ENVIRONMENTAL SCIENCE AND TECHNOLOGY

Air Pollution Levels and Lockdown Measures: A Meta Analysis

Harshita Sharma(20171099)
Jalees Jahanzaib (2018101001)
Ojaswi Binnani(20161006)

Project Objectives

- **AQI Analysis**
  - Analysis of Air Quality Index before and after the lockdown in different cities in India
- **Find Reasons**
  - Finding the reasons for the changes in the Air Quality
- **Predicting Future Air Quality**
  - Using ML models to predict Future Air Quality
- **Data Visualisation**

Air Quality Index

- The **air quality index** (**AQI**) is an index for reporting air quality on a daily basis. It is a measure of how air pollution affects one's health within a short time period.
- A web-based system is designed to provide AQI on real time basis. It is an automated system that captures data from continuous monitoring stations without human intervention, and displays AQI based on running average values (e.g. AQI at 6am on a day will incorporate data from 6am on previous day to the current day).
- For manual monitoring stations, an AQI calculator is developed wherein data can be fed manually to get AQI value.

Air Quality Index

- There are 6 categories of the air have been created in this air quality index.

Good (0–50)	Minimal Impact	**Poor (201–300)**	Breathing discomfort to people on prolonged exposure
Satisfactory (51–100)	Minor breathing discomfort to sensitive people	**Very Poor (301–400)**	Respiratory illness to the people on prolonged exposure
Moderate (101–200)	Breathing discomfort to the people with lung, heart disease, children and older adults	**Severe (>401)**	Respiratory effects even on healthy people

Calculation of AQI

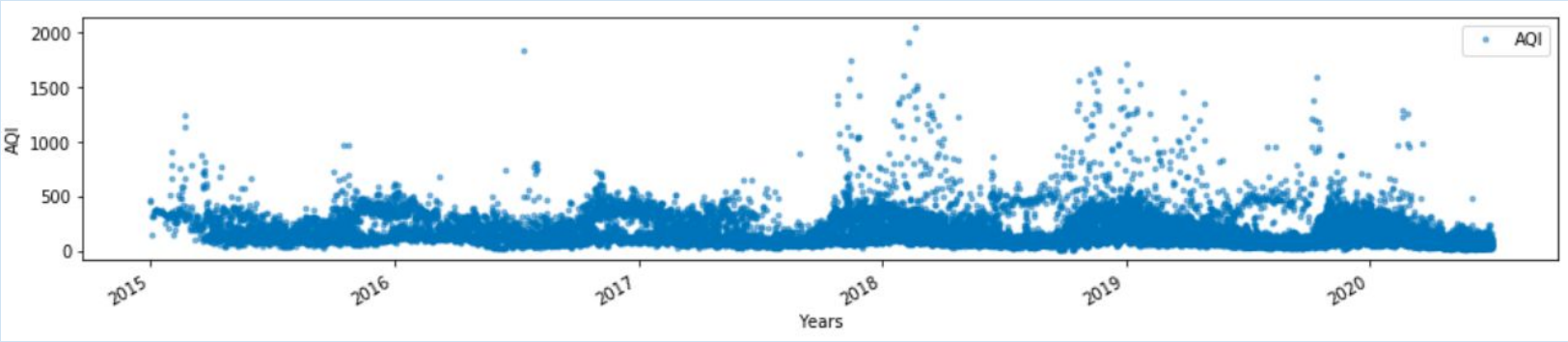
- The air quality index (AQI) is defined as ratios of the measured concentration of the atmospheric pollutants to their standard prescribed values. A general formula to compute an AQI is the following:

$$AQI_{\text{pollutant}} = \left(\frac{\textit{pollutant concentration reading}}{\textit{Standard Concentration}} \right) x\ 100$$

- The AQI calculation uses 7 measures: PM2.5(Particulate Matter 2.5-micrometer), PM10, SO2, NOx, NH3, CO and O3(ozone). Sometimes measures are not available due to lack of measuring or lack of required data points.
- Final AQI is the maximum Sub-Index with the condition that at least one of PM2 and PM10 should be available and at least three out of the seven should be available.

AQI Analysis

1. Visualising AQI over the years 2015-2020(mid-july)



AQI Analysis

2. Visualising AQI - Most Polluted Cities

- AQI mean over the years to find the most polluted cities(on right)
- AQI of some other major cities:

14	Chennai	114.500000
15	Hyderabad	109.210000
16	Mumbai	105.350000
20	Bengaluru	94.320000

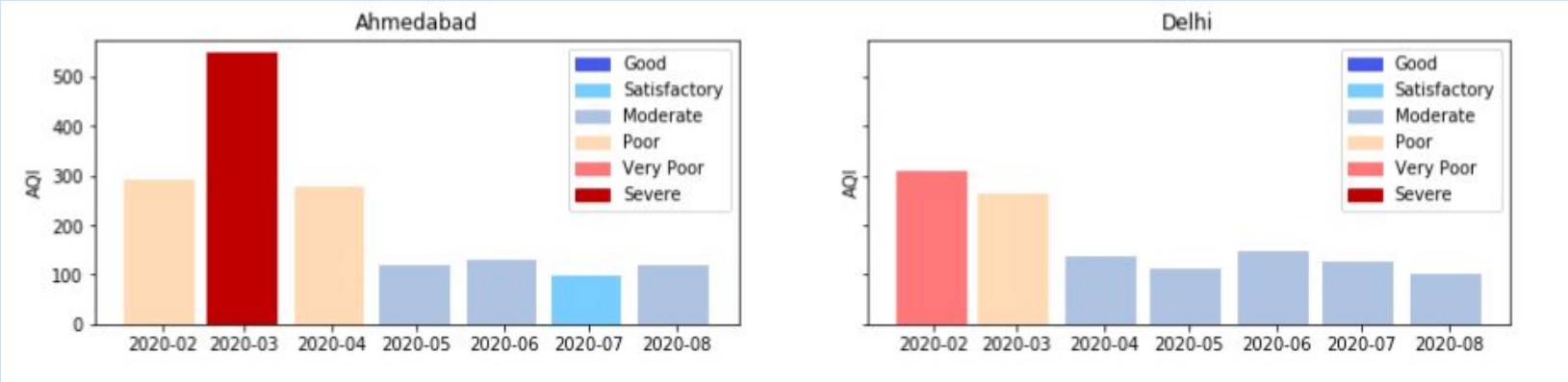
Good	0-50
Satisfactory	51-100
Moderate	101-200
Poor	201-300
Very Poor	301-400
Severe	>400

	City	AQI
0	Ahmedabad	452.120000
1	Delhi	259.490000
2	Patna	240.780000
3	Gurugram	225.120000
4	Lucknow	217.970000
5	Talcher	172.890000
6	Jorapokhar	159.250000
7	Brajrajnagar	150.280000
8	Kolkata	140.570000
9	Guwahati	140.110000

AQI Analysis

3. Some major cities and their AQI: Ahmedabad, Delhi, Mumbai, Kolkata, Hyderabad, Chennai

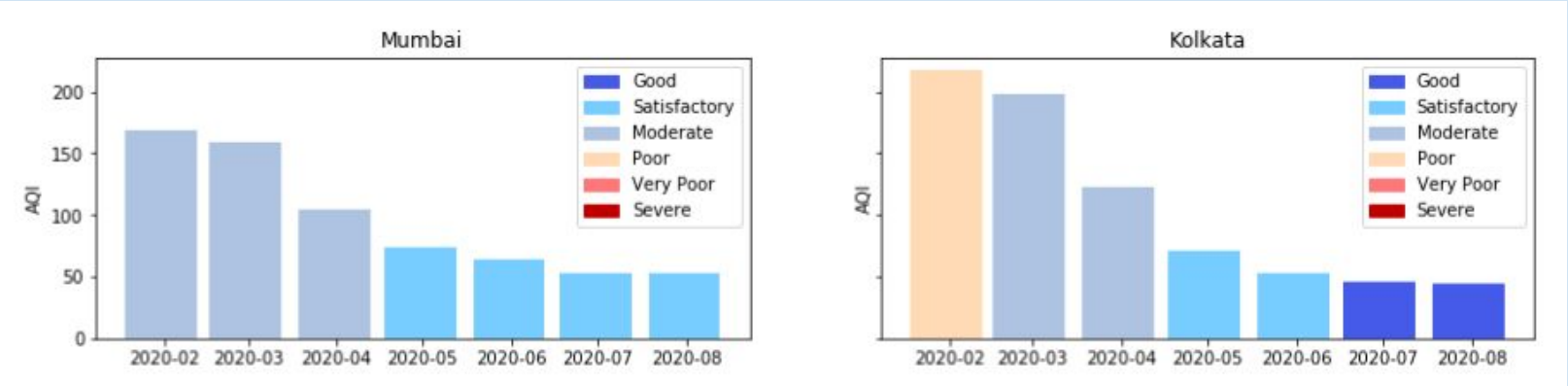
- AQI in the year 2020



AQI Analysis

3. Some major cities and their AQI: Ahmedabad, Delhi, Mumbai, Kolkata, Hyderabad, Chennai

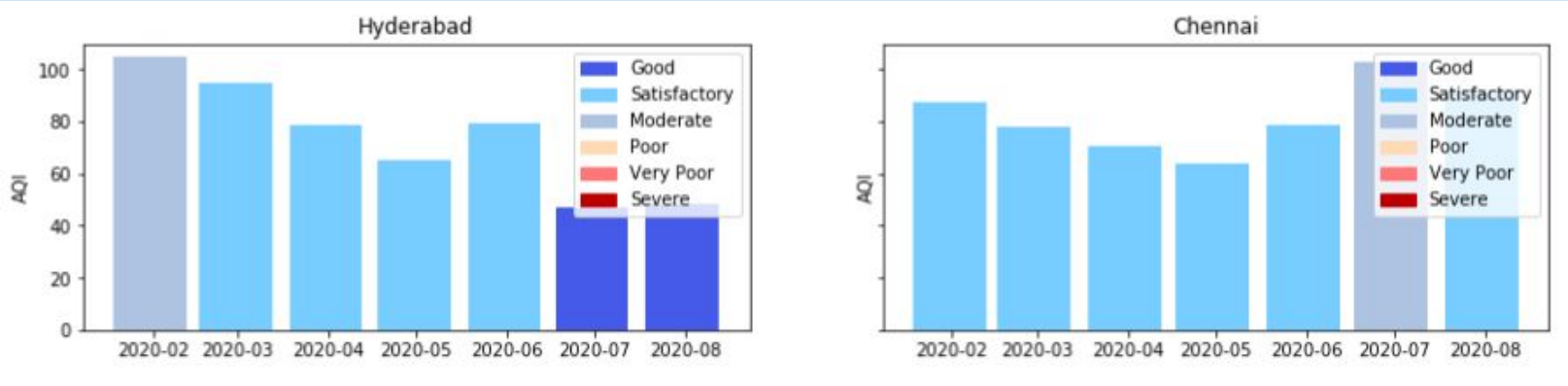
- AQI in the year 2020



AQI Analysis

3. Some major cities and their AQI: Ahmedabad, Delhi, Mumbai, Kolkata, Hyderabad, Chennai

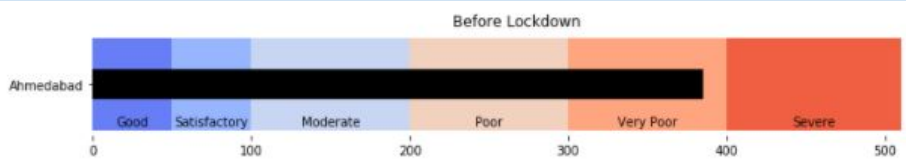
- AQI in the year 2020

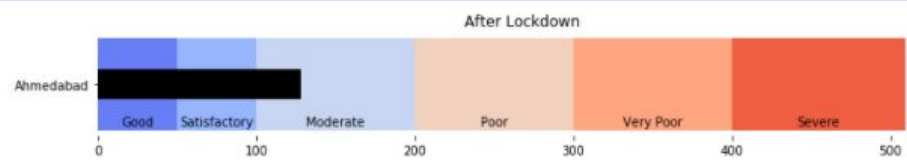


AQI Analysis

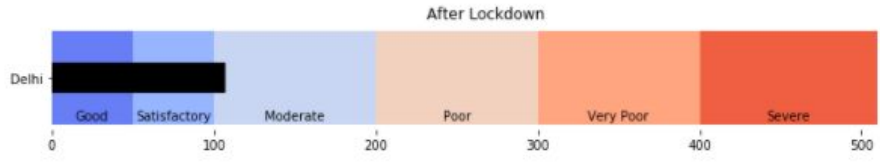
3. Some major cities and their AQI: Ahmedabad, Delhi, Mumbai, Kolkata, Hyderabad, Chennai

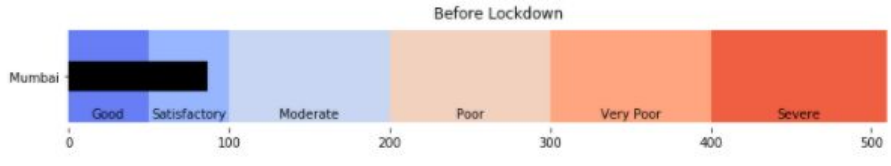
- Before and during Lockdown
  - Analysed using Bullet Graphs
  - Based on the six AQI categories - Good, Satisfactory, Moderate, Poor, Very poor, Severe

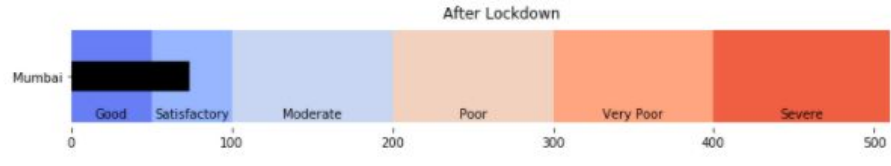

Before Lockdown
Ahmedabad
Good
Satisfactory
Moderate
Poor
Very Poor
Severe
0
100
200
300
400
500


After Lockdown
Ahmedabad
Good
Satisfactory
Moderate
Poor
Very Poor
Severe
0
100
200
300
400
500

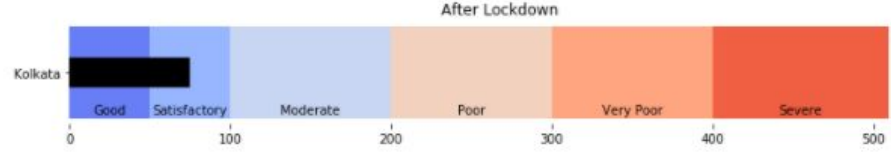

Before Lockdown
Delhi
Good
Satisfactory
Moderate
Poor
Very Poor
Severe
0
100
200
300
400
500


After Lockdown
Delhi
Good
Satisfactory
Moderate
Poor
Very Poor
Severe
0
100
200
300
400
500

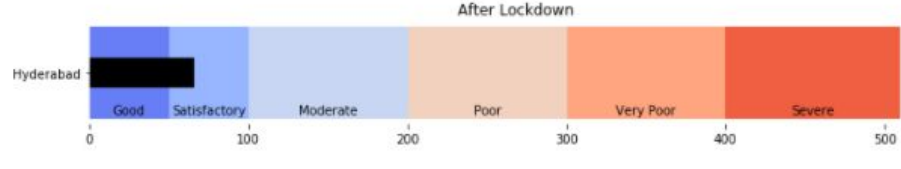

Before Lockdown
Mumbai
Good
Satisfactory
Moderate
Poor
Very Poor
Severe
0
100
200
300
400
500


After Lockdown
Mumbai
Good
Satisfactory
Moderate
Poor
Very Poor
Severe
0
100
200
300
400
500

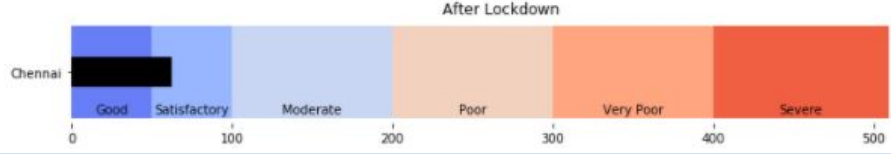

Before Lockdown
Kolkata
Good
Satisfactory
Moderate
Poor
Very Poor
Severe
0
100
200
300
400
500


After Lockdown
Kolkata
Good
Satisfactory
Moderate
Poor
Very Poor
Severe
0
100
200
300
400
500


Before Lockdown
Hyderabad
Good
Satisfactory
Moderate
Poor
Very Poor
Severe
0
100
200
300
400
500


After Lockdown
Hyderabad
Good
Satisfactory
Moderate
Poor
Very Poor
Severe
0
100
200
300
400
500


Before Lockdown
Chennai
Good
Satisfactory
Moderate
Poor
Very Poor
Severe
0
100
200
300
400
500


After Lockdown
Chennai
Good
Satisfactory
Moderate
Poor
Very Poor
Severe
0
100
200
300
400
500

Data Preprocessing

1. Handling Missing Values

- Deleting rows with NULL values

Pros and Cons:

- Complete removal of data with missing values results in robust and highly accurate model. Deleting a particular row or a column with no specific information is better, since it does not have a high weightage.
- Loss of information and data and works poorly if the percentage of missing values is high (say 30%), compared to the whole dataset

Data Preprocessing

1. Handling Missing Values

- Deleting rows with NULL values: The error generated was less but the data got distorted and the AQI values strayed away from the trend.
- Another issue with the approach was that some months did not have any values at all - so when we grouped the data month-wise(next step) - it still increased the NULL values - hence the model failed for most of the cities.
- So this approach was discarded.

Data Preprocessing

1. Handling Missing Values

- Filling Missing Values in Database with most frequent values (i.e. Mode)
- **Filling Missing Values in Database with average of all the values (i.e. Mean)**
- Filling Missing Values in Database with values separating the higher half from the lower half of a data sample (i.e. Median)

Pros and Cons:

- Prevent data loss which results in removal of the rows and columns
- Imputing the approximations add variance and bias

Data Preprocessing

2. Changing data structure to City-wise Monthly data

For ease in: Observing Data for a particular city and observing Data for a particular year

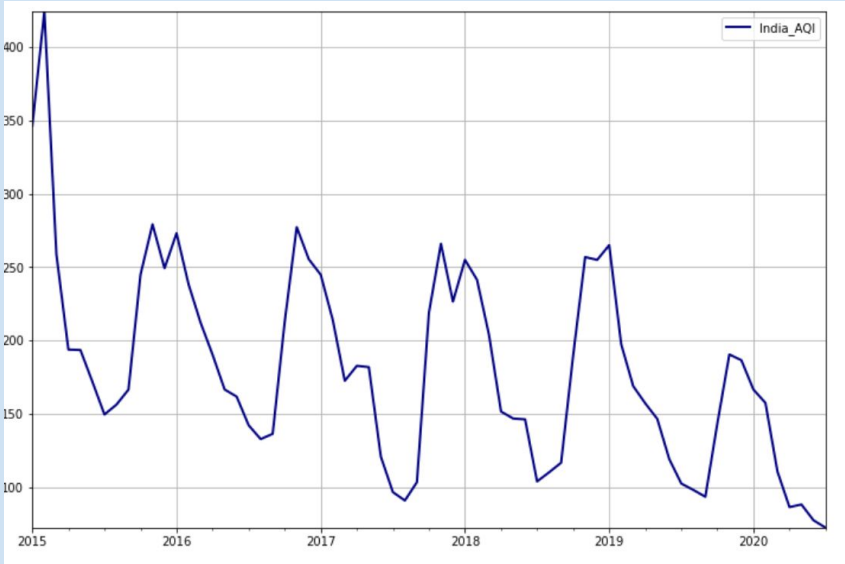
	Ahmedabad_AQI	Aizawl_AQI	Amaravati_AQI	Amritsar_AQI	Bengaluru_AQI	Bhopal_AQI	Brajrajnagar_AQI	Chandigarh_AQI	Chennai_AQI	Coimbat
2020-03-01	344.645161	66.105263	52.548387	95.387097	90.741935	122.709677	152.354839	57.096774	70.290323	85
2020-04-01	121.900000	40.300000	44.400000	59.866667	68.533333	108.400000	140.666667	44.233333	63.500000	82
2020-05-01	129.774194	24.193548	59.096774	77.677419	73.161290	104.451613	144.161290	74.129032	78.677419	57
2020-06-01	98.066667	20.800000	47.866667	101.533333	55.166667	71.666667	113.033333	66.500000	103.066667	39
2020-07-01	119.000000	20.000000	54.000000	78.000000	43.000000	69.000000	76.000000	66.000000	92.000000	33

Methodology

Which Models should we use?

Two Observations from the AQI plot of India:

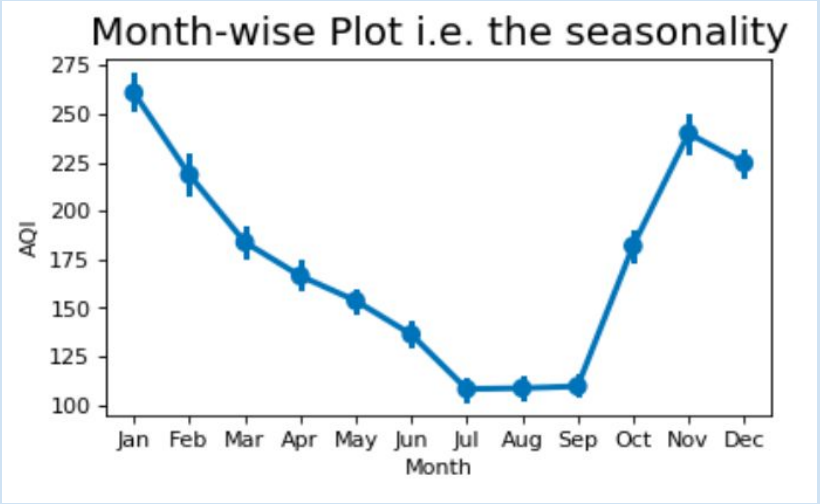
- Presence of trend in AQI over the years
- Seasonal Component plays a Important Role

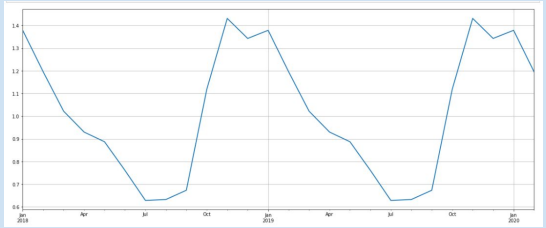


Methodology

What is Seasonal Effect?

Seasonality: The repeating short-term cycle in the data. There is an abnormal rise in the AQI values during the winter because of Winter inversion, Valley effect and other factors such as dust storms, crop fires, burning of solid fuels for heating and firecracker-related pollution during diwali or stubble burning.





Methodology

What is Seasonal Effect?

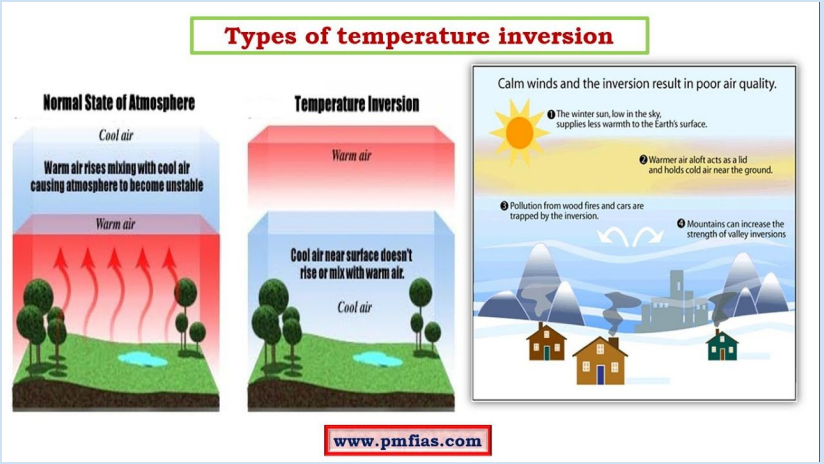
Winter Inversion: In summer, air in the planetary boundary layer is warmer and lighter, and rises upwards more easily. This carries pollutants away from the ground and mixes them with cleaner air in the upper layers of the atmosphere however during winter, the planetary boundary layer is thinner as the cooler air near the earth's surface is dense. The cooler air is trapped under the warm air above it which forms a kind of atmospheric 'lid'. This phenomenon is called Winter Inversion.

Methodology

What is Seasonal Effect?

Valley Effect: When concentration of pollutants increases in low lying areas such as valleys because of certain weather conditions such as winter when cold air containing pollutants generated from vehicular emission becomes trapped by a layer of warmer air above the valley. This phenomenon is known as Valley Effect.

The longer this air inversion lasts the worse the quality of air gets

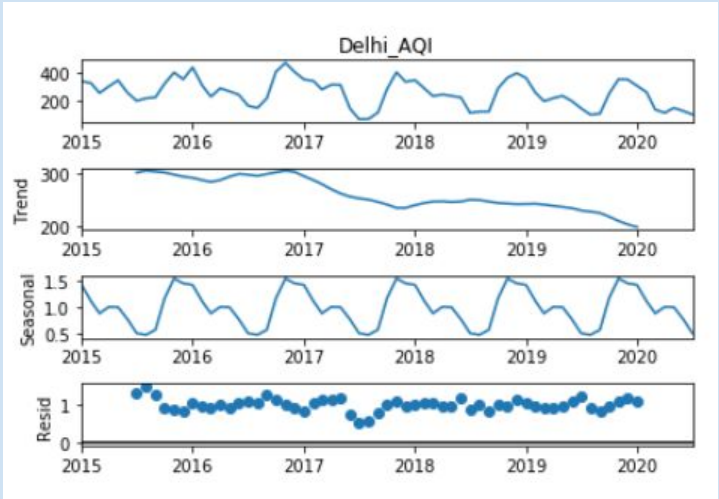


Methodology

What is Seasonal Decomposition?

To check if the data has seasonality throughout. It also describes the trends in the Data. The Idea here was to use a **Time Series Model**. We have Identified all required parameter such as

- **Level**: The average value in the series.
- **Trend**: The increasing or decreasing value in the series.
- **Seasonality**: The repeating short-term cycle in the series.
- **Noise/Residue**: The random variation in the series.



Methodology

Multiplicative Model for Seasonal Decomposition

- It suggests that the components are multiplied together as follows:

```
y(t) = Level * Trend * Seasonality * Noise
```

- Nonlinear, such as quadratic or exponential.
- Changes increase or decrease over time.
- **Why not linear model?** If linear model, dimensions will increase hence complexity of the model will also increase

Methodology

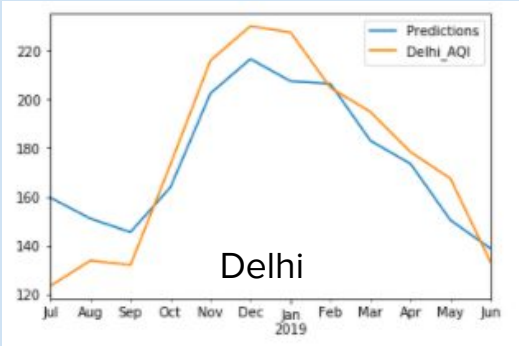
Model 1: SARIMAX

- Autoregressive Integrated Moving Average, or ARIMA, is one of the most widely used forecasting methods for univariate time series data forecasting.
- Although the method can handle data with a trend, it does not support time series with a seasonal component.
- An extension to ARIMA that supports the direct modeling of the seasonal component of the series is called SARIMA.
- SARIMAX is an extension of the SARIMA model that also includes the modeling of exogenous variables (covariates and can be thought of as parallel input sequences that have observations at the same time steps as the original series).

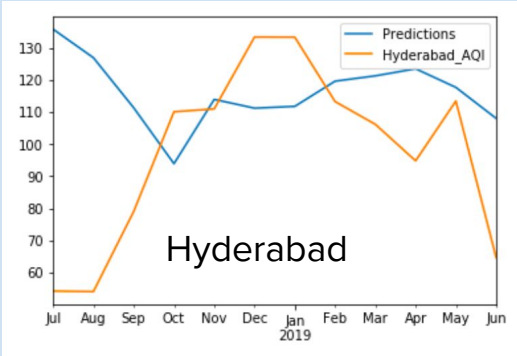
Predicting Values

Model 1: SARIMAX

- Using data ranging from 2015-2018 as training data, we predict the AQI values for 2019.



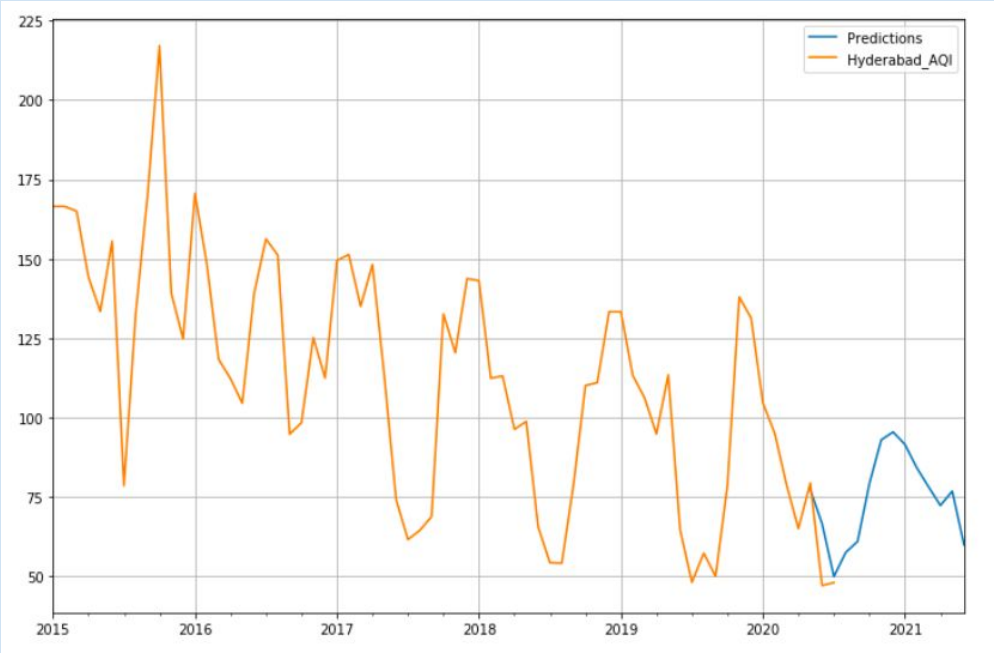

```
Root Mean Squared Error:  16.125559620875286
Mean AQI: 176.2514063089236
Mean Absolute Error: 13.660681
Bias: 1.327311
```




```
Root Mean Squared Error:  37.85122379249102
Mean AQI: 97.27746415770609
Bias: -18.947500
```

Predicting the Unknown(Year 2021):

- We can see the predictions plotted in continuation with 2020
- One thing we note is the highly optimistic prediction.
- That is purely due to the fact that 2020 is such an outlier. So if lockdown were to continue, the air will continue to have better quality.
- When lockdown is removed: chances are, the pollution levels will follow the trend pre 2020 which would mean a bump in the AQI levels unless the city decides to keep the restrictions etc as is which is highly unlikely.



Methodology

Model 2: Recurrent Neural Network: Sequential Model

A type of Neural Network which is used for time based/frequency based/memory based data like text data, speech, time series etc. We will be using a particular cell type LSTM (Long Short term memory).

Advantages:

- LSTM networks are particularly meant to keep particular information for a longer term as compared to regular RNN's.

Disadvantages:

- RNN is a black box method, which means there is little transparency in the model and how it trains.
- Also there is the high complexity of hyperparameters(initial parameters before training,more).

Methodology

Model 2: Recurrent Neural Network

Min Max Scaler: (scaler = MinMaxScaler()): Transform features by scaling each feature to a given range. Here we are scaling and translating each feature individually such that it comes inside the range zero and one.

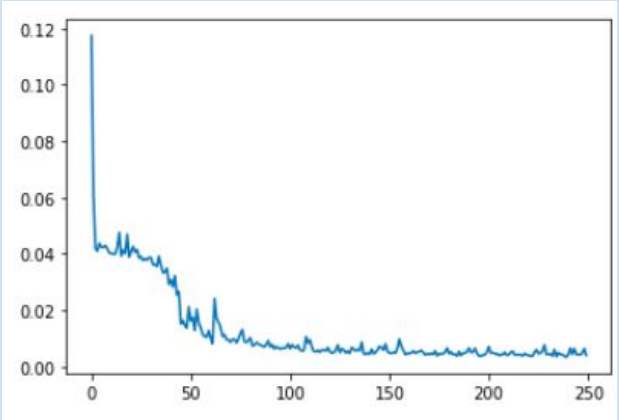
Time Series Generator: (generator = TimeSeriesGenerator(scaled_train, scaled_train, length=n_input, batch_size=1): Using it we are creating an instance of the class where input and output series are same which is scaled_input. Now **we will use it to train a our neural network model**. This class both informs what the model will learn and how we intend to use the model in the future when making predictions.

Methodology

Model 2: Recurrent Neural Network

Fitting the Model: model.fit_generator(generator,epochs=250)

- Here we are fitting our model with the timeseries generator created earlier. Here epochs represent the number of passes of the entire training dataset that our model has done to get trained.
- As each epoch gets over we observed that loss (error over the training set) reduces.

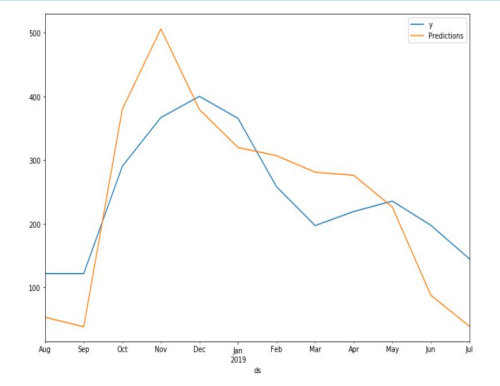


Predicting Values

Model 2: RNN

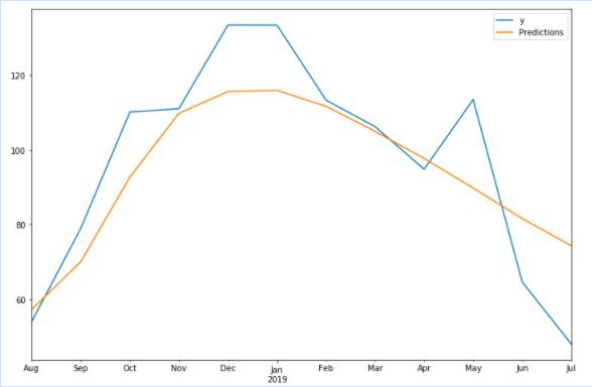
- Using data ranging from 2015-2018 as training data, we predict the AQI values for 2019.

Delhi



RMSE = 82.7404628036889

Hyderabad



RMSE = 16.625953834298937

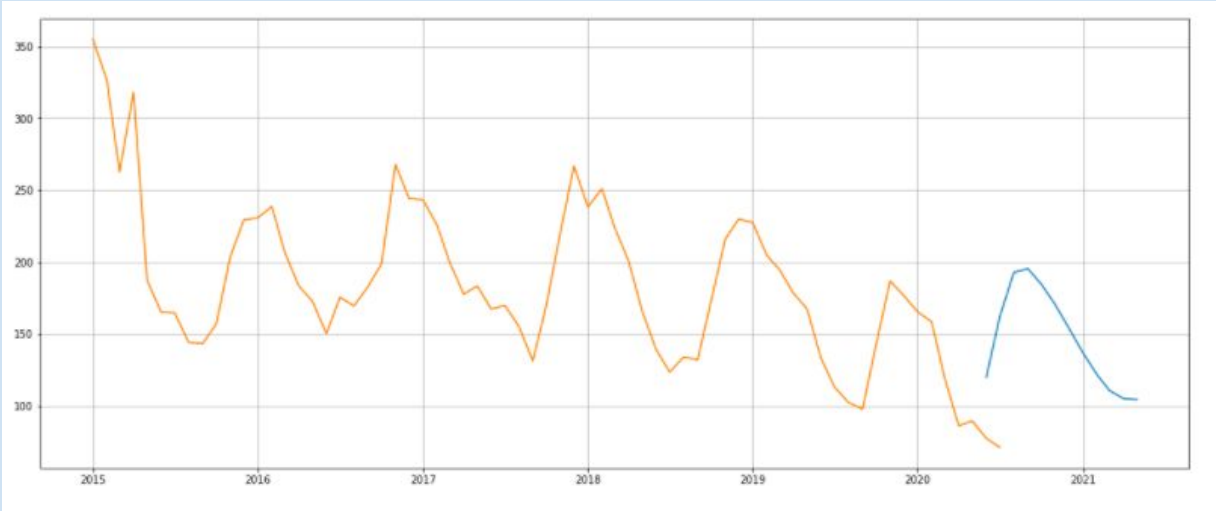
Predicting Values

Model 2: RNN

- It is obvious that RMSE for Delhi is very high using RNN but if apply the model on a different city perform better than SARIMAX.
- Ex. Bangalore
  - SARIMAX RSME = 25
  - RNN RSME =22
- Hyderabad
  - SARIMAX RSME = 37
  - RNN RSME = 16

Predicting the unknown - year 2021

Model 2: Recurrent Neural Network: For Delhi



SARIMAX vs RNN

Which one performed better?

- For Smaller Ranges RNN Model performs better
- For Big Ranges SARIMAX performs better

Conclusion

Thank you!