STAT 538A Final Project

Due: On Canvas at 5pm Dec 8, 2020

For the final project, you (independently; this is not a collaborative assignment) must find a dataset, conduct a comprehensive data analysis, write a statistical report, and make a 8 minute presentation about the analysis. You will need to find a dataset by yourself, e.g., from internet, books, data from your own projects, etc. For example, Statlib: `http://lib.stat.cmu.edu/` has many datasets available, as well as many statistical journal websites such as Biometrics and JASA. Once you find a suitable dataset, you will analyze the dataset using the models and methods learned in this course and report your results and conclusions. That is, at least one GLM/Bayesian model must be used in your analysis (not a linear model).

Your written report (35% of the 45% total for the project) should be less than 15 pages (12-point font size), including figures and tables. Your report should be typeset using LaTeX. Your report should contain the following materials:

- Objectives of the study.
- Data description (e.g., how and where the data were collected, what variables were measured, sample size, data summaries, etc).
- Exploratory data analysis (e.g., summary statistics, graphical displays of the data such as boxplots, and preliminary conclusions. No models are assumed in this section).
- Confirmatory data analysis (e.g., GLM, model selection, model diagnostics, justifications of models, and interpretations of results). It is not a bad idea to consider two or more different approaches/models to see if the results agree and explain why or why not.
- Final conclusions and discussions (e.g., discussions of limitations/advantages of different models/methods used in the analysis).

Your oral presentation (10% of the 45% total for the project) should be less than 8 minutes long, and consist of slides that summarize the same material that appears in your report. This will be presented in the last two lectures of class.

Remarks:

- Do not use the same dataset as someone else in class. Post the dataset that you will be using on the course discussion board, with a URL. First come first served.
- Tables and figures should be selective and informative (i.e., do not include too many tables or figures).
- Computer outputs should be summarized (i.e., do not include raw computer code/output in your report) and should be interpreted. Python code should be put in the Appendix.

Evaluation criteria:

- The report should be concise, with a focus on important materials such as motivations and justifications of the chosen models and methods, interpretations of analysis results.
- Scientific writing is important. Please make sure you write clearly and logically.
- Evaluation will be based on both statistical criteria and scientific criteria, as well as the depth of your work. Try to avoid a routine analysis without much insight.
- The presentation will be graded based on its clarity and organization

The graded final report will not be returned to you.