STAT410 Course Project

In this project, you will conduct a full analysis of a dataset to answer a question of interest to science/industry/public or simply yourself. You should use the tools we have learnt throughout this semester; tools beyond this class are allowed and welcomed but not required. The goal of the project is to rehearsal your debut as a statistician to apply statistical skills in your own field of application.

**Timeline:**

11/21: Final report due (no more than 10 pages)

11/1: proposal due (no more than 1 page) – early submission strongly recommended

11/15: report due (almost final version) & presentation due

**Guideline:**

The project is default to be individual; a group of 2-3 is allowed with prior permission.

The final report is a docx/pdf file – no more than 10 pages (Appendix, References, and Self-reflection do not count toward this page limit). It should include the following key components:

1. Statement of the Problem: Purpose of your project.  
   What problem(s) or question(s) did you set out to solve? What are the goals of the study? Are there any effects of particular interest?
2. Introduction (background and plan):
   * Proposal - Describe how you plan out your project. What type of background reading did you do? What information did you use in order to better conceptualize your project and frame a design? How might your goals be answered, i.e. tests / confidence intervals?
   * Final report – Description of the background knowledge and your analysis plan.
3. Data collection. You should describe the data set such as the resources and variables, with basic summary statistics and/or visualization.
4. Data analysis and results. In this section, you should develop and diagnose models for the data that will allow them to answer some of the specific goals of the study, and report their results obtained by fitting the proposed models. Loads of R output would, in general, not be acceptable. Plots and well-organized tables are good things to have in this section. Possible questions to be addressed here are the following:
   * What is the final regression model for the data?
   * How was this model obtained, i.e. forward stepwise search (to be seen in class)?
   * Using the standard diagnostic tests, does the model appear to fit the data well?
   * What are the final confidence intervals for the effects of interest mentioned in the study section? Do these intervals seem very sensitive to the choice of model (i.e. do they vary widely for different choices of variables in the model)?
5. Summary & Discussion. Briefly summarize your data analysis results and findings and state any future work if applicable.
6. Appendix: In this section, you should attach an executable version of R code used in the analysis. Ideally, there will be comments in the file, i.e. lines beginning with “#” to clarify what each part of the code is doing.
7. References (and/or Acknowledgements): If you consult outside sources that refer to this data set, you should cite these as references, and describe what you used from each source. Sources include material found on the internet, journal articles and books.
8. Self-reflection: How much time you devoted into the project? What you learned about the process of doing your project. What went wrong? What would you do differently next time? What advice would you give future students in this class?

**Alternatives to data analysis:**

Other possibilities include: investigation of some properties in linear regression (e.g., validation of QQ plots, theoretical properties of estimators under model misspecification), cool visualizations, to develop an interactive app, etc.

**Evaluation of projects:**

* Following the time line
* Problem to be solved (novel? Interesting?)
* Clarity and completeness of the final report
* Technical advancement and correctness from data collection to results
* Presentation
* Highlights such as data collection, visualization, interactive app, etc.

**Examples of ideas:**

1. Rice/Houston related problems.
2. Data science platform (kaggle, e.g.)
3. Sports statistics.
4. Website for datasets: Go to http://www.infoplease.com/us Choose the U.S. statistics link and pick a topic that is interesting to you.
5. Financial statistics (yahoo finance, e.g.)
6. Dean De Cock. Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. Journal of Statistical Education, 19(3), 2011.
7. A sample of past projects:
   1. What Houston Floods Tell Us: An Analysis on Memorial Day Flood and Tax Day Flood
   2. Determining the Relationship between Psychological Distress and Drug Abuse/Sociodemographic Factors
   3. Course Ideals among Rice Students
   4. Predicting Voter Turnout in Rice University Presidential Elections to Improve Campaign Directed Voter Targeting
   5. Weather Elements that Influence San Francisco Bay Area Bike Usage
   6. Why Do More and More People Choose to Live Alone?
   7. An Investigation of the Perceptions of Password Security
   8. Are We Able to Predict How Many Opioid Prescriptions a Doctor Will Give Annually?
   9. Do Liberal Arts Students Earn More Than STEM Students in the Long Run?
   10. (Statistically) Significant Others: The Science of Speed Dating
   11. Multivariate Linear Regression Model on Departure Delay Time
   12. ​Did Harvey Hit on Stock Price: What to Look?
   13. Relationship between Nutrition and Food and GPA in Students at Mercyhurst University
   14. From Film to Data: MLR models for predicting IMDB scores and Gross Revenue