HW 1

SDS348 Spring 2021

2021-01-31

Rose Hedderman rrh2298

This homework is due on Feb 1, 2021 at 8am. Submit a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

Question 1: (2 pts)

The dataset `faithful` contains information about eruptions of the Old Faithful geyser in Yellowstone National Park. The first few observations are listed below.

```
head(faithful)
```

```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```

```
View(faithful)
```

How many observations are there of each variable (i.e., how many rows are there)? What exactly do these variables measure? *Use a command to get information about the dataset.*

```
# The nrow() function will tell the number of rows in the dataset
nrow(faithful)
```

```
## [1] 272
```

```
?faithful
```

There are 272 observations of each of the 2 variables. One variable is called 'eruptions' which is the eruption time in minutes. The data was originally taken in seconds and is rounded ot minutes. The other variable is called 'waiting' which is the time until the next eruption in minutes.

Question 2: (7 pts)

2.1 (5 pts) What are the minimum, maximum, mean, and median values for each variable? Write sentences to describe these values. Note that there are many functions that can be used to answer this question.

```
summary(faithful$eruptions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.600   2.163   4.000   3.488   4.454   5.100
```

```
summary(faithful$waiting)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    43.0    58.0    76.0    70.9    82.0    96.0
```

The minimum eruption time is 1.6 minutes, the maximum eruption time is 5.1 minutes, the median eruption time is 4.0 minutes, and the mean eruption time is 3.488 minutes. The minimum waiting time is 43.0 minutes, the maximum waiting time is 96.0 minutes, the median waiting time is 76.0 minutes, and the mean waiting time is 70.9 minutes.

2.2 (2 pts) Create a table (using - and |) to display the statistics calculated previously for each variable.

|data type|eruptions(min)|waiting(min)| |—Min—|—-1.600—–|—-43.0—-| |–1st Q–|—-2.163—–|—-58.0—-| |—Med—|—-4.000 —–|—-76.0—-| |—Mean–|—-3.488—–|—-70.9—-| |–3rd Q–|—-4.454—–|—-82.0—-| |—Max—|—-5.100—–|—-96.0—-| —

Question 3: (6 pts)

Recall how logical indexing of a dataframe works in R. To refresh your memory, in the example code below I ask R for the number of rows in the dataset where the variable `waiting` takes on values greater than 60. Then I ask for the average of the variable `eruptions` when the variable `waiting` is above 60.

```
nrow(faithful[faithful$waiting>60,])
```

```
## [1] 189
```

```
mean(faithful[faithful$waiting>60,]$eruptions)
```

```
## [1] 4.138587
```

3.1 (1 pt) What is the comma doing in the code above (i.e., why is it necessary)?

The comma is commonly used in indexing and slicing. It is necessary for making something into a subset of rows.

3.2 (1 pt) What is the standard deviation of the variable `eruptions` ?

```
sd(faithful$eruptions)
```

```
## [1] 1.141371
```

The standard deviation of eruption time was 1.141371 minutes.

3.3 (2 pts) What is the mean of the variable `eruptions` when `waiting` is *less than* 1 hour?

```
mean(faithful[faithful$waiting < 60,]$eruptions)
```

```
## [1] 1.998273
```

The average eruption time when the waiting time is less than an hour is 1.998273 minutes.

3.4 (2 pts) What is the standard deviation of the variable `eruptions` when `waiting` is *greater than* the median?

```
sd(faithful[faithful$waiting > median(faithful$waiting),]$eruptions)
```

```
## [1] 0.3730518
```

The standard deviation of the eruption time when the waiting time is greater than the median of the eruption time is 0.3730518 minutes.

Question 4: (3 pts)

Both variables are measured in minutes. Create two new variables named `eruptions_h` and `waiting_h` that give each variable **in hours rather than minutes** and add them to the dataset `faithful`. To help get you started, I have given you code that creates both variables but fills them with `NA` values. Replace `NA` below with code on the left-hand side that computes the requested transformation. Print out the first few rows of the updated dataset using `head()`.

```
faithful$eruptions_h<-faithful$eruptions/60
faithful$waiting_h<-faithful$waiting/60
head(faithful)
```

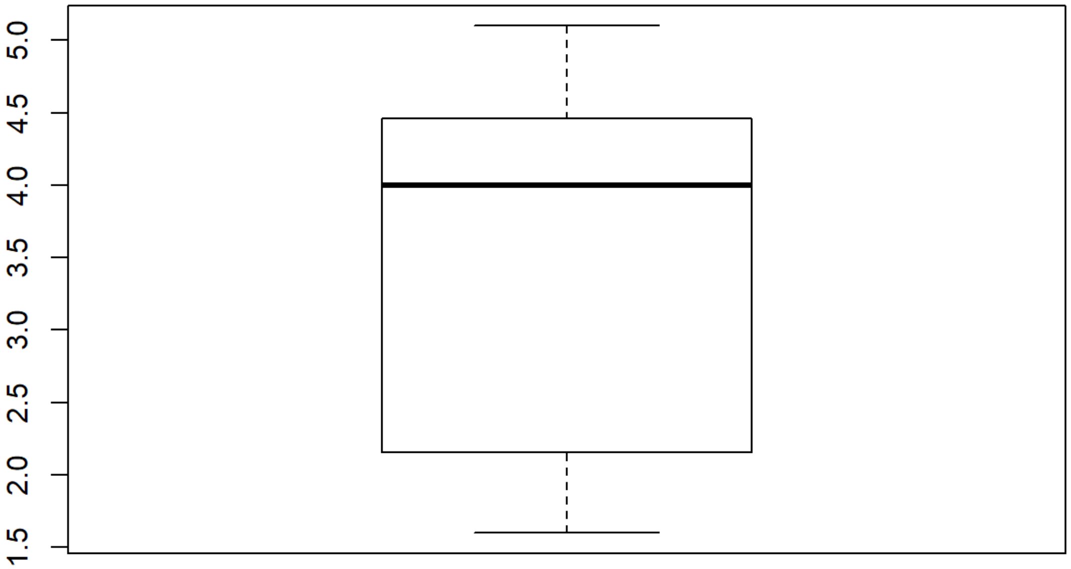
```
##   eruptions waiting eruptions_h waiting_h
## 1     3.600      79     0.06000 1.3166667
## 2     1.800      54     0.03000 0.9000000
## 3     3.333      74     0.05555 1.2333333
## 4     2.283      62     0.03805 1.0333333
## 5     4.533      85     0.07555 1.4166667
## 6     2.883      55     0.04805 0.9166667
```

Question 5: (7 pts)

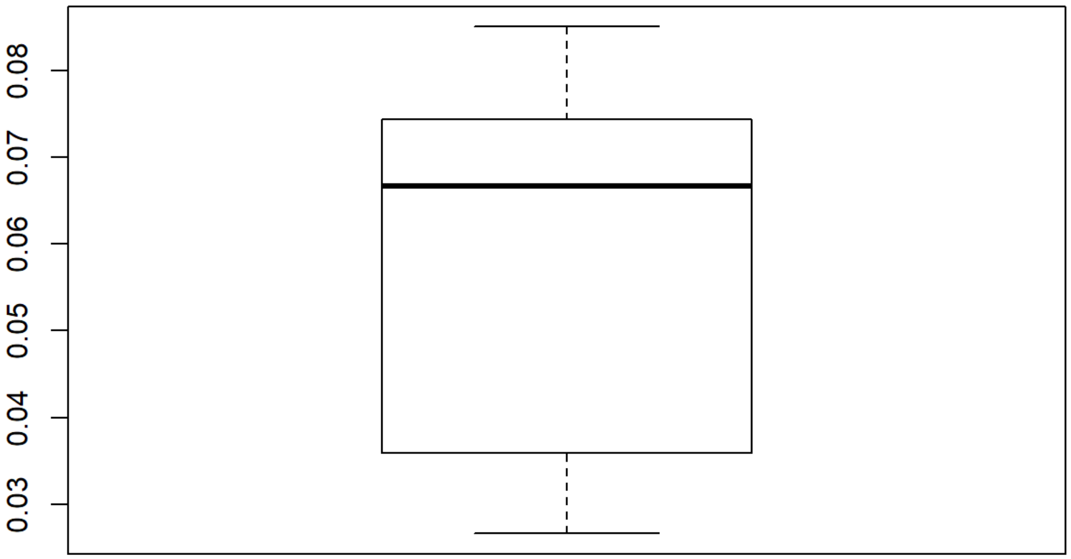
Let’s make some plots in base R.

5.1 (2 pts) Create a boxplot of each variable using the `boxplot()` function and describe the distribution of each variable (e.g., use the words symmetric, skewed, the center is around _, …). Make sure to label axes and give a title to the graph.

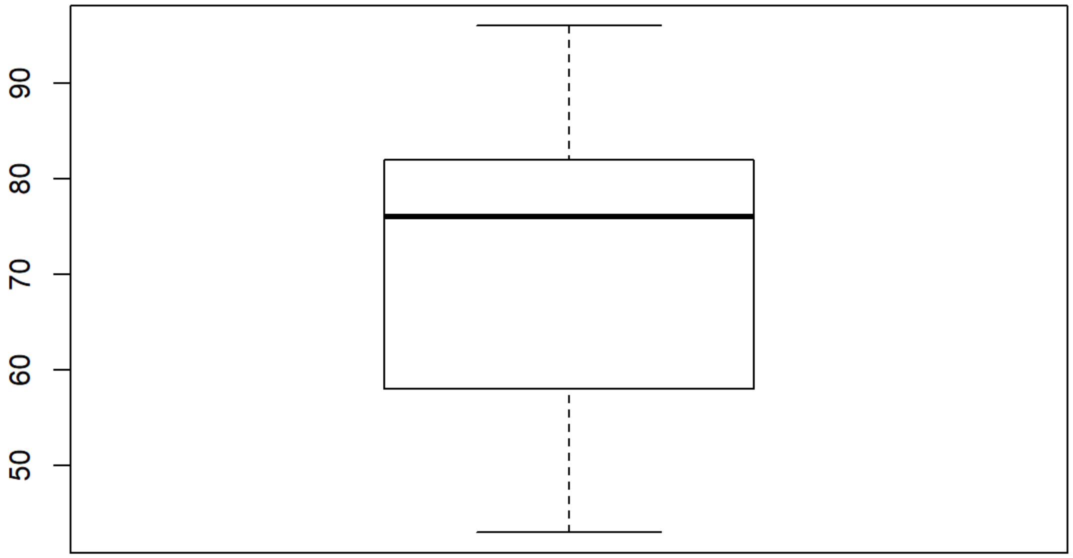
```
boxplot(faithful$eruptions)
```



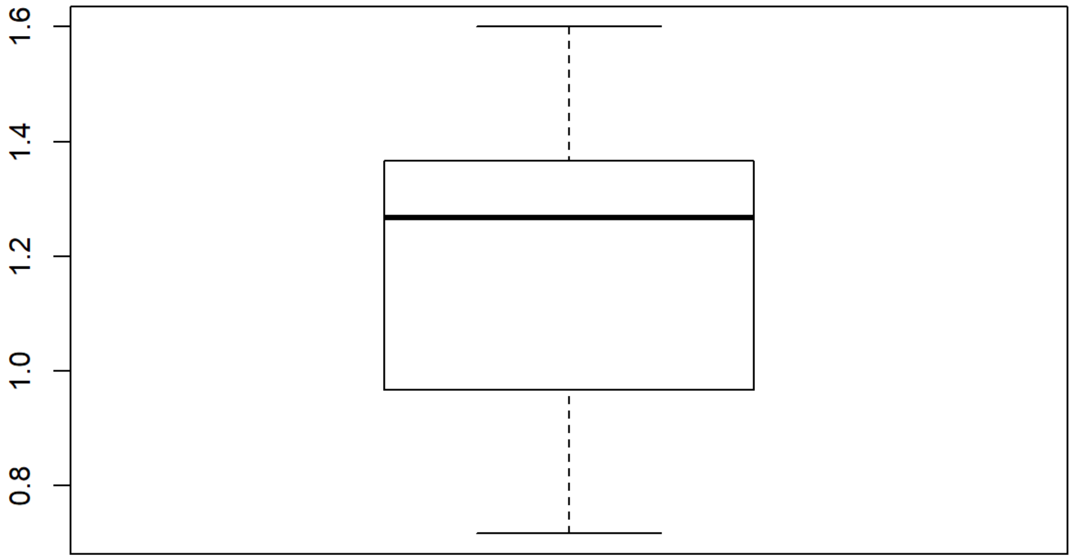
```
boxplot(faithful$eruptions_h)
```



```
boxplot(faithful$waiting)
```



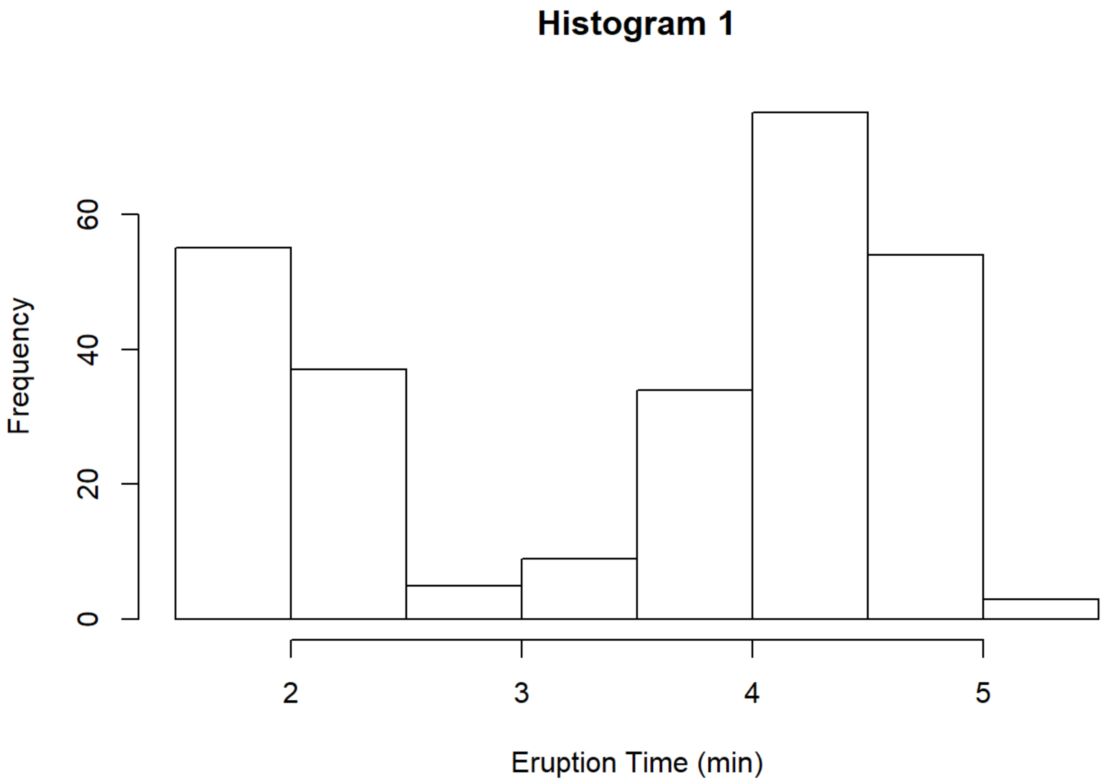
```
boxplot(faithful$waiting_h)
```



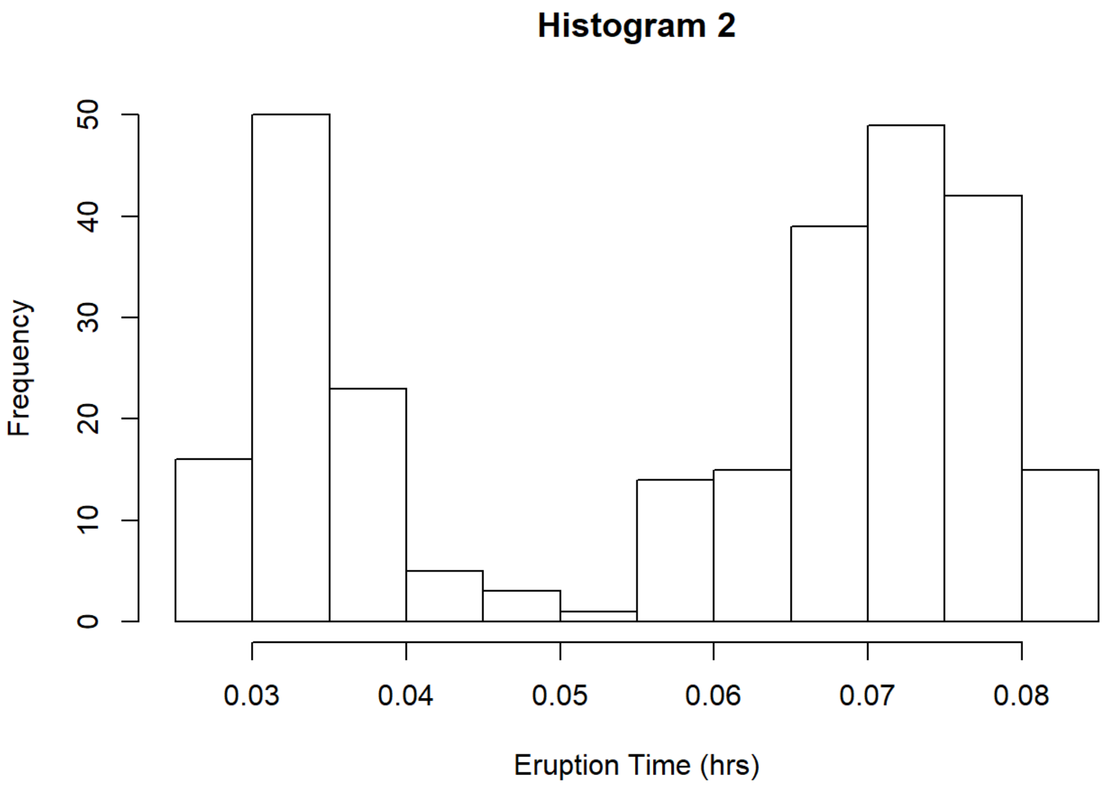
The boxplot for eruptions and eruptions_h have the same shape but different scales; the median of these plots is skewed to the right. The boxplot for waiting and waiting_h have the same shape with different scales; the median of these plots is also skewed the right.

5.2 (2 pts) Create a histogram of each variable using the `hist()` function and describe the distribution of each variable (e.g., use the words symmetric, skewed, the center is around _, …). Make sure to label axes and give a title to the graph.

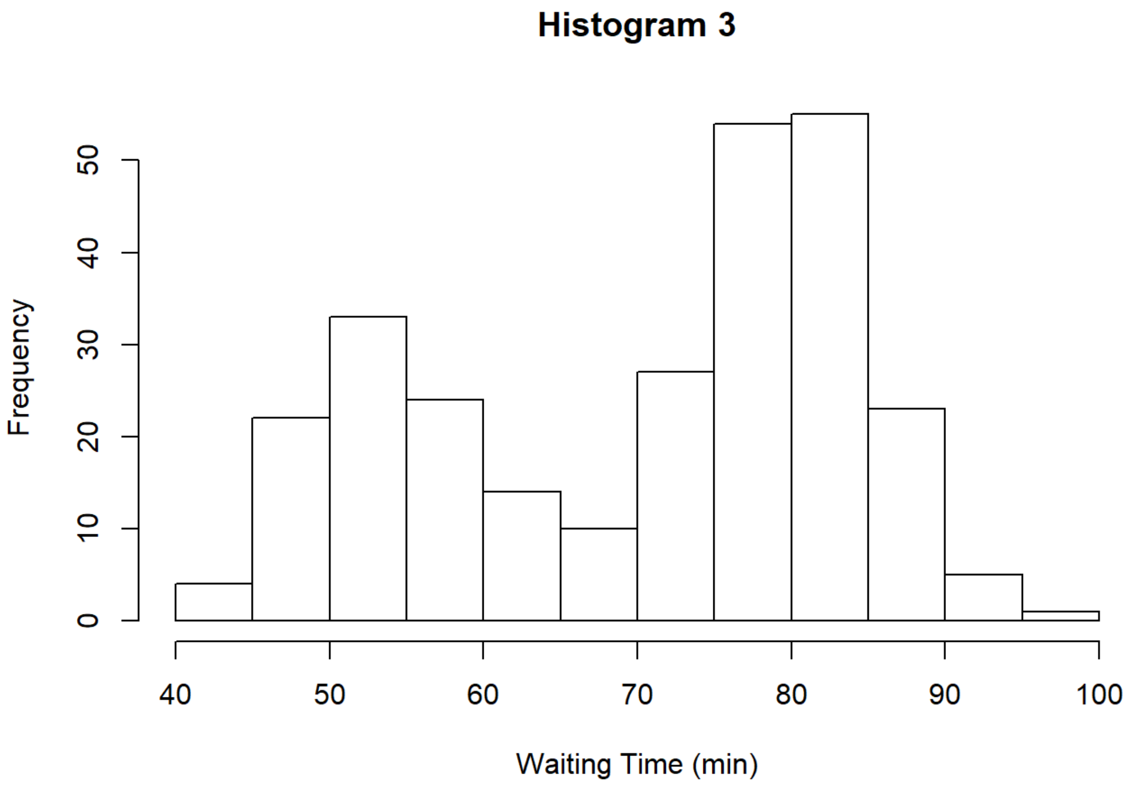
```
hist(faithful$eruptions, main = "Histogram 1", xlab = "Eruption Time (min)")
```



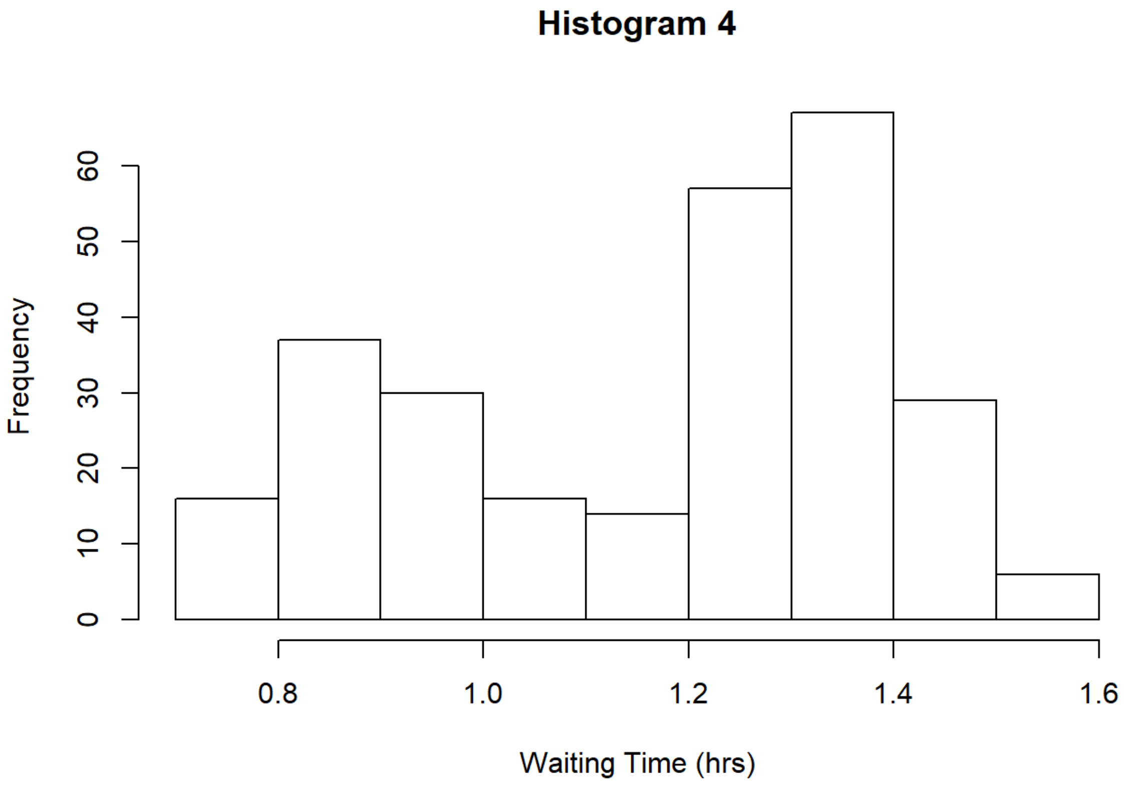
```
hist(faithful$eruptions_h, main = "Histogram 2", xlab = "Eruption Time (hrs)")
```



```
hist(faithful$waiting, main = "Histogram 3", xlab = "Waiting Time (min)")
```



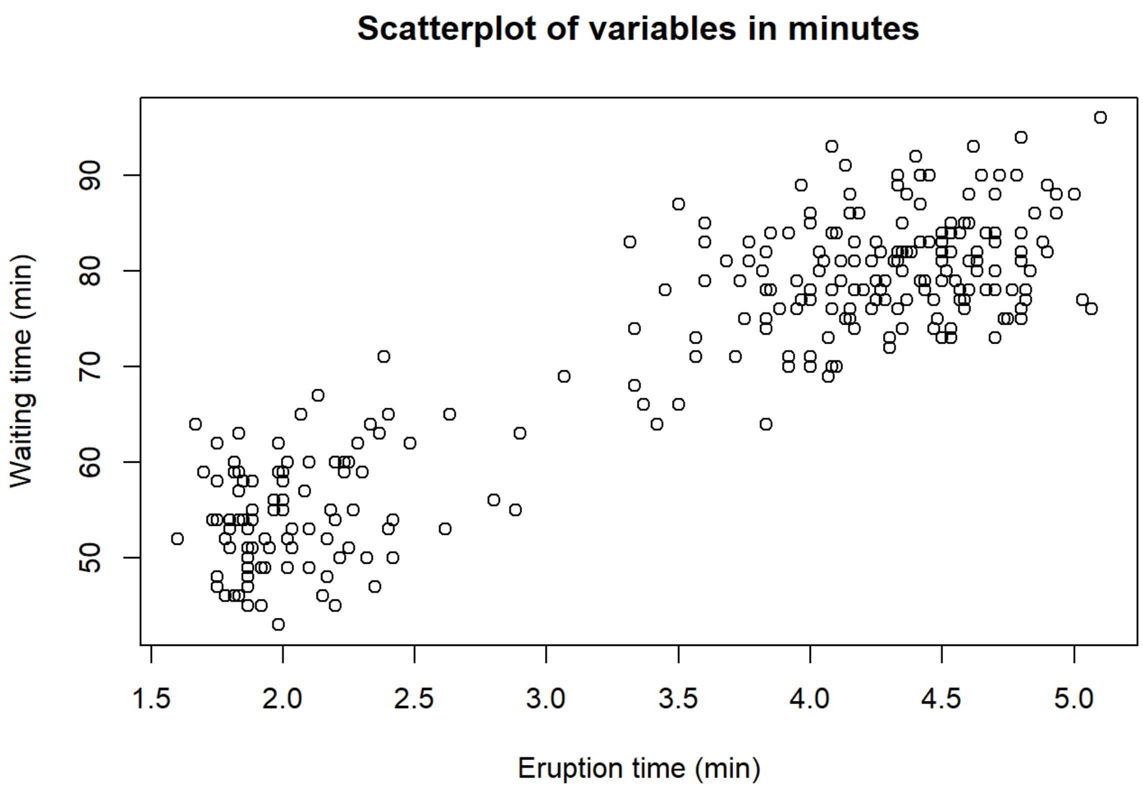
```
hist(faithful$waiting_h, main = "Histogram 4", xlab = "Waiting Time (hrs)")
```



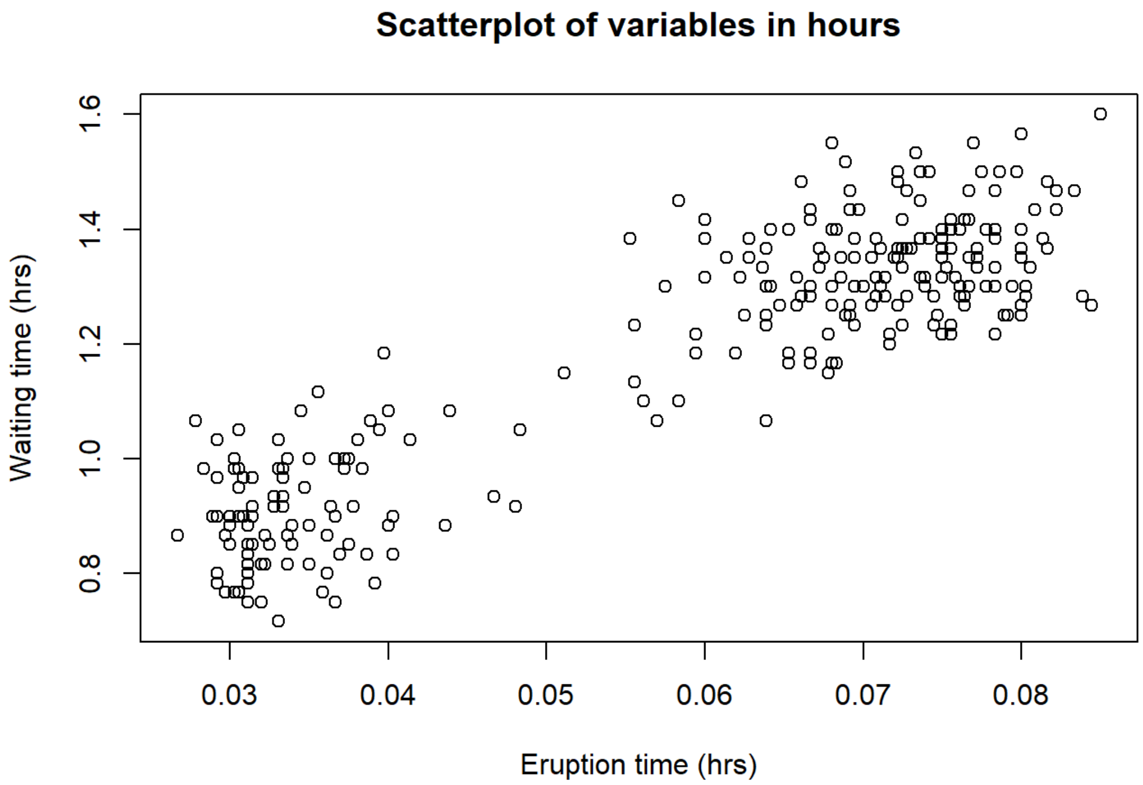
The distribution of the histogram for eruption time in minutes and in hours are both bimodal. The distribution of the histogram for waiting time in minutes and in hours are also both bimodal.

5.3 (1 pt) Create a scatterplot by plotting both variables against each other using the `plot()` function. Make sure to label axes and give a title to the graph.

```
plot(faithful$eruptions, faithful$waiting, main = "Scatterplot of variables in minutes", xlab = "Eruption t
ime (min)", ylab = "Waiting time (min)")
```



```
plot(faithful$eruptions_h, faithful$waiting_h, main = "Scatterplot of variables in hours", xlab = "Eruption
time (hrs)", ylab = "Waiting time (hrs)")
```



5.4 (2 pts) What can you see from the scatterplot that you cannot see from the histograms? What can you see from the histogram that you cannot see from the boxplots?

The scatterplots put the two variables on the same plot which a histogram cannot do. The histograms better compare the difference in frequencies within a variable more closely than the scatterplot.

```
##        sysname        release        version       nodename        machine 
##      "Windows"       "10 x64"  "build 18363"     "ROSE-XPS"       "x86-64" 
##          login           user effective_user 
##        "roseh"        "roseh"        "roseh" 
```