Project #1

Machine Learning 1

Project 1

The first project this semester will focus on doing data wrangling, exploratory data analysis, and visualization using a publicly available survey data set called the General Social Survey.

The **General Social Survey** is a panel data survey about Americans' social and economic views conducted since 1972, available at https://gss.norc.org/. Since it is a panel, the same people are surveyed over time, which is very useful and interesting, since we can see changes at the individual level. On the other hand, it means that within a year, only about 3000 people are actually surveyed. Despite that, the data are about 40MB zipped and about 520MB unzipped, because the survey has been conducted for so many years.

Another key feature of the GSS is that some questions are constant and provide context and continuity, while other questions might only appear for a few years. This means that you might ask a question about change over time, but you'll have to use a question with a long history of being asked, or you can focus on a question about the cross section of opinions at a given moment, but you won't be able to look at dynamics.

0. Look at the Codebook. It is big, but the pdf is searchable and there are a lot of interesting questions asked.
1. Pick a research question. What are you interested in understanding? At this point, we don't have many analytic tools available, so it's OK for the question to be simple or vague. At this point you have no tools for asking "why"/causal inference questions, so it's best to keep the question simple.
   Question: Is there a relationship between age and political preferences?
2. Get appropriate data from your data source. Clean it, documenting the choices you make and referring often to the codebook. **POLVIEWS & AGE**
3. The core of your project will probably be a grouped kernel density plot or histogram, grouped scatterplots, or some other kind of visualization. Try to think about getting that piece decided as quickly as possible, since that means the rest of the project will quickly fall into place.
4. Instead of being too ambitious about data wrangling or analysis, think instead about iterating on steps 2-3 until you have good results, and then get to work on writing. If your project loses focus or becomes too ambitious, you will spend a lot of time agonizing over choices without knowing the consequences. This is how projects get out of control, in general.
5. Once you have results that your group is satisfied with, arrange the code according to the format below and write around the code chunks to complete the paper. The .ipynb file you submit that includes your work should be something I can compile (i.e. your work should be reproducible).

Paper format

The format of the paper should be:

• Summary: A one paragraph description of the question, methods, and results (about 350 words). • Data: One to two pages discussing the data and key variables in the analysis, and any challenges in reading, cleaning, and preparing them for analysis.

• Results: Two to five pages providing visualizations, statistics, and a discussion of your findings. If you have a lot of plots or tables, that's OK, but try to focus on a few key pieces of evidence rather than doing every single pairwise comparison of some set of variables.

• Conclusion: One to two pages summarizing the project, defending it from criticism, and suggesting additional work that was outside the scope of the project.

• Appendix: If you have a significant number of additional plots or tables that you feel are essential to the project, you can put any amount of extra content at the end and reference it from the body of the paper.

Group Work and Submission

For each group, I will create a private GitHub repo under DS3001 that only your group members can access.

Unlike the homework, you will be working with other people. We will cover how to branch and merge in Git,

so that each group member can start work and contribute to the project in their own branch or by branching off of previous members' work. We will discuss how to use more advanced features of Git as we start work on the project.

Half of each student's grade is based on their commits to the repo. Each student is expected to do something specific that contributes to the overall project outcome. Since commits are recorded explicitly by Git/GitHub, this is observable. A student can contribute by cleaning data, creating visualizations, or writing about results, but everyone has to do something substantial. A student's work doesn't need to make it into the final report to be valuable and substantial, and fulfill the requirement to make a contribution to the project.

The other half of each student's grade is based on the report. Groups will work together on combining results and writing up findings in a Jupyter notebook, using code chunks to execute Python commands and markdown chunks to structure the paper and provide exposition. The notebook should run on Colab or Rivana from beginning to end without any errors.

Notebook Criteria

The grading for the project notebook is graded based on five criteria:

- Project Concept: What is the research question? Is it well-defined? Are the data appropriate to address the question? References to data documentation or codebooks are essential in this part of the project. What is the research strategy, and is it appropriate to research question?
- Wrangling: How are are missing values handled? For variables with large numbers of missing values, to what extent do the data and documentation provide an explanation for the missing data? If multiple data sources are used, how are the data merged?
- Exploratory Data Analysis and Visualization: For the main variables in the analysis, are the relevant data summarized and visualized through a histogram or kernel density plot where appropriate? Are basic quantitative features of the data addressed and explained? How are outliers characterized and addressed? Are tools for visualizing the relationships between variables used appropriately and interpreted correctly? Are statistics like the mean, variance, median, quantiles, and correlation used appropriately?
- Analysis: What are the main findings of the research? Do the plots and statistics support the conclusions? Is the research strategy carried out correctly? If the research strategy succeeds, are the results interpreted correctly and appropriately? If the research strategy fails, is a useful discussion of the flaws of the data collection process or the research strategy discussed?
- Replication/Documentation: Is the code appropriately commented? Can the main results be replicated from the code and original data files? Are significant choices noted and explained?

Each of the five criteria are equally weighted (10 points out of 50).