Statistical Methods for Data Science

Mini Project 2

Mini Project Group No # 11

Names of group members: Parashar Parikh, Aakash Prajapati

Contribution of each group member:

Both worked together to solve and discuss all the problems. Aakash implemented second question and Parashar implemented first question. Both did documentation together.

---

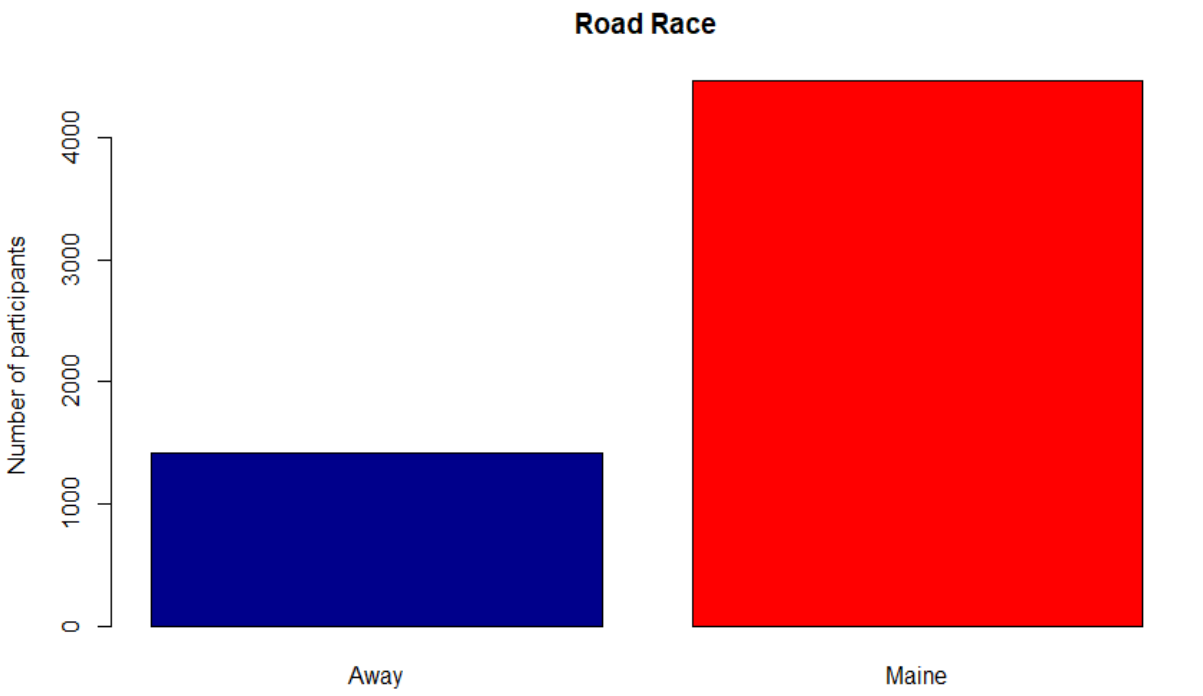
1. **(12 points) Consider the dataset roadrace.csv posted on eLearning. It contains observations on 5875 runners who finished the 2010 Beach to Beacon 10K Road Race in Cape Elizabeth, Maine. You can read the dataset in R using read.csv function.**

a) 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.
Source Code:

```
data = read.csv('C:/Users/Parashar Parikh/Desktop/UTD/Sem3/stats/Miniprojects/Miniproject_2/roadrace.csv')
barplot(table(data$Maine),main="Road Race",ylab="Number of participants",col=c("darkblue","red"))
summary(data$Maine)
```

Summary:

```
> summary(data$Maine)
 Away Maine
 1417  4458
```



Observation:

From the above bar plot, we can observe that the Away participants are in lower numbers compared to the number of participants of Maine. That also reflects in the summary statistics.

b) **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.**

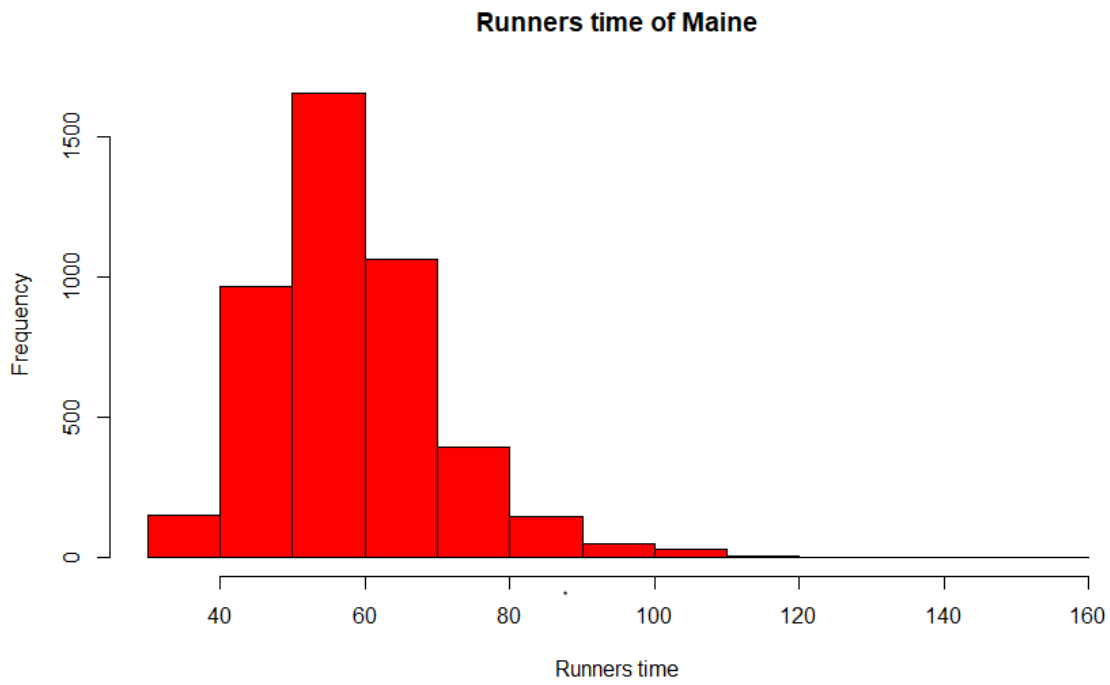
Source code:

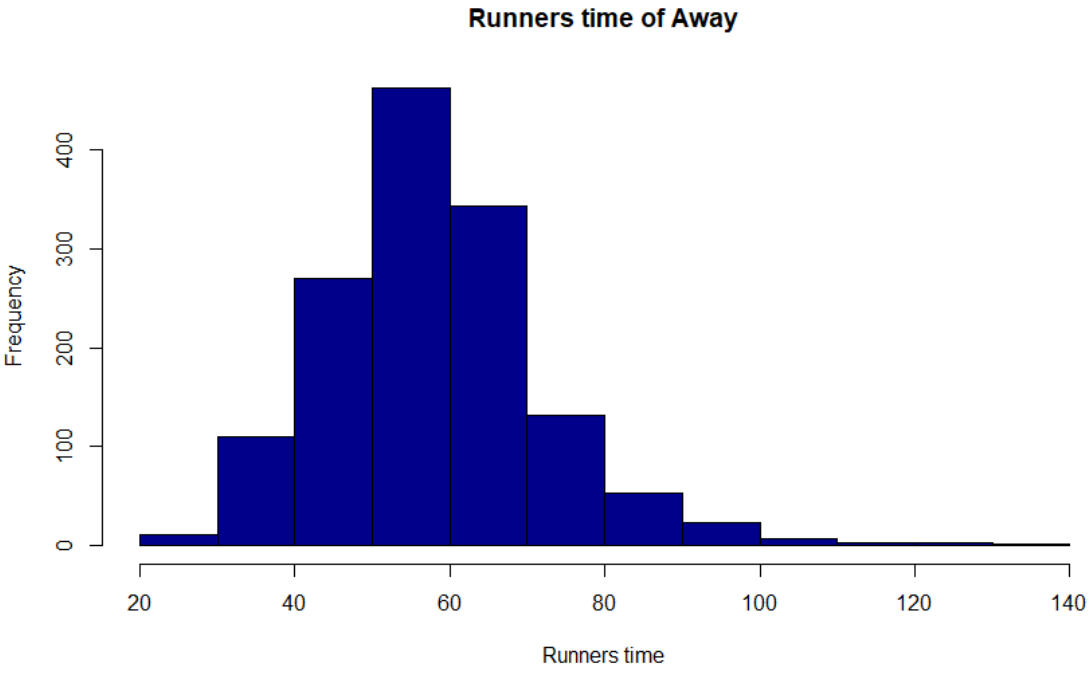
```
#selecting maine runners
maine = subset(data, Maine == "Maine")

#histogram
hist(maine[,12],xlab = "Runners time", main="Runners time of Maine",col=c("red"))

#Summary statistics of runtimes of Runners from Maine
summary(maine[,12])

#Inter quatertile range of runtime of runners from Maine
IQR(maine[,12])

#Standard Deviation of runtime of runners from Maine
sd(maine[,12])

#Range of the runtimes of runners from Maine
range = max(maine[,12])-min(maine[,12])


#selecting away candidates
away = subset(data, Maine == "Away")

#histogram
hist(away[,12],xlab = "Runners time", main="Runners time of Away",col=c("darkblue"))

#Summary Statistics of runtimes of runners for away runners
summary(away[,12])

#Inter quatertile range of runtime of runners for away runners
IQR(away[,12])

##Standard Deviation of runtime of runners for away runners
sd(away[,12])

##Range of the runtimes of runners foraway runners
range = max(away[,12])-min(away[,12])
```

Summary:

```
> #Summary statistics of runtimes of Runners from Maine
> summary(maine[,12])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.57   50.00   57.03   58.20   64.24  152.17
>
> #Inter quatertile range of runtime of runners from Maine
> IQR(maine[,12])
[1] 14.24775
>
> #Standard Deviation of runtime of runners from Maine
> sd(maine[,12])
[1] 12.18511
>
> #Range of the runtimes of runners from Maine
> range = max(maine[,12])-min(maine[,12])
> range
[1] 121.6
>
```

```
>
> #Summary Statistics of runtimes of runners for away runners
> summary(away[,12])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.78   49.15   56.92   57.82   64.83  133.71
>
> #Inter quatertile range of runtime of runners for away runners
> IQR(away[,12])
[1] 15.674
>
> ##Standard Deviation of runtime of runners for away runners
> sd(away[,12])
[1] 13.83538
>
> ##Range of the runtimes of runners foraway runners
> range = max(away[,12])-min(away[,12])
> range
[1] 105.928
>
```





Observation:
Both the histograms are in normal distributions. Both Away and Maine have same range and approximately equal mean and median and this is backed up by the summary statistics of the data.

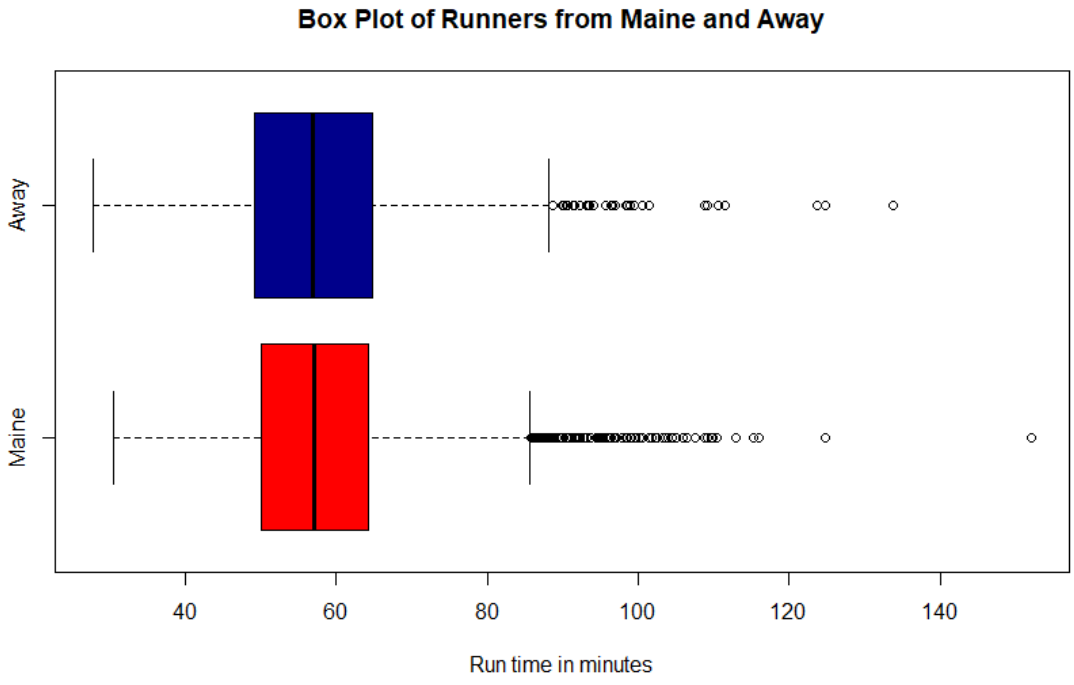
c) **Repeat (b) but with side-by-side boxplots.**

Source Code:

```
#box plot for Maine and Away side by side
boxplot(maine[,12],away[,12],names=c("Maine","Away"),col=c("red","darkblue"),horizontal=TRUE,
        main="Box Plot of Runners from Maine and Away",xlab="Run time in minutes")
```

Summary:

```
> #Summary statistics of runtimes of Runners from Maine
> summary(maine[,12])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.57   50.00   57.03   58.20   64.24  152.17
>
> #Inter quatertile range of runtime of runners from Maine
> IQR(maine[,12])
[1] 14.24775
>
> #Standard Deviation of runtime of runners from Maine
> sd(maine[,12])
[1] 12.18511
>
> #Range of the runtimes of runners from Maine
> range = max(maine[,12])-min(maine[,12])
> range
[1] 121.6
>
```



Observation:

From the above side by side box plot and from the summary we can conclude that both Away and Maine are approximately equal.

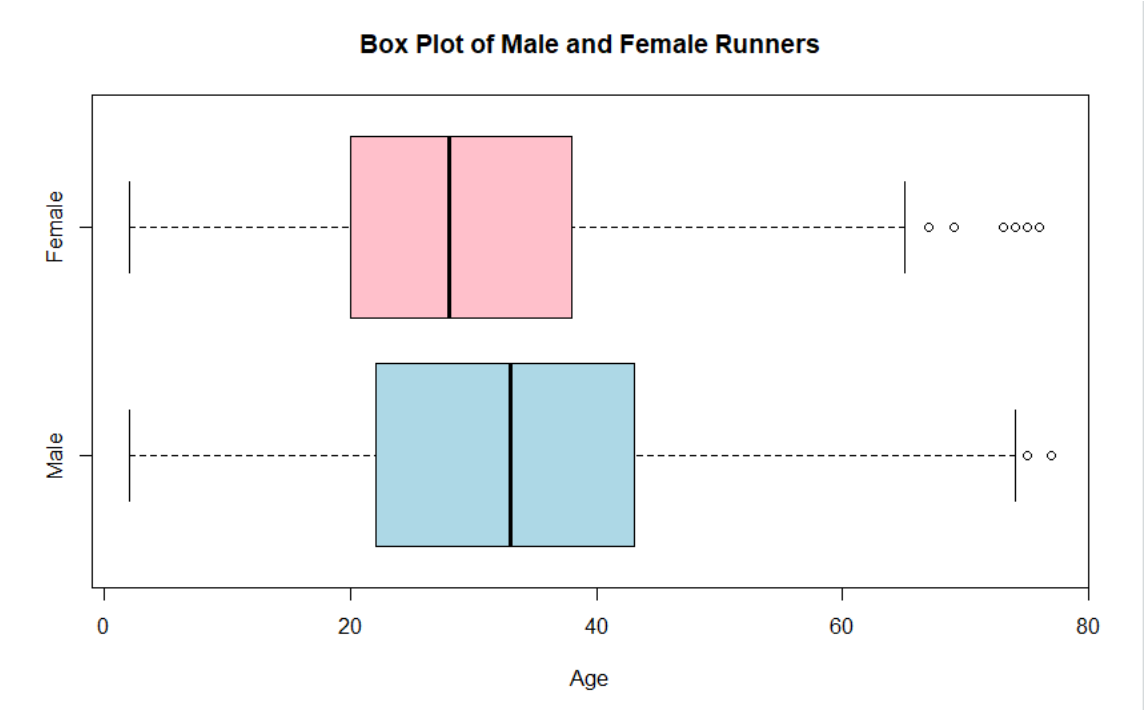
d) **Create side-by-side boxplots for the runners' ages (given in years) for male and female runners. What can we conclude about the two distributions? Back up your conclusions with relevant summary statistics, including mean, standard deviation, range, median, and interquartile range.**

Source Code:

```
#selecting male  and female runners
male = subset(data, Sex== "M")
female = subset(data, Sex== "F")

#histogram
boxplot(as.numeric(male[,5]),as.numeric(female[,5]),names=c("Male","Female"),col=c("lightblue","pink"),horizontal=TRUE,
        main="Box Plot of Male and Female Runners",xlab="Age")

#Summary statistics of Male runner
summary(as.numeric(male[,5]))

#Inter quatertile range of Male runners
IQR(as.numeric(male[,5]))

#Standard Deviation of Male runners
sd(as.numeric(male[,5]))

#Range of the runtimes male
range = max(as.numeric(male[,5]))-min(as.numeric(male[,5]))
range
#Summary statistics of female runner
summary(as.numeric(female[,5]))

#Inter quatertile range of female runners
IQR(as.numeric(female[,5]))

#Standard Deviation of female runners
sd(as.numeric(female[,5]))

#Range of the runtimes female
range = max(as.numeric(female[,5]))-min(as.numeric(female[,5]))
range
```

Summary:

```
> #Summary statistics of Male runner
> summary(as.numeric(male[,5]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00   22.00   33.00   32.56   43.00   77.00
>
> #Inter quatertile range of Male runners
> IQR(as.numeric(male[,5]))
[1] 21
>
> #Standard Deviation of Male runners
> sd(as.numeric(male[,5]))
[1] 14.07031
>
> #Range of the runtimes male
> range = max(as.numeric(male[,5]))-min(as.numeric(male[,5]))
> range
[1] 75
> #Summary statistics of female runner
> summary(as.numeric(female[,5]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00   20.00   28.00   29.26   38.00   76.00
>
> #Inter quatertile range of female runners
> IQR(as.numeric(female[,5]))
[1] 18
>
> #Standard Deviation of female runners
> sd(as.numeric(female[,5]))
[1] 12.28545
>
> #Range of the runtimes female
> range = max(as.numeric(female[,5]))-min(as.numeric(female[,5]))
> range
[1] 74
>
```



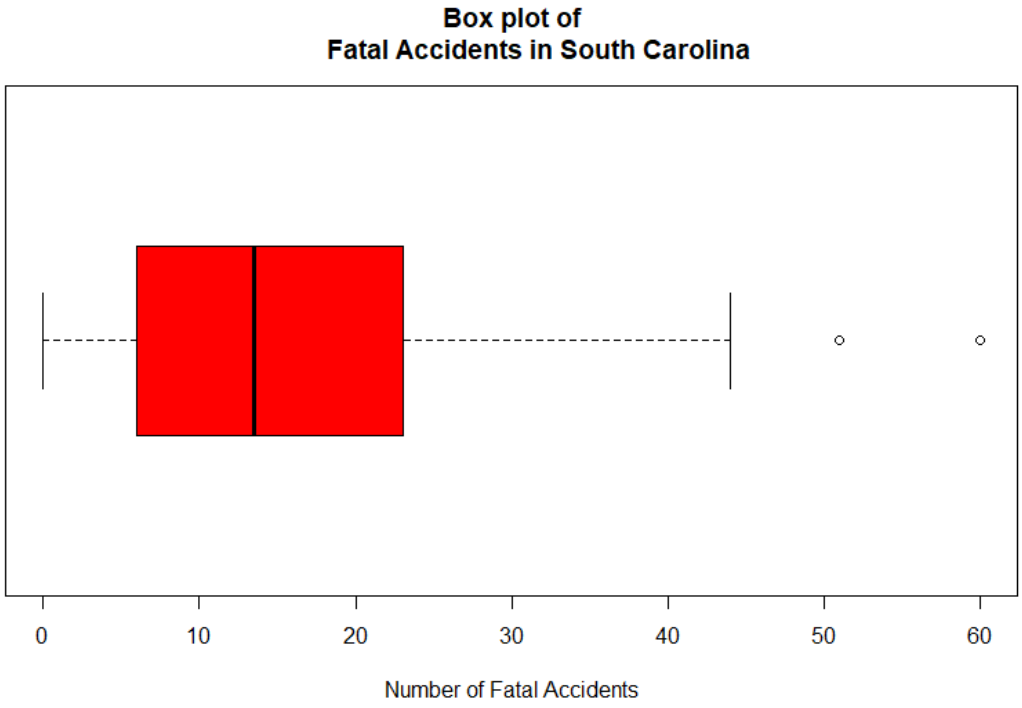
Observation:

From the above boxplot, we can observe that there is age difference between male runner and female runner. Average age of male runner is more than the average female runner and same goes for the range of ages. Both the conclusions are backed up by the summary statistics.

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

Source Code:

```
data = read.csv('C:/Users/Parashar Parikh/Desktop/UTD/Sem3/stats/Miniprojects/Miniproject_2/motorcycle.csv')
#Summary statistics of fatal number of accidents
summary(data[,2])

#Plotting box plot for Fatal number of accidents
boxplot(data[,2],horizontal=TRUE,xlab="Number of Fatal Accidents",main="Box plot of
        Fatal Accidents in South Carolina",col='red')

#Summary statistics of fatal number of accidents
summary = fivenum(data[,2])
summary
IQR=IQR(as.numeric(data[,2]))
#Calculating the outlier range
outlierLower = summary[2] - 1.5 * IQR
outlierHigher = summary[4] + 1.5 * IQR
##Finding the outlier
outlier = subset(data, data[,2] < outlierLower | data[,2] > outlierHigher)
outlier
```



```
>
> #Summary statistics of fatal number of accidents
> summary = fivenum(data[,2])
> summary
[1]  0.0  6.0 13.5 23.0 60.0
> IQR=IQR(as.numeric(data[,2]))
> #Calculating the outlier range
> outlierLower = summary[2] - 1.5 * IQR
> outlierHigher = summary[4] + 1.5 * IQR
> ##Finding the outlier
> outlier = subset(data, data[,2] < outlierLower | data[,2] > outlierHigher)
> outlier
       County Fatal.Motorcycle.Accidents
23 GREENVILLE                         51
26      HORRY                         60
>
```

Observation:

From the above box plot, we can observe that the data distribution is right skewered and most of the counties have fewer number of fatal accidents.
From the above box plot, we can say that 2 counties, Green Ville and Horry are outliers with 51 and 60 accidents, respectively.
Reasons for high numbers of accidents in these 2 counties:

- Poorly maintained highways / roads
- Volume of traffic on the highways
- Negligent drivers
- Travelling patterns of people
- Weather conditions
- Busy roads on pick hours