# MA256 Course Guide

## Task

Develop critical statistical thinkers who can generate precise research questions; identify, collect, and analyze relevant data; and translate this analysis to a complete and correct response that answers the original question.

## Purpose

Our daily lives are full of question whose answers should be informed by the growing amount of data surrounding us. Politicians, scientists, sports teams, and military leaders are all looking toward quantitative analysis to provide them with unbiased information with which to make the correct decision. In this course you will gain valuable experience using data in an investigation process to collect, analyze, and report your results to help answer relevant research questions.

You will gain experience in not only completing statistical analysis but also asking the important follow-up questions about why you are getting these results and how you can expect them to change if the experiment was repeated. You will also gain experience in reading, comprehending, and reproducing the results of published scientific papers. All of these tools will help to equip you to be successful in the course project where you are provided the opportunity to do research on a question that interests you. Perhaps you will discover, like some of your fellow cadets that, indeed, more cadets do report to sick call on Mondays.

## Grading

Below is the point break-down for this course:

| Category | Points |
|---|---|
| **Homework** | **350** |
| Block 1 | 50 |
| Block 2 | 125 |
| Block 3 | 125 |
| Block 4 | 50 |
| **Midterm** | **150** |
| **Course Project** | **250** |
| Proposal | 25 |
| IPR 1 | 25 |
| IPR 2 | 25 |
| Presentation | 75 |
| Report | 100 |
| **TEE** | **250** |
| **Total** | **1000** |

# Block 1: Question Formulation and Data Exploration

## Lesson 1: Course Introduction and RStudio Familiarization

**Objectives:**

- Introduction to MA256 and course expectations

- Installation and familiarization with RStudio and R
- Tidyverse Tutorial

**Before Class Activities:**

- Read: *Introduction to Statistical Investigations* - Preliminary 1 (P.1)
- Watch: WileyPlus - Videos P.1.1 - P.1.4 (Optional but recommended)
- Do: Consult *Introduction to RStudio and Tidyverse* (AKA *Tidyverse Tutorial*) and follow the "Download R and RStudio" instructions.
- Do: Take a look at Hadley Wickham's book *R for Data Science* which is available at the link below. This is a great resource for your coding questions as it was written by the guy who wrote a large part of the *Tidyverse*. https://r4ds.had.co.nz/

**After Class Activities:**

- Ensure that R/RStudio is installed and functional on your computer.
- Carefully go through the *Tidyverse Tutorial*. Attention now will save you a great deal of time later.
- Consider the dataset that we discussed in class today, what sort of research questions could you ask based on the data and variables present in this dataset. No need to answer the questions, but develop three good research questions before next class.

## Lesson 2: Research Questions

**Objectives:**

- Understanding types of and relationships between variables
- Formulation of specific and focused research questions
- Investigation of the "state of the art"
- Presentation of research in a literature review

**Before Class Activities:**

- Read: *Introduction to Statistical Investigations* - Chapter 4
- Watch: WileyPlus - Videos 4.1.1 - 4.1.5 and Videos 4.2.1 - 4.2.3
- Do: Exploration 4.2 - Problems 1 - 14
- Do: Prepare the three research questions referenced in **Lesson 1 - After Class Activities**.

**After Class Activities:**

- Pick one of your group's research questions and prepare a literature review for at least two scholarly articles. One submission per group. This will be due at the start of Lesson 4.

## Lesson 3: Data Exploration - 1

**Objectives:**

- Understand terminology used in describing distributions
- Understand the information conveyed in a boxplot and five-number summary
- Create appropriate data visualizations in *R*
- Understand reasons for and conclusions drawn from data exploration

**Before Class Activities:**

- Read: *Introduction to Statistical Investigations* - Preliminary 2 (P.2)
- Read: *Introduction to Statistical Investigations* - Section 6.1
- Watch: WileyPlus - Videos P.2.1 - P.2.5 (Optional but recommended)
- Do: Work on the literature review referenced in **Lesson 2 - After Class Activities**.

- Do: Review and replicate the code examples shown below.

### Lesson 3 - Example Code

There is a lot of example code and 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)
```

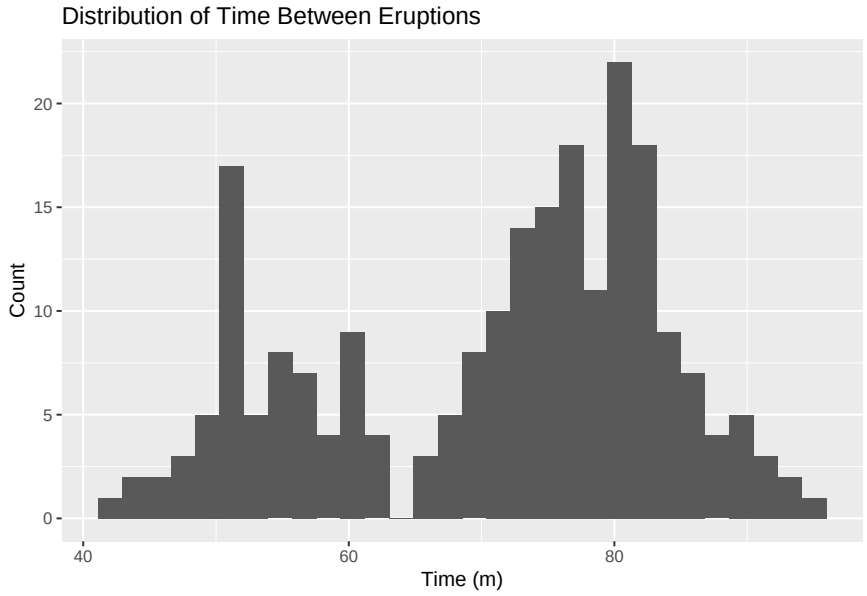
```
## # A tibble: 6 x 2
##   EruptionType TimeBetween
##   <chr>              <dbl>
## 1 short                 55
## 2 short                 58
## 3 short                 56
## 4 short                 50
## 5 short                 51
## 6 short                 60
```

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.

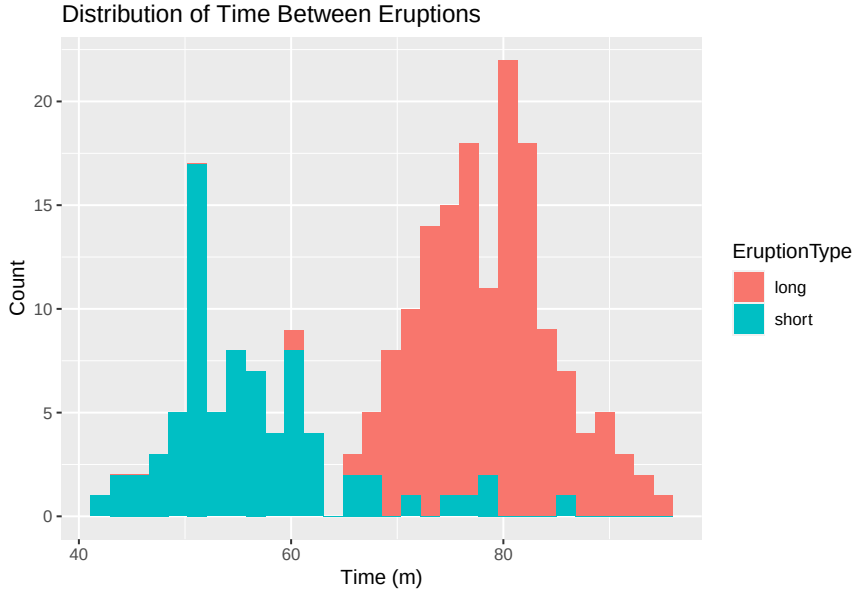
```
geyser %>%
  ggplot(aes(x = TimeBetween)) +
  geom_histogram() +
  labs(x = "Time (m)", y = "Count", title = "Distribution of Time Between Eruptions")
```



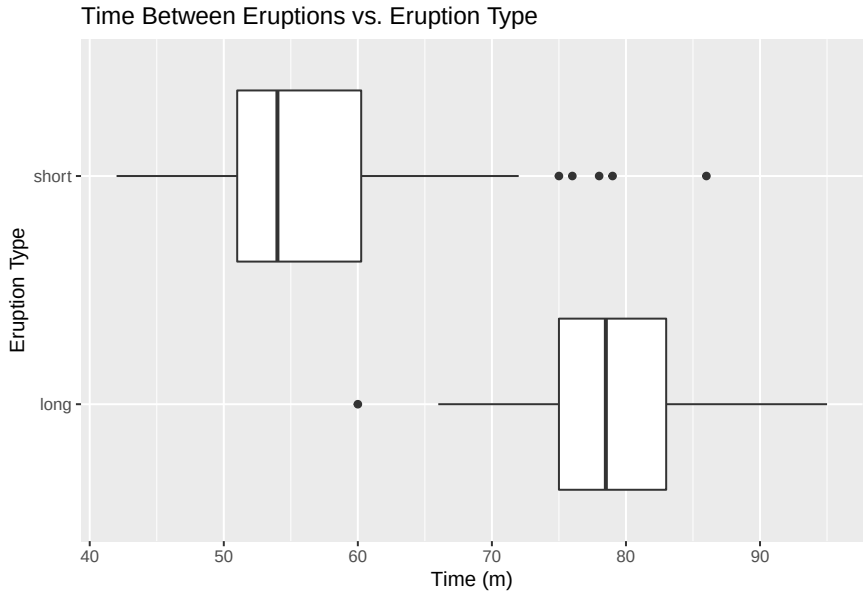

We can get an idea of the shape, center, variability, and unusual observations from this histogram. We can see that we are dealing with a *bi-modal* dataset, a conclusion that is reinforced by the obvious split in the data. We know that we have this other variable *EruptionType* but imagine that we didn't 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.

**After Class Activities:**

- Finalize and submit the literature review referenced in **Lesson 2 - After Class Activities**.

## Lesson 4: Data Exploration - 2

**Objectives:**

- Demonstate proficiency of data exploration skills

**Before Class Activities:**

- Read: *Introduction to Statistical Investigations* - Preliminary 3 (P.3)
- Watch: No videos for this lesson.
- Do: Ensure you have the *NYPD Arrest* data loaded for this lesson.
- Do: Literature from **Lesson 2** is due before the start of class.

**After Class Activities:**

- Finalize your work from class. Submission is due before the start of **Lesson 5**. One submission per group.