Mini Project – 2: Report

CS 6313.001 Statistical Methods for Data Science

Mini Project #2
Group Members: Harshali Dube, Arpita Kumane
Contribution: Both the team members collaborated and worked on the project together to write the R scripts. We analyzed, discussed, and efficiently worked to submit the two questions. We both checked the correctness of the output and analyzed the values to verify the results.

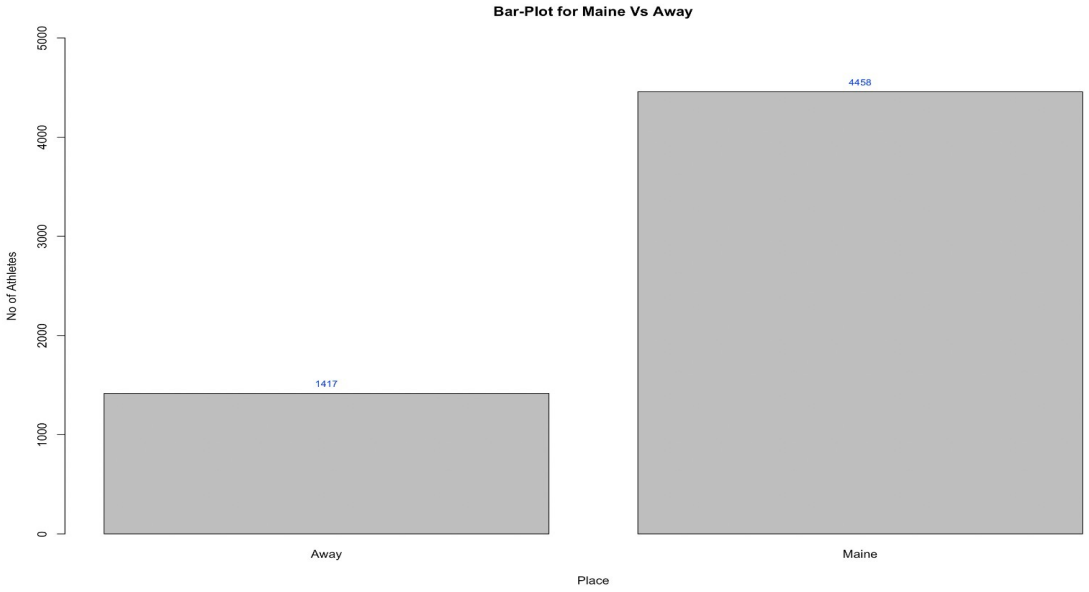
Question 1: Consider the dataset roadrace.csv posted on eLearning. It contains observations on 5875 runners who finished the 2010 Beach to Beacon 10K Road Race in Cape Elizabeth, Maine. Youcan read the dataset in R using read.csv function.

1.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 bar plot function for this. What can we conclude from the plot? Back up your conclusions with relevant summary statistics.

Solution:
R code:

```
> a = read.csv("Documents/SemIII/Stats/Harshali/roadrace.csv")
> plot = barplot (table(a$Maine),main = "Bar-Plot for Maine Vs Away",ylim = c(0,5000),xlab = "Place", ylab = "No of Athletes")
> text(x = barplot, y = table(a$Maine), label = table(a$Maine), pos = 3, cex = 0.8, col = "blue")
>
```

Bar graph:



- By observing the bar graph, we can conclude that majority of the runners are fromMaine (**75.88%**) and rest of them from other states (**24.12%**).

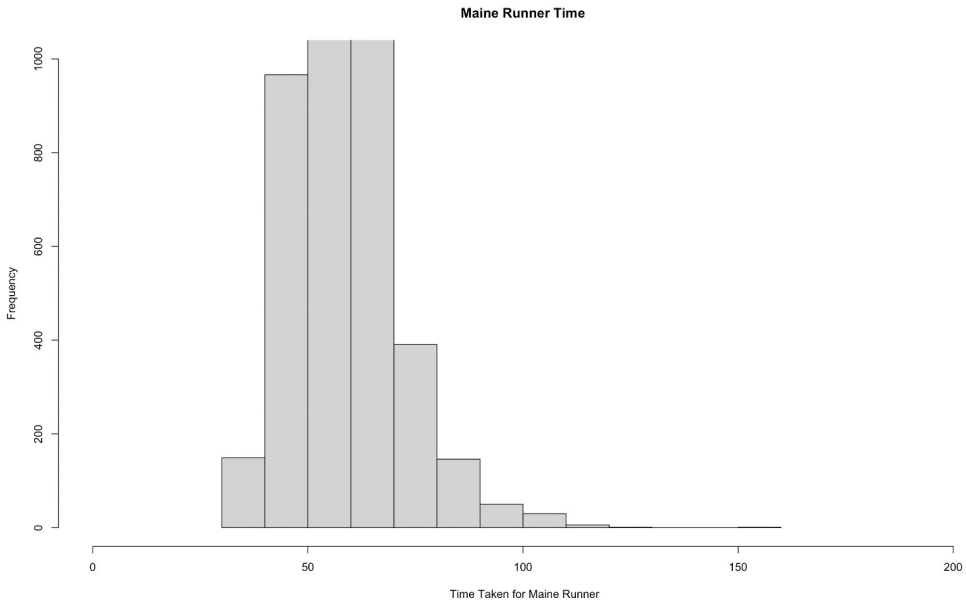
1.b Create two histograms the runners' times (given in minutes) | one for the Maine group and the second for the Away group. Make sure that the histograms on the same scale. What can we conclude about the two distributions? Back up your conclusions with relevant summary statistics, including mean, standard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

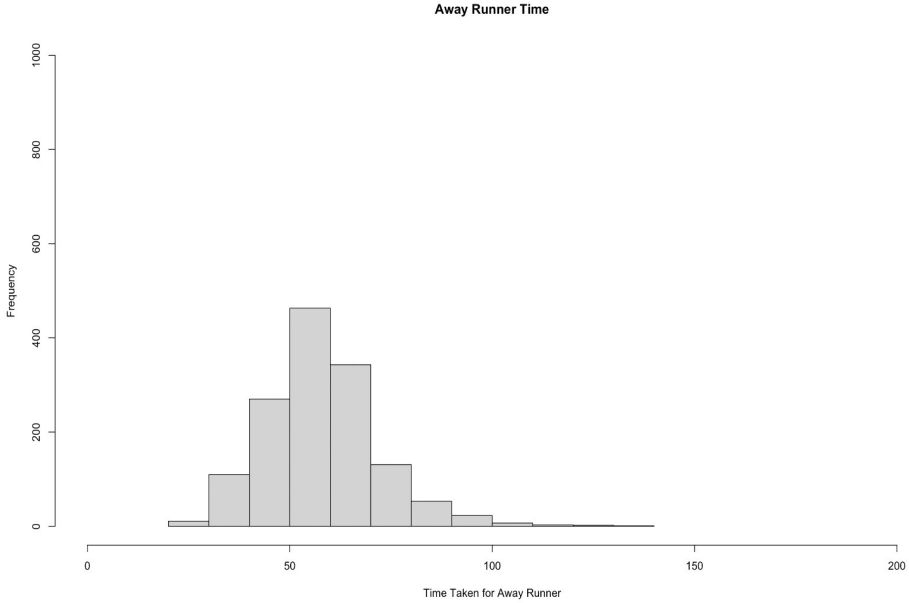
Solution:

R code: -

```
> b = read.csv("Documents/SemIII/Stats/Harshali/roadrace.csv")
> t  = a$Time..minutes.
> mt = subset(t, a$Maine == "Maine")
> at = subset(t, a$Maine == "Away")
> hist(mt, main="Maine Runner Time",ylim = range(0,1000),xlim= range(0,200),xlab=" Time Taken for Maine Runner")
> hist(at, main="Away Runner Time",ylim = range(0,1000),xlim= range(0,200),xlab="Time Taken for Away Runner")
> summary(mt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.57   50.00   57.03   58.20   64.24  152.17
> summary(at)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.78   49.15   56.92   57.82   64.83  133.71
> range(mt)
[1]  30.567 152.167
> range(at)
[1]  27.782 133.710
> IQR(mt)
[1] 14.24775
> IQR(at)
[1] 15.674
> sd(mt)
[1] 12.18511
> sd(at)
[1] 13.83538
```

Histogram:





Observation:
From the above histogram it is evident that both the distributions are similar and right skewed. Both Away and Maine groups possess almost same mean and median. It is difficult to predict which group the runner belongs to given his/her running time.

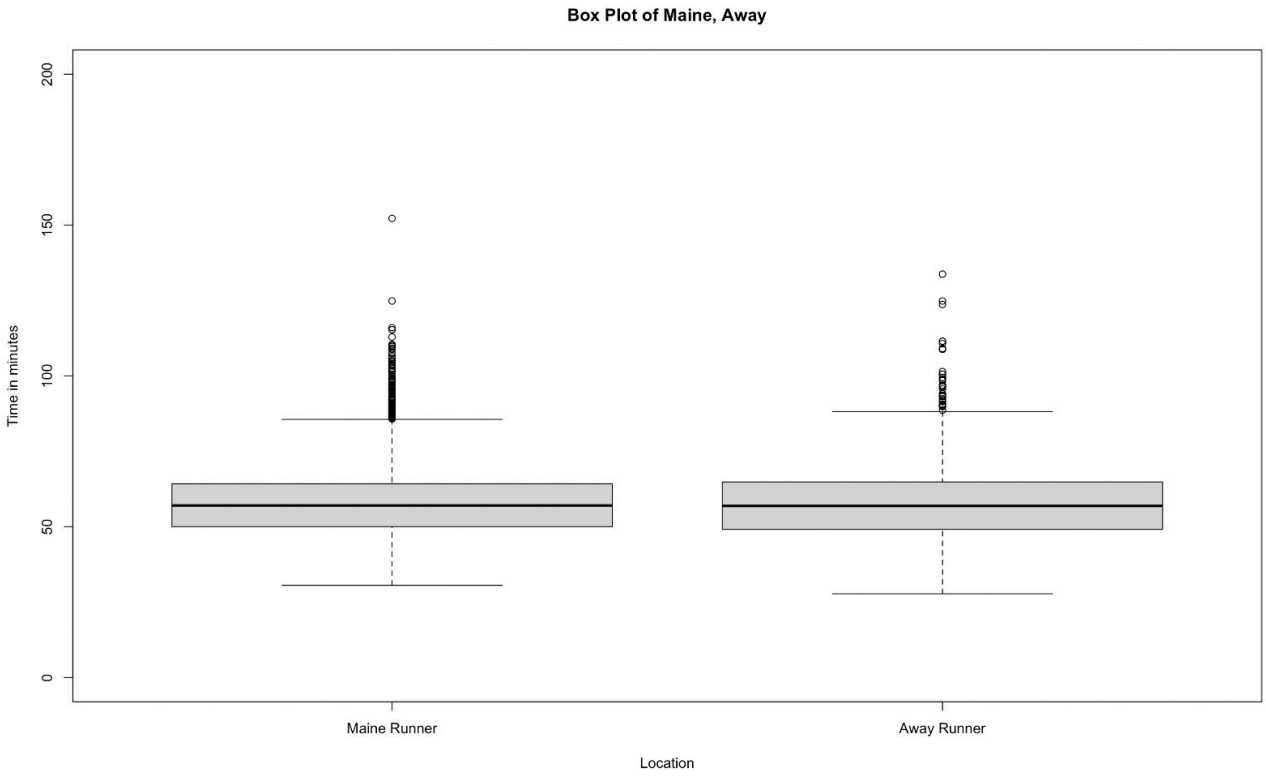
1.c Repeat (b) but with side-by-side boxplots.

Solution:

R code:

```
> c = read.csv("Documents/SemIII/Stats/Harshali/roadrace.csv")
> t  = c$Time..minutes.
> mt = subset(t, c$Maine == "Maine")
> at = subset(t, c$Maine == "Away")
> boxplot(mt,at,  ylim=range(0,200), xlab="Location", ylab="Time in minutes",main="Box Plot of Maine, Away",names=c("Maine Runner","Away Runner"))
>
```

Boxplots:



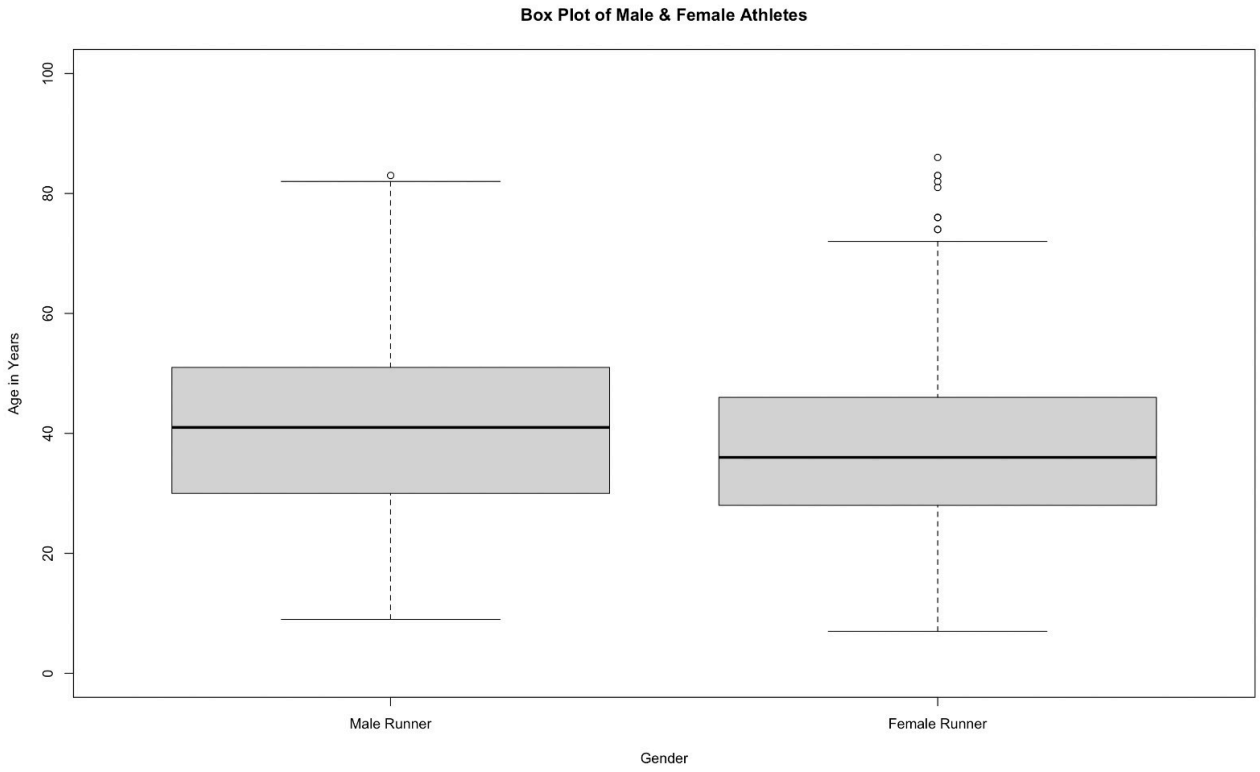
Observation:
It is evident that box plots of both Maine and Away location data is similar with the median nearly 59 and Q1 and Q3 nearly 50 and 65. They both possesses similar distribution except that Maine location has more outliers because there are more runners from Maine location.

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.
Solution:

R code:

```
> d = read.csv("Documents/SemIII/Stats/Harshali/roadrace.csv")
> age  = d$Age
> m= subset(age, d$Sex == "M")
> f= subset(age, d$Sex == "F")
> boxplot(as.numeric(m), as.numeric(f),ylim = range(0,100), xlab="Gender", ylab="Age in Years", main="Box Plot of Male & Female Atheletes",names=c("Male Runner","Female Runner"))
>
```

Boxplot:



```
d= read.csv("Documents/SemIII/Stats/Harshali/roadrace.csv")
age=d$Age
m= subset(age, d$Sex =="M")
f= subset(age, d$Sex =="F")
boxplot(as.numeric(m), as.numeric(f), vlim =range(0,100),xlab="Gender", ylab="Age", main="Box Plot of Male & Female Athletes", names=c("Male Runner", "Female Runner"))
summary(as.numeric(m))
summary(as.numeric(f))
range(as.numeric(m))
range(as.numeric(f))
IQR(as.numeric(m))
IQR(as.numeric(f))
sd(as.numeric(m))
sd(as.numeric(f))
```

```
> d= read.csv("Documents/SemIII/Stats/Harshali/roadrace.csv")
> age=d$Age
> m= subset(age, d$Sex =="M")
> f= subset(age, d$Sex =="F")
> boxplot(as.numeric(m), as.numeric(f), vlim =range(0,100),xlab="Gender", ylab="Age", main="Box Plot of Male & Female Athletes", names=c("Male Runner", "Female Runner"))
> summary(as.numeric(m))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   30.00   41.00   40.45   51.00   83.00
> summary(as.numeric(f))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   28.00   36.00   37.24   46.00   86.00
> range(as.numeric(m))
[1]  9 83
> range(as.numeric(f))
[1]  7 86
> IQR(as.numeric(m))
[1] 21
> IQR(as.numeric(f))
[1] 18
> sd(as.numeric(m))
[1] 13.99289
> sd(as.numeric(f))
[1] 12.26925
>
```

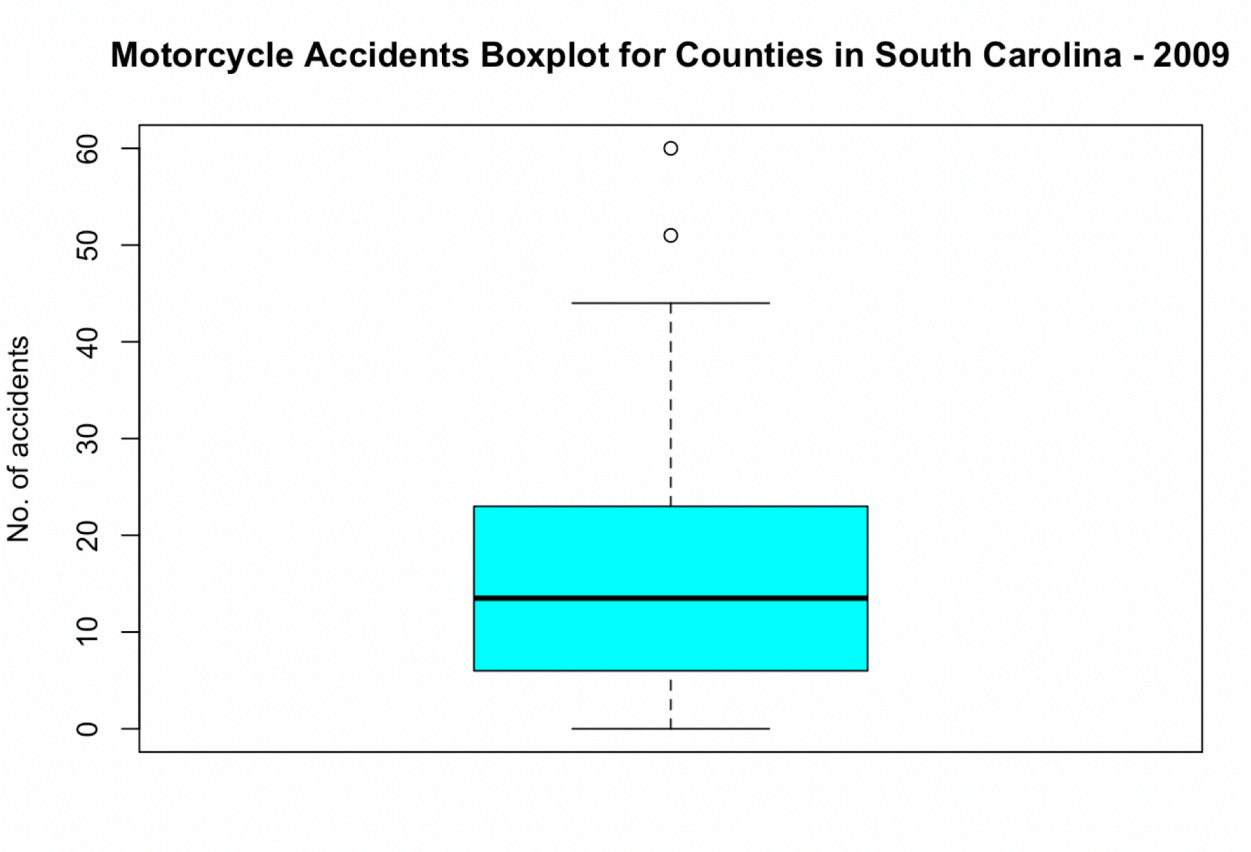
Observation:
From the above statistics female distribution is little bit right skewed whereas male distribution is little left skewed. We can see more outliers in female age and there are only two outliers in male age. Range of female age is higher than that of male age. In contrast, IQR and standard deviation of male age is higher than that of female age.

Question 2. 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?

Solution:

Boxplot:

```
> motorcycle <- read.csv("/Users/arpitakumane/Desktop/motorcycle.csv")
> boxplot(motorcycle$Fatal.Motorcycle.Accidents, col = "cyan",
+         ylab = "No. of accidents",
+         main = "Motorcycle Accidents Boxplot for Counties in South Carolina - 2009")
>
```



Summary Statistics:

```
> a = read.csv("/Users/arpitakumane/Desktop/motorcycle.csv")
> summ = summary(a$Fatal.Motorcycle.Accidents)
> print(summ)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    6.00   13.50   17.02   23.00   60.00
> IQR(a$Fatal.Motorcycle.Accidents)
[1] 17
> range(a$Fatal.Motorcycle.Accidents)
[1]  0 60
> sd(a$Fatal.Motorcycle.Accidents)
[1] 13.81256
>
```

Outliers:

```
> a= read.csv("/Users/arpitakumane/Desktop/motorcycle.csv")
> FatalAccidents = a$Fatal.Motorcycle.Accidents
> LowerBound = max(quantile(FatalAccidents, prob=0.25)- 1.5*IQR(FatalAccidents),
+                  min(FatalAccidents))
> UpperBound=min(quantile(FatalAccidents, prob=0.75) + 1.5*IQR(FatalAccidents),
+                max(FatalAccidents))
> FatalCounty=a$County[which(a$Fatal.Motorcycle.Accidents<LowerBound |
+                              a$Fatal.Motorcycle.Accidents > UpperBound)]
> FatalCounty
[1] "GREENVILLE" "HORRY"
>
```

- **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 find the outliers, we must calculate the 25th and 75th quantiles. To achieve this, quantile () function with default type 7 value is chosen.
- Upon calculating the probabilities of 25% quantile and 75% quantile they come out to be 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. There can be various reasons whythese counties have the highest number of fatalities. Without understanding features in data, we cannot exactly say what reasons contributed to this number. Since only final number is given without extracting features, we are assuming the following reasons which might be responsible for the higher number:
- Geographical terrain might be rough like hills and mountains.
- Weather conditions might be extreme during the recorded period.
- Poor road and highway maintenance.
- Negligent Drivers.