



**CCT College Dublin**

**Assessment Cover Page**

| **Module Title:** | Strategic Thinking CA2 |
| --- | --- |
| **Assessment Title:** | CA2 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:** | 14/11/2023 |
| **Date of Submission:** | 14/11/2023 |

**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 3](#_heading=h.gjdgxs)

[Table of Figures 5](#_heading=h.39kk8xu)

[1. Introduction 6](#_heading=h.30j0zll)

[2. Pollutants monitored 7](#_heading=h.1fob9te)

[2.1. Fine Particulate Matter (PM2.5) 7](#_heading=h.3znysh7)

[2.2. Black Carbon 8](#_heading=h.17dp8vu)

[2.3. Nitric Oxide (NOx) 9](#_heading=h.3rdcrjn)

[2.4. Nitrogen Dioxide (NO2) 10](#_heading=h.26in1rg)

[2.5. Sulphur Dioxide (SO2) 11](#_heading=h.lnxbz9)

[2.6. Ozone 12](#_heading=h.35nkun2)

[3. General goal of this Project 13](#_heading=h.44sinio)

[4. Technologies used in Semester 1 13](#_heading=h.z337ya)

[4.1. Libraries 13](#_heading=h.3j2qqm3)

[4.2 Models 13](#_heading=h.1y810tw)

[5. What has been accomplished so far for NYC pollutants dataset 14](#_heading=h.2xcytpi)

[5.1. Datasets 14](#_heading=h.2bn6wsx)

[5.2. Source 14](#_heading=h.qsh70q)

[5.3. Dimensions 14](#_heading=h.3as4poj)

[5.4. Descriptive statistics 14](#_heading=h.1pxezwc)

[5.5. Data visualisation 15](#_heading=h.49x2ik5)

[5.6. Data preparation 15](#_heading=h.2p2csry)

[6 Machine Learning Algorithms 16](#_heading=h.xnsqryluldj8)

[6.1 Random Forest Regressor with default hyperparameters 17](#_heading=h.vx1227)

[6.2 GridSearchCv for best hyperparameters 17](#_heading=h.37m2jsg)

[6.3 Random Forest Regressor with tuned hyperparameters 17](#_heading=h.46r0co2)

[6.4 KNeighbors Regressor 17](#_heading=h.111kx3o)

[6.5 Decision Tree Regressor 18](#_heading=h.3l18frh)

[7. Project Management Collaboration Tools 19](#_heading=h.206ipza)

[7.1 Version Control – GIT 21](#_heading=h.4k668n3)

[7.2 So what is GIT. 22](#_heading=h.2zbgiuw)

[Git for our CA 22](#_heading=h.1egqt2p)

[8. Hospital data 28](#_heading=h.2iq8gzs)

[9. CRISP-DM Methodology 33](#_heading=h.xvir7l)

[10. Challenges encountered. 35](#_heading=h.3hv69ve)

[10.1 Challenge 1: Datasets 35](#_heading=h.1x0gk37)

[10.2 Challenge 2: Team members 35](#_heading=h.4h042r0)

[10.3 Challenge 3: Target variable and Machine Learning Model 35](#_heading=h.2w5ecyt)

[11. Analysis of the machine learning model results 36](#_heading=h.1baon6m)

[12. Next Steps 37](#_heading=h.3vac5uf)

[13. Conclusion 38](#_heading=h.2afmg28)

[14. References 39](#_heading=h.pkwqa1)

### Abstract

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
* Make a prediction using machine learning algorithms for the Air Quality Index value, improving on previous findings through cross validation.
* Access the New York Health data and determine if there has been any trend in respiratory illness that may be attributed to Air Quality Index trends.

Summarising from the previous semester we managed to make a prediction based on the AQI. We plan to explore other machine learning algorithms and use cross validation techniques in an attempt to improve on this result. In tandem to this we decided to look at other data available in the public domain and see if there was anything that we could build a story to tell.

# The site: ‘<https://health.data.ny.gov/>’ contains hospital discharge data for all of New York state broken down by facility, county and diagnosis among other variables. We took the years 2016 to 2021 inclusive to run a comparison on AQI versus any respiratory illness. The data is very large on the health side and would be a substantial standalone project to attempt to merge and analyse the full set. A graphical comparison will be made and inferences thereof.

# 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.

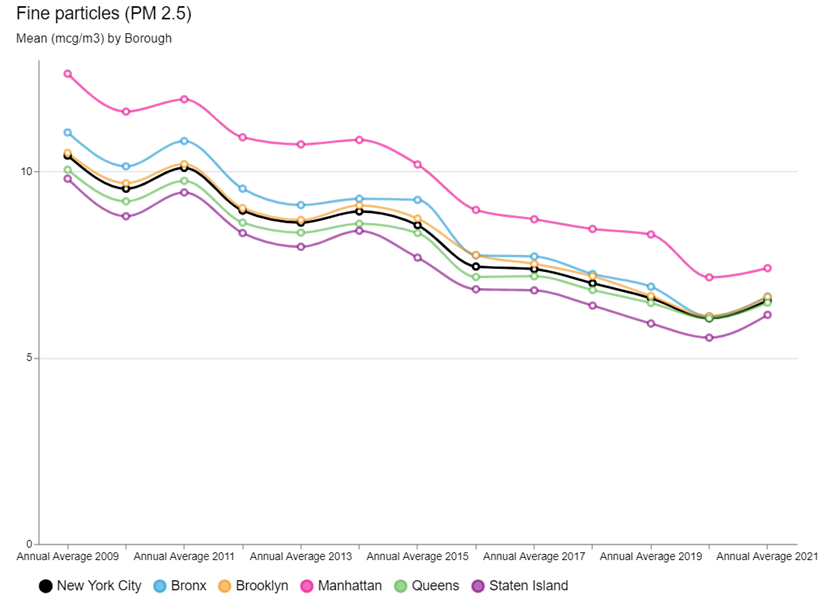
# Pollutants monitored

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)

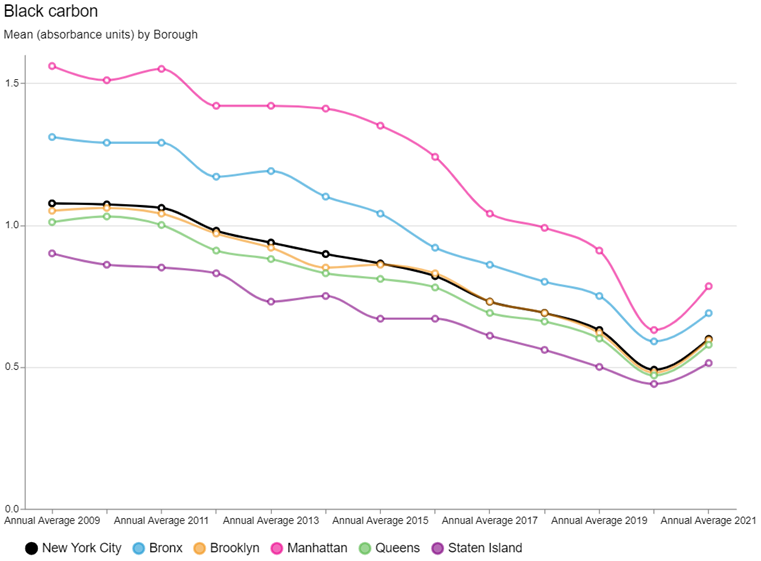
Fine particulate matter consists of small, airborne particles with a diameter of 2.5 micrometres or less. These particles form in the atmosphere because of interactions with chemicals such as sulphur dioxide and nitrogen oxides, which are emitted from power plants, industries, and automobiles. (USEPA,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



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

## Black Carbon

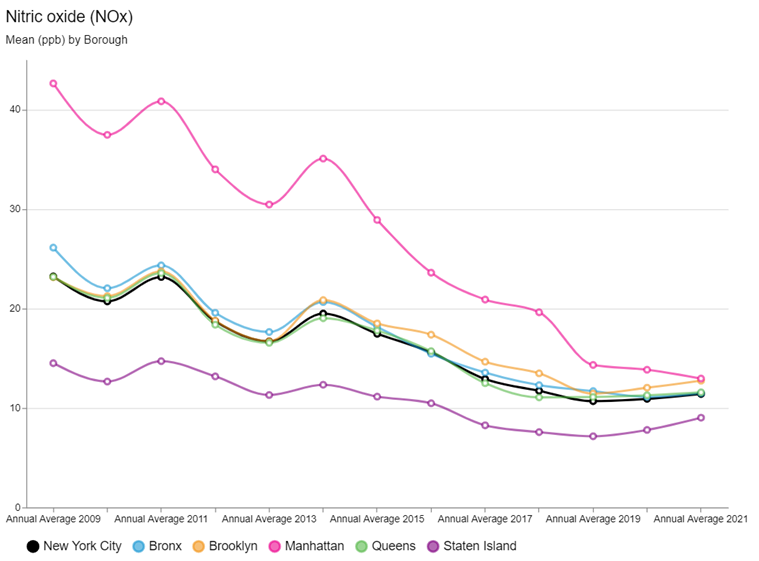
Black carbon is the soot like material emitted from combustion engines, coal-fired power plants, and other sources that burn fossil fuel. It comprises a significant portion of particulate matter. (USEPA, 2022)



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

## Nitric Oxide (NOx)

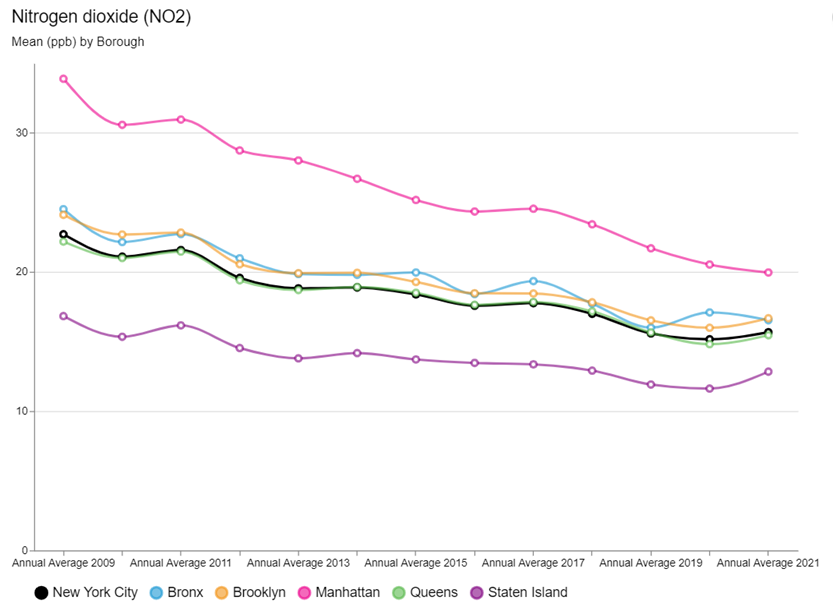
A gas formed by combustion under high temperature and high pressure in an internal combustion engine; it is converted by sunlight and photochemical processes in ambient air to nitrogen oxide which is a precursor of ground-level ozone pollution, or smog. (U.S EPA, 2018)



*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)

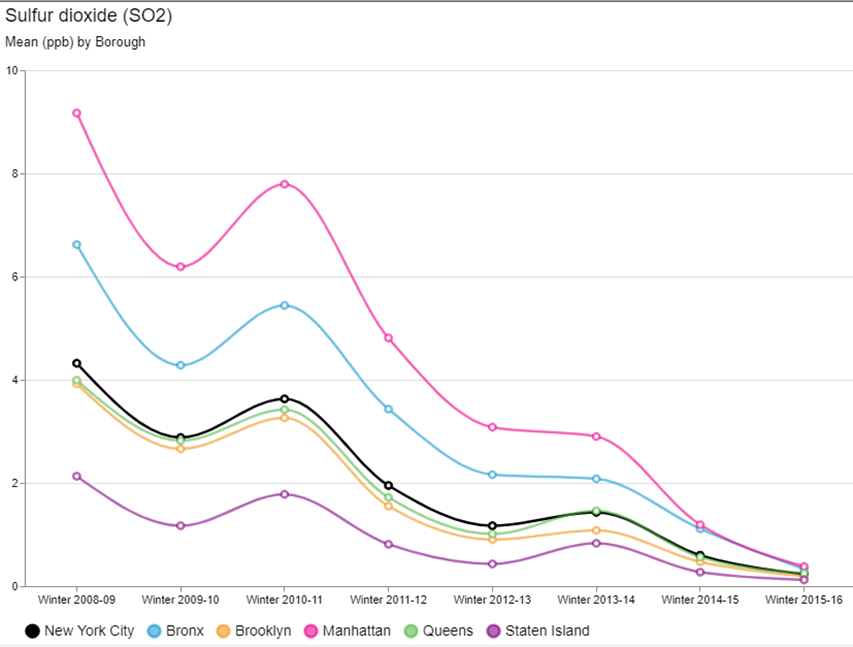


*Figure 4 Nitrogen Dioxide (NYC.Gov, 2023)*

## 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

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

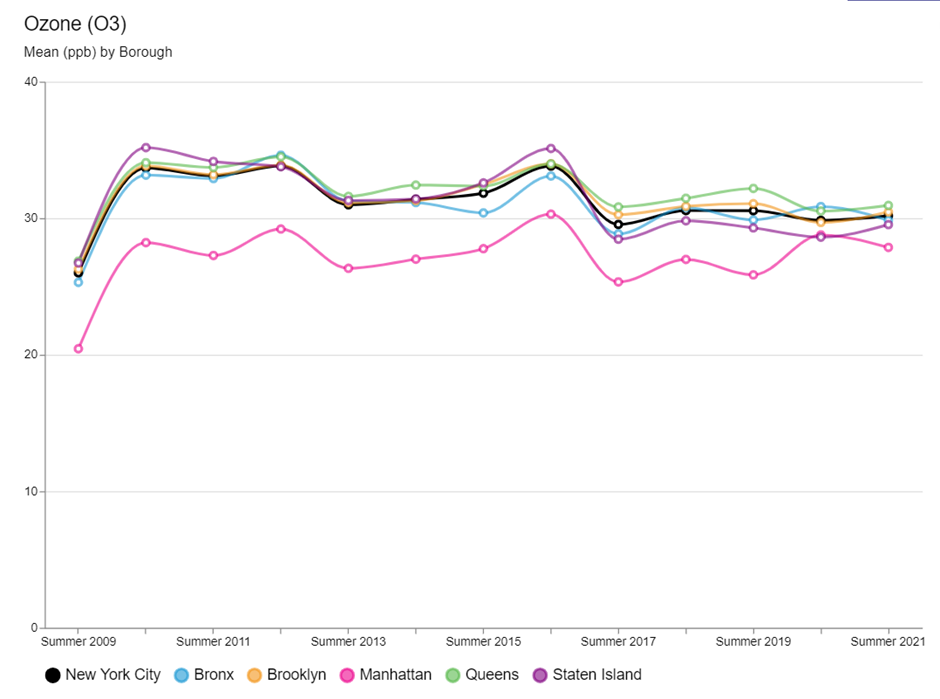
Warmer temperatures and increased daylight hours result in increased ozone production. (U.S EPA, 2018)

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 algorithms of the Air Quality Value
* Tune in hyperparameters
* Access the New York Health data and determine if there has been any trend in respiratory illness that may be attributed to Air Quality Index trends.

# Technologies used

## Libraries

We used a Jupyter Notebook with Python as the coding language for this project and utilised 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

## Models

To predict AQI values, we used a combination of Random Forest Regressor with GridSearchCV for hyperparameter tuning, KNN, and Decision Trees models. These models are well-suited for this task due to their ability to handle non-linear relationships. Random Forest, with its ensemble approach is effective in reducing overfitting and increasing accuracy, KNN is beneficial for capturing local patterns and Decision Trees offer interpretability.

# What has been accomplished so far for NYC pollutants dataset

## Datasets

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

## Source

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

## 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.

## 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.

A graph of two people

Description automatically generated

*Figure 10 Daily mean PM 2.5 and daily max 8 hour ozone concentration outliners*A graph of a bar and a bar of a graph

Description automatically generated with medium confidence

## 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.

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

A graph of different colored bars

Description automatically generated

*Figure 7 Daily Max SO2 Concentration vs Years*

## 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 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.

# Machine Learning Algorithms

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.

A screenshot of a computer

Description automatically generated

*Figure 13 Correlation matrix of a cleaned data*

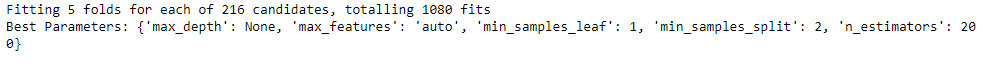
## Random Forest Regressor with default hyperparameters

A close-up of a number

Description automatically generated

The cross-validation scores are high, with a mean above 0.9, indicating consistency. The R-squared value is close to 1, showing models predictions closely match the observed data. The errors are

## GridSearchCv for best hyperparameters



## Random Forest Regressor with tuned hyperparameters

A screenshot of a computer

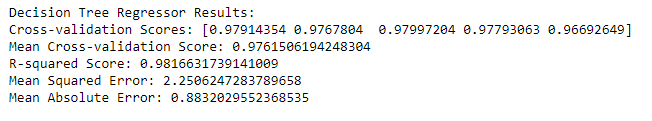
Description automatically generated

## KNeighbors Regressor

A number on a white background

Description automatically generated

## Decision Tree Regressor





# Project Management Collaboration Tools



Taking a quote from the famous basketballer Michael Jordan illustrates that to win we need teamwork. To make a 'Team Work' you have to have the right tools to effectively facilitate the team and manage the project with the minimum of distraction.

There is no end to the amount of applications that will ‘power’ a team to greater heights and picking the right mix for our team evolved rather than being dictated from any team member.

All three team members are from very different backgrounds and some are very new to the collaboration tools we needed to use (for example git) and some were already 'indoctrinated' by the tools they have spent a working lifetime with. We were also guided and excited to use some new tools that are new to the project management landscape.

In essence we needed project collaboration tools that could simplify our teamwork tasks with the following characteristics:

* online with web interface
* easy to use and familiar
* fast to use with minimal onboarding
* store files with version control
* meets the requirements of the Assessment

The following table outlines the project management collaboration tools we selected grouped into the following functionality

1. Real-time collaboration
2. Instant messaging
3. Video conferencing
4. Task management
5. File transfer and sharing
6. Version control
7. Project monitoring and management

| **Functionality** | **Selected Tool** | **Details and Examples of use** |
| --- | --- | --- |
| Real-time collaboration:  sharing and editing documents on the fly with the 3 team members | Google Documents | 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.  Writing the report we would all view the shared document and collaborate together as edits were being done and agree what was missing, need clarification etc. |
| Instant messaging | Whatsapp, mobile and desktop. | Whatsapp was our go to for daily updates and status updates and to be honest to get us all focussed on the outstanding tasks.  We started out the project with a whats app group for the 3 members very early in the project.    The whatsapp group kept us in contact from its inception was great for passing back and forth messages and getting to know each other's working style and availability. |
| Video conferencing | Zoom and Discord | We started out with zoom, staying online after Strategic Thinking class ended in a meeting room.  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.  Kate suggested we move to discord and we setup a channel on <https://discord.com/channels/>.    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.  It was exciting to use discord in this way and our code reviews were easily done as a team with the screen sharing features. |
| Task management | Whatsapp, discord | We did not select a formal PM tool to manage our tasks but logged them in discord after our team meetings.  During the meets ups we recorded what were the next steps by each team member.  Furthermore a lot of tasks were done together as a team when we held code reviews. For example setting up GIT and discussing datasets were all dont collaboratively when we were joined together on discord. |
| File transfer and sharing | Google Drive | In the initial stages we used google drive as our shared file tool. As the team got more comfortable with Github we migrated all files to our git repository and google drive became less important. |
| Version control | Git, Github | We started our Git repository with an initial commit on 2nd May 2023.    It was not until CA2 we started to use it effectively as a team and all members got the hang of it very quickly. See the action on Git below. |
| Project monitoring and management | Whatsapp, discord, Git | As our project evolved we used Whatsapp, discord & Git to manage the project and successfully completed the tasks with these tools. |

We believe that the project management collaboration tools we chose and used really helped our teamwork and helped us to effectively deliver the CA project. Our choice of tools helped

* to solve our challenge
* make it easy and be responsive for the team
* meet our practical needs(ie no custom install of software)
* meet the assessment requirements

## Version Control – GIT

Part of the requirements for the assessment is to use GIT.

## So what is GIT.

Git is a tool used for source code management. It is free and open-source. Some major companies have developed applications based on GIT (ie Altassian) but the free version is still the most used.   
In fact stackoverflow have claimed that GIT has won the race for the most popular version control system and they claim that the branches feature is the reason for it overwhelming success.

Git was created in 2005 by Linus Torvalds when he was working on the development of the Linux kernel. Linus is legendary in the world of Linux, to read more about this incredible man visit his wiki page here <https://en.wikipedia.org/wiki/Linus_Torvalds>.   
Linus needed something to manage the development of the then nascent operating system which had people all over the world contributing and decided that he would build his own ‘source-control management (SCM) system’ to manage all the developers who wanted to contribute to Linux. The story goes that Linus (who called Linux after himself) called it ‘git’ to be sarcastic and refer to himself as an unpleasant person. I suppose you need to have a finnish sense of humour to get it but the unix world has a great history of funny name for commands, for example GNU.

Put simply Git is used to track changes in computer source code, enabling multiple developers work together on development.

Git allows

* Every developer to have a copy of the code on their local system
* Any changes that are made to the source code can be tracked by others
* Regular communication between developers
* Rules/mechanisms to allow source code changes proceed to the main/sub branch

### Git for our CA

Our use of Git and GitHub was to manage the Jupyter notebooks and the datasets for our CA. We did upload our code as the project started to take shape for CA1 and did some commits at that time. A lot of work was done in the initial stages of CA1 that we are still using today in CA2 but we did not have the level of commits back then.

### Git setup for the team

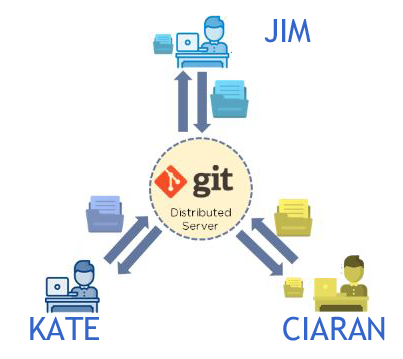
The way we setup GIT was as follows:

* We created a blank repo (the cool way to say repository) in github.com under the personal account of ciaranq. The Git repo called ny-air-quality was made available @ the git url of ciaranq/ny-air-quality. (this account is separate to my CCT git account)
* We assigned this a Public access setting and got the URL of this repository.
* I (ciaran) opened up our local GitHub Desktop application and used our CCT Git login that we created in Class with James in the GIT tutorial
* To access the repository, I selected File/Clone repository, enter https://github.com/ciaranq/ny-air-quality.git in the URL option
* I 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,

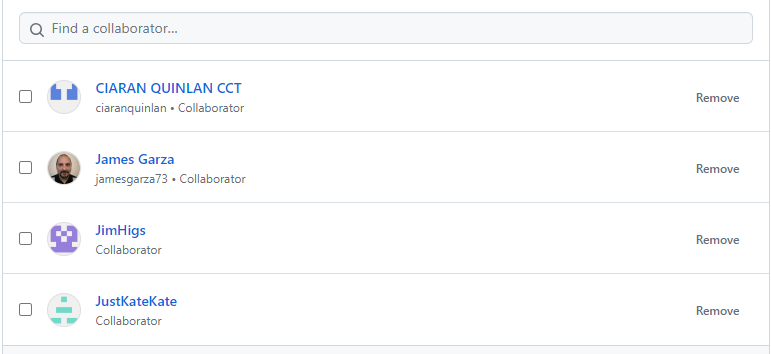
#### Adding collaborators

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

* Go to the git account that owns the repository, ciaranq
* Select settings, collaborators, Enter the emails of the collaborators and send an invite to @ciaranquinlan, @JimHigs, @JustKateKate.
* The team members get an email and are asked to accept the invite, this is done by logging into their git account or their github desktop and accepting the invite.
* In github the collaborators (team members) can then select a local folder to store the repository and where they will run the notebooks from. This starts the cloning process of the repo to the local directory.
* When We all accepted the invites out Git looked like this



#### Adding James as a collaborator

As our CA 2 was progressing we invited James, our esteemed lecturer to join our GIt as a collaborator. James thankfully accepted and now our Git setup looks like this:

#### Using Git

We had a number of files ready for upload when we all joined Git and after the initial commit we started to push up some files and get the hang of running code on the local machine and seeing the changes we made in Github Desktop and commit the changes to the main branch

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

* Make whatever changes to the local file and add a reference for the commit in the commit message dialog
* Select fetch/pull request from the repository menu in GitHub desktop to sync with the current repo
* Commit the changes.
* Select push from the repository menu in GitHub desktop.

To do the same on the linux or Virtual machines we use for Storage Solutions for big data that dont have the Github Desktop, a shell script was provided by Ciaran to upload to main called dogit.

| #!/bin/bash  echo "Enter 1 to push the git files for this CA: "  echo "Enter 2 to do a pull from git: "  echo "Enter 3 to exit: "  read opt  if [[ $opt -eq 1 ]]; then  echo "git is now updating. Be patient!"  git pull origin main  git add ./  echo "Enter commit message: "  read commit\_message  git commit -m "$commit\_message"  git push https://ghp\_rHoM91YEKcD8ZloD7AMJhw7GB7gReR3FAK7y@github.com/ciaranq/ny-air-quality.git main  echo "All gitted!" |
| --- |

This all worked well when we were all working away on separate tasks and we did not run into any difficulties but as we all started to use and change different files we needed to start to use branching so we could update and later merge with the main branch.

#### Branching.

As mentioned above StackOverflow thinks that branching is the killer app within GIt and why it is so popular. Branches allow you to keep different versions of code cleanly separated. This is ideal when collaborating in a team with other people and also for developers who want to work on a new feature but aren't quite ready to show the rest of the world or their boss probably.

These are the commands we used to get branchind setup for our team

* To create a branch for you individual coding go to this github link https://github.com/ciaranq/ny-air-quality/
* On the main dropdown select main,
* click New branch, under "Branch name", type a name for the branch ie. Jim, Kate etc
* Under "Branch source", choose a source for your branch, this will be main for our project .
* Click Create branch. Now go back to your github desktop and select your branch to start working/adding your code.

When we tried this first we had to put our name in using the git config command so we can be identified for our specific branch. The commands at the dos or ubuntu prompt was as follows:

| git config --global user.email "sbs23098@student.cct.ie"  git config --global user.name "CIARAN QUINLAN CCT" |
| --- |

We were then able

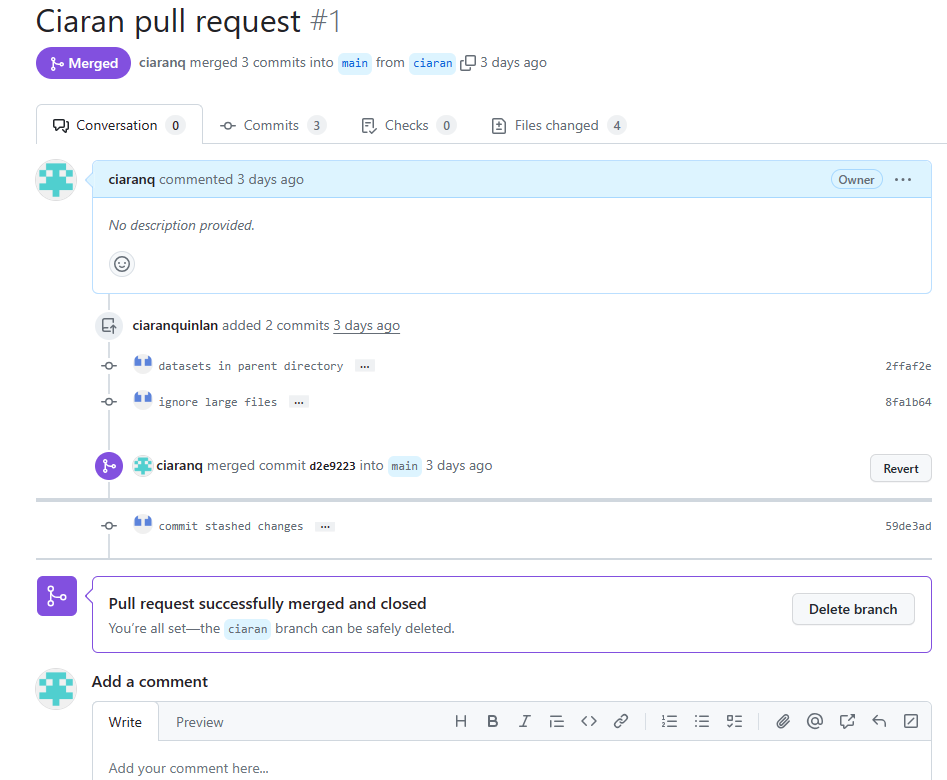
#### Viewing Branches

If you want to see if changes have been made to main branch while you have been working on you branch you can view the main branch on the github desktop by changing branch.

This will give you the option to 'stash' your changes or to commit them. (Git stash is a powerful Git feature for temporarily shelving and reverting local changes, allowing you to re-apply them later on. )

#### Rebasing / Merging Branches

All three members created branches in the Git repository and successfully merged the branches into the main branch. Rebasing and merging branches takes a bit of time to understand the process but all team members got the hang of it and with some more time and project collarration we will all be merge experts. Here are som screen shots of the rebasing and merging process.



#### Examples of re-basing



#### Rebasing commands

#### Checkout via command line

If the conflicts on this branch are too complex to resolve in the web editor, you can check it out via command line to resolve the conflicts.

HTTPS

SSH

Patch

Step 1: Clone the repository or update your local repository with the latest changes.

git pull origin main

Step 2: Switch to the head branch of the pull request.

git checkout ciaran2

Step 3: Merge the base branch into the head branch.

git merge main

Step 4: Fix the conflicts and commit the result.

See [Resolving a merge conflict using the command line](https://docs.github.com/pull-requests/collaborating-with-pull-requests/addressing-merge-conflicts/resolving-a-merge-conflict-using-the-command-line) for step-by-step instructions on resolving merge conflicts.

Step 5: Push the changes.

git push -u origin ciaran2

#### Gitignore

We have very large datasets and Git cant handle >100mb files, one way we dealt with these is to put them in a directory that is not part of the git repo and refer to them in a custom path or specify what files you want to ignore in the .gitignore file and they will not be pushed/pulled etc.

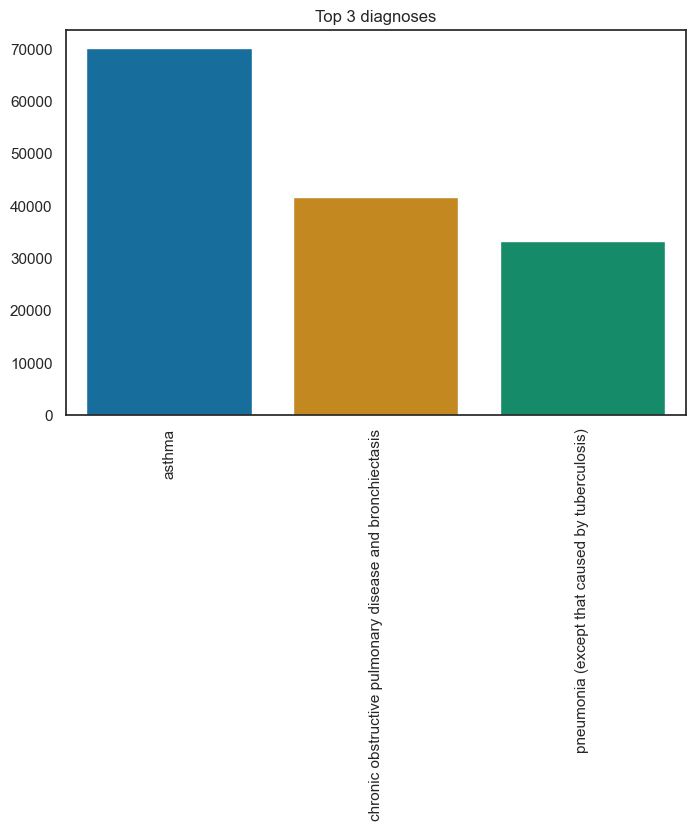
# Medical facility data analysis

Generally it can be assumed that the quality of the air around us would have an impact on respiratory illness and disease. (Dominici F, 2006) concluded that there was an increase in risk for hospital admission for respiratory diseases with short term exposure to fine particulate matter of 2.5 nanometres(PM2.5). New York has seen a reduction in PM2.5 since the implementation of their ‘One New York’ plan. Earlier in this project we could see this was true by performing analysis on the published data sets from the Environmental Protection Agency and determining that the Air Quality Index (AQI) does depend on the PM2.5 density level. So based on this assumption we want to see if this is reflected in the health data. We started an investigation into the data surrounding the admission to the emergency rooms of medical facilities in New York available from ‘<https://health.data.ny.gov/>’

Data is anonymised to preserve data privacy. We took data from 2016, 2017, 2018, 2019, 2020 and 2021 to see if there was any trend in the admission of respiratory illnesses that could be compared to the trending of AQI in the same time frame. The annualised data is quite comprehensive and large, more than 2 million rows with 35 columns per year. Concatenating the data in the original format was not an option as the memory on the standard machine was unable to free up the space required to process.

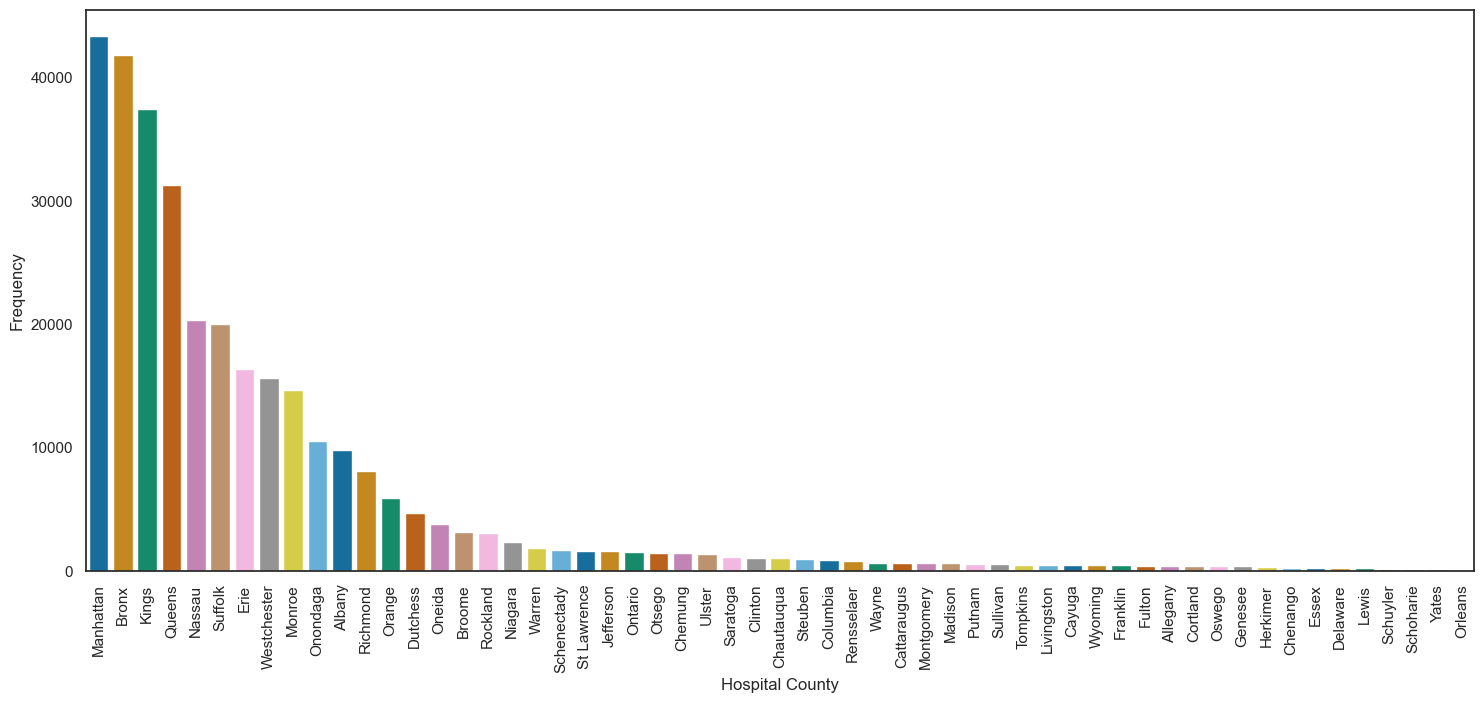
Each data year was therefore pre-processed individually to include only respiratory illnesses and disease that had a minor threat of mortality. Other columns regarding payment, insurance, ethnicity, and sex were also removed. This resulted in a dataset of typically 65k rows with 15 columns. Concatenating the six pre-processed sets gave a dataset with 320960 rows and 15 columns.

Analysing this data, we found that the top 3 minor respiratory illness diagnoses over this 6-year period across all New York counties were ‘Asthma’, ‘Chronic Obstructive Pulmonary Disease and Bronchiectasis,’ ‘Pneumonia’. We can see in Fig.1 that ‘Asthma’ is almost twice the frequency of the next most common diagnosis.



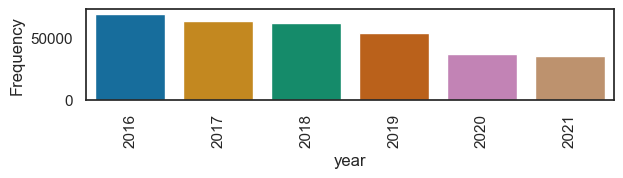
*Figure 14 Top 3 diagnoses*

Looking at the distribution of the data we can see that the largest volume of presented respiratory illnesses to emergency rooms in medical facilities is related to the higher populated areas, namely the counties in the centre of New York; Manhattan, Bronx, Kings, and Queens. (Madani, 2023) deduced that emergency room visits were elevated in counties with higher levels of air pollution. Which would corroborate with what we are seeing in the graph below.



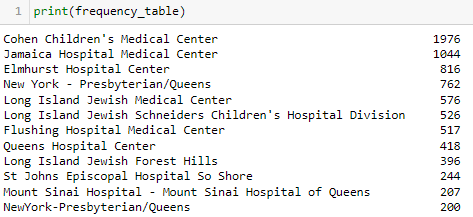
*Figure 15 Hospital County frequencies*

Breaking down the annual frequencies we can see that there has been a reduction year on year. Considering the pause in 2020 due to Covid -19, we purposefully set the threat to minor to remove the Covid-19 data from the set as it adds a skew to the data in 2020 and subsequent years.



*Figure 16 year to year frequency*

Focusing our attention on the hospital county of Queens there are 11 facilities listed.



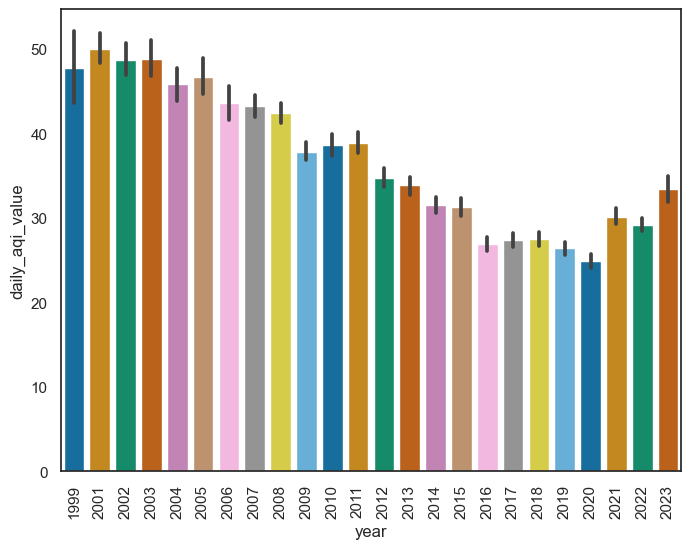
*Figure 17 List of medical facilities*

On the map these facilities are well spread across the county of Queens. Map generated with the folium library. Latitude and longitude data for the medical facilities was generated with GeoPy library.



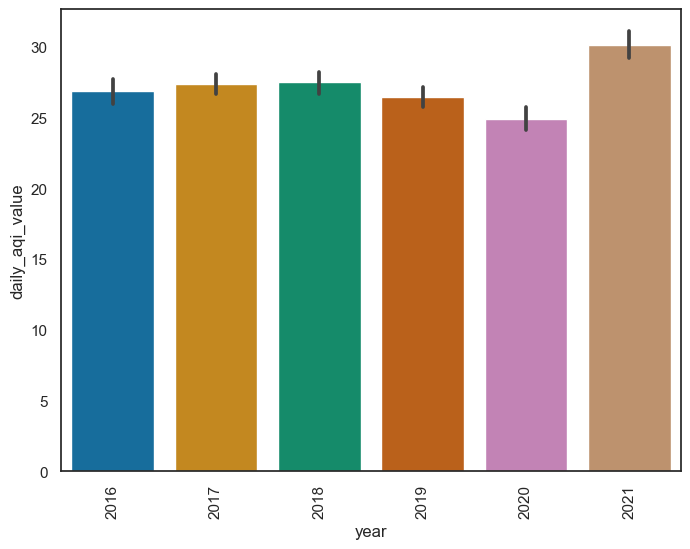
*Figure 18 Medical facilities in Queens County*

Looking back at the Air Quality Index data set we can see that the AQI value had been dropping year to year, and had stabilised between 25 and 30 after 2015.

**

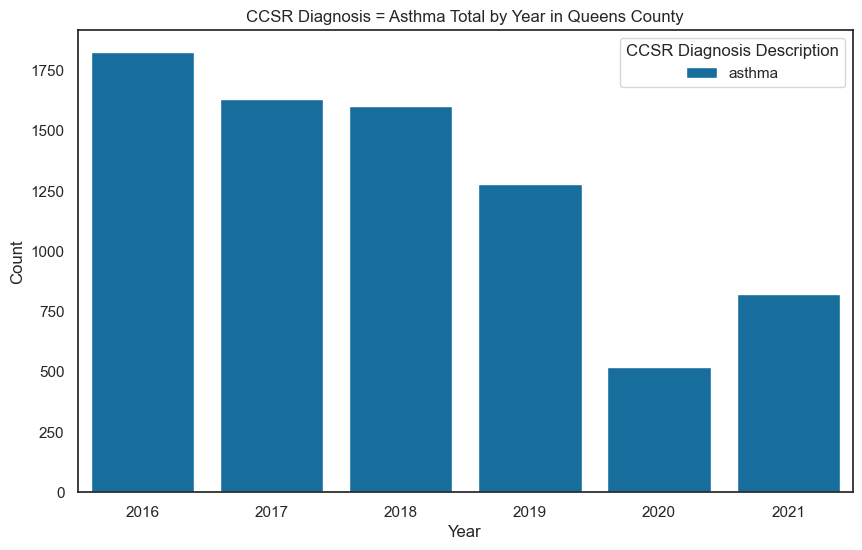
*Figure 19 AQI value by year*

We trimmed the data to capture the years 2016 to 2021 in line with the emergency room admission data. We can see the stabilisation clearly with an increase in 2021 as the pause in 2020 was ended and there were elevated activities as life got back to normal.

**

*Figure 20 AQI data 2016 to 2021*

Trending the highest frequency of respiratory illness diagnosed in Queens County over the six years; ’Asthma’. We can see in the figure below that asthma count of emergency room diagnoses has been decreasing over the duration.

**

*Figure 21 Asthma 2016 to 2021*

There was a noticeable drop in 2020 with the Covid-19 pandemic. Increasing again in 2021 as lockdowns were lifted in the latter half of the year. Comparing the graphs on the surface it looks like the reduction in patients presenting at medical facility emergency rooms with asthma may be linked to the reduction in AQI across the same duration. Further analysis would need to be completed by taking data going back to when the data was logged in the database, earliest date for the data in the repository online is 2009.

(Madani, 2023) had done extensive modelling and found a strong correlation between elevated air pollutants and emergency room attendance for asthma. They also investigated other factors including smoking and poverty but could not document the known effects on any other respiratory illness except for asthma.

Further study could be done in this area considering sex, age, and ethnicity to determine other factors contributing to the respiratory illnesses other than air quality index.

# 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:

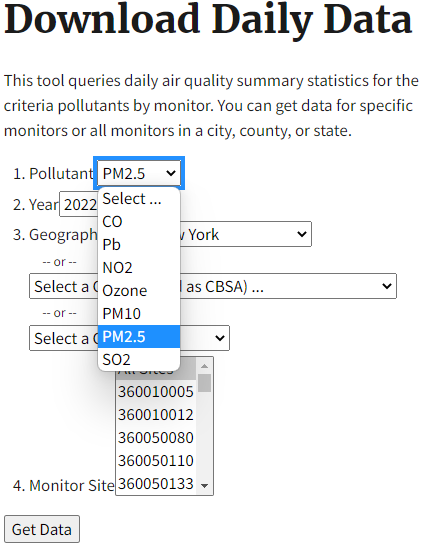
* Actual AQI value for 2008/11/04 = 28.25
* Predicted AQI value for 2008/11/04 = 28.50

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.

# Challenges encountered.

## 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>

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 several times to get a solid dataset to work with and we found alternative ways to get data that was easier to use.

The hospital data set we were using was also very large when attempting to concatenate, and overcame this challenge by processing the data for each year in its own notebook and then concatenating afterwards. If we were more confident in the handling of the data through a data file sharing system, such as Hadoop and storing in its own database for analysis we might have had a less convoluted method of processing.

## 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 members up to date and this just slowed progress a little.

## Challenge 3: Target variable and Machine Learning Model

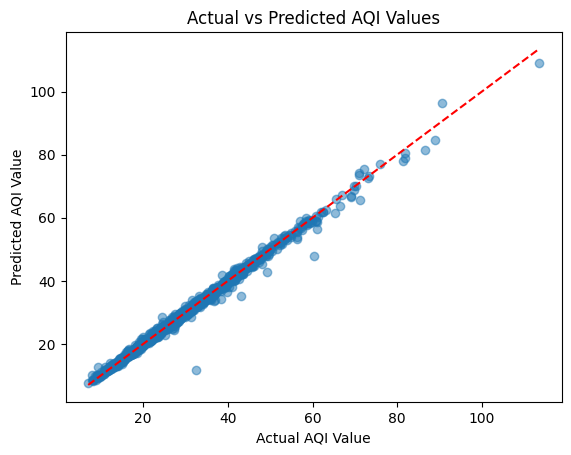
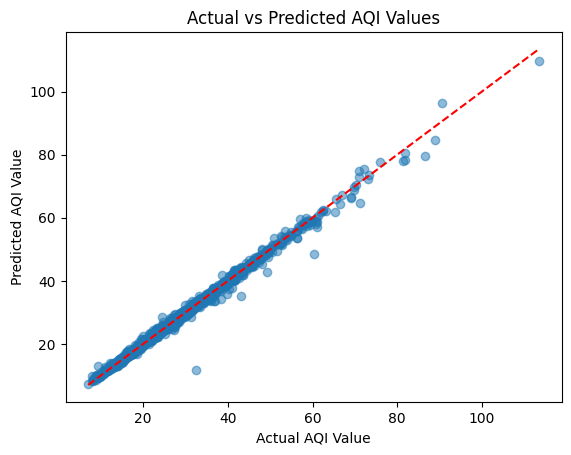
Our first attempt was to predict an air quality index using a classification model. The model had very little data to work on as most of the pollutants were within one index. After a thorough discussion, we decided to predict an air quality index value and use a regression model.

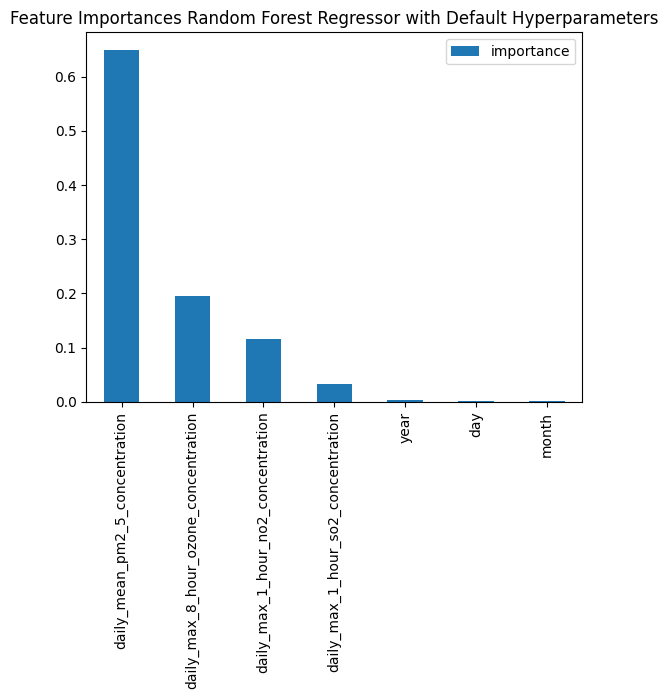
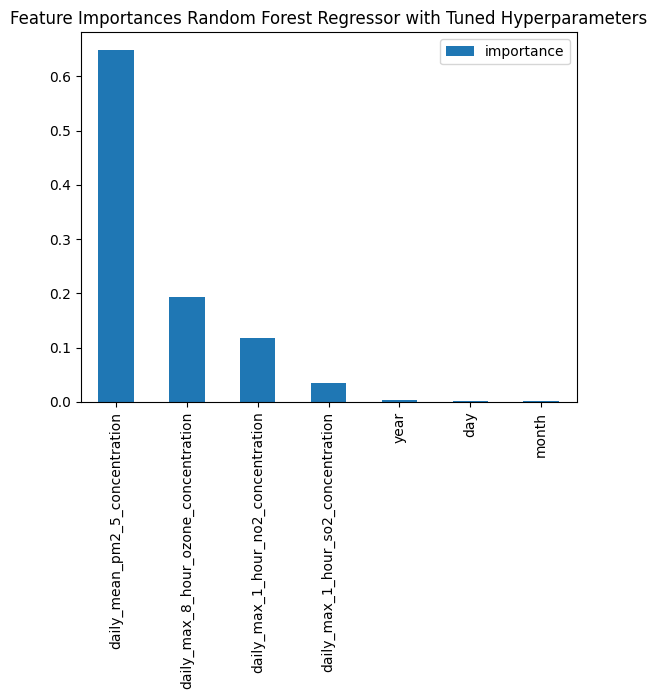
# Analysis of the machine learning model results

To analyse the results of our models we decided to use plots along with feature importance graphs and test it on seen data.

The first machine learning model we chose to employ was a Random Forest Regressor with default hyperparameters. We then used GridSearchCV to look for the best hyperparameters to get better results for the model, which after tuning the Random Forest Regressor, gave us a slightly worse results than with the default parameters.

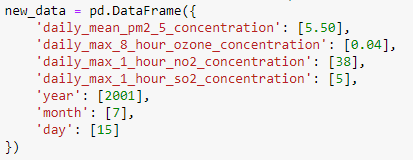
The results on the scatter plot were quite promising with the predicted and actual aqi values very close to the red diagonal line. There are a few points that are further away from the line of best fit which might be outliners where the model’s predictions were less accurate.



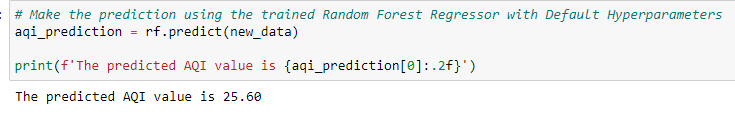
For both models the feature importance were the same, where daily mean pm2.5 triumphing over the others.

The testing on the seen data gave us an interesting results, we have decided to create a data frame with one of the data rows from the dataset, and try if the models predict the aqi value.

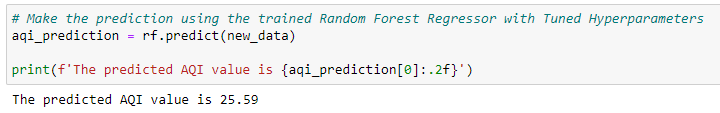
The new dataframe is as follows with a aqi value of 26.25



And the following results were achieved:

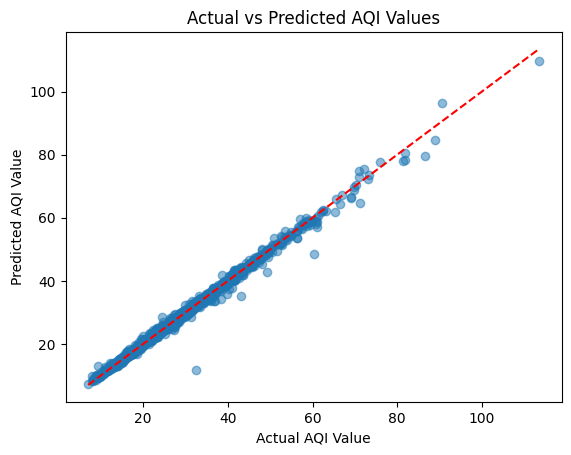


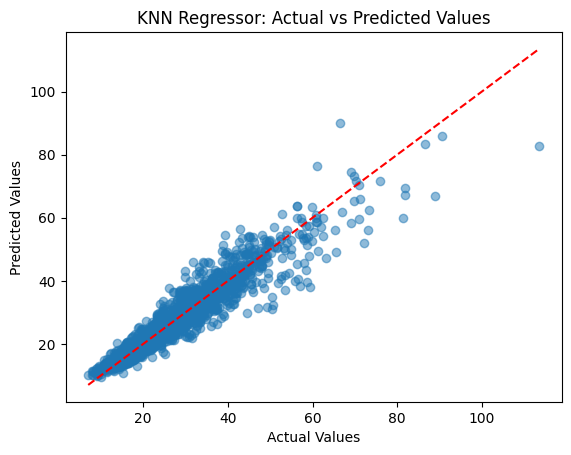
The random forest regressor with default hyperparameters managed to predict the aqi value very well, with only a slight error.



For both models the results were really close to each other, we decided to take best model and compare it to the rest of the models. We decided to with with the Random Forest Regressor with default hyperparameters as it had a better results.

KNN Regressor comparing to our Random Forest

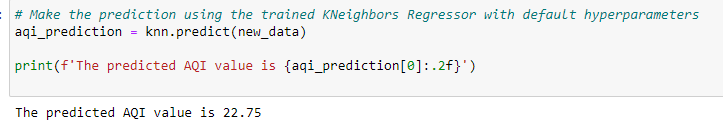




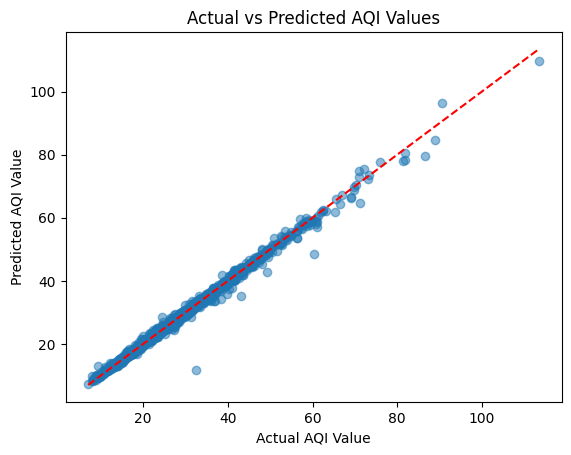
KNN compared to Random Forest

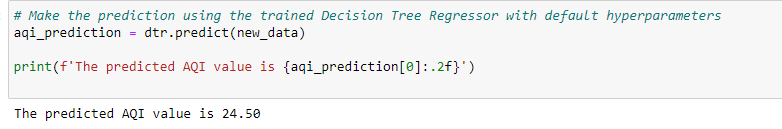
And comparing to predicting aqi value on the seen data where aqi should be 26.25

KNN model is underfitting.

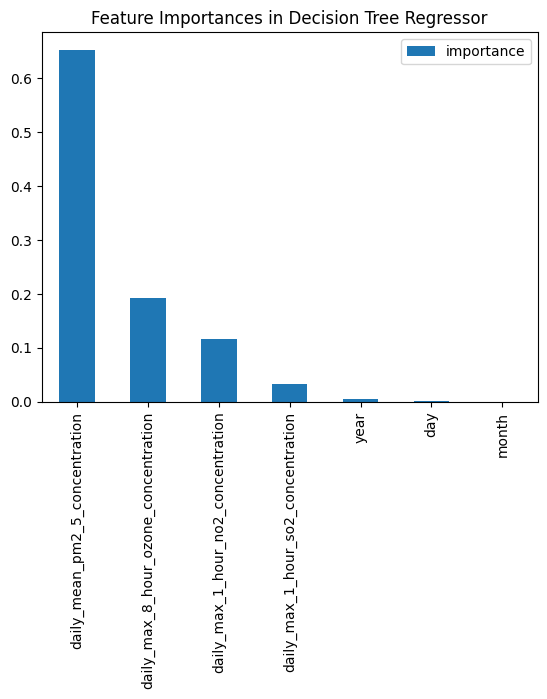
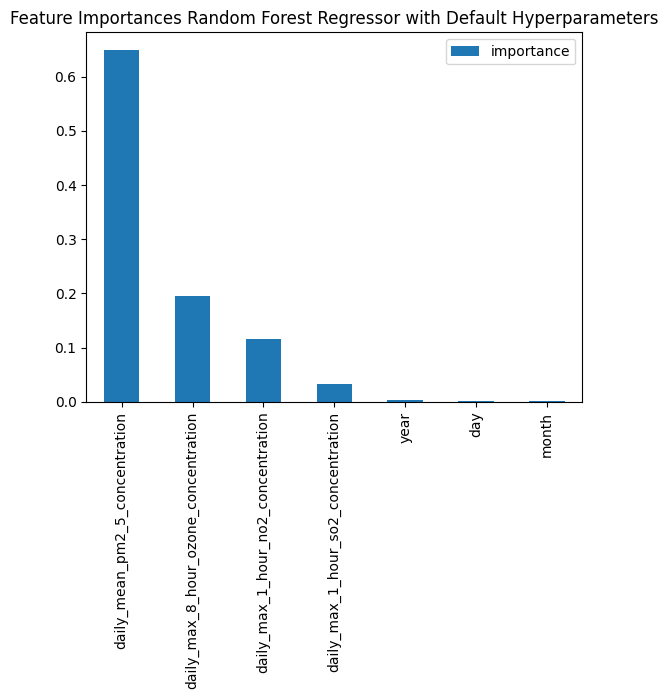


Decision Tree Regressor compared to Random Forest





The output on the seen data looks like it is underfitting too.

And interesting discovery was in the feature importance, the results slightly differ, where Random Forest took month into the consideration where Decision Tree did not:  


# Potential Next Steps

We have developed a model that can make a prediction of the air quality index by selecting the best results based on various algorithms. Our next step could possibly to delve further into the medical facility discharge data and merge it with the AQI data file. This might give insights into predictions of the quantity of patients presenting with respiratory illness diagnosis based on air quality index values.

This can lead to further analysis into how seasonal AQI may affect age ranges, sex and ethnicity of patients and can potentially make a prediction of the quantity of patients that may present to the emergency room with symptoms of a respiratory illness. A distribution analysis may allow for the allocation of resources to facilities where there could

A distribution analysis of facilities may allow the reallocation of resources during times of elevated AQI to help relieve pressure on medical systems locally. This can be broken down to a county level or the wider state.

# Conclusion

In conclusion, the goal of our project was to examine NYC air quality, particularly in the vicinity of Queens College. 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 generalises 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 as the air quality index starts to rise again post covid-19.

Comparing the impact of the air quality on respiratory illness presentation at hospital emergency rooms there is some correlation between the improving air quality and the presentation of patients with asthma. Further investigation can be done considering age, ethnicity, and sex to determine if there are other contributory factors relating to respiratory illnesses, but this is outside the scope of this assignment.

# 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.

“A look under the hood: how branches work in Git - Stack Overflow.” The Stack Overflow Blog, 5 April 2021, https://stackoverflow.blog/2021/04/05/a-look-under-the-hood-how-branches-work-in-git/. Accessed 10 November 2023.

“Git.” Wikipedia, https://en.wikipedia.org/wiki/Git#History. Accessed 13 November 2023.

Appendix 1

Filenames and directories

| ├───.github  │ └───workflows | Github files |
| --- | --- |
| ├───Archive | Files no longer used, testing and reference files |
| ├───CA1 | CA 1 files. report etc |
| ├───CA2 | CA 2 files. report etc |
| ├───code\_and\_dataset  │ ├───.ipynb\_checkpoints  │ └───mglearn  │ └───\_\_pycache\_\_ | main code and datasets on NY air quality  main notebook = queens\_rev05 |
| ├───Hospital data  │ └───.ipynb\_checkpoints  └───index\_files | Hospital data notebooks and datasets |
| └───index\_files | Github Pages files |