Introduction

For the final project, you will conduct your own exploratory data analysis and create an RMD file that explores the variables, structure, patterns, oddities, and underlying relationships of a data set of your choice.

The analysis should be almost like a stream-of-consciousness as you ask questions, create visualizations, and explore your data.

This project is open-ended in that we are not looking for one right answer. As John Tukey stated, "The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data." We want you to ask interesting questions about data and give you a chance to explore. We will provide some options of data sets to explore; however, you may choose to explore an entirely different data set. You should be aware that finding your own data set and cleaning that data set into a form that can be read into R can take considerable time and effort. This can add as much as a day, a week, or even months to your project so only adventure to find and clean a data set if you are truly prepared with programming and data wrangling skills.

Now, on to the details!

Step One - Choose your Data Set

First, you will choose a data set from the **Data Set Options** document. You should choose a data set based on your prior experiences in programming and working with data. The data set you choose will not increase or decrease your chances of passing the final project. In general, **tidy data sets** are easier to work with since each variable is a column and each row is an observation; there's no data cleaning or wrangling involved. We offer guidance below for choosing your data set. Time estimates include reading all of the project instructions and rubric, conducting the analysis, and submitting the final project.

Step Two - Get Organized

Eventually you'll want to submit your project (and share it with friends, family, and employers). Get organized before you begin. We recommend creating a single folder on your desktop that will eventually contain:

1. The **RMD file** that contains the analysis, final plots and summary, and reflection (in that order). A template can be found on the next page to help you put this together.
2. The **HTML file** that will be knitted from your RMD file
3. The **data set** you used (which you will only submit if you found your own data set)

Step Three - Explore your Data

This is the fun part. Start exploring your data! Keep track of your thoughts as you go (in an RMD file). Please refer to the **Example Project** that we have provided. Your report should look similar!

Step Four - Document your Analysis

You will want to document your exploration and analysis in an **RMD file** which you will submit. That file should be formatted in markdown and should contain (in order):

1. **A stream-of-consciousness analysis and exploration of the data.**
   a. Headings and text should organize your thoughts and reflect your analysis as you explored the data.
   b. Plots in this analysis do not need to be polished with labels, units, and titles; these plots are exploratory (quick and dirty). They should, however, be of the appropriate type and effectively convey the information you glean from them.
   c. You can iterate on a plot in the same R chunk, but you don't need to show every plot iteration in your analysis.

2. **A section at the end called "Final Plots and Summary"**
   You will select three plots from your analysis to polish and share in this section. The three plots should show different trends and should be polished with

appropriate labels, units, and titles (see the **Project Rubric** for more information).

3. **A final section called "Reflection"**
   This should contain a few sentences about your struggles, successes, and ideas for future exploration on the data set (see the **Project Rubric** for more information).

Step Five - Knit your RMD file

Your knitted RMD file should not be one long chunk of R code. It should contain text and plots interspersed throughout. The goal is to give the person reading the file insight into what you were thinking as you explored your data.

Step Six - Document your Data (if you chose your own data set)

The data set you submit (only if you chose your own) should include a text file, like those in the R documentation (e.g. ?diamonds) that describes the source of your data and an explanation of the variables in the data set (definition of any variables, units, levels of categorical variables, and the data generating process, such as how data was collected if possible).