Project Instructions

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

- Linear Regression (Weeks 2 and 3)
- Logistic Regression (Week 4)
- Ridge Regression (Week 5)
- LASSO (Week 5)
- k Nearest Neighbors (Week 6)
- Clustering (Week 7)

You are encouraged to find a dataset that is interesting, even if you are only able to explain a tiny part of it. Students should work in groups of 2 or 3.

After applying the methods we have learned about, you should write a report that includes:

- An investigation and 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? A footnote or 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 review of what other people have found looking at this data or data of this sort. This requires research outside the dataset. Your additional sources should also have adequate footnotes or bibliographic entries.
- One or more applications of data modeling/data reduction techniques from the class to summarize/understand/predict something about the dataset. The methods must be appropriate for your chosen dataset.
- Summaries of the data and your model of the data as visualizations. Visualizations must be appropriate and informative. Visualizations must also be 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.). Each figure needs a caption explaining the figure.

Your grade will depend on 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 page limit for your report is 6 pages including visualizations (not including bibliography or submitted code). Some people can do a good job in 4 pages.

Group projects require a statement as to who did what work (though it will not be scrutinized the way the visualization and bibliography will be).

General Timeline

The project is not due for quite some time; however, you should start working on it now. A general description of the work you should be able/should have been able to do after each week of the course appears below–please note that you do not have to attempt every method in the table!

Week 2	Dataset selection, explanation of dataset origins, initial lit review, some initial data cleaning, exploratory data analysis including data visualizations.
Week 3	Selection of response variable, more exploratory data analysis including correlation matrices and identification of possible collinear/multicollinear variables based on dataset documentation. Optional: Project Assignment A due Wednesday, October 18 (no credit).
Week 4	Identification of appropriate type of model (classification vs. regression), initial passes at multiple regression model fits. Analysis of residuals if using linear regression. Optional: Project Assignment B due Wednesday, October 25 (no credit).
Week 5	Summaries of model fits (e.g., r^2), analysis of important features (significance tests). Optional: Project Assignment C due Wednesday, November 1 (no credit).
Week 6	Initial passes at ridge regression or LASSO (classification or regression).
Week 7	Initial passes at kNN classification or regression models. Optional: Project Assignment D due Wednesday, November 15 (no credit).
Week 8	Initial passes at clustering.
Week 9	Project due 11:59pm on Wednesday, November 29.

Datasets

You can find interesting and complex datasets all over! 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 (US Census Bureau, US Department of Agriculture, US Department of Energy, US Department of Labor, etc.). When looking for your data, check out the following sites:

- 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)
- Washington Post's DEA Pain Pills Database