Post Midterm Progress Report

Date: October 24, 2021

Accomplishments:

- Uploaded an EPA air quality sites csv file to the GitHub project repository. I joined this file to my datasets to obtain state names of state codes.
- Split the PM2.5 studies into two Jupyter notebook parts. The first part looks at the datasets for 1999 and 2012 and compares the results with Dr. Peng's study. The second part expands the study to a 2020 dataset. Cleaned 2012 and 2020 datasets for the second part. Removed excluded values under the events feature, removed all 1-hour samples, removed 24-hour blk average samples if they were present on the same site and on the same day as a 24-hour sample, removed observations from Mexico. (The 1999 dataset did not need cleaning.) Re-wrote and re-organized content within the Jupyter notebooks, to include the use of headings. Completed the PM2.5 study and visualizations for the questions: (1) How does the level of PM 2.5 in the U.S. compare between 1999, 2012, and 2020? (2) Which states have the highest and lowest levels of PM 2.5?
- Cleaned all datasets for the ozone studies. Removed all excluded values under the events feature. Removed all observations from Canada and Mexico. Completed the ozone study and visualizations for the questions: (1) How does the level of ozone in the U.S. compare between 1999, 2012, and 2020? (2) Which states have the highest and lowest levels of ozone?
- Started the AQI studies and visualizations. Completed initial analysis on four of the five target questions at the county level: (1) Which states have reports of hazardous air quality index (AQI)? (2) Which states have reports of very unhealthy AQI? (3) Which states have reports of unhealthy AQI? (4) Which states have reports of unhealthy for sensitive individuals AQI?
- Uploaded the revised PM2.5 Jupyter notebook file to the GitHub project repository.
- Uploaded the revised ozone Jupyter notebook file to the GitHub project repository.
- Uploaded the new PM2.5 part2 Jupyter notebook file to the GitHub project repository.
- Uploaded the new AQI studies Jupyter notebook file to the GitHub project repository.
- Uploaded the post midterm progress report to the GitHub project repository.
- Uploaded a CDC Chronic Disease Indicators: Asthma dataset to the GitHub project repository.

Current Activities: I am currently working on the AQI datasets and visualizations.

Challenges: Although I am combining values from different monitoring sites within a state into state values, I realize it is not as accurate as comparing data from the same site. Furthermore, some states are not represented at all. There are monitoring sites in U.S territories, and in Canada and Mexico. Cleaning data that I initially did not clean and then re-running the analyses on the cleaned data has been the most time consuming.

Work to be Completed: For the next project milestone, I will finish what was to be completed in this milestone: the AQI studies and visualizations and cleaning the asthma dataset. I will also finish the asthma exploratory analysis and visualizations.

Answers to Questions from Last Progress Report: You wrote: "I see the document reference on acceptable values. But I'm puzzled. If it's a concentration or quantity, then a negative value is not possible (unless it's relative to a baseline)."

The email response I received from the EPA was: "Thanks for your message and reaching out. We have a short write up on acceptable values in EPA's Air Quality System (AQS) here: https://aqs.epa.gov/aqsweb/documents/about_aqs_data.html#_acceptable_values, but since that explanation is very short I can elaborate.

Every instrument has an allowable uncertainty, and occasionally as you've noted monitors can yield small negative hourly values. Say it's +/- 10 ppb for whatever substance. If the instrument reads 100 ppb, that means the real concentration will be somewhere between 90 and 110. If the instrument reads -3 that means the real value can be anywhere between 0 and 7 (negative concentrations not being possible). We allow reporting of the negative values to capture valid, quality assured readings that are valid members of the sample set. With PM2.5 monitors, negative hourly concentrations for PM2.5 down to -4.99 ug/m3 (the default QC range check) are used in computing 24-hour averages so as not to bias that average."

I had included the email excerpt in the Jupyter Notebook PM2.5 studies, but I will include it in the final report, as well.