**STAT 632, Project Paper**

**Instructions**:

- You may work individually, or in a group of up to 3 members. If working in a group, each member should write at least one section of the paper.
- The paper should be about 3-5 pages with tables and figures (but not including code). Include the R code for your analysis as an appendix at the end of your paper.
- Please submit your completed project as a PDF to Blackboard by the deadline. Late submissions will not be accepted.
- The paper should focus on multiple linear or logistic regression modeling. However, you may also consider other modeling techniques (e.g., LASSO, random forests).
- The paper should be organized as follows:

  1. **Title Page**. Include a title and the names of each author.
  2. **Introduction**. Describe the main research questions and goals of your data analysis and statistical modeling. 1-2 paragraphs should be sufficient.
  3. **Data Description**. Briefly describe your data. What is the source? What are the response and potential predictor variables? What is the dimension (number of rows and columns)? In this section, also present and discuss relevant summary statistics and graphical displays of your data set (e.g., scatterplots, box plots). Be selective about the descriptive statistics that you decide to include.
  4. **Methods and Results**. Describe the methods used to estimate and select your regression model, and present the major modeling results (e.g., regression summary table, diagnostics plots, cross-validation results). Provide a concise description of the results and your interpretation.
  5. **Conclusion**. Summarize the major conclusions and findings of your regression analysis. In this section, you can also provide ideas for future work.
  6. **Code Appendix**. You may also provide a supplementary R script, or link to a GitHub repository.

  All tables and figures should be numbered and referred to by number (e.g., Figure 1, Figure 2, Table 1, etc.).

**Grading**: A list of specific expectations are provided below.

- The research questions and goals of the analysis are clearly described.
- The source of the data set is provided, and the relevant variables are listed and described.
- The selected descriptive statistics illustrate important aspects of the data set.
- The statistical methods considered are appropriate for the data set.
- The methods and results are thoroughly and clearly described using precise notation.
- The paper is well-formatted and organized. There are very few typos or grammatical mistakes.
- Figures and tables are well-formatted with appropriate labels.
- The R code in the appendix is easy to follow and reproducible.

Papers that meet these expectations will receive an A. Papers with minor flaws, that mostly address the above expectations, will receive an A-. Papers that fail to address several of the above expectations in critical ways will receive a B or B-. For example, papers that have poor formatting, organization, and/or writing will receive a B or B-. Papers that are incomplete, plagiarized, and/or demonstrate little interest or effort will not receive a passing grade.

**Data Sources**
Here are some potential sources for data sets. You do not need to limit yourself to these.

- Kaggle: `https://www.kaggle.com/datasets`
- UCI machine learning repository: `https://archive.ics.uci.edu/ml/datasets.php`
- DataSF: `https://datasf.org/opendata/`
- FiveThiryEight: `https://data.fivethirtyeight.com/`
  R package: `library(fivethirtyeight)`
- United Nations data: `http://hdr.undp.org/en/data`
- Google data set search: `https://datasetsearch.research.google.com/`

You can also use a data set from one of the regression textbooks cited in this class. However, you cannot use a data set that has already been used in lecture or HW.

- *A Modern Approach to Regression*:
  `https://gattonweb.uky.edu/sheather/book/data_sets.php`
- *OpenIntro*: `https://www.openintro.org/data/`
  R package: `library(openintro)`
- *Stat2*: `http://www.stat2.org/datapage.html`
  R package: `library(Stat2Data)`
- *An Introduction to Statistical Learning*
  `https://www.statlearning.com/`
  R package: `library(ISLR2)`
- *Linear Models with R*
  R package: `library(faraway)`

To get a list of the data sets in an R package you can run the command `data(package = "name")`. For example, run the following command to get a list of the data sets in the `openintro` package:

```
data(package = "openintro")
```

**Presentation**

- Presentations will occur during the last two weeks of the course
- Presentations will be randomly ordered (I will use R in class to determine this order)
- Presentations will be in 20-minute chunks but the presentations should be no more than 15 minutes to allow for transitions and Q&A
- Each member of the group must contribute verbally to the presentation

**Timeline**

- 3/25, an email (one per group) with the members of the group
- 4/7, an email (one per group) with a link to the data set and its description
- weeks of 4/25 and 5/2; presentation of projects