# BMI 500 Week 3 Lab- Geyser Behavior and Better Data Management

Frank Chien

September 8, 2021

## 1 Methods

Data on the eruption behavior of Old Faithful Geyser in Yellowstone national park was obtained through professor Reyna's lab website. Data included interval to between eruptions given in minutes as well as duration of eruption, also expressed in minutes. The data was imported as a csv file into R. The first 30 lines imported were skipped as they were comments about the data, rather than the data itself. Within the resulting table, there exists two rows consisting only of the with the character string "NaN", which typically stands for "Not a number" in MatLab. These rows were removed from the table.

Subsequent data cleaning and organizing took an omission-centered approach on ambiguous data. This method has the advantage of not presuming intentions of researchers who entered the data. Two examples of ambiguity include data entered as "7L" or "5.O33" where the capital letter "Oh" was entered instead of the numeric zero. While it's reasonable to presume 71 and 5.033 were the intended inputs, this involves conjecture on the part of the analyst and may be a source of error.

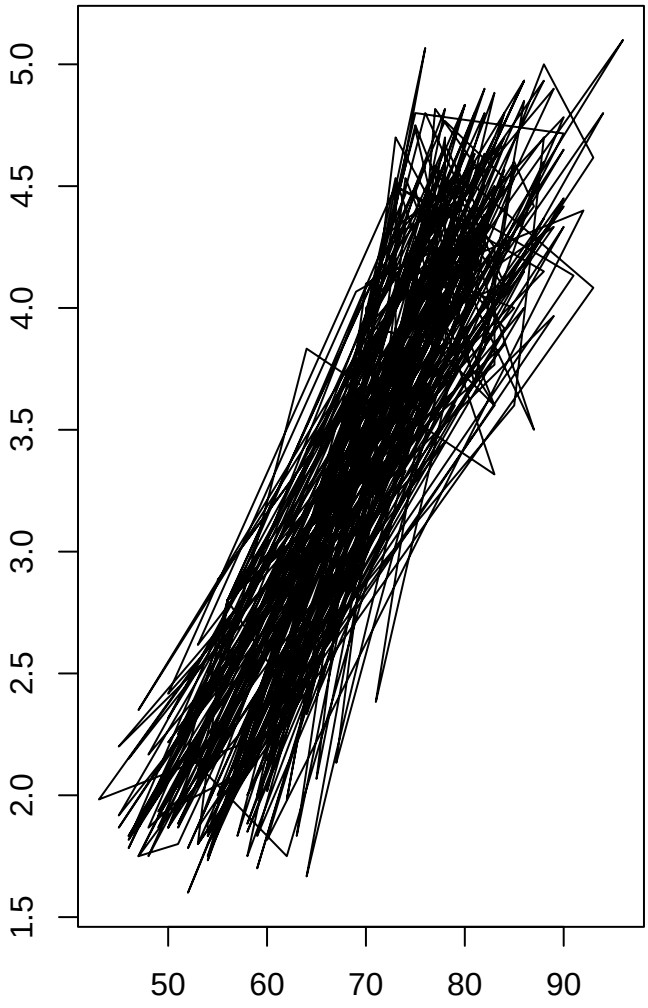
Negative numbers were removed from the data set, as these are nonsensical given the premise. That is to say, both variables measured a time duration, which cannot be negative. If either the eruption duration or interval between eruptions was entered as a negative number, the entire row was removed from the data set. Two such negative numbers were detected, and thus their respective rows were omitted. Finally, extremely outlying data that are greater than three standard deviations away from the mean were removed. Three standard deviations was chosen because under a normal distribution, 99% of the data are still included within three standard deviations of the mean. Similar to negative numbers, if either eruption duration or interval between eruptions fell outside the three standard deviations, the entire row was omitted. Two rows were removed under this exclusion criteria. One illustrative example is when two thousand eight hundred minutes was entered for eruption duration. This was likely a result of a mistakenly omitted decimal point.

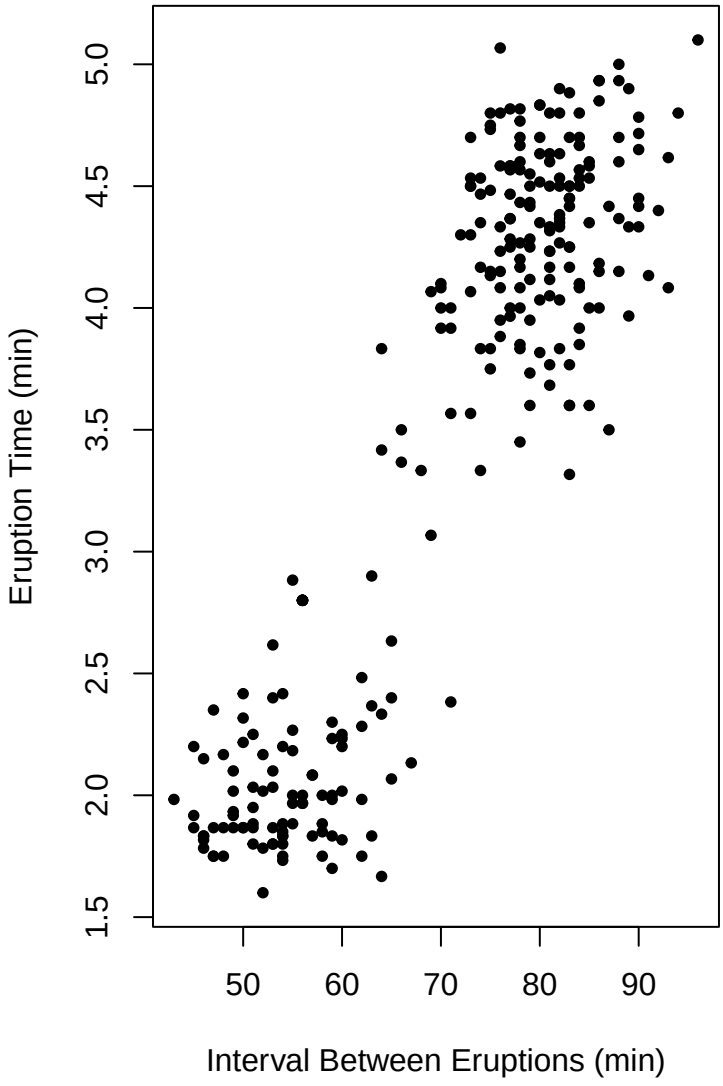
## 2 Results

The first figure below is an example of bad plot. This graphic is confusing because it plots a scatter plot with lines connecting the data points invoking a time-series relationship that does not exist in the data. Additional, the excess of lines are busy and confusing. Neither axis are labelled, and the title asserts a relationship between intervals time and eruption duration which may or may not actually exist. The second graphic is an improvement, as data points are represented by dots and not connected with lines. The axis this time are labelled with informative names and units, and there is no assertion of a relationship that is insufficiently supported with data. The absence of lines reveals two separate clusters of data points, which were not readily apparent in the first graphic.

Whereas the first graphic may suggest a straightforward positive correlation between the two data series, the second graphic reveals two populations of data points. This suggests that both the duration between eruptions and eruptions times follow a bi-modal distribution. Eruption times are likely to cluster around 2 minutes or 4.5 minutes, and likewise, intervals between eruptions clusters around 50 minutes or 80 minutes. Geologically, this may be explained by there being two different kinds of eruptions. Alternatively, there may be two research teams obtaining the data with systematic differences in obtaining data. Of note, the national park services website on Yellowstone National Park supports there being two modes of eruptions that old faithful exhibits, a long eruption with a long interval, and a short one with a short interval.

https://www.nps.gov/yell/learn/nature/hydrothermal-features.htmonthisPage-4


Wait Times Effect Eruption Duration
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
50
60
70
80
90


Eruption Time (min)
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
50
60
70
80
90
Interval Between Eruptions (min)