DATA203 Foundational Python (Prof. Maull) / Fall 2024 / HW3

Points Possible	Due Date	Time Commitment (estimated)
25	Monday, November 11	*up to* 15 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

- Explore Pandas
- Explore Problem Background Information
- Explore Data Sources for Air Quality
- Obtain Data Sources using Python

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

(0%) Explore Pandas

The following tasks are not worth any points and are recommended but *not required*.

§ Task: Study the official documentation on subsetting Pandas data in Series and DataFrames.

Go to the following section in the Pandas documentation:

- How do I select a subset of a DataFrame?: Pandas official documentation → https://pandas.pydata.org/docs/getting_started/intro_tutorials/03_subset_data.html

§ Task: Study the official documentation on summary statistics in Pandas.

Go to the following section in the Pandas documentation:

- How to calculate summary statistics: Pandas official documentation → https://pandas.pydata.org/docs/getting_started/intro_tutorials/06_calculate_statistics.html

(70%) Explore Problem Background Information

Now that we have some new skills and are on our way to becoming full-blown Pythonistas, it is time to take a small breather and consider the life of a Data Scientist.

In this part of the assignment, we are going to assume that you are part of an analytics team that is interesting in discovering the relationships between datasets for air quality measurement.

Specifically, you are going to be diving in to the Mossville, Louisiana community, which has been identified as a community at risk for toxic pollutants in the air.

Of specific threat are *dioxins*, which are carcinigenic chemical compounds you will learn about in this assignment. Detection of these compounds is complex and expensive, but tools are getting better and more accessible as technology improves. However, inexpensive tools exist today to detect air particles which may contain dioxins, thus providing a mechanism to quantify the latent and potential threat of dioxin exposure (potential because not all air particles contain dioxins and dioxins exist in water and other substances, latent because the impact of these chemicals may be dose dependant over long time scales).

§ Task: Learn about the Mossville community.

One of the most *imperative* things you can do as a data scientist on a project, is to get background information about what you are analyzing. Not only is it helpful in painting the picture of the data, but it is also a guide for understanding how to conduct an analysis of the data – sometimes knowing which questions to prioritize come through this initial understanding.

Using the **web** (ChatGPT/LLMs are **not advised** due to poor training data specifically about places and the inability to understand model bias), **learn about** Mossville, LA and **answer the following questions**:

1. List 3-5 facts about Mossville, LA. You may (but it is not required) include historical information about the town.
2. Provide a basic demographic summary of the community – use the US Census from 2020.
   Your summary should minimally include: population size, income, race, gender and age stratification.
3. What are the major sources of income for the community?
4. Look at the satellite map of Mossville using your favorite map tool. List the 5 closest towns to Mossville.
5. What is your general thought about this community within the context of demographics?
   You might include in your answer, thoughts about the demographics of the state at large, or even surrounding communities you listed in (4).

§ Task: Learn about community health surveillance and dioxins in Mossville.

Using the CDC ATSDR (Agency for Toxic Substances Registry), explore the data which has been "officially" produced by the collaboration between federal, state and local agencies:

- ATSDR Overview of Mossville

You are encouraged to read as many of the documents here as you can.

Read the 2002 document ATSDR HEALTH CONSULTATION, and **answer the following questions**:

1. Please provide your summary of this document. Be brief in 3-5 sentences.

2. What is your opinion of the following statement?

   > "Mossville community health is at significant threat from air and waterborne toxins. Community health surveillance is a state and federal problem.".

   Use evidence from the document to support your opinion.

3. What is your opinion of the recommendations made in the document? Answer as if you or a family member lived directly in the community.

4. Imagine in 2002, you were a scientist on the evaluation and production team of this report. Imagine still, that in 2003, you were launched into space and put into a deep sleep for 20 years. Upon your return, you checked in on the progress that was made on those recommendations you tirelessly worked on. Provide evidence for whether **those recommendations were implemented**, and whether there is **evidence that there was any improvement in health outcomes** within the Mossville community. You will need to use "official" federal, local and state sources **only**. The reason for this is that in a legal capacity, these sources are the one's which will stand as the "highest" level of evidence (or lack thereof) in court.

§ Task: Learn more about the current state of the art in dioxin detection techniques considered by the EPA.

Read the interview: *A Step in the Right Direction for Dioxin Detection*, **March 25, 2021** by Karen Steward, PhD.

There are technical terms in the article which may be unfamiliar to you. We are not interesting in becoming overnight chemists, but we should still be able to read the article and glean the main points from it (an essential skill of any data scientist).

Answer the following questions:

1. What 2-3 things did you learn from this interview?

§ Task: Learn about PM2.5.

Read the entire resource and all sub-pages:

- Particulate Matter (PM) Pollution

There are no questions to answer for this part.

(20%) Explore Data Sources for Air Quality

Now that we have a little background information about the Mossville community, we are going to familiarize ourselves with data that we are going to eventually analyze in the **final course assignment**.

There are a variety of Federal sources for air quaity information, the most authoratative being the data produced by EPA monitoring stations around the country. Please study the following online sources:

- AirNow
- OpenAQ

There are further state sources of air quality data including the Louisiana Department of Environmental Quality:

- Louisiana Department of Environmental Quality

The LA/DEQ produces air quality forecasts which allow for the public to understand their *potential* threats for common pollutants found in PM2.5. Unfortunately, the old forecasts are not stored, but luckily, there are

- the Internet Archive, and
- CommonCrawl

If you do not already know about these services, now you know – and you would be advised to bookmark them as they preserve the precious history of the Internet (albeit incompletely) which as we know has been prone to alteration and removal. **Familiarize yourself with these services**.

There are "un-authoritative" sources of "community science" air quality data which you must know about. One of the most important of which is the PurpleAir air quality monitor. Study it well:

- PurpleAir

Finally, you will need to familiarize yourself with the EJScreen Tool, which is a data aggregator which provides essential connective tissue to the underlying raw data, demographics and other Census information for the US:

- EJ Screen Tool

§ Task: Obtain a Purple Air developer account.

Go to PurpleAir.com and obtain a free developer account. You may want to use your `howard.edu` email.

1. Take a screen shot of your account dashboard (logged in).

NOTE: DO NOT attached your API key in this answer in your assignment.

§ Task: Obtain the dates of the 2024 Internet Archive LA/DEQ Forecast page crawls.

1. What are the dates of the 2024 crawl data.
2. What was the Saturday Feb 24, 2004 PM25 forecast for Lake Charles? (HINT: Use the Internet Archive search)
3. How many crawls were done in 2024 (Jan-present)?

§ Task: Find the Purple Air sensor "CORE Futures 2 Quarkume" on the PurpleAir map.

1. Take a current screenshot of the current PM2.5 reading and insert that into the homework.
2. Take a current screenshot of the nearby "Westlake, LA" sensor.
3. Compare and contrast those two readings. What do you observe?

§ Task: Use the OpenAQ map to find the closest EPA Reference PM2.5 monitor to "CORE Futures 2 Quarkume".

1. What are the lat/lon coodinates of the sensor (e.g. 21.3456 N, 56.345 W)?
2. Which PurpleAir sensor is closer: the "Westlake, LA" or "CORE Futures 2 Quarkume"?

3. What is the next closest EPA monitor that produces PM2.5 data?

§ Task: Use the EJScreen Tool https://ejscreen.epa.gov/mapper/ to answer the questions below.

1. What (National) percentile is Mossville in using the Environmental Justice Index for Particulate Matter 2.5?
2. What (National) percentile is Mossville in using the Environmental Justice Index for Toxic Release to Air?

3. Compare Mossville to Lake Charles, Westlake and points *east* of Westlake. What do you observe? Provide screenshots to support your observations.

(10%) Obtain Data Sources using Python

As you saw during the lecture, data can be easily obtained from the web if there is a CSV (or JSON) at the end if the URL to `DataFrame.read_csv()`. The ability to grab data directly and use it in your Python code, greatly reduces the work required to begin analyzing data, often taking many hours away from the data engineering portion of your work, at the very least, it optimizes it when you know what to do.

A lot of data on the web is not sitting at then end of a neatly packaged CSV (though, indeed, a lot of it is!). Instead, however, there are machines whose sole purpose is to respond to data requests and return CSV or JSON data so that it can be consumed as part of a pipeline for analysis.

The EJ Screen Tool provides a URL for getting data directly into your programs – you may not have noticed it, but on the mapping tool page, once you create a shape, and click `Submit (GET)` the system is being invoked to return the data so it can be consumed by other programs.

These are called APIs or "Application Programming Interfaces". Believe it or not, APIs run the web.

You can learn more about the EJ Screen API and you will need to study the EJ Screen API documentation here:

- https://ejscreen.epa.gov/mapper/EJAPIinstructions.pdf

You will also need to learn about the *data dictionary* here:

- https://ejscreen.epa.gov/mapper/ejsoefielddesc1.html

§ Task: Go to Github and find the helper module for HW3 (`hw3_demo.ipynb`). Use it to answer the questions below.

You will need to download, study and execute the file to the same folder as your homework (or run on Jupyter Hub).

1. Explain what the `load_ejs_dataframe()` function does. Make sure to describe the inputs and outputs (return values).

2. Explain what the API DATA LOADING section of the notebook does.

3. Provide a 2-3 sentence reaction to the following statement:

   > *In the data provided, the life expectancy of Mossville is higher than Westlake or Lake Charles, but the number of air pollution facilities is much higher, but the raw PM2.5 values are very similar between all three locations.*

   You will need to look at the final cell of the demo notebook to answer the question. Use evidence from the final notebook cell in your answer.