# Mini Project 2

Mohammed Allama Hossain, Lavanya Gopal We worked on the questions together. we collaborated and completed the coding part of the assignment. Allama then ran the simulations and Lavanya compiled the results and completed the documentation.

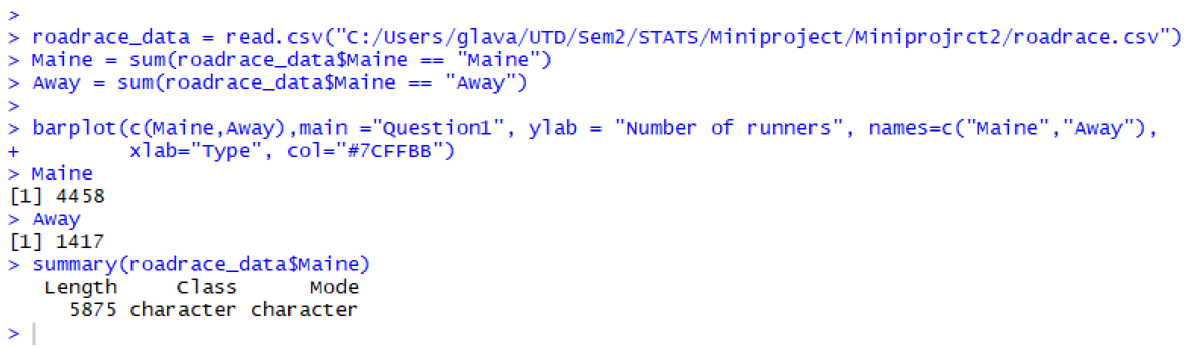
1. Consider the dataset roadrace.csv posted on eLearning. It contains observations on 5875 runners who finished the 2010 Beach to Beacon 10K Road Race in Cape Elizabeth, Maine. You can read the dataset in R using read.csv function.

(a) Create a bar graph of the variable Maine, which identifies 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.

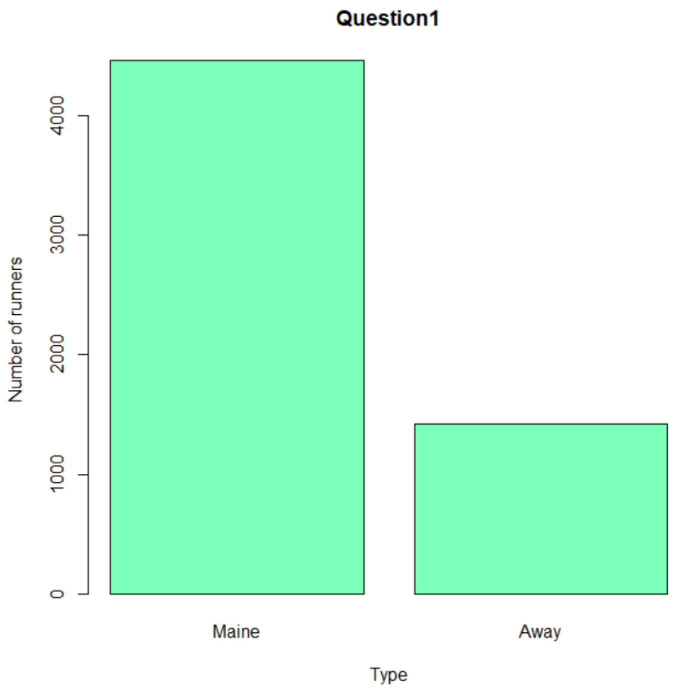
**Code:**

```
# The csv file is read using the built in function “read.csv” and stored the result in the variable called “roadrace_data”
roadrace_data =read.csv("C:/Users/glava/UTD/Sem2/STATS/Miniproject/Miniprojrct2/roadrace.csv")

# Get the Maine and Away column values and add them and store the sum in the variables
Maine = sum(roadrace_data$Maine == "Maine")
Away = sum(roadrace_data$Maine == "Away")

# using the built in function barplot(), a barplot is drawn with Maine and Away data we have already filtered from the csv file
barplot(c(Maine,Away),main ="Question1", ylab = "Number of runners", names=c("Maine","Away"),
    xlab="Type", col="#7CFFBB")
Maine
Away

# printing the summary of the Maine column, it displays the count of the different values in this column, which shows the count of away and maine as shown below.
summary(roadrace_data$Maine)
```

**Output:**



```
>
> roadrace_data = read.csv("C:/Users/glava/UTD/Sem2/STATS/Miniproject/Miniprojrct2/roadrace.csv")
> Maine = sum(roadrace_data$Maine == "Maine")
> Away = sum(roadrace_data$Maine == "Away")
>
> barplot(c(Maine,Away),main ="Question1", ylab = "Number of runners", names=c("Maine","Away"),
+         xlab="Type", col="#7CFFBB")
> Maine
[1] 4458
> Away
[1] 1417
> summary(roadrace_data$Maine)
   Length     Class      Mode 
     5875 character character 
> 
```



**Conclusion:**

- The graph shows that the Maine has more values than the Away and Main comprises of approximately 75% of the total count and Away has around 25% of the total.
- By the stats summary, we can verify by seeing the summary results which shows that, out of the 5875 runners, Maine runner count is 4458, and Away runners are 1417, which are 75.88% and 24.12% of the total respectively.

**(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.

**Code:**

```
# As we have already stored the data from the csv file in the previous question, we are filtering the count column Maine which
has the value as "Away" and converting it to minutes and storing it in the variable called "Away"
away_runner_time = roadrace_data$Time..minutes.[which(roadrace_data$Maine == 'Away')]

# After storing the required values in the variable, we are using a built-in function hist() to get the histogram foe the data and
we have taken the xlim and ylim as same for Away and Maine runners as mentioned in the question.
hist(away_runner_time, main = "Histogram - Runner Time (Away)", xlim = range(0,200), ylim = range(0,2000),
   xlab = "Away Runner Time (minutes)", ylab = "Frequency", border = "black", col = "yellow" )

# Similar to the Away data, we have repeated the same process with the data of Maine and plotted the histogram as shown
below.
```

```
maine_runner_time = roadrace_data$Time..minutes.[which(roadrace_data$Maine == "Maine")]
hist(maine_runner_time, main = "Histogram - Runner Time (Maine)", xlim = range(0,200), ylim = range(0,2000),
   xlab = "Maine Runner Time (minutes)", ylab = "Frequency" , border = "black", col = "yellow")
```

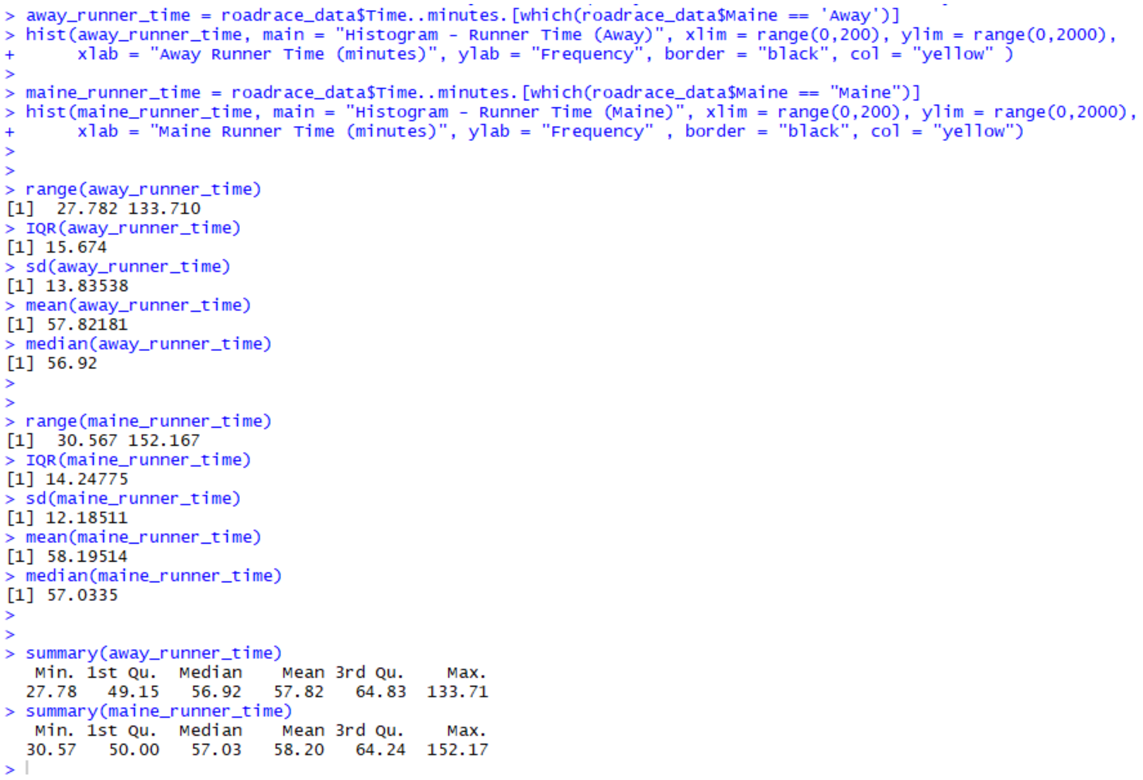
# Next we have print the range, Interquartile range and Standard deviation, mean and median of both the data.

```
range(away_runner_time)
IQR(away_runner_time)
sd(away_runner_time)
mean(away_runner_time)
median(away_runner_time)


range(maine_runner_time)
IQR(maine_runner_time)
sd(maine_runner_time)
mean(maine_runner_time)
median(maine_runner_time)
```

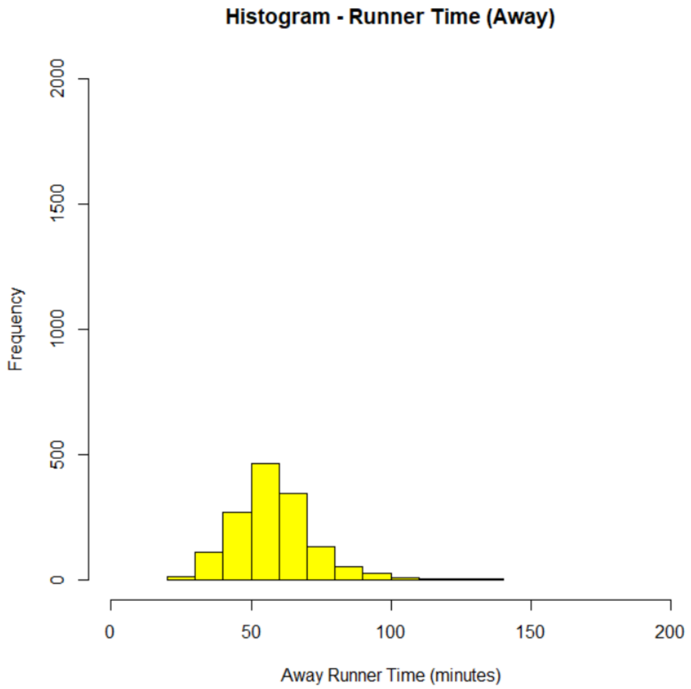
# The summary consisting of mean, median, minimum, maximum and the 1st and 3rd quartile are printed for both the Maine and Away data

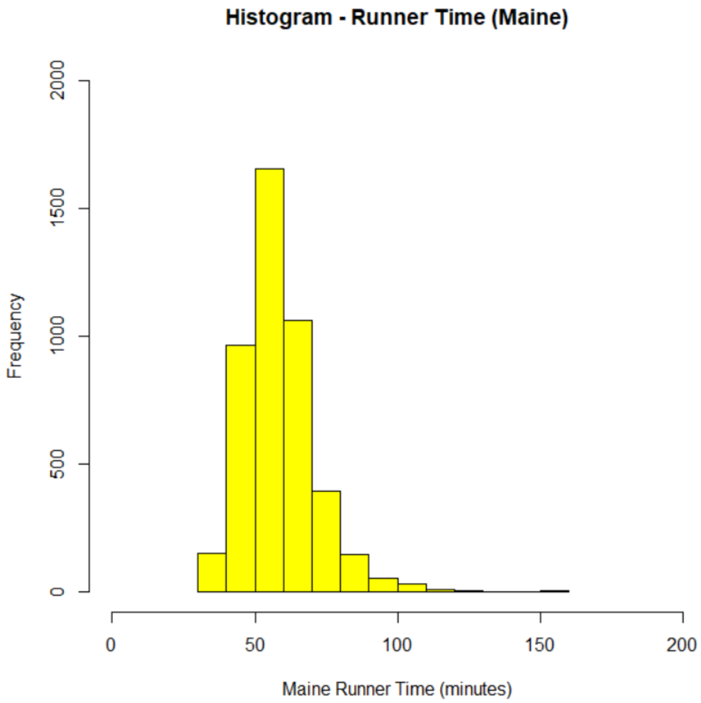
```
summary(away_runner_time)
summary(maine_runner_time)
```

## Output:



```
> away_runner_time = roadrace_data$Time..minutes.[which(roadrace_data$Maine == 'Away')]
> hist(away_runner_time, main = "Histogram - Runner Time (Away)", xlim = range(0,200), ylim = range(0,2000),
+      xlab = "Away Runner Time (minutes)", ylab = "Frequency", border = "black", col = "yellow" )
>
> maine_runner_time = roadrace_data$Time..minutes.[which(roadrace_data$Maine == "Maine")]
> hist(maine_runner_time, main = "Histogram - Runner Time (Maine)", xlim = range(0,200), ylim = range(0,2000),
+      xlab = "Maine Runner Time (minutes)", ylab = "Frequency" , border = "black", col = "yellow")
>
>
> range(away_runner_time)
[1]  27.782 133.710
> IQR(away_runner_time)
[1] 15.674
> sd(away_runner_time)
[1] 13.83538
> mean(away_runner_time)
[1] 57.82181
> median(away_runner_time)
[1] 56.92
>
>
> range(maine_runner_time)
[1]  30.567 152.167
> IQR(maine_runner_time)
[1] 14.24775
> sd(maine_runner_time)
[1] 12.18511
> mean(maine_runner_time)
[1] 58.19514
> median(maine_runner_time)
[1] 57.0335
>
>
> summary(away_runner_time)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  27.78   49.15   56.92   57.82   64.83  133.71 
> summary(maine_runner_time)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  30.57   50.00   57.03   58.20   64.24  152.17 
> 
```


Histogram - Runner Time (Away)
Frequency
0
500
1000
1500
2000
0
50
100
150
200
Away Runner Time (minutes)


Histogram - Runner Time (Maine)
Frequency
0
500
1000
1500
2000
0
50
100
150
200
Maine Runner Time (minutes)

**Conclusion:**

- Through the histogram we can see that the number of runners are more in Maine than Away, however the distribution of the runners is both are similar.
- As per statistics, the mean and median of the Away runners are little less than the ones from the Maine, which indicates that the Runners from Away take less time than Maine runners. Also mean and median are close values for Maine and Away runners.
- The range is bit higher for Maine runners as per the statistics, which can also contribute for some runners taking more time in Maine and thus increasing the mean to incline towards it, however, overall the runners distribution is almost similar in Away and Maine as per both stats and the histogram.

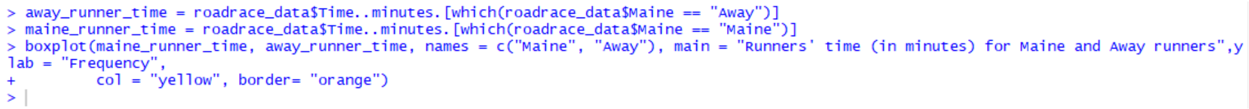
(c) Repeat (b) but with side-by-side boxplots.

**Code:**

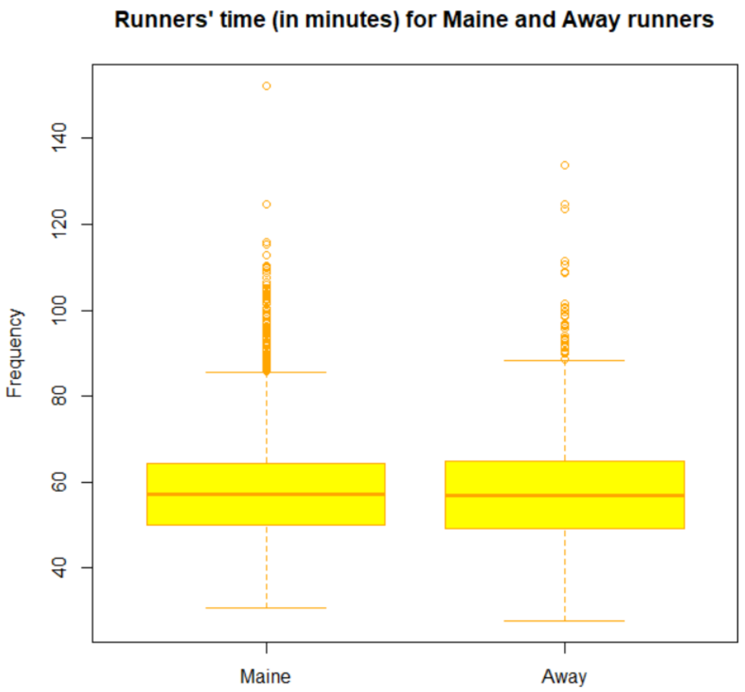
# After filtering the same data of Away and Maine as previous question, we have used the function “boxplot” to plot the graph. We have given both the data to get the plot side by side as shown in the diagram.

```
away_runner_time = roadrace_data$Time..minutes.[which(roadrace_data$Maine == "Away")]
maine_runner_time = roadrace_data$Time..minutes.[which(roadrace_data$Maine == "Maine")]
boxplot(maine_runner_time, away_runner_time, names = c("Maine", "Away"), main = "Runners' time (in minutes) for Maine and Away runners",ylab = "Frequency",
      col = "yellow", border= "orange")
```

**Output:**



```
> away_runner_time = roadrace_data$Time..minutes.[which(roadrace_data$Maine == "Away")]
> maine_runner_time = roadrace_data$Time..minutes.[which(roadrace_data$Maine == "Maine")]
> boxplot(maine_runner_time, away_runner_time, names = c("Maine", "Away"), main = "Runners' time (in minutes) for Maine and Away runners",y
lab = "Frequency",
+         col = "yellow", border= "orange")
> |
```



**Conclusion/Analysis:**

- We have drawn the boxplot for the same data as in previous question (b), so there are no changes to the stats however, we can visualize the different way of graph as shown above.
- This graph also shows that the distribution of both Maine and Away are similar, and mean, median are also very close for Maine and Away we can see as per box plot as well.

(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.

**Code:**

# Get the age values for Male and Female from the file using the which() as shown below. And store the values in the variables.

```
MaleAge = roadrace_data$Age[which(roadrace_data$Sex=='M')]
FemaleAge = roadrace_data$Age[which(roadrace_data$Sex=='F')]
```

# convert the values to the numeric values using the function as.numeric() as shown below to convert the data suitable to plot a boxplot.

```
MaleAge = as.numeric(as.character(MaleAge))
FemaleAge = as.numeric(as.character(FemaleAge))
```

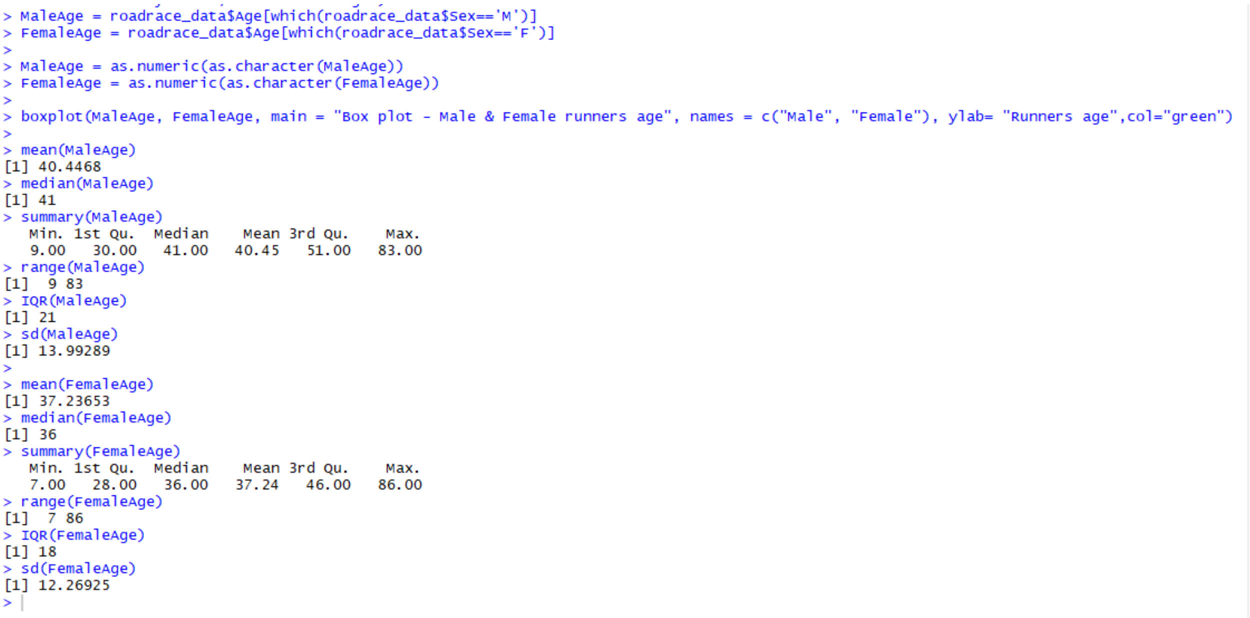
# After getting the required data plot the box plot using the function “boxplot()” as shown below.

```
boxplot(MaleAge, FemaleAge, main = "Box plot - Male & Female runners age", names = c("Male", "Female"), ylab= "Runners age",col="green")
```

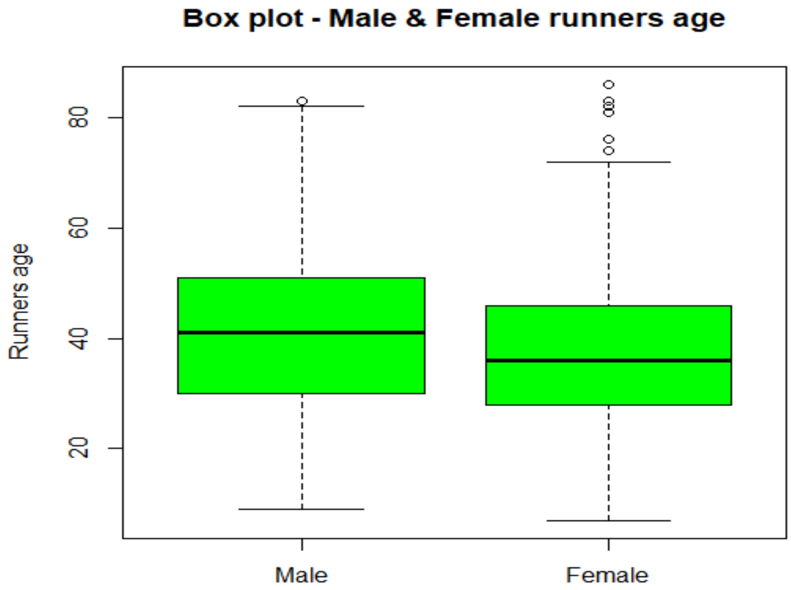
# Use the following functions to get the statistics of the data: like, mean, median, summary, range, IQR and standard deviation.

```
mean(MaleAge)
median(MaleAge)
summary(MaleAge)
range(MaleAge)
IQR(MaleAge)
sd(MaleAge)

mean(FemaleAge)
median(FemaleAge)
summary(FemaleAge)
range(FemaleAge)
IQR(FemaleAge)
sd(FemaleAge)
```

**Output:**



```
> MaleAge = roadrace_data$Age[which(roadrace_data$Sex=='M')]
> FemaleAge = roadrace_data$Age[which(roadrace_data$Sex=='F')]
>
> MaleAge = as.numeric(as.character(MaleAge))
> FemaleAge = as.numeric(as.character(FemaleAge))
>
> boxplot(MaleAge, FemaleAge, main = "Box plot - Male & Female runners age", names = c("Male", "Female"), ylab= "Runners age",col="green")
>
> mean(MaleAge)
[1] 40.4468
> median(MaleAge)
[1] 41
> summary(MaleAge)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   30.00   41.00   40.45   51.00   83.00
> range(MaleAge)
[1]  9 83
> IQR(MaleAge)
[1] 21
> sd(MaleAge)
[1] 13.99289
>
> mean(FemaleAge)
[1] 37.23653
> median(FemaleAge)
[1] 36
> summary(FemaleAge)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   28.00   36.00   37.24   46.00   86.00
> range(FemaleAge)
[1]  7 86
> IQR(FemaleAge)
[1] 18
> sd(FemaleAge)
[1] 12.26925
> 
```



**Conclusion/analysis:**

- By looking at the various metrics we can conclude that the male population participating in the race is on average older than the female population participating in the race with outliers in the female population as seen from the fact that three out of the five oldest participants are female. The mentioned inference can be delineated by the mean and median of the male population which is 40.45 and 41 respectively as compared to the mean and median of the female population which is 37.24 and 36 respectively.
-
- The inter-quartile range is also higher for male population. The first quartile of male population participating in the race is 30 as compared to 28 for the female population participating in the race, and it 51 and 46 respectively for the third quartile of the male and female participants.
-
- The range of female participant age is higher than that of the male participant age as the youngest and oldest participants were both female. The range for female age is 79 and the range for male age is 74.

2. (8 points) 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?

## Code:

# Read the data from the csv file using the “read.csv()” function. Then filtering the required accidents count data and stored in a variable

```
data1 = read.csv("C:/Users/glava/UTD/Sem2/STATS/Miniproject/Miniprojrct2/motorcycle.csv")
FatalAccidents = data1$Fatal.Motorcycle.Accidents
```

# using the built-in boxplot() function, we have plotted the details of the FatalAccidents as shown in the diagram below.

```
boxplot(FatalAccidents, main = "BoxPlot - Motorcycle Accidents", xlab = "Fatal Accidents", ylab = "Accidents count",
col="orange")
```

# Calculated the lower bound and upper bound and got the county details as shown in the output.

```
LowerBound = max(quantile(FatalAccidents, prob=0.25)- 1.5*IQR(FatalAccidents),min(FatalAccidents))
UpperBound=min(quantile(FatalAccidents, prob=0.75) + 1.5*IQR(FatalAccidents),max(FatalAccidents))
```

#Outlier details

```
FatalCounty=data1$County[which(data1$Fatal.Motorcycle.Accidents<LowerBound | data1$Fatal.Motorcycle.Accidents >
UpperBound)]
FatalCounty
```

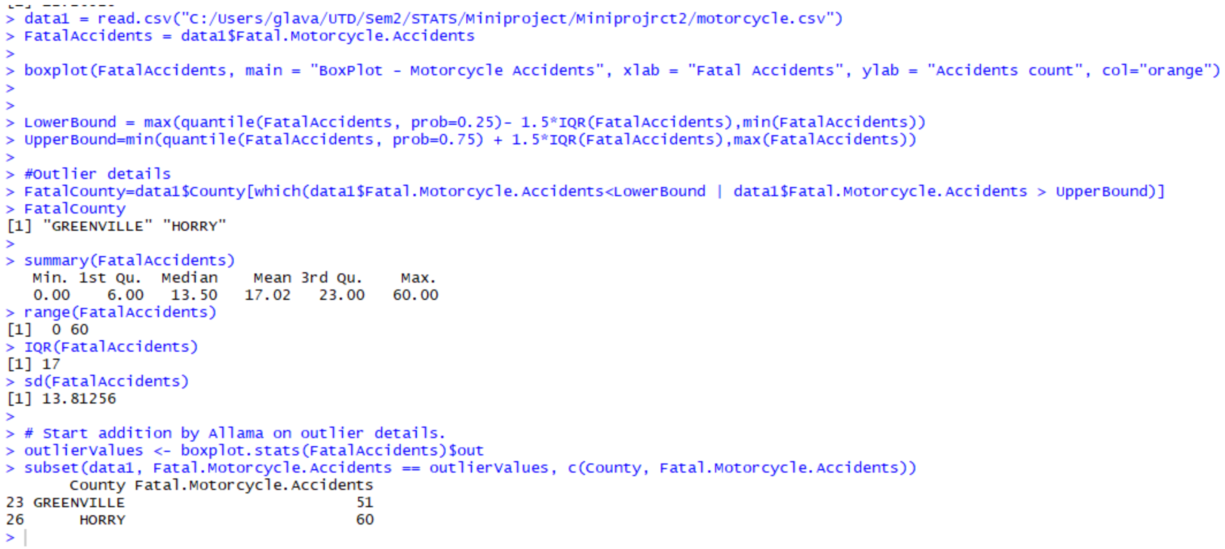
# summary, range, IQR and standard deviation of the Fatal Accidents are calculated.

```
summary(FatalAccidents)
range(FatalAccidents)
IQR(FatalAccidents)
sd(FatalAccidents)
```

# Start addition by Allama on outlier details.

```
outlierValues <- boxplot.stats(FatalAccidents)$out
subset(data1, Fatal.Motorcycle.Accidents == outlierValues, c(County, Fatal.Motorcycle.Accidents))
```

## Output:

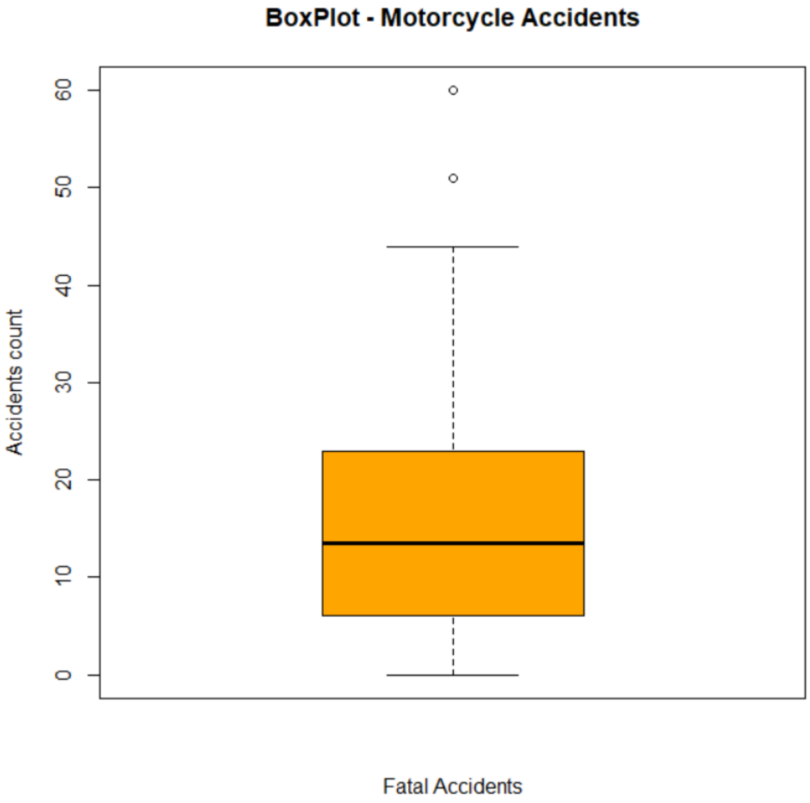


```
> data1 = read.csv("C:/Users/glava/UTD/Sem2/STATS/Miniproject/Miniprojrct2/motorcycle.csv")
> FatalAccidents = data1$Fatal.Motorcycle.Accidents
> 
> boxplot(FatalAccidents, main = "BoxPlot - Motorcycle Accidents", xlab = "Fatal Accidents", ylab = "Accidents count", col="orange")
> 
> 
> LowerBound = max(quantile(FatalAccidents, prob=0.25)- 1.5*IQR(FatalAccidents),min(FatalAccidents))
> UpperBound=min(quantile(FatalAccidents, prob=0.75) + 1.5*IQR(FatalAccidents),max(FatalAccidents))
> 
> #Outlier details
> FatalCounty=data1$County[which(data1$Fatal.Motorcycle.Accidents<LowerBound | data1$Fatal.Motorcycle.Accidents > UpperBound)]
> FatalCounty
[1] "GREENVILLE" "HORRY"
> 
> summary(FatalAccidents)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    6.00   13.50   17.02   23.00   60.00 
> range(FatalAccidents)
[1]  0 60
> IQR(FatalAccidents)
[1] 17
> sd(FatalAccidents)
[1] 13.81256
> 
> # Start addition by Allama on outlier details.
> outlierValues <- boxplot.stats(FatalAccidents)$out
> subset(data1, Fatal.Motorcycle.Accidents == outlierValues, c(County, Fatal.Motorcycle.Accidents))
       County Fatal.Motorcycle.Accidents
23 GREENVILLE                         51
26      HORRY                         60
> 
```

## Conclusion/analysis:

- We have found the 1st and 3rd quartile to find the outliers( if any), using the formulas.

- Lower bound : probability(0.25) – (1.5* IQR)
- Upper bound : probability(0.75) – (1.5* IQR)
- After that we have found the outlier which is Greenville and Horry, having 51 and 60 accidents which is the greatest among all counties. We have also listed the max, min, quartiles and Standard deviation of the data.


BoxPlot - Motorcycle Accidents
Accidents count
Fatal Accidents