



**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 |
| **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 |

**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

[1. Introduction 3](#_heading=h.1ci93xb)

[Why is air quality important? 3](#_heading=h.9dzoqn4rykqy)

[Pollutants monitored. 4](#_heading=h.3whwml4)

[General goal of this Project 10](#_heading=h.147n2zr)

[2. Technologies used 11](#_heading=h.3o7alnk)

[2.1. Models 11](#_heading=h.23ckvvd)

[2.2. Libraries 12](#_heading=h.ihv636)

[2.3. Machine Learning Algorithm (at least one implementation required!) 14](#_heading=h.32hioqz)

[2.4. Plotting and Visualisations 15](#_heading=h.q9ivldz1iw2h)

[2.5. Team Collaboration tools 16](#_heading=h.u9ih6ynqcf3a)

[2.6. Version Control – GIT 16](#_heading=h.hm3mqkeiwovp)

[3. What has been accomplished so far 18](#_heading=h.1hmsyys)

[3.1. Datasets 18](#_heading=h.41mghml)

[3.2. Source 18](#_heading=h.2grqrue)

[3.3. Attributes 18](#_heading=h.vx1227)

[3.4. Dimensions 18](#_heading=h.3fwokq0)

[3.5. Descriptive statistics 18](#_heading=h.1v1yuxt)

[3.6. Data visualisation 19](#_heading=h.4f1mdlm)

[3.7. Data preparation 19](#_heading=h.2u6wntf)

[3.8. Models 20](#_heading=h.19c6y18)

[4. Challenges encountered (including strategies used to overcome them) 21](#_heading=h.3tbugp1)

[5. Results and analysis and next steps 22](#_heading=h.28h4qwu)

[6. Conclusion 24](#_heading=h.nmf14n)

[7. References 25](#_heading=h.4i7ojhp)

[**References 25**](#_heading=)

# 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 implemented to minimise idling engine pollution.
* Reduce emissions from buildings. In 2015 the City stopped issuing permits for use of #6 oil as a heating oil.

## Why is air quality important?

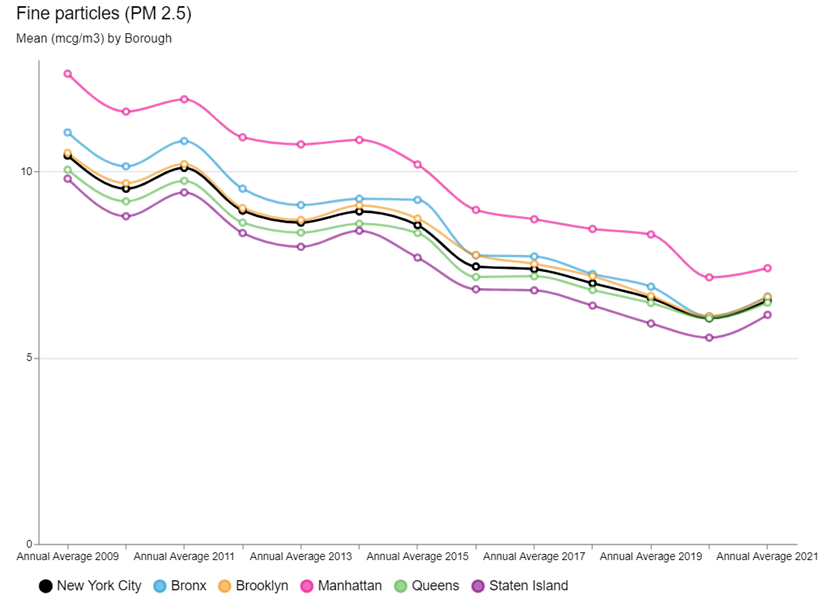
Measures taken by cities to improve Air Quality not only reduce greenhouse gas emissions (good for the planet!) but provide cleaner, healthier air for every resident. Air pollution is a huge threat to public health and it is estimated that particulate matter (PM2.5) contributes to more than 2,000 deaths and over 6,000 emergency visits and hospitalizations for cardiovascular and respiratory disease each year. (*Air Quality*, 2022)

# 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.”

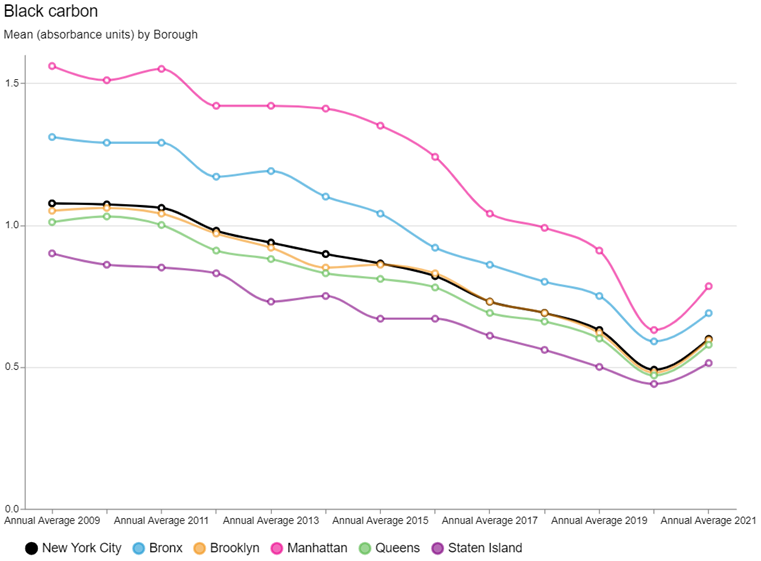
More up to date data shows they have continued to fall with slight increases in PM2.5 and Black Carbon after the NYC Pause due to Covid-19 (NYC.Gov, 2023)

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)



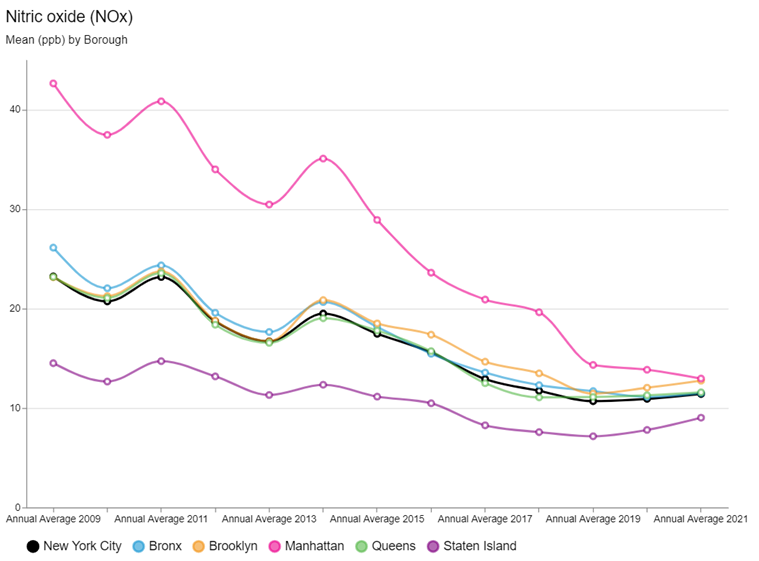
*Figure 1 PM2.5 (NYC.Gov, 2023)*

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)



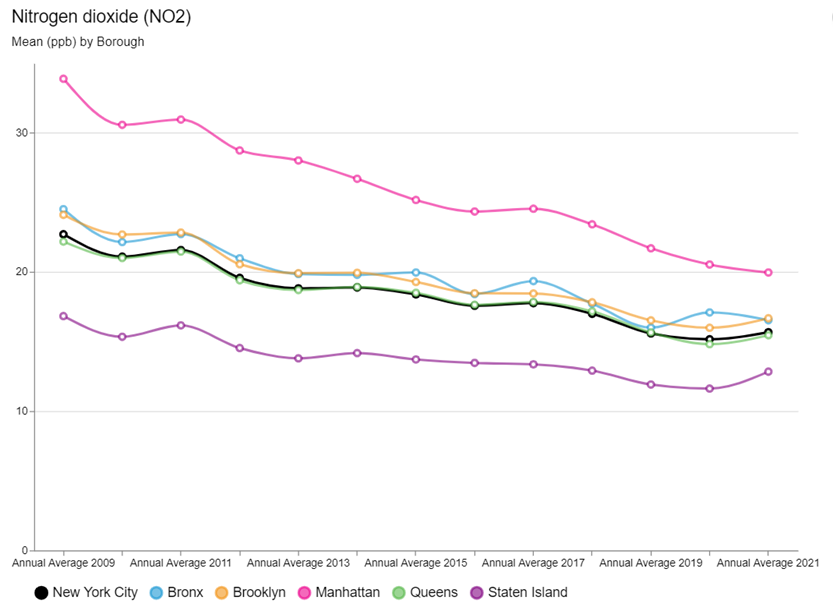
*Figure 2 Black Carbon (NYC.Gov, 2023)*

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)*

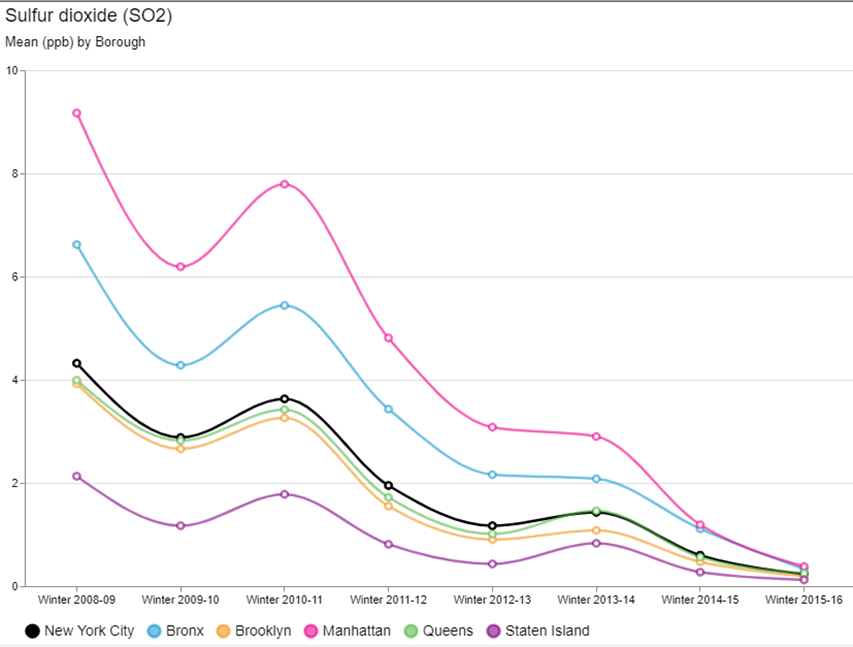
Nitrogen Dioxide (NO2) A reactive gas which is present in urban atmospheres. This gas is formed in the atmosphere from emissions of oxides of nitrogen (NOX). As discussed previously NOX is produced by fuel combustion sources, vehicles and industrial boilers. (U.S EPA, 2018)



*Figure 4 Nitrogen Dioxide (NYC.Gov, 2023)*

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)*

Ozone is formed in the atmosphere through the reaction of other pollutants (oxides of nitrogen and volatile organic compounds) in the presence of sunlight.

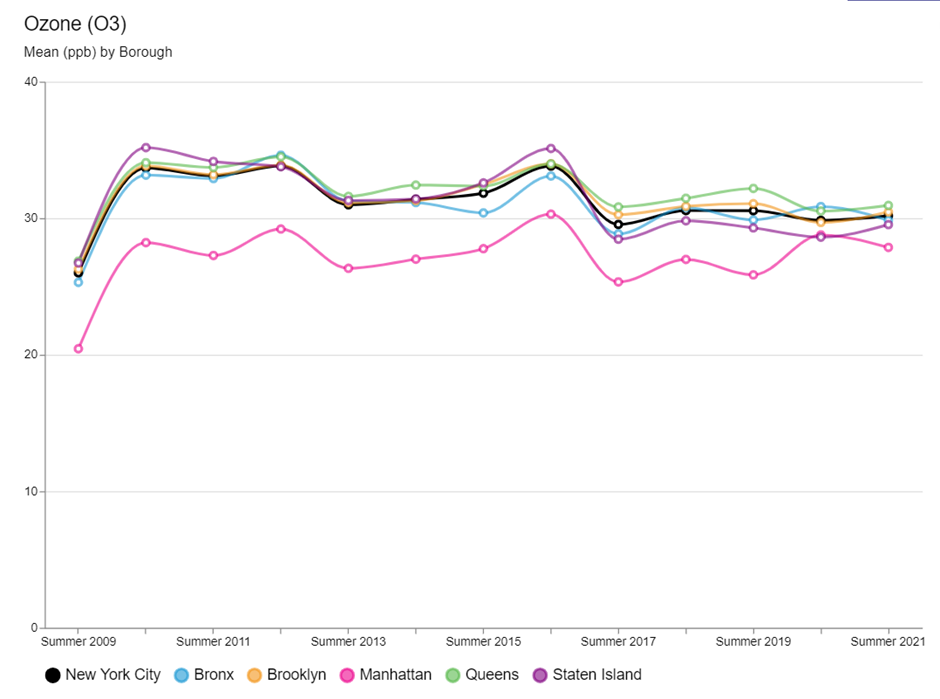
Warmer temperatures and increased daylight hours result in increased ozone production. (U.S EPA, 2018)

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)*

# 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 next steps to improve on the model as knowledge of the programming language improves.

# Technologies used

## Libraries

Python has heaps of open-source libraries and we included a number of libraries so we have access to all sorts of functions and features that can help us to work and manipulate the datasets we planned to analyse.

The list of libraries that we included and what functions we used are as follows:

| LIBRARY/FUNCTIONS USED | DESCRIPTION | HOW WE ARE USING IT |
| --- | --- | --- |
| Pandas | Pandas is a well-known library used for analysis, manipulation, and cleaning of data | We use the dataframe (df) functions to import and store our dataset and to view, manipulate and sort the contents. |
| NumPy | NumPy (numerical python)makes it easy to work with large matrices and multidimensional data | We used the np functions to declare and initialise matrices. |
| matplotlib   * .pyplot as plt * .dates as mdates | The plotting of numerical data is the responsibility of this library. | We plotted the ‘Scatter Plot of Daily Mean PM2.5 Concentration’ using this library by feeding in the dataframe containing ‘dates’ and ‘Daily Mean PM2.5 Concentration’ values. |
| Seaborn   * .barplot * .heatmap | This library is used for the visualisation of statistical models. | We used the barplot function to map air quality by year. We use the ‘Correlation Matrix’ function. |
| Scikit-learn   * sklearn.model\_selection import train\_test\_split, cross\_val\_score * sklearn.ensemble import RandomForestRegressor | Scikit is a famous Python library that works with complex data. | We used several functions to build our RandomForestRegressor model and to create the train\_test\_split and cross\_val\_score value. |
| warnings | Used to suppress warnings in the notebook. | We used this to make it easier to run the notebook without constant error and warnings. |
| datetime   * datetime | The Datetime library contains classes for manipulating dates and times. | Used this in merging date and time values for our measurements. |

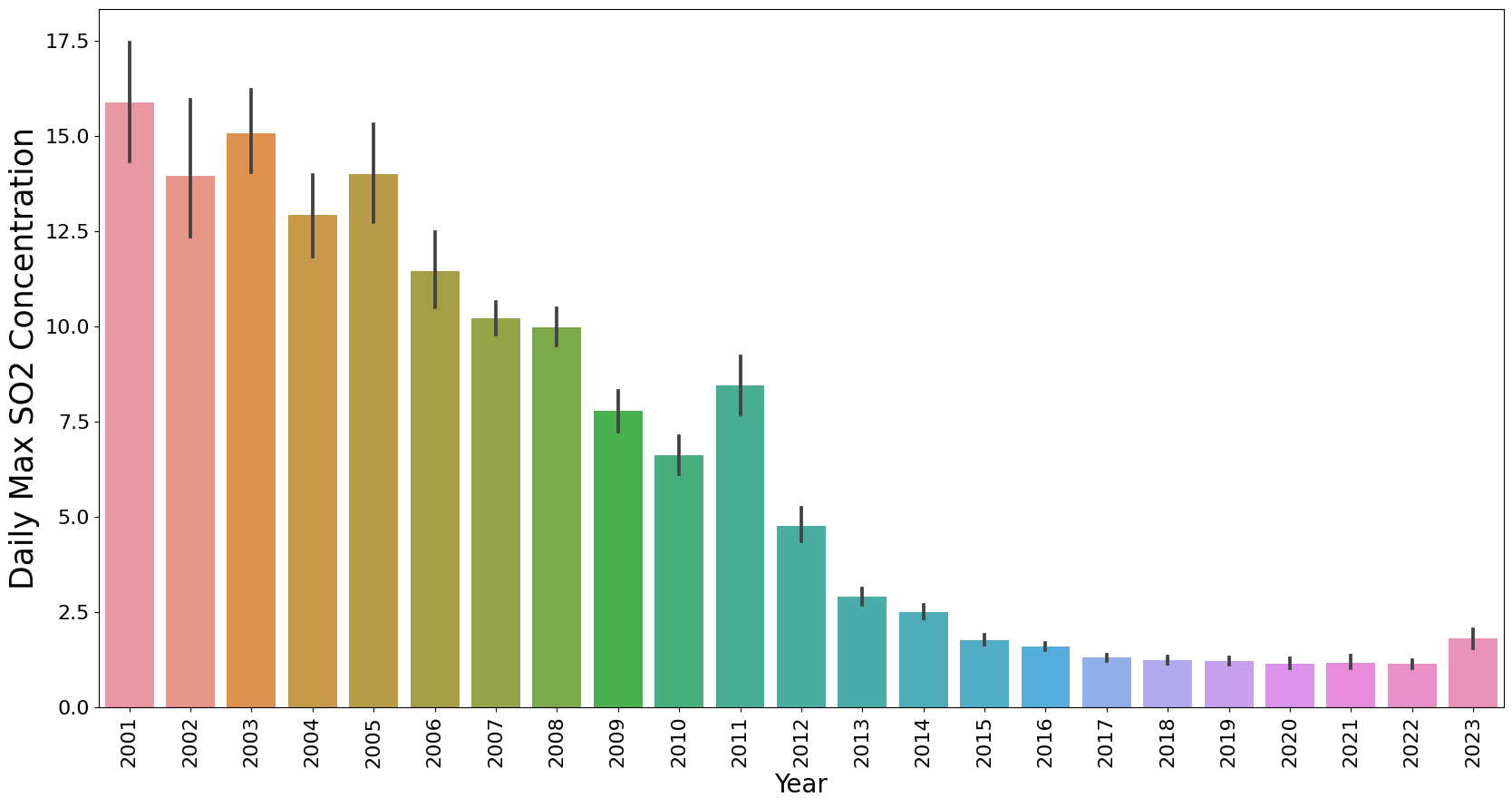
## Machine Learning Algorithm (at least one implementation required!)

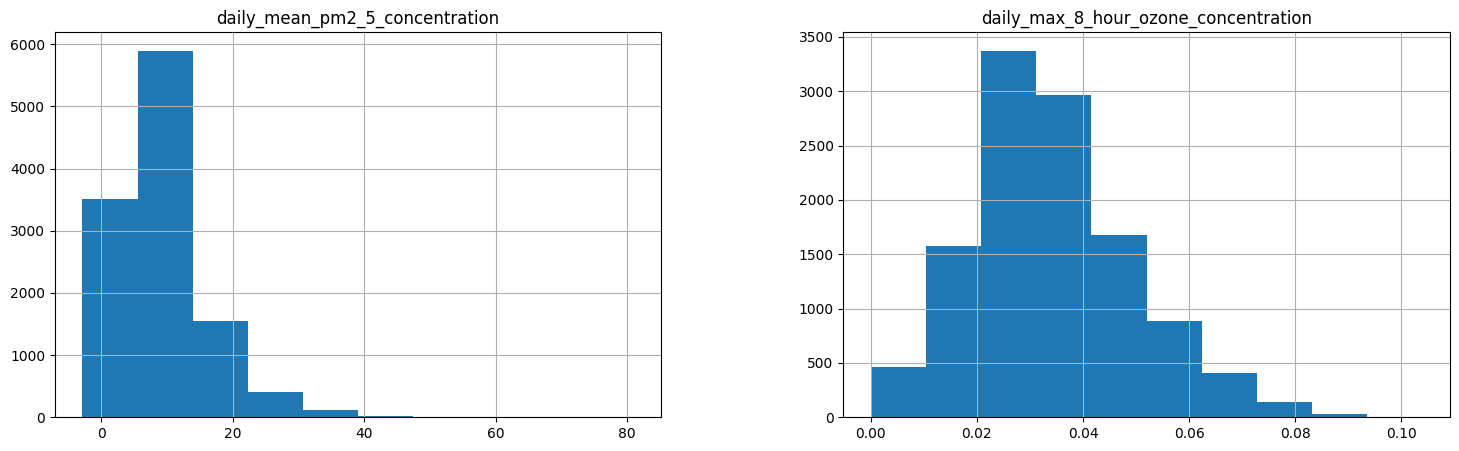
For predicting continuous numerical values and where the data is labelled we have decided to use a supervised learning algorithm with a Random Forest Regressor model.

During the project and while we got to understand the characteristics of the dataset we reviewed what would be the best ML models that we could use with the data.

## Plotting and Visualisations

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





## 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.

As we progressed and the deadline loomed closer 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.

## 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.   
Git is a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency, in summary and we are ready to start to use it in the next phase of the CA.

The way we used GIT was as follows:

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:

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

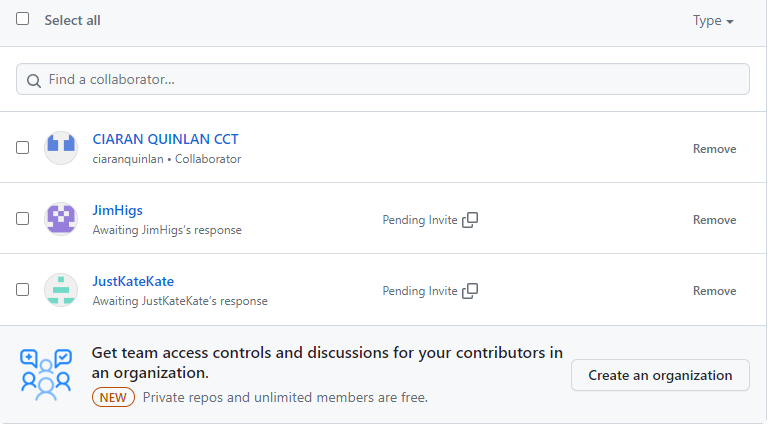
For the team members to get access to the repo the following steps were followed.

1. Go to the git account that owns the repository, ciaranq
2. Select settings, collaborators, Enter the emails of the collaborators and send an invite to @ciaranquinlan, @JimHigs, @JustKateKate.
3. 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.

This is the screen shot inviting members to collaborate on the repository.

# What has been accomplished so far

## Datasets

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

## Source

Reliable source of United States Environmental Protection Agency (<https://www.epa.gov/outdoor-air-quality-data/air-data-concentration-plot>)

## CRISP-DM Methodology

The Cross Industry Standard Process for Data Mining CRISP-DM is a process model that serves as the base for a data science process. We have followed the early phases of the CRISP-DM methodology so far in the project as we selected and prepared our dataset. Data Science Process Alliance. 2023. What is the Data Science Process?

CRISP-DM has 6 phases , We have spent most of CA1 on phases 1-3.

Phase 1. **Business understanding** – What does the business need?  
In our case we did not have a business need but we did Identify a project idea and defined a number of project goals. During our first ‘sprint’ call we developed a high-level project plan and the roles each person would focus on, Jim evaluated the traffic dataset, Kate evaluated the Air Quality dataset and Ciaran worked on the documentation and the project collaboration tools to support the team

Phase 2. **Data understanding** – What data do we have / need? Is it clean?  
We explored ways to get the data, for example we used Kaggle to get the public NY traffic dataset and went directly to the EPA website for the the air quality. These were both very reliable sources of data which we could use at the data at no cost and and comply with the licence to use the data.

Phase 3. **Data preparation** – How do we organize the data for modeling?  
We had a team call when we had all viewed and explored the datasets. The Traffic dataset was >3GB and we discussed ways to use such a large dataset. When we had explored the Air Quality dataset we focused down on PM 2.5 for a number of location but we soon realised that we needed to expand the number of columns of data (ie add more pollutants) to use this dataset with an ML model  
As we developed our datasets for ML modeling we did a number of data cleaning exercises and constructed new datasets to achieve our project goals. Throughout the data preparation phase we met via discord a number of times to share and discuss the datasets and to refine the project goals and plan.   
It is said that over 80% of the CRISP-DM process can be spent on phases 2 and 3 and certainly we have spent the bulk of this project on these phases.

Phase 4. **Modeling** – What modeling techniques should we apply?  
We are just getting to understand how we are going to work through this phase and while we have selected some models to work on with our datasets and we will be spending more time on these phases as we start on CA2.

Phase 5: **Evaluation** – Which model best meets the business objectives?  
We were able to evaluate some aspects of our dataset but more work needs to be completed on the coding aspects of the project.

Phase 6: **Deployment** – How do stakeholders access the results?

## Dimensions

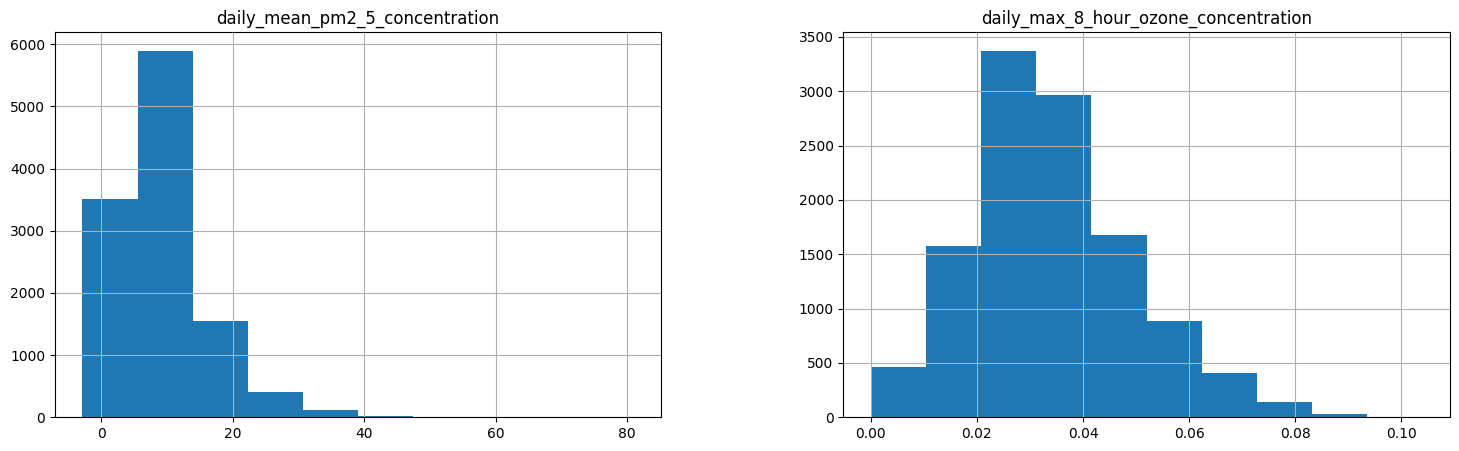
To merge the datasets successfully we had to use specific columns when reading the csv files. After thorough investigation of the four datasets, we decided to use the values of the pollutants, units the pollutants were measured in, date, daily aqi value and site latitudes/longitudes.

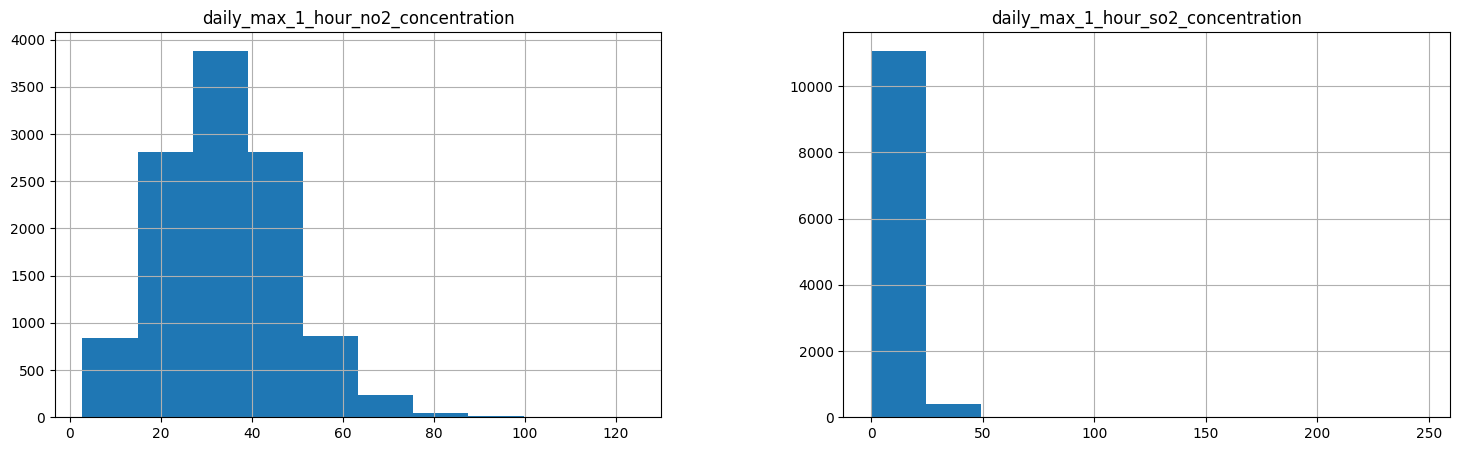
We merged the datasets on date to ensure dates matched.

The dimension of the dataset came out to be 11,637 rows with 15 columns before data cleaning.

## 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.





## 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.

## 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 lower case 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.

A picture containing text, device

Description automatically generated

https://www.deq.ok.gov/air-quality-division/ambient-monitoring/aqi-ozone-watches-alerts-and-health-advisories/

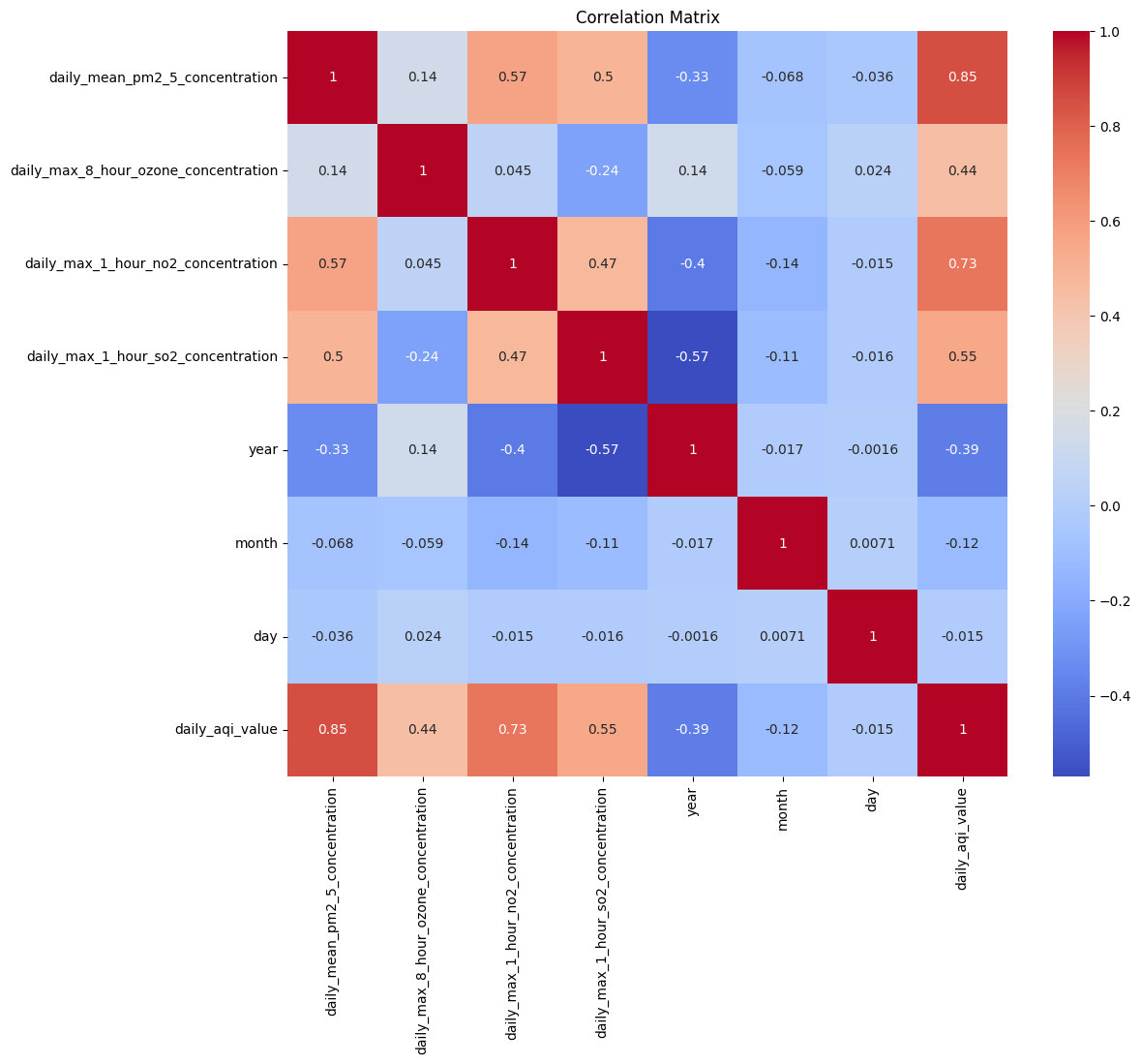
We also had to check for duplicates, missing values, and incorrect data.

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

## Models

Before 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 me 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.



# Challenges encountered (including strategies used to overcome them)

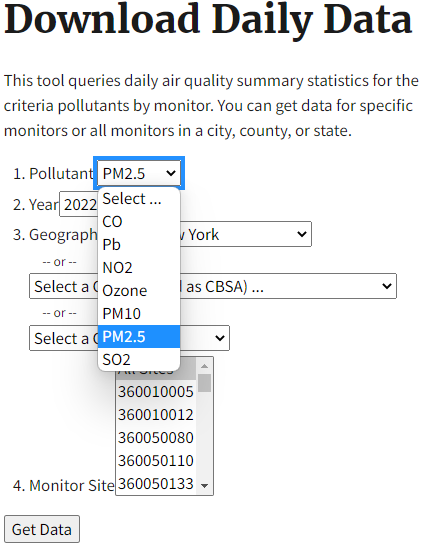
## Challenge 1: Datasets

The biggest challenge is finding the right dataset and framing the question to best use the dataset. 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 for the years 2010-2016. This dataset from the New York City Department of Transportation listed traffic sample volume counts at bridge crossings and roadways throughout NY. We also looked at some other dataset but they were also ruled out.

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>, using this interface



The website was not that user friendly in downloading data as we had to download 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.

## Challenge 2: Team members

One of the team members was unable to continue with the project and some time had been spent on getting this member to contribute. 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.

Challenge 2: selecting a ML Model

Having decided on the dataset and spending so much time cleaning the dataset it started to dawn on us that not all ML models would suit our dataset. We ruled out linear regression due to our selection of concentrating on the PM2.5 variable and KNN (K-Nearest Neighbor Regression) was also not an option based on our dataset findings.

# Results and analysis and next steps

(**Analysis of results and conclusions are sensible.)**

Results we got from the model were are follows:

R2 : 0.99378

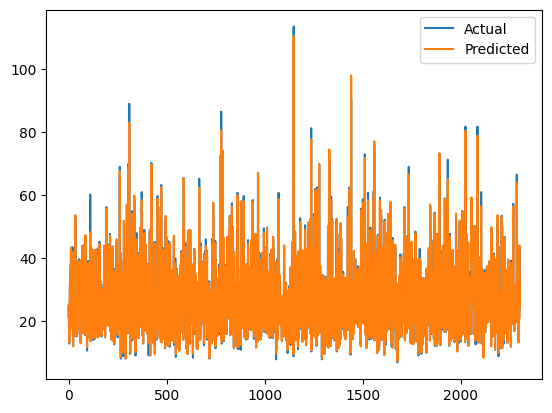
MAE: 0.4458

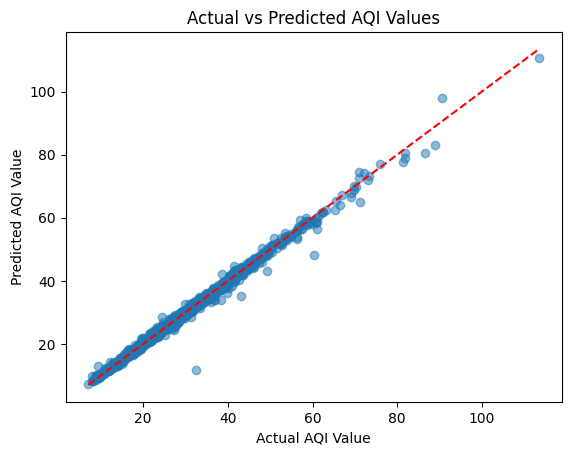
Mean cross-validation score: 0.9917

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.





# Conclusion

As per the CRISP DM methodology we worked from, we have deployed our code and it works well. We will come back to business understanding in the next step as we still want to add other features to the model. And tune in hyperparameters. Compare results to other regression models.

Diagram

Description automatically generated

# References

(***Referencing and citations have been done correctly.)***

# References

Anon., 2023. *Air Pollution & Regulations.* s.l.:City of New York..

Anon., n.d. *Libraries in Python.* s.l.:s.n.

Bruinius, H., 2013. *Cleanest air in 50 years! How did New York do it?.* s.l.:The Christian Science Monitor..

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].

How do I cite data and graphics obtained from the AirData website. US Environmental Protection Agency. Air Quality System Data Mart [internet database] .   
Available at: https://www.epa.gov/outdoor-air-quality-data   
[Accessed: May 7, 2023].

Data Science Process Alliance. 2023. What is the Data Science Process? [Online]   
Available at: https://www.datascience-pm.com/data-science-process/  
[Accessed 06 May 2023].