Mini Project #2

Group 14.

Group Members:

Manav Gupta [MXG220027]

Shalin Ronakkumar Kaji [SXK220263]

Shivani Talatam [SXT220047]

Contribution:

Manav : Solved Question 2.

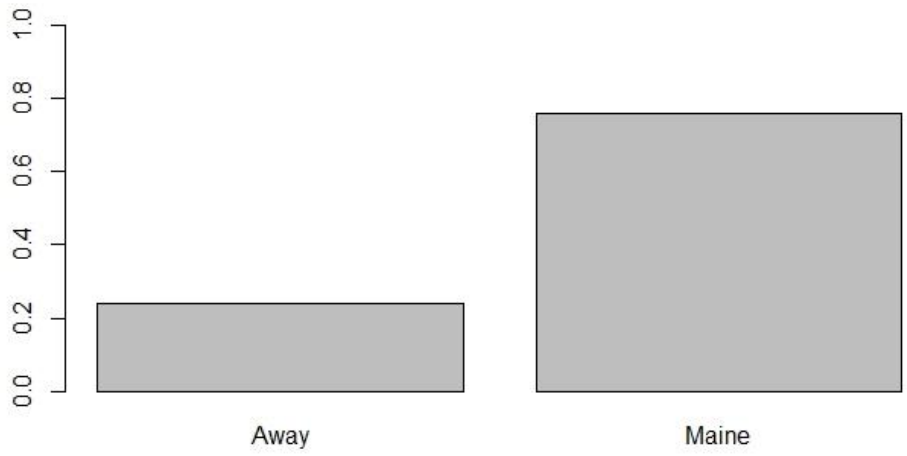
Shalin : Solved Question 1 – a, b, d and prepared the final draft of the documentation.

Shivani : Solved Question 1 – c.

Question 1:

a) Creating a Bar Plot for Runners on basis of the categorical variable of Maine attribute: Maine/Away.

```
> barplot(table(rdrace$Maine)/length(rdrace$Maine), ylim = c(0,1))
```



We can conclude after analysing the Bar Plot:

> To conclude: No. of runners from Maine is thrice the no. of runners from other cities.

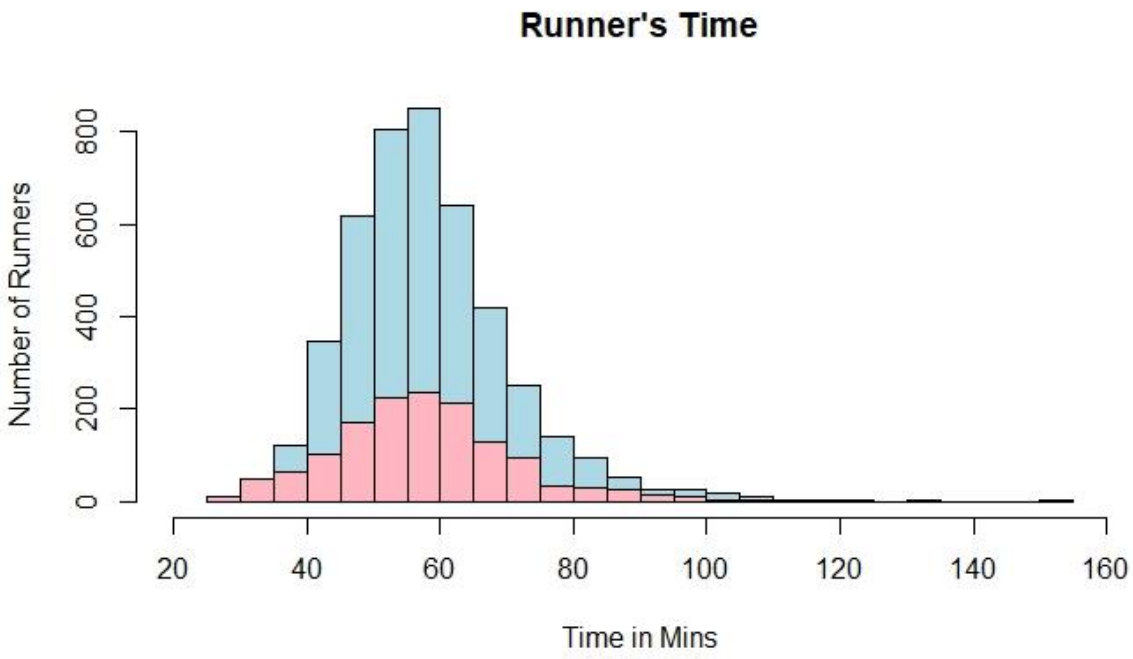
The following summary statistics support our conclusion:

```
> print('Probabilities of runners from Maine/Away: ')
table(rdrace$Maine)/length(rdrace$Maine)
```

Away	Maine
0.2411915	0.7588085

b) On creating Histogram plots for both the 'Maine' and 'Away' group of runners we get:

```
> p1 <- hist(fr_Maine$Time..minutes., breaks=20)
> p2 <- hist(fr_Away$Time..minutes., breaks=20)
```




```
> desc_Maine <- summary(fr_Maine$Time..minutes.)
> desc_Maine
```

```
> print(paste("Range: ",desc_Maine["Max."]-desc_Maine["Min."]))
> print(paste("Standard Deviation: ", sd(fr_Maine$Time..minutes.)))
> print(paste("IQR: ",(desc_Maine["3rd Qu."]-desc_Maine["1st Qu."])))
```

Minimum	1st Quartile	Mean	Median	3rd Quartile	Maximum
30.57	50.00	58.20	57.03	64.24	152.17

```
[1] "Range:  121.6"
[1] "Standard Deviation:  12.1851105531497"
[1] "IQR:  14.24775"
```

Statistics for 'Away' runners:

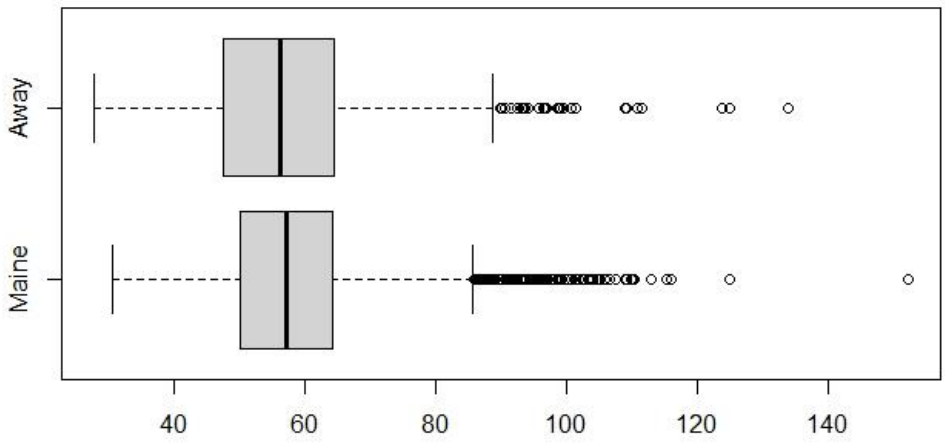
```
> desc_Away <- summary(fr_Away$Time..minutes.)
> desc_Away
> print(paste("Range: ",desc_Away["Max."]-desc_Away["Min."]))
> print(paste("Standard Deviation: ", sd(fr_Away$Time..minutes.)))
> print(paste("IQR: ",(desc_Away["3rd Qu."]-desc_Away["1st Qu."])))
> print("In conclusion: The running times of Runners from Maine & Away
follow Symmetric distribution. ")
```

Minimum	1st Quartile	Mean	Median	3rd Quartile	Maximum
27.78	49.15	57.82	56.92	64.83	133.71

```
[1] "Range:  105.928"
[1] "Standard Deviation:  13.8353842414778"
[1] "IQR:  15.674"
[1] "In conclusion: The running times of Runners from Maine & Away
follow Symmetric distribution. "
```

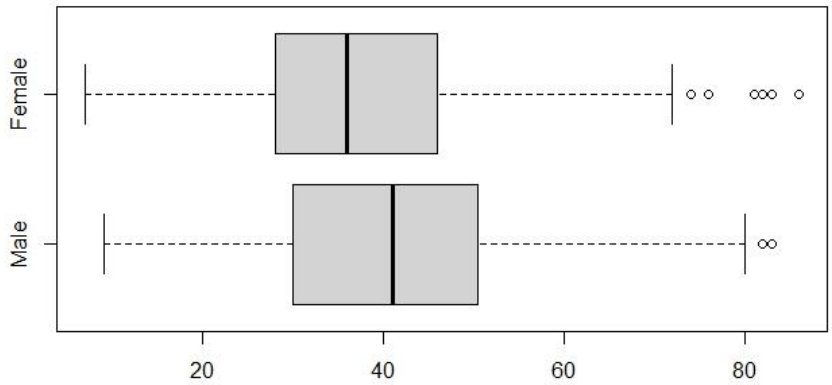
c) > # Creating Box plots for both Runner categories.

```
> bplt_runners <-
cbind("Maine"=(fr_Maine$Time..minutes.),"Away"=(fr_Away$Time..minut
es.))
> boxplot(bplt_runners, beside=T, horizontal = TRUE)
```



d) > # Question 1-d

```
> Run_M <- rdrace[rdrace$Sex=='M',]
> Run_F <- rdrace[rdrace$Sex=='F',]
> mfrunner <-
cbind("Male"=(type.convert(Run_M$Age)),"Female"=(type.convert(Run_F
$Age)))
> boxplot(mfrunner, beside=T, horizontal = TRUE)
```



```
> summary(mfrunner)
> sd_run1 <- sd(mfrunner[,"Male"],na.rm = T)
> sd_run1
> sd_run1 <- sd(mfrunner[,"Female"],na.rm = T)
> sd_run1
```

```
> summary(mfrunner)
      Male            Female
 Min.   : 9.00   Min.   : 7.00
 1st Qu.:30.00   1st Qu.:28.00
 Median :41.00   Median :36.00
 Mean   :40.31   Mean   :37.24
 3rd Qu.:50.50   3rd Qu.:46.00
 Max.   :83.00   Max.   :86.00
```

[1] 14.00449 # male sd

[1] 12.26925 # female sd

R Code for MP-2-Question-1:

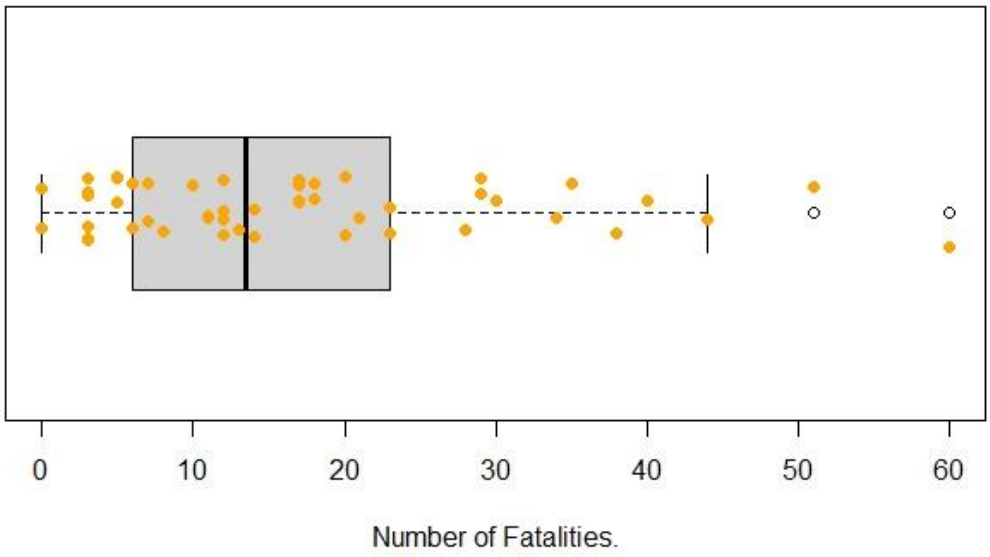
```
# STATS-Mini-Project-2

# Setting working directory to Proj-2 folder.

setwd("C:/Users/Shalin Kaji/Desktop/UT-Dallas-Spr23/STATS-DS-Min.Chen/Mini-Proj-2")

getwd()


# Observing our .csv file

rdrace <- read.csv("roadrace.csv")

head(rdrace,10)


# Question 1-a

# Create a barplot of Maine, what does it conclude?

barplot(table(rdrace$Maine)/length(rdrace$Maine), ylim = c(0,1))

print('Probabilities of runners from Maine/Away: ')

table(rdrace$Maine)/length(rdrace$Maine)

print('To conclude: No. of runners from Maine is thrice the no. of runners from other cities.')


# Question 1-b

# Creating histograms of running times for Runners from Maine/Away.

fr_Maine <- rdrace[rdrace$Maine=='Maine',]

fr_Away <- rdrace[rdrace$Maine=="Away",]

p1 <- hist(fr_Maine$Time..minutes., breaks=20)

p2 <- hist(fr_Away$Time..minutes., breaks=20)

plot(p1,col="light blue",xlim = c(20,160), main="Runner's Time" , xlab = "Time in Mins" , ylab = "Number of
Runners")

plot(p2,col="light pink",add=T)


legend("topright", c("Maine","Away"), col=c("light blue","light pink"),lwd = 5)

# Statistics for Maine Runners

desc_Maine <- summary(fr_Maine$Time..minutes.)

desc_Maine

print(paste("Range: ",desc_Maine["Max."]-desc_Maine["Min."]))

print(paste("Standard Deviation: ", sd(fr_Maine$Time..minutes.)))

print(paste("IQR: ",(desc_Maine["3rd Qu."]-desc_Maine["1st Qu."])))
```

```
#Statistics for Away Runners
desc_Away <- summary(fr_Away$Time..minutes.)
desc_Away
print(paste("Range: ",desc_Away["Max."]-desc_Away["Min."]))
print(paste("Standard Deviation: ", sd(fr_Away$Time..minutes.)))
print(paste("IQR: ",(desc_Away["3rd Qu."]-desc_Away["1st Qu."])))
print("In conclusion: The running times of Runners from Maine & Away follow Symmetric distribution. ")

# Question 1-C
# Creating Box plots for both Runner categories.
bplt_runners <- cbind("Maine"=(fr_Maine$Time..minutes.),"Away"=(fr_Away$Time..minutes.))
boxplot(bplt_runners, beside=T, horizontal = TRUE)

# Question 1-d
Run_M <- rdrace[rdrace$Sex=='M',]
Run_F <- rdrace[rdrace$Sex=='F',]

mfrunner <- cbind("Male"=(type.convert(Run_M$Age)),"Female"=(type.convert(Run_F$Age)))
boxplot(mfrunner, beside=T, horizontal = TRUE)
summary(mfrunner)
sd_run1 <- sd(mfrunner[,"Male"],na.rm = T)
sd_run1
sd_run1 <- sd(mfrunner[,"Female"],na.rm = T)
sd_run1
```

Question 2:

Plotting Box Plot for number of fatal motorcycle accidents along with jittered data points to avoid over-plotting the outliers.

```
> ctyacc <- motor$Fatal.Motorcycle.Accidents
> boxplot(ctyacc, horizontal = TRUE)
> stripchart(ctyacc, method = "jitter", pch = 19, add = TRUE, col = "orange")
```



Analysing the motorcycle dataset, we obtain the following statistics:

```
> dstats <- summary(ctyacc)
> dstats
> print(paste("Standard Deviation: ", sd(ctyacc)))
> print(paste("IQR: ", dstats[5]-dstats[2]))
```

Minimum	1st Quartile	Mean	Median	3rd Quartile	Maximum
0.00	6.00	17.02	13.50	23.00	60.00

```
[1] "Standard Deviation:  13.8125591683852"
```

[1] "IQR: 17"

Identifying the outliers in the given dataset, using Boxplot:

> motor[motor$Fatal.Motorcycle.Accidents>((dstats[5]-dstats[2])*1.5+dstats[5]),]

	County	Fatal.Motorcycle.Accidents
23	GREENVILLE	51
26	HORRY	60

These counties might have the highest number of motorcycle fatalities in South Carolina:

> Due to many underage riders.

R Code:

```
getwd()


# Observing our .csv file
motor <- read.csv("motorcycle.csv")
ctyacc <- motor$Fatal.Motorcycle.Accidents
boxplot(ctyacc, horizontal = TRUE, xlab = 'Number of Fatalities.')
stripchart(ctyacc, method = "jitter", pch = 19, add = TRUE, col = "orange")
dstats <- summary(ctyacc)
dstats
print(paste("Standard Deviation: ", sd(ctyacc)))
```