401 HW#3

Sylvia (Shuyuan) Ding Shuyuand

9/17/2019

Problem 5

Data

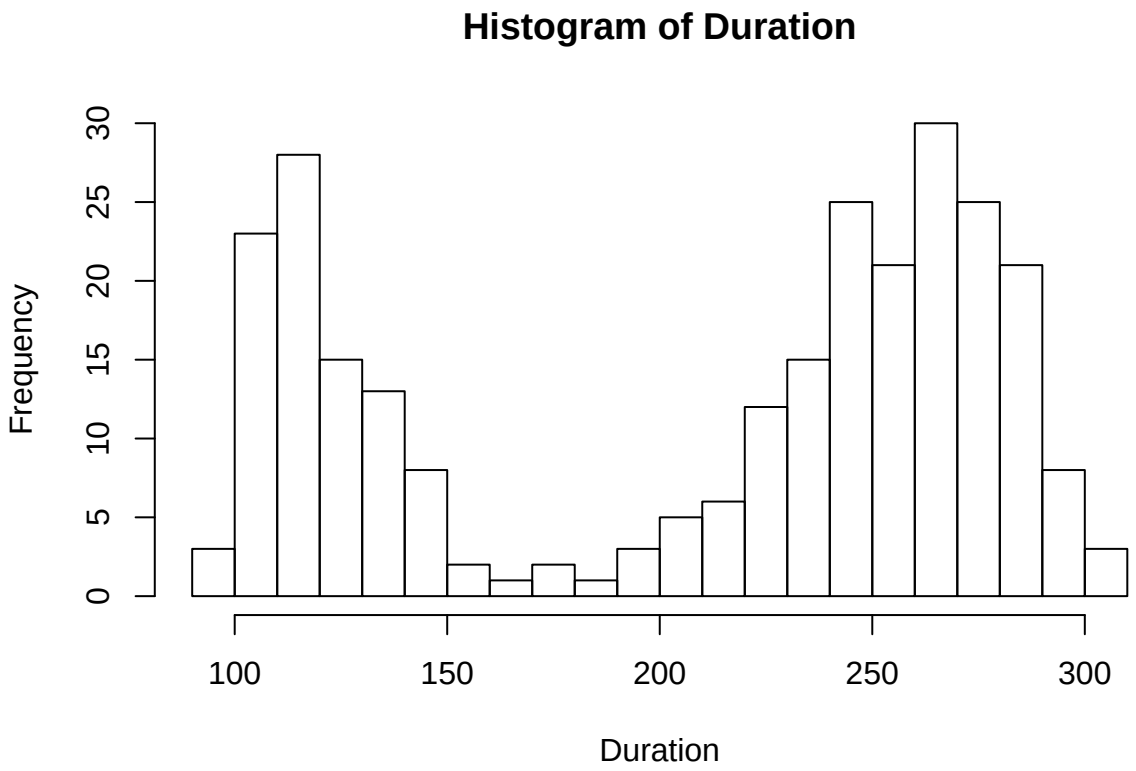
```
setwd("/Users/silviading/Desktop/CMU\U0001f33c/Fall 2019/36401/HW/")
oldfaith = read.table("oldfaith.txt",header = TRUE)
```

(a) It seems like Duration is the explanatory variable and Interval is the response variable. Both variables are quantitative.

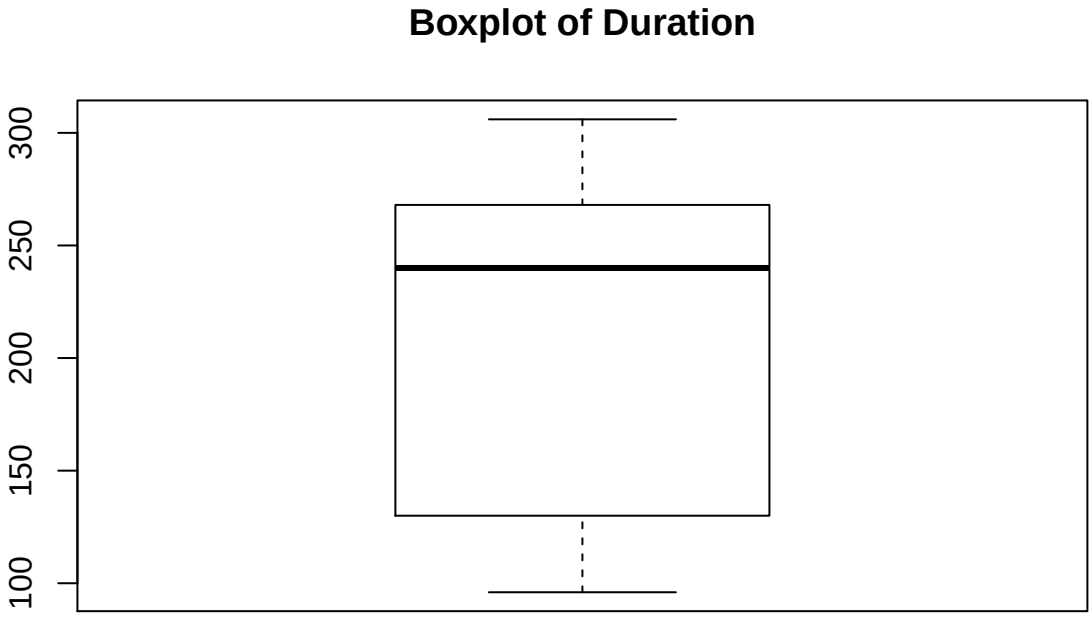
Plots

(b)

```
hist(oldfaith$Duration, xlab = "Duration", main = "Histogram of Duration", breaks = 20)
```

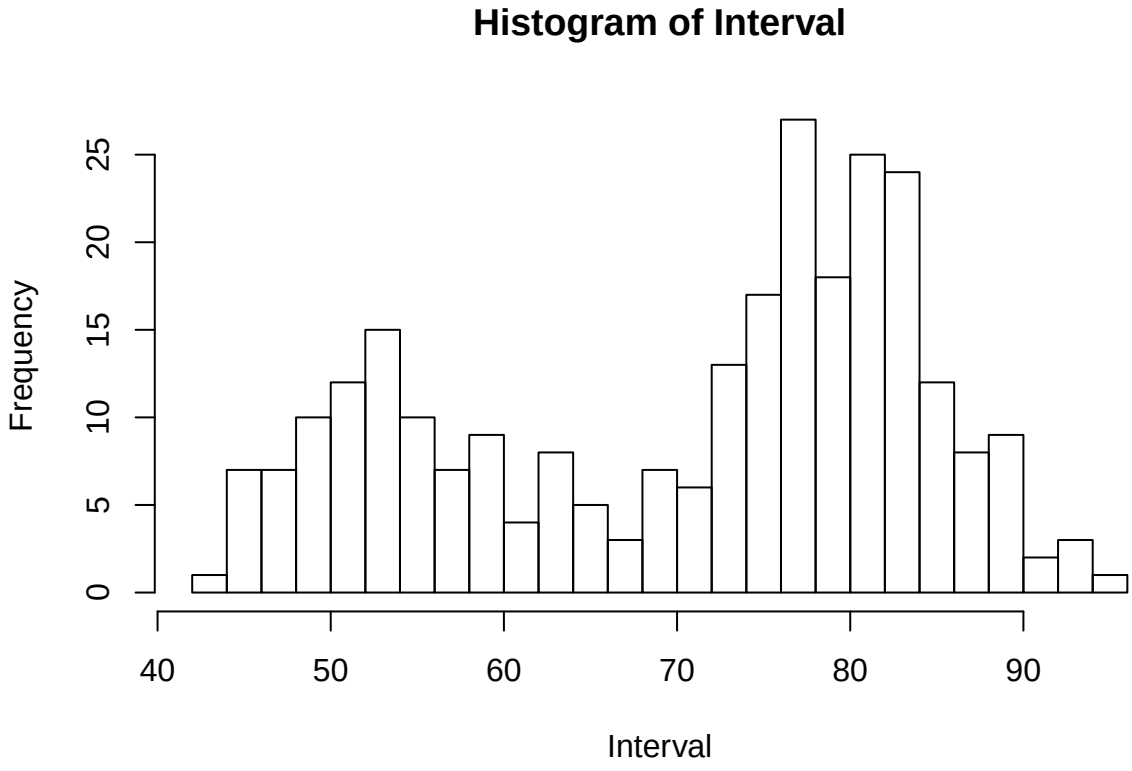



```
boxplot(oldfaith$Duration, main = "Boxplot of Duration")
```

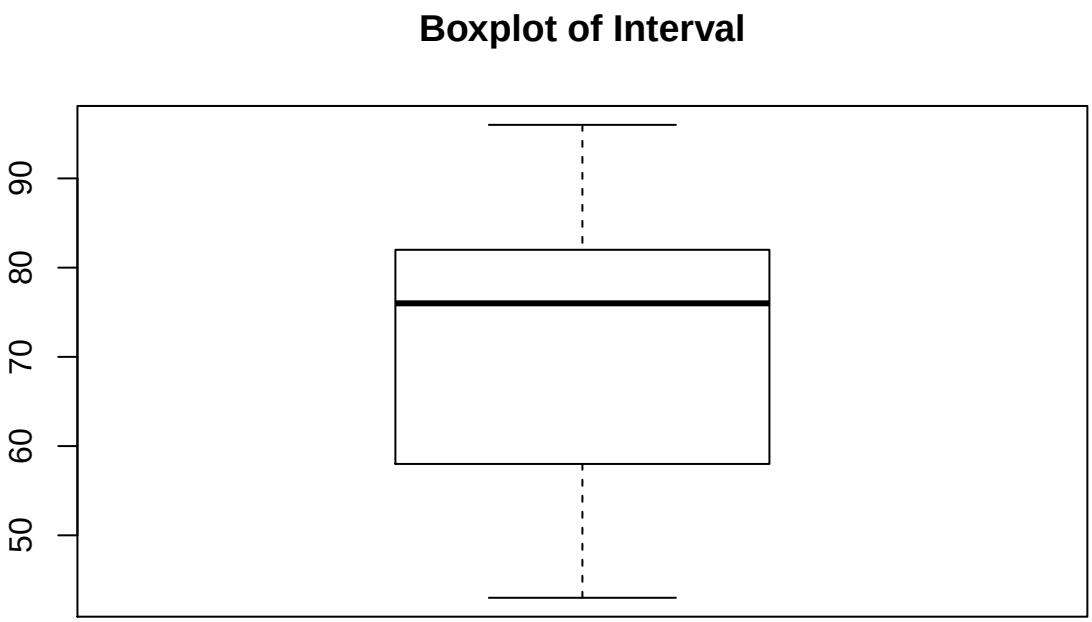


From the histogram, we can see that the distribution of Duration is bimodal, with two peaks at around 120 and 270, which means it happens most frequent that the duration in seconds of the current eruption is at 120 seconds and 270 seconds. Also it means we should probably look into the data as two separate groups, one with highest frequency of duration at 120 seconds, and the other with 270 seconds. Moreover, from the boxplot, we see that both range and IQR are really big. Range is about 200 (from 100 to 300), and IQR is about 250 as well. Thus, the median would not be really informative. And there does not seem to be any outlier.

```
hist(oldfaith$Interval, xlab = "Interval", main = "Histogram of Interval", breaks = 20)
```



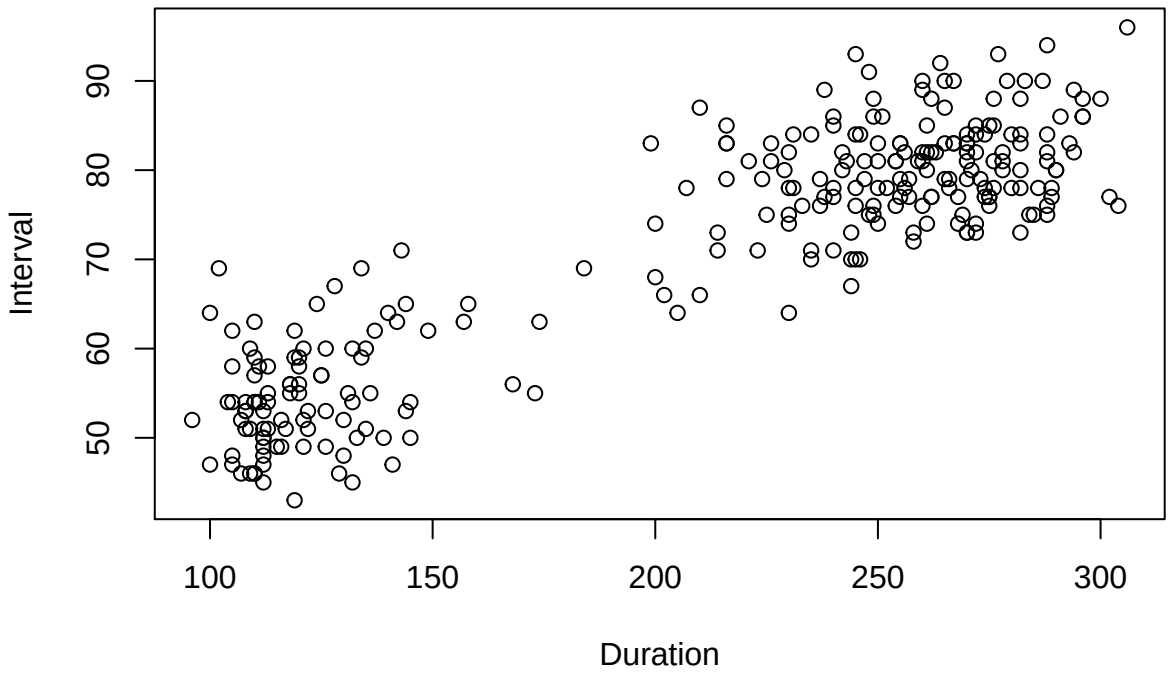
```r
boxplot(oldfaith$Interval, main = "Boxplot of Interval")
```



For interval, the histogram is also bimodal with peaks at around 52 and 76. Thus we know it is most frequent for one group of the next erruption to happen in 76 minutes, and another group for the next eruption to happen in 68 minutes. Then from the boxplot, we can see that the range of interval is about 50, and IQR is about 30. Also just like Duration, IQR and range are big, so median is not that informative. Also there does not seem to be any outlier.

(c)

```r
plot(Interval~Duration, data = oldfaith)
```



From the graph, we can see that there is positive correlation between the variable Duration and Interval. Interval increases as duration increases. Also there seems to be two subgroups of data, one on the left bottom

corner, one at top right corner, and there is a gap in the middle. It shows the information that we do not have information of the eruption with duration of around 180 seconds and interval of around 70 minutes. Also it seems like the correlation is pretty strong, but also there is some variability as well, because it seems like if we form a linear regression, the points are not all really close to the line.

(d) Abstract

Old Faithful is a famous geyser located in Yellowstone's Upper Geyser Basin. Discovered in 1870, Old Faithful geyser was named for its frequent and somewhat predictable eruptions. We are interested in discovering the relationship between Old Faithful's duration and interval time period. Duration is measured in seconds of the current eruption, and interval is measured in minutes of the time to the next eruption. We will be conducting histogram and boxplots to firstly do univariate data analysis to see the distribution of each, and then later do a scatterplot for bivariate EDA to test if there is any correlation between two variables.