Project_2_description

March 31, 2021

1 W200 Introduction to Data Science Programming, UC Berkeley MIDS

2 Project 2

The final project will comprise an Exploratory Data Analysis of a dataset of your choosing using the numpy and pandas tools.

- Push your proposal to your team's GitHub repository by 11:59PM PST the day before class, **week 12**
- Your group will give a 10 minute presentation during your final class, **week 14**
- Push your report to your team's GitHub repositiory by 11:59PM PST the day **after** class, **week 14**.

Some past showcase projects are here: https://github.com/UCB-INFO-PYTHON/MIDS_python_showcase

2.1 Instructions

Team

- Form a team of 2 to 3 students. Record your team here.
- Create a new GitHub repository for your team in the UCB-INFO-PYTHON organization. Name this folder as Project2 plus each last name of the students in the group like: **Project2_Foster_Kleemann_Llop**; please make this repo PRIVATE and invite the Instructor team / the rest of your team members.
- You can choose your team members from among students in your section, or in exceptional cases from among students in other sections. (Please email **both** section instructors to get an OK)

Data Analysis

- This project is relatively 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 pre-approved. You may join several datasets together. For example, you might combine two datasets to answer questions like, "Is felony crime higher in NYC after 2 days of rain?"
- This 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.

Proposal (10%) With your team members, prepare a 1 to 2 page proposal about the questions that you intend to ask of the data.

The proposal should describe:

- Names of team members,
- Name of your team's GitHub repository,
- A primary dataset you intend to analyze,
- Initial plots, figures, or tables,
- Some of the variables (column names) you intend to explore and what kind of insights you expect to glean,
- Supplemental datasets, if any, to complement your primary dataset - this means links, columns that you'll join on, etc.,
- What you plan to cover in the final report and how you plan to organize it.

In-Class Presentation (20%) The final class will include a 10 minute presentation per group with a 5 minute question/answer period. The presentation should include: - Your overall question - The steps your group took to analyze the question - Any assumptions you made in the analysis - The key is to aim for clarity and telling a story with the data - Organize your argument clearly - Guide the listeners through the evidence in the data - Include any key figures/plots/charts or graphs - You do not need to show any code but it might be handy to have the code ready in case there are some questions on it

Report (70%) Grading Breakdown: - 10% Questions - 20% Data Cleaning / Sanity Checks - 20% Compelling Text and Data Stories - 20% Compelling Figures

Your report should be 8-10 pages (including appropriately sized figures) and describe 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 - rather, all code should be included in a sub-folder in either plain python files or in jupyter notebooks. If your report contents are over 8-10 pages, please edit to include the most important material in the first 8-10 pages and any extra material in the Appendix.

For the report, any figure/plot/chart/graph should be annotated with descriptions about why they are included.

Note: Using GitHub to share jupyter notebooks has proven problematic in the past. Be sure to create copies of notebooks when you make an edit. As your code becomes increasingly stable, you may find it easier to just put it into a python file and import it into a jupyter notebook as needed.

Points for the report: - Your analysis is a written argument. - The key is to aim for clarity and exposition. - Organize your argument clearly - Guide the reader through the evidence in the data. - Proofread! - Every figure/plot/chart/graph in the report should be mentioned in your writing, explain what it means - If you don't have something to say about your figure/plot/chart/graph, don't display it at all. - No output dumps - Document decisions - If you decide that observations should be removed, state which ones. - If values are suspicious, but you leave them in, state

that too. - If you transform a variable, for example, by taking the logarithm, state that. - Your justification can often be very brief (just a sentence), but make sure the reader can follow your logic. - Characterize relationships between variables - Keep in mind the purpose of the analysis. - For example, if you're interested in explaining the price of a house, look to see what kind of relationship that variable has with the other variables. - Use only descriptive statistics - Descriptive statistics summarize a particular sample of data. - In the field of inferential statistics (which you'll learn in w203) you'll see how to create a model that represents the population (or process) from which the sample came from, and to make assertions about that population. - Since we haven't taught you any inference, please don't use it. - Beware of the word 'significant' - This has a technical meaning, implying that you've performed a statistical test

2.2 Instructor Pre-Approved Datasets

- Titanic passenger data
- Political Ad Archive
- Global warming data (Berkeley Earth) - use the *time series section*
- US Gov't Web Logs
- Parks graffitti report (St. Paul)
- The National UFO Reporting Center Online Database will be challenging but fun since the reports are not uniform and are split across multiple tables 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 performed in async would not be appropriate.

2.3 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 are some pointers on how to get at it. This dataset is pre-approved as a **supplemental** dataset, but see your instructor if you would like this to be the primary dataset.
- Data is Plural
  - The pre-approved datasets come 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 proposing one to your instructor. There are some awesome datasets here like the NYC taxi dataset.
- Data.gov
  - All government data, which is good!
- Reddit/r/Datasets
  - Community-generated and definitely some potential for good datasets.

To get a dataset approved ...

1. Send your instructor a description of the data, including a link to the exact dataset or a place from where it can be downloaded with one click without 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).
- 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 and that there will be enough interesting questions to ask of the data. This should include column names and information about any missing data.