Individual Assignment I
DNSC 6305 Sections 80/81
Fall 2022

Write one Jupyter notebook with your solutions to all of the following problems. Document each step of your process in a reproducible manner, including downloads and other file changes. When you are finished, your instructor and anyone else should be able to run your notebook from start to finish without errors.

Add text notes on your process as appropriate, documenting any assumptions and explaining key decisions you make along the way. Use markdown cells for this text, formatting your notes so they are easy to read.

Be sure to answer each question directly and precisely, using the data to justify your answers, and showing all of your work along the way.

This is an individual assignment. As I explained during first lecture, you are welcome to seek and give assistance to others who might become stuck along the way. Please acknowledge any assistance you receive. At the same time, each student must perform and submit her/his own work, in accordance with the GWU Code of Academic Integrity.

This assignment is due on Monday, Sep 19, at 4 pm. Please name your Jupyter file according to the following format:

DNSC6305_Assignment1_firstname_lastname.ipynb

Problem 1 - Word counts (30 points)

Solve parts A and B

Part A. Characters in The Hound of the Baskervilles (15 points)

Use the text available at https://s3.amazonaws.com/dmfa-2020/project-1/hound.txt for this part.

How many times are each of the following characters mentioned by name in the text of The Hound of the Baskervilles?

Holmes, Watson, Barrymore, Mortimer

Hint: Be careful where the names matches both the exact name and a plural relating to a family name. For example, Watsons vs. Watson. What do you observe?

Part B. Characters in Hamlet (15 points)

Use the text available at https://s3.amazonaws.com/dmfa-2020/project-1/hamlet.txt for this part.

How many times do each of the following characters have speaking lines in Hamlet? Keep in mind that this is the text of a play.

Hamlet, Polonius, Ophelia, and Horatio

A speaking line is usually start with an Upper-case name with a trailing ".”
For example,
HAMLET. Not so, my lord, I am too much in the Sun

Problem 2 - Capital Bikeshare (40 points)
Use the data available at https://s3.amazonaws.com/dmfa-2020/project-1/2018-capitalbikeshare-tripdata.zip for this problem.

Part A (20 points)
1. Unzip the data and combine the two inflating CSV files using csvstack. Name your combined file “biketrip.csv”. Compare the total number of lines in each file and in the new combined file using wc command. Comment on your findings.

2. List the labels for the heading line.

3. Which 10 Capital Bikeshare stations were the most popular departing stations in July and August 2018 in terms of number of rides? provide the full name of the station, not just the station number.

4. Which 10 stations were the most popular destination stations in July and August 2018 in terms of number of rides? provide the full name of the station, not just the station number.

5. Which 10 station-pairs (Departing-Destination) are most popular in July and August 2018 in terms of number of rides? provide the full name of the station, not just the station number.

6. Comments on your finding for Q3, Q4 and Q5.

Part B (20 points)
1. For the most popular departure station, which 12 bikes were used most in trips departing from there? provide the full name of the station, not just the station number.

2. Which 12 bikes were used most in trips ending at the most popular destination station? provide the full name of the station, not just the station number.

Problem 3 – advanced csvkit command (30 points)

For this problem use the following two files
https://www.fec.gov/files/bulk-downloads/data_dictionaries/cm_header_file.csv
https://www.fec.gov/files/bulk-downloads/2022/cm22.zip

This data contains basic information for each committee registered with the US Federal Election Commission. It supposed to have one record per committee. Description of the columns representing this dataset are provided here

The first file above represents the header and contains the labels for all the columns in the second file. The second file represents the committee data. The first file is a csv file, the second file is "|" delimited text file.

Questions

1. Upload the two files into your root directory, combine the first file (header) with the unzipped version of the second file (data) and name the combined file as "candidates.csv" (7 points)

   Hint: you may need to reformat your second file from "|" delimited to csv before you combine it with the first file. You can use csvformat command.

2. Provide a list of the label heading for the combined file. Display the first 7 columns and 15 rows of the combined file (3 points)

3. Provide in a table a list of all Republican committees with an office in NY. The list should contain the Committee name, Committee city or town, and Connected organization's name. Sort the finding by Committee name. (10 points)

4. Provide the names of the committees that appear more than two times in this dataset? If you find any investigate whether these records are duplicates and suggest an action (e.g. remove) (10 points).