Mini Project 2

Names of group members: Karan Risbud(KSR190005), Shubham Vartak(SXV200115)

Contribution of each group member: Both the Project group members worked together on the project. Collaborated to solve the problem and implementation of R programming.

1) (12 points) 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.

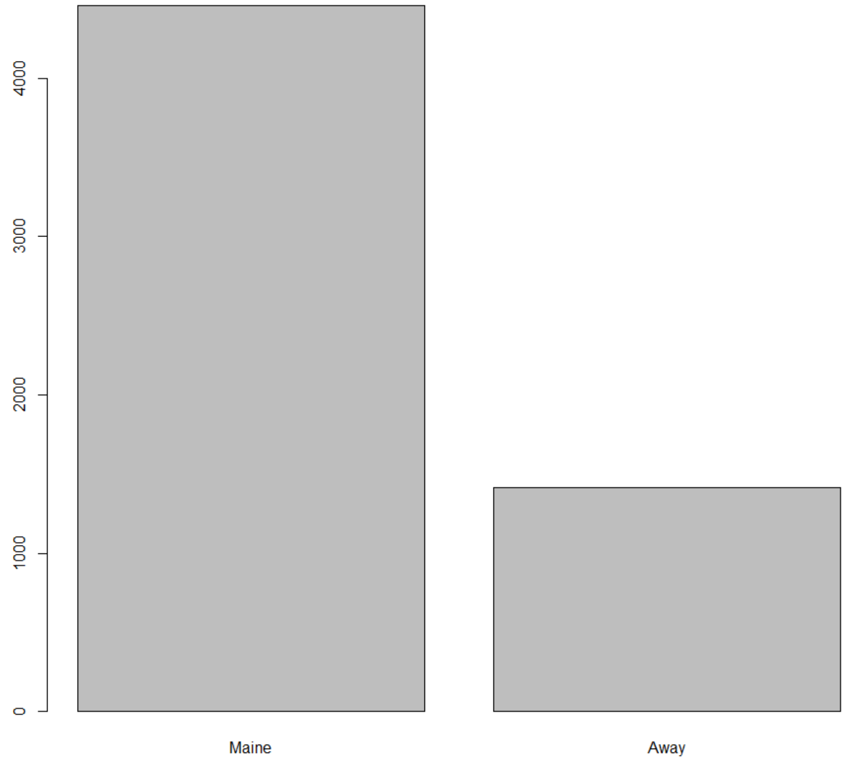
Code:

```
roadrace = read.csv(file="/Users/shubh/Downloads/STATS/mini_project_2/roadrace.csv")
```

1a) 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:

```
height = c(sum(roadrace$Maine=='Maine'),sum(roadrace$Maine == 'Away'))
barplot(height,names.arg=c('Maine','Away'))
```

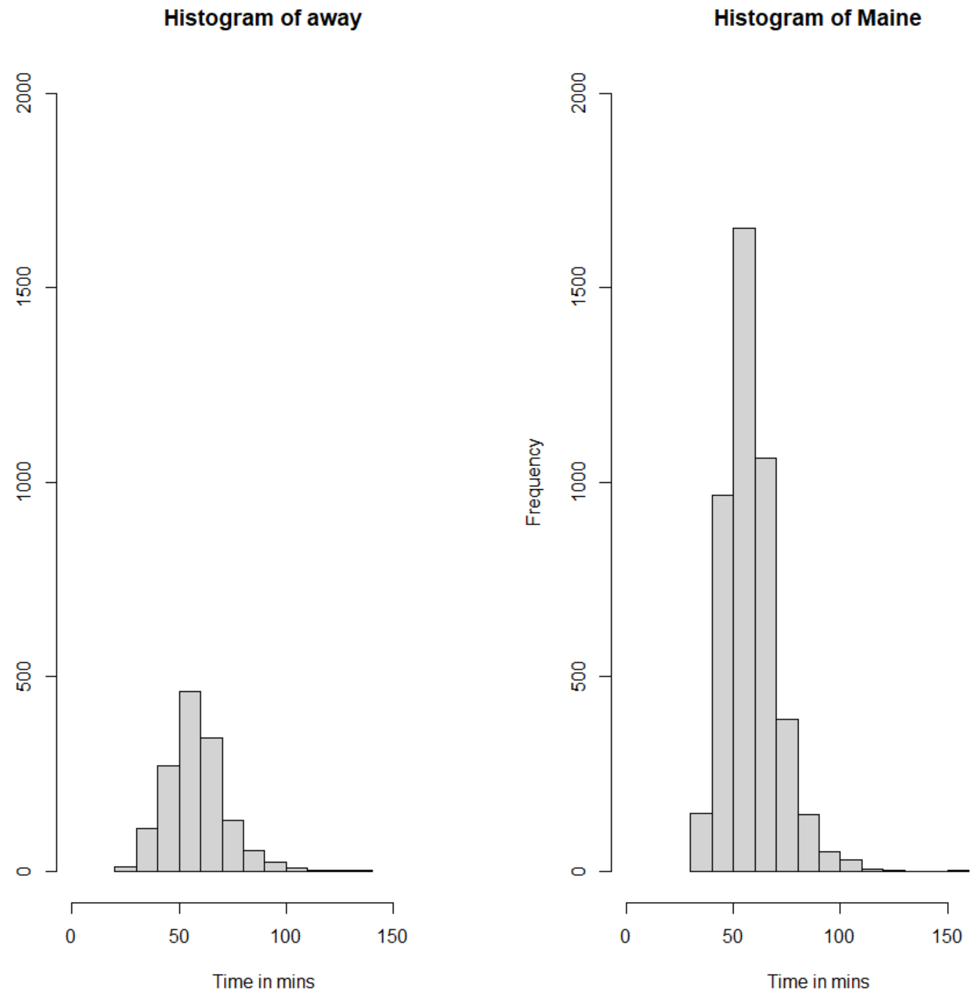


Observation
There are more runners who are from Maine as compared to the number of runners away from Maine.

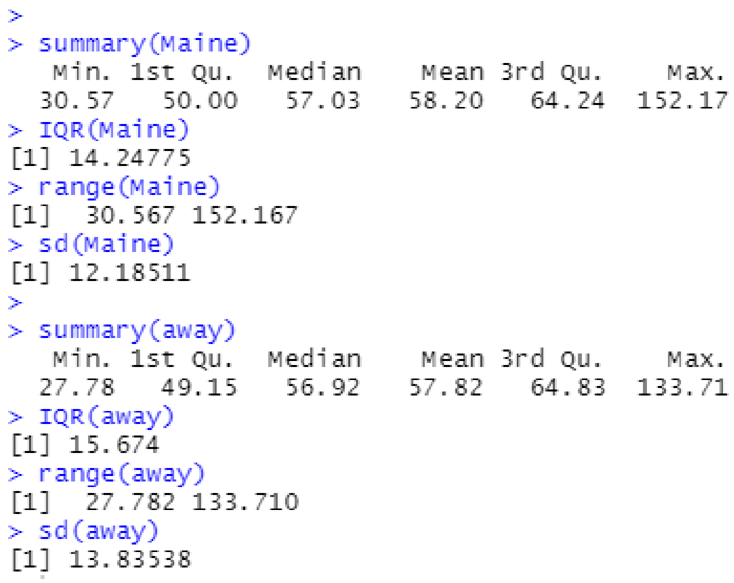
1b) Create two histograms the runners' times (given in minutes) — one for the Maine group and the second for the Away group. Make sure that the histograms on the same scale. What can we conclude about the two distributions? Back up your conclusions with relevant summary statistics, including mean, standard deviation, range, median, and interquartile range.

Code:

```
away = (roadrace$Time..minutes[which(roadrace$Maine=='Away')])
Maine = (roadrace$Time..minutes[which(roadrace$Maine=='Maine')])
par(mfrow=c(1,2))
hist(away,xlim=c(0,180),ylim = c(0,2000),xlab="Time in mins")
hist(Maine,xlim=c(0,180),ylim = c(0,2000),xlab="Time in mins")

summary(Maine)
IQR(Maine)
range(Maine)
sd(Maine)

summary(away)
IQR(away)
range(away)
sd(away)
```



Console Output:



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

Observation:

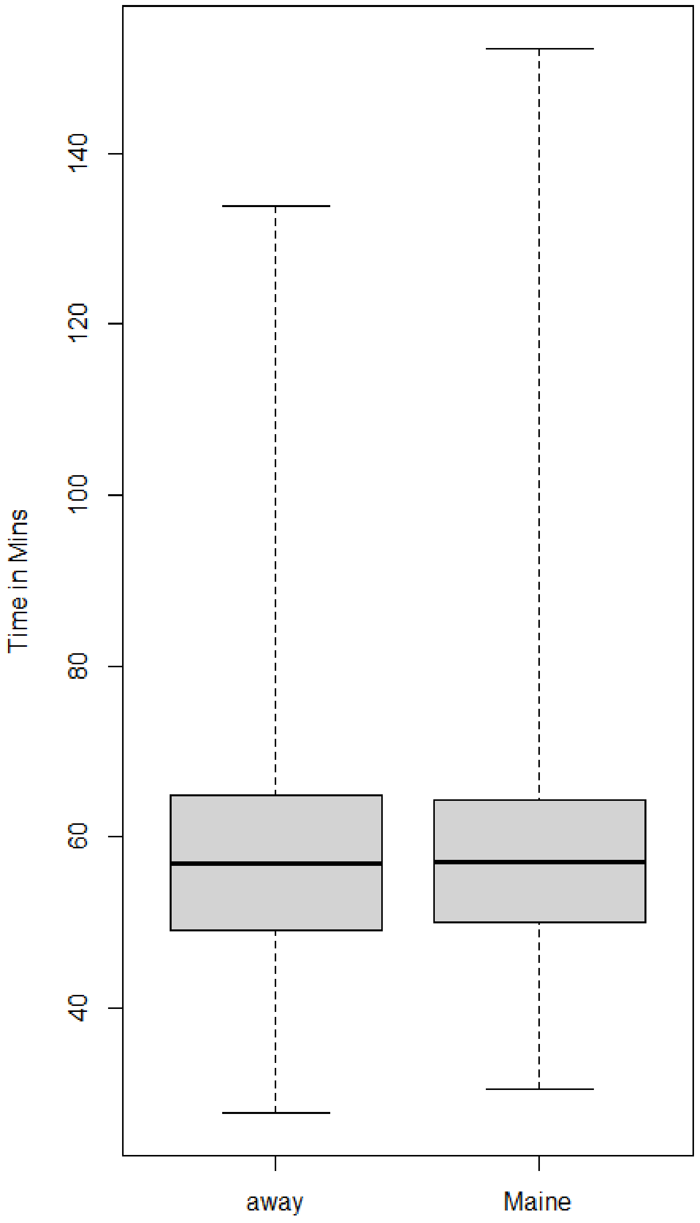
Based on the summary of both Maine and Away. We can observe that both the data have approximately same mean and median thus they are approximately normally distributed. Since deviation of away is more the dispersion of data is more as observed in the histogram.

1c) Repeat (b) but with side-by-side boxplots.

Code:

```
boxplot(ylab ="Time in Mins",away,Maine,names=c("away","Maine"),range = 0)
```

Output:

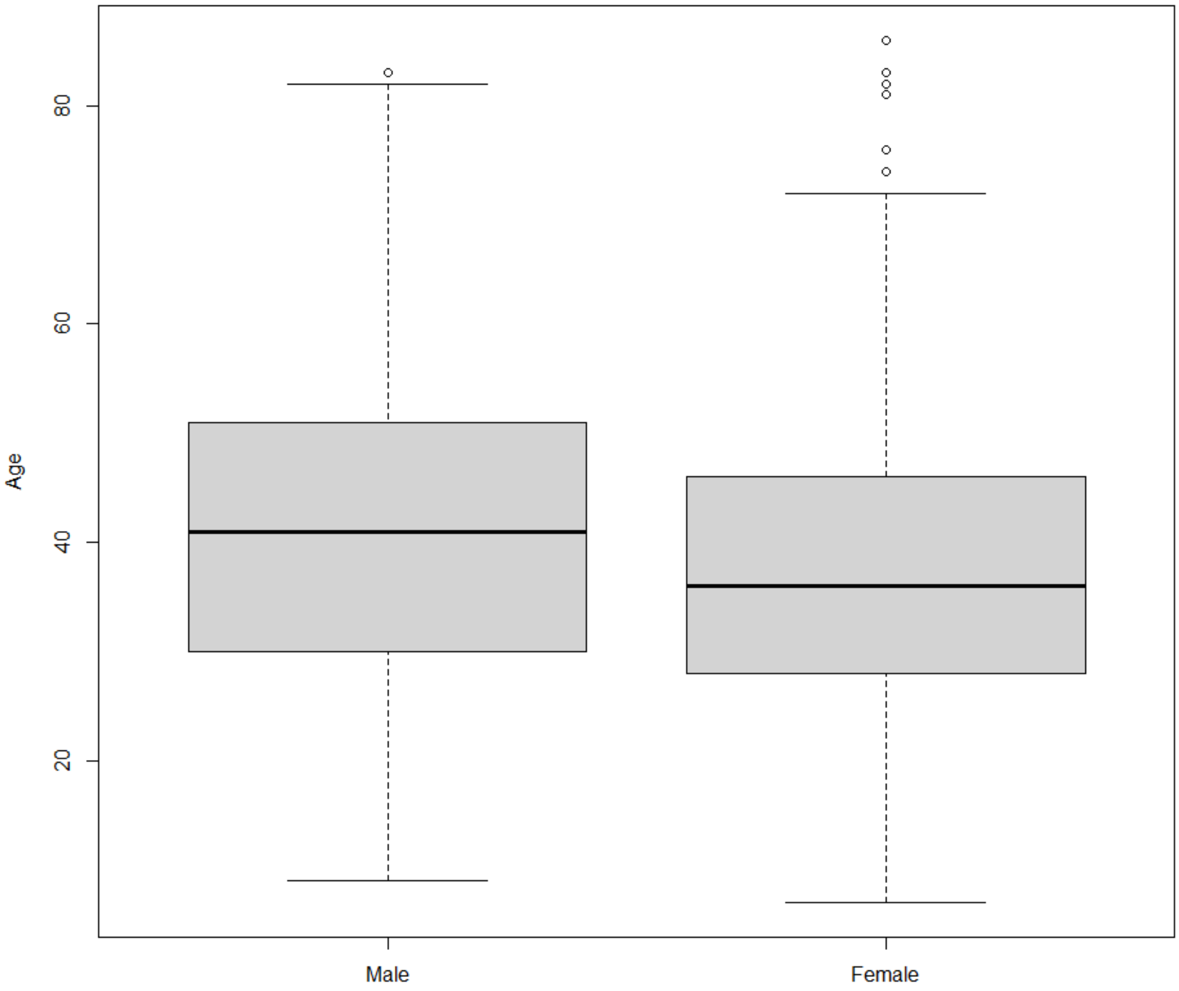


1 d) Create side-by-side boxplots for the runners' ages (given in years) for male and female runners. What can we conclude about the two distributions? Back up your conclusions with relevant summary statistics, including mean, standard deviation, range, median, and interquartile range.

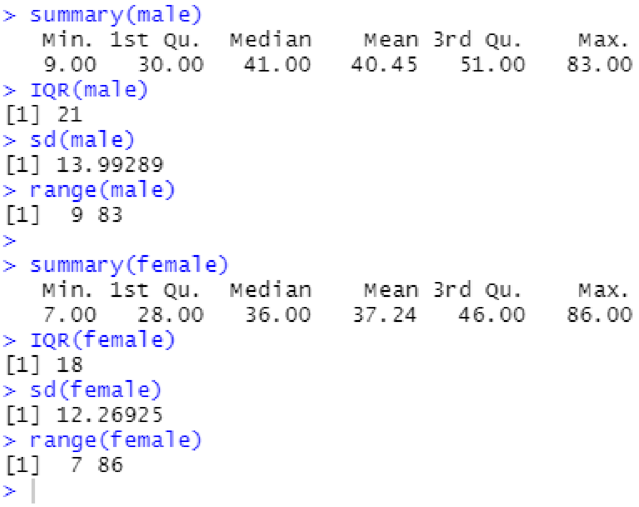
Code:

```
male = (roadrace$Age[which(roadrace$Sex=='M')])
female = (roadrace$Age[which(roadrace$Sex=='F')])


male = as.numeric(male)
female = as.numeric(female)

boxplot(ylab ="Age",male,female,names=c("Male","Female"),range = 1.5)

summary(male)
IQR(male)
sd(male)
range(male)

summary(female)
IQR(female)
sd(female)
range(female)
```

Output:



Console Output:



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

Observation:

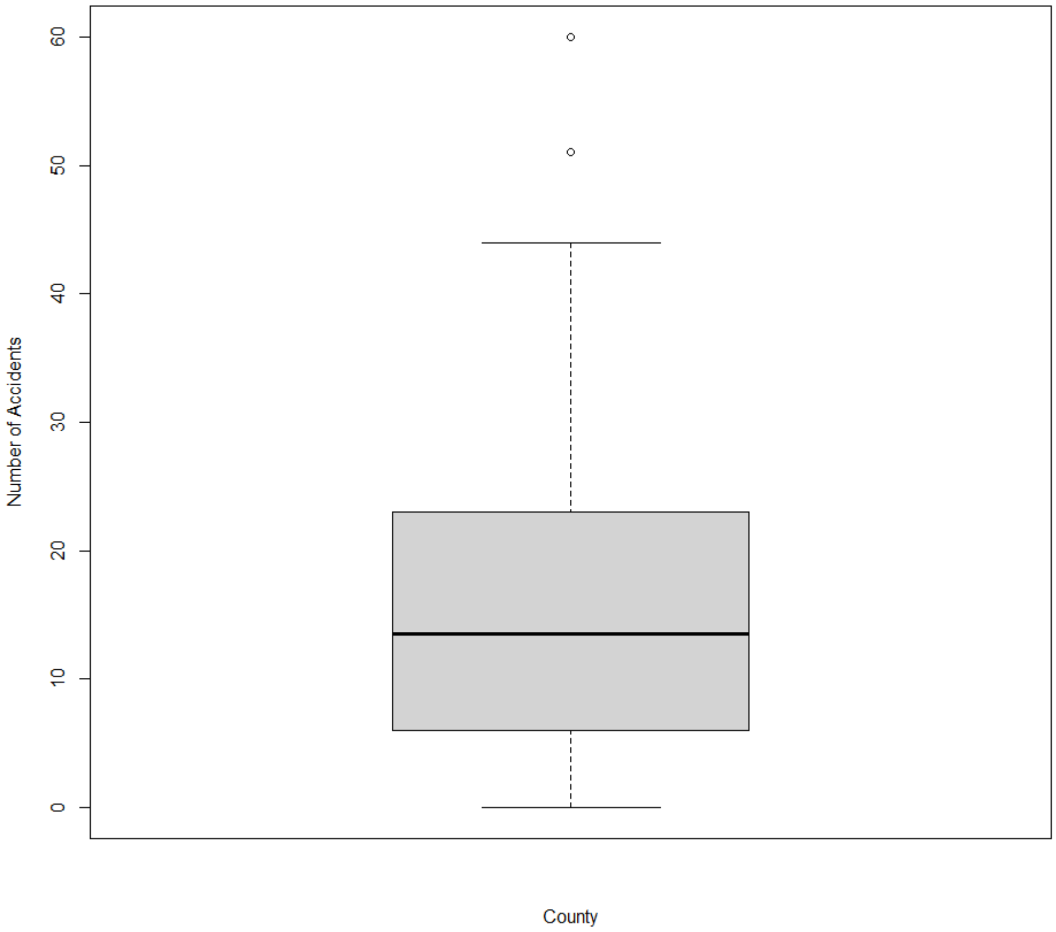
1. Median age of male runners is greater than that of female.
2. Female runners have a lot of outliers.
3. Age for female runners is slightly right skewed.

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 Car olina during 2009. Create a boxplot of data and provide relevant summary statistics. Discuss the features of the data distribution. Identify which counties may be con sidered outliers. Why might these counties have the highest numbers of motorcycle fatalities in South Carolina?

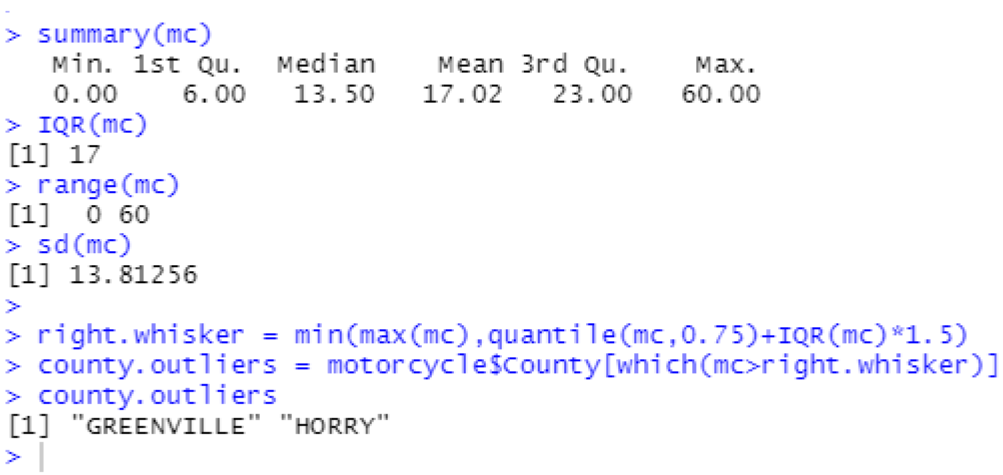
Code:

```
motorcycle = read.csv(file = "/Users/shubh/Downloads/STATS/mini_project_2/motorcycle.csv")
mc = motorcycle$Fatal.Motorcycle.Accidents
boxplot(mc,ylab ="Number of Accidents",xlab="County")

summary(mc)
IQR(mc)
range(mc)
sd(mc)

right.whisker = min(max(mc),quantile(mc,0.75)+IQR(mc)*1.5)
county.outliers = motorcycle$County[which(mc>right.whisker)]
county.outliers
```

Output:



Console Output:



```
> summary(mc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    6.00   13.50   17.02   23.00   60.00 
> IQR(mc)
[1] 17
> range(mc)
[1]  0 60
> sd(mc)
[1] 13.81256
> 
> right.whisker = min(max(mc),quantile(mc,0.75)+IQR(mc)*1.5)
> county.outliers = motorcycle$County[which(mc>right.whisker)]
> county.outliers
[1] "GREENVILLE" "HORRY"
> 
```

Observations:

1. Most of the counties have number of accidents ranging from 0-48.5
2. The range of accidents is from 0-60.
3. Median is around 13.5 and mean is 17.02
4. We infer that ‘Greenville’ and ‘Horry’ are the outliers because the number of accidents lie above the right whisker(Q3+1.5*IQR).
5. Insufficient Data to determine why these counties has higher accidents.