Kyle Gustke
Todd Graham

Midpoint Report

CS 410: Explorations in Data Science

Objective:

Researching and visualizing the relationship between air pollution levels and COVID-19 cases around the world.

Approach:

We will begin by researching the connection between air pollution levels and COVID-19 cases and possible metrics for quantitative analysis. Once the metrics for analysis are identified, we will acquire and process the data. We will then analyze the data to determine the best possible way to visualize the connection between our topics. This visualization may be done using both Python and Tableau. In the end, we will evaluate the connection between air pollution levels and COVID-19 cases based on our analysis and visualizations.

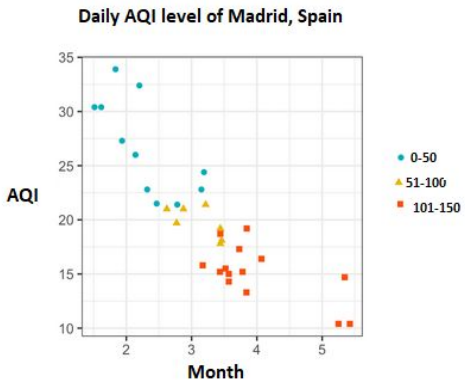
Team Structure:

For each milestone our team will discuss how to split the work appropriately to match both team member's strengths and interests. For example, during the processing of data milestones we will have two distinct data sets (Air Pollution and COVID-19) so each team member will be responsible for cleaning one data set. For milestones without definitive demarcations, we will work together to complete the task.

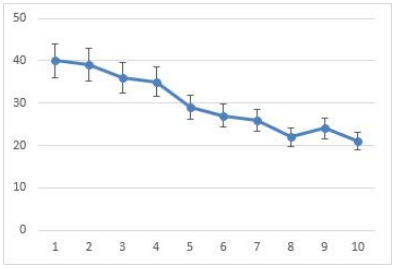
Analysis and Visualization Plan:

These tasks are ordered in precedence. We believe accomplishing tasks 1 & 2 would create a valid representation of our data while the subsequent tasks would be supplementary. Tasks 3 & 4 would be used as comparisons to the references that we discussed in our Research Report Paper.

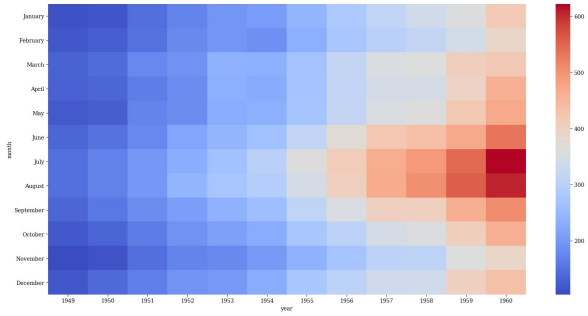
1. 2020 Lockdown to Present (China, India, USA, Spain, Italy)
    a. Scatter plot of daily metrics for chosen cities (color coded dependent on AQI rating [see Research Report Paper for color references])



b. Line plot with week-to-week standard deviation



c. STRETCH GOAL heat map for AQI throughout a time period



2. Date matched comparison of 2019 of $PM_{2.5}$ values to 2020 values (China, India, USA, Spain, Italy)
    a. Scatter plot of daily metrics for chosen cities
    b. Line plot with week-to-week standard deviation
3. Date matched comparison of 2015-2018 of $PM_{2.5}$ values to 2019 and 2020 values (China, India, USA, Spain, Italy)
    a. Scatter plot of daily metrics for chosen cities
    b. Line plot with week-to-week standard deviation

4. Expand visualizations from Tasks 1-3 beyond selected countries to cover full range of countries in the database
5. Import the data into Tableau, overlay the $PM_{2.5}$ concentrations over a map to display the data in a time-series plot.

Project Milestones:

1. Research and gather data
    a. Air quality data was collected for 2015 - Present with global values and daily timepoints. Specific data includes temperature, humidity, PM10, PM2.5, CO, Ozone and NO2. We plan to focus on PM2.5 due to its health impacts.
2. Process data for specific use case
    a. A parsing tool has been created in Python using pandas. Data for Analysis and Visualization task 1 has been parsed and is ready for visualization.
3. Complete Midpoint Report
    a. This report is now completed!
4. Analyze data
    a. See above for analyzation/ visualization plan
5. Visualize and model data
    a. Matplotlib, plotly, seaborn
    b. Tableau
6. Assess validity of models/ results
    a. Compare tasks 1 and 2 to the analysis/visualization plan to references used in the report.
7. Final project presentation
    a. Make a powerpoint.

Scheduled Midpoint Meeting:

Tuesday 7/21 at 1:00pm