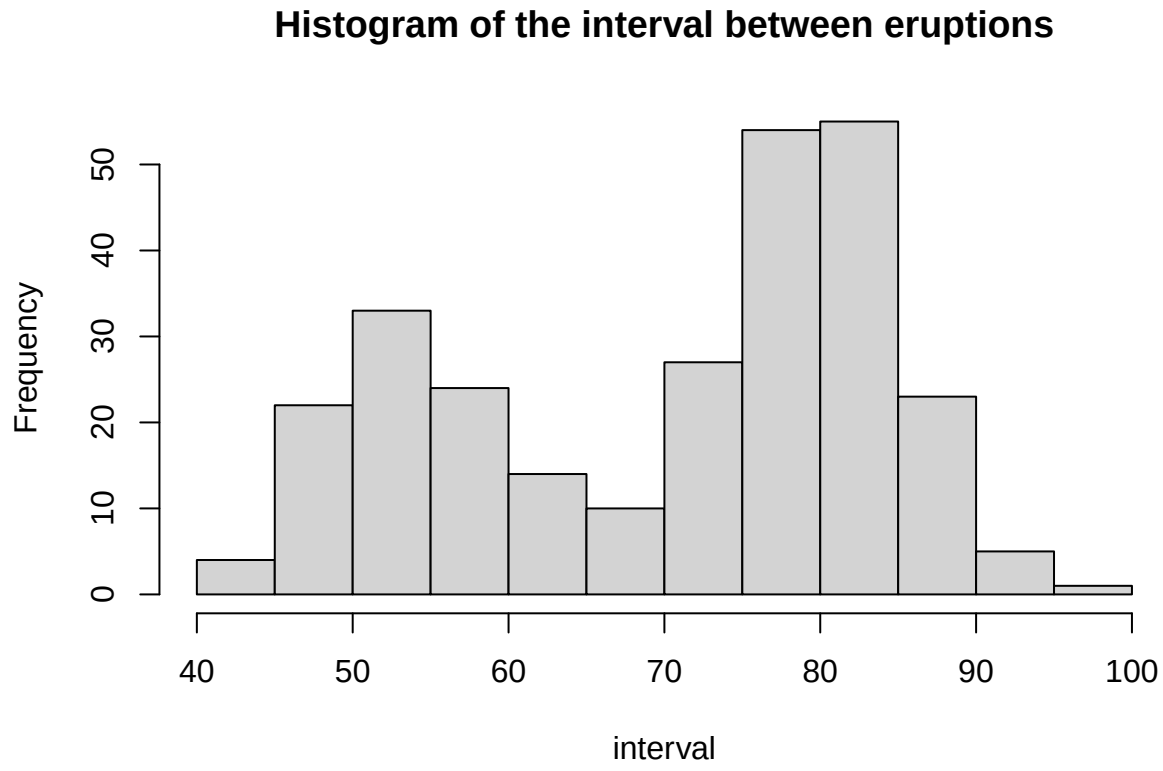
Old Faithful Geyser

```r
# import geyser_data
geyser_data <- read.csv("geyser.csv",header=TRUE)
```

a).

```r
# Construct a histogram of the interval between eruptions
hist(geyser_data $ waiting,
     main = "Histogram of the interval between eruptions",
     xlab = "interval")
```



It is a bimodal histogram, centered at around 55 minutes and 80 minutes.

I would say it is faithful, since it will erupt between 40 minutes and 100 minutes; moreover, it will erupt around 55 minutes and 80 minutes most likely which means visitors are most likely wait around 55 minutes or 80 minutes if they just missed the last eruption.

b).

```
mean(geyser_data $ waiting)
```

```
## [1] 70.89706
```

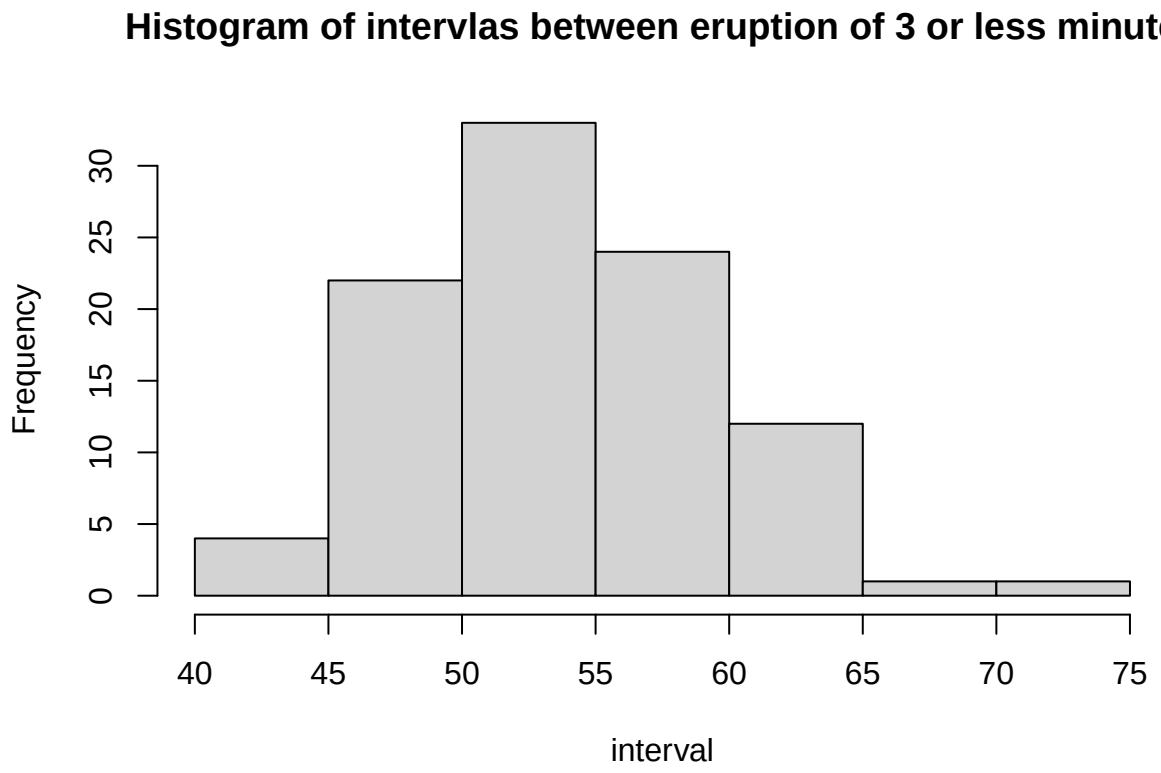
```
sd(geyser_data $ waiting)
```

```
## [1] 13.59497
```

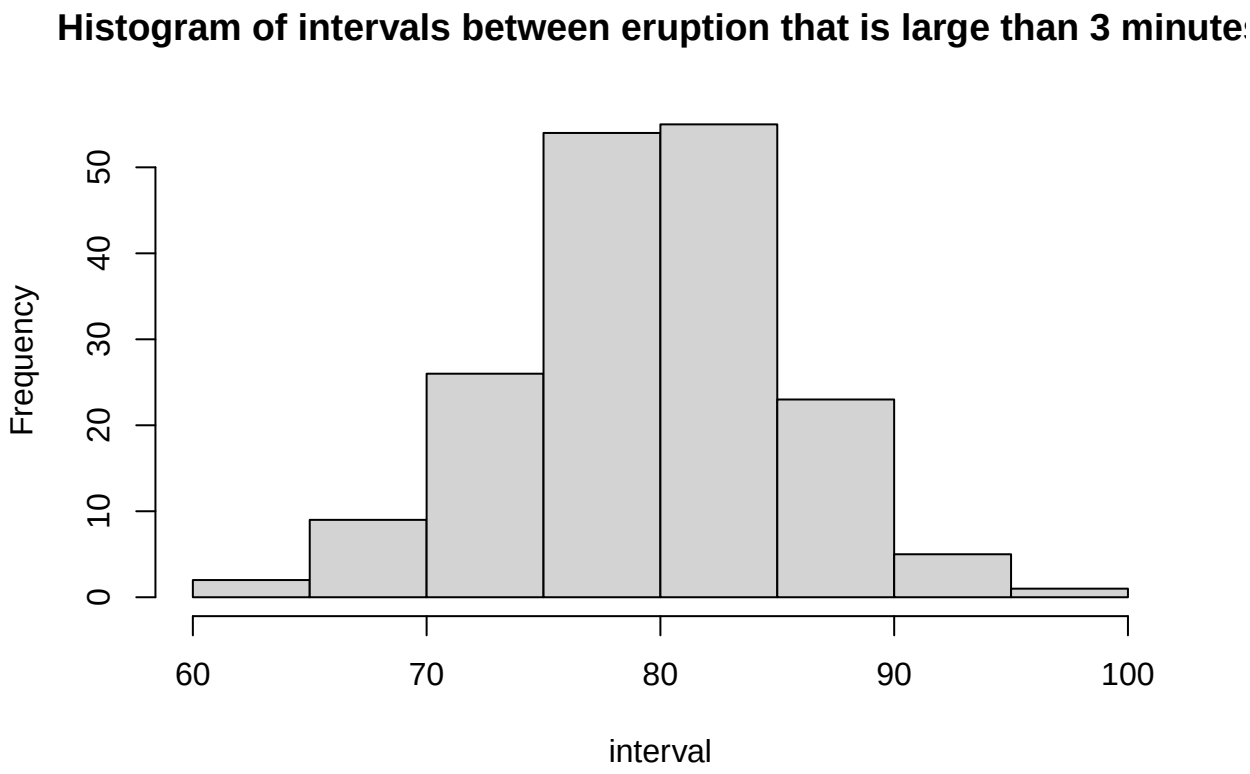
I would not use the standard descriptive statistics for the interval data(mean,standard deviation). The histogram is not bell-shaped, so the center and spread are not a good summary of the data.

c).

```
# Divide the data into two parts
geyser_data_3_or_less <- subset(geyser_data, eruptions <= 3)
geyser_data_larger_than_3 <- subset(geyser_data, eruptions >3)
# Construct histograms of intervals between eruptions for both sets of data
hist(geyser_data_3_or_less $ waiting,
     main = "Histogram of intervlas between eruption of 3 or less minutes",
     xlab = "interval")
```




```
hist(geyser_data_larger_than_3 $ waiting,
     main = "Histogram of intervals between eruption that is large than 3 minutes",
     xlab = "interval")
```




```
# Find the mean and standard deviation in two groups.

# eruption 3 minutes or less
mean1=mean(geyser_data_3_or_less $ waiting)
sd1=sd(geyser_data_3_or_less $ waiting)

# eruption greater than 3 minutes
mean2=mean(geyser_data_larger_than_3 $ waiting)
sd2=sd(geyser_data_larger_than_3 $ waiting)
```

Use the 68-95-99.7 empirical rule to construct a rule based on the length of the previous eruption (3 minutes or less or greater than 3 minutes) to estimate the interval between eruptions.

```
mean1-sd1
```

```
## [1] 48.65475
```

```
mean1+sd1
```

```
## [1] 60.33494
```

If the length of the previous eruption is 3 minutes or less, then 68% of visitors will wait around 49 minutes to 60 minutes.

```
mean1-2*sd1
```

```
## [1] 42.81465
```

```
mean1+2*sd1
```

```
## [1] 66.17504
```

If the length of the previous eruption is 3 minutes or less, then 95% of visitors will wait around 43 minutes to 66 minutes.

```
mean1-3*sd1
```

```
## [1] 36.97455
```

```
mean1+3*sd1
```

```
## [1] 72.01514
```

If the length of the previous eruption is 3 minutes or less, then 99.7% of visitors will wait around 37 minutes to 72 minutes.

```
mean2-sd2
```

```
## [1] 73.99433
```

```
mean2+sd2
```

```
## [1] 85.98281
```

If the length of the previous eruption is greater than 3 minutes, then 68% of visitors will wait around 74 minutes to 86 minutes.

```
mean2-2*sd2
```

```
## [1] 68.00009
```

```
mean2+2*sd2
```

```
## [1] 91.97705
```

If the length of the previous eruption is greater than 3 minutes, then 95% of visitors will wait around 68 minutes to 92 minutes.

```
mean2-3*sd2
```

```
## [1] 62.00585
```

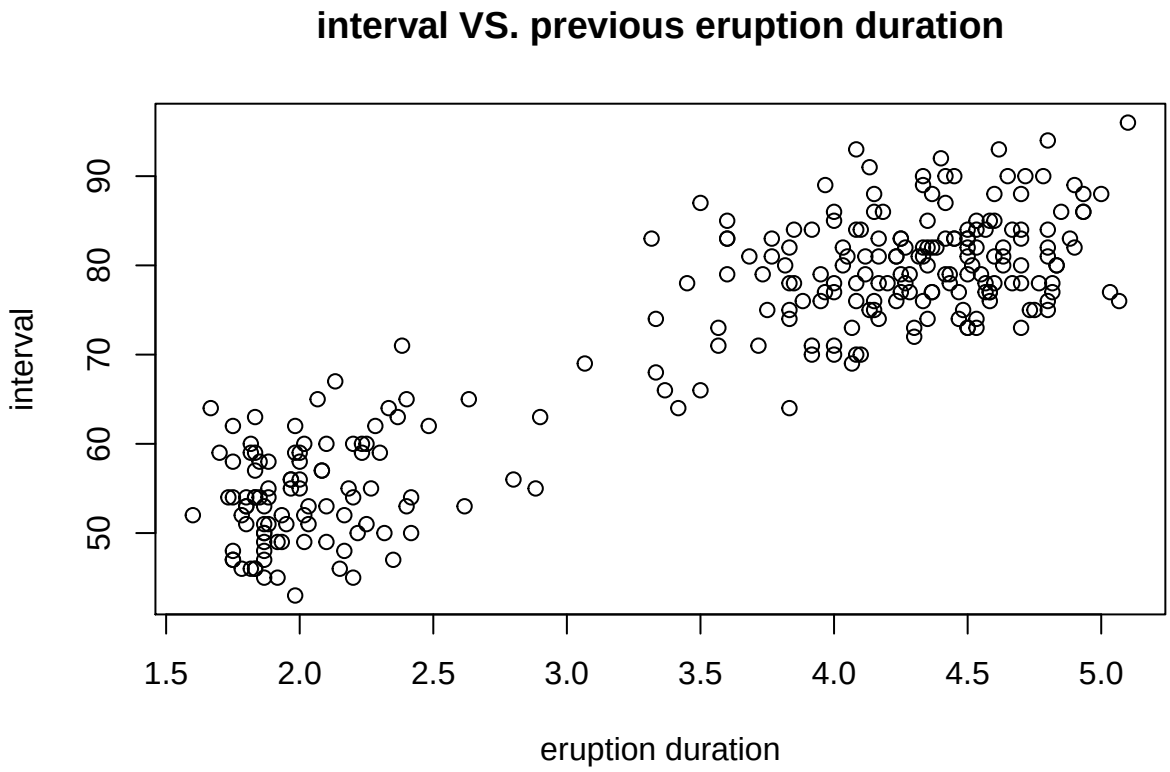
```
mean2+3*sd2
```

```
## [1] 97.97129
```

If the length of the previous eruption is greater than 3 minutes, then 99.7% of visitors will wait around 62 minutes to 98 minutes.

d).

```
# Construct a scatter-plot
plot(formula = waiting ~ eruptions,
     data = geyser_data,
     main = "interval VS. previous eruption duration",
     xlab = "eruption duration",
     ylab = "interval")
```

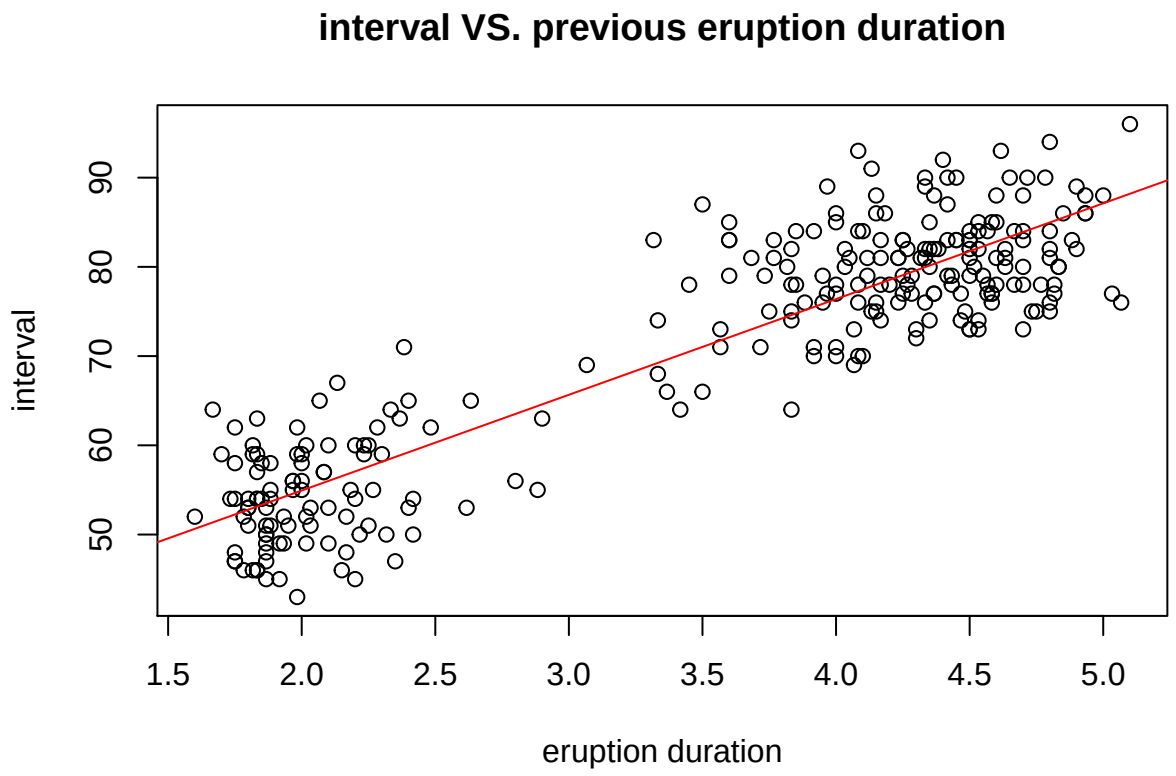


According to the scatter plot, we can see there are two subgroups just like we discussed earlier(eruption duration 3 minutes or less and greater than 3 minutes). we can also conclude that there is a positive relationship between eruption duration and interval. As the eruption duration increased, interval will increase as well.

A straight line appear to provide a reasonable approximation of the relationship between interval and eruption duration, so linear regression could be used here.

e).

```
reg <- lm(formula = waiting ~ eruptions,
          data = geyser_data)
plot(formula = waiting ~ eruptions,
     data = geyser_data,
     main = "interval VS. previous eruption duration",
     xlab = "eruption duration",
     ylab = "interval")
abline(reg,
       col = "red")
```




```
reg $ coefficients
```

```
## (Intercept)   eruptions
##    33.47440    10.72964
```

We can conclude that for each additional minutes of eruption duration, interval increases around 11 minutes.

f).

```
range(geyser_data$eruptions)
```

```
## [1] 1.6 5.1
```

```
predict(reg, data.frame(eruptions = 2))
```

```
##        1
## 54.93368
```

If the length of the previous eruption was 2 minutes, I would expect to wait around 55 minutes until the next eruption.