INFO 348 – Fall 2023

Final Project

Proposal due Friday, December 1st at 11:59pm
Project due Wednesday, December 13th at 11:59pm

Bringing It All Together

For the final project, you will be conducting exploratory data analysis and writing up your findings. The goal for the final project is to give you an opportunity to apply some of the techniques we've covered over the course of the semester while asking and answering your own questions from the data. You may work individually, or collaborate with another student on your project.

The project is intended to be open-ended, in that it's up to you to decide where to focus your efforts. You can work with any of the data sets from the attached list, or you may use your own (*with instructor approval!*).

Your exploration should consist of applying several of the techniques we studied over the course of the semester. For instance, you may choose to cluster your examples, and offer some interpretation of the output. You could create a series of plots or maps that bring some aspect of your data set into specific relief. Alternatively, you could build a predictive model to try and classify your data, or compare the results of multiple classifiers. The goal is to find a question (or set of questions) about the data that interest you, and attempt to answer them through your analysis.

Requirements

For full credit, you must complete a minimum number of the items from the following list of data and analysis tasks. If you're working alone, you must complete at least *three*; if you're working with a partner, you must complete at least *five* (partners may do one of the tasks twice, as long as the effort required is at least double).

Furthermore, the tasks fall into two categories: *data tasks* and *analysis tasks*. Single-person projects must include at least one analysis task, while two-person projects must complete at least two. The tasks for each category are:

Data Tasks

- *Data Wrangling*
  Create a unique data set by transforming and combining multiple existing sources of data into a cohesive whole. Must involve a non-trivial amount of code to manipulate

data into something usable for analysis.

- *API Access*
  Construct a data set using an API. To receive full credit for this task, the data manipulation must be not trivial: you must make calls to at least two different API endpoints and join the results together into a cohesive data set (simply using a single API call to pull down a CSV is not enough).

- *Databases*
  Programmatically store and query your data using a relational database. While we focused on reading from RDBMSs, you may elect to set up a SQLite database using `CREATE TABLE` statements and loading in your data. Once your data is loaded, you can demonstrate the use of queries to answer questions, summarize, or understand the data set.

- *Other, with permission*
  Come up with your own data task. Just make sure you include it in your project proposal to get approval.

Analysis Tasks

- *Statistical Analysis*
  Summarize and explain particular aspects of your data using summary statistics or other calculations. Simply reporting on means or modes is not enough, you must instead seek to identify and quantify interesting findings in your data set as a whole or particular subpopulations.

- *Visualization*
  Create a series of plots that bring some aspect of your data set into specific relief. The types of plots you use should be determined by what you want to convey, and what questions you hope to answer.

- *Geocoding*
  Perform some spatial analysis using a geocoder to calculate locations and distances between data instances or other points of interest. You might also use location data to plot on a map.

- *Clustering*
  Use the unsupervised learning functionality from scikit-learn to perform clustering on your data set. Be sure to explain what types of groupings you're hoping to capture, along with sample outputs and descriptions.

- *Predictive Modeling*
  Build a classifier (or regressor) to predict some feature of your data using scikit-learn. Include a performance analysis, along with any explanation of modeling choices you made (e.g., model types, parameter values for the models, attributes included or excluded).

- *Statistical Associations*
  Identify statistical associations found in your data, showing quantitative and visual proof of the relationships between variables. Discuss the meanings of these associations

- *Causal Hypotheses*
  Take your association analysis one step further by discussing possible causal mechanisms that may be in effect, and describe possible experiments that might clarify the causal dynamics, or additional data that would be helpful. (Note: you can also test for conditional independence using the pingouin module).

- *Other, with permission*
  Come up with your own analysis task. Just make sure you include it in your project proposal to get approval.

Data Sets

Your analyses should be based on one of the data sets from the list in this document. You are free to augment that data with other publicly available data if useful. You can also use your own data set *with prior approval*.

Project Proposal

As a first step to completing your project, you must fill out the proposal form here and submit a pdf to Gradescope. You do not need to write a lot for the proposal form, but you should have a look at the data you want to work with and do some thinking about your approach. While you are not required to stick to the tasks listed in the proposal, if you want to make major changes to your project (such as switching data sets, adding a partner, etc.) you should resubmit the form.

Format

You must submit all code, along with a writeup (see below). The level of effort required for your project should be equivalent to 1.5-2x a normal homework assignment. Additionally, if you're working with a partner, the scope of your work and writeup is expected to be more substantial than if you're working by yourself.

You are encouraged to leverage existing tools (pandas, scikit-learn, etc.) wherever possible. You are free to base your code off of any of the examples covered in class or homework assignments. While your code doesn't have to be pristine, you should include comments that help a reader to understand how everything functions. You are free to use Python scripts (.py) or notebooks (.ipynb) or both. The use of large language models (e.g. ChatGPT, Copilot) is prohibited, as is replicating previous analysis.

The Writeup

The writeup should include 2-4 pages worth of text describing your work, and should address the "*why*" behind your efforts in addition to the "*what*". It's especially important to articulate the questions about the data you're hoping to answer through analysis. How you structure the writeup is up to you, but in general you should include:

- The names of all team members, along with a brief overview of how each person contributed
- A description of the data set, including any preprocessing you did to get the data into a usable format
- A short writeup for each task, summarizing the techniques you used, as well as any conclusions you were able to draw
- An overview of the code you wrote and existing tools you used, along with instructions on how to run the code
- Description of challenges you encountered when working with the data, and how you were able to overcome them (or not!)
- Descriptions of any insights into the data or domain that you obtained through your work
- Ideas for future exploration of the data, including interesting questions raised by your analysis

Grading

Your project will be graded holistically, taking into account effort, creativity, degree of difficulty, technical proficiency, quality of the writeup, etc. Note that "negative results" are totally acceptable for this assignment (for example, "here's several things we tried along with a theory of why none of them worked").

What to Submit

You should submit single zip file containing:

- A file writeup.pdf, with your writeup (described above)
- All code used to generate results
- Data files used in the project (if they are > 5MB in size, you can include a text file called data.txt containing links to the data sets)

Appendix: Pe-Approved Data Sets

The following data sets have all been vetted by course staff. You are free to use another data set with permission.

- **FEC 2021-2022 Political Donations Data**
  https://www.fec.gov/files/bulk-downloads/2022/weball22.zip
  (other years can be found here: https://www.fec.gov/data/browse-data/?tab=bulk-data)
  Documentation:
  https://www.fec.gov/campaign-finance-data/all-candidates-file-description

- **Eviction data from Eviction Lab**
  https://data-downloads.evictionlab.org
  There are several data sets here, along with data dictionaries and codebooks explaining what the different columns mean.

- **Economic Mobility Data from Opportunity Insights**
  https://data.humdata.org/dataset/85ee8e10-0c66-4635-b997-79b6fad44c71/resource/ec896b64-c922-4737-b759-e4bd7f73b8cc/download/social_capital_county.csv
  https://data.humdata.org/dataset/85ee8e10-0c66-4635-b997-79b6fad44c71/resource/ab878625-279b-4bef-a2b3-c132168d536e/download/social_capital_zip.csv
  Documentation:
  https://data.humdata.org/dataset/85ee8e10-0c66-4635-b997-79b6fad44c71/resource/fbe5b0b9-e81c-41c7-a9f2-3ebf8212cf64/download/data_release_readme_31_07_2022_nomatrix.pdf

- **Genius Music API**
  https://docs.genius.com
  (requires authorization with an API key)

- **FoodData Central API**
  https://fdc.nal.usda.gov/api-guide.html
  (requires authorization with an API key)

- **CFPB Financial Well-Being Survey**
  https://www.consumerfinance.gov/data-research/financial-well-being-survey-data/

- **StatsAmerica Census Downloads**
  There are many different data sets here, ranging from population characteristics to economic and well-being factors.
  https://www.statsamerica.org/downloads/default.aspx