Mini Project # 2

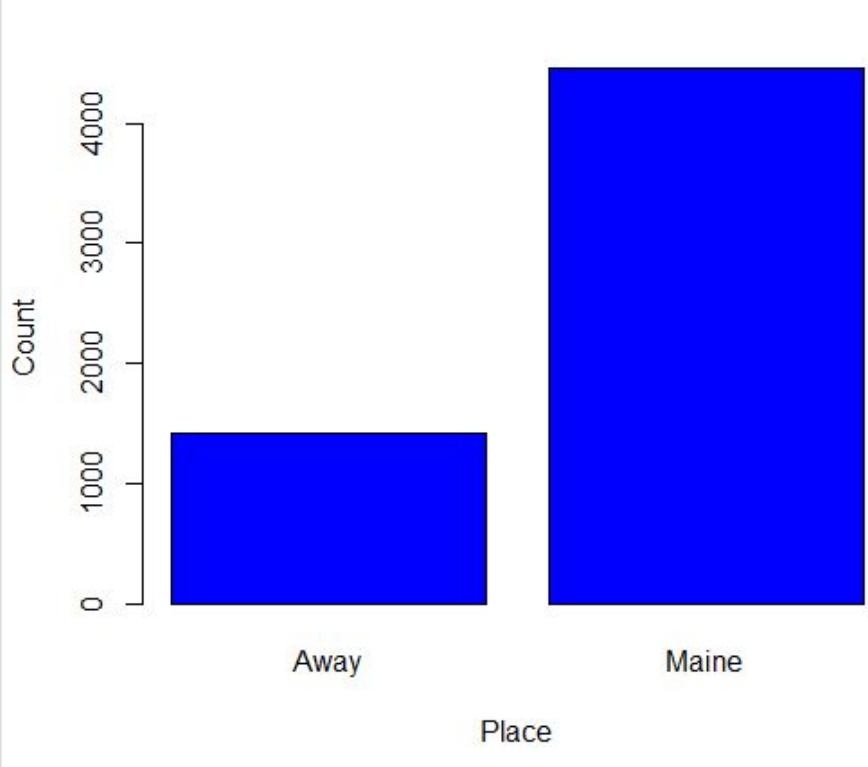
Qingyu Lan, Lakshmi Priyanka Selvaraj

Contribution of each group member :

Both members worked on the questions together.

Section 1. Answers to the specific questions asked

1. Question 1
   A.

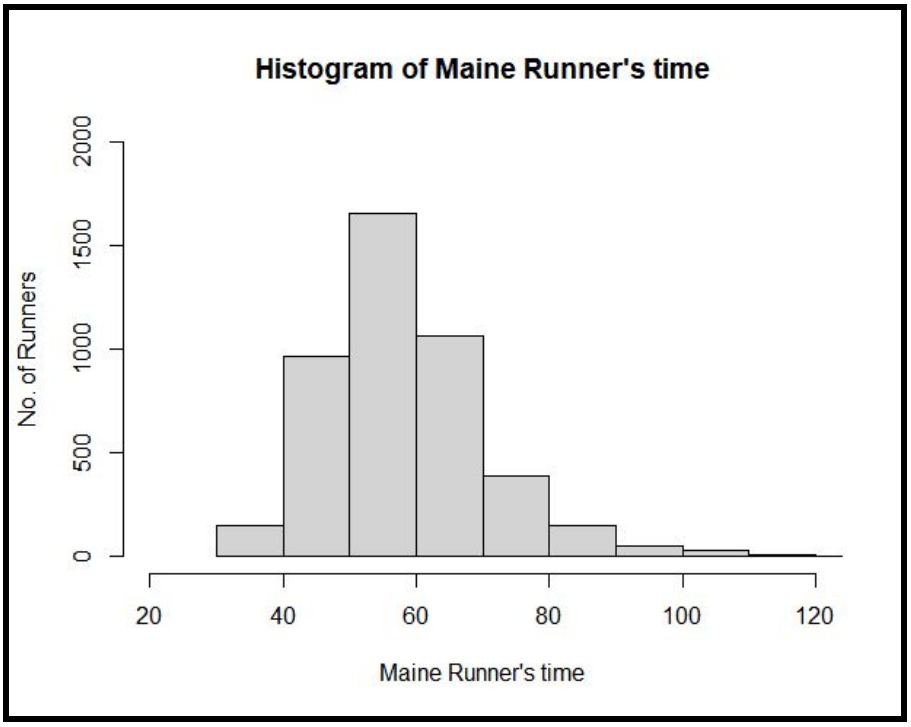


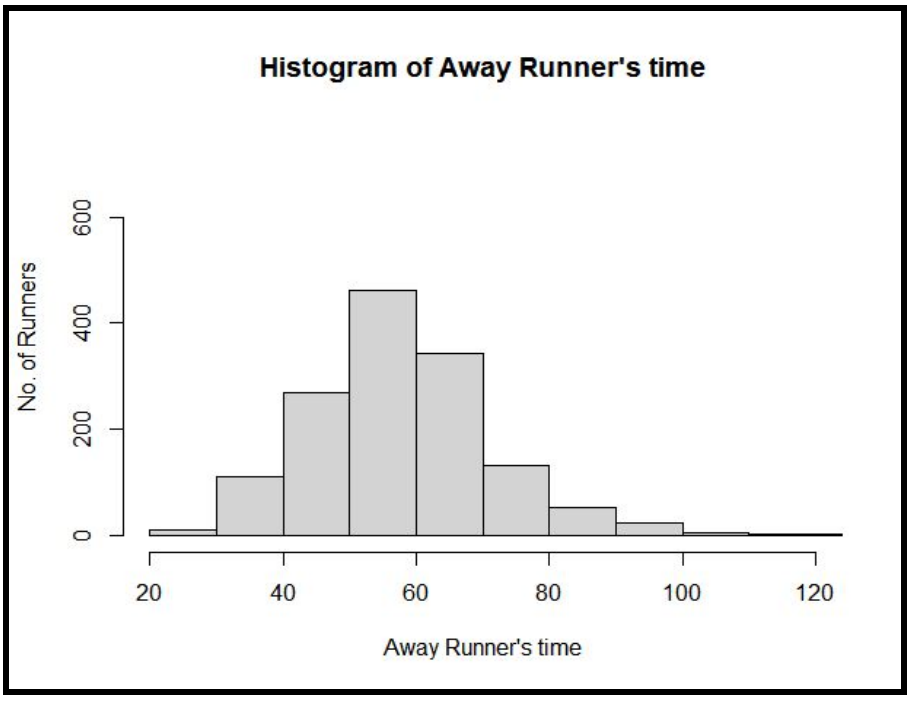

```
> roadrace %>% count(Maine, sort = TRUE)
  Maine    n
1 Maine 4458
2  Away 1417
```

From the barplot, we can conclude that there are around 3000 more participants from Maine than from Away. This is backed up by the count of the participants, there are 4458 in the Maine group and 1417 from the Away group. Out of the

total number of participants, approximately 75% of the participants are from Maine.

B.





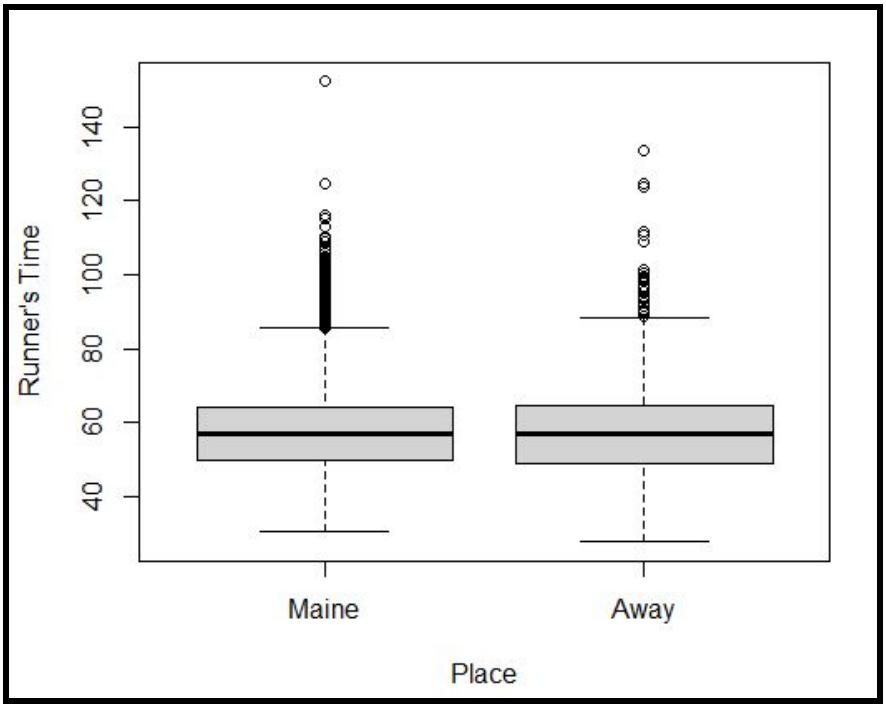
```
> summary(maine_runner_time)
  Runner_time
 Min.   : 30.57
 1st Qu.: 50.00
 Median : 57.03
 Mean   : 58.20
 3rd Qu.: 64.24
 Max.   :152.17
> summary(away_runner_time)
  Runner_time
 Min.   : 27.78
 1st Qu.: 49.15
 Median : 56.92
 Mean   : 57.82
 3rd Qu.: 64.83
 Max.   :133.71
```

```
> sd(as.numeric(unlist(maine_runner_time)))
[1] 12.18511
> sd(as.numeric(unlist(away_runner_time)))
[1] 13.83538
```

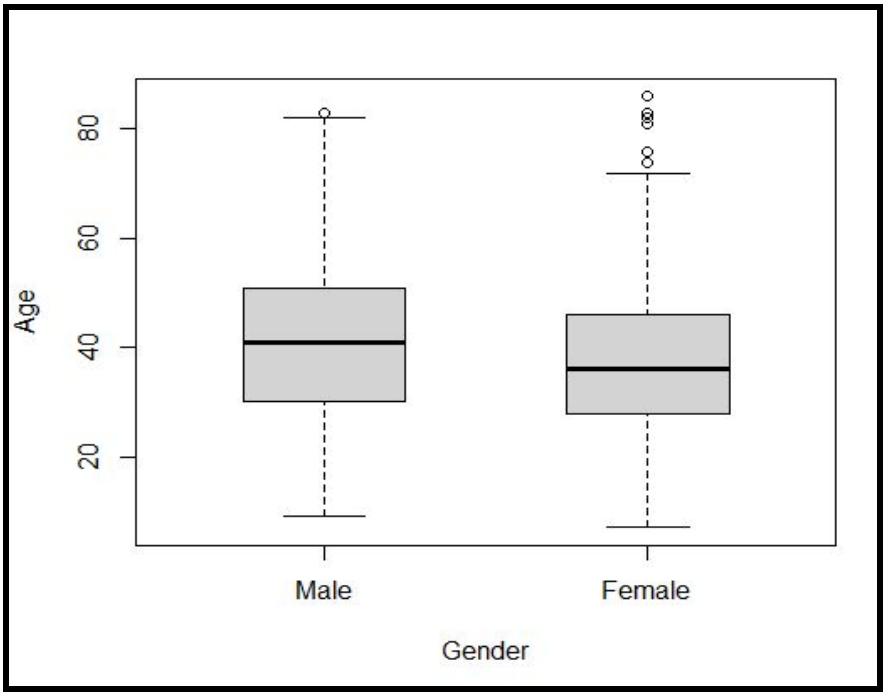
```
> IQR(as.numeric(unlist(maine_runner_time)))
[1] 14.24775
> IQR(as.numeric(unlist(away_runner_time)))
[1] 15.674
```

From the histograms we can conclude that both the Maine runners and the Away runners share a similar distribution of runner time. This is supported by the summary shown above that the Maine runner time distribution and Away runner time distribution have very similar mean, range, standard deviation, median, and interquartile range. The Maine_runner time has a wider range(121.6) when compared to Away runner time(105.9) This might be due to the reason that the number of participants is very high when compared to the latter.

C. Boxplot based on Place of the participant



D. BOXPLOT based on Gender

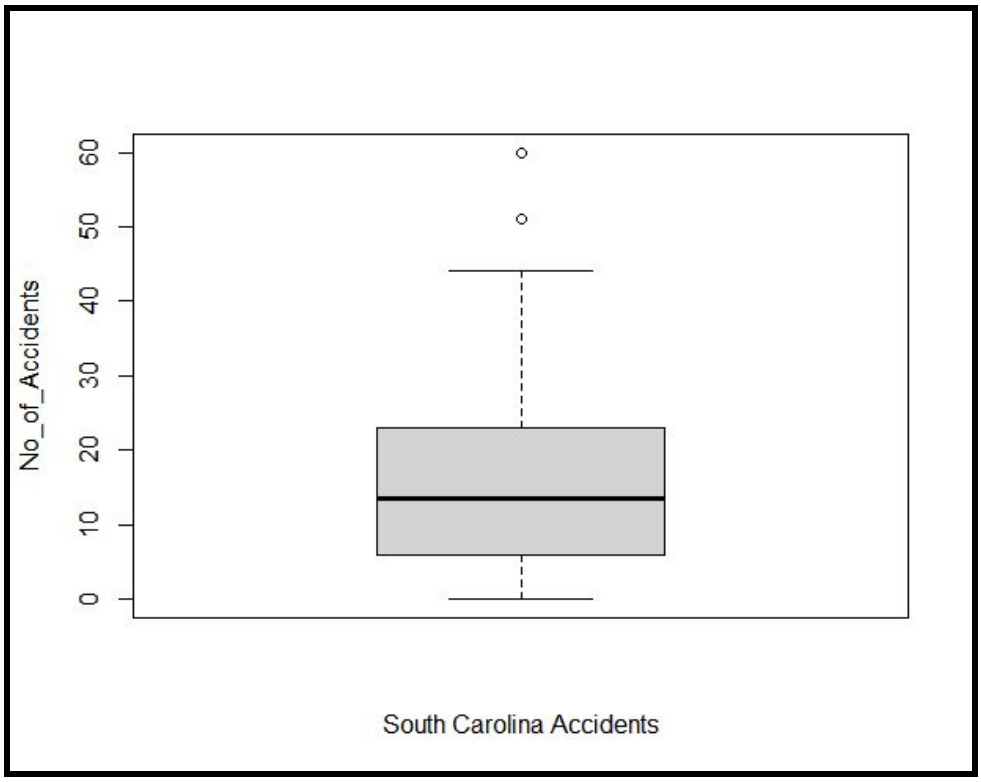


```
> summary(male_runners)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   30.00   41.00   40.45   51.00   83.00
> summary(female_runners)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   28.00   36.00   37.24   46.00   86.00
> sd(as.numeric(unlist(male_runners)))
[1] 13.99289
> sd(as.numeric(unlist(female_runners)))
[1] 12.26925
> IQR(as.numeric(unlist(male_runners)))
[1] 21
> IQR(as.numeric(unlist(female_runners)))
[1] 18
```

From the two distributions, we can conclude that the male participants are on average older than the female participants with a few outliers on the female side. This is shown by the male runners having a higher mean of 40.45 years compared to the female mean of 37.24 years, and male runners having a median of 41 years compared to female median of 36 years. The male runners also have higher interquartile range, male 1st quartile is 30 years compared to female 28 years, and male 3rd quartile is 51 years compared to 46 years.

Also, the age of female participants(Range:79) has a higher range than male participants(Range:74). The youngest and the oldest participant of the race are both female.

2. Question 2




```
> summary(motorcycle$Fatal.Motorcycle.Accidents)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    6.00   13.50   17.02   23.00   60.00
> sd(as.numeric(unlist(motorcycle$Fatal.Motorcycle.Accidents)))
[1] 13.81256
> IQR(as.numeric(unlist(motorcycle$Fatal.Motorcycle.Accidents)))
[1] 17
```

The distribution of fatal motorcycle accidents in each country of South Carolina during 2009 has a mean of 17.02 with standard deviation of 13.81256. It has a median of 13.50. The first quartile is 6, and the 3rd quartile is 23. The lowest fatal motorcycle accidents in a county is 0, and the highest fatal motorcycle accidents in a county is 60.

```
> subset(motorcycle, Fatal.Motorcycle.Accidents == outlier_values,
+        c(County, Fatal.Motorcycle.Accidents))
       County Fatal.Motorcycle.Accidents
23 GREENVILLE                         51
26      HORRY                         60
```

The two counties that are considered outliers are Greenville with 51 accidents and Horry with 60 accidents. Those counties might have the highest numbers of motorcycle fatalities in South Carolina due to multiple factors. Some are Population, Type of terrain, More defaulters, Accuracy of reports and so on. Since the data doesn't provide other factors, we can't assume for sure what was the reason behind these increased numbers of accidents.

Section 2: R code.

R code for question 1

```
#Q1:Read CSV roadrace.csv

rm(list=ls()) #removes the variables and cleans the environment
roadrace <- read.csv("roadrace.csv")

maine_data <- roadrace$Maine  #this contains the column data of main

barplot(table(maine_data),
        names.arg=c("Away","Maine"),
        xlab= "Place", ylab="Count",
        col= "blue",
        border=TRUE)

install.packages("dplyr")
library(dplyr)
roadrace %>% count(Maine, sort = TRUE)


#b
maine_runner_time <- subset(roadrace,Maine=="Maine",c(Time..minutes.))
away_runner_time <- subset(roadrace,Maine=="Away",c(Time..minutes.))
names(maine_runner_time)[names(maine_runner_time)=='Time..minutes.'] <- 'Runner_time'
names(away_runner_time)[names(away_runner_time)=='Time..minutes.'] <- 'Runner_time'

hist(maine_runner_time$Runner_time,
     main = paste("Histogram of Maine Runner's time"),
     xlim=c(20,120),
     ylim=c(0,2000))

hist(away_runner_time$Runner_time,
     main = paste("Histogram of Away Runner's time"),
     xlim=c(20,120),
     ylim=c(0,750))

summary(maine_runner_time)
summary(away_runner_time)
```

```
sd(as.numeric(unlist(maine_runner_time)))
sd(as.numeric(unlist(away_runner_time)))

IQR(as.numeric(unlist(maine_runner_time)))
IQR(as.numeric(unlist(away_runner_time)))

#range
range_maine <- max(maine_runner_time$Runner_time)- min(maine_runner_time$Runner_time)
range_away <- max(away_runner_time$Runner_time)-min(away_runner_time$Runner_time)


#c

boxplot(maine_runner_time$Runner_time,away_runner_time$Runner_time,
     xlab = "Place", ylab = "Runner's Time",
     names = c("Maine", "Away"))


#d

male<- subset(roadrace, Sex == "M", c(Age))
male_runners <- as.numeric(male$Age) #since age is saved as character we convert it to
numeric
female <- subset(roadrace, Sex == "F", c(Age))
female_runners <- as.numeric(female$Age)

boxplot(male_runners, female_runners,
       xlab="Gender", ylab="Age",
       names = c("Male","Female"),
       boxwex = 0.5)


summary(male_runners)
summary(female_runners)

sd(as.numeric(unlist(male_runners)))
sd(as.numeric(unlist(female_runners)))

IQR(as.numeric(unlist(male_runners)))
IQR(as.numeric(unlist(female_runners)))

male_range <- max(male_runners)- min(male_runners)
female_range <- max(female_runners)-min(female_runners)
```

R code for question 2

#2

```
motorcycle <- read.csv("motorcycle.csv")
head(motorcycle)

boxplot(motorcycle$Fatal.Motorcycle.Accidents, xlab= "No_of_Accidents")
summary(motorcycle$Fatal.Motorcycle.Accidents)

sd(as.numeric(unlist(motorcycle$Fatal.Motorcycle.Accidents)))


#Outlier values
outlier_values <- boxplot.stats(motorcycle$Fatal.Motorcycle.Accidents)$out

subset(motorcycle, Fatal.Motorcycle.Accidents == outlier_values,
       c(County, Fatal.Motorcycle.Accidents))
```