sep = ',')
> boxplot(data$Fatal.Motorcycle.Accidents, col = 'yellow')
> summary(data$Fatal.Motorcycle.Accidents)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    6.00   13.50   17.02   23.00   60.00
> IQR(data$Fatal.Motorcycle.Accidents)
[1] 17
hist(data$Fatal.Motorcycle.Accidents, xlab='Fatal Accidents', main='Histogram of Fatal Motorcycle Accidents', col = ' yellow')
```


0
10
30
50


Histogram of Fatal Motorcycle Accidents
Frequency
0
5
10
15
0
10
20
30
40
50
60
Fatal Accidents