DATA203 Foundational Python (Prof. Maull) / Fall 2024 / HW4

Points Possible	Due Date	Time Commitment (estimated)
30	Sunday, December 01	*up to* 30 hours

- **GRADING:** Grading will be aligned with the completeness of the objectives.
- **INDEPENDENT WORK:** Copying, cheating, plagiarism and academic dishonesty *are not tolerated* by University or course policy. Please see the syllabus for the full departmental and University statement on the academic code of honor.

OBJECTIVES

- Use Pandas for data engineering of AQ data
- Use Pandas for API data extraction of AQ data
- Understand background knowledge to understand function implementation
- Understand data merging and analysis using Pandas
- Explore data plots of AQ data using Pandas

WHAT TO TURN IN

You are being encouraged to turn the assignment in using the provided Jupyter Notebook. To do so, make a directory in your Lab environment called `homework/hw1`. Put all of your files in that directory. Then zip or tar that directory, rename it with your name as the first part of the filename (e.g. `maull_hw1_files.zip`, `maull_hw1_files.tar.gz`), then download it to your local machine, then upload the `.zip` to Canvas.

If you do not know how to do this, please ask, or visit one of the many tutorials out there on the basics of using zip in Linux.

If you choose not to use the provided notebook, you will still need to turn in a `.ipynb` Jupyter Notebook and corresponding files according to the instructions in this homework.

ASSIGNMENT TASKS

(25%) Use Pandas for data engineering of AQ data

In the previous assignment, we learned that the EJ Screen tool provided a wealth of data that we could use to both visualize and analyze data from the EPA.

We also learned that there were other sources of data, namely data from citizen science sources like Purple Air and also from historical sources like the Internet Archive.

In the prior assignment, you obtained a sense of the work involved in getting this data. Luckily, this can be automated with a little effort.

Instead of having you do that automation, you will be able to see it in action if you wish. There are two notebooks titled `pre_nb001` and `pre_nb002`, which contain the code to get data from Internet Archive Common Crawl files, and from Purple Air. You will notice the Purple Air data extraction uses a machine API (application programmer interface) which makes getting data exceedingly easy in Python and other languages. The data from Common Crawl is not as easy to obtain – it requires a bit more complexity by requiring the archive pages to be parsed from the HTML source. This technique is often *required* when there is no API, furthermore, since the LDEQ data forecasts are not archived anywhere, we must rely on what we can get from the IA/Common Crawl files.

The fact that these files exist and can be easily accessed by anyone with a decent Internet connections is critical and important – you can contemplate the profundity of this reality.

§ Task: Study the `pre_nb001` and `pre_nb002` notebooks. You will not need to do anything with them, not will you have to run them, but know that a working data scientist has some knowledge about how this process works (when using a tool to perform these tasks), if not a basic knowledge of how to do it manually when called to do so.

§ Task: Please open `hw4_001_extract_explore_explain.ipynb` and follow the instructions and answer the questions *in* the notebook.

(25%) Use Pandas for API data extraction of AQ data

Often once data is extracted from their sources additional transformation needs to be done, for example to convert data, average it or perform other important pre-processing steps on the data.

The data is often in multiple files and needs to be combined for further processing.

You will learn a bit more about this process in the task below.

§ Task: Please open `hw4_002_api_explore_explain.ipynb` and follow the instructions and answer the questions *in* the notebook.

(15%) Understand background knowledge to understand function implementation

Often you will need to read the conversations of others to understand how code was implemented (or how to implement something yourself).

In this part you will look at another notebook and understand why some decisions were made in code that is being included in the existing base (I already wrote it, so you do no have to do so).

§ Task: Please open `hw4_003_transform_explore_explain.ipynb` and follow the instructions and answer the questions *in* the notebook.

(20%) Understand data merging and analysis using Pandas

After you have data that you need, the final work that needs to be done often requires final cutting, labeing and merging, with the output being your data of interest that you will analyze, plot, run statistical tests or even build models from (or get inspiration for building such models).

§ Task: Please open `hw4_004_merge_explain.ipynb` and follow the instructions and answer the questions *in* the notebook.

(15%) Explore data plots of AQ data using Pandas

The fun of analysis is in the pretty pictures. While partially fecetious, the reality is no one except the data geeks want to see the numbers – things come alive and stories get interesting when data is visualized.

In this part, you will explore a plot and answer a simple question about it.

§ Task: Please open `hw4_005_plot_explain.ipynb` and follow the instructions and answer the questions *in* the notebook.