Mini Project #2
Nikhil Pareek, Tanushri Singh
Contribution of each group member:
Both worked together to finish the two questions. Collaborated to learn R and then write the scripts. Nikhil worked to check the accuracy of the script and Tanu worked to Document and report all the findings. Both partners worked efficiently to complete the project requirements!

Question 1:
Initially started of by downloading csv files and reading the values of data that is stored in each file. Now we began coding, read in the files and stored them in designated variables. It is important to note that if one has already set up a working directory and placed the csv file in that, there is no need to mention the file path separately.

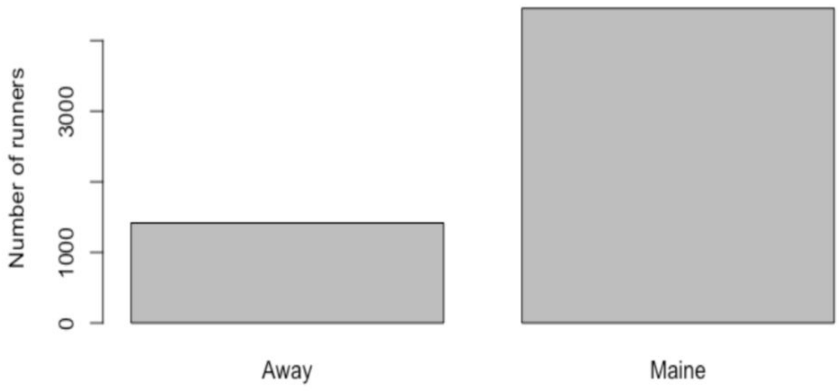
a) To create a bar graph first the “maine” column must be extracted from the csv in a new variable. We used the which function to accomplish the filtration that is necessary. Then used the barplot() function to graph the bar graph.

   Analytically computing the number values for maine and away results in:
   Maine group = 4458 Away group = 1417

   Code Snippet:

```
> data =read.csv("/Users/TanushriSingh/Desktop/CS6313/MiniProject 2/roadrace.csv")
> barplot(c(sum(data$Maine == 'Away'), sum(data$Maine == 'Maine')), names.arg = c('Away', 'Maine'), space = 0.25, ylab = 'Number of runners')
> sum(data$Maine == 'Away')
[1] 1417
> sum(data$Maine == 'Maine')
[1] 4458
```

   Bar Graph:



From the bar graphs it can be concluded that the Maine group is greater than the total number of runners from the away group. Also, it can be concluded that the Maine group account for 75.8% of the portion while the away group account for 24.2% of the portion out of a total 5875 runners.

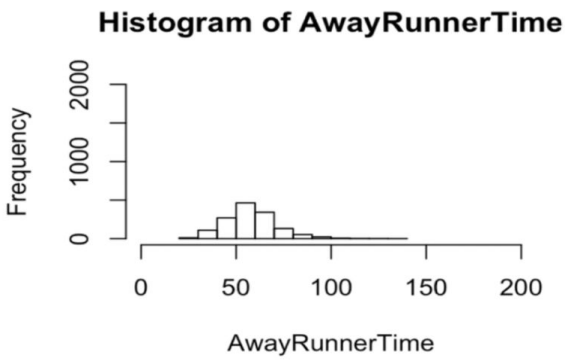
b) In order to handle this problem firstly the required columns must be extracted using the which condition. This will be stored in variables we will call MaineRTime and AwayRTime. Then Histograms are drawn using the hist function.

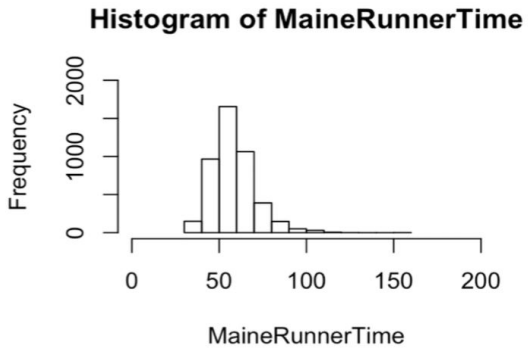
Code Snippet:

> AwayRunnerTime = data$Time..minutes.[which(data$Maine == 'Away')] > hist(AwayRunnerTime, xlim = range(0,200), ylim = range(0,2000))
> MaineRunnerTime = data$Time..minutes.[which(data$Maine == 'Maine')] > hist(MaineRunnerTime, xlim = range(0,200), ylim = range(0,2000))

Calculations for the Maine and Away groups:

```
> summary(MaineRunnerTime)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.57   50.00   57.03   58.20   64.24  152.17
> summary(AwayRunnerTime)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.78   49.15   56.92   57.82   64.83  133.71
> IQR(AwayRunnerTime)
[1] 15.674
> IQR(MaineRunnerTime)
[1] 14.24775
> range(MaineRunnerTime)
[1]  30.567 152.167
> range(AwayRunnerTime)
[1]  27.782 133.710
> sd(AwayRunnerTime)
[1] 13.83538
> sd(MaineRunnerTime)
[1] 12.18511
```





From the graphs it is evident that both distributions and skewed to the right. An R function called Summary() can be used to record the min, 1st Q, median, mean, 3rd Q and max values. Here are the generated values for both the Maine and Away groups in a tabulated form.

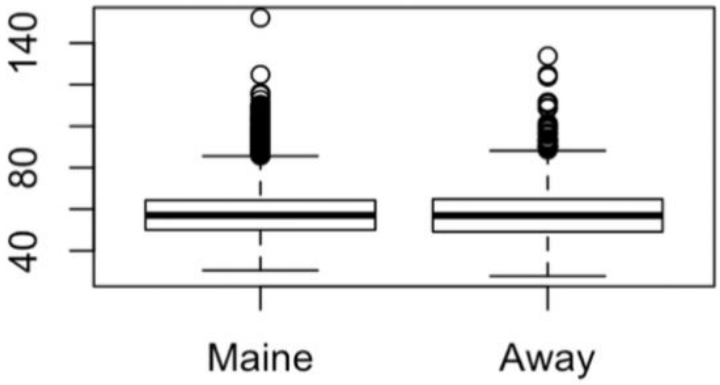
Groups	Min	1st Qu.	Median	Mean	3rd Qu.	Max	IQR	Range	Var	SD
Maine Group	30.57	50	57.03	58.20	64.24	152.17	14.25	30.567 152.17	148.47	12.19
Away Group	27.78	49.15	56.92	57.82	64.83	133.71	15.67	27.78 133.71	191.41	13.84

Some observations that are made right away include how the values of min, 1st Q, median, mean and max are higher for the maine group but 3rd Q, IQR and Standard deviation are higher for the away group.

c) To create a box-plot the boxplot() function is used in R.

Code Snippet:
> **boxplot(MaineRunnerTime, AwayRunnerTime, names = c('Maine', 'Away'))**

Box Plot:



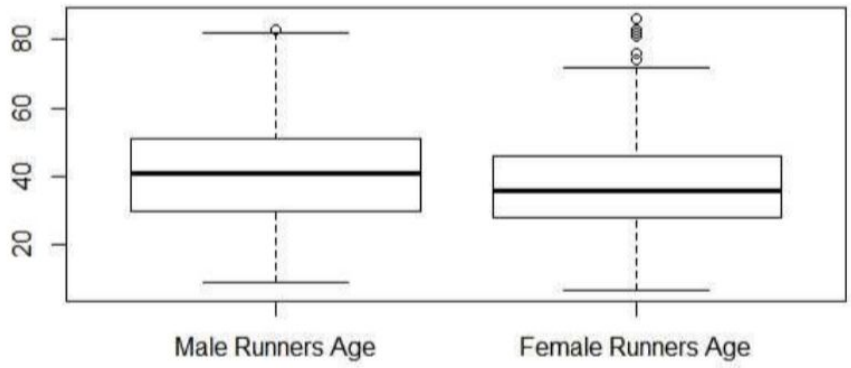
d) Now a box-plot needs to be created that is gender specific.

Code Snippet:
> **MaleRunnersAge = roadrace$Age[which(roadrace$Sex=='M')]**
> **FemaleRunnersAge = roadrace$Age[which(roadrace$Sex=='F')]**
> **boxplot(MaleRunnersAge, FemaleRunnersAge, names = c('Male Runners Age', 'Female Runners Age')**

Calculations for the Maine and Away groups:

```
> summary(MaleRunnersAge)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   30.00   41.00   40.45   51.00   83.00
> IQR(MaleRunnersAge)
[1] 21
> range(MaleRunnersAge)
[1]  9 83
> sd(MaleRunnersAge)
[1] 13.99289
> summary(FemaleRunnersAge)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   28.00   36.00   37.24   46.00   86.00
> IQR(FemaleRunnersAge)
[1] 18
> range(FemaleRunnersAge)
[1]  7 86
> sd(FemaleRunnersAge)
[1] 12.26925
```

Boxplot:



Required statistics in a tabular format:

Groups	Min	1st Qu.	Median	Mean	3rd Qu.	Max	IQR	Range	SD
Male Group	9	30	41	40.45	51	83	21	9 83	13.99
Female Group	7	28	36	37.24	46	86	18	7 86	12.26

Observations: the statistical values of males is higher than females. However ladies who are of age 86 and up seem to be active participants in races.

Question 2:
Start off by using the which condition in order to separate the suitable values. Then the boxplot() function must be used to create the graph.

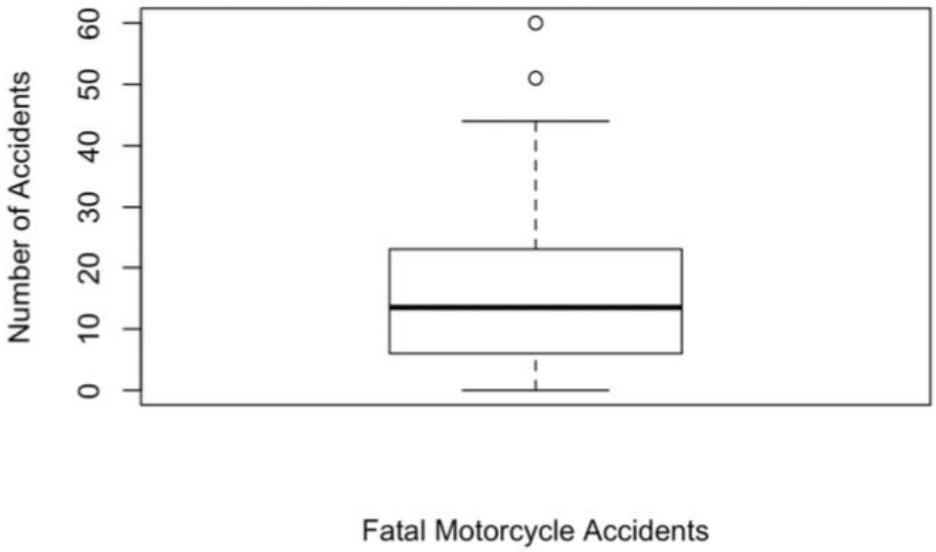
Code Snippet:

```
#Read in Data
>  data1= read.csv("/Users/TanushriSingh/Desktop/CS6313/MiniProject 2//motorcycle.csv")
>  FatalAccidents = data1$Fatal.Motorcycle.Accidents
#Now create boxplot for Fatal.Motorcycle.Accidents values
> boxplot(FatalAccidents, xlab = 'Fatal Motorcycle Accidents', ylab = 'Number of Accidents') #For calculating the required statistics, we are executing the following code

#Now calculate the Upper and Lower  bounds
> 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))

 #Now check for outliers and display counties that are outliers
> FatalCounty=data1$County[which(data1$Fatal.Motorcycle.Accidents<LowerBound | data1$Fatal.Motorcycle.Accidents > UpperBound)]
```

Resulting Value for FatalCounty:

```
> FatalCounty
[1] GREENVILLE HORRY
```

Boxplot:



In order to calculate the required statistics the following lines are fun:

```
> summary(FatalAccidents)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    6.00   13.50   17.02   23.00   60.00
> IQR(FatalAccidents)
[1] 17
> range(FatalAccidents)
[1]  0 60
> sd(FatalAccidents)
[1] 13.81256
```

Tabular representation of the statistical summary:

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	IQR	Range	SD
Accidents	**0**	**6**	**13.5**	**17.02**	**23**	**60**	**17**	**0 60**	**13.81**

Now, in order to discover the outliers we must calculate the 25th and 75th quantiles. For which the quantile() function with default type 7 value is chosen.
Upon calculation the probabilities of 25% quantile and 75% quantile are 0.25 and 0.75 respectively. A value is an outlier if it's more than 1.5*IQR() away from the 25th and 75th quantiles.
Hence lower bound: 25th percentile - 1.5IQR and upper bound: 75th percentile + 1.5IQR
Now, any value in the lower or upper bound are outliers.

The counties highest number of motorcycle fatalities are in South Carolina in Greenville Horry. The main reason for this in 2009 could be due to the poor road and highway maintenance as well as negligent drivers.