MT 143, Final Project, Spring 2020

The purpose of this project is to select a real-world data set and perform some statistical explorations. You should select a real-world data set that has sufficient data to explore thoroughly. It should contain at least three variables (Subject ID is not a variable you can analyze, so three in addition to that). If you are having difficulty finding a real-world data set, please contact me. I can send you some suggested websites to look through. You may combine multiple datasets on similar subjects to reach the three-variable minimum, or to expand your analysis.

To perform the statistical explorations, you should utilize a technology of your choice. This can include software like Minitab or Excel, you can use various other online tools, or the R programming language. (If you use R, I have some tutorials for making graphs of numerical data on my website, but I suggest that this is for the ambitious with some prior programming experience.) You may use more than one tool. Minitab is good for some things that Excel can't do, but Excel can also do some things that Minitab can't, and so forth, so you are welcome to mix and match as you see fit.

After selecting your data set, your analysis will need to include the following information:

- Identify the variables you are going to perform your analysis on and explain why they are interesting to you.
- Summarize the analysis you did in the mini-project: include any graphs or summary statistics that will be useful for the rest of your analysis. If you have more than three variables of potential interest in your data set, you may use this section to reject variables that don't seem meaningful, or if you have too few variables, to justify bringing in more data. If you did your mini-project on a different data set, perform the same calculations and analysis under those instructions and summarize the results thoroughly.
- Determine whether it is beneficial to perform regression analysis on the data.
- The main purpose of this project is to perform a hypothesis on your data. To do this you can do any of the following:
  - You may research the variable and determine background assumptions/knowledge about the data to determine if your data agrees with this information or not. (One sample hypothesis test for means or proportions.) If you do this, you must, of course, cite your source.
  - You may split your data into groups based on a categorical variable and perform a two-sample test (of means or proportions) or an ANOVA test (three categories or more).
  - You may do a regression analysis with hypothesis testing (linear, nonlinear, or multiple). This should include residual analysis.
  - If you have multiple categorical variables, you may do a χ^2-test of independence.
  - Somewhat more challenging would be to do a goodness-of-fit test. Research may be able to tell you what kind of distribution to expect and you could test to see if your data matched (if so, cite your source).

Optionally: You may extend your analysis further by setting up a simulation based on your data or perform a bootstrap analysis. These are a little more advanced, but you are welcome to attempt these methods if you want. YouTube is a great place to find out more about how to set these up in your chosen technology.

Your analysis should be about 3-5 pages. Remember that the graphs and output of the analyses will take up quite a bit of space, so don't be too daunted by this. If you are tending toward the short end of this range, do not count a title page or reference pages. (Hint: make some more graphs!) You may also make a PowerPoint if you prefer but expect 12-20 slides (counting title slide and references).

The last thing you will need to produce is a short, roughly 5-minute video, presenting your results. Initially, we were going to do this in class, with a video as an option, but now it is a requirement since we don't have class anymore. You may use your PowerPoint to make a movie of your presentation with voiceover, you may use your phone to record yourself doing the presentation, or you can use a screencast software (Screencast-o-matic is free for videos shorter than 15 minutes and is linked in Blackboard), or any other software that you are familiar with. I strongly encourage you to keep the video as close to 5 minutes as possible. I am okay if you go a little over, by if you go over 10 minutes, I may have to take off points.

This project is worth **150 points**. Due date is 5/13/2020.