Assignment 2

DSCI 222 – Data Science Workflow using Python
School of Mathematical and Data Sciences
West Virginia University

Instructions

This is a homework assignment to be solved in a small group (3-4). It is a great opportunity to show your teamwork skills. Enjoy it! All material needed for the assignment can be found in Github

- Even though you will be working in groups, each student should submit their own, albeit identical, set of deliverables. All deliverables must be in a folder created in your Google Drive account. The folder name should follow this format: `DS_X_LastName_FirstName`, where **X** represents the assignment number.

  Files included in your deliverables folder

  - Deliverable 1 Video Recording
  - Deliverable 2 Colab Notebook
  - Deliverable 3 Report
  - Anything additional or supplementary that contributed to the assignment (this includes .csv files, .json files, etc.)

- Include your **full name** as a Python comment at the top of the notebook and at the top of your report.

- All data manipulation should occur within the Python script. No manipulation of the supporting file(s) should occur prior to importing the file(s) into your script. You may check your work by manually performing data analysis.

- Include your report in PDF format, written in LaTeX.

- Set share folder permissions so chood@mix.wvu.edu can access and run every file and notebook.

- Everything counts! Include as much as you want in your deliverables, even if the activity is not fully complete by the deadline. *Important:* Review the grading policy and course policies in the online syllabus.

- **Total: 100 points.**

Activity 1: Video Recording (50 points)

Task

Create a video tutorial on **"How to create a US Population Choropleth Map"**

In class, we have learned about ways to use python to visualize data. With this assignment, we will be expanding on this by learning how to create choropleth map that can overlay a map of the United States. You will need to learn about geopandas, a Python library for working with geospatial data. It extends the popular pandas library to make it easier to handle geometric data types.

You will be also be using a JSON file containing state geometric data. A JSON file is a plain text file that stores data using the JavaScript Object Notation (JSON) format. It's one of the most common ways to exchange data between applications, APIs, and programming languages. Further knowledge of JSON files is not necessary for this assignment.

Be sure to also compare and contrast GeoDataFrames and DataFrames somewhere in your recording.

Requirements

1. In the Assignment 2 folder in Github, you will find three files `choropleth_map.ipynb`, `population_data.csv`, and `us-states.json`.
2. Create a video explaining the .ipynb file.
3. Your video should clearly explain what each line of code is doing in its entirety. It is expected for every aspect of the script to be addressed and nothing should be ignored.
4. You may choose which of the two choropleths you would like to explain. You do not need to explain the code of both.
5. Be sure to address what a GeoDataFrame is and how it differs from a pandas DataFrame (this may require some research on your part).
6. Your video must be a screen recording of the code as it is being explained, to help follow along. You may include yourself in the video as well, but you do not have to.
7. You may add code and text to the .ipynb file to aid in your explanation if you would like.
8. All members of the group must speak in the video

Deliverable 1:

Video recording

Activity 2: Python Source Code (50 points)

Scenario

Imagine you are a data scientist working for an organization or government advisory office of your choice. You have been tasked with gathering and analyzing data to persuade a political leader to take action on an important issue. This issue could relate to public health, education funding, environmental policy, transportation, or economic development.

Deliverable 2:

Google Colab Notebook

- Select a topic and collect relevant data.
- Create a choropleth map that clearly shows patterns, trends, or disparities in your data. Your data can represent the U.S., a different country, or the world.
- Perform some sort of transformation to the raw data. This means your choropleth map should not be a direct representation of the data you collect. You should manipulate your data to make your argument more convincing. (Ethical considerations aside, this is a common practice in data storytelling.)
- Make sure your choropleth map includes attributes that make it easy and quick to interpret for the viewer.
- Include in the notebook printouts of the raw data, the data after it has been cleaned, and the data after transformation.
- Before each block of code, include text explaining how that block works.
- Within each block of code, include comments explaining how each component of the code works and why it was used.
- Any libraries, functions, methods, etc., used that are not explicitly discussed in class must be clearly explained.

Deliverable 3:

Write a 2-3 paragraph persuasive story by explaining your findings to the political figure of your choosing in a way that helps them understand the urgency or importance of your topic. Highlight key states or trends that support your argument, and use your visualizations to strengthen your message.