Statistical Methods for Data Science: Mini Project 2 Solved

Mini Project #: 2
Group #: 9
Names of group members: Bhargaw Rajnikant Patel, Nikita Ahuja
Contribution of each group member: Both team members worked together to solve the two problems. Both members reviewed the datasets from the problems given, worked out the solutions in R, wrote the code and finished the report in a timely manner. Both partners worked equally to complete the Mini Project 2 requirements.

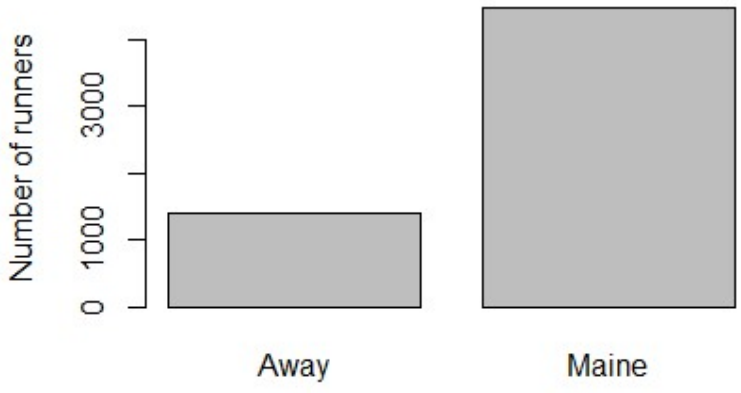
Problem 1:
First, downloaded the csv file and analysed all the field and columns. Then, read the file and stored them into variables and started programming.

a. We extracted csv file into variable, applied filtering functions to filter details and used barplot function to create bar graph of Maine and Away groups.

Also, computed the number of values for Away and Maine.
Away = 1417 Maine = 4458

Code:

```
> roadrace_data =read.csv("D:/UTD/Fall 21/6313 Stats/Mini Projects/Mini Project 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
>
```



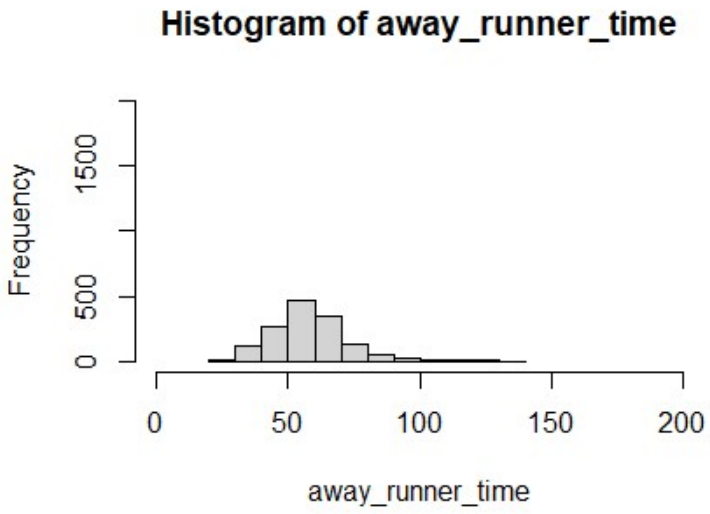
It can be concluded from the bar graph that Maine group is greater than the Away group. Out of 5875 runners, 75.8% runners are in Maine group, however that Away group hold only 24.2% portion.

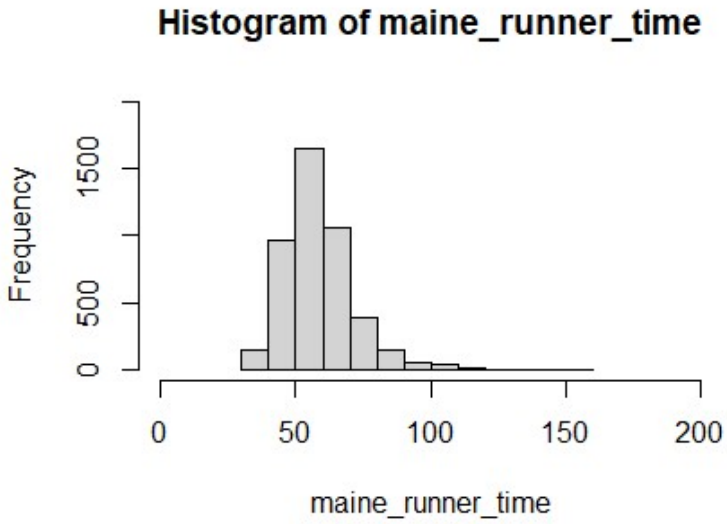
b. First, extract column using which condition and stored in variables called away_runner_time and maine_runner_time. Then use hist function to draw the histogram

Code:

```
> away_runner_time = roadrace_data$Time..minutes.[which(roadrace_data$Maine == 'Away')]
> hist(away_runner_time, xlim = range(0,200), ylim = range(0,2000))
> maine_runner_time = roadrace_data$Time..minutes.[which(roadrace_data$Maine == 'Maine')]
> hist(maine_runner_time , xlim = range(0,200), ylim = range(0,2000))
> 
```

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

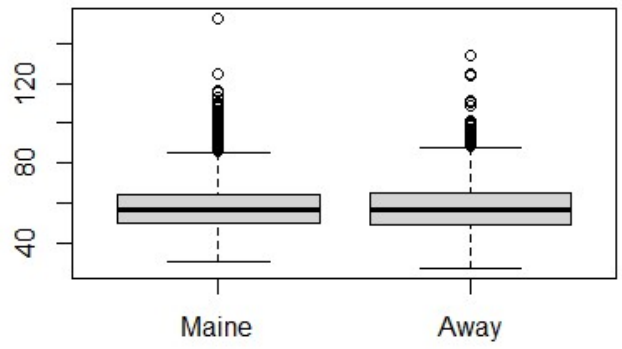




- Both distributions are tilted to the right, as shown in the graphics. The min, 1st Q, median, mean, 3rd Q, and max values can be recorded using the R function Summary (). In a tabulated format, below are the calculated values for both the Maine and Away groups.

Gender	Min	1st Quartile	Median	Mean	3rd Quartile	Max	IQR	Range	SD
Male	9	30	41	40.45	51	83	21	9 – 83	13.99289
Female	7	28	36	37.24	46	86	18	7 – 86	12.26925

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



d. Given: The dataset roadrace.csv that contains observations on 5875 runners who finished a road race in Cape Elizabeth, Maine.

```
> # Read the data from the file
> my_data = read.csv("C://Users/Nikita/Desktop/New folder/Personal Information/Statist
ical Methods for Data Science/Mini Project 2/roadrace.csv")
```

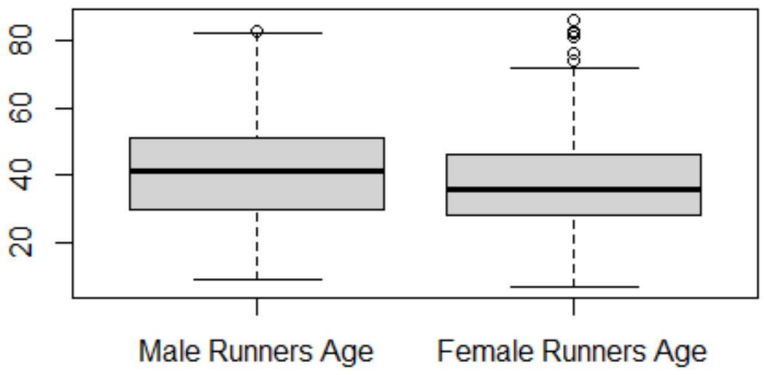
my_data	5875 obs. of 12 variables

```
> # We have to create two side by side boxplots for the runners' ages under the constraint
 that the ages are sorted by gender - i.e. Male and Female
> male_runners = as.integer(my_data$Age[which(my_data$Sex=='M')])
> female_runners = as.integer(my_data$Age[which(my_data$Sex=='F')])
> boxplot(male_runners, female_runners, names = c('Male Runners Age','Female Runners Age'))
```

male_runners	int [1:2923] 25 21 28 25 21 22 29 28 25…

female_runners	int [1:2951] 25 25 23 26 32 29 31 30 29…

Box Plot:




```
> # Let's find the summary statistics for all males
>
> summary(male_runners)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   30.00   41.00   40.45   51.00   83.00
>
> IQR(male_runners)
[1] 21
>
> range(male_runners)
[1]  9 83
>
> sd(male_runners)
[1] 13.99289
```

```
> # Let's find the summary statistics for all females
>
> summary(female_runners)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   28.00   36.00   37.24   46.00   86.00
>
> IQR(female_runners)
[1] 18
>
> range(female_runners)
[1]  7 86
>
> sd(female_runners)
[1] 12.26925
```

Derived statistical results:

Gender	Min	1st Quartile	Median	Mean	3rd Quartile	Max	IQR	Range	SD
Male	9	30	41	40.45	51	83	21	9 – 83	13.99289
Female	7	28	36	37.24	46	86	18	7 – 86	12.26925

We can conclude many observations from the statistical results for the two distributions. Firstly, the median age for the male and female runners who completed the race is 41 and 36 respectively. The mean of ages for male and female runners is 40.45 and 37.24 respectively. The statistical results derived for males in terms of ages are higher than those derived for the females as seen from the tabular form of results. However, there are female runners who were even older than the males taking part in the race – and the oldest female runner is 86. Therefore, the results have been derived for this problem.

Problem 2:

Given: The dataset motorcycle.csv that contains the number of fatal accidents that occurred in each county of South Carolina during 2009.

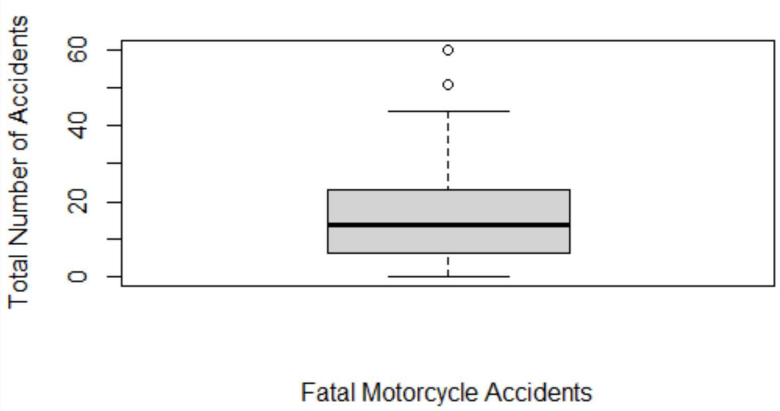
```
> # Read the data from the file
> dataset = read.csv("C://Users/Nikita/Desktop/New folder/Personal Information/Statistical Methods for Data Science/Mini Project 2/motorcycle.csv")
```

dataset	48 obs. of 2 variables

```
> # Generate the box plot for all the fatal motorcycle accidents
> fatal_accidents = dataset$Fatal.Motorcycle.Accidents
> boxplot(fatal_accidents, xlab = 'Fatal Motorcycle Accidents', ylab = 'Total Number of Accidents')
```

fatal_accidents	int [1:48] 3 28 3 35 3 7 13 38 6 44 ...

Box Plot:



It is important to calculate the 25th and 75th quantiles in order to find the outliers in the dataset. The quantile() function in R serves this purpose and is the solution to finding the upper and lower bounds. The probabilities of the 25th and 75th quantiles are 0.25 and 0.75 respectively. A county is considered to be an "outlier" if it is 1.5 times away from the interquartile range, i.e. 1.5 * IQR() is used in R.

Lower bound = 25th percentile – 1.5IQR

Upper bound = 75th percentile + 1.5IQR

```
> # In order to find which counties are the outliers, i.e. accidents < lower bound and acciden
ts > upper bound, we first calculate the lower and upper bounds
> lower_bound = max(quantile(fatal_accidents, prob=0.25)- 1.5*IQR(fatal_accidents),min(fatal_a
ccidents))
> upper_bound = min(quantile(fatal_accidents, prob=0.75) + 1.5*IQR(fatal_accidents),max(fatal_
accidents))
```

lower_bound	0
upper_bound	48.5

```
> fatal_county = dataset$County[which(dataset$Fatal.Motorcycle.Accidents < lower_bound | dat
aset$Fatal.Motorcycle.Accidents > upper_bound)]
```

fatal_county	chr [1:2] "GREENVILLE" "HORRY"

From the dataset, it is evident that Greenville County is an outlier since the number of fatal accidents here is 51, i.e. greater than the upper bound. Horry County is also an outlier since the number of fatal accidents here is 60, i.e. greater than the upper bound.

```
> # Summary Statistics
> summary(fatal_accidents)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    6.00   13.50   17.02   23.00   60.00

> # Interquartile range
> IQR(fatal_accidents)
[1] 17
```

```
> # Range
> range(fatal_accidents)
[1]  0 60

> # Standard deviation
> sd(fatal_accidents)
[1] 13.81256
```

Tabular Representation of Summary Statistics:

	Min	1st Quartile	Median	Mean	3rd Quartile	Max	IQR	Range	SD
Accidents	0	6	13.5	17.02	23	60	17	0 – 60	13.81256

The results for the problem have been derived. The Counties – Greenville and Horry are reported to have the highest numbers of motorcycle fatalities back in 2009 probably due to rash driving by negligent drivers, improper construction of roads or poor maintenance of the roads by the authorities.