Air Quality

Strategic Thinking - AI Team 2

Presented by

Katarzyna Gogolka
Jim Higgins
Ciaran Quinlan

Lecturer: James Garza

Project Goal

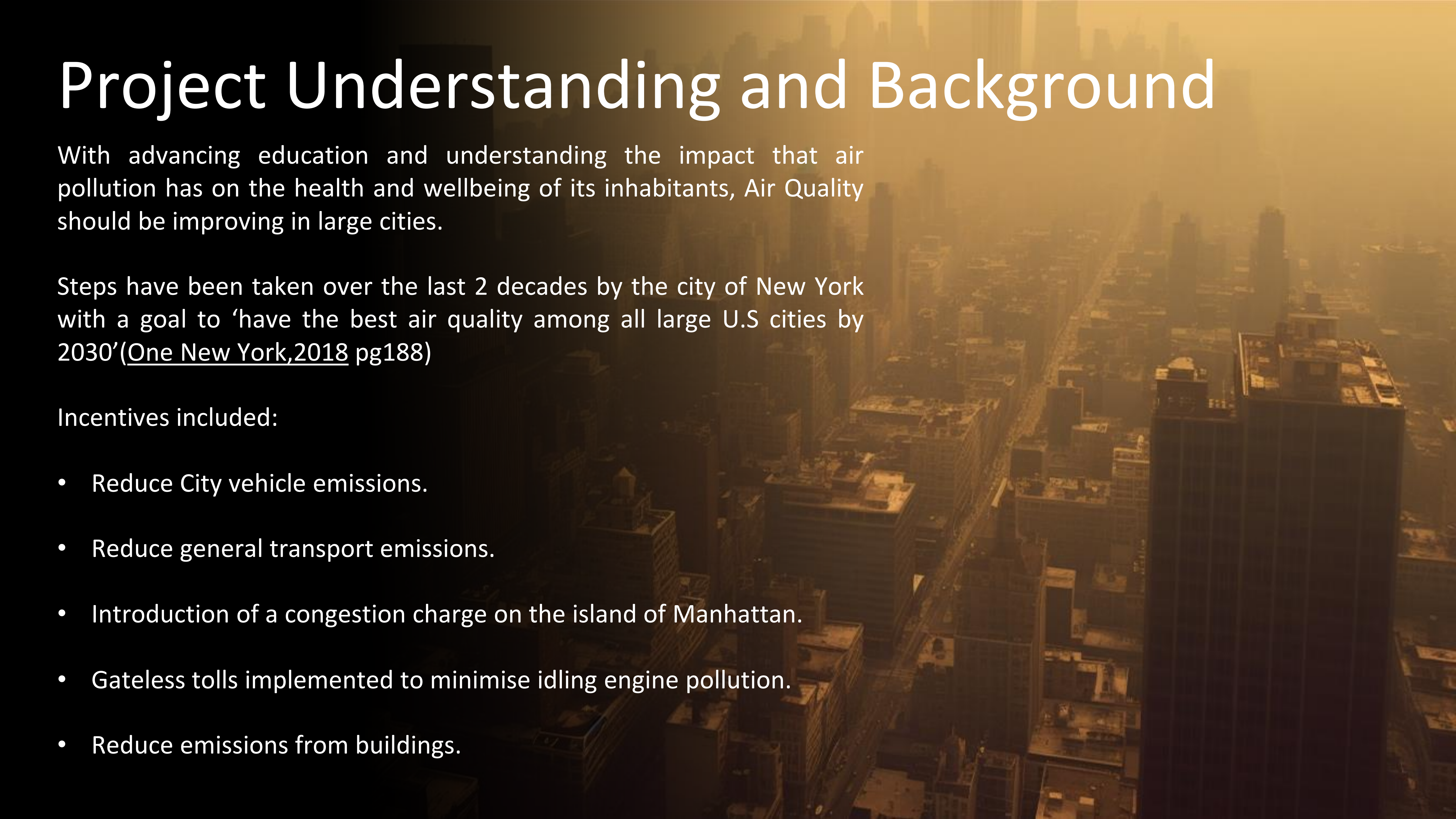
The goal of this project:

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

Make a prediction with a machine learning algorithm for the Air Quality Value

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



Project Understanding and Background

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.

Steps have been taken over the last 2 decades by the city of New York with a goal to ‘have the best air quality among all large U.S cities by 2030’(One New York,2018 pg188)

Incentives included:

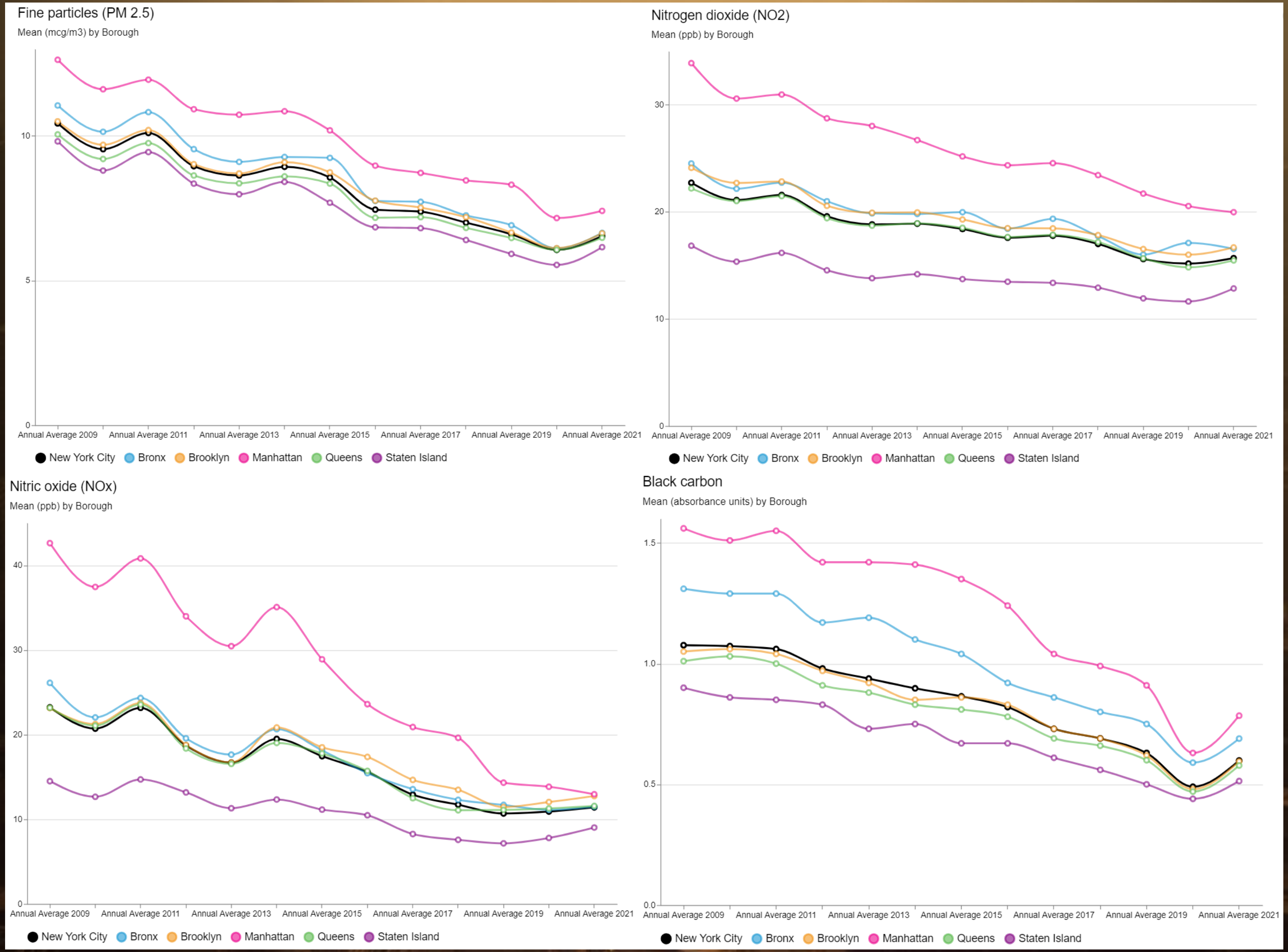
- Reduce City vehicle emissions.
- Reduce general transport emissions.
- Introduction of a congestion charge on the island of Manhattan.
- Gateless tolls implemented to minimise idling engine pollution.
- Reduce emissions from buildings.

Project Understanding and Background

The NYC Community Air Survey Findings Between 2009 and 2017:

- Fine particulate matter (PM 2.5) ↓30%
- Nitrogen Dioxide (NO2) ↓26%
- Nitric oxide (NOx) ↓44%
- Black Carbon ↓30%

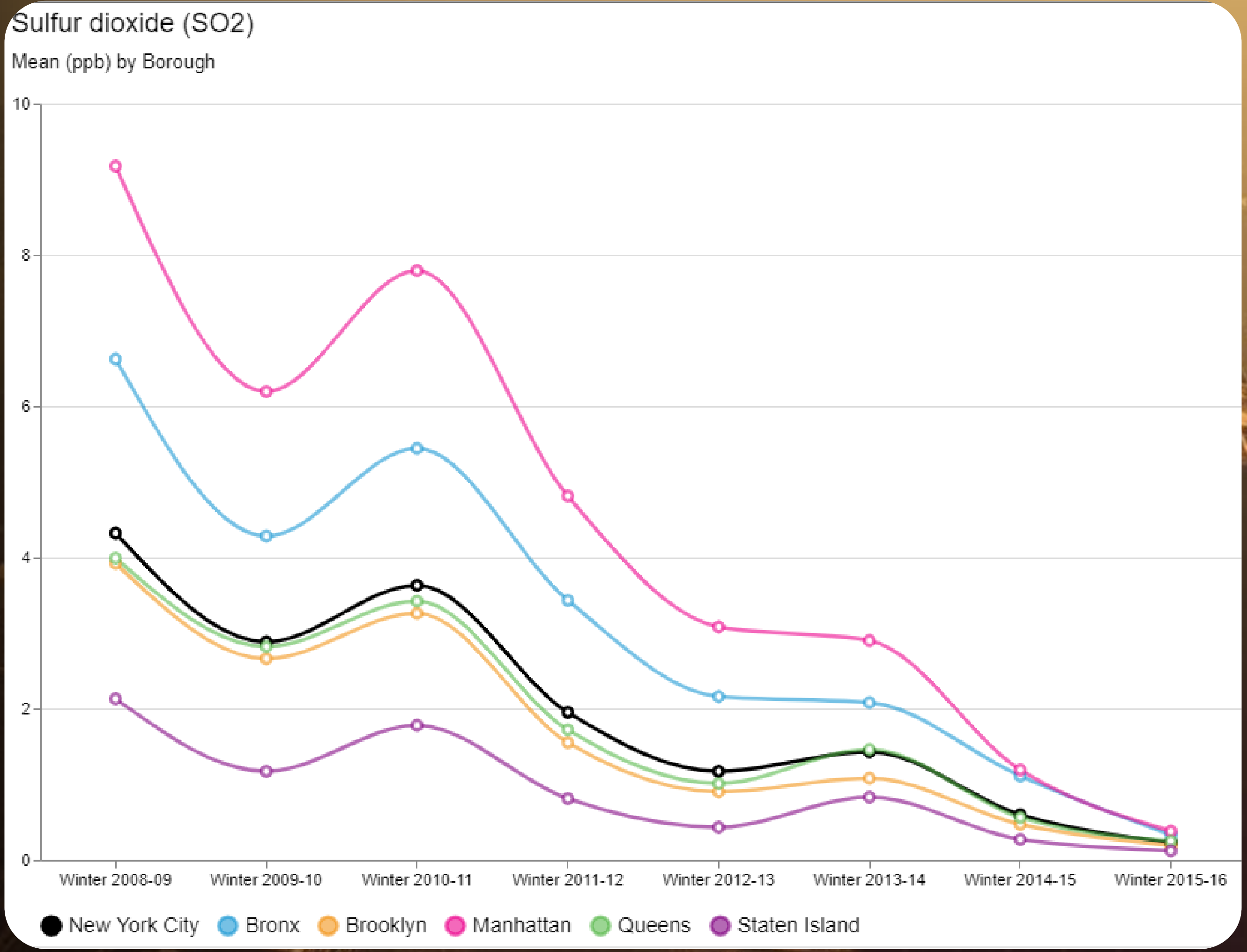
(NYC.gov, 2019)



Project Understanding and Background

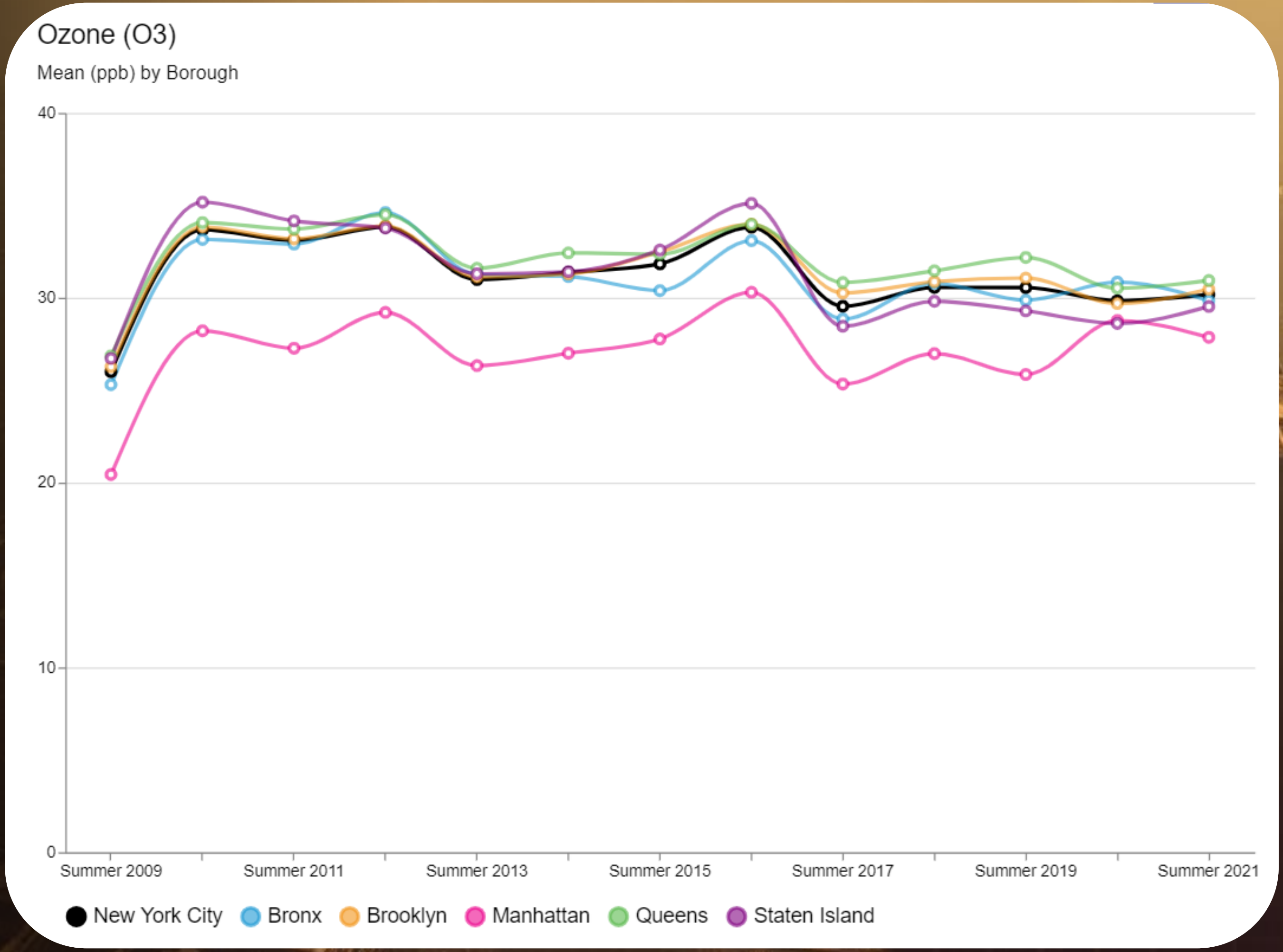
- Sulphur Dioxide (SO2) ⬇ 96%

The largest declines have been observed for Sulphur Dioxide (SO2), mostly due to an implementation of heating oil regulations.



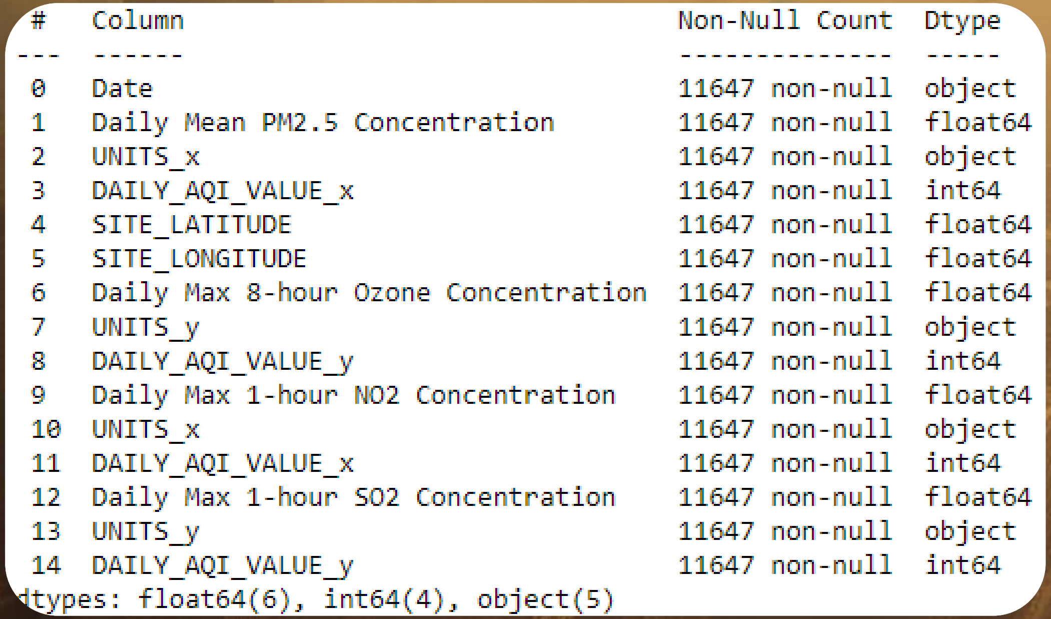
Project Understanding and Background

The overall Summer Ozone concentration remained stable and is consistent by Borough.



Data Understanding

- Queens College 2 Station
- Years 2001-2023
- Four different pollutants
- Daily AQI Value per Pollutant
- Total 11,647 Rows
- Total 15 Columns
- 5 Categorical
- 10 Numerical

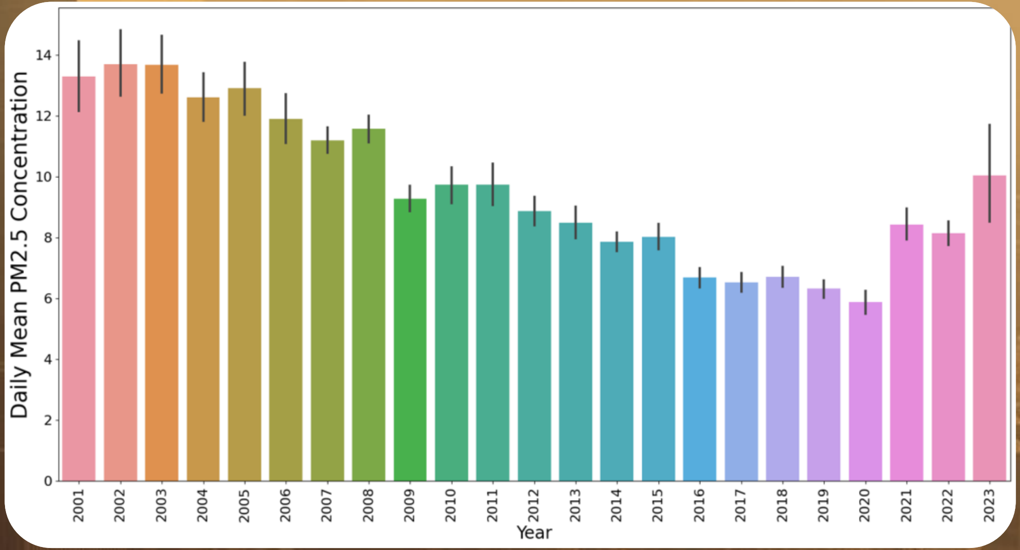


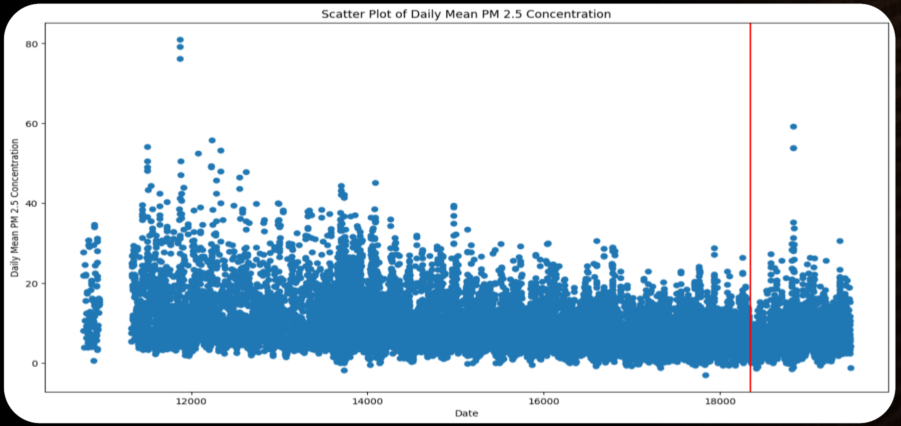
```
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Date                               11647 non-null  object
 1   Daily Mean PM2.5 Concentration     11647 non-null  float64
 2   UNITS_x                            11647 non-null  object
 3   DAILY_AQI_VALUE_x                  11647 non-null  int64
 4   SITE_LATITUDE                      11647 non-null  float64
 5   SITE_LONGITUDE                     11647 non-null  float64
 6   Daily Max 8-hour Ozone Concentration 11647 non-null  float64
 7   UNITS_y                            11647 non-null  object
 8   DAILY_AQI_VALUE_y                  11647 non-null  int64
 9   Daily Max 1-hour NO2 Concentration 11647 non-null  float64
 10  UNITS_x                            11647 non-null  object
 11  DAILY_AQI_VALUE_x                  11647 non-null  int64
 12  Daily Max 1-hour SO2 Concentration 11647 non-null  float64
 13  UNITS_y                            11647 non-null  object
 14  DAILY_AQI_VALUE_y                  11647 non-null  int64
dtypes: float64(6), int64(4), object(5)
```

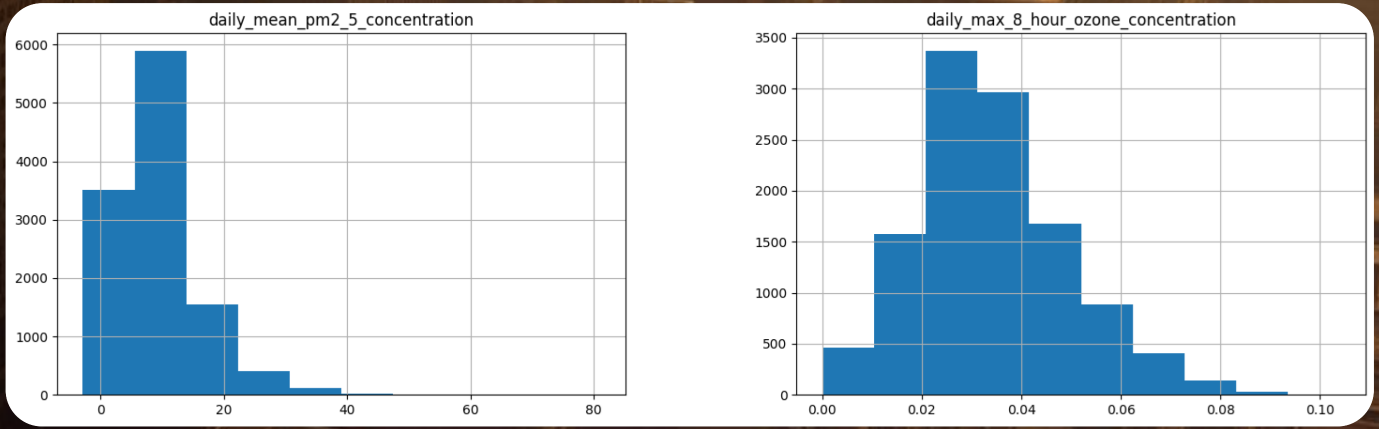
Dataset taken from U.S Environmental Protection Agency (epa.gov)

Basic Data Exploration & Visualisation

- Standardize column names ✓
- Deal with missing values ✓
- Check for duplicates ✓
- Check for unique values ✓
- Extract useful features ✓
- Inspect outliners ✓
- Visualise pollutants ✓

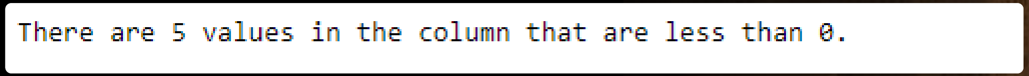




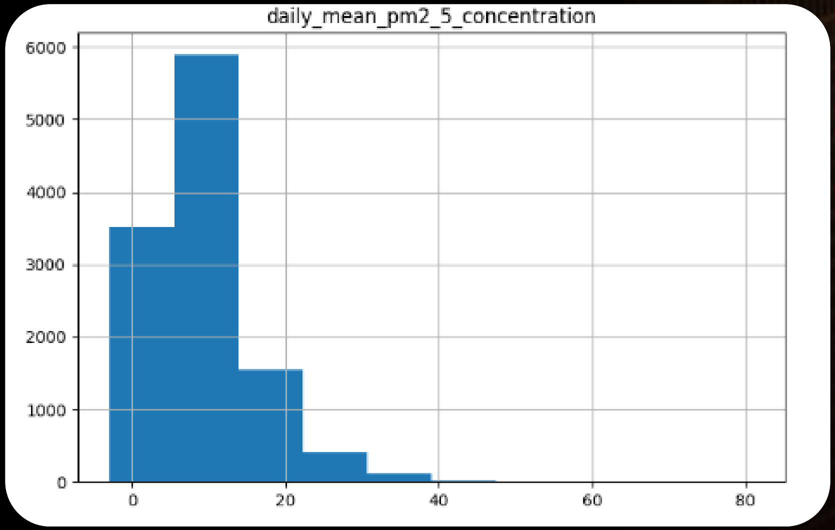


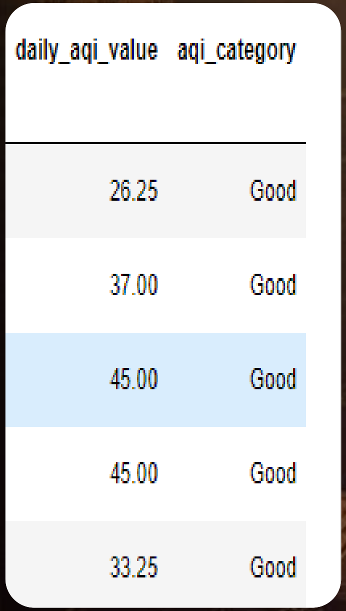
Feature Engineering

- Delete not needed columns✓
- Deal with outliers✓
- Create a column with an average AQI✓
- Create AQI Index column✓
- Deal with incorrect features✓

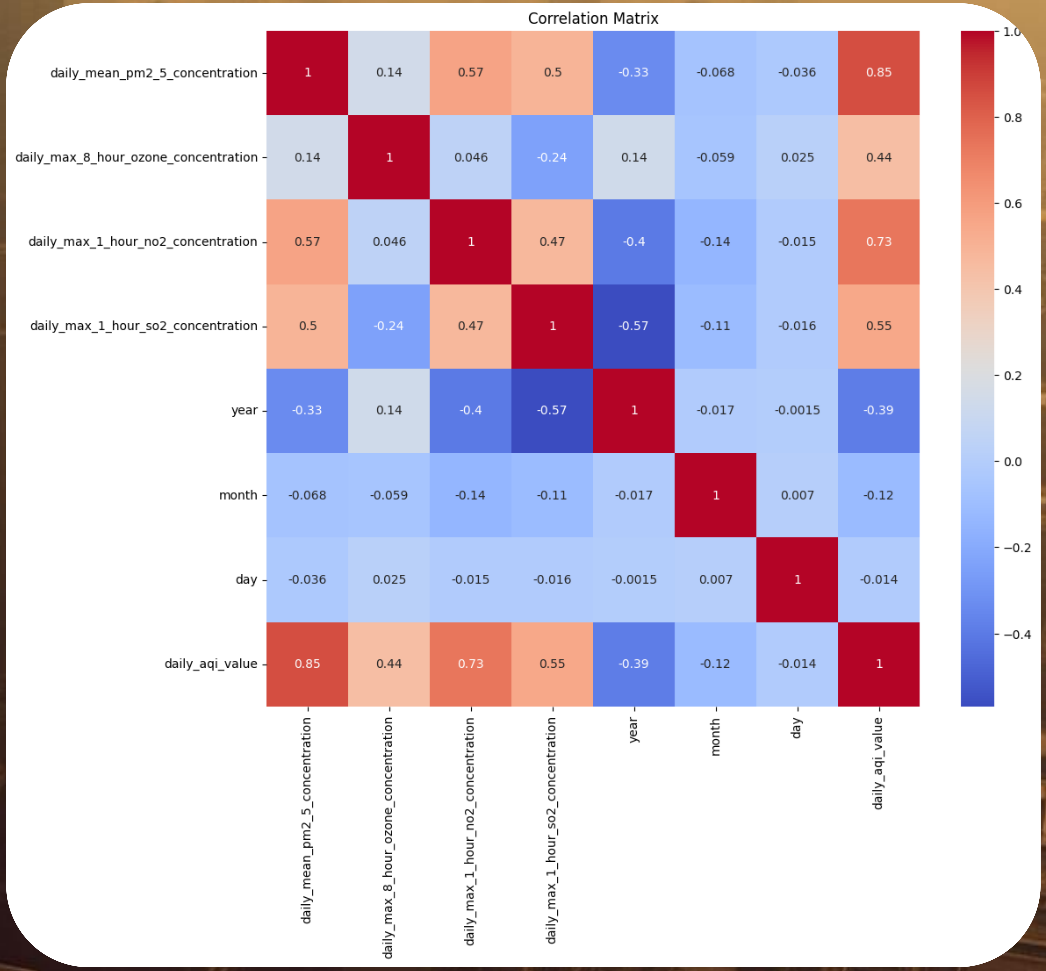


```
There are 5 values in the column that are less than 0.
```





daily_aqi_value	aqi_category
26.25	Good
37.00	Good
45.00	Good
45.00	Good
33.25	Good

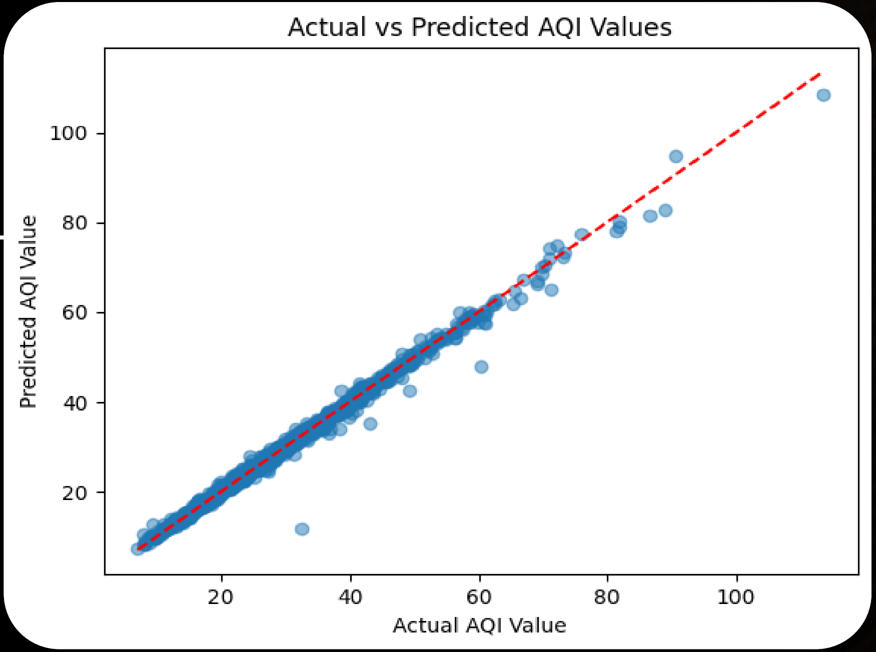


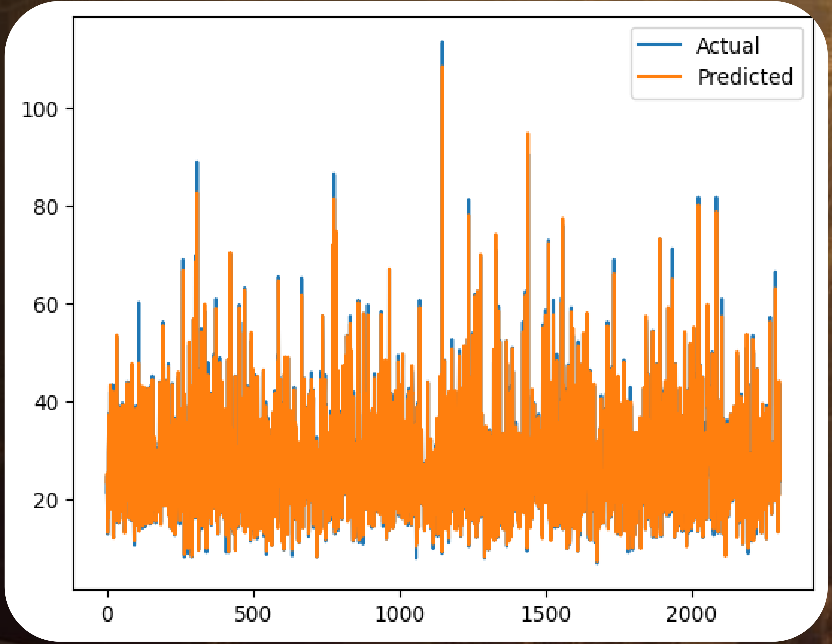
Correlation Matrix

	daily_mean_pm2_5_concentration	daily_max_8_hour_ozone_concentration	daily_max_1_hour_no2_concentration	daily_max_1_hour_so2_concentration	year	month	day	daily_aqi_value
daily_mean_pm2_5_concentration	1	0.14	0.57	0.5	-0.33	-0.068	-0.036	0.85
daily_max_8_hour_ozone_concentration	0.14	1	0.046	-0.24	0.14	-0.059	0.025	0.44
daily_max_1_hour_no2_concentration	0.57	0.046	1	0.47	-0.4	-0.14	-0.015	0.73
daily_max_1_hour_so2_concentration	0.5	-0.24	0.47	1	-0.57	-0.11	-0.016	0.55
year	-0.33	0.14	-0.4	-0.57	1	-0.017	-0.0015	-0.39
month	-0.068	-0.059	-0.14	-0.11	-0.017	1	0.007	-0.12
day	-0.036	0.025	-0.015	-0.016	-0.0015	0.007	1	-0.014
daily_aqi_value	0.85	0.44	0.73	0.55	-0.39	-0.12	-0.014	1

Machine Learning Model

- Random Forest Regressor Results:
  - R squared = 0.9938 - Model fits the data very well
  - Mean Absolute Error = 0.44 - Mean model’s prediction is off by 0.44 units
  - Mean Cross - Validation Score = 0.9917 - Performed well on test data and likely to generalize well to a new data





Challenges Encountered

Data challenges

- Selecting relevant datasets

Deciding on the relevant parameters in the final dataset

- Per pollutant?
- Per borough?
- Per station?
- What years?

Deciding on the machine learning model

- Regression or Classification model?
- Air Quality Index or Air Quality Index Value?

Team Composition

- Losing and gaining a member late in project
- Splitting up the workload

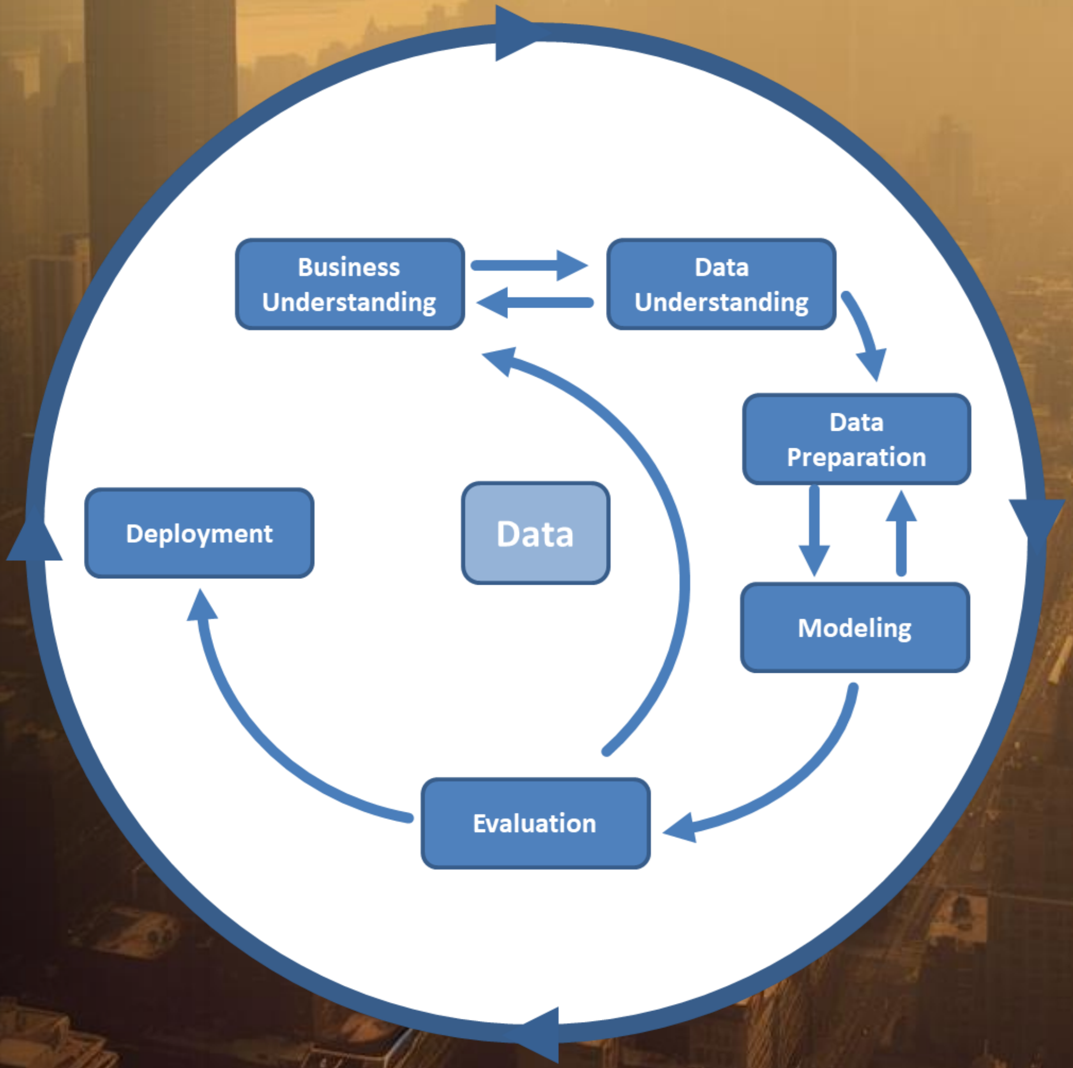
The CRISP-DM Methodology and Next Steps

80% of our effort to date has been on data understanding and data preparation

We have reached to the deployment stage of CRISP DM strategy, as our model predicts very well on existing and new data.

We need to re-track back to the evaluation phase as we work further as a team on the code.

We still believe the code could be improved in the future by choosing more features for the model.



Conclusion

Our model for prediction of AQI value worked really well.
Model had a very little error in the prediction

Next steps:

- Move the project to Github for further collaboration
- Merge traffic data in the region of Queens College
- Assess the AQI v traffic volume
- Check for seasonality variations in both sets
- Hypothesise if traffic is the biggest influence on the AQI

End

Thank you for your attention.

Questions?