W200 Project 2

James Kajdasz

July 22, 2018

W200 Python Fundamentals for Data Science, UC Berkeley MIDS

Project 2

The final project will consist of an analysis of a dataset, using the numpy and pandas tools we covered in class.

The Setup & Instructions

- 2 to 3 people (no exceptions)

The Data Analysis

- The exercise is unguided. You will come up with your own interesting questions that you can answer using the variables in your dataset, then perform an analysis based on those questions.
- You can analyze either an *instructor approved dataset of your finding* or one that we have preapproved. You may join several datasets together. For example, you might combine two datasets to answer questions like, after 2 days of rain, is felony crime higher in NYC?
- The project emphasizes the exploratory and descriptive techniques covered in class. Although some of you have a background in statistics, we ask that you avoid any statistical inference or other advanced techniques. In particular, this means that you should confine your analysis to the sample of data you have, and avoid making statements about the population that the sample comes from.

The Proposal

With your group come up with a 1 - 2 page proposal about the questions that you intend to ask of the data. This should include:

- Initial plots, figures or tables.
- References to column names and the analysis that they may provide.
- Additional datasets that you plan on including in your analysis like the weather data. This means links, columns that you'll join on, etc.
- What you plan to cover in the final report and how you plan on organizing it.

The Report

The report will be 8+ pages (including appropriately sized figures) and will be a report on what you found out from the data. This should focus on telling stories and explaining the narrative of the exploration and challenges associated with that. The report should not include any code - all code should be included in a sub-folder in either plain python files or in jupyter notebooks.

For the report, any graph, table, or figure should be annotated with why it is included. This is really to enforce just slapping graphs in your report that have no meaning.

One thing to note, using github to share jupyter notebook is a **nightmare**. Be sure to create copies of notebooks when you make an edit and as code gets stable, you'll likely find it easier to just put it into a python file and import it.

Instructor Pre-Approved Datasets

- Titanic passenger data
- Political Ad Archive
- Global warming data (Berkeley Earth) - use the *time series section*
- Whale tracking data - use the *raw* tracking data
- Shipwreck database (NOAA wrecks and obstructions)
- US Gov't Web Logs
- Parks graffitti report (St. Paul)
- NYPD 7 Major Felony Incidents
- [The National UFO Reporting Center Online Database] (http://www.nuforc.org)
  - The UFO data set will be challenging but fun since the reports are not uniform and is split across multiple tabled that you may have to scrape.
- The airlines dataset that we explored in async and any subsets thereof (so you may choose to analyze several years for example). Obviously a repeat of the analysis that performed in async would not be appropriate :).

Potential Locations for Other Datasets

REMEMBER NONE OF THESE HAVE BEEN APPROVED, YOU HAVE TO TALK TO YOUR INSTRUCTOR TO GET THEM APPROVED.

- Historical Weather Data
- This can be a bit difficult to figure out what to download. In this discussion there seem to be some pointers on how to get at it. This is dataset is pre-approved as a **supplemental** dataset, but please come to me if your group would like this to be their focus. I imagine there are some good resources for how best to download this data across the web but I just haven't had the time to search for them.
- Data is Plural
- I got the pre-approved datasets primarily from here. Pretty extensive list.
- Awesome Public Datasets
- Awesome list that's worth saving for future reference. Be sure to actually download and explore the data before sending me the link because sometimes it's harder than it appears. There are some awesome datasets here like the NYC taxi dataset.
- Data.gov
- All gov't data, which is good!
- OpenDataCache.com
- Cache of the above link - can sometimes speed up downloads.
- Reddit/r/Datasets
- Community generated and definitely some potential for good datasets. Worth searching.

To get a dataset approved:

1. Send your section instructor a description of the data including a link to the exact dataset or a place where they can download it in one click (w/o registering or anything like that).
   - The data needs to be in a raw text file format like json or csv (or at least very easy to convert to that format).
   - Please report the size of the dataset or the subset you plan on looking at. This is **not** a big data project.

2. Show that you've done some preliminary research that there will be enough interesting questions to answer in the data. This will include column names, information about missing data