Data Project

Spring 2020

During the second half of the course, you'll work on a data processing project to answer a quantitative question of interest. The purpose of the project is to help you connect this course content directly with a professional or personal interest of your own, and to perform a complete start-to-finish piece of data analysis using the tools and techniques we've been developing (and will continue to develop).

The question

So, what makes a good question? For our purposes, a good question is one that you

- are interested in,
- have data about (preferably in CSV format), and
- can answer by applying a straightforward algorithm to that data.

What makes an algorithm straightforward is, to a first approximation, something you can write as one or more accumulator loops through the data. We'll talk more about that in the next week or two.

What's likely to be the more important issue is finding data. Two excellent sources of relatively clean data are CORGIS[1] and data.gov[2]. You might also ask some of your professors in your major if they have any interesting data sets to work with—I'm happy to try to help convert them into a format that you'll be able to process. Browse through these data sources and find something that appeals to you. It's great if it connects to your major, but it could also connect to a hobby or just plain sound interesting to you. Look at the columns of the data and think about what you might ask about.

[1] `https://think.cs.vt.edu/corgis/csv`

[2] `https://catalog.data.gov` , look for CSV data sets

Tell us about it

After you've got an idea of the question (or perhaps a few related questions) that you want to ask about your data, you need to write it up. Your initial proposal will include a link to the data set, explain how the data represents information of interest, and formulate your proposed question(s) in terms of the algorithm you intend to use to solve it. The document should be about a page long (12pt, double spaced, 1" margins). There should be enough background that a reader can understand why the question is interesting, and at least a little detail on how the data will be processed to answer it.

I'll respond to the proposal and (if necessary) help you refine it into something workable.

Then, after it's approved, you'll write a much briefer writeup to stir up interest (details forthcoming), which you'll post on the Canvas site for everyone in the class to read.

Answer it

The core of the assignment will be writing the actual program that implements your chosen algorithm to perform the data analysis you need. Your program will read in your data file and produce a value or values that can be used to answer the question(s). (The actual *answer* to the question(s) may require further interpretation.) The program should be written such that if the data set were updated or traded for a related data set in the same format, the program would still run (and produce an updated answer).

Tell us about it

Once your program is running reliably and you're sure it's running correctly, you'll need to wrap up the project by using the program's output to actually answer the question; an important step here is to understand enough about the problem to know whether your answer is plausible (i.e. not obviously the result of a bug!), and to situate the answer in the context of the original question.

In the lead-up to handing in the final writeup, you'll do two more short posts for the class to read about your project. The first will have a technical slant, explaining to the class in formal terms what your algorithm does and

how it works to answer your question. The second will return to overview mode, and actually *answer* the question.

The final writeup should include the content from the original proposal and both later updates, edited to clearly illustrate the quantitative reasoning process, plus instructions on how to run your program (so I can duplicate your results). The format is, again, 12pt, double-spaced, with 1" margins. If you make statements that require citation, you should cite them (obviously!), but you can use APA, MLA, or whatever format you're most familiar with in your own discipline.

Your submission will be, via canvas, a .zip file containing

- The .py file you're running
- The data file(s) that you can run it on
- The writeup (probably a .doc or .docx file)

Scoring summary

Proposal writeup	12
Post about question	8
Program	40
Post about process	16
Post about answer	16
Final writeup	8
Total	100

Deadline calendar

Note that the first deadline are a change from the original syllabus and the Canvas posts are replacing the speeches and talks laid out in the syllabus.

March 23 (Monday)	Proposal writeup
April 1 (Wednesday)	Question post
April 24 (Friday)	Process post
April 29 (Wednesday)	Answer post, final writeup, program