CMSC 201

Section 60

Fall 2024

Project #3: Data Analysis using Jupyter Notebooks

Project Value:

100 points

Due Date:

Completed Project: Monday, December 9, before midnight

Overview:

The purpose of this project is to gain expertise in data analysis and representation using common tools Jupyter Notebooks and a plotting library, matplotlip.

One of the most common and powerful uses of Python is to do active research, including analyzing data and presenting the results of the analysis in easy-to understand graphics. Python provides a number of tools that make this easy.

When you have completed this project, you will demonstrate sufficient mastery of these tools to use them in your future work in non-Computer Science courses and contexts..

Design and constraints:

You must implement this project using a Jupyter Notebook. Other than that, you are given many fewer constraints on the design of this project than of project 1, and thus you have much more freedom to express your own coding style. Be wary, though, because that gives you the freedom to do things wrong. Hopefully we've taught you better than that!

You must use proper design and modularity. Implement a function where you need it. If you do something multiple times, it should probably be a function. If you do something that's separate from the rest of the code, you should probably isolate it by putting it in a function. If you need to interact with the user or the system in some way, in different parts of the program, you should probably implement a helper function to take care of that.

You must also provide appropriate markdown cells that describe what the following code cells will do, and where appropriate, describe what the analysis from the preceding cell shows.

Data:

You will be analyzing the results of Presidential elections in the United States from 1976 to 2020, which was the most recent Presidential election. The original source of the data is MIT's Election Lab; the data file was downloaded from https://electionlab.mit.edu/data

To simplify the project, you will not need to download the data yourself. Prof. Arsenault has downloaded the data and placed it in the class GitHub repo. The data exists in two formats: comma-separated value (.csv) and tab-separated value (.tab). YOU ONLY NEED TO WORK WITH ONE OF THESE FILES: YOU DO NOT NEED BOTH. The files contain the identical information. You can choose the format with which you feel the most comfortable.

Assignment:

Step 1:

Download Jupyter Notebook to your laptop. In your favorite browser, go to https://jupyter.org/ Scroll down the page to the Jupyter Notebook section; do NOT choose "Jupyter Lab."

In the Jupyter Notebook section, click "Install the Notebook." It will take you to an instruction page.

On that page, it will give you the commands to type. If you are on a Windows machine, open a Powershell window and type

```
pip install notebook
```

If you are on a Mac, open a terminal window and type

pip install notebook

“pip” stands for the “Package Installer for Python” and it is the most common way new software can be installed and run on your Python environment.

If you get an error saying that “pip cannot be found” or similar, contact Prof. Arsenault.

If you cannot or do not want to install the Jupyter Notebook app onto your computer, that is fine. You will run the notebook in a browser tab. In a browser tab, go to https://jupyter.org/try-jupyter/retro/notebooks/?path=notebooks/Intro.ipynb. In the top left of hte page, click “file” and then “new.” Select “notebook” from the menu, and a new Jupyter Notebook will open in your browser. Select “Python” as the kernel.

Step 2:

Download the project 3 data file from the class Github repo to your laptop. Save it in a directory in which you can read it.

Step 3:

Open a new Jupyter Notebook. Call it “project_3.ipynb”

Step 4:

Create a new markdown cell. In that cell, write:

- Your name
- Your section (61 or 62)
- The date
- The semester: Fall 2024
- The project name: Project 3
- A summary of this project

Step 5:

Create a code cell. In that code cell, write Python code to read in your chosen data file. Convert the data file into a 2-dimensional list of words. Print out the first 20 lines of that list to ensure that this cell works. Once it does, move on to Step 6.

Step 6:

Create a markdown cell to describe what you're doing in this step.

Then create a code cell. In the code cell, create a new, 2D list that is a subset of the list you created in Step 5. The new list should ONLY contain the following.

- A row for the DEMOCRAT candidate in each year, for each state; and
- A row for the REPUBLICAN candidate in each year, for each state

Each row in the new list should contain ONLY these columns from the original list:

- The year
- The state name
- The party
- The votes that candidate received; and
- The total votes cast in that state

Print out the first 20 lines of this new list to show that the code cell works properly. Once it does, move on to Step 7.

Step 7

Start with another new markdown cell to describe what's happening.

Then create another code cell. In this cell:

- For each state, count the number of times the DEMOCRAT candidate beat the REPUBLICAN candidate in that state. For example, find the two rows for 1976 for ALABAMA. The DEMOCRAT vote count is higher than the REPUBLICAN vote count, so that's one win for the DEMOCRAT. For every other year in the database, you will find that the REPUBLICAN got more votes in ALABAMA. So you wind up with 1 win for DEMOCRAT, 11 wins for REPUBLICAN
- Do this for each of the 50 states.
- Print out the results - number of wins for each party - for the following states:
    - ALABAMA
    - CALIFORNIA
    - INDIANA
    - MARYLAND
    - VIRGINIA

Once you have completed that, you will move on to Step 8.

Step 8:

Now it is time to start creating graphics to show your results. Create a markdown cell to describe what you are going to do. Then create a code cell:

- Import the Pyplot library from matplotlip
- Create a list of the number of times the DEMOCRAT candidate won each state. For example, the DEMOCRAT won in ALABAMA once, so the first entry in the list is 1. For Arizona the DEMOCRAT candidate won 2 times, so the next entry in the list is 2. Create the list for every state.
- Now use the pyplot library to create a histogram of this list. This histogram will represent the number of states in which the DEMOCRAT won once; the number of states in which the DEMOCRAT won twice, and so on up to the number of states in which the DEMOCRAT won all 12 times.

When you have this histogram complete, move on to Step 9.

Step 9

Create a markdown cell to explain what you're doing. Then create a code cell.

In the code cell, create a list showing the number of states the DEMOCRAT candidate won in each election. This list should have 12 elements in it when you are done, one for each of the 12 elections in the dataset. Each element should be the number of states won by the DEMOCRAT in that election.

Use the Pyplot library to draw a scatterplot of this list. There should be 12 dots in your scatterplot.

Step 10

Create a markdown cell to describe what you're doing. Then create a code cell. For the state of MARYLAND, calculate the percentage of votes the DEMOCRAT got in each election in the dataset. That is, take the votes for the DEMOCRAT, divided by the total number of votes cast in the state.

Use matplotlib to plot a line drawing showing these results. Make sure to label the X and Y axes of this plot. Then move on to Step 10.

Submitting:

The submission process for this file is as follows:

When you have your project working, upload it to gl.umbc.edu. Your Jupyter notebook will NOT run on gl because gl doesn't have the necessary graphical user interface. But you will submit on gl because that's the easiest way to do this.

After you have submitted, let Prof. Arsenault know that you are done. Prof. Arsenault will fetch all the .ipynb files from gl.umbc.edu and grade the assignments on a separate machine that exists for this purpose.

The submit command is:

```
submit cmsc201 PROJECT3 project_3.ipynb
```

Note that the file you submit will NOT have a .py extension; it will be a .ipynb extension.