Project Instructions

For this assignment, you are asked to select a dataset and explore it using some of the tools we have talked about this quarter:

- Naive Bayes Classifier (Week 2)
- Linear Regression (Week 3)
- Logistic Regression (Week 4)
- LASSO or Regularization, Classification or Regression (Weeks 3 and 4)
- Random Forests, Classification or Regression (Week 5)
- Bagging or Boosting, Classification or Regression (Week 5)
- K-Means or other Clustering (Week 6)
- Principal Component Analysis, t-SNE, or some other dimensionality reduction method (Week 6)
- Neural Networks (Weeks 7, 8, 9)

You are encouraged to find a dataset that is interesting, even if you are only able to explain a tiny part of it. We will note that many students find that they get out of projects what they put into it–**choosing a more challenging dataset results in a deeper understanding of the methods as they are applied to real datasets, as well as what it means carry out a data analysis from start to finish**. Feel free to use methods not discussed in class, even if you don't understand the theory or details. After applying the methods, you should write a report that includes:

- An abstract summarizing your entire project (no more than 300 words). The abstract should appear on its own page.
- A paragraph with a brief explanation of the origin and the meaning of the dataset. When and where is the data from, and why might we care about it? An entry in the Works Cited must indicate the origin of the dataset in sufficient detail to permit it to be found, with dataset name, author, and revision number in addition to the URL.
- A paragraph describing what others have found with your data or data like it. This requires research outside the dataset. Additional sources should have adequate bibliographic entries.
- A paragraph with summaries of initial exploratory data analysis, including no more than one or two visualizations leading to the main exploration. Visualizations must be appropriate, informative, relevant, and free of correctable flaws (everything that needs a label must be labeled, fonts must be no smaller than half the font size of the text in the report, included images must not be grainy or illegible, etc.). If you include a figure, it should be accompanied by a caption as well as discussed in the text. The origins of the dataset, descriptions of what others have found, and exploratory data analysis should be no more than one page.
- Two applications of the techniques listed above to summarize/understand/predict something about the dataset. The methods must be appropriate for your chosen dataset.
- A comparison of the applications of the techniques. How are they different from each other? Is one of them more appropriate for the dataset? Does one of them outperform the others? **Model performance must be summarized with appropriate model performance metrics and one figure each that communicates the main finding**. The applications and comparisons should be no more than one page.

Your grade will depend on three things:

- The quality of your data analysis;
- The quality of your data reporting. Explaining what is in a dataset is difficult, and doing it in a way that is easy to read is even more so.
- The challenge you have set for yourself in terms of risk taking or innovative thinking (see rubric for more details).

Students should work in groups of 2 or 3. Your project requires a statement as to who did what work, which should be included below the abstract (on the same page). In total, your submission should be four pages (one for the abstract, one for the exploration, one each for the analysis methods).

Datasets

There are many places/organizations which collect and index datasets–sometimes because the data is cool, because it has been analyzed before, or because data collection is part of the organization's mission. You may use any source you like, but if you need a starting point:

- Awesome public datasets
- Centers for Medicare and Medicaid Open Payments database (pharma compensation to prescribers, dubbed by Pro publica "Dollars for Docs")
- Chicago City Data portal. You can find multiple municipal datasets of various sorts, including city finances, communication, and law enforcement.
- Guttmacher Institute Public-Use Datasets on Abortion and Fertility by Geography and Year
- Indicators of Gender Equality from the World Bank 1960-2017
- Tidy Tuesday Data Repository–multiple datasets for different weeks, going back to 2018 (can also be used with Python, all data in .csv format)
- United States Census Bureau and other government sites.
- Washington Post's DEA Pain Pills Database
- Any other source you find interesting.