RNN STAGE B
PREMIERE PROJECT PRESENTATION

February 25, 2022
HDSC WINTER 2022

Project Overview - AIR QUALITY INDEX

Using a time series analysis, we analysed the air quality index of Madrid, Spain using dataset from the year 2001 to 2018.

With the analysis, factors contributing to the quality of air in Madrid were analysed and model developed for possible prediction

Understanding the problem

1

Authorities in Madrid have been faced with the challenge of combating the deterioration of air quality in the city.

2

Prediction of air pollution at different elevation and station proximity

Project objective
To answer the following questions;

- How do different gases correlate their levels?
- Are there any changes in trends?
- Can they be mapped to the recent decisions made by the city council, or do they relate to rainy dates?
- What is the best model to predict pollution levels?
- How do the levels interpolate between the location of the stations?
- Are some gases more common at different elevations?

Our Approach

Data profiling and cleaning, exploratory data analysis, and Linear Regression to forecast the Air Quality Index in the future

Dataset used for Analysis

The dataset contains information about Air quality in Madrid (2001 - 2018) collected from Kaggle website taking into consideration different pollution levels in Madrid from (2001 to 2018).

Data Cleaning

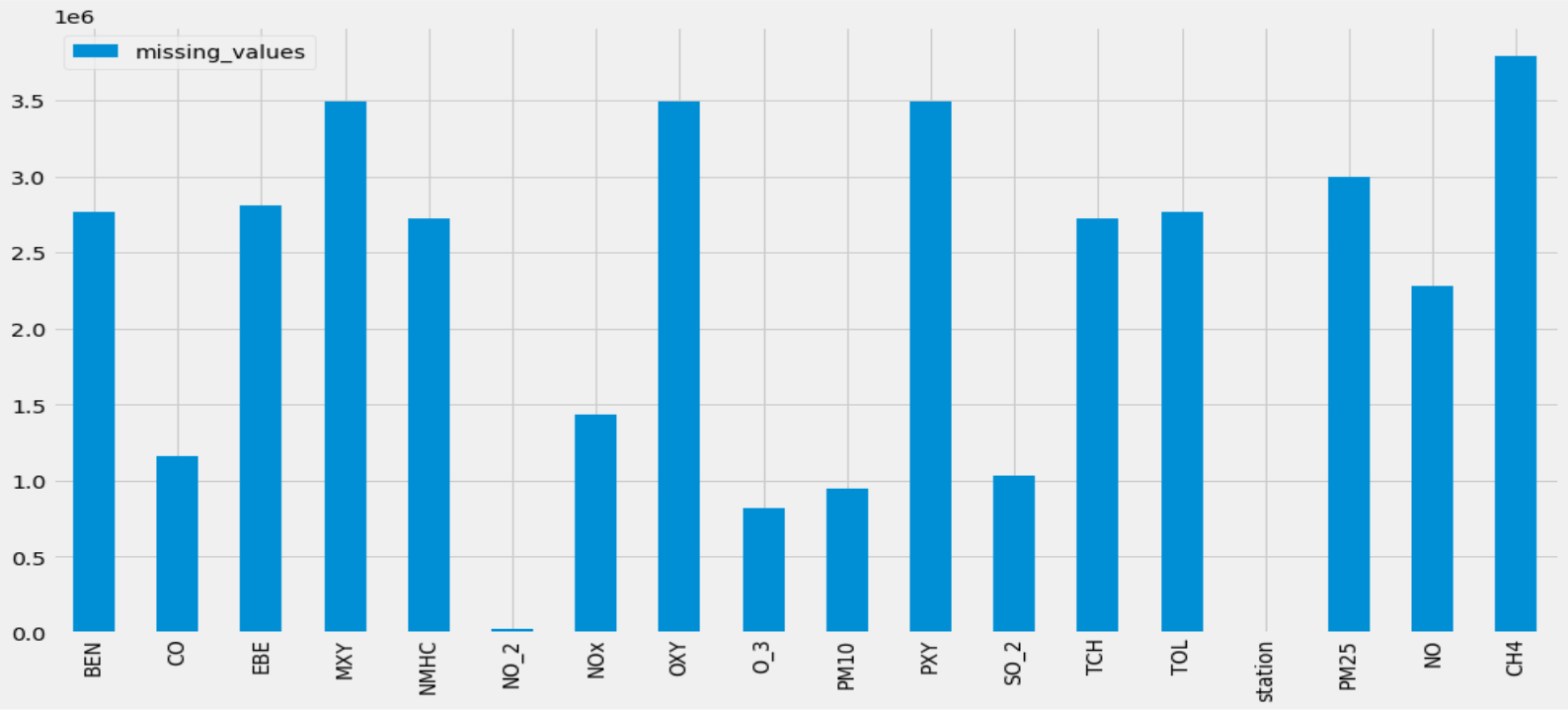
The missing values in columns of interest are CO, NMHC, NO_2, NOx, O_3, PM10, SO_2, TCH, station, NO which were filled with zeros (0) due to the unavailability of exact information

Methods:

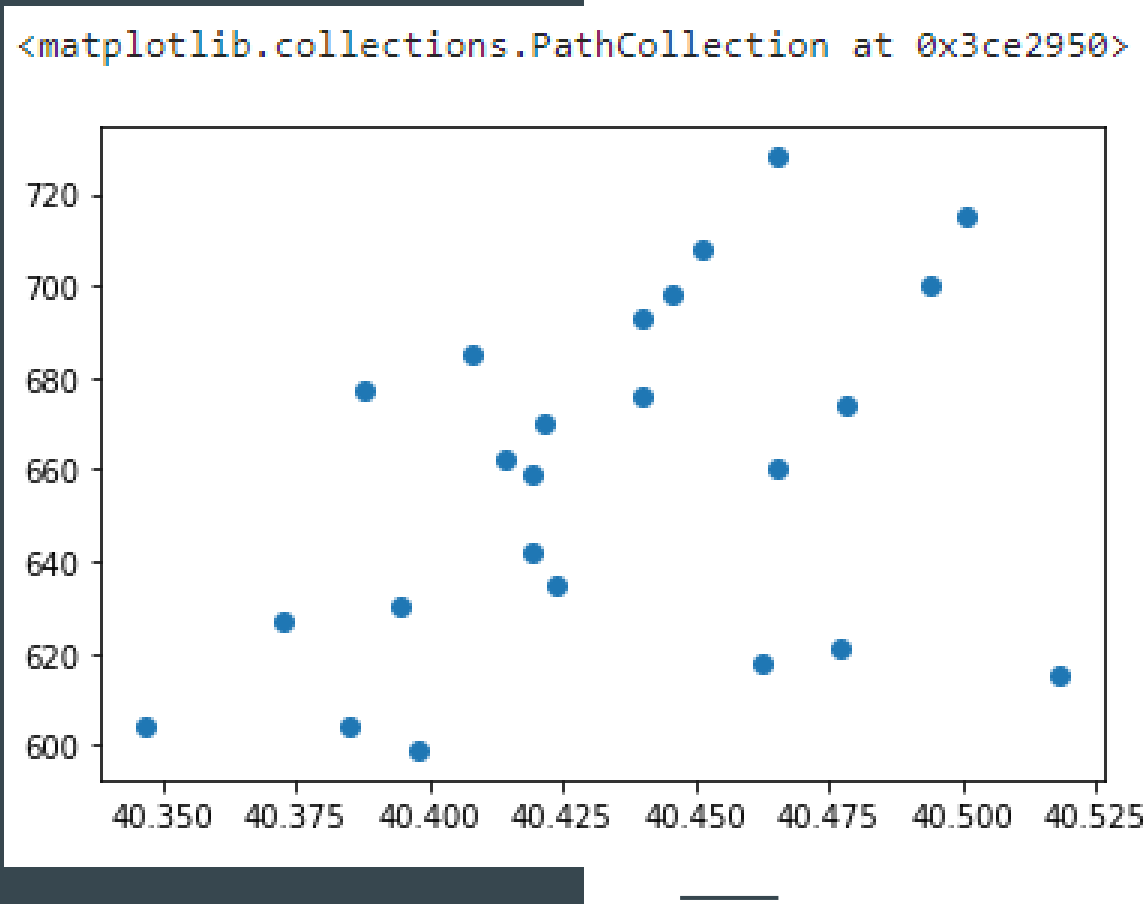
1. Missing values in other features will be replaced with respect to stations with their median values.
2. Left-over Missing values will be replaced with respect to month with their mean values.
3. Missing values in other features will be replaced with their median values

VISUALIZATION

Out[9]: <AxesSubplot:>



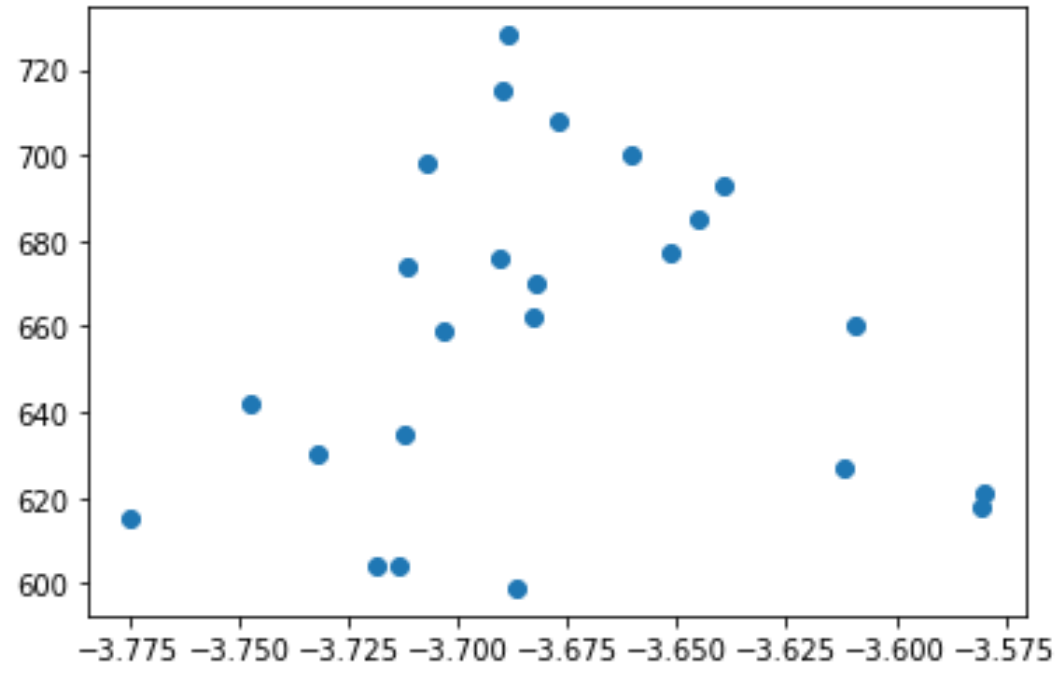
Missing Values



Scatterplot showing the relationship between Latitude and Elevation

```
plt.scatter(station_df['lon'], station_df['elevation'])
```

```
<matplotlib.collections.PathCollection at 0x3d3c3f0>
```

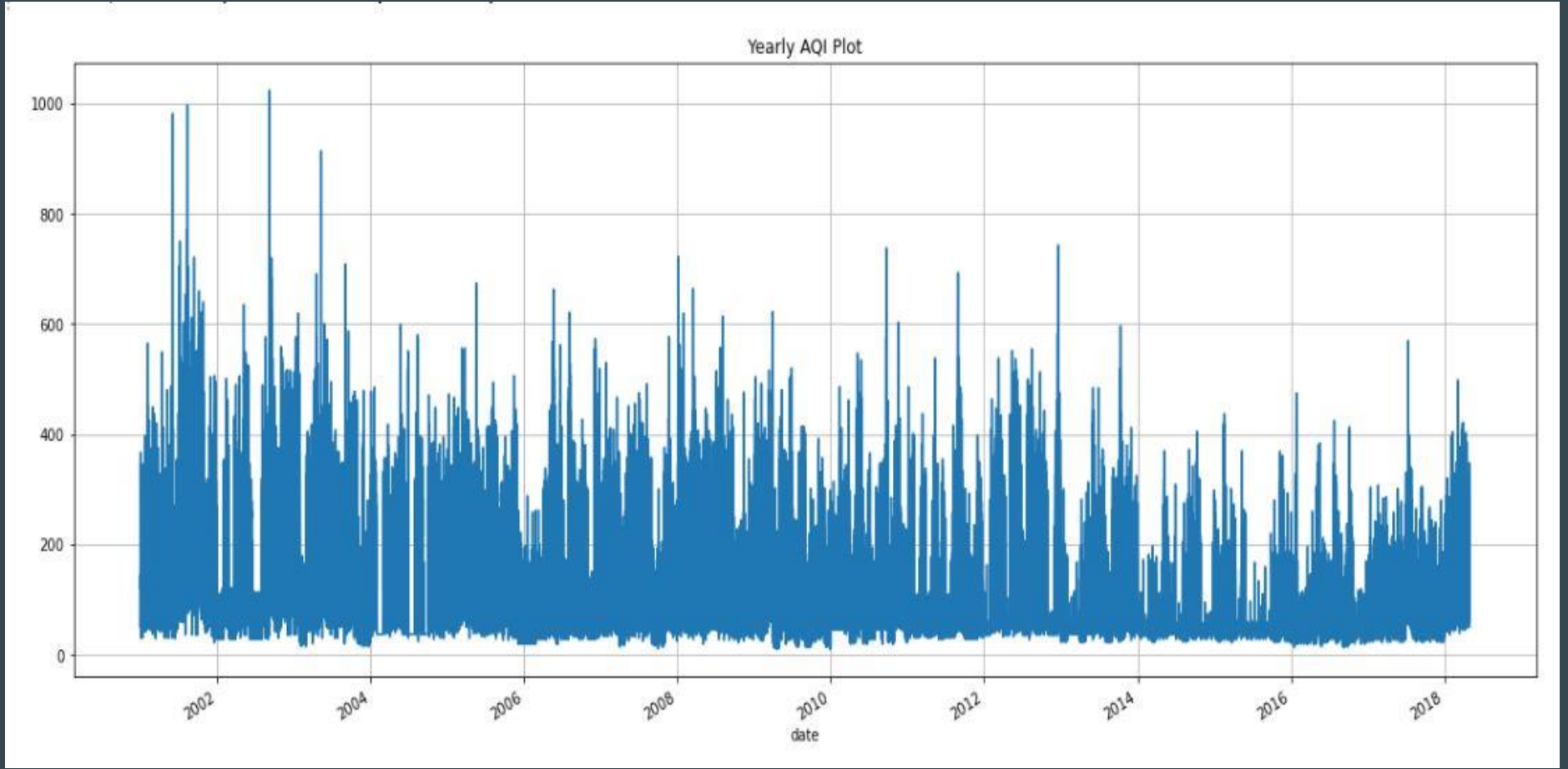


Scatterplot showing the relationship between Longitude and Elevation

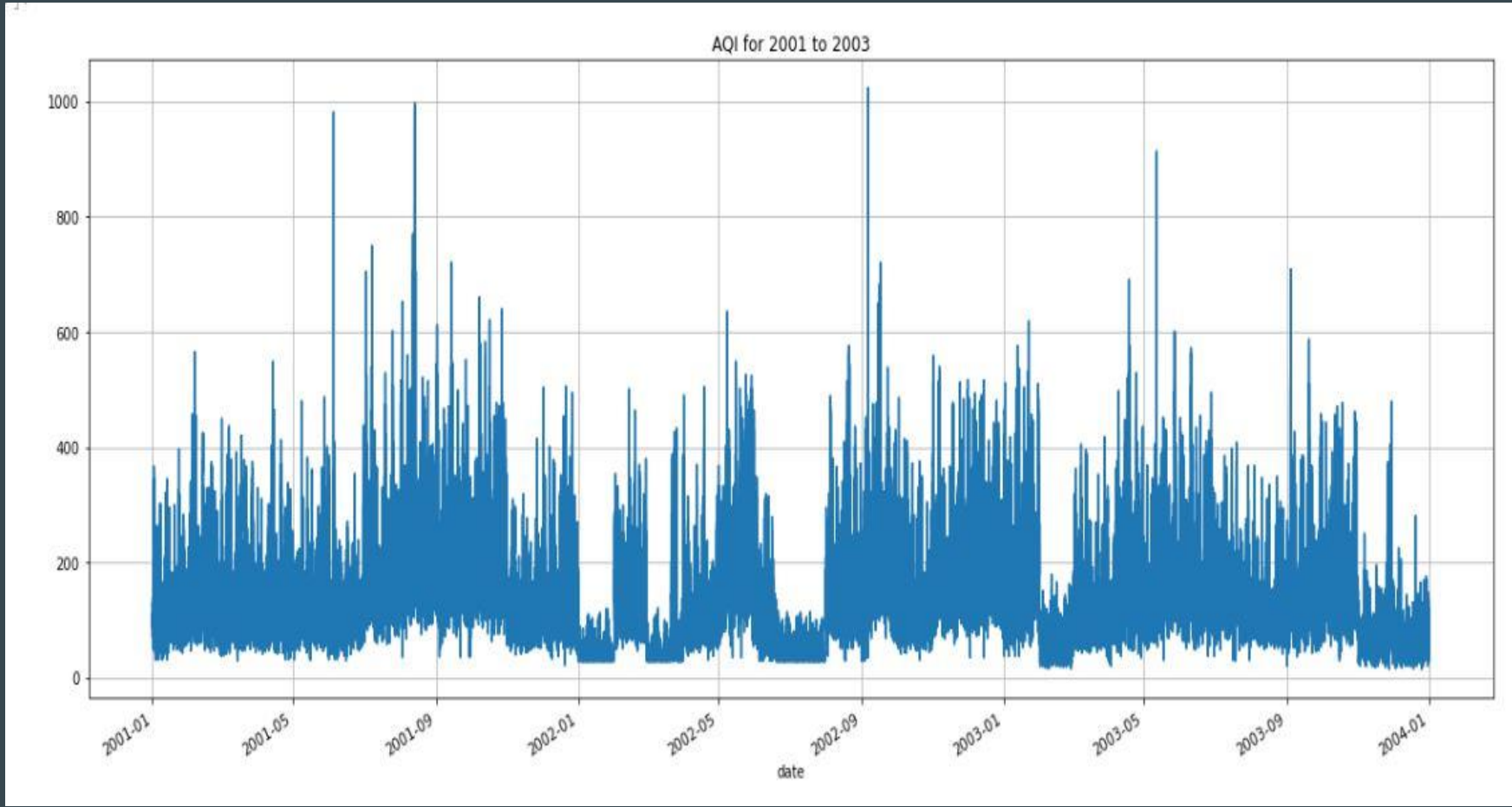
Inference

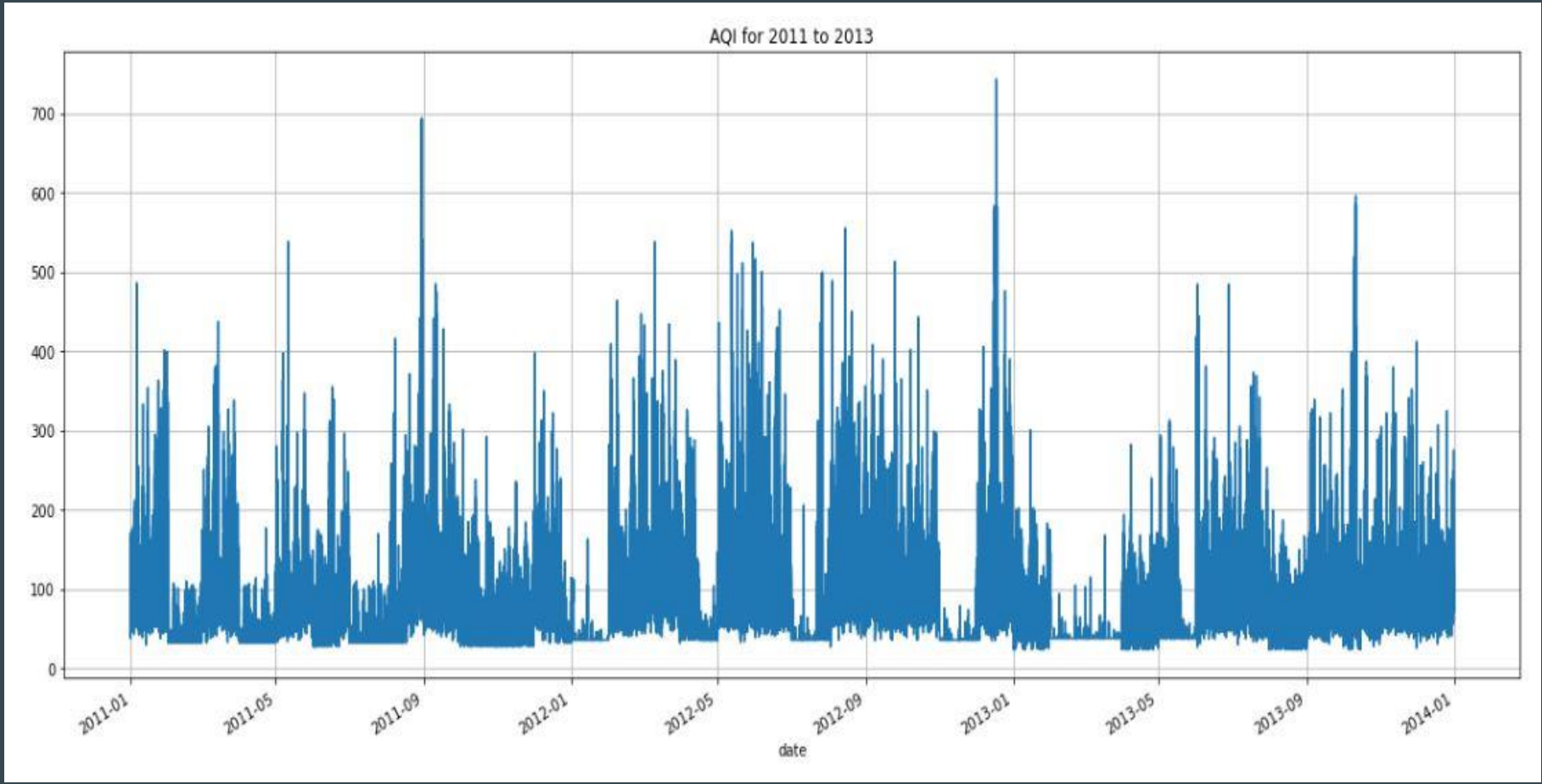
There is a weak positive correlation between the latitude and the elevation (they tend to rise together) meaning that areas with high latitude can be characterized by air quality such as lower oxygen, strong winds, frigid temperatures etc

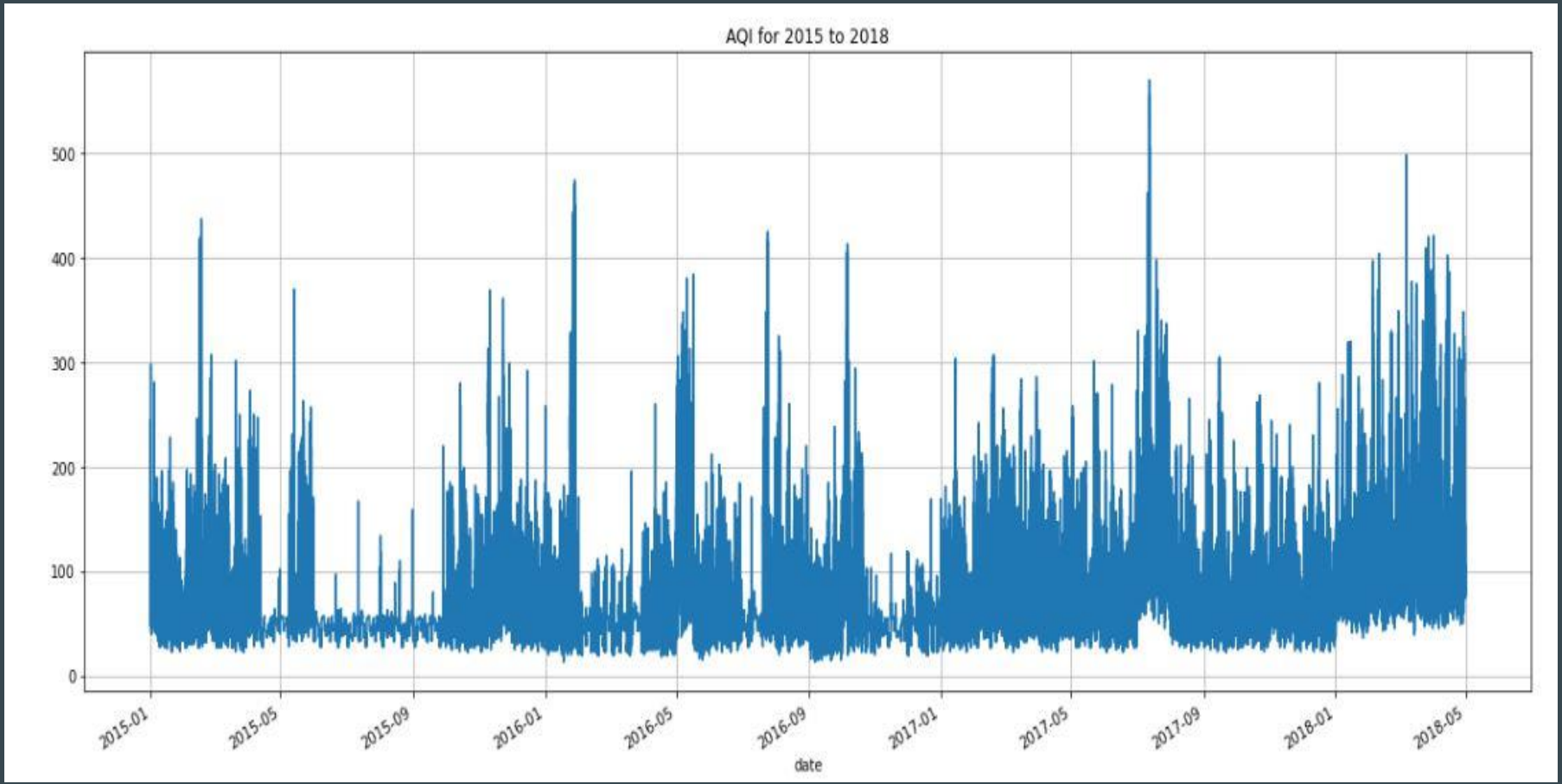
The above statement is also true for the longitude but the relationship between latitude and elevation is stronger than that of longitude and elevation.

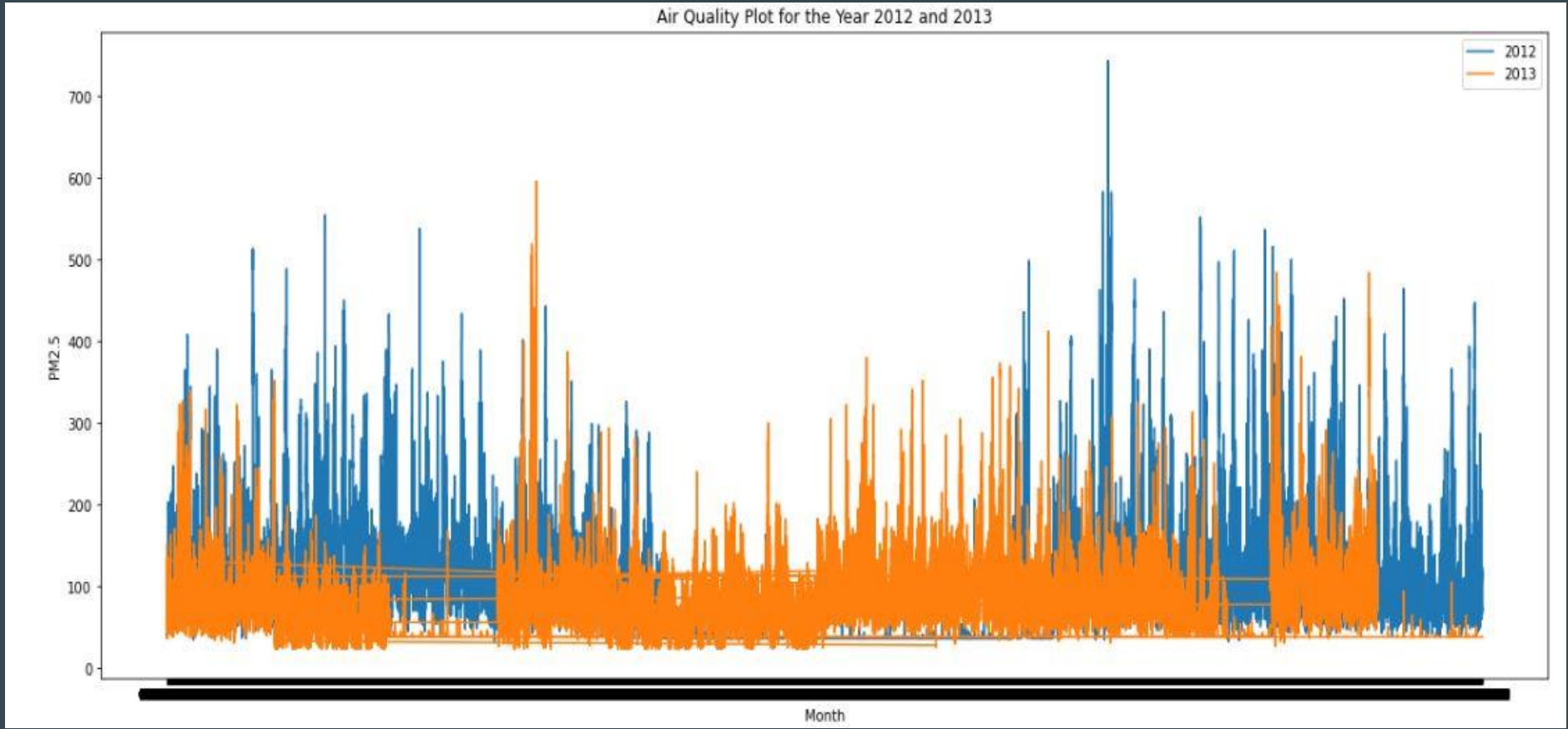


This graph shows trend of the Air Quality Index across the years

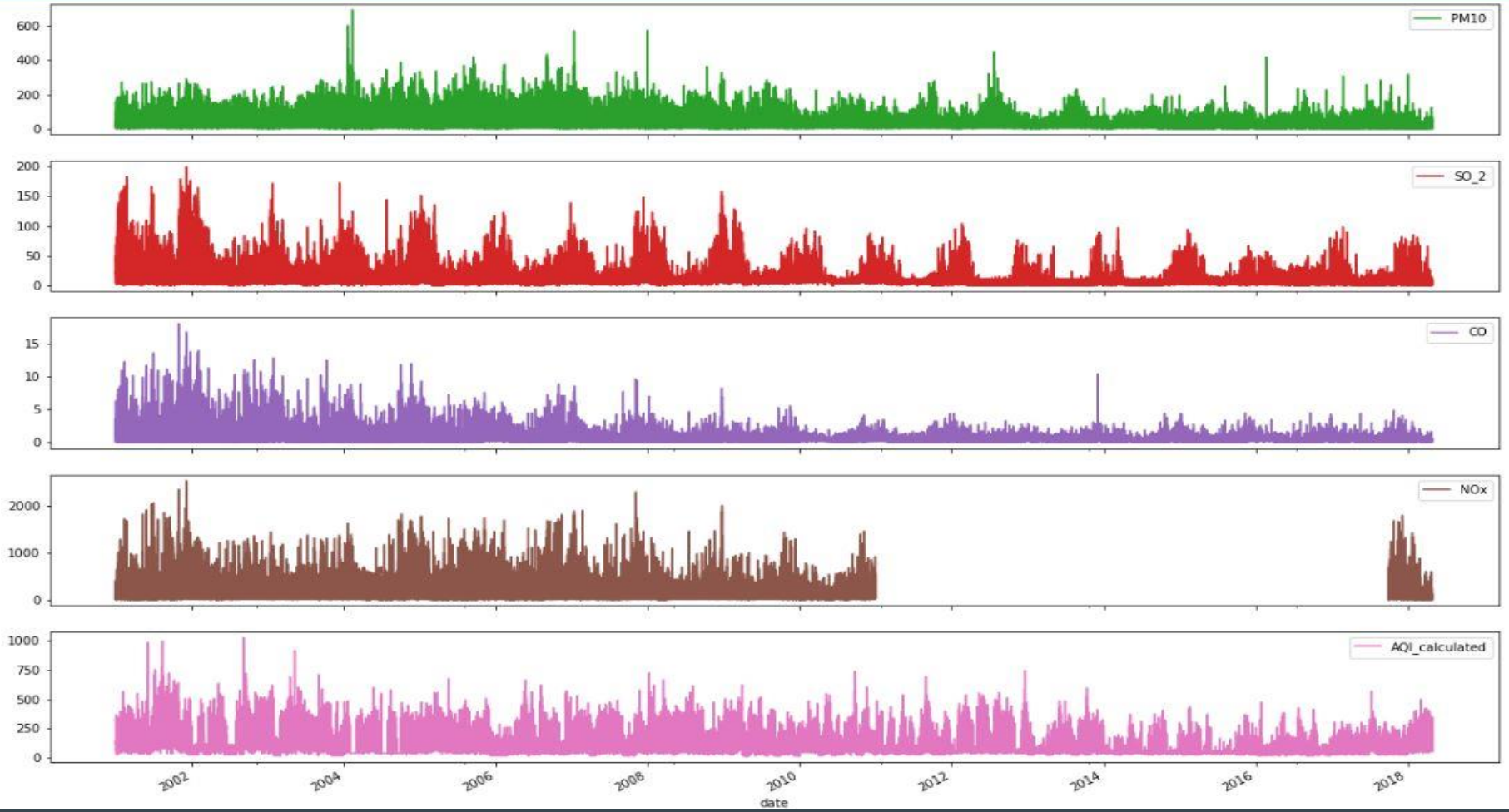

AQI for 2001 to 2003
1000
800
600
400
200
0
2001-01
2001-05
2001-09
2002-01
2002-05
2002-09
2003-01
2003-05
2003-09
2004-01
date


AQI for 2011 to 2013
700
600
500
400
300
200
100
0
2011-01
2011-05
2011-09
2012-01
2012-05
2012-09
2013-01
2013-05
2013-09
2014-01
date


AQI for 2015 to 2018
500
400
300
200
100
0
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
date

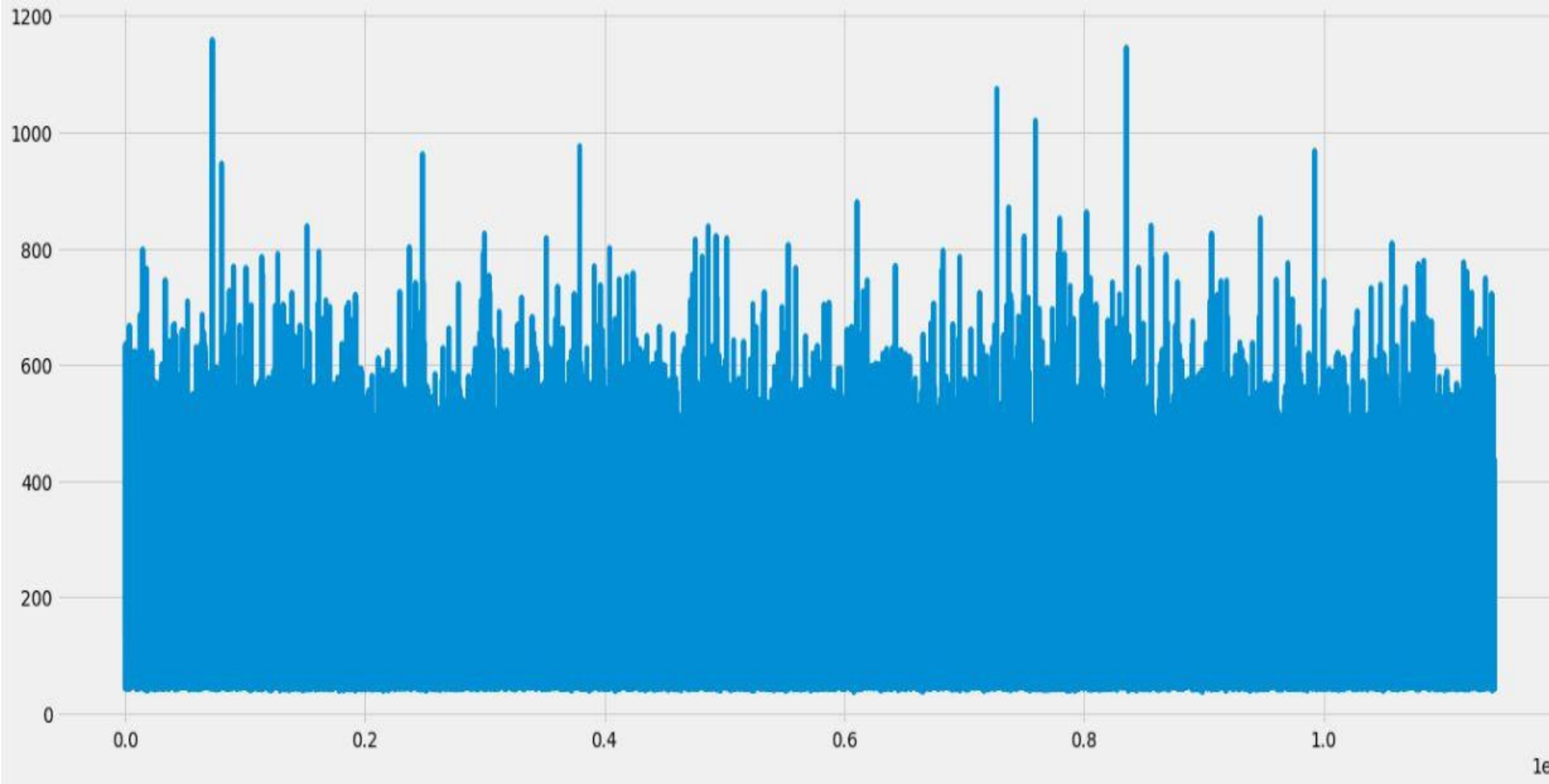


This graph shows the difference trend of the air quality index in both 2012 and 2013



Here is the trend of different gas(and the AQI) across the years

We then train a Linear Regression model with Air quality index as the target variable, we want to predict the trend of the Air Quality Index in the future



Here is the forecast of Air Quality Index in the future i.e. years after 2018, say next 5 year

Conclusion

Our linear regression model prediction showed a rise and fall in the Air Quality Index, all above 100. Research has shown that an Air Quality Index above 100 is unhealthy. As a result of this, the Government should enact laws to mitigate activities responsible for releasing these gasses into the atmosphere.