Statistical Methods for Data Science

Mini Project 2 (Solution)

1. (a) Figure 1 represents the barplot for the variable Maine and table 1 displays the corresponding frequencies and proportions for each category in Maine variable. These suggest that there are more than three times as many runners form Maine than somewhere else.

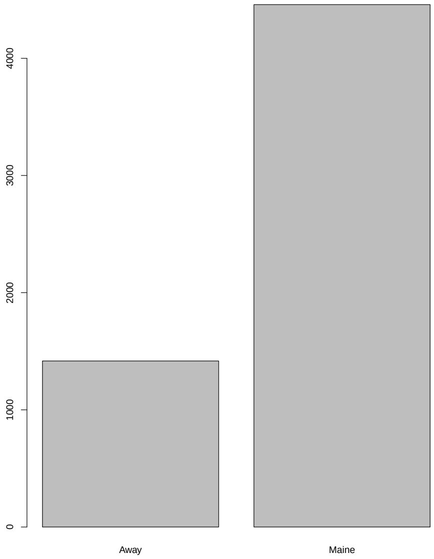

Figure 1: Barplot for Maine variable

	Away	Maine
Count	1417	4458
Proportion	0.24	0.76

Table 1: Summary statistics for Maine

(b) Figure 2 show the histograms for times for runners who are from Maine and somewhere else. We can see that both distribution are symmetric. Moreover, there are more runners from Maine than the other places. The summary statistics in table 2 confirm these findings.

	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	50.00	57.03	58.20	64.24	152.17	14.24

Table 2: Summary statistics for runner's time by Maine

(c) Figure 3 represents side-by-side boxplots for runner's times of Maine and away. Both categories have the similar values for all three quartiles — Q1, median, and Q3 implying that the two distributions are similar. Runners from both areas have unusually high running time. Both distributions seem symmetric.

(d) Table 3 shows the summary statistics of age by Sex and Figure 4 displays their side by side boxplots. We see that the estimates for all three quartiles — Q1, median, and Q3 — for Male

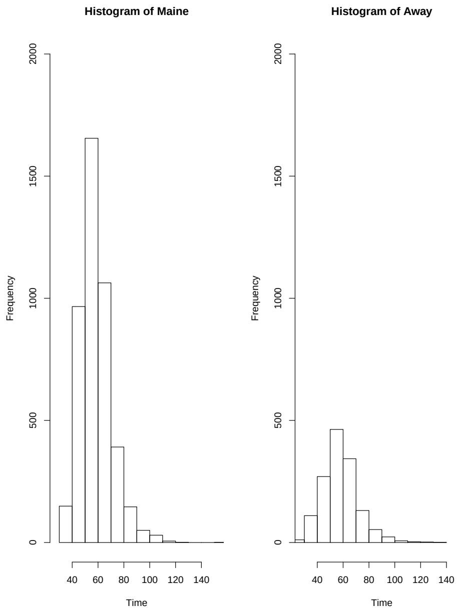

Figure 2: Histograms for Maine and Away categories

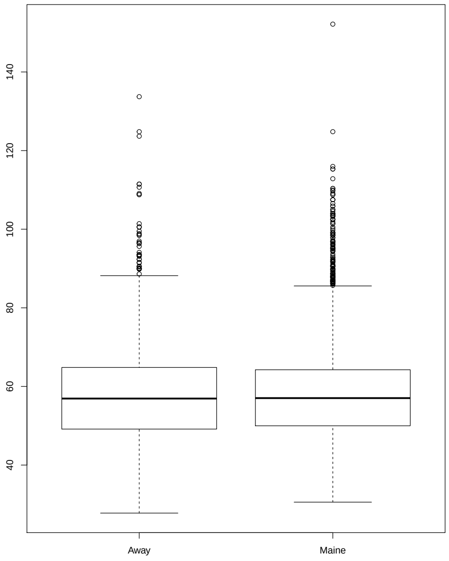

Figure 3: Side by side boxplots for runner's time by Maine

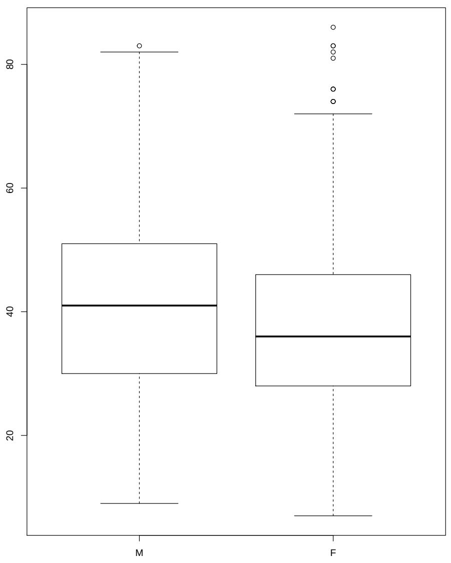

Figure 4: Side by side boxplots for runner's age by Sex

are larger than those for female, implying that the distribution of male's age may different than the female. Also male runner's age seem to have a larger variability than female runners. The distribution of age of male runners seem to be left skewed while the distribution of age of female runners is right skewed.

	Min	**Q1**	**Median**	**Mean**	**Q3**	**Max**	**IQR**
Male	9.00	30.00	41.00	40.45	51.00	83.00	21.00
Female	7.00	28.00	36.00	37.24	46.00	86.00	18.00

Table 3: Summary statistics for runner's age by Sex

2. Figure 5 represents boxplot of motorcycle accidents. We can see that about 75% of motorcycle accidents is above 6. There are some states where there is no accidents and two have unusually high number of motorcycle accidents. The distribution of motorcycle accidents seem to be right skewed.

   There are two outliers in the data, Greenville and Horry. There may be several reasons for having the highest numbers of motorcycle fatalities in these counties. Having higher number of highways, high population density, condition of weather and condition of roads are some of them.

Min	**Q1**	**Median**	**Mean**	**Q3**	**Max**	**IQR**
0.00	6.00	13.50	17.02	23.00	60.00	17.00

Table 4: Summary statistics for motorcycle accidents

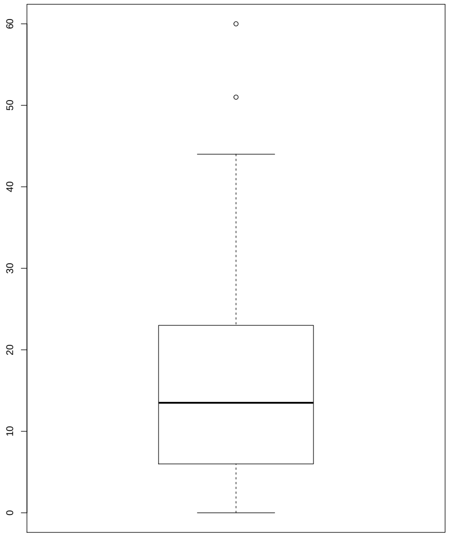

Figure 5: Boxplot for motorcycle accidents

R code:

```
########################################
# Exercise 1 #
########################################

#read data
roadrace <- read.csv("roadrace.csv", na.strings = "*")
attach(roadrace)
colnames(roadrace)

#barplot and summary statistics
barplot(table(Maine), main = "Barplot for Maine")
table(Maine)
prop.table(table(Maine))


#histrogram for runner’s time
maine <- subset(roadrace, Maine == "Maine")$Time..minutes.
away <- subset(roadrace, Maine == "Away")$Time..minutes.

#sumamry statistics
summary(maine)
IQR(maine)

summary(away)
IQR(away)

#histograms
hist(maine, xlim = c(min(a), max(m)), ylim = c(0, 2000), xlab = "Time", main = "Histogram of Maine
hist(away, xlim = c(min(a), max(m)), ylim = c(0, 2000), xlab = "Time", main = "Histogram of Away")

#side by side boxplot
boxplot(Time..minutes.˜Maine)

#side by boxplot and summary for sex
male <- Age[Sex == "M"]
female <- Age[Sex == "F"]
boxplot(male, female, names = c("M", "F"))

summary(male)
summary(female)


###################################################
# Exercise 2 #
##################################################

motor <- read.csv("motorcycle.csv")
attach(motor)
```

```
#boxplot
boxplot(Fatal.Motorcycle.Accidents)

#outliers
box <-boxplot(Fatal.Motorcycle.Accidents)
box$out
tail(motor[order(Fatal.Motorcycle.Accidents), ], 2)

#summary statistics
summary(Fatal.Motorcycle.Accidents)
```