Mini Project 02

Mini Project Group 16

Group Members

Dhairya Pinakin Desai: DPD140130

Jaya Venkata Sai Krishna Kosana : JXK220036

Krishna Sai Gandi : KXG220038

Contribution: The three team members collaborated seamlessly, each putting in an equal amount of effort to complete the assigned mini project.

Q1)

We are reading the csv file using read.csv function of R and then storing it in a variable "rdrace".

```
> rdrace <-
read.csv("C:\\Users\\dpd140130\\OneDrive - The
University of Texas at Dallas\\CS
6313\\Projects\\02\\Mini-Proj-2\\roadrace.csv")
```

a) Now according to the question to plot the bar graph of the column Maine from the csv file. For that we will use the barplot function of R to show a bar graph of the players who are from Maine and away from Maine.

```
> b_plot <- barplot(maine, main = "Runners
Distribution", xlab = "Maine or Away", ylab =
"Number of Players",ylim = c(0,5000),
col="skyblue")
```

- From the bar graph, we can conclude that the majority of the runners are from Maine. Approximately, 75.88% of the runners are from Maine and the rest are from different (Away) places.
- We can make this conclusion based on the following summary:

```
> rdrace_maine = table (rdrace$Maine)
> rdrace_maine

 Away Maine
 1417  4458
```

So we can conclude that bar chart matches the summary statistics of the variable Maine.

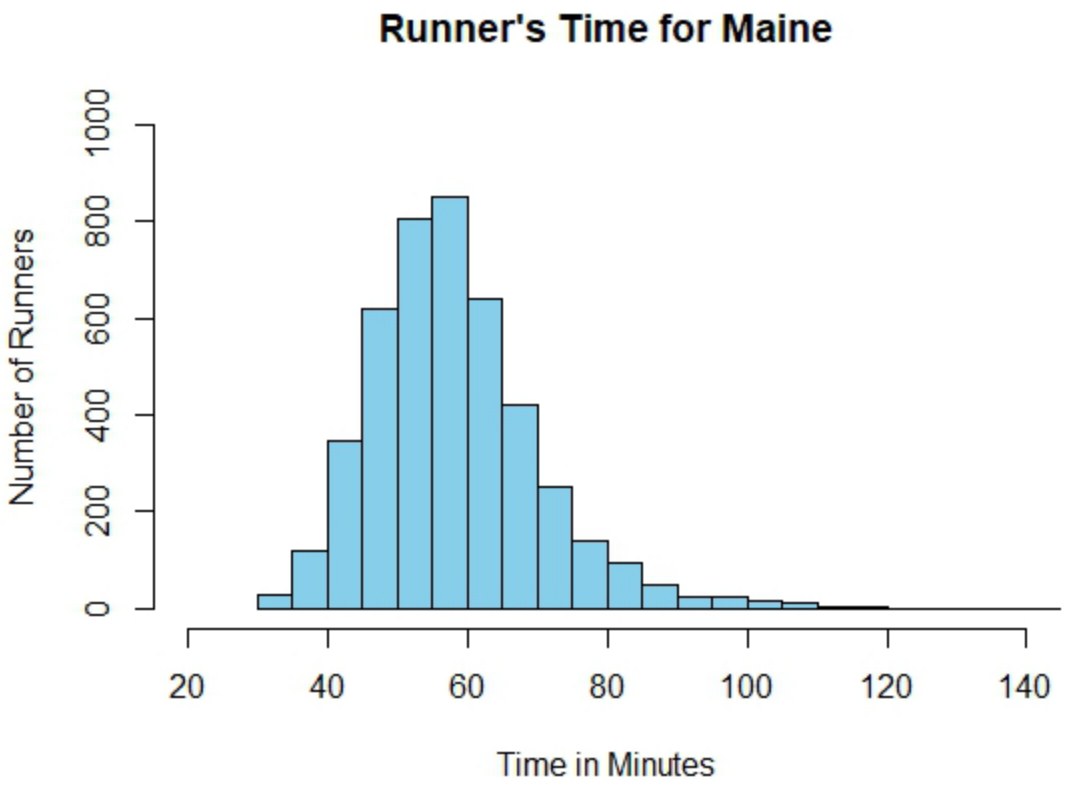
R Code:

```
> #Read the csv file
> rdrace <- read.csv("C:\\Users\\dpd140130\\OneDrive
- The University of Texas at Dallas\\CS
6313\\Projects\\02\\Mini-Proj-2\\roadrace.csv")
>
> #Q1.a - Create a barplot of Maine
> b_plot <- barplot(maine, main = "Runners
Distribution", xlab = "Maine or Away", ylab = "Number
of Players",ylim = c(0,5000), col="skyblue")
> rdrace_maine = table (rdrace$Maine)
> rdrace_maine
```

b) Now to calculate the runner's time in minutes for Maine we use the following commands:

```
> rt_maine <- rdrace[rdrace$Maine=='Maine',]
```

```
> hist(rt_maine$Time..minutes.,
breaks=20,col="skyblue", main="Runner's Time for
Maine" , xlab = "Time in Minutes" , ylab = "Number of
Runners", xlim=c(20,140),ylim=c(0,1000))
```

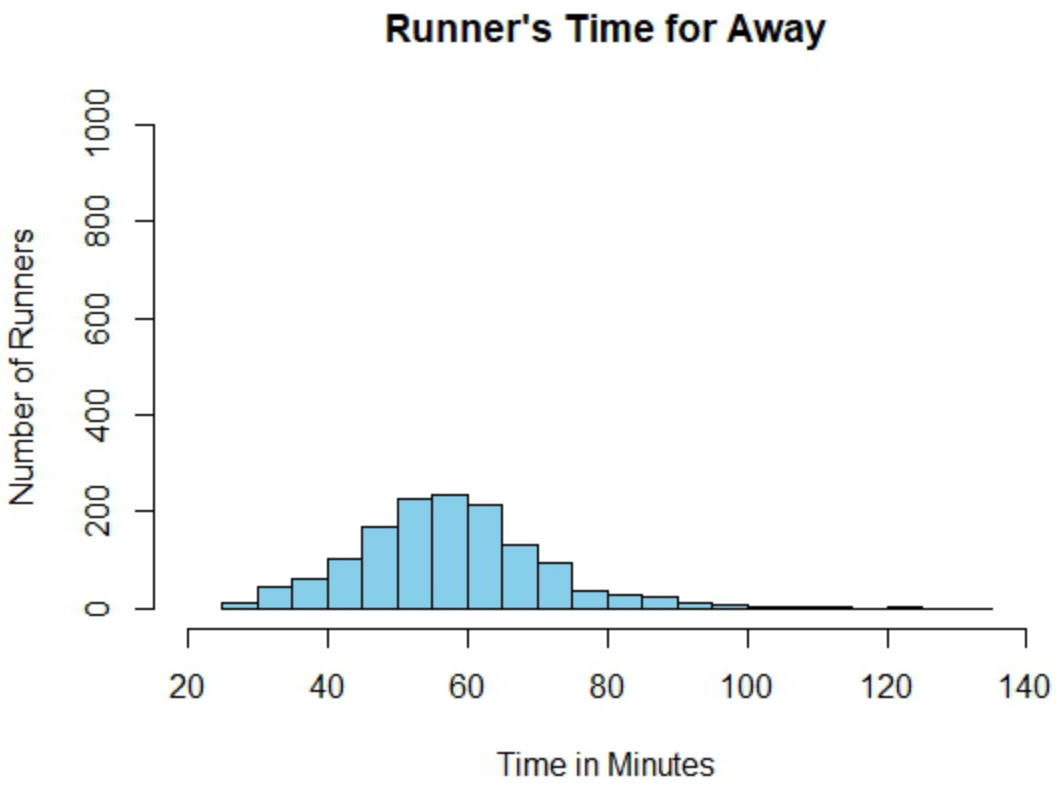


- From the histogram plot, we can see that the distribution is slightly skewed to the right; so, we can conclude that it is a **Right – Skewed Distribution** because its peak is slightly off center and its tail is stretching towards the right.

Now to calculate the runner's time in minutes for "Away from Maine" we use the following R commands:

```
> rt_away <- rdrace[rdrace$Maine=='Away',]

> hist(rt_away$Time..minutes., breaks=20,col="skyblue",
main="Runner's Time for Away" , xlab = "Time in
Minutes" , ylab = "Number of Runners", xlim=c(20,140),
ylim=c(0,1000))
```



- From the histogram plot, we can see that the plot for the time taken by runners who are not from Maine (Away), has a much better symmetry than the one above and forms a **bell-shaped normal distribution curve**. However, it's not entirely bell shaped, instead it is slightly skewed to the right as the tail is stretching to the right. So, it is a **slightly right skewed distribution**.

To back up the above conclusions the following summary can be helpful:

Summary Statistics:

- Summary statistics containing times taken by all runners who are from Maine:

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

- Summary statistics containing times taken by all runners who are not from Maine (Away):

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

- Standard deviation for maine runners

```
> sd(rt_maine$Time..minutes.)
```

[1] 12.18511

- Standard deviation for away runners

```
> sd(rt_away$Time..minutes.)
```

[1] 13.83538

- Gives range for maine runners

```
> diff(range(rt_maine$Time..minutes.))
```

[1] 121.6

- Gives range for away runners

```
> diff(range(rt_away$Time..minutes.))
```

[1] 105.928

- Gives inter quartile range for maine runners

```
> IQR(rt_maine$Time..minutes.)
```

[1] 14.24775

- Gives inter quartile range for away runners

```
> IQR(rt_away$Time..minutes.)
```

[1] 15.674

We can see that the difference between Min and Median from the summary statistics is less than the difference between Max and Median. So, the conclusion about the right skewed distribution is verified.

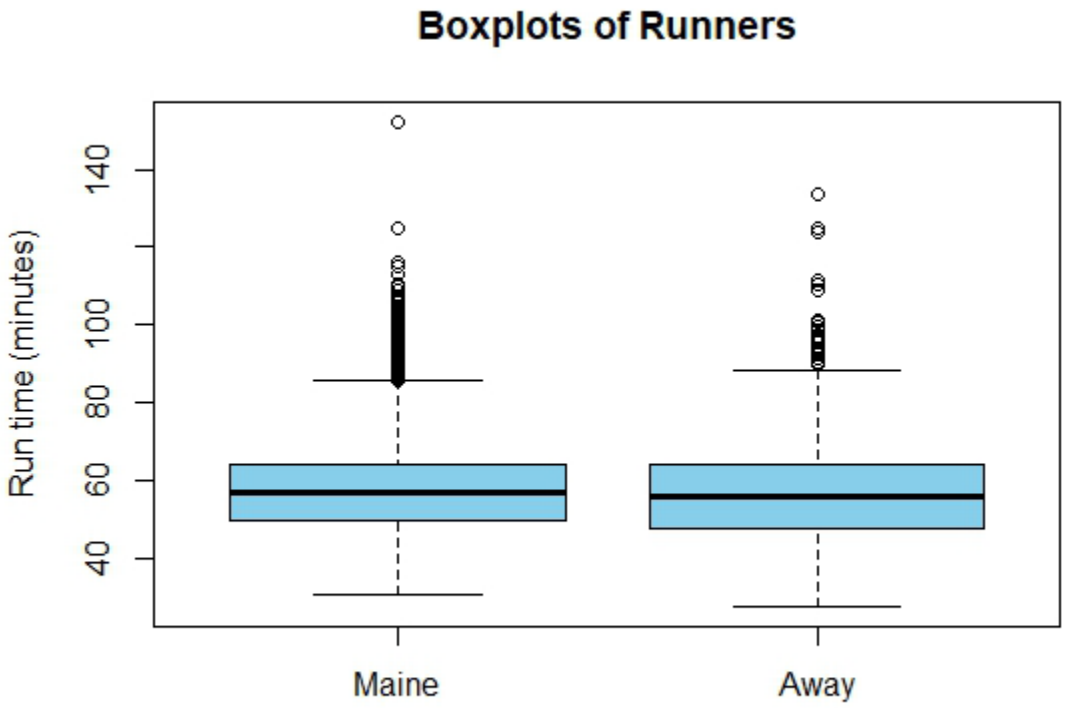
R Code:

```
> #Q1.b - Creating histograms of running times for
Runners from Maine/Away.
> rt_maine <- rdrace[rdrace$Maine=='Maine',]
> hist(rt_maine$Time..minutes.,
breaks=20,col="skyblue", main="Runner's Time for
Maine" , xlab = "Time in Minutes" , ylab = "Number of
Runners",xlim=c(20,140),ylim=c(0,1000))
> rt_away <- rdrace[rdrace$Maine=='Away',]
> hist(rt_away$Time..minutes.,
breaks=20,col="skyblue", main="Runner's Time for
Away" , xlab = "Time in Minutes" , ylab = "Number of
Runners",xlim=c(20,140),ylim=c(0,1000))

> #Q1.b - Summary Statistics
> summary(rt_maine$Time..minutes
> summary(rt_away$Time..minutes.)
> sd(rt_maine$Time..minutes.)
> sd(rt_away$Time..minutes.)
> diff(range(rt_maine$Time..minutes.))
> diff(range(rt_away$Time..minutes.))
> IQR(rt_maine$Time..minutes.)
> IQR(rt_away$Time..minutes.)
```

(c) Repeating the step (b) using side by side box plots

```
> bxplt_runrs <-
    cbind("Maine"=(rt_maine$Time..minutes.),"Away"=(rt_away$Tim
    e..minutes.))
> boxplot(bxplt_runrs, beside=T, col="skyblue",
    horizontal=T,xlab="Run time (minutes)",main="Boxplots of
    Runners")
```

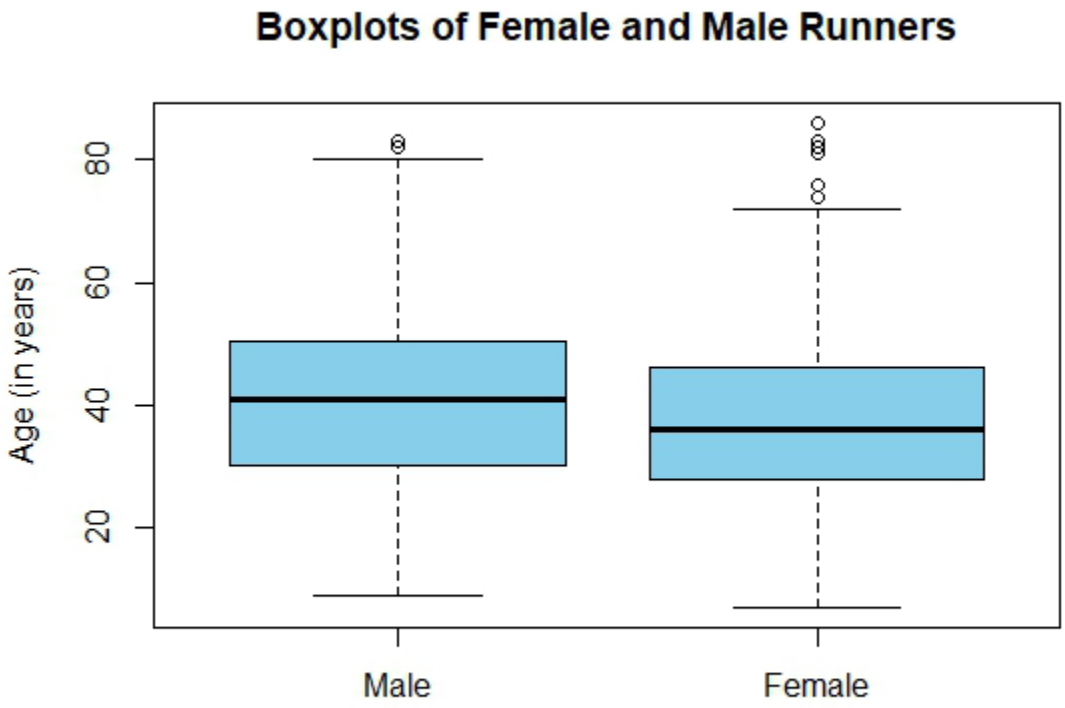


R Code:

```
> #Q1.c - Creating Box plots for both Runner
categories.
> bxplt_runrs <-
cbind("Maine"=(rt_maine$Time..minutes.),"Away"=(rt_aw
ay$Time..minutes.))
> boxplot(bxplt_runrs, beside=T, col="skyblue",
horizontal=T,xlab="Run time (minutes)",main="Boxplots
of Runners")
```

(d) Now following is the side by side boxplot for the runner's ages for male and female runner.

```
> male_run <- rdrace[rdrace$Sex=='M',]
> female_run <- rdrace[rdrace$Sex=='F',]
> age_run <-
      cbind("Male"=(as.numeric(male_run$Age)),
      "Female"=(as.numeric(female_run$Age)))
> boxplot(mfrunner, beside=T, col="skyblue",
      horizontal=T,xlab="Age (in
      years)",main="Boxplots of Female and
      Male Runners")
```



- The above side by side boxplot shows that both distributions are Right-Skewed Distribution as both have (Max-Median) > (Median-Min).

Summary Statistics:

```
> summary(as.numeric(male_run$Age))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   30.00   41.00   40.45   51.00   83.00

> summary(as.numeric(female_run$Age))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   28.00   36.00   37.24   46.00   86.00

> sd(as.numeric(male_run$Age))
[1] 13.99289

> sd(as.numeric(female_run$Age))
[1] 12.26925

> diff(range(as.numeric(male_run$Age)))
[1] 74

> diff(range(as.numeric(female_run$Age)))
[1] 79

> IQR(as.numeric(male_run$Age))
[1] 21

> IQR(as.numeric(female_run$Age))
[1] 18
```

- From the statistics calculated above, we can conclude that for male runners,
  Median – Min = 41-9 = 32
  Max – Median = 83-41 = 42

  Therefore, (Max - Median) > (Median - Min) and thus the male runner's data is a Right Skewed Distribution.

- Similarly, for the female runnres,

Median – Min = 36 -7 = 29
Max – Median = 86 – 36 = 50

Therefore, (Max – Median) > (Median - Min) and thus the female runner's data is also a Right Skewed Distribution.

R Code:

```
> #Q1.d - side by side boxplot for the runner's ages
for male and female runner
> male_run <- rdrace[rdrace$Sex=='M',]
> female_run <- rdrace[rdrace$Sex=='F',]
> age_run <-
cbind("Male"=(as.numeric(male_run$Age)),"Female"=(as.
numeric(female_run$Age)))
> boxplot(age_run, beside=T, col="skyblue",
horizontal=T,xlab="Age (in years)",main="Boxplots of
Female and Male Runners")

> #Q1.d - Summary Statistics
> summary(as.numeric(male_run$Age))
> summary(as.numeric(female_run$Age))
> sd(as.numeric(male_run$Age))
> sd(as.numeric(female_run$Age))
> diff(range(as.numeric(male_run$Age)))
> diff(range(as.numeric(female_run$Age)))
> IQR(as.numeric(male_run$Age))
> IQR(as.numeric(female_run$Age))
```

Q2)

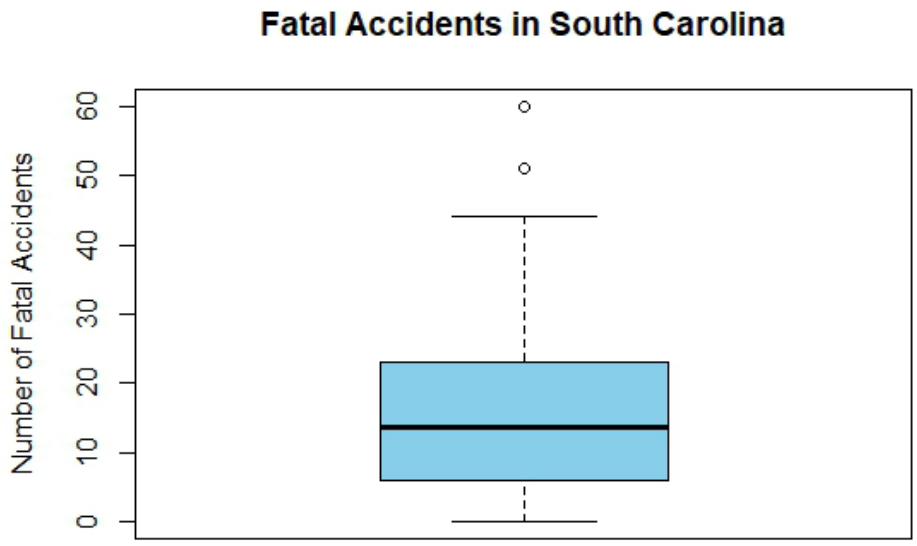
The dataset motorcycle.csv contains the number of fatal motorcycle accidents that occurred in each county of South Carolina during 2009.

First, we need to read the csv file using read.csv function of R and then storing it in a variable "mtcycle".

```
> mtcycle <- read.csv("C:\\Users\\dpd140130\\OneDrive - The University of Texas at Dallas\\CS 6313\\Projects\\02\\Mini-Proj-2\\motorcycle.csv")
```

Now to plot the data for the number of accidents by each county we are using boxplot function in R.

```
>boxplot(mtcycle$Fatal.Motorcycle.Accidents,col="skyblue", horizontal=T,xlab="Number of Fatal Accidents",main="Fatal Accidents in South Carolina")
```



- From the boxplot we can see that the distribution is **Right Skewed**.

To support the above conclusion following is the summary statistics.

```
> summary_stat <-
summary(mtcycle$Fatal.Motorcycle.Accidents)

> summary_stat

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    6.00   13.50   17.02   23.00   60.00
```

- Here we can see that the (Median – Min) < (Max – Median) which verifies the Right Skewed Distribution of data.
- Moreover, summary statistics shows that that on an average(median), there are 13.50 motorcycle fatal accidents in South Carolina.

Now we need to find the names of the countries which are outliers. For that we need to find the inter quartile range and from that we can find the outliers.

```
> int_qurt_rng <-
IQR(mtcycle$Fatal.Motorcycle.Accidents)

> int_qurt_rng

[1] 17

>mtcycle[mtcycle$Fatal.Motorcycle.Accidents>((int_qur
t_rng)*1.5+summary_stat[5]),]

       County Fatal.Motorcycle.Accidents

23 GREENVILLE                         51

26      HORRY                         60
```

After applying the outlier detection, we can find that the counties 'Greenville' and 'Horry' can be considered outliers as the number of fatal motorcycle accidents in these counties are at extreme (above 50).

- As per the data, 'Greenville' and 'Horry' counties might have the highest number of fatalities in South Carolina because traffic rules might not be followed strictly in those counties or there might be administrative hinderances.

R Code:

```
> #Read the csv file
> mtcycle <- read.csv("C:\\Users\\dpd140130\\OneDrive
- The University of Texas at Dallas\\CS
6313\\Projects\\02\\Mini-Proj-2\\motorcycle.csv")

> #Box plot of the fatal accidents in South Carolina
> boxplot
(mtcycle$Fatal.Motorcycle.Accidents,col="skyblue",
horizontal=T,xlab="Number of Fatal
Accidents",main="Fatal Accidents in South Carolina")

> #Summary Statistics
> summary_stat <-
summary(mtcycle$Fatal.Motorcycle.Accidents)
> summary_stat

> #To find outliers
> int_qurt_rng <-
IQR(mtcycle$Fatal.Motorcycle.Accidents)
> int_qurt_rng
> mtcycle
[mtcycle$Fatal.Motorcycle.Accidents>((int_qurt_rng)*1
.5+summary_stat[5]),]
```