Abhishek Gupta
Harsh Sugandh
Michael Morrison
Priyanshu Tripathi
Vrishti Jain

Data Science

Assignment 4 Report (Team 10)

December 17th, 2020

1. Investigation/Goal

Choose an investigation and identify pre-existing source of data that can address a particular data science goal (7%)

a) Choose, and state, the goal and reasons why the datasets were chosen and how they were found and managed, Min 3-4 sentences.

Our goal is to verify if there is correlation between Covid-19 data and the air quality Index of an area and analyze the trend if the correlation exists, specifically in some counties of New York State.

To capture the Covid-19 data, We collected the data from the New York State Govt NY website. For air pollutant data we used the Air Quality Index Daily Values Report from the the United States Environmental Protection Agency website.

The reason for choosing these datasets is that it is a comprehensive collection in both cases, with varied columns about cobid-19 testing and results as well as each pollutant has a specific data sheet with various columns of AQI, latitude, longitude of the county and so on.
Also, this data is as authentic for New York state as possible, since both these belong to Government agencies.

Since, we wanted to get these authentic datasets, we searched upon various Github repositories and blogs to locate and capture them. The dataset files are downloaded in .csv and stored on Google Drive for a shared medium as well on Github for version control, analysis and project development.

b) Document and discuss the data formats and any metadata standards/ conventions in use, and the method(s) of discovery and access and how they helped or hindered the process, Min 3-4 sentences.

SInce both these datasets have been captured from the Government websites, the data formats and conventions have been imbibed without changes in our research process.
The date format is the US standard of mm/dd/yyyy, and other descriptive information about the data and its format conventions have been uploaded as files and links to the readme.md section of the Github repository.

The metadata was easily discoverable on both the websites, as it gets updated and maintained regularly. In the case of Air Data, the link to the original data formats saved on the .html file is always the revised and updated version. Whereas, in case of the NY State Health data, the format and description in the form of data dictionary remains as the .pdf files embedded on their website.

The process of using the data in given formats for our analysis was actually easier because the formats and conventions are ready to be adopted and worked upon for analysis through python libraries. We could focus on the outliers and the columns required for the data analysis more because the downloaded data was clean and workable.

2. Data Analysis

a) Develop and state two particular questions/hypotheses related to the goal of the investigation and that can be answered using the datasets under consideration. Design an analysis study (preliminary, full and post) to answer these questions and document the analysis design, Min 3-4 sentences (3%)

Our hypothesis is that there is some type of relation between pollutants in the air before and after the lockdown took place in NYS. So in our first hypothesis we aim to verify this using the pollutants dataset and the covid dataset for the NYS. In our second hypothesis we aim to find out how the concentration of pollutants varied this year compared to the previous years?. Along with this we wanted to see if the size and population of the county had an effect on the concentration of pollutants or not?

We started off by exploring our covid dataset for new positives in various counties in NYS. We looked into the general distribution of the data to start off. We saw that in general, there was an increasing trend in the number of new positives all throughout the year of 2020. We then performed our exploratory analysis on the pollutants dataset where we explored the data for all the pollutants namely CO, NO2, Ozone, PM2, and SO2 (refer Fig 1-4). After exploring each of the two datasets individually, we merged these two datasets using the basis of counties and pollutants present in that county. We joined our two datasets using an inner join to get a dataframe.

After merging and performing initial preprocessing on our data, we looked into the daily concentration of various pollutants corresponding to various counties, using box-plots. We then selected a particular pollutant, namely NO2, for our further analysis. We calculated the daily AQI value and average NO2 concentration for all the months from January to February for various counties (refer Fig5). We found that there was a decreasing trend in NO2 concentration from January to December.

To further validate our hypothesis, we looked into the NO2 concentration for previous years and found out that there was a decrease in the amount of pollutants from previous years (refer Fig 6-9). This helped us conclude that lockdown coming into effect from around March in NYS helped reduce the concentration of various pollutants in the air. We were also able to verify that lockdown had a greater impact on larger and more populated counties. The smaller counties did not see a significant change in the pollutants concentration throughout the year.

b) Provide a description of the choices of tools/methods used or a description of any code or scripts written, and describe how your results were stored and managed, Min 3-4 sentence. Submit your code to course GitHub repository for evaluation (3%)

We used the Pandas package in Python to read the data and convert it into dataframes. We also made use of NumPy for some preprocessing and exploration on the data. In addition to these two, we used matplotlib and Seaborn libraries in Python to perform our visualizations and

exploratory analysis. Overall, the workflow involved reading the data, cleaning and preprocessing the data, merging the data, and finally visualizing and drawing conclusions based on our analysis. All the ipynb files, data files and python scripts are stored on GitHub for easy replication. We also stored the poster and our final report on GitHub for long term access.

c) Perform the analysis in a form that can be validated and describe the steps and results your group took to ensure this validation, Min 3-4 sentences (4%)

Given our initial hypothesis, our goal was to verify the impact of lockdown caused by COVID on air quality. We started off by exploring our covid dataset for new positives in various counties in NYS. We looked into the daily concentration of various pollutants corresponding to various counties, using box-plots. We then selected a particular pollutant, namely NO2, for our further analysis. We calculated the AQI value and average NO2 concentration for all the months from January to February for various counties. We found that there was a decreasing trend in NO2 concentration from January to December.

To further validate our hypothesis, we looked into the NO2 concentration for previous years and found out that there was a decrease in the amount of pollutants from previous years. This helped us conclude that lockdown coming into effect from around March in NYS helped reduce the concentration of various pollutants in the air. We were also able to verify that lockdown had a greater impact on larger and more populated counties. The smaller counties did not see a significant change in the pollutants concentration throughout the year.

3. Presentation/Visualization (8%)

a) Prepare presentation / visualization of both the data (and any metadata, information) and the results of the analysis and describe them, Min 2-3 sentences. (3%)

Following are the figures and plots that show the correlation between the Air Quality data and Covid-19 data.
Fig. 1,2 3, 4 and 5 show concentration of various air pollutants against the counties of New York State.
Fig. 6 - 9 show monthly concentration levels of the air pollutants.
Fig. 10 - 14 present Daily change in AQI value in various counties after removing outliers, creating cleaned dataset for analysis and plots.

Thus, the following plots build up gradually how Air Quality and the pollutants vary with areas, their sizes, their population as well as monthly changes in number of positive Covid-19 cases.

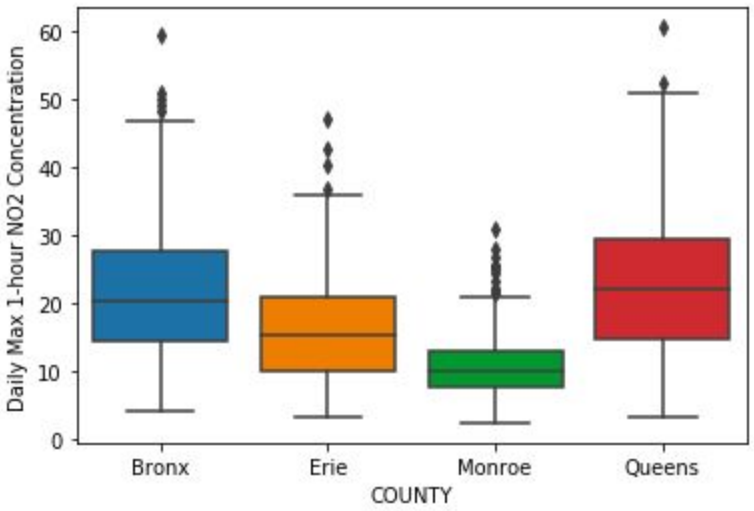

Fig1: Daily NO2 concentration in various counties in NYS

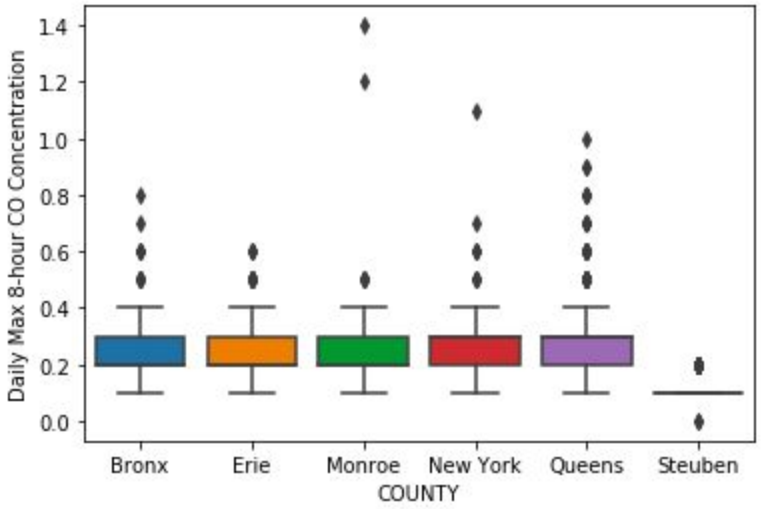

Fig2: Daily CO concentration for various counties in NYS

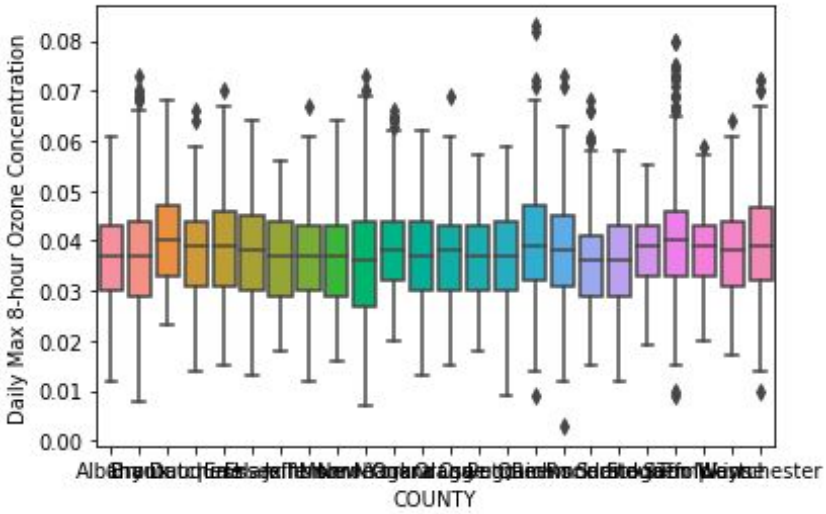

Fig3: Daily Ozone concentration for various counties in NYS

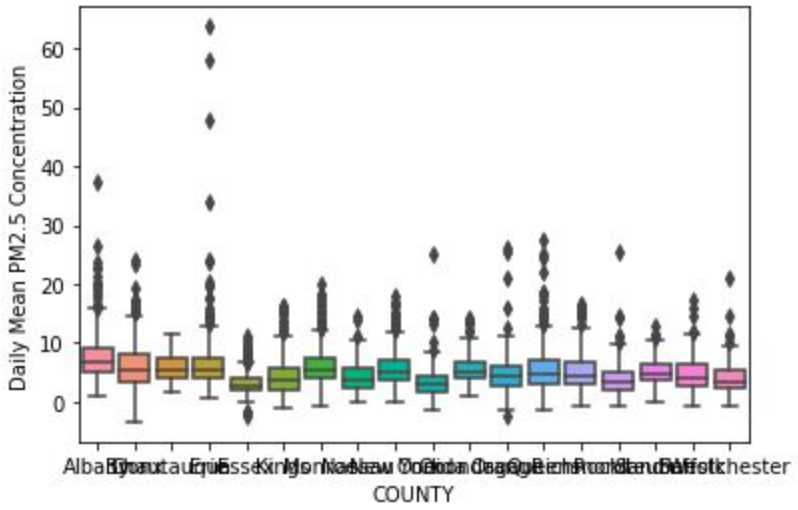

Fig4: Daily PM2.5 concentration for various counties in NYS

For Albany AVG by month

	Month	AVG DAILY_AQI_VALUE	Avg Ozone
0	January	26.344828	0.028448
1	February	31.896552	0.034517
2	March	35.516129	0.038452
3	April	39.300000	0.042400
4	May	37.032258	0.039968
5	June	38.933333	0.041200
6	July	35.612903	0.038484
7	August	33.064516	0.035742
8	September	31.466667	0.033967
9	October	26.838710	0.029065
10	November	27.566667	0.029800
11	December	25.333333	0.027333

Fig5: AQI value and average NO2 concentration for Albany

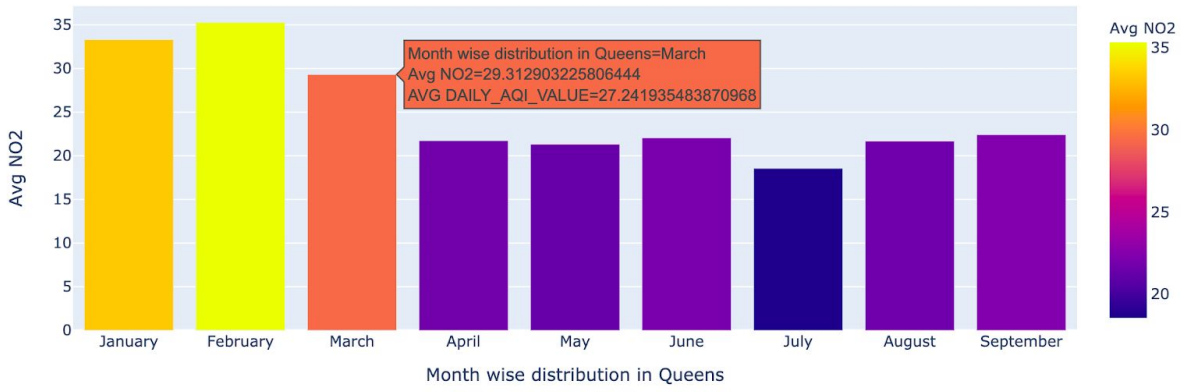

Fig6: Month wise concentration of NO2 in Queens

Fig7: Month wise concentration of NO2 in Bronx

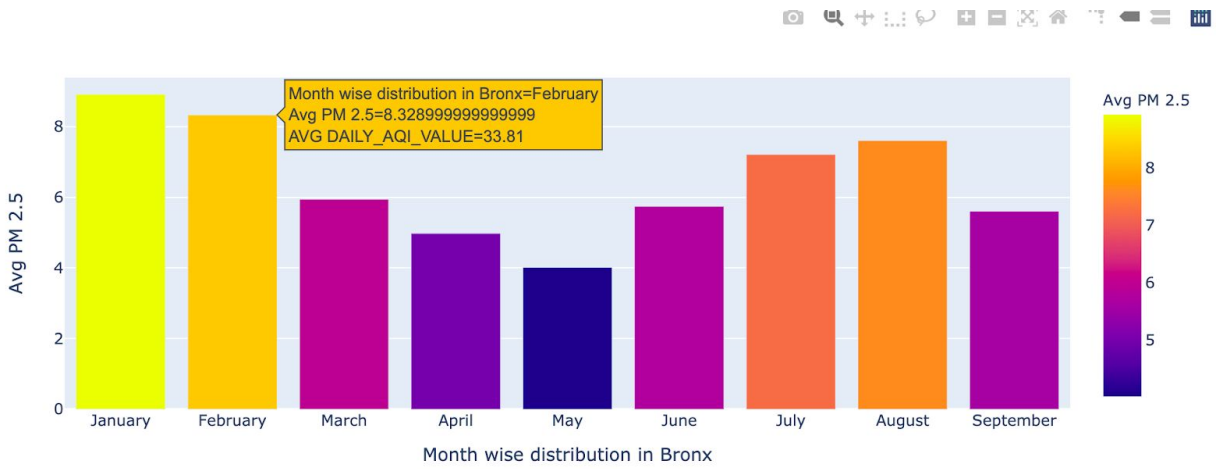

Fig8: Month wise concentration of PM2.5 in Bronx

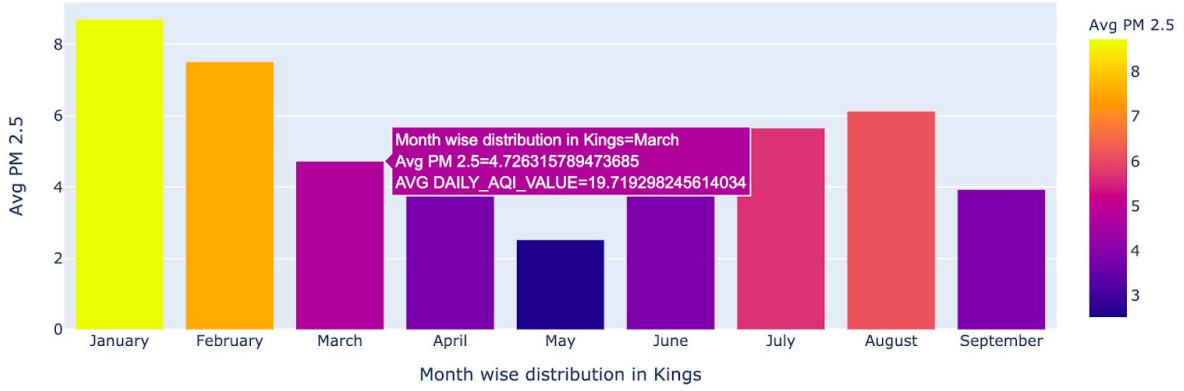

Fig9: Month wise concentration of PM2.5 in Kings

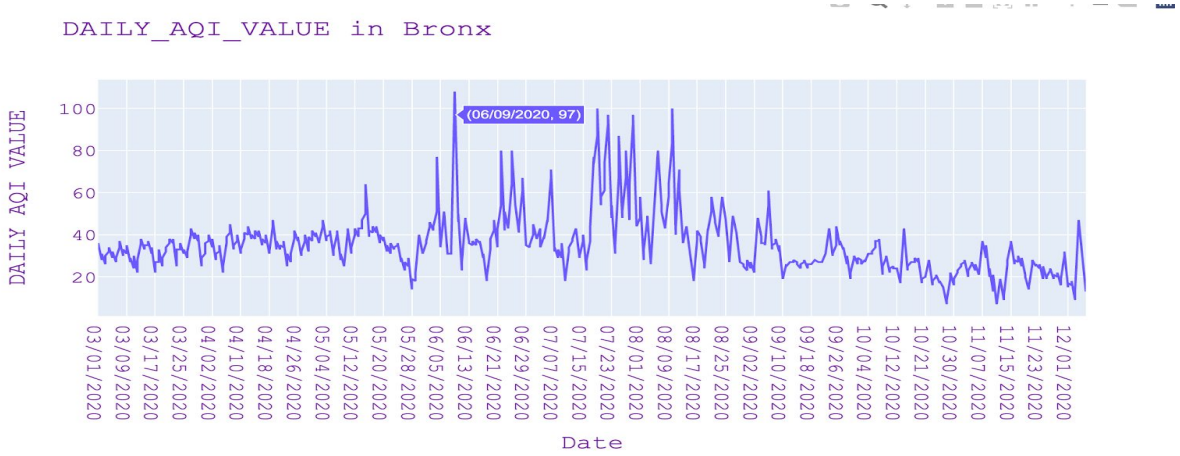

Fig10: Daily change in AQI value in Bronx before removing outliers

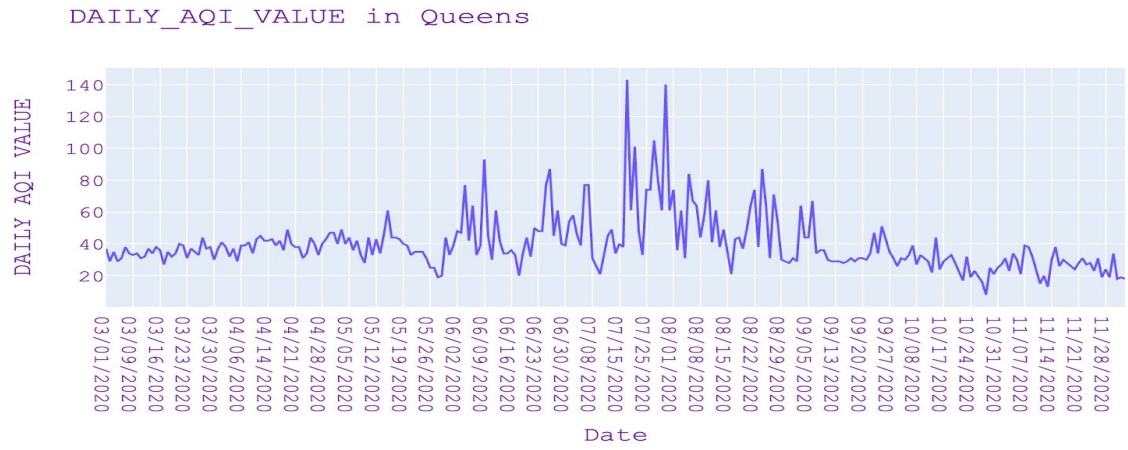

Fig11: Daily change in AQI value in Queens before removing outliers

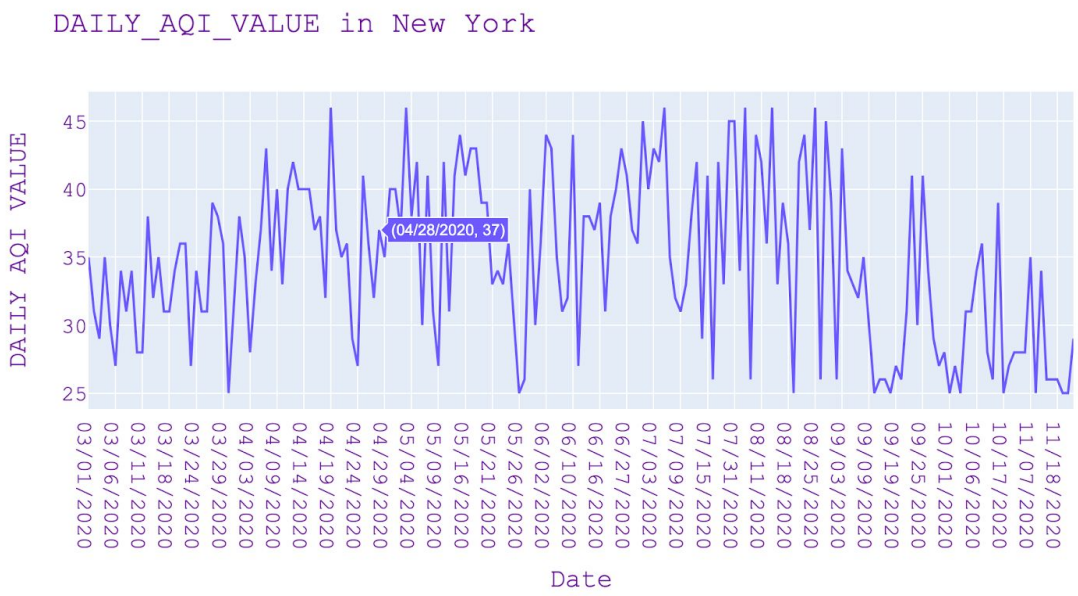

Fig12: Daily change in AQI value in New York after removing outliers

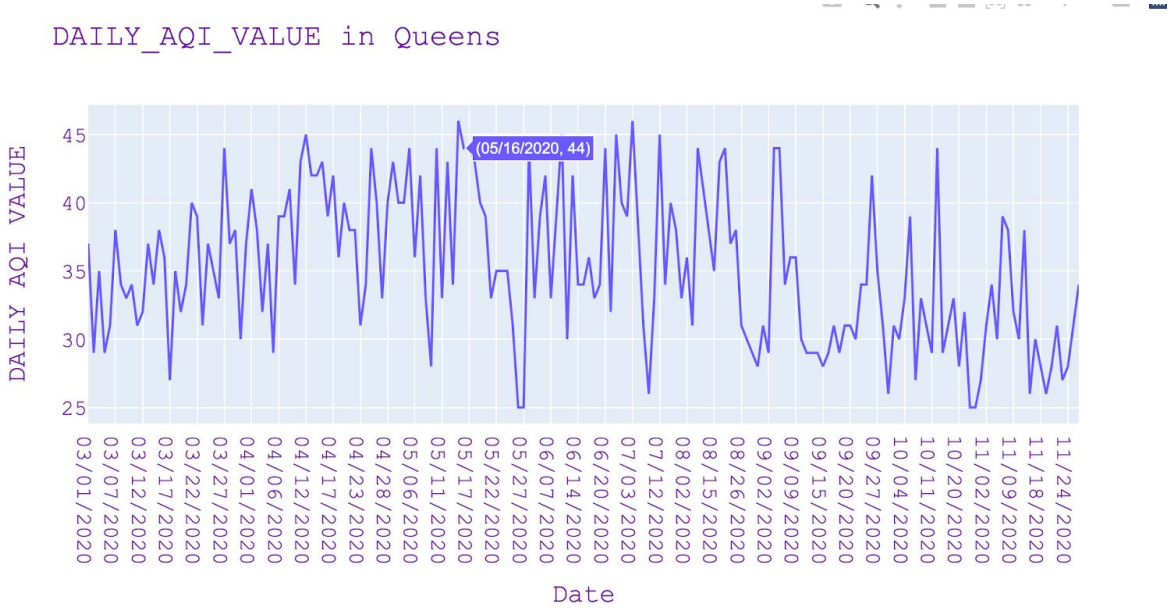

Fig13: Daily change in AQI value in Queens after removing outliers

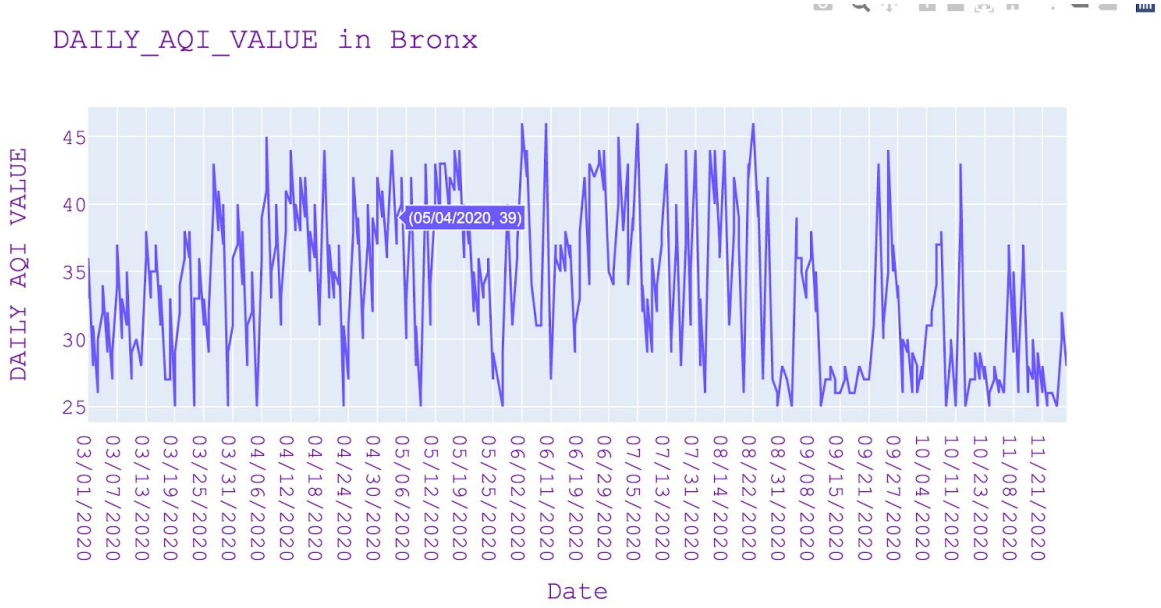

Fig14: Daily change in AQI value in Bronx after removing outliers

b) Document the management of the presentation / visualization products and any associated metadata, etc. Min 2-3 sentences (2%)

We stored both the poster, the associated plots and this final report in our course's GitHub repository. The plots are self-describing; each of them have titles associated with them which describe what they show, and the legends show how to correctly interpret them. For example, from the seaborn box-plots we can see that the pollutants concentration varies across different counties. The labels in these plots describe the counties and the pollutant for which the plot is drawn.

Long term storage of these visualizations ensures that anyone in the future can easily recognize the findings and can then take it further from there. The storage on GitHub ensures that the data, visualizations and analysis are archived for future use. This detailed report also makes sure that the person with the interest in using our work has proper documentation of our work and so that he/she can utilize it further.

The metadata and formats are uploaded in the Github repository's readme.md file for learning about the description of datasets used for this research and analysis.

For Air Quality is located in : https://aqs.epa.gov/aqsweb/airdata/FileFormats.html
For Covid-19, the descriptive metadata and dictionary are located in pdf files called: NYSDOH_COVID19_Data_Dictionary.pdf and NYSDOH_COVID19_Overview.pdf

c) Describe how your presentation/visualization supports the goal of the data science investigation and highlight any value that was gained, Min 3-4 sentences (3%)

We answered our hypothesis mostly through visualizations and a thorough exploratory analysis of the data. After this analysis, a correlation between the covid tests and pollutants in the air (AQI value) can be clearly seen and verified. We decided not to go for more in-depth analysis because it was clear that we would not learn anything new. Our presentation and our poster will clearly explain our results. Through the exploratory analysis of both our datasets we were able to identify the patterns in the data and prove our hypothesis. Furthermore, taking into consideration more factors such as population, population density, and other pollutants data will further improve our analysis and underscore our point.

4. Data management plan

Describe your overall data management plan for the results for questions 1,2, and 3 using the 9 categories from assignment 2, Min 1-2 sentences for each category (5%)

- **Creation of logical collections:**
  The data was captured from two different sources and there are two main logical collections- the Air Quality Index data and the Covid-19 data for New York State. These two main collections contain separate pre-processed data sheets that contain cleaned data after the elimination of outliers. Then, both the collections are merged into one for data analysis.

- **Physical data handling:**
  Both datasets have been collected as spreadsheets and .CSV files were downloaded from two different sources.The spreadsheets have been uploaded to Github in the “data” folder accessible through this link, as well as a backup has been created on Google Drive. It is used as a sharing medium and a backup as well.

- **Interoperability support:**
  Both the datasets as well as the processed ones are in the .csv formats which is easily machine- readable through MS Excel, OpenOffice Calc, or Google Sheets. Hence, it is easily transferable and accessible. These sheets can be used for data analysis in R or Python, just like the pandas library that we used for data analysis.

- **Security support:**
  The data, python .ipynb files, and results with the visualizations are all saved on Github as a comprehensive project. Apart from this, the datasets and cleaned/processed data sheets are saved on Google Drive’s shared folder. There, the project information is accessible only by the team members, securing it from any breach or external access; whereas the Github repository is accessible by the team members, the professor and anyone who can find the repository. Although, anyone who is able to access the repository cannot make any changes to it. Contributors can be added by group members only, making it secure from manipulation.

- **Data ownership:**
  Covid-19 data is captured from the Github repository owned by the New York State Govt and Air Quality Index dataset is owned by the United States Environmental Protection Agency. Apart from this, the data after analysis and all other project information in terms of code, ipynb files, processed .csv files etc., are owned by the project team. It will be accessible to the professor and RPI after submission through LMS.

- **Metadata collection, management and access:**
  The Metadata and description about the data was captured from both the sources. Since, the datasets were downloaded and then worked upon, the metadata standards and conventions were adopted as they were originally in the given Government website sources. These standards were easily usable for our research through pandas and other libraries of python. The standards and important descriptive information that we filtered to use for our analysis is captured in the readme.md file on Github repository that anyone accessing our repository for cloning, can easily locate and read.

- **Persistence:**
  The project data will be available on the Github repository until one or more members decide to manipulate or delete it. It does not have an expiration date at all.

- **Knowledge and information discovery:**
  The project data can be accessed through Github repository and it can be cloned by any data science enthusiasts to build up on the knowledge and interpretation. The information discovery on the lines of correlation between Air data and Covid-19 can be further built upon for a comparative study with data in future as well.

- **Data distribution and publication:**
  Anyone researching on the effects of air quality or the level of pollution with Covid-19 data (especially for New York state), can build more on this analysis and this information is highly useful for the correlation drawn above. In this era of the pandemic and later on, research on the correlation between AQI and Covid-19 cases can be of extensive help for journals, articles and research articles. The project data can be cloned through the Github repository.

5. Poster

Create a poster (poster templates are available on LMS). Please submit your poster on LMS using the same naming scheme mentioned above. Mandatory peer-evaluation form must be submitted within 12 hours of the final project submission on LMS in order to receive the presentation and participation grade. (10%)

Done and submitted on to GitHub and LMS