**Statistics 101 F2019 Final Project**

Due: December 18th at 11:59 pm via Sakai

Key requirements:

* 2000 words
* Paper should have 1) title page 2) main body and 3) works cited page. Graphs and tables should be in-line with the text.
* Font: 12 pt Times New Roman
* Spacing: double spaced
* Citations: Chicago
* A title page with your name, paper title, word count and date
* Subject headings for each major section of your paper
  + Suggested structure:
    - Introduction
    - Hypotheses & Literature Review
    - Summary Statistics
    - Hypothesis Testing & Inference
    - Conclusion
* Maximum 10 graphs in the Appendix
* Maximum 5 tables in the Appendix
* Attach the data set you used to your Sakai submission

Dataset Resources:

* <https://www.kaggle.com/>
* <https://www.gapminder.org/data/>
* <http://mlr.cs.umass.edu/ml/>
* <https://data.world/>
* <https://www.icpsr.umich.edu/icpsrweb/ICPSR/search/studies>
* <https://www.visualdata.io/>
* <https://guides.library.cmu.edu/machine-learning/datasets>
* <https://toolbox.google.com/datasetsearch>
* Other datasets you may be interested in

Dataset Requirements:

1. Must have a response variable (you have to try to predict something)
2. Response variable must be quantitative (exceptions available with my permission)
3. Must have at least one quantitative predictor variable (though more is better)
4. Must have at least 30 observations
5. Must have at least three total predictor variables
6. Must not exactly replicate an already published paper’s statistical analysis
7. Response variable and predictor variables cannot be an index (such as happiness index). You can use a dataset that has an index in it, but you cannot use the index variable.

Project Description

In this assignment, I want you to independently locate a dataset of interest to you. The data can come from anywhere, it can be via a survey you conduct, a dataset you find online, or a dataset you located in text and entered into the computer.

**If you find a dataset that has already been used by an author to base a paper on, you need to conduct a different analysis than the author of the paper conducted**. That means if there is a dataset is used by an author predicted infant mortality across countries, you can use the same dataset but need to choose another response variable (for example, medical care level) or add in your own, new, predictor variables.

After finding a dataset that interests you, I want you to first briefly explore the dataset, take a look at some summary data and graph, for your own use, some simple relationships you might be interested in. This is called *exploratory data analysis*. This is not information you need to record or enter into your final paper but it is useful for assessing which hypotheses are probably reasonable and which are not.

Next, I want you to develop some hypotheses about the data. You can find some hypotheses about the data from scholarly literature (<http://scholar.google.com> is a good place to start), from more serious journalistic pieces (articles from news websites or magazines), or other advanced analysis on the topic. In your hypothesis and literature review section, you need to explain the general state of the literature on the subject of your research and then detail the hypotheses that you want to test with your dataset, including specific predictions of how you think the general hypotheses will apply to your dataset. This section should take about 500 words.

I want you to then spend about 500 words summarizing the data in your dataset. First, you should describe how the data in your dataset was collected and in what units the key variables are measured. Then, you should describe your variables distributions and their correlations as well as any important two-variable relationships. These discussions can be based on summary tabular data, distribution plots, or any other basic display of data. In your discussion, you may want to compare any demographic variables (if the observations are humans) or other key summary data and consider how comparable your data is to previous work in the field, including whether these differences might impact your hypotheses.

In your third section, please discuss the results of more sophisticated tests you conduct with your dataset. Hypothesis testing and inference can be conducted using the results of t-tests (where appropriate), but must include at least one multiple regressions. You may use more advanced statistical significance tests but you should seek approval before using anything not presented in class. Also discuss the magnitude of your findings – not only if they are significant but how large of an impact do the predictor variables have on the response variable? Additionally, you will also want to make sure the conditions necessary for regression use are valid – how do the residuals look? How about the partial residual plot between your key predictor variable and your response variable? Then, evaluate whether or not you can accept or reject your hypotheses presented at the start of your paper.

What I do not want in the paper is large amounts of graphs and tables to examine. Focus on the key results and relationships, do not try to analyze every single relationship in the data. What are the key variables for your hypotheses? Which statistical tests are most informative? Describe and graph those relationships. The fewer graphs and tables you need to arbitrate between your hypotheses the better! As always, make sure graphs are well labeled, appropriately scaled, of high quality, and relevant to the questions raised by your paper.