MSDS 5043 - HW1

Olivia Samples, Lipscomb University

September 30, 2019

Load Data

Load the data here

```
hw1 <- tbl_df(faithful)
summary(hw1)
```

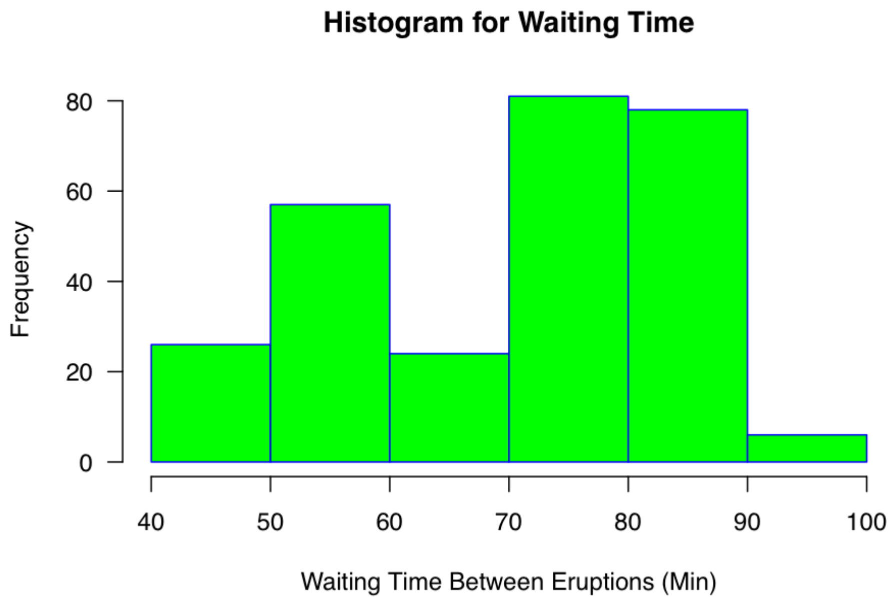
```
   eruptions        waiting    
 Min.   :1.600   Min.   :43.0  
 1st Qu.:2.163   1st Qu.:58.0  
 Median :4.000   Median :76.0  
 Mean   :3.488   Mean   :70.9  
 3rd Qu.:4.454   3rd Qu.:82.0  
 Max.   :5.100   Max.   :96.0  
```

```
 ?faithful
```

Question 1

Plot a histogram or other summary plot which meaningfully describes the distribution of the waiting time. Be sure it is very clearly labeled.

```
hist(hw1$waiting,
     main="Histogram for Waiting Time",
     xlab="Waiting Time Between Eruptions (Min)",
     border="blue",
     col="green",
     xlim=c(40,100),
     las=1,
     breaks=6)
```



Question 2

What appears to be a typical waiting time? Compare the mean, median and 80% trimmed mean (mean of the middle 80% of the observed waiting times.)

The mean of the waiting time is:

```
mean(hw1$waiting)
```

```
[1] 70.89706
```

The median of the waiting time is:

```
median(hw1$waiting)
```

```
[1] 76
```

The 80% trimmed mean of the waiting time is:

```
mean(hw1$waiting, trim = 0.8)
```

```
[1] 76
```

And so, because the trimmed mean and median were equivalent, it appears that a typical waiting time is 76 minutes.

Question 3

What is the inter-quartile range, and how does it compare to the standard deviation?

```
IQR(hw1$waiting)
```

```
[1] 24
```

```
sd(hw1$waiting)
```

```
[1] 13.59497
```

The inter-quartile range (24) is nearly double the standard deviation (13.59497), which would imply that there are approximately two standard deviations within the inter-quartile.

Question 4

Is the distribution skewed (and if so, in which direction) or is it essentially symmetric? How do you know?

From Question 2, we know that the mean is less than the median ($70.89706 < 76$), which implies that the data set is skewed left.

Question 5

Are there any unusual (outlier) values in the distribution, and if so, what are they?

```
OutVals = boxplot(hw1$waiting, plot = FALSE)$out
OutVals
```

```
numeric(0)
```

```
OutVals2 = boxplot(hw1$eruptions, plot = FALSE)$out
OutVals2
```

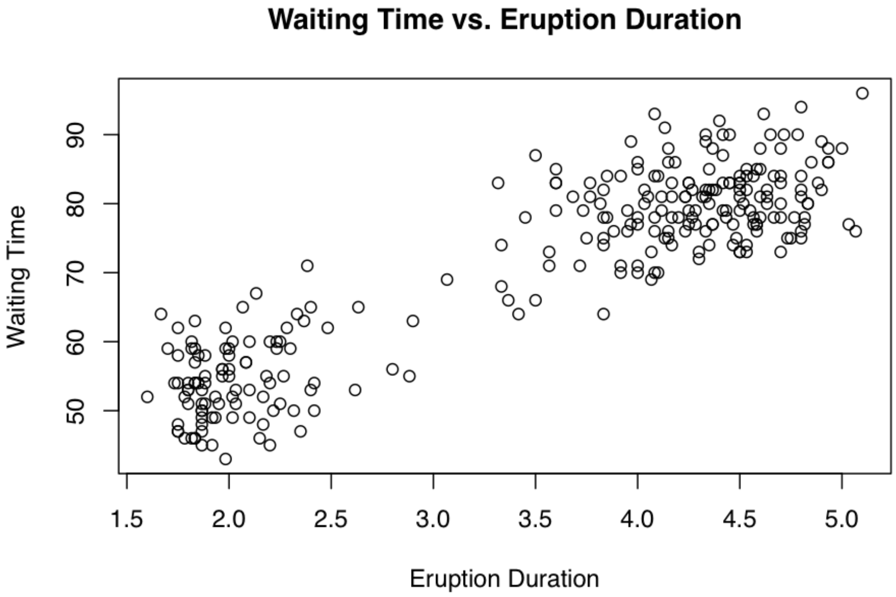
```
numeric(0)
```

And so, there are no outliers for either the waiting times or the eruption duration.

Question 6

Plot a scatterplot of the waiting times (y-axis) vs. the eruption durations (x-axis) and be sure your plot is very clearly labeled. Describe your general impression of the plot: what sort of relationship do you see?

```
plot(hw1$eruptions,hw1$waiting, main="Waiting Time vs. Eruption Duration",
     xlab="Eruption Duration", ylab= "Waiting Time")
```



The plot shows that the eruption duration is positively correlated to the waiting time. When the eruption duration is less, the waiting time is less. When the eruption duration is longer, the waiting time tends to be longer as well. Also, there appears to be a time frame for both x (2.5,3.5) and y (65,75) variables that is not strongly represented. This could imply that eruptions and waiting times tend to not fall within this gap, or we could just not have enough data to represent reality correctly.

Question 7

What is the correlation of waiting time with eruption duration? How would you interpret this result?

```
cor(faithful$eruptions, faithful$waiting)
```

```
[1] 0.9008112
```

Because the correlation is nearly 1, we know that the waiting time and eruption duration are very strongly correlated with one another in a positive direction. This is consistent with the graph.

Part 2

The second dataset is from Titanic. Please load the data and answer questions 8-10, 10 points each.

```
hw1d2 = read.csv("/Users/osamples/Desktop/MSDS5043/Titanic.csv", head = TRUE, sep = ",")
```

Question 8

What is the contingency table between Survived and Class?

```
co.table <- table(hw1d2$Survived, hw1d2$Class)
co.table
```

```
       Crew First Second Third
  Alive  212   203    118   178
  Dead   673   122    167   528
```

Question 9

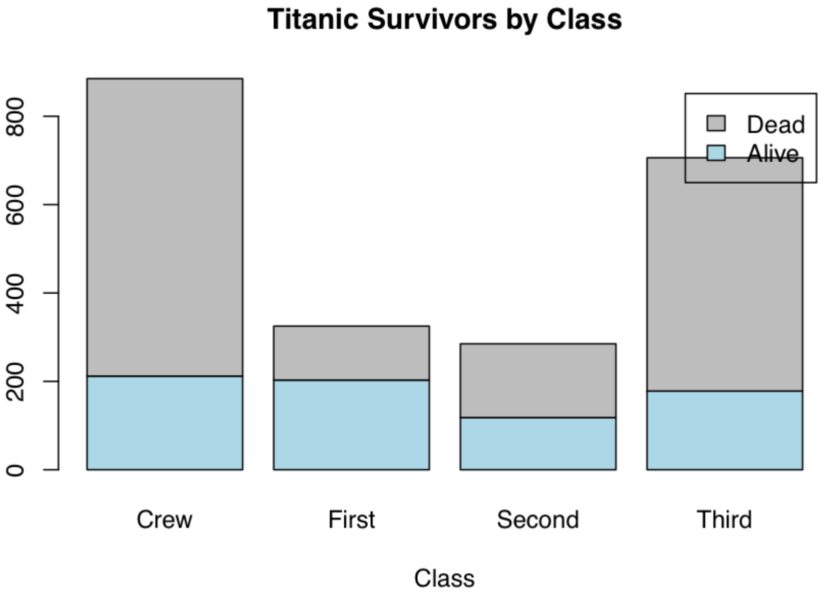
How do you interpret the contingency table in question 8?

The contigency table provides a summary of how many passengers were classified as alive or dead with respect to their class. For instance, 212 crew members survived, and 528 third class passengers died. The other columns can be read respectively.

Question 10

Bar plot for Survived and Class

```r
barplot(co.table, main="Titanic Survivors by Class",
  xlab = "Class", col=c("light blue","grey"),
  legend = rownames(co.table))
```


Titanic Survivors by Class
Dead
Alive
Crew
First
Second
Third
Class