**Mini Project 2 (Solution)**

**Mini Project Duo Group # 12**

**Contribution of each group member**

Chetan Siddappareddy – 50%

Ankit Sahu – 50%

Both of us have contributed equally to the project. We learnt R through collaboration and then write the R scripts for the corresponding and report all the findings.

**Section 1**

**Problem 1**

**a)** The below represents the bar plot for the variable Maine and figure 2 displays the corresponding count and proportions for each Away and Maine in Maine variable. Concluding from the data, there are 3 times more Maine runners than the away runners.

**Chart, bar chart, box and whisker chart

Description automatically generated**

**Figure 1: Bar plot for Maine Variable**

Count Maine

**Text

Description automatically generated**

Proportion Maine

**Text

Description automatically generated**

**Figure 2: Summary Statistics for Maine**

**b)** Below figure shows the histograms for the Maine and Away. They look symmetric.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Figure 2: Histogram for Away and Maine

Summary and IQR for Away and Maine are in the below Table:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Min** | **Q1** | **Median** | **Mean** | **Q3** | **Max** | **IQR** |
| **Away** | 27.78 | 49.15 | 56.92 | 57.82 | 64.83 | 133.71 | 15.67 |
| **Maine** | 30.57 | 50 | 57.03 | 58.20 | 64.24 | 152.17 | 14.24 |

**Table 1: Summary and IQR of Maine and Away**

**c)** The side-by-side plot is shown on the below figure 3. It compares the runner’s times of Maine and away. Quartile 1, Median, Quartile 3 have similar values for both Maine and away, and distributions seem to be symmetric.

Chart, box and whisker chart

Description automatically generated

**Figure 3: Side by Side boxplots for runner’s time**

**d)** Summary and IQR for male and female in shown in below table 2, and side by side box plots are shown in figure 4. It shows that all three quartiles Q1, median, and the Q3 are larger for male than the female, it shows that the distribution of male age may be different than that of female. The male runners age has larger variability than the female’s age. Also, the male runners seem to be left skewed while female runner is right skewed.

Min. 1st Qu. Median Mean 3rd Qu. Max. IQR

Male 9.00 30.00 41.00 40.45 51.00 83.00 21

Female 7.00 28.00 36.00 37.24 46.00 86.00 18

**Table 2: Summary and IQR for male and female**

**Chart, box and whisker chart

Description automatically generated**

**Figure 4: Side-by-Side boxplots for runner’s age by sex**

**PROBLEM 2:**

Figure 5 shows the box plot of motorcycle accidents. It can clearly be seen that the 75% of motorcycle accident is above 6. Although 2 states have very high number of motorcycle accidents but there are some states with no motorcycle accidents also. The distribution of motorcycle is right skew. Greenville and Horny are two outliers in the given data. The reason for motorcycle accidents is higher number of accidents are high population density, condition of weather and road, higher number of roads are few of them.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Min** | **Q1** | **Median** | **Mean** | **Q3** | **Max** | **IQR** |
| **0.00** | **6.00** | **13.50** | **17.02** | **23.00** | **60.00** | **17** |

**Table 3: Summary for Motorcycle accidents**

**Chart, box and whisker chart

Description automatically generated**

**Figure 5: Boxplot for motorcycle accidents**

**Section 2**

**####################################**

**R CODE FOR PROBLEM 1:**

**####################################**

**# Solution for Problem 1**

**# Part a**

**# Load data**

**roadrace = read.csv("/Users/sahuankit010/Desktop/Repo/CS-6313-Stats/Mini Projects/MP2/roadrace.csv")**

**# Read Data**

**print(summary(roadrace))**

**print(colnames(roadrace))**

**attach(roadrace)**

**# Bar Graph**

**barplot(table(Maine), main = "Bar graph Maine")**

**# summary of Maine**

**t <- table(Maine)**

**m <- prop.table(m)**

**print(t)**

**print(m)**

**# Part b**

**maine <- subset(roadrace, Maine == "Maine")$Time..minutes.**

**away <- subset(roadrace, Maine == "Away")$Time..minutes.**

**# Summary for both "maine" and "away"**

**summary(maine)**

**summary(away)**

**IQR(maine)**

**IQR(away)**

**# Histograms**

**hist(maine, xlim = c(min(away), max(maine)), ylim = c(0, 2000), xlab = "Time", main = "Histogram of Maine")**

**hist(away, xlim = c(min(away), max(maine)), ylim = c(0, 2000), xlab = "Time", main = "Histogram of Away")**

**# Part c Side by Side Plot**

**boxplot(Time..minutes.~Maine)**

**# Part d Male and Female Runnner Part**

**ml <- Age[Sex == "M"]**

**fl <- Age[Sex == "F"]**

**ml = strtoi(ml)**

**fl = strtoi(fl)**

**boxplot(ml, fl, names = c("M", "F"))**

**summary(ml)**

**summary(fl)**

**IQR(ml)**

**IQR(fl)**

**Table

Description automatically generated with medium confidenceGraphical user interface, text, application

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, application

Description automatically generated**

**####################################**

**R CODE FOR PROBLEM 2:**

**####################################**

**# Solution for Problem 2**

**# Read the data**

**mc <- read.csv("/Users/sahuankit010/Desktop/Repo/CS-6313-Stats/Mini Projects/MP2/motorcycle.csv")**

**attach(mc)**

**mc**

**#boxplot**

**boxplot(Fatal.Motorcycle.Accidents)**

**#outliers**

**box <-boxplot(Fatal.Motorcycle.Accidents)**

**box$out**

**tail(mc[order(Fatal.Motorcycle.Accidents), ], 2)**

**#summary statistics**

**summary(Fatal.Motorcycle.Accidents)**

**IQR(Fatal.Motorcycle.Accidents)**

**Graphical user interface

Description automatically generated with medium confidence**

**Graphical user interface, application, Word

Description automatically generated**