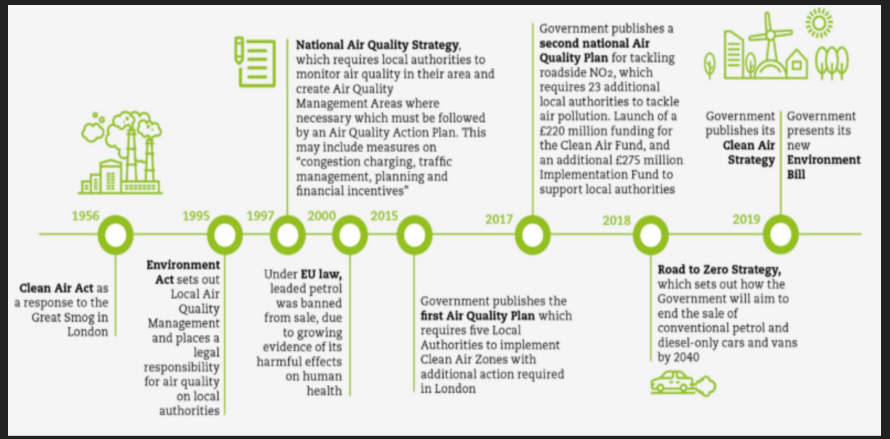
Real Time Air Quality Monitoring and Prediction

Alan Uthuppan

Objective & Significance

- Develop predictive models for real-time monitoring and forecasting of PM2.5 and PM10 levels
- Enhance air quality management and public health response through accurate predictions
- Facilitates proactive measures and policies by environmental agencies to mitigate pollution effects.

Background



- Air quality is a critical public health issue influenced by both natural and man-made sources.
- Particulate matter (PM2.5 and PM10) are among the most harmful air pollutants due to their ability to penetrate deep into the respiratory tract,
- Predicting these pollutants' levels can be essential for preventing adverse health effects, especially for those with sensitive health conditions.

AQI Category and Color	Index Value	Description of Air Quality
Good Green	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Moderate Yellow	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups Orange	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Unhealthy Red	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy Purple	201 to 300	Health alert: The risk of health effects is increased for everyone.
Hazardous Maroon	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

$$I_p = \frac{I_{Hi} - I_{Lo}}{BP_{HI} - BP_{Lo}}(C_p - BP_{Lo}) + I_{Lo}.$$

Where I_p = the index for pollutant p

C_p = the truncated concentration of pollutant p

BP_{Hi} = the concentration breakpoint that is greater than or equal to C_p

BP_{Lo} = the concentration breakpoint that is less than or equal to C_p

I_{Hi} = the AQI value corresponding to BP_{Hi}

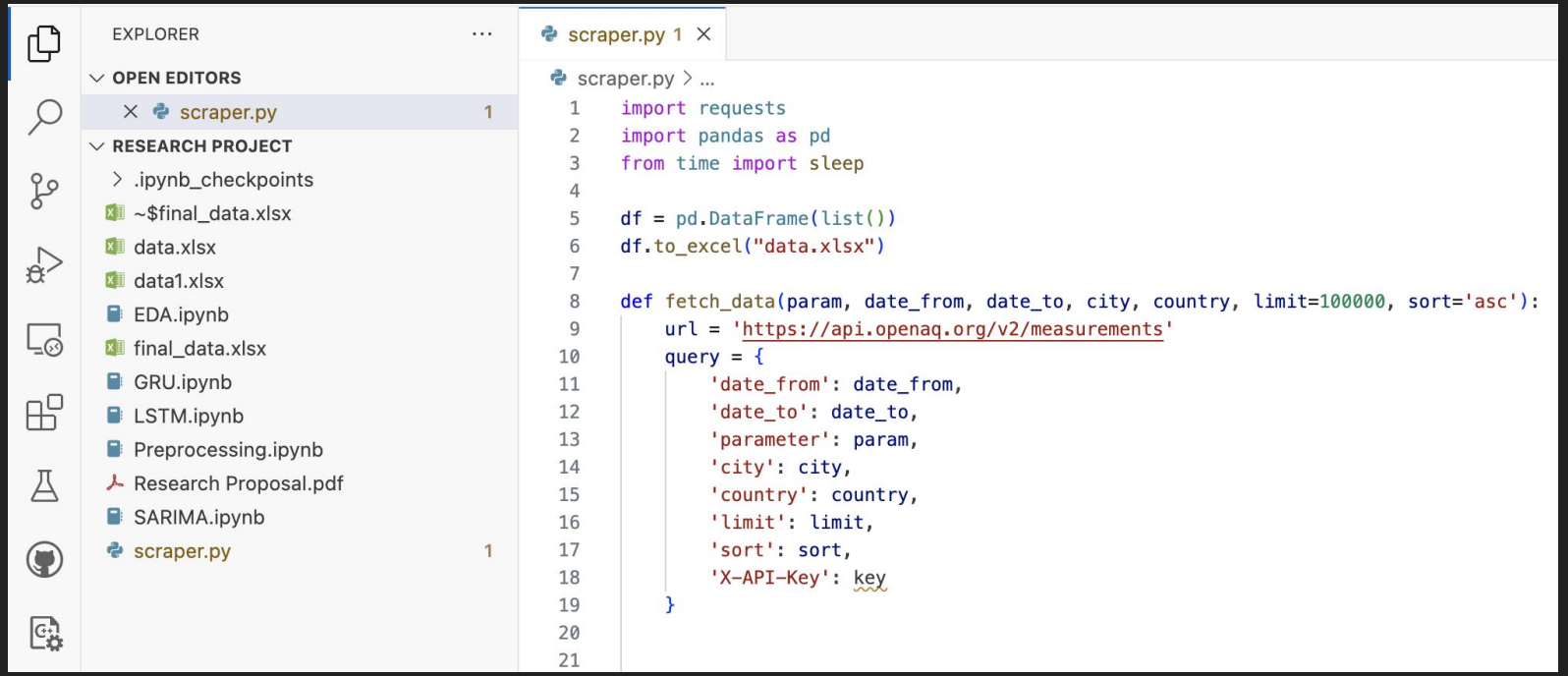
I_{Lo} = the AQI value corresponding to BP_{Lo}

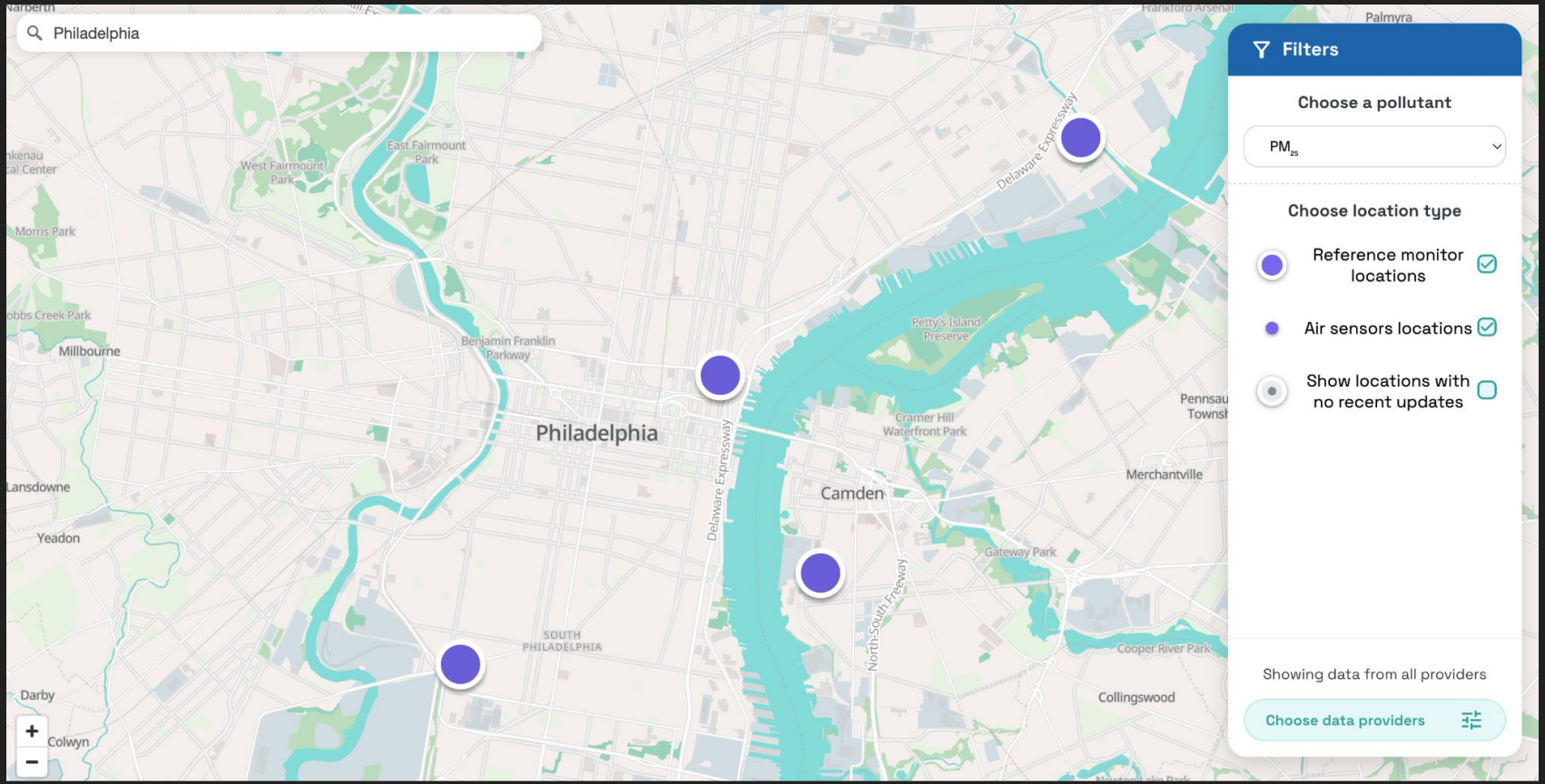
Proposed Approach

- Scrape historical air quality data from OpenAQ REST API.
- Conduct data cleaning, preprocessing, and exploratory analysis (pandas, sklearn, seaborn)
- Train SARIMA, LSTM, and GRU models for forecasting.
- Compare models based on their ability to handle seasonal variations and long-term dependencies in data.

Data

- Dataset consists of over 7,300 hourly observations from multiple air quality sensors.
- Features include PM2.5, PM10 readings, air quality indices, time stamps, and derived features like hourly and daily averages.
- Issues getting more data due to rate limit for API requests and size limit for downloads.

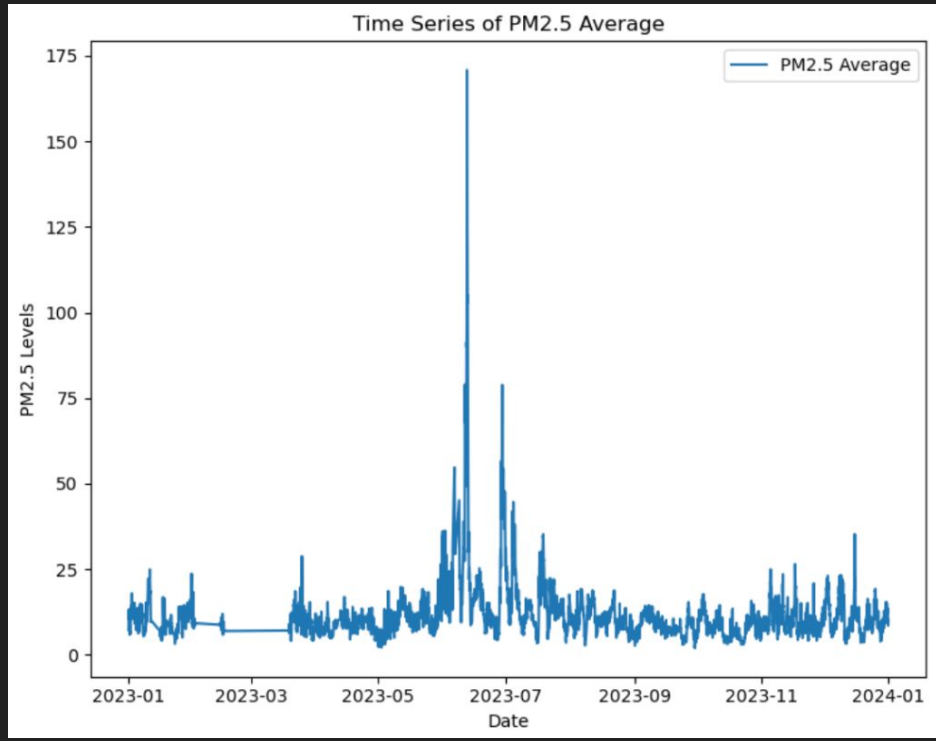

EXPLORER
OPEN EDITORS
scraper.py 1
RESEARCH PROJECT
.ipynb_checkpoints
~$final_data.xlsx
data.xlsx
data1.xlsx
EDA.ipynb
final_data.xlsx
GRU.ipynb
LSTM.ipynb
Preprocessing.ipynb
Research Proposal.pdf
SARIMA.ipynb
scraper.py 1
scraper.py 1
scraper.py > ...
1 import requests
2 import pandas as pd
3 from time import sleep
4
5 df = pd.DataFrame(list())
6 df.to_excel("data.xlsx")
7
8 def fetch_data(param, date_from, date_to, city, country, limit=100000, sort='asc'):
9 url = 'https://api.openaq.org/v2/measurements'
10 query = {
11 'date_from': date_from,
12 'date_to': date_to,
13 'parameter': param,
14 'city': city,
15 'country': country,
16 'limit': limit,
17 'sort': sort,
18 'X-API-Key': key
19 }
20
21

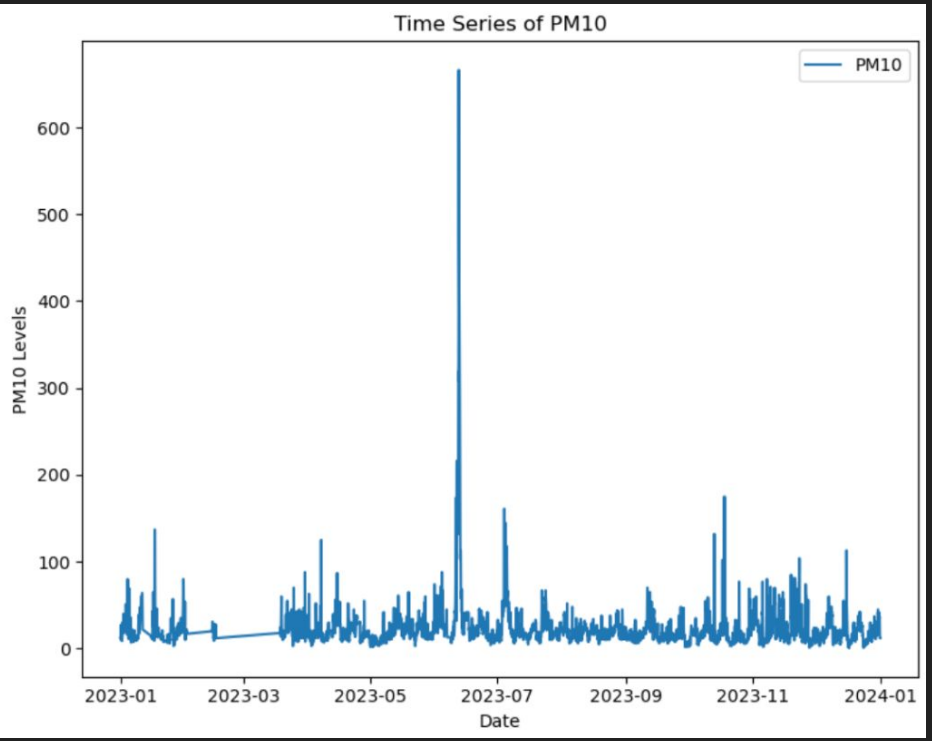

Philadelphia
Filters
Choose a pollutant
PM2.5
Choose location type
Reference monitor locations
Air sensors locations
Show locations with no recent updates
Showing data from all providers
Choose data providers
Philadelphia
Camden
Delaware Expressway
Benjamin Franklin Parkway
East Fairmount Park
West Fairmount Park
Morris Park
Millbourne
Lansdowne
Yeadon
Darby
Colwyn
SOUTH PHILADELPHIA
Petty's Island Preserve
Cramer Hill Waterfront Park
Gateway Park
North-South Freeway
Cooper River Park
Collingswood
Merchantville
Palmyra

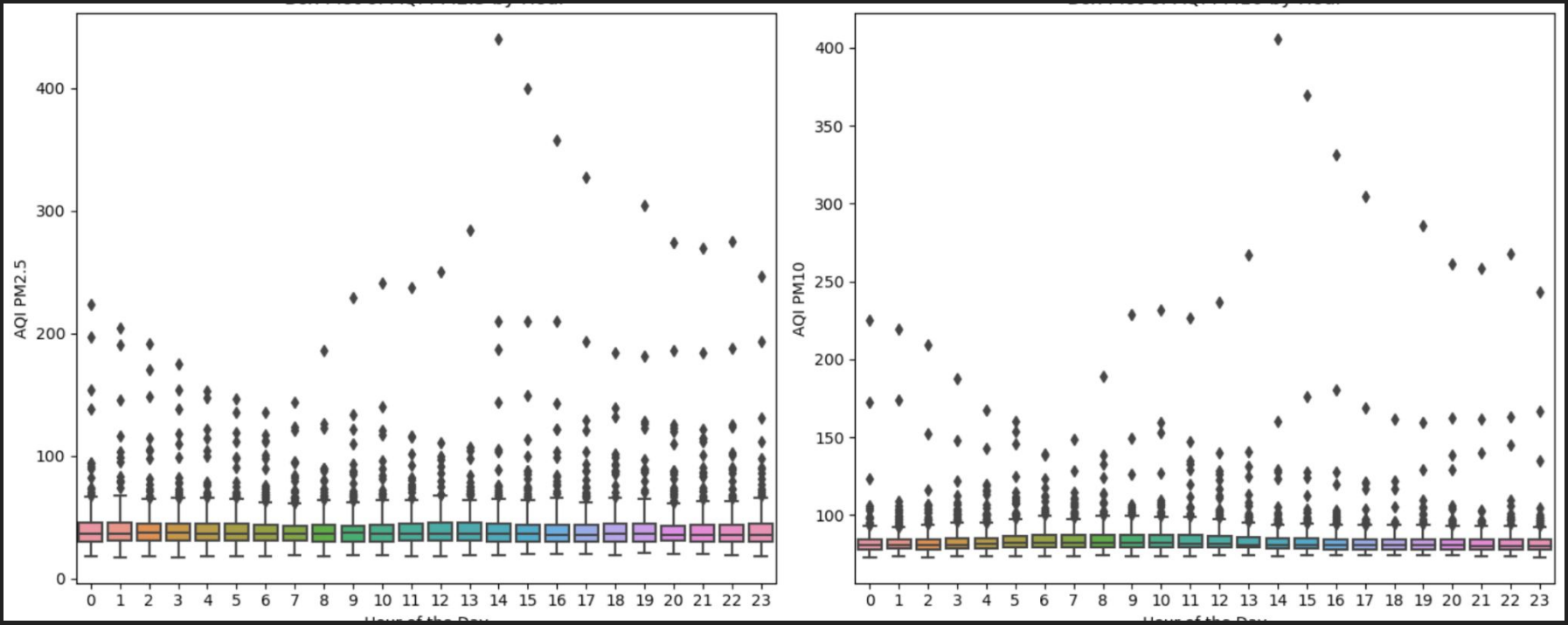
	Timestamp	pm25_1	pm25_2	pm25_3	pm25_AVERAGE	pm10	hour	day_of_week	month	AQI_pm25	AQI_pm10	pm25_24hr_avg	pm10_24hr_avg	pm25_lag1	pm10_lag1	pm25_change	pm10_change
0	2023-01-01 01:00:00	9.3	6.2	6.1	7.2	12	1	6	1	30.5	78.5						
1	2023-01-01 02:00:00	8.4	6.7	6.6	7.233	12	2	6	1	30.5825	78.5			7.2	12	0.033	0
2	2023-01-01 03:00:00	10.2	7.6	8.3	8.7	14	3	6	1	34.25	79.5			7.233	12	1.467	2
3	2023-01-01 04:00:00	11.4	7.7	8.5	9.2	15	4	6	1	35.5	80			8.7	14	0.5	1
4	2023-01-01 05:00:00	22.4	9.1	8.3	13.267	27	5	6	1	45.6675	86			9.2	15	4.067	12
5	2023-01-01 06:00:00	17.6	8.9	7.7	11.4	22	6	6	1	41	83.5			13.267	27	-1.867	-5
6	2023-01-01 07:00:00	14.9	10.4	7.2	10.833	19	7	6	1	39.5825	82			11.4	22	-0.567	-3
7	2023-01-01 08:00:00	13.3	10.7	6.8	10.267	18	8	6	1	38.1675	81.5			10.833	19	-0.566	-1
8	2023-01-01 09:00:00	14.3	11.5	6.8	10.867	18	9	6	1	39.6675	81.5			10.267	18	0.6	0
9	2023-01-01 10:00:00	14.7	11.6	7.2	11.167	19	10	6	1	40.4175	82			10.867	18	0.3	1
10	2023-01-01 11:00:00	13.7	12	9.4	11.7	18	11	6	1	41.75	81.5			11.167	19	0.533	-1
11	2023-01-01 12:00:00	12.1	11.3	8.7	10.7	16	12	6	1	39.25	80.5			11.7	18	-1	-2
12	2023-01-01 13:00:00	12.2	9.9	7.7	9.933	16	13	6	1	37.3325	80.5			10.7	16	-0.767	0
13	2023-01-01 14:00:00	7.4	5.8	7.1	6.767	10	14	6	1	29.4175	77.5			9.933	16	-3.166	-6
14	2023-01-01 15:00:00	5.5	4.4	7.7	5.867	9	15	6	1	27.1675	77			6.767	10	-0.9	-1
15	2023-01-01 16:00:00	6.1	4.3	10.2	6.867	10	16	6	1	29.6675	77.5			5.867	9	1	1
16	2023-01-01 17:00:00	6.5	4.5	9.3	6.767	10	17	6	1	29.4175	77.5			6.867	10	-0.1	0
17	2023-01-01 18:00:00	5.7	4.7	8.4	6.267	9	18	6	1	28.1675	77			6.767	10	-0.5	-1
18	2023-01-01 19:00:00	6	5	7.4	6.133	10	19	6	1	27.8325	77.5			6.267	9	-0.134	1
19	2023-01-01 20:00:00	6.7	5.6	7.3	6.533	11	20	6	1	28.8325	78			6.133	10	0.4	1
20	2023-01-01 21:00:00	7.6	6.7	7.9	7.4	12	21	6	1	31	78.5			6.533	11	0.867	1
21	2023-01-01 22:00:00	9.4	9.1	6.6	8.367	14	22	6	1	33.4175	79.5			7.4	12	0.967	2
22	2023-01-01 23:00:00	13.4	10.5	6.2	10.033	21	23	6	1	37.5825	83			8.367	14	1.666	7
23	2023-01-02 00:00:00	15.5	11.5	7.3	11.433	25	0	0	1	41.0825	85	8.954208333	15.29166667	10.033	21	1.4	4
24	2023-01-02 01:00:00	18.4	11.5	7.6	12.5	30	1	0	1	43.75	87.5	9.175041667	16.04166667	11.433	25	1.067	5
25	2023-01-02 02:00:00	18.8	13.3	8.3	13.467	28	2	0	1	46.1675	86.5	9.434791667	16.70833333	12.5	30	0.967	-2
26	2023-01-02 03:00:00	18.7	13.2	8.4	13.433	27	3	0	1	46.0825	86	9.632	17.25	13.467	28	-0.034	-1
27	2023-01-02 04:00:00	16.1	13.4	7.5	12.333	22	4	0	1	43.3325	83.5	9.762541667	17.54166667	13.433	27	-1.1	-5
28	2023-01-02 05:00:00	16.5	15.3	8.6	13.467	22	5	0	1	46.1675	83.5	9.770875	17.33333333	12.333	22	1.134	0
29	2023-01-02 06:00:00	16.6	18.2	8.1	14.3	22	6	0	1	48.25	83.5	9.891708333	17.33333333	13.467	22	0.833	0
30	2023-01-02 07:00:00	16.9	16.1	9.5	14.167	23	7	0	1	47.9175	84	10.030625	17.5	14.3	22	-0.133	1
31	2023-01-02 08:00:00	19.4	14.2	8.7	14.1	26	8	0	1	47.75	85.5	10.19033333	17.83333333	14.167	23	-0.067	3
32	2023-01-02 09:00:00	21.4	13.6	7.6	14.2	30	9	0	1	48	87.5	10.32920833	18.33333333	14.1	26	0.1	4
33	2023-01-02 10:00:00	19.2	15.2	8.5	14.3	29	10	0	1	48.25	87	10.45975	18.75	14.2	30	0.1	-1
34	2023-01-02 11:00:00	20	15.4	7.7	14.367	29	11	0	1	48.4175	87	10.570875	19.20833333	14.3	29	0.067	0
35	2023-01-02 12:00:00	21.6	16.2	6.9	14.9	31	12	0	1	49.75	88	10.745875	19.83333333	14.367	29	0.533	2
36	2023-01-02 13:00:00	22.4	18.7	9.2	16.767	34	13	0	1	54.4175	89.5	11.030625	20.58333333	14.9	31	1.867	3
37	2023-01-02 14:00:00	25.7	20.5	7.7	17.967	40	14	0	1	57.4175	92.5	11.49729167	21.83333333	16.767	34	1.2	6
38	2023-01-02 15:00:00	25	18.2	6.8	16.667	33	15	0	1	54.1675	89	11.94729167	22.83333333	17.967	40	-1.3	-7
39	2023-01-02 16:00:00	25.9	15.5	7	16.133	34	16	0	1	52.8325	89.5	12.333375	23.83333333	16.667	33	-0.534	1
40	2023-01-02 17:00:00	19.8	14.6	7.5	13.967	28	17	0	1	47.4175	86.5	12.633375	24.58333333	16.133	34	-2.166	-6

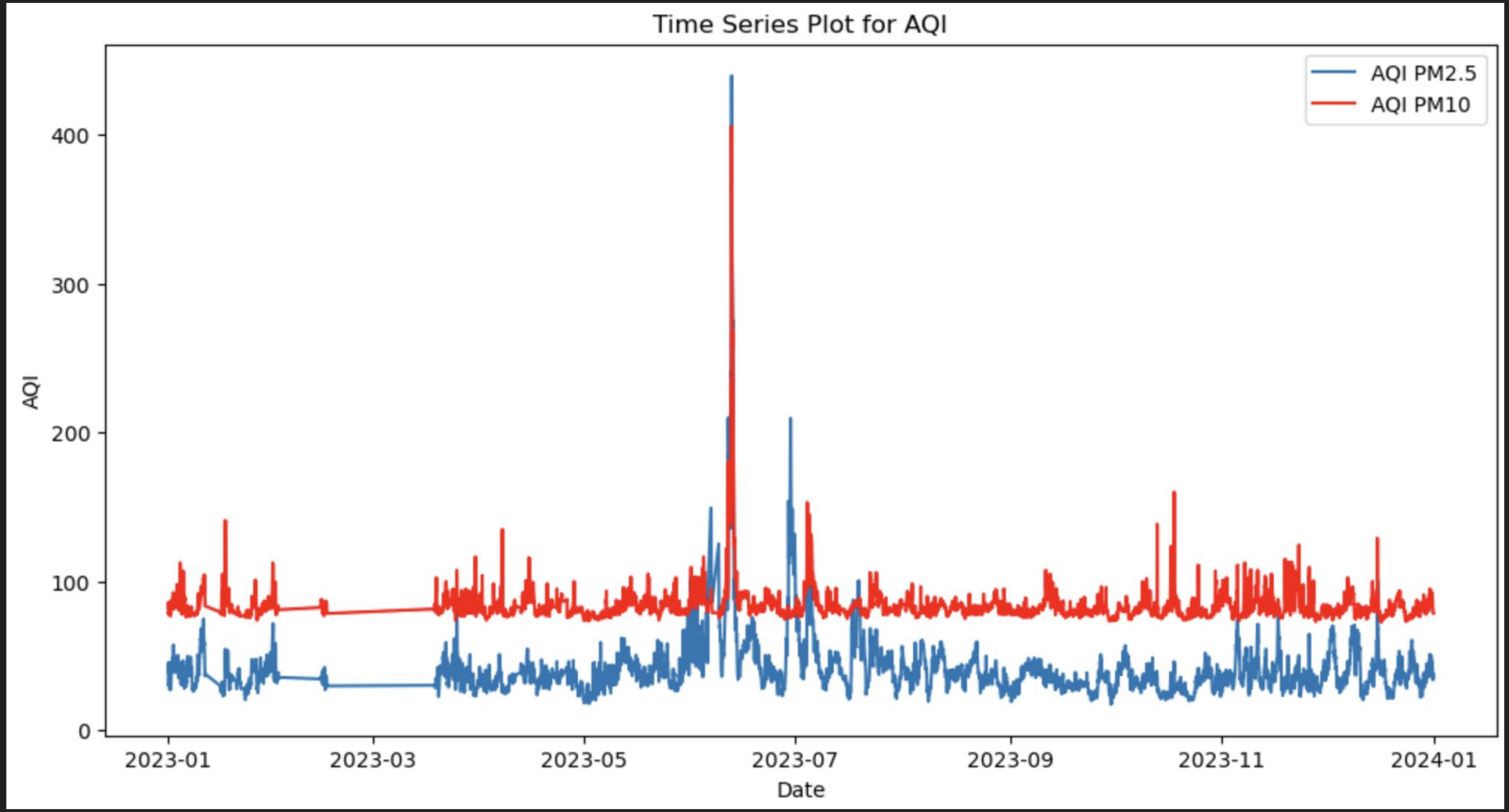
Preliminary Results & Discussion

- Unfortunately because of issues gathering the data, building the models was delayed, and thus I am still in the process of interpreting those results.
- Data standardization successfully implemented, improving input consistency for modeling.
- Initial exploratory data analysis revealed strong seasonal patterns and correlations between features.


Time Series of PM2.5 Average
PM2.5 Average
PM2.5 Levels
175
150
125
100
75
50
25
0
2023-01
2023-03
2023-05
2023-07
2023-09
2023-11
2024-01
Date


Time Series of PM10
PM10
PM10 Levels
600
500
400
300
200
100
0
2023-01
2023-03
2023-05
2023-07
2023-09
2023-11
2024-01
Date


AQI PM2.5
0
100
200
300
400
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
AQI PM10
100
150
200
250
300
350
400
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23


Time Series Plot for AQI
AQI PM2.5
AQI PM10
AQI
0
100
200
300
400
2023-01
2023-03
2023-05
2023-07
2023-09
2023-11
2024-01
Date

Evaluation Plan

- Models will be evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared metrics.
- Perform backtesting to assess model robustness over different time periods and conditions (hopefully with more data).
- Use cross-validation to validate model predictions and prevent overfitting.

Discussion & Moving Forward

- Complete model training and initial evaluations
- Refine the models by tuning hyperparameters and possibly incorporating additional data as needed (more points and/or more features)
- Finalize the report, ensuring all components (methodology, results, and discussions) are adequately covered