R Lab Assignment: Field Trip to Yellowstone Park

Natalie Lee

Fall 2023

1 Introduction

Yellowstone National Park has a popular tourist attraction called Old Faithful. Old Faithful is a hot water geyser that has regular, spectacular eruptions that can last minutes.

The goal of this analysis is to provide valuable insight into the behavior of the geyser that serves to help people who are visiting the attraction get the most out of their experience (i.e. spend less time looking at a non-erupting hole in the ground).

To analyze the patterns of Old Faithful's eruptions we will be using the `faithful.csv` data set. In the data set, 272 cases of eruptions are examined. There are two observation variables. The first one, called eruptions, is the duration in minutes of the geyser eruption. The eruption duration in this data set range from 1.6 to 5.1 minutes! The second variable, called waiting, is the length of the subsequent waiting period in minutes until the next eruption. These waiting periods range from 43 to 96 minutes.

In the following sections, various statistical analysis and visualization methods will be utilized to explore the waiting periods in this data set.

2 Frequency Distribution of Eruption Waiting Periods

Frequency distribution is used to see the number of a data value in a given interval. In the case of waiting time, the frequency distribution can be used to count how many eruptions in the data set occurred at after specific time interval.

In the table below, the first column indicates the time interval in minutes, and the second column indicates the frequency of data points that occurred in that interval. For example, at the 40 to 50 minute time interval, 21 of the subsequent eruptions began.

```
##          waiting.freq
## [40,50)            21
## [50,60)            56
## [60,70)            26
## [70,80)            77
## [80,90)            80
## [90,100)           12
```

3 Duration Sub-Interval with the Most Eruptions

The table in the previous section can provide information about which waiting time interval most of the eruptions occur after. The figure of "[80,90)" below indicates that most of the eruptions occur between 80 to 90 minutes after the previous eruption.

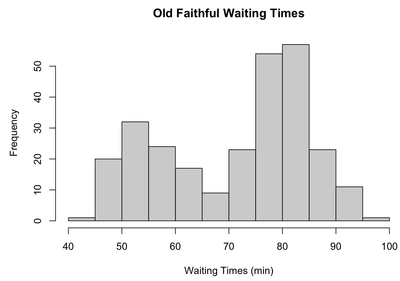
Knowing this, a tourist can expect that most likely an eruption will take place about an hour and twenty minutes to an hour and a half after the previous eruption. That being said, if one is just arriving after a eruption, they should feel free to take a detour to the gift shop.

```
## [1] "[80,90)"
```

4 Histogram of Eruption Waiting Period

To get a more visual look at the frequency distribution of the waiting periods, a histogram is helpful. The figure below shows on the x-axis the waiting time interval and the y-axis indicate the frequency of data points in that time interval.

In this way of looking at the data, the peak with the highest frequency is above the range we indicated previous at 80 to 90 minutes. Additionally, a second sub-interval of time in which eruptions occur the most is around 50-60 minutes after an eruption becomes clear.



5 Relative Frequency Distribution of Eruption Waiting Periods

While knowing the count of occurrences of eruptions at specific waiting periods is helpful, putting these values in a proportions can provide a figure that is more easily framed in the context of the entire data set. To calculate these proportions the frequency at a specific time interval was divided by the total count of eruptions.

In the table below, the first column indicates the time interval in minutes, and the second column indicates the proportion of data points that occurred in that interval. For example, at the 40 to 50 minute time interval, 8% of the subsequent eruptions began.

```
##          waiting.relfreq
## [40,50)             0.08
## [50,60)             0.21
## [60,70)             0.10
## [70,80)             0.28
## [80,90)             0.29
## [90,100)            0.04
```

6 Cumulative Frequency Distribution of Eruption Waiting Periods

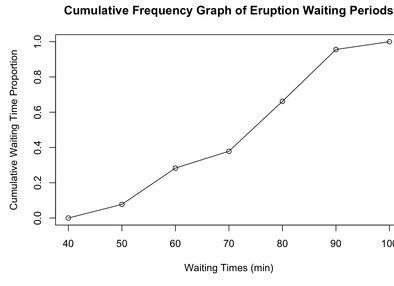
Cumulative frequencies show how the geyser's activity accumulates across the waiting time periods. Using this metric the number of data points that lie above and below a range of waiting times can be examined.

In the table below, the first column indicates the time interval in minutes, and the second column indicates the number of eruptions that have occurred up to that point. For example, at the 50 to 60 minute time interval, 77 out of the 272 eruptions in the data set have happened.

```
##          waiting.cumfreq
## [40,50)               21
## [50,60)               77
## [60,70)              103
## [70,80)              180
## [80,90)              260
## [90,100)             272
```

7 Cumulative Frequency Graph of Eruption Waiting Periods

The graph below provides a visual of the table in the previous section. This graph showcases well how most of the eruptions in the data set accumulated around the waiting times from 70 to 90 minutes.



8 Stem-and-Leaf Plot of Eruption Waiting Periods

In the plot below, each digit on the right hand side represents the last digit of a data point for waiting times. The digits on the left hand side are the stems of plot and represent all but the last digit of a data point.

A stem-and-leaf plot can provide a visual of data density. If there are many leaves for a single stem, it indicates a high frequency of eruptions occurred after that waiting period. The figure below shows a high density of waiting times with the '8' digit stem or the 80-89 minute waiting period range.

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   4 | 3
##   4 | 555666677778889999
##   5 | 000011111222223333333444444444
##   5 | 555555666677788889999999
##   6 | 00000022223334444
##   6 | 555667899
##   7 | 000111112333333444444
##   7 | 555555556666666677777777777888888888888889999999999
##   8 | 000000011111111111222222222233333333333344444444444
##   8 | 555555666666777888888999
##   9 | 00000012334
##   9 | 6
```

9 Scatter Plot of Eruption Durations and Waiting Intervals

Waiting periods are impacted by the duration of eruption. The scatter plot below examines the relationship between the duration of the geyser's eruption on the x-axis and the subsequent waiting time on the y-axis.

There is a positive linear relationship between the two variables, meaning that when the duration of a eruption is longer the subsequent wanting period is also longer.

Knowing this, tourist who have arrived after an eruptions with a longer duration can expect to wait longer than is they have arrived after a shorter eruption.