NYC Air Quality - Air Quality Prediction

HDip in Science in Artificial Intelligence Applications

Kate Gogolka(sbs23076@student.cct.ie)
Jim Higgins(sbs23021@student.cct.ie)
Ciaran Quinlan(sbs23098@student.cct.ie)

Word count: 3102

Lecturer: James Garza

CCT College Dublin

Assessment Cover Page

Module Title:	Strategic Thinking CA1
Assessment Title:	CA1 Capstone Project
Lecturer Name:	James Garza
Student Full Name:	Katarzyna Gogolka, Jim Higgins, Ciaran Quinlan
Student Number:	Katarzyna Gogolka (sbs23076@student.cct.ie) Jim Higgins (sbs23021@student.cct.ie) Ciaran Quinlan (sbs23098@student.cct.ie)
Assessment Due Date:	07/05/2023
Date of Submission:	07/05/2023

Katarzyna Gogolka Jim Higgins Ciaran Quinlan

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Table of Contents

Table of Contents 2

Table of Figures 3

1. Introduction 4

2. Pollutants monitored 5

   2.1. Fine Particulate Matter (PM2.5) 5

   2.2. Black Carbon 6

   2.3. Nitric Oxide (NOx) 7

   2.4. Nitrogen Dioxide (NO2) 8

   2.5. Sulphur Dioxide (SO2) 9

   2.6. Ozone 10

3. General goal of this Project 11

4. Technologies used 11

   4.1. Libraries 11

   4.2. Machine Learning Algorithm 12

   4.3. Plotting and Visualisations 12

   4.4. Team Collaboration tools 13

   4.5. Version Control – GIT 13

5. What has been accomplished so far 14

   5.1. Datasets 14

   5.2. Source 14

   5.3. Dimensions 14

   5.4. Descriptive statistics 14

   5.5. Data visualisation 15

   5.6. Data preparation 15

   5.7. Models 16

6. CRISP-DM Methodology 17

7. Challenges encountered 18

   7.1. Challenge 1: Datasets 18

   7.2. Challenge 2: Team members 18

   7.3. Challenge 3: Target variable and Machine Leaning Model 18

8. Results and analysis and next steps 19

9. Conclusion 20

10. References 21

Table of Figures

Figure 1 PM2.5 (NYC.Gov, 2023) 5
Figure 2 Black Carbon (NYC.Gov, 2023) 6
Figure 3 Nitric Oxide (NYC.Gov, 2023) 7
Figure 4 Nitrogen Dioxide (NYC.Gov, 2023) 8
Figure 5 Sulphur Dioxide (NYC.Gov, 2023) 9
Figure 6 Average Summer Ozone concentration (NYC.Gov, 2023) 10
Figure 7 Daily Max SO2 Concentration vs Years 12
Figure 8 Daily Mean PM2.5 and Daily max 8 hour ozone concentration plots 12
Figure 9 Invitation of members to collaborate on the repository 13
Figure 10 Daily mean PM 2.5 and daily max 8 hour ozone concentration outliners 14
Figure 11 Daily max 1 hour NO2 and daily max 1 hour SO2 concentration outliners 14
Figure 12 Air Quality Index classification (Oklahoma Department of Environmental Quality, n.d.) 15
Figure 13 Correlation matrix of a cleaned data 16
Figure 14 Interface of datasets source 18
Figure 15 Actual vs Predicted values with a diagonal line for perfect prediction indication. 19
Figure 16 Actual vs Predicted Values plot 19

1. Introduction

With advancing education and understanding the impact air pollution has on the health and wellbeing of its inhabitants Air Quality should be improving in large cities.
Steps have been taken over the last two decades by the city of New York with a goal to 'have the best air quality among all large U.S cities by 2030' (NYC, 2018)

Incentives included:

- Reduce vehicle emissions from roads by retrofitting the City's vehicles' diesel engines to meet better emission standards. In the (NYC, 2018) report it is estimated that replacing or retrofitting a vehicle to 2007 standards reduces emissions by approximately 90 percent over the previous standard.
- Reduce general transport emissions (including ferries, trains and so on) and convert public transport to greener fuel sources. Adopting Electric Vehicles for some of the transit network and offering rebates for the upgrade and conversion of trucks to carbon neutral gas and scrappage allowances for the removal of 'dirtier' diesel trucks.
- Introduction of a congestion charge in the financial area of Manhattan to promote the use of public transport in and out of the area. Gateless tolls are implemented to minimise idling engine pollution.
- Reduce emissions from buildings. In 2015 the City stopped issuing permits for the use of #6 oil as a heating oil.

2. Pollutants monitored

The NYC Community Air Survey Findings Between 2009 and 2017 (NYC.gov, 2019) showed that the annual average levels of fine particulate matter (PM 2.5), Nitrogen Dioxide (NO2), Nitric oxide (NOx) and black carbon "have declined 30%, 26%, 44% and 30%, respectively."
More up to date data shows they have continued to fall with slight increases in PM2.5 and Black Carbon after the NYC Pause due to Covid-19 (NYC.Gov, 2023)

2.1. Fine Particulate Matter (PM2.5)

Fine particulate matter consists of small, airborne particles with a diameter of 2.5 micrometres or less. These particles form in the atmosphere because of interactions with chemicals such as sulphur dioxide and nitrogen oxides, which are emitted from power plants, industries, and automobiles. (USEPA, 2022)

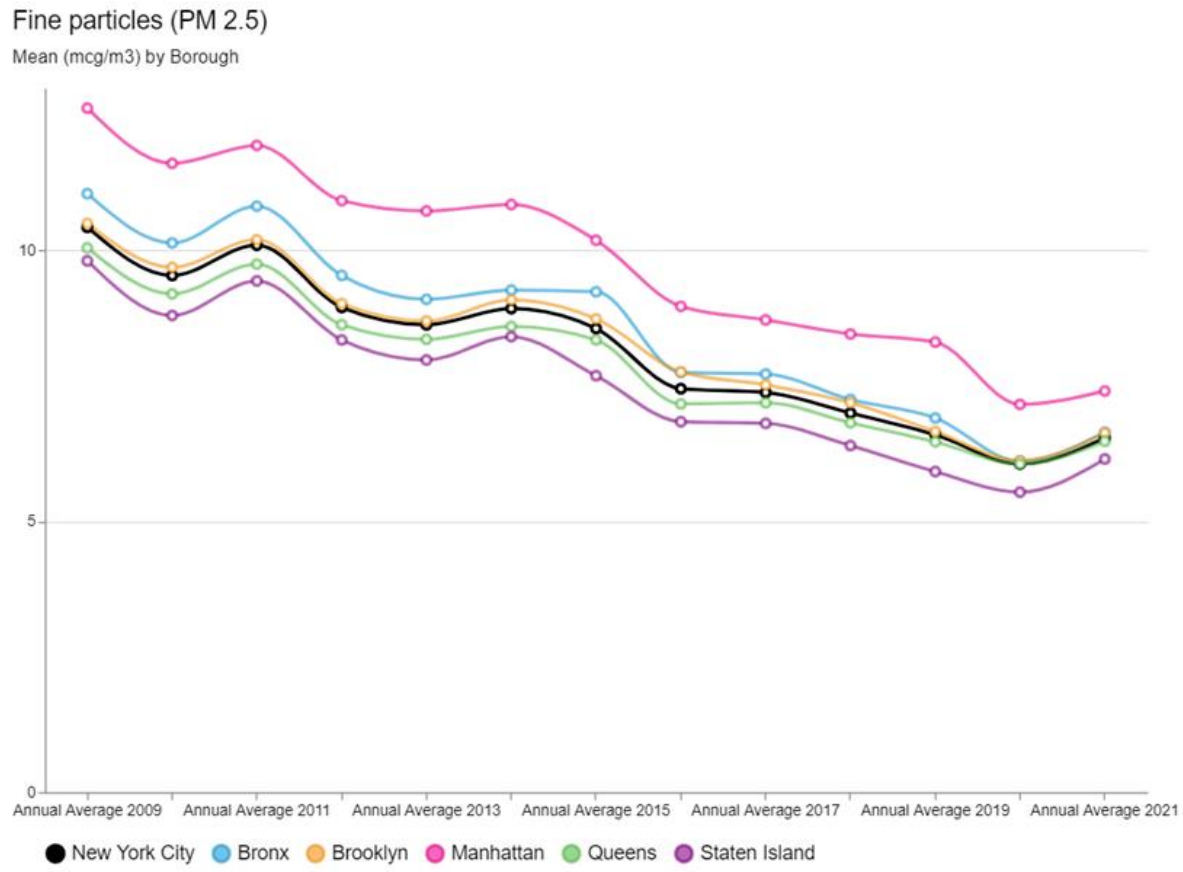

Figure 1 PM2.5 (NYC.Gov, 2023)

2.2. Black Carbon

Black carbon is the soot like material emitted from combustion engines, coal-fired power plants, and other sources that burn fossil fuel. It comprises a significant portion of particulate matter. (USEPA, 2022)

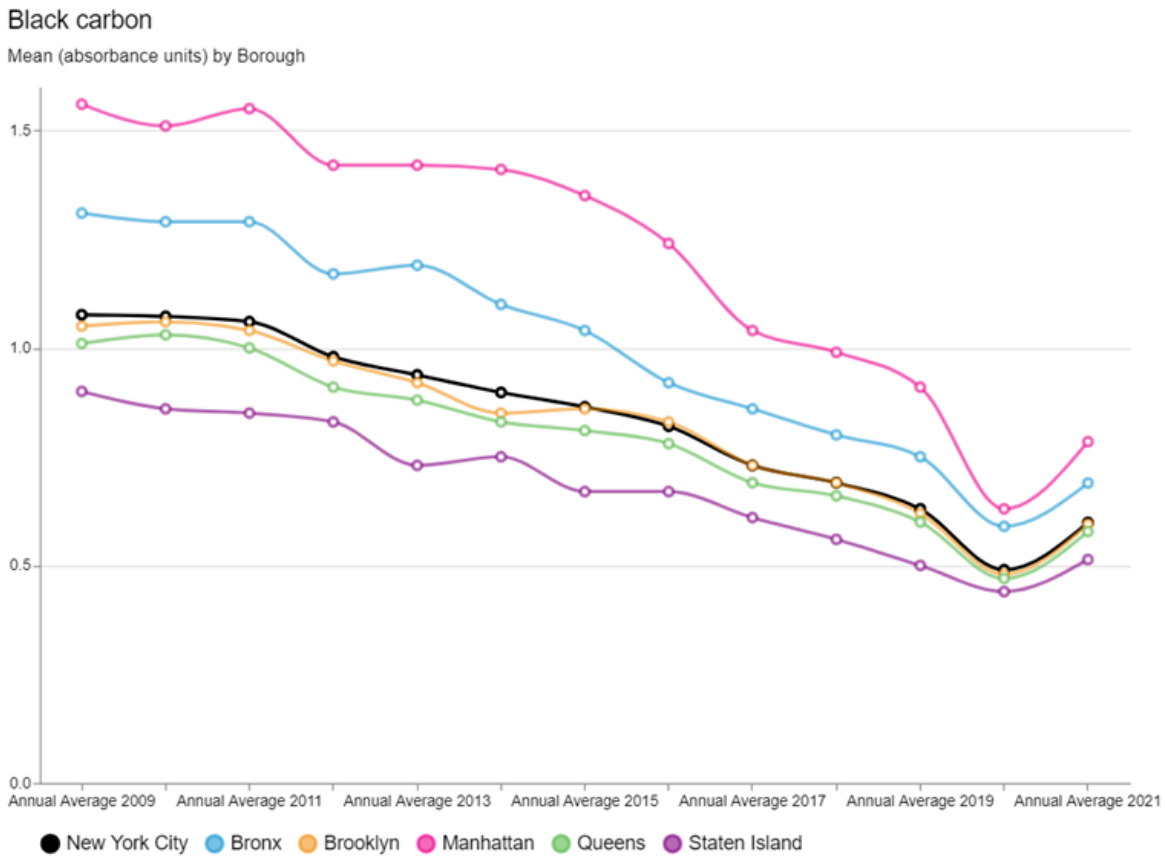

Figure 2 Black Carbon (NYC.Gov, 2023)

2.3. Nitric Oxide (NOx)

A gas formed by combustion under high temperature and high pressure in an internal combustion engine; it is converted by sunlight and photochemical processes in ambient air to nitrogen oxide which is a precursor of ground-level ozone pollution, or smog. (U.S EPA, 2018)

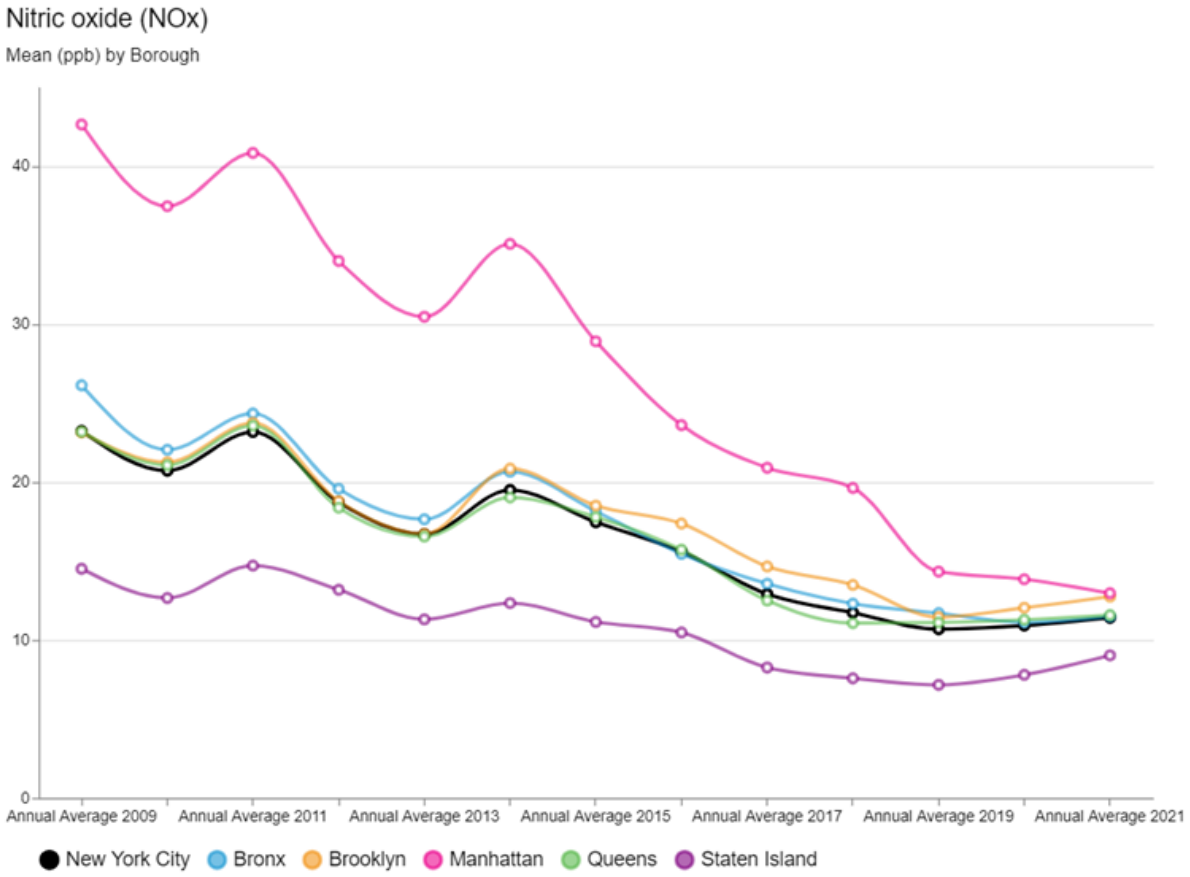

Figure 3 Nitric Oxide (NYC.Gov, 2023)

2.4. Nitrogen Dioxide (NO2)

A reactive gas which is present in urban atmospheres. This gas is formed in the atmosphere from emissions of oxides of nitrogen (NOX). As discussed previously NOX is produced by fuel combustion sources, vehicles and industrial boilers. (U.S EPA, 2018)

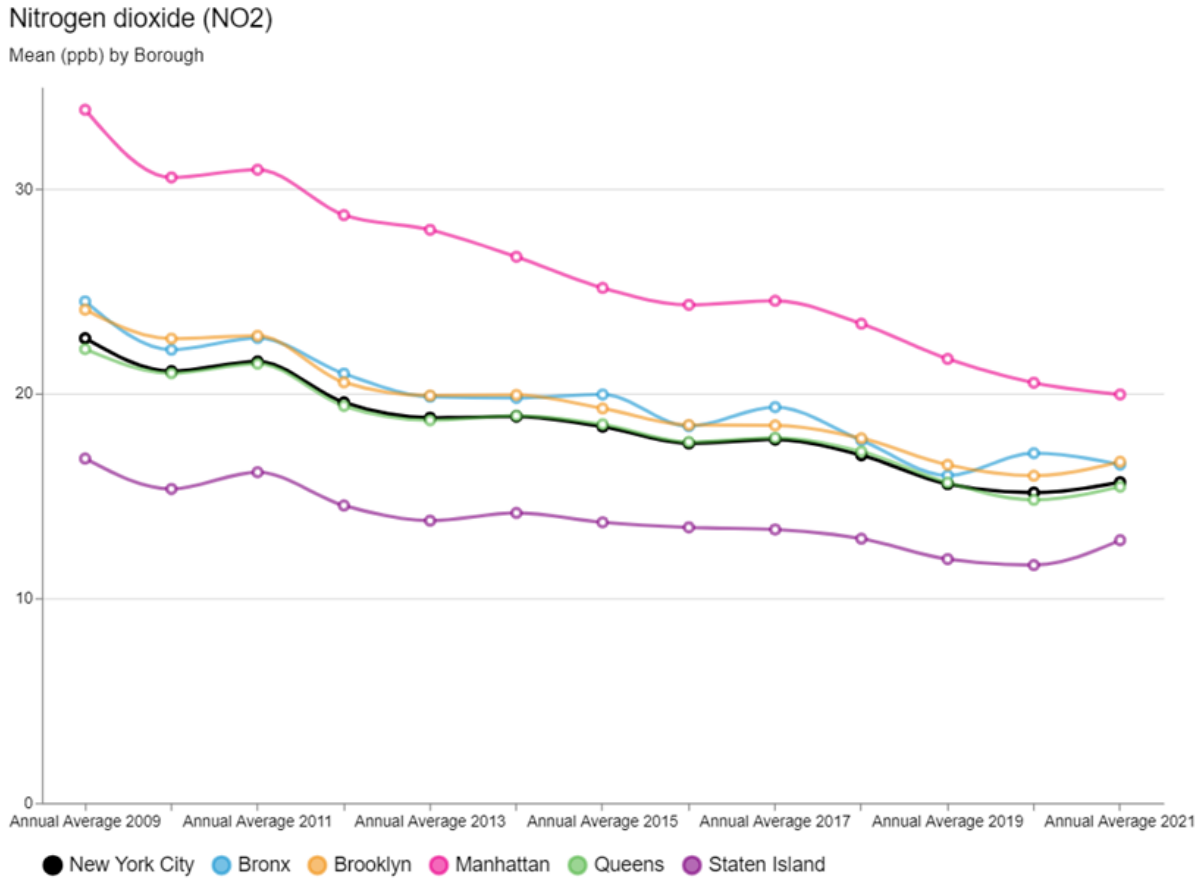

Figure 4 Nitrogen Dioxide (NYC.Gov, 2023)

2.5.Sulphur Dioxide (SO2)

Sulphur dioxide is a naturally occurring gas that causes acid rain. The burning of fossil fuels such as coal and gas releases SO2 into the atmosphere. (U.S EPA, 2018)

The largest declines have been observed for Sulphur Dioxide (SO2), due largely to an implementation of heating oil regulations. As of July 2015, there were no more permits issued for the use of the #6 fuel oil, a heavy burner oil used primarily for maritime and oil-fired heating systems. All boilers had to switch to gas or less harmful oils #2 or #4. With targeted enforcement and support approximately 90% of boilers had been converted over from #6 oil with thousands more to switch over from #4 oil to less polluting #2 oil by 2030. (NYC, 2018)

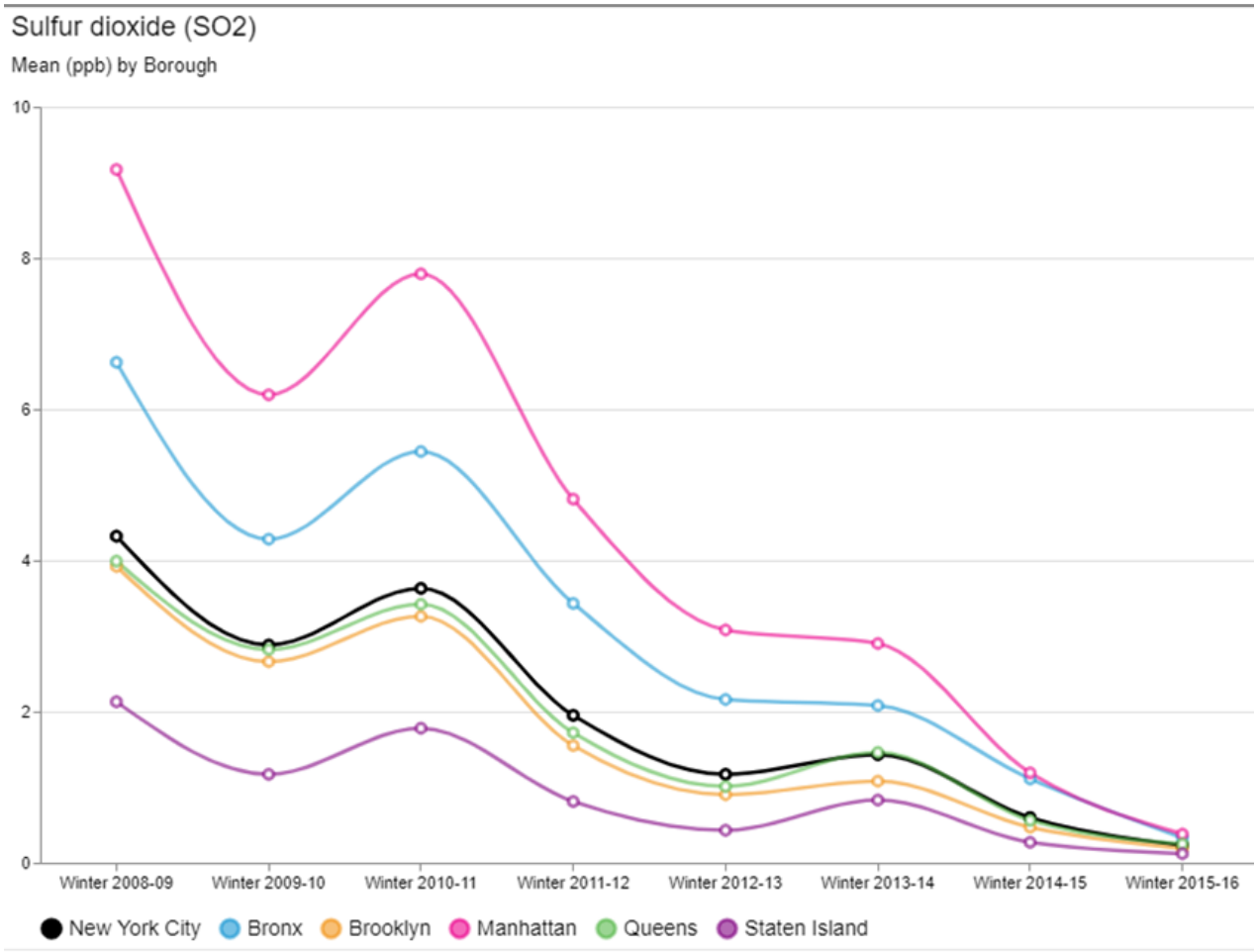

Figure 5 Sulphur Dioxide (NYC.Gov, 2023)

2.6. Ozone

Ozone is formed in the atmosphere through the reaction of other pollutants (oxides of nitrogen and volatile organic compounds) in the presence of sunlight.
Warmer temperatures and increased daylight hours result in increased ozone production. (U.S EPA, 2018)

The overall Summer Ozone concentration remained stable and is consistent by Borough.
In areas of high density of pollution sources, emissions tend to react with ozone and reduce concentrations.
A report summary on the ozone trends 2009 to 2015 by NYC Community Air survey (NYC.Gov, 2016) concludes that areas of high traffic density tend to have lower ozone concentrations.
This would explain the increase in Summer Ozone concentration in Manhattan during the 'New York Pause of 2020' due to Covid-19

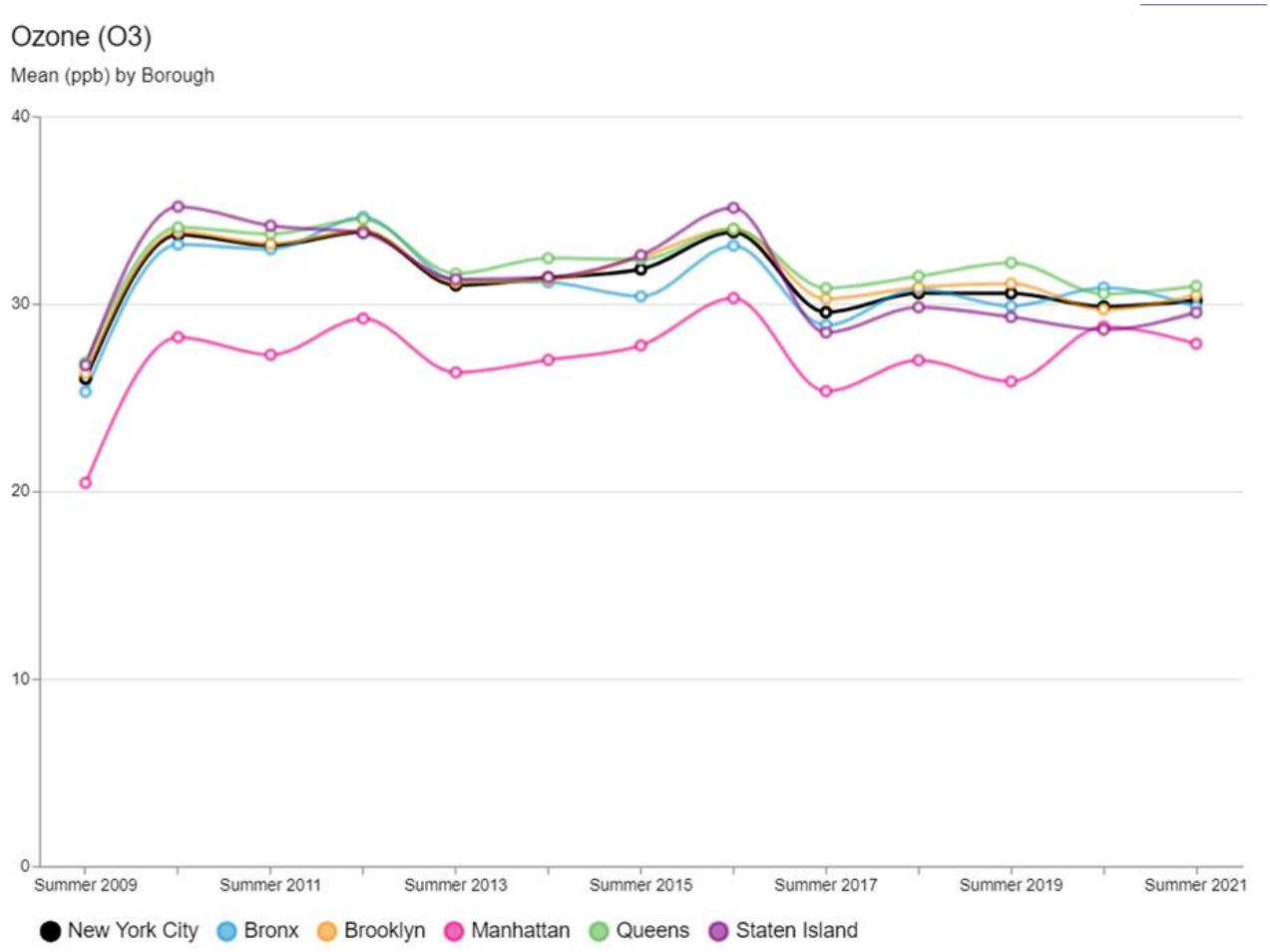

Figure 6 Average Summer Ozone concentration (NYC.Gov, 2023)

3. General goal of this Project

The goal of this project:

- Access the data available for the Air Quality Value for the city of New York from the Environmental Protection Agency
- Assess the current trend and determine if it is in line with the expectations published by the Mayor's Office in 2018 (NYC, 2018)
- Make a prediction with a machine learning algorithm of the Air Quality Value
- Determine the next steps to improve the model as knowledge of the programming language improves.

4. Technologies used

4.1. Libraries

We used a Jupiter Notebook with Python as the coding language for this project and utilized the following libraries:

- import pandas as pd
- import numpy as np
- import matplotlib.pyplot as plt
- import seaborn as sns
- from sklearn.model_selection import train_test_split, cross_val_score
- from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
- from sklearn.metrics import confusion_matrix
- from sklearn.ensemble import RandomForestRegressor
- from datetime import datetime

4.2. Machine Learning Algorithm

For predicting continuous numerical values and where the data is labelled, we have decided to use a supervised learning algorithm with a Random Forest Regressor model with default hyperparameters.

4.3. Plotting and Visualisations

With only basic data cleaning we were able to plot some simple graphs to help us visualise the dataset we had.

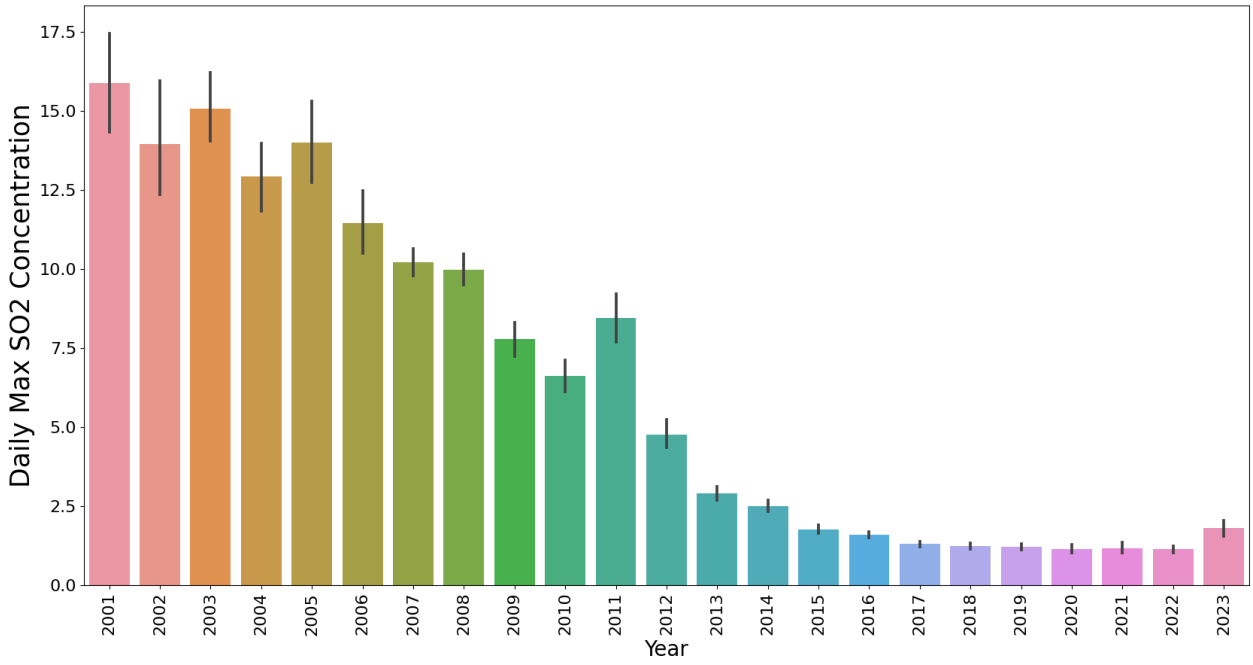

Figure 7 Daily Max SO2 Concentration vs Years

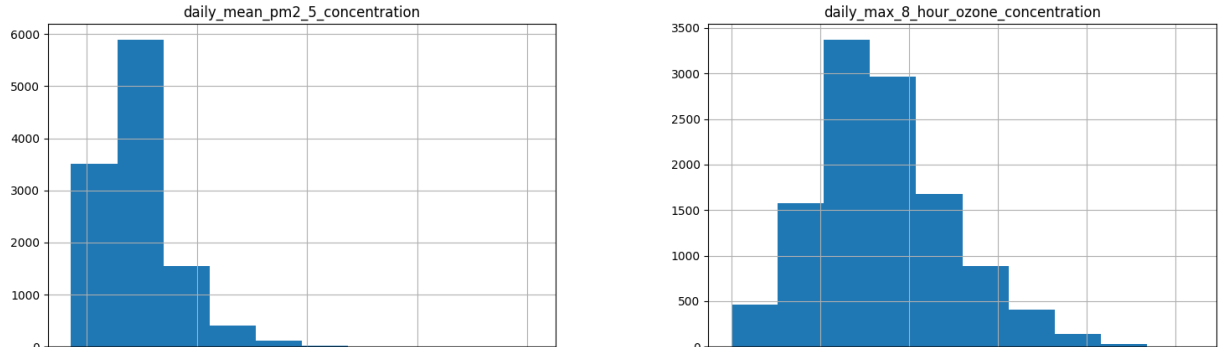

Figure 8 Daily Mean PM2.5 and Daily max 8 hour ozone concentration plots

4.4. Team Collaboration tools

We started out the project with a **whats app group** for the 3 members and this was great for passing back and forth messages and getting to know each other's working style and availability. James organised a **zoom meeting** for our first sprint and we stayed in the meeting room to meet Face to Face and put a plan together for the project.

We used the **google drive** on our CCT accounts to share documents and datasets and continued to use this throughout the project as a shared project drive.

Kate introduced us to **discord** (https://discord.com/channels) and we got a meeting room there and we used this to share ideas and content for the CA. As we all got to understand how discord would support our virtual working we became very comfortable with the platform and it was excellent to 'jump in' on our Discord channel when we had some time or some feedback for the group.

4.5. Version Control – GIT

Since we covered it in class we started to investigate the use of Git and GitHub to manage the Jupyter notebook and the datasets. While we did not use it extensively, we did upload some of our code as the project started to take shape. We are ready to start to use it in the next phase of the CA. The way we used GIT was as follows:

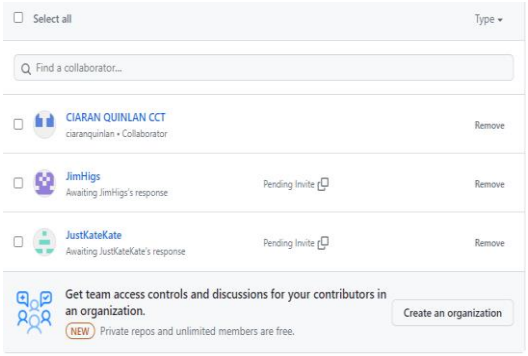

Figure 9 Invitation of members to collaborate on the repository

1. We set a blank public repo in github.com under the personal account of ciaranq and created a Git repo under my personal Git account called ciaranq/ny-air-quality
2. Assigned this a Public access setting and got the URL of this repository.
3. We the opened up our local GitHub Desktop application and used our CCT Git login that we created in Class with James in the GIT tutorial
4. To access the repository, we selected File/Clone repository, enter https://github.com/ciaranq/ny-air-quality.git in the URL option
5. We then selected a local folder to store the repository and where we could run the notebooks from and use the code that was shared by Selecting Clone at the end of the dialog. This cloned the repo to the local directory,

To use the Git to push your changes: we did this:

6. Make whatever changes to the local file and add a reference for the commit, Commit to the main branch

For the team members to get access to the repo the following steps were followed.

7. Go to the git account that owns the repository, ciaranq
8. Select settings, collaborators, Enter the emails of the collaborators and send an invite to @ciaranquinlan, @JimHigs, @JustKateKate.
9. The team members get an email and are asked to accept the invite, this is done by logging into their git account.

Team members can complete steps 3-6 above and the repository of our CA will be downloaded to their local folder.
Team members can complete steps 7-8 to make a commit to the main branch of the repository.
To Pull or fetch the repo to update the repo with the main branch use a pull request from the repository menu in GitHub desktop. ("GitHub.com Help Documentation")

5. What has been accomplished so far

5.1. Datasets

Four pollutants of air quality were taken from (epa.gov) for NYC borough, Queens, and for only one station Queens College 2.

5.2. Source

A reliable source of the United States Environmental Protection Agency (https://www.epa.gov/outdoor-air-quality-data/air-data-concentration-plot)

5.3. Dimensions

To merge the datasets successfully we had to use specific columns when reading the csv files. After a thorough investigation of the four datasets, we decided to use the values of the pollutants, units the pollutants were measured in, date, daily aqi value and site latitudes/longitudes.
We merged the datasets on date to ensure dates matched.
The dimension of the dataset came out to be 11,637 rows with 15 columns before data cleaning.

5.4. Descriptive statistics

We plot the four pollutants on a histogram to see the outliers. We decided that the outliers were not as extreme and could be useful so we decided to leave them in when working on our model.

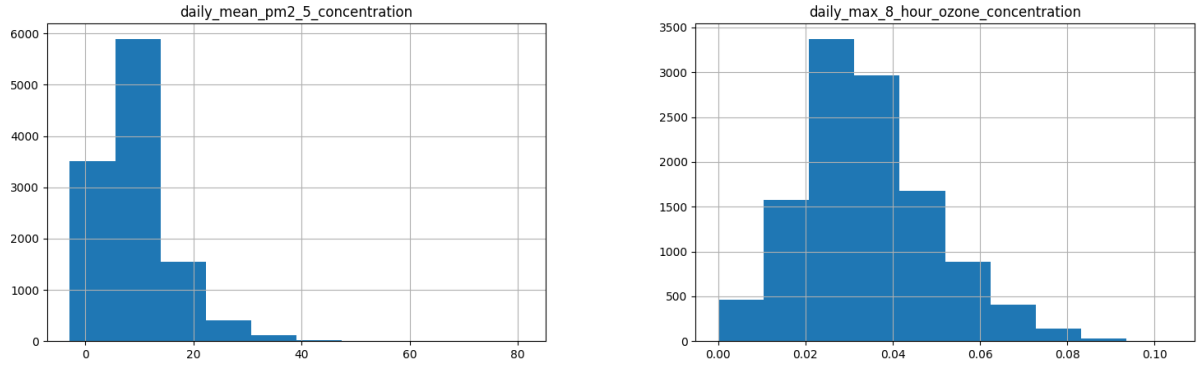

Figure 10 Daily mean PM 2.5 and daily max 8 hour ozone concentration outliners

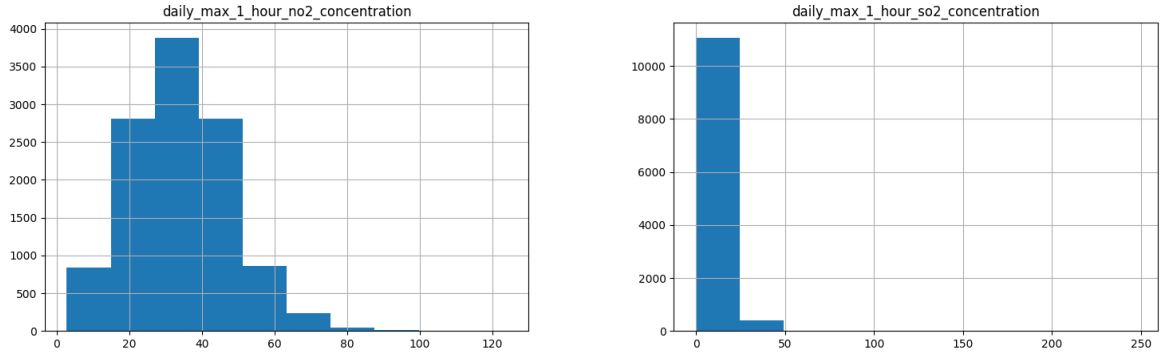

Figure 11 Daily max 1 hour NO2 and daily max 1 hour SO2 concentration outliners

5.5. Data visualisation

Once we changed the date to datetime and extracted useful features from it we were able to plot the pollutants on the graphs to better visualise their behaviour.

5.6. Data preparation

Before any machine learning model could be trained, we had to prepare and clean the data.

We have standardised the column names to improve computational time and ease of transferring the names we made them all lowercase without spacings or dots but a "_".

Changed date to datetime and extracted useful features.

The AQI values were different for each pollutant, so we decided to take an average of the four and create a new column with a main AQI value and predict our AQI based on it.

We have created a column that takes the AQI value and shows the AQI index as per the US AQI guidelines.

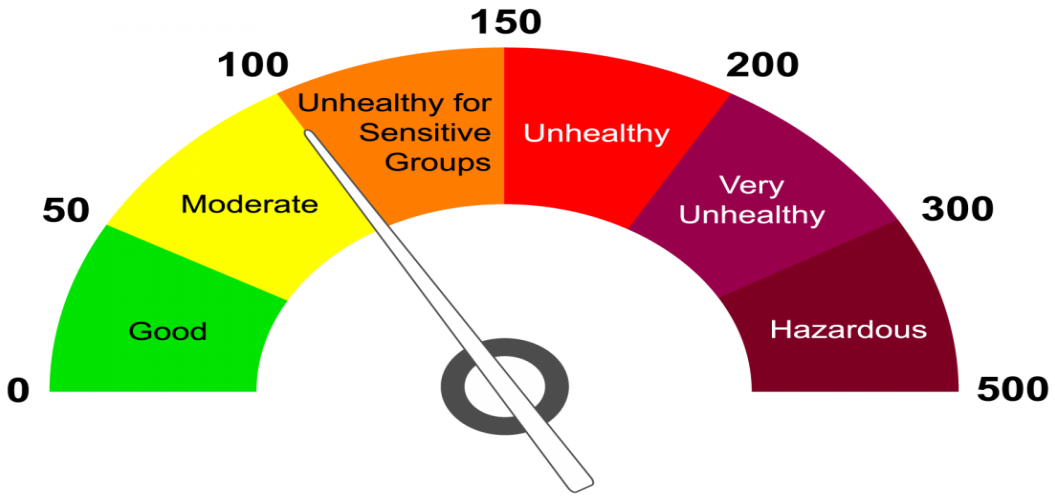

Figure 12 Air Quality Index classification (Oklahoma Department of Environmental Quality, n.d.)

We also had to check for duplicates, missing values, and incorrect data.

Since pollutants can't be a negative number, we had to delete all negative numbers from our pollutant's columns.

5.7. Models

Before the machine learning model, we had to decide which columns are unnecessary for the model and drop them. We decided to drop units, and site location, as the site location is the same for all the stations and units can't be matched with the corresponding value.

We plotted the correlation matrix and decided to use all the listed columns for our model. Even though the day/month correlates weakly, we wanted to predict the AQI value by day/month too.

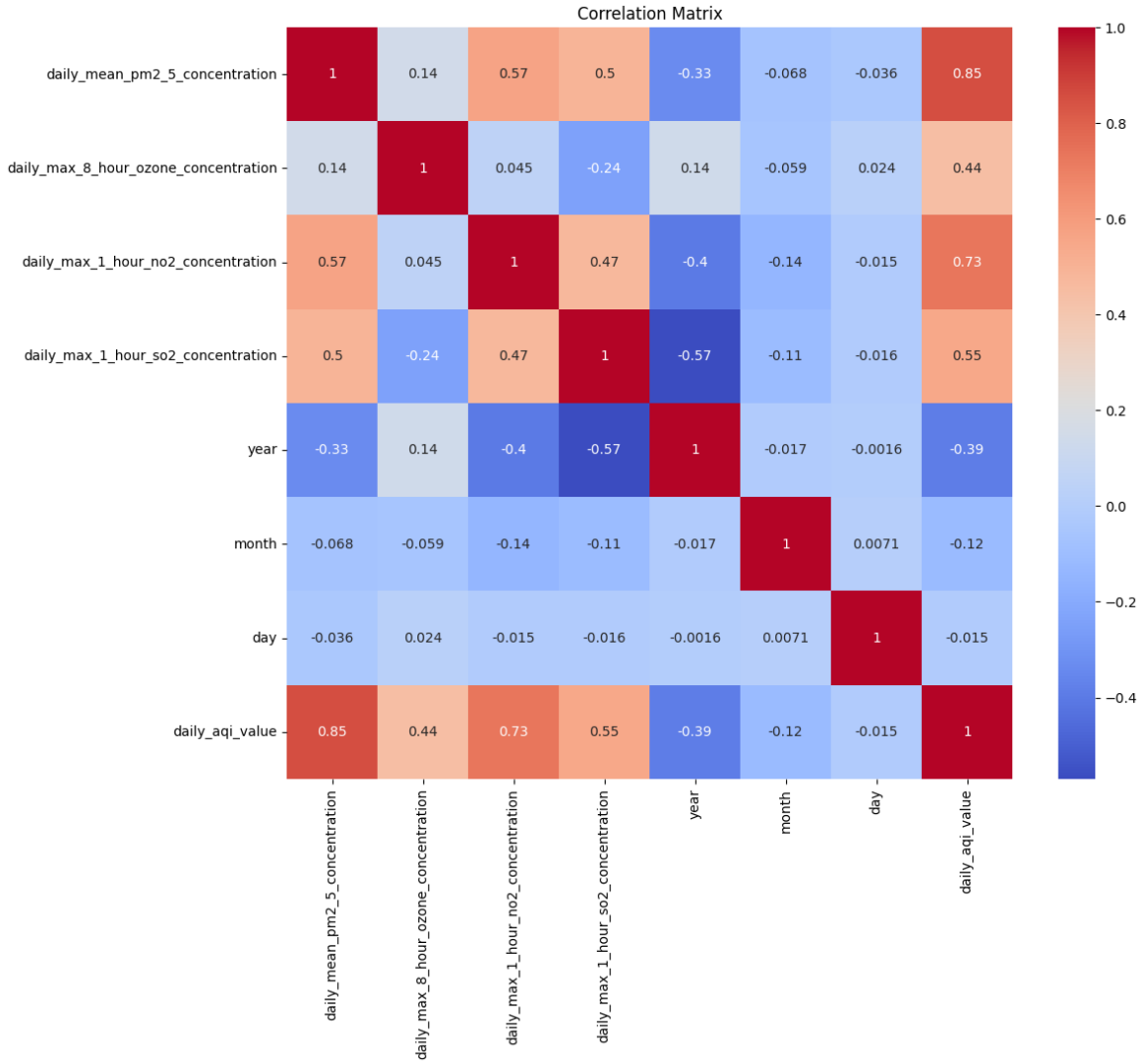

Figure 13 Correlation matrix of a cleaned data

6. CRISP-DM Methodology

Phase 1. **Project understanding**

With advancing education and understanding the impact that air pollution has on the health and wellbeing of its inhabitants, Air Quality should be improving in large cities.

We decided we would investigate the Air Quality in New York, a large city with a growing population. Steps have been taken over the last two decades by the city of New York to improve air quality.

Incentives included:

- Reduce City vehicle emissions.
- Reduce general transport emissions.
- Introduction of a congestion charge on the island of Manhattan.
- Gateless tolls are implemented to minimise idling engine pollution.
- Reduce emissions from buildings.

The project we have taken on is to review the data available for Air Quality and assess if New York is making meaningful progress towards its self-assigned goals. The press releases from NY Mayors office were showing data up to 2019, we would get the up-to-date data and assess if the trend was holding true. We would attempt a prediction model using machine learning.

Phase 2. **Data understanding**

We searched through the U.S. EPA website and the NYC.Gov website for data that might be linked to air quality. The NYC.Gov site had links to various datasets, some linking directly to sets in the EPA site. Initially, we looked at a traffic volume dataset with a plan to link it to an Air Quality dataset.

Phase 3. **Data preparation**

The initial traffic data set was large and had many data points and had information every five minutes. The Air Quality data set had information hourly, but we had difficulty matching the data to location to get accurate correlations. We parked the traffic data and concentrated on the air quality data. We found there was a lot of inconsistencies in the data, a lot of missing data and gaps of several years in some of the reporting locations. We decided then we would focus on one location that had the most consistent data on all the pollutants. This location was Queens College.

Phase 4. M**odelling**

Initially, we looked at a linear regression as we were dealing with a single continuous variable(2.5pm) but, since we had added other variables such as other pollutants. We selected the Random Forest model as we had several features that we could leverage to get the prediction we needed.

Phase 5: **Evaluation**

- The random forest model we applied was successful with R2 score 0.99
- Mean cross-validation score: 0.9917,

99% of the variability in the target variable is explained.

Phase 6: **Deployment**

The model is in deployment, we were able to make an accurate prediction:

- Actual AQI value for 2008/11/04 = 28.25
- Predicted AQI value for 2008/11/04 = 28.50

Next steps would be to merge this data with another set, possibly tailor the traffic data for the Queens College region and hypothesise if the traffic volume is the primary source of the pollutants measured.

7. Challenges encountered

7.1. Challenge 1: Datasets

The biggest challenge is finding the right dataset and deciding on the right target variable. Choosing the topic of air quality in New York was easy but getting and understanding what has been measured and the time stamps for these measurements took a lot of 'digging'.

In addition to the Air Quality dataset, Jim had started to work on a dataset about NY traffic. This dataset from the New York City Department of Transportation listed traffic sample volume counts at bridge crossings and roadways throughout NY.

We had planned to combine the Air and traffic datasets but the project was getting too complicated and we decided as a group to focus just on Air Quality.

When we started to use the Air Quality dataset we downloaded the datasets from the following link at the EPA website, https://www.epa.gov/outdoor-air-quality-data/download-daily-data

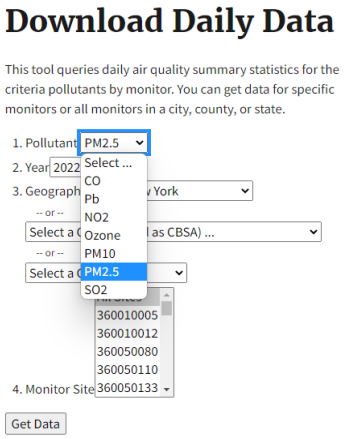

Figure 14 Interface of datasets source

We had to download datasets by year and by pollutant in excel format and then merge the excel files together to get our dataset. As we parsed the dataset and worked to 'frame' our question we had to re-visit the data source a number of times to get a solid dataset to work with and we found alternative ways to get data that was easier to use.

7.2. Challenge 2: Team members

One of the team members was unable to continue with the project. Another team member joined (Ciaran) and by this stage, the datasets were selected and work had started to progress. The challenge was getting the division of work. Time then had to be spent on getting new member up to date and this just slowed progress a little.

7.3. Challenge 3: Target variable and Machine Learning Model

Our first attempt was to predict an air quality index using a classification model. The model had very little data to work on as most of the pollutants were withing one index. After a thorough discussion, we decided to predict an air quality index value and use a regression model.

8. Results and analysis and next steps

The results we got from the model are as follows:

- R2 : 0.99378
- MAE: 0.4458
- Mean cross-validation score: 0.9917

The R2 result indicates 99% of the variability in the target variable is explained.

The MAE results indicate that on average the model is predicting AQI value within 0.4458 units of the actual value.

Mean cross-validation score tells us that the model generalises well to unseen data and that it is not overfitting.

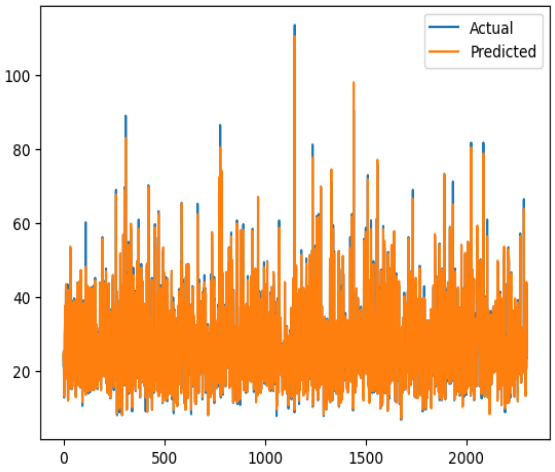

Figure 16 Actual vs Predicted Values plot

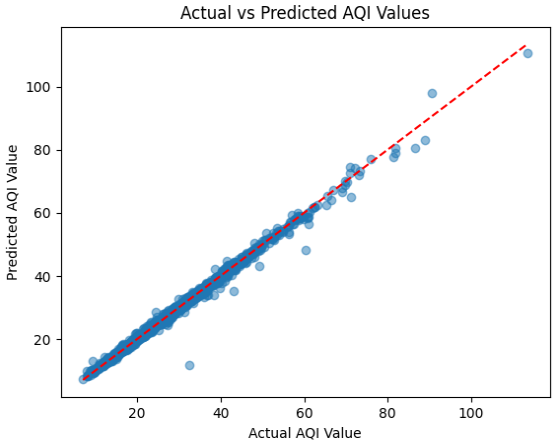

Figure 15 Actual vs Predicted values with a diagonal line for perfect prediction indication.

9. Conclusion

In conclusion, the goal of our project was to examine NYC air quality, particularly in the vicinity of Queens College. In order to forecast air quality levels, we acquired data from the Environmental Protection Agency and employed a supervised learning algorithm using a Random Forest Regressor model. An R2 score of 0.99378, which indicates that 99% of the variability in the target variable is explained, and a mean cross-validation score of 0.9917, which indicates that the model generalizes well to unseen data, respectively, indicating that the model is effective.

Finding the appropriate datasets and choosing the target variable for our model proved difficult for our team. We also had to change the way we went about things, concentrating entirely on air quality data rather than mixing it with traffic data. However, we were able to efficiently collaborate thanks to platforms like GitHub, Zoom, and Discord.

Our findings show that there has been progress in NYC's attempts to decrease air pollution, but more work needs to be done. In order to determine if the volume of traffic is the primary source of the raising air quality index value, our next steps will involve integrating our data with those from traffic volume.

10. References

NYC.Gov, 2016. *NYC.Gov.* [Online]
Available at: https://www.nyc.gov/assets/doh/downloads/pdf/eode/nyccas-6yrtrend-o3.pdf#:~:text=Warmer%20temperatures%20and%20increased%20daylight%20hours%20result%20in,with%20a%20slight%20decline%20from%202012%20to%202013.
[Accessed 06 May 2023].

NYC.gov, 2019. *NYC.gov.* [Online]
Available at: https://www.nyc.gov/site/doh/about/press/pr2019/health-department-releases-report-on-air-quality.page
[Accessed 05 May 2023].

NYC.Gov, 2023. *Environment and Health Data Portal.* [Online]
Available at: https://a816-dohbesp.nyc.gov/IndicatorPublic/beta/data-explorer/air-quality/
[Accessed 5 May 2023].

NYC, 2., 2018. *One New York: The Plan for a Strong and Just City.,* New York: City of New York Mayors Office.

U.S Environmental Protection Agency, Office of Research and Development, 2011. *EPA.Gov.* [Online]
Available at: https://www.epa.gov/sites/default/files/2013-12/documents/black-carbon-fact-sheet_0.pdf
[Accessed 06 May 2023].

U.S EPA, 2018. *EPA.Gov.* [Online]
Available at:
https://ofmpub.epa.gov/sor_internet/registry/termreg/searchandretrieve/glossariesandkeywordlists/search.do?details=&glossaryName=Acid%20Rain%20Glossary
[Accessed 06 MAy 2023].

USEPA, 2022. *EPA.Gov.* [Online]
Available at: https://www.epa.gov/pm-pollution/particulate-matter-pm-basics
[Accessed 06 May 2023].

Air Quality. 2022. NYC Office of Climate and Environmental Justice. [Online]
Available at: https://www.nyc.gov/site/sustainability/initiatives/air-quality.page.
[Accessed 06 May 2023].

"GitHub.com Help Documentation." Docs.github.com, docs.github.com/en.
Accessed 7 May 2023.

Oklahoma Department of Environmental Quality. (n.d.). Air Quality Index, Ozone Watches/Alerts, and Health Advisories. [online] Available at: https://www.deq.ok.gov/air-quality-division/ambient-monitoring/aqi-ozone-watches-alerts-and-health-advisories/. Accessed 7 May 2023.