

Progress Report 2

Date: October 10, 2021

Accomplishments:

- Updated GitHub project ReadMe File and began tracking local Jupyter notebook files to push to the remote repository for version control.
- Began using markdown for Jupyter notebook textual content in cells rather than the default code format.
- Read in the 2020 PM2.5 daily summary csv file, extracted the arithmetic mean column, and printed a summary in a Jupyter notebook file.
- Delved deeper into negative PM2.5 values by contacting the EPA via email.
- Uploaded the revised PM2.5 Jupyter notebook file to the GitHub project repository.
- Read in the 1999, 2012, and 2020 ozone daily summary csv files, extracted the first maximum value columns, and printed brief summaries in a Jupyter notebook file.
- Uploaded the new ozone Jupyter notebook file to the GitHub project repository
- Uploaded the progress report 2 to the GitHub project repository.

Current Activities: I am currently working on the PM2.5 and ozone AQ datasets and visualizations.

Challenges: Deciding how to move forward when I was unable to re-create similar boxplot visualizations to Dr. Peng's PM2.5 studies has been challenging. I looked more closely at all the datasets because of the negative values in the 2012 and 2020 datasets. The 1999 csv file used 24-hour sample durations only and had no negative values. In the 2012 and 2020 datasets, some sites reported two or three different values on a given day: 1-hour samples, 24-hour block average samples, and 24-hour samples. Re-reading the EPA's description of the data, multiple records for a monitor may be present if the sites report them. All negative values were either 1-hour samples or 24-hour block average samples. Due to correspondence from the EPA and their listed acceptable values at https://aqs.epa.gov/aqsweb/documents/about_aqs_data.html#_acceptable_values, I will not remove the negative values from the datasets as I was initially planning to do. I am planning to remove observations when sites report more than 1 value on a given day. How to prioritize which values to use and how to accomplish removal of multiple records for a monitor using Python has been the most time consuming.

Trying to reproduce visualizations using different programming languages is more complicated than I thought it would be. Deciding how best to visualize the data is also time consuming.

Work to be Completed: For the next project milestone, I will finish what was to be completed in this milestone: the PM2.5 studies and visualizations, the ozone visualizations, and the AQI studies and visualizations. I will also begin the asthma portion of the project by locating and cleaning the dataset.