SI 330 Final Project
Winter 2020

Project Overview

In the SI 330 final project you'll use the skills from class to tackle a messy problem of your own design. You will find two datasets from different sources including different types, merge and manipulate them, and then gain an insight that you could not have gained with a single dataset. The manipulation could involve filtering, format conversation, dealing with missing/noisy data, match records and so on. Your manipulation must make use of some of the techniques we've covered in class. Because data manipulation work can be both time-intensive and under-appreciated, your project will involve explaining your work and why you did each step. Each person will have their own project, because human data manipulation is difficult work to do in parallel. Also, I encourage, but don't require, you to find a project that might have some application towards social good.

The project grade will consist of:

- Project Work Plan (15%)
    - First draft (via Google Docs) due ~~Tuesday, March 17~~ Thursday, March 19th
    - Final version (via Google Form) due ~~Sunday, March 22~~ Tuesday, March 24
- Peer Reviews (5%)
    - Due ~~Thursday, March 19~~ Sunday, March 22
- Final Project (80%)
    - Due April 22nd at 11:59 pm

Learning Outcomes:

- Understand the real-world challenges involved in working with data sets
- Use data manipulation skills in a messy, ill-defined situation
- Understand the real-world application of the skills taught in SI 330
- Combine skills used in a variety of contexts to analyze data
- Be able to describe the work you've done and the value you've added to a project

SI 330 Final Project
Winter 2020

Project Work Plan

Before you jump into any data manipulation task, you should *write down* a work plan that tells you (a) what you will be doing, and (b) why you are doing it. In most cases, your work plan will change as you progress. However, having the work plan written down will help you understand what your goals are. It will also help keep you on track when you start to get lost in the weeds. You can adjust your work plan and project later as your progress proceeds.

A good plan includes:

1. What are you doing and why?
2. Where will you get the data from? (Make sure you can get it!)
3. What data manipulation/analysis steps do you expect you'll have to do?
4. How long do you think this will take? (This one can be super-hard.)
5. If this works like you expect, what does a finished product look like?

These are just some basic pieces. Your work plan will be different, depending on who you're writing it for. For this assignment, we have specific questions to make sure your project stays on track.

Directions:

- Make a local copy of the Project Work Plan Template form (Google Doc).
- Fill in each section before class on ~~Tuesday, March 17.~~ (See changed dates on first page.)
- Come together to work on this on ~~Tuesday, March 17.~~
- Submit it via a Google Form link by ~~11:59 pm on Sunday, March 22.~~

Peer Reviews

It's worthwhile getting feedback on your work plan. ~~So bring your work plan to class on Tuesday, March 17. We'll discuss the work and give each other feedback.~~ **Your grade in this portion will be based on the reviews that you *write*.** If you cannot attend class on that date, it will be up to you to arrange on Slack to find a group of people that you can exchange work plans with. You will need to upload pdfs of your reviews to Canvas by ~~Thursday, March 19 at 11:59 pm~~

Directions:

- Copy (don't write over) at least two other peoples' work plans in Google Docs.
- Give them feedback by making comments or marking up your copies. (~~We'll do this in class on Tuesday, March 17.)~~ (See changed dates on first page.)
- Send them a pdf of your review. ~~(We'll do this in class on Tuesday, March 17.)~~
- Upload the pdf reviews *that you did* (not the reviews of your project) to Canvas by the end of the day on ~~Thursday, March 19th.~~ (See changed dates on first page.)

Final Project

The final project will be a written summary of the work you've done. It should be roughly 4-5 pages and include the following labeled sections, though it can include more. Each section should address the questions listed.

Motivation

a) Briefly state the nature of your project and why you chose it.
b) What specific question or goal did you try to address?

Data Sources Describe the properties of the two dataset(s) or API services you used. Be specific. You should include (at minimum):

- How/where did you access the datasets or API resources? (Include URLs, if appropriate.)
- What formats were returned/used?
- Which variables did you think were important and which did you use? (Note that you need to use multiple data types.)
- How many records did you retrieve/use?
- What time periods did they cover (if relevant)?

For example, give URLs. I should be able to easily find and access the resources you used.

Data Processing This section should have two parts:

1) A summary, in paragraph form, of the data processing steps you did to get the data workable. This should give enough information that someone with your skill level could reproduce your work and get roughly the same results.
2) A bulleted list. Each bullet should have a problem or challenge in working with the data and a sub-bullet(s) that describe your solution. All of the steps you took should be included in the list. This should give enough information that someone with your skill level could reproduce your work and get *exactly* the same results. (Example below.)

Example:

- *We need to have similar keys to merge the data on, but the time stamps were in different formats.*
    - *I used the datatime module to format both time stamps in the same way.*
- *Some times were in Dataset A, but not in Dataset B.*
    - *I used an inner join, which only kept the entries on both datasets.*

Analysis and Visualization You brought the two datasets together to a new piece of information. This is the place to explain the answer to your question or topic (found in Motivation).

1) Describe the data analysis process. Be specific about the variables you used and how you analyzed them.
2) What interesting relationships or insights did you get from your analysis?
3) What didn't work, and why?
4) Include at least one visualization that summarizes your analysis. This should, at least in part, use your skills with Matplotlib and/or Seaborn.

Attributions Clearly list anyone who helped you with your project, any ideas or code that you used that came from someone else, and any other classes where you're currently working on the same data. If you're using your project data elsewhere, be specific about how you're using that data in that other class.

To Turn In:

- Your written, nicely-formatted project report as a pdf.
- Your code, in .ipynb format.
- Your code, exported from Jupyter into html.