Print to PDF

Project 3

Contents

- Objective
- Intro
- Process
- Prompt
- Analysis requirements

General Project information

Objective

Prove or disprove a hypothesis using skills learned in this class.

Intro

You can think of this as similar to the Project 2 requirements, but expanded. Examples of Final Projects for Python for Public Policy - the result of this Project will be similar.

Process

1. Find a dataset that seems interesting.
    - To meet the requirement that your project "not be trivial," you probably want a dataset that's large enough that you can't understand it at a glance.
        - If you're only using one dataset, you probably want it to have 500+ rows.
2. Load the data into a DataFrame.
3. Inspect the data a bit.
4. Fill out the prompt (below).
    - Work backwards: On a piece of paper / whiteboard, draw the visualization you imagine producing.
5. Use the data to answer the question.
6. If you end up answering your initial research question easily (haven't met the requirements), that's fine. Ask and answer follow-up question(s).
    - Go deep, not broad.

Prompt

Put the following in a Markdown cell in your notebook and fill it out:

- **Dataset(s) to be used:** [link]
- **Analysis question:** [question]
- **Columns that will (likely) be used:**
    - [Column 1]

    - [Column 2]
    - [etc]
- (If you're using multiple datasets) **Columns to be used to merge/join them:**
    - [Dataset 1] [column]
    - [Dataset 2] [column]
- **Hypothesis**: [hypothesis]

Raw Markdown

```
- **Dataset(s) to be used:** [link]
- **Analysis question:** [question]
- **Columns that will (likely) be used:**
  - [Column 1]
  - [Column 2]
  - [etc]
- (If you're using multiple datasets) **Columns to be used to merge/join them:**
  - [Dataset 1] [column]
  - [Dataset 2] [column]
- **Hypothesis**: [hypothesis]
```

The question should be:

- Specific
- Objectively answerable through the data available
- *Just the right amount* of ambitious (non-trivial)

If you want help/feedback, don't hesitate to ask on Ed or come to office hours.

Tips

- Don't overthink it; getting up through filling out the prompt shouldn't take more than a few hours.
- Your question/hypothesis doesn't need to be something novel; confirming something you read / heard about is fine.
- The point of the prompt is to ensure you've dug into the data and that your project scope is reasonable. Think of it as a guide rather than something you're locked into.
- Even the question can bake in assumptions.
    - Example: "What ZIP code has the highest number of food poisoning cases?" assumes a relationship between food-borne illness and geography.
    - What assumptions does your question make?

Analysis requirements

Your submission must:

- **Meet the general Project information**
- **Not be trivial** (or -30 points) - requiring:
    - At least 40 lines of code to come to a conclusion
        - That code should be relevant to answering your question. In other words, having 40 lines of `print("hello world")` wouldn't count.
        - If you meet all the other requirements, you will likely be well over this number.
        - You can count them automatically using a tool like tokei.
    - Operations that aren't easily done in a spreadsheet.

- **Transform data through grouping, merging, and/or reshaping** of DataFrames (or -15 points)
- **Have a visualization** (chart or map) of some kind (or -5 points)
    - General requirements
- **Have the prompt filled out**

Submit.