Lesson 3 Companion

There is a lot of example code and are important points made in the *Tidyverse Tutorial* so please review the **ggplot2** section again if you have any questions. For this example code I'm going to use the geyser data referenced in the P.2 section of your book. This data is available from the book website at the link included in the code.

```
library(tidyverse)

geyser = read_table2("http://www.isi-stats.com/isi/data/prelim/OldFaithful2.txt")

head(geyser)
```

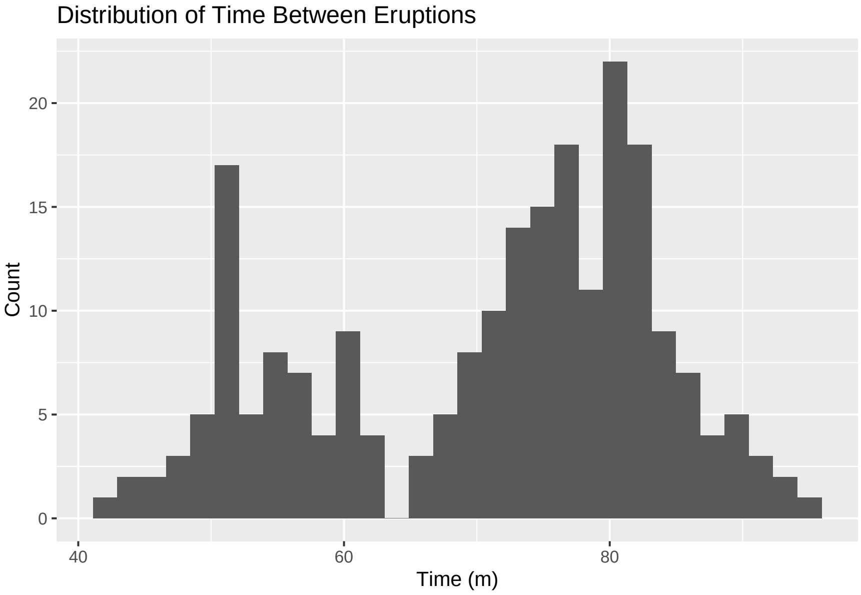
```
## # A tibble: 6 x 2
##   EruptionType TimeBetween
##   <chr>              <dbl>
## 1 short                 55
## 2 short                 58
## 3 short                 56
## 4 short                 50
## 5 short                 51
## 6 short                 60
```

The code above turns on the awesome (loads the tidyverse library) and then reads in our data from the book's website. Notice we use the **read_table2** command here (as opposed to the **read_csv**) covered in the *Tidyverse Tutorial*. I'm doing this because we aren't dealing with a comma-seperated values file but rather a file in which the values for each variable are seperated by the tab character. I find that **read_table2** is a good way to deal with these sort of files.

The last bit of code offers a quick glimpse into the first six rows of the dataframe. It also allows us to ensure that the variables have been "read in" correctly. The *EruptionType* variable was read in as a character (chr) and the *TimeBetween* variable was read in as a double (dbl) so we are good. Occasionally your categorical variables will be read in as some type of number (integer, double, float, long, etc.) which can cause problems later on.

As part of the data exploration process for this dataset, I want to create a visualization that will allow me to get an idea of the distribution of times between eruptions. There are two way to consider this: with regard to the *EruptionType* and without. Let's start with without considering it.

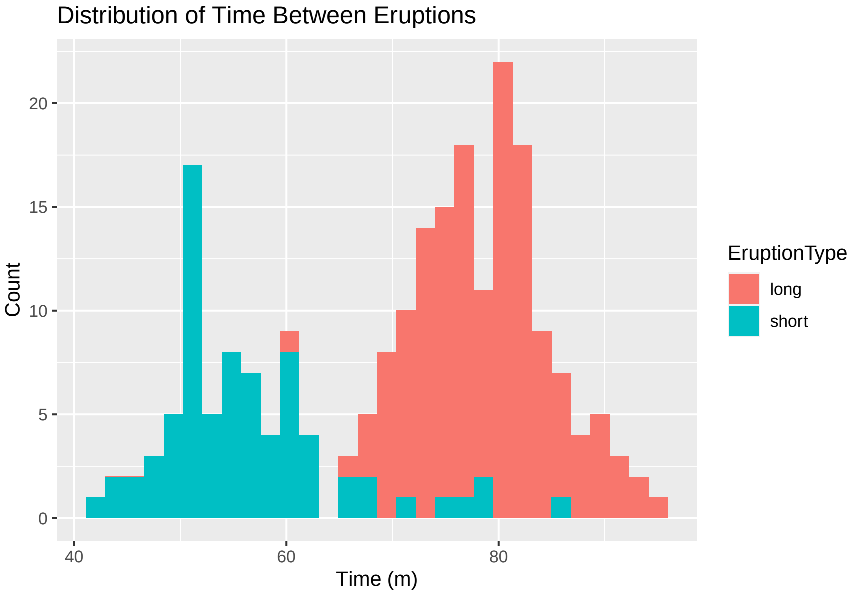
```
geyser %>%
  ggplot(aes(x = TimeBetween)) +
  geom_histogram() +
  labs(x = "Time (m)", y = "Count", title = "Distribution of Time Between Eruptions")
```



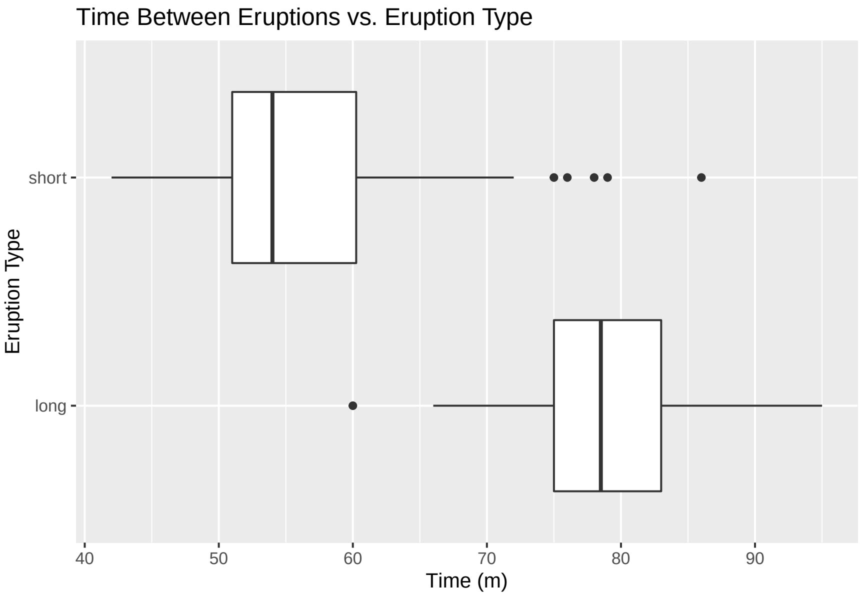
We can get an idea of the shape, center, variability, and unusual observations from this histogram. We can see that we are dealing with a *bi-modal* dataset, a conclusion that is reinforced by the obvious split in the data. We know that we have this other variable *EruptionType* but imagine that we didn't know about that: seeing this split should lead you to question whether there is some additional data you should be collected to try to explore why there is a bifurcation of our data. Furthermore, remember that our second variable is *EruptionType* and this is a split in the times **between** eruptions so, even though we have this second variable to explore, there may still be more information our there we would like to gather.

Now let's look at a couple visualizations that take into account our second variable of *EruptionType*.

```
geyser %>%
  ggplot(aes(x = TimeBetween, fill = EruptionType)) +
  geom_histogram() +
  labs(x = "Time (m)", y = "Count", title = "Distribution of Time Between Eruptions")
```



```r
geyser %>%
  ggplot(aes(x = TimeBetween, y = EruptionType)) +
  geom_boxplot() +
  labs(x = "Time (m)", y = "Eruption Type", title = "Time Between Eruptions vs. Eruption Type")
```



These two visualizations allow us to gain some information about the relationship between this two variables in our sample. We can now describe the shape, center, variability, and unusual observations for each group in the histogram. The side-by-side boxplot can provide very similar information but in a different, and perhaps easier to understand form. Please reference Section 6.1 for more information about the specific parts of the boxplot.