Lab 1

Author: Melina Padron

This assignment is due by the end of the lab. Only one student in the group submits 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.

In this lab, you will explore the dataset `faithful`. It contains information about eruptions of the Old Faithful geyser in Yellowstone National Park. The first few observations are listed below.

```
library(dplyr)
library(datasets)
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
```

Question 1: (2 pts)

How many rows are there in this dataset? How many columns? Try using the `glimpse()` function from the `tidyverse`. In which units are the variables reported? You will need more information about the dataset to answer that last question: run `?faithful` **in the console**. *Note: using `?` in your Markdown document might prevent you from knitting or will open the documentation in a new tab.*

```
#Found number of rows and columns using the glimpse function
glimpse(faithful)
```

```
## Rows: 272
## Columns: 2
## $ eruptions <dbl> 3.600, 1.800, 3.333, 2.283, 4.533, 2.883, 4.700, 3.600, 1.95~
## $ waiting   <dbl> 79, 54, 74, 62, 85, 55, 88, 85, 51, 85, 54, 84, 78, 47, 83, ~
```

```
#Found variable units using '?'
?faithful
```

Answer: There are 272 rows and 2 columns. Both variables, eruption and waiting time, are reported in minutes.

Question 2: (2 pts)

Using the function `summary()` for each variable, find the minimum, maximum, mean, and median values of each variable. Write a sentence to interpret the mean eruption duration and the mean waiting time.

```
#Minimum, maximum, mean, and median for 'eruptions' and 'waiting' time variable
summary(faithful)
```

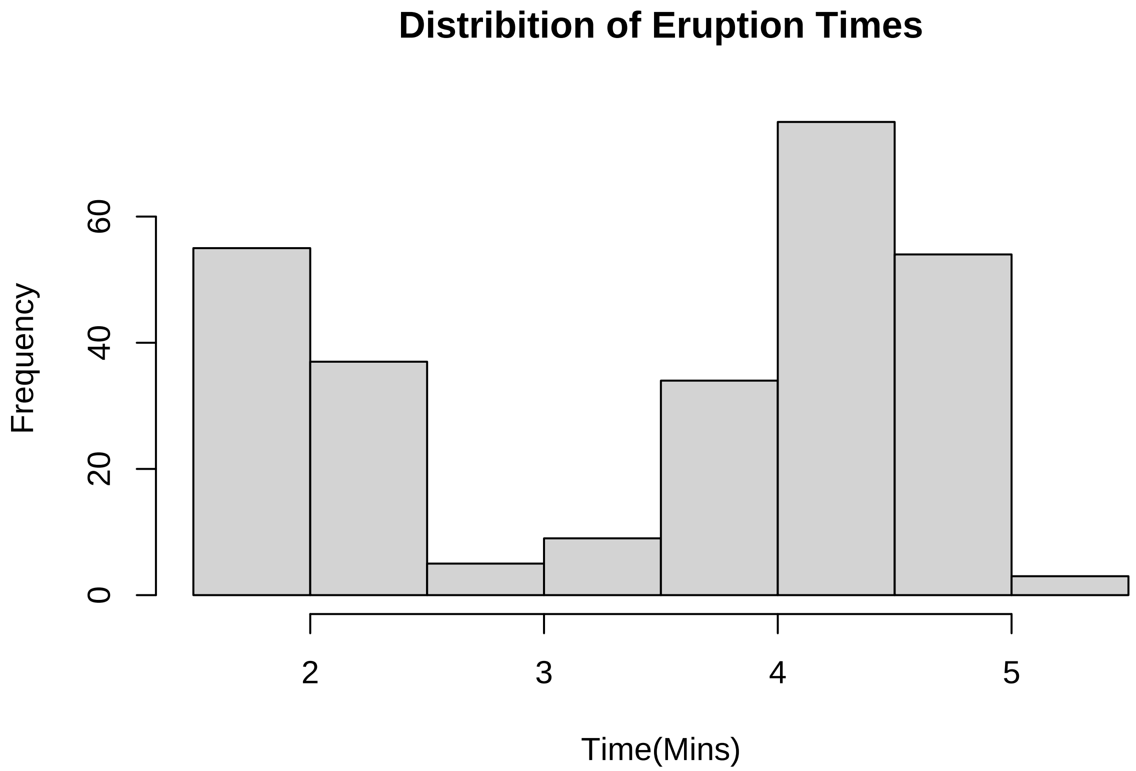
```
##    eruptions        waiting
##  Min.   :1.600   Min.   :43.0
##  1st Qu.:2.163   1st Qu.:58.0
##  Median :4.000   Median :76.0
##  Mean   :3.488   Mean   :70.9
##  3rd Qu.:4.454   3rd Qu.:82.0
##  Max.   :5.100   Max.   :96.0
```

The mean eruption duration is the average time in minutes that an eruption lasted in this dataset. The mean waiting time is the average time in minutes that occured between the end of one eruption to the start of the next eruption.

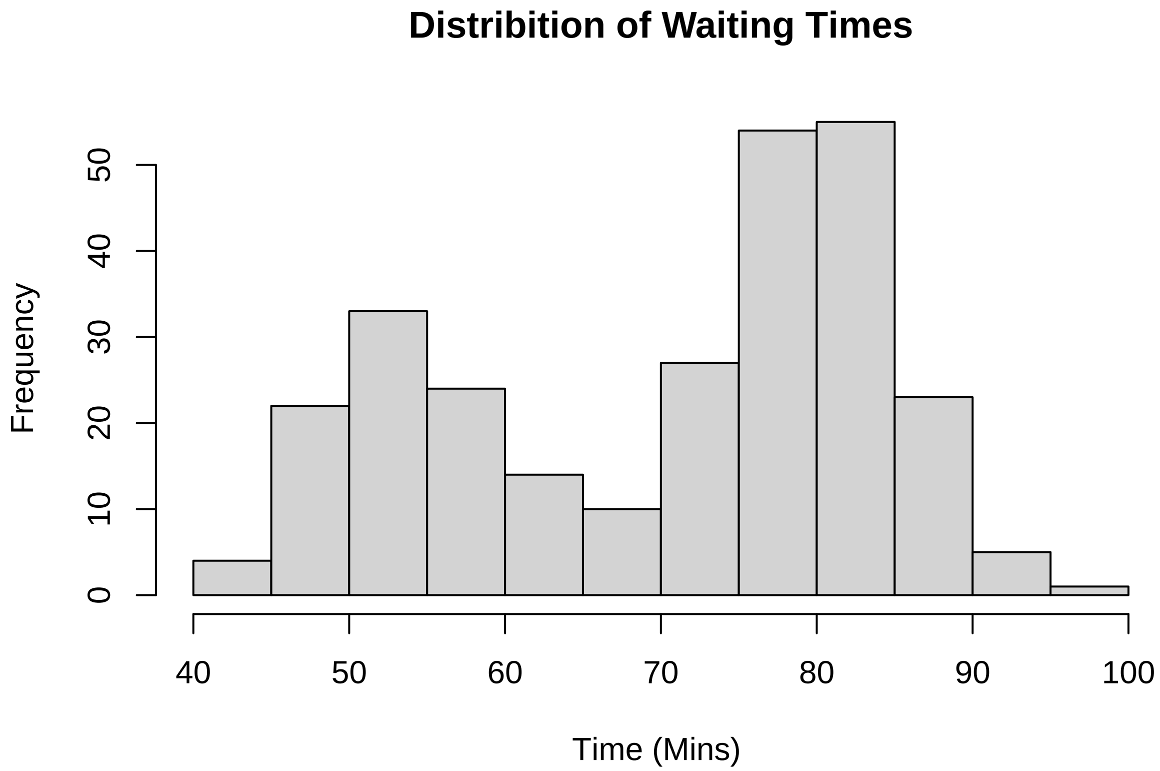
Question 3: (2 pts)

Create a histogram of each variable with the function `hist()`. (You can find the help page for `hist()` by calling `?hist` at the console.) Make sure to label axes (`xlab=`) and give a title to the graph (`main=`).

```
#Histogram of eruption time
hist(faithful$eruptions,
     main = 'Distribition of Eruption Times',
     xlab = 'Time(Mins)'
     )
```



```
#Histogram of waiting time
hist(faithful$waiting,
     main = 'Distribition of Waiting Times',
     xlab = 'Time (Mins) '
     )
```



Question 4: (2 pts)

Let's do some filtering and logical indexing. What does the code below do?

NOTE: The %>% symbol is equivalent to the |> symbol for piping the output of functions to other functions.

```
faithful %>%
        filter(waiting > 60)
```

Answer: The code above filters the Faithful Dataset and returns all of the columns and rows where the 'Waiting' variable is above 60 minutes.

Question 5: (2 pts)

Using filtering and logical indexing and the function `mean()`, find the mean of the variable `eruptions` when `waiting` is **less than or equal to** 1 hour and the mean of the variable `eruptions` when `waiting` is **more than** 1 hour. Compare the two means.

```
#The mean of the variable `eruptions` when `waiting` is less than or equal to 1 hour
faithful %>%
        filter(waiting <= 60) %>%
        summarize(mean(eruptions))
```

```
##   mean(eruptions)
## 1        2.005831
```

```
#The mean of the variable `eruptions` when `waiting` is more than 1 hour
faithful %>%
        filter(waiting > 60) %>%
        summarize(mean(eruptions))
```

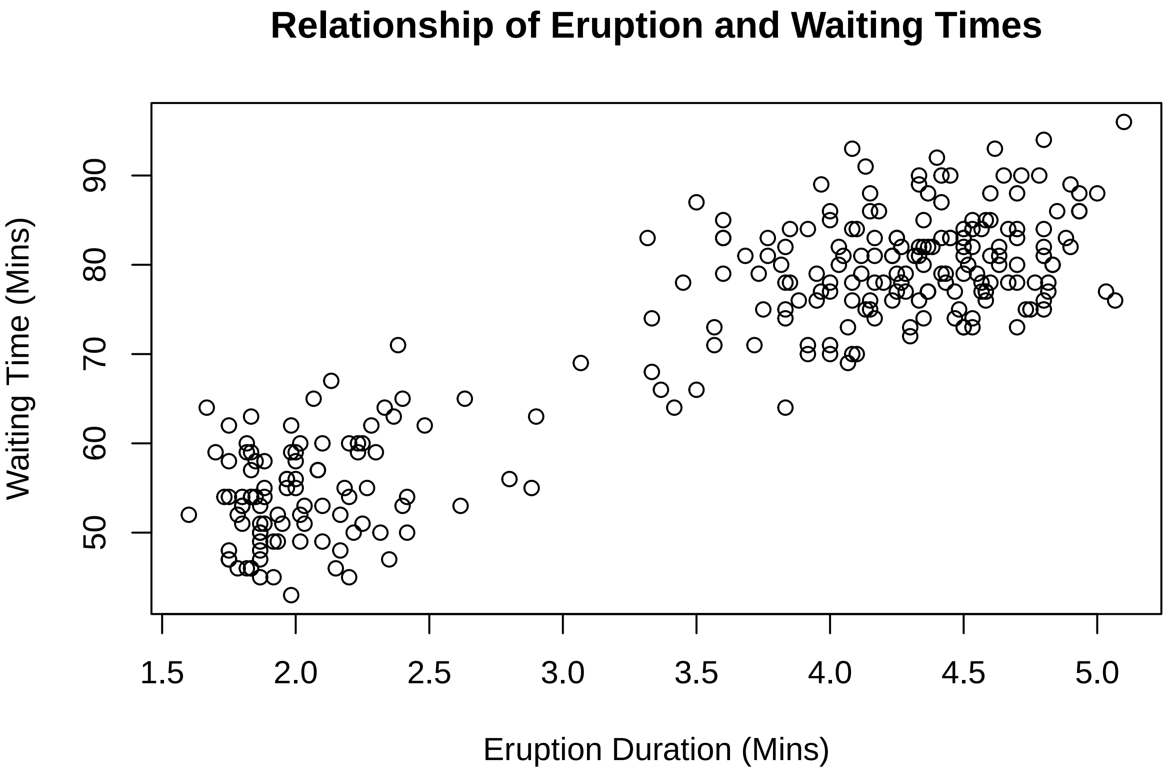
```
##   mean(eruptions)
## 1        4.138587
```

Answer: The mean of the variable **`eruptions`** *when* **`waiting`** *is less than or equal to 1 hour is 2.005831 minutes, while the mean of the variable* **`eruptions`** *when* **`waiting`** *is more than 1 hour is 4.138587 minutes. Thus, when the waiting time is more than 1 hour, you will get a higher average 'eruption' time comapred to when the waiting time is less than or equal to an hour.*

Question 6: (2 pts)

Create a scatterplot using the `plot()` function to explore how the waiting time might affect the eruption duration. Make sure to label axes (`xlab=`, `ylab=`) and give a title to the graph (`main=`). Briefly describe what you notice in this graph.

```
# Created scatter plot of eruption and waiting times
plot(faithful,
     main = 'Relationship of Eruption and Waiting Times',
     xlab = 'Eruption Duration (Mins)',
     ylab = 'Waiting Time (Mins)')
```



Answer: Based on the scatterplot above, we can see that there is a positive relationship between the waiting time between eruptions and the eruption duration.

Question 7: (2 pts)

How does the scatterplot that you made in Question 6 compare to the one you selected in the pre-quiz? Does it look similar or different? If the plot looks different from what you selected in the pre-quiz, how would you explain the difference?

Answer: The scatterplot that I made in Question 6 is similar to the one I selected in the pre-quiz as I hypothesized that there was a positive relationship between both variables.