Data Engineering Homework 2

The following exercises will be graded. The due date is Saturday June 1, by 17:00.

After you attempt the exercises yourself (using Google search as you see fit), if you continue to have problems or questions, I highly encourage you to first consult with your peers and then with me or other instructors (either email or gather questions and bring them to office hours). You may also use GenAI technology, however, it is your responsibility to ensure that you have received the correct answer from the GenAI tool and more importantly that you learn the material the homework is designed to help you practice.

Code Overview

1. Download the DataEngineering-II-NumPy.ipynb file located in the path ~\Dropbox\DataEngineering-Class\Python file to your computer (copy paste from Dropbox to a local folder), then load it to your Google drive and finally open it with Google Colab. This file includes the Python code we went over together during class. Go over all the code again (i.e., run all the cells one by one, while reading the comments) to familiarize yourself with NumPy basic syntax.
2. Find a YouTube tutorial video on Pandas that you like. In addition to the YouTube channels I have already suggested in the DataEngineering-I-IntroToPython.ipynb and DataEngineering-II-NumPy.ipynb you can also check Rob Mulla's channel (see, for example, here). You must then replicate the code of the video i.e., create your own ipynb file with the video's commands and with appropriate comments (use the right verbs). You will need to email me your ipynb file with your name in the title as: PandasTutorial_Stefanos_Kechagias.ipynb. I will run the ipynb file In Google Colab and I will grade it based on how well (I believe) a novice Python student can learn Pandas from it. Again, if you don't like the channels or videos I have suggested you can select your own resource.

Data Engineering

The Paycheck Protection Program (PPP) is a $1 trillion business loan program established by the US federal government in 2020 to help business and sole proprietors continue paying their workers. For this assignment assume that you are an analyst for a US government bureau, and you are asked to create a report on financial aid provided to US business during COVID. Ideally, the report must focus on all available geography levels (country, state, etc.) and if possible, provide insights on demographics, industries etc. Prior to the final report, your manager has asked for some intermediate deliverables:

1. Download all the PPP data from the US Small Business Administration website (link).
2. Download all information on job postings from the Opportunity Insights (here).
3. Download the Latest Population estimates of US Counties from US Census Bureau
4. Create a folder called CovidRecovery where you will save all the relevant files for this project.
5. Save all the downloaded files in one folder called RawData.
6. Load the data into memory using Python or R (Python is preferable but if you do not yet feel comfortable with it then feel free to use R).
7. Perform EDA and clean the data as you see fit, given the time that is available to you.

8. Merge the information by geography and time so the data can be used in further analysis. All cleaned data should be saved into a folder called output and all code should be saved into a folder called code. All code must be well-commented.
9. Create a csv file with two columns: the State Name and the "Total Loan Amount per 100k residents". Save it in a folder called CSV. Although, the manager hasn't asked you for a map that shows this information, anytime they have seen one in the past they loved it!
10. Your code should be organized into the following sections: Preamble, Load Data, Clean Data, Analyze Data, Output Data and should have detailed an clear instructions for me to replicate it.
11. Connect the CovidRecovery folder to a GitHub Repository.
12. Anytime you face a difficulty during this project document it is in a ppt file.

Notes:

1. In the past your manager has asked you to perform tasks that do not always make complete sense. However, in instances where you showed up at the follow up meetings empty-handed, instead of admitting their erroneous "ask" your manager blamed you for not taking initiative by adapting and working on something similar to what they asked that would in fact make sense. In that spirit, if you get stuck on an issue for a long period or if you do not have enough time to find the exact answer, or if a deliverable is not properly/adequately defined, you are expected to think of workaround solutions or to make assumptions that allow you to proceed and present progress.
2. The long-term goal of your unit is to understand how financial aid has helped the US geographies/industries/demographics and perhaps where economic policy should focus in the future. Naturally, you expect that in the following week your manager will give you additional work to help achieve this goal. Seeing how much your manager appreciates (and expects) initiative, if you complete the assigned work of this week and have additional free time, it would be great to have prepare some ideas to share with your manager during your next meeting.