Air Quality analysis and prediction using Machine Learning on Delhi

- Abstract:

Examining and protecting air quality has become one of the most essential activities for the government in many industrial and urban areas today. Air quality compromise is one of the burning issues in the present scenario. Air is contaminated by the arrival of dangerous gases into the climate from the industries, vehicular emissions, etc. Nowadays, air pollution has reached critical levels, and the level of pollution in many large cities has exceeded the government's air quality index value. It has a significant impact on human health. Due to the availability of environmental sensing networks and sensor data, many academics have recently begun to apply Big Data Analytics. Machine learning technology has advanced to the point that it is now able to anticipate contaminants based on historical data. In this research, we have used two Machine learning approaches: the Prophet Model and SARIMAX Model to forecast the Air pollution in Delhi for the year 2020. Initially we have combined the related data of many cities but found the best scope for analysis in Delhi.The parameters like pollutants (PM2.5 ,NOx, NO, NO2, SO2, NH3, SO2, CO, O3, C6H6 and PM10) and the respective date-time is tracked from 2015 to 2020. This time series data is then studied in detail and made observations and conclusions based on the seasonality and trends seen in the data.

- Keywords:

Air Quality, Machine Learning, Big Data Analytics, Prophet, SARIMAX

- Introduction:

Air pollution is a major issue around the world, particularly in smart cities. Air pollution is defined as any material from any atmospheric source that exists in any form, whether liquid, solid, or gas phase, and causes destruction or the ability to alter the usual characteristics of the atmosphere, thereby increasing the health risk to living things or disrupting the environment and ecosystem.Accelerated expansion, urbanisation, and improved lifestyles have significantly increased air pollution in urban areas in recent years. The regulation of pollution has attracted public attention. New Delhi, India's capital, is one of the most polluted cities on the planet. Various studies have been undertaken to examine air pollution trends and their potentially dangerous repercussions, such as one recent study that found that pollutant concentrations in Delhi are significantly higher than the legal levels. As a result, the residents' life expectancy has decreased by six years. Another study looked at the effects of traffic pollution on human health.These findings indicate that air quality monitoring and regulation will be required in the future. The air quality index, or AQI, is a numerical representation of how dirty the air is at any particular time. A timely AQI forecast would also reveal the likely air quality trend, allowing the government to take more effective and efficient corrective measures. Machine learning is the science of creating programmes that 'learn' from their surroundings and adjust accordingly. The release of hazardous suspended particulates into the atmosphere causes air pollution. In most highly populated locations, it is a big worry. Every day, new diseases are discovered, and each year, more deaths are documented as a result of air pollution. The rise in air pollution levels can be attributed to a

number of factors, including industrialization and globalization. Many air quality monitoring stations have been built up around the world. In recent years, air quality monitoring utilising soft computing approaches has yielded a wealth of study analysis.A machine learning algorithm learns from its past experiences, improvises, adapts to changes, and improves the efficiency of the work at hand. As a result, machine learning approaches can be quite useful in constructing prediction models for air pollution forecasting.Air is one of the most important and life-sustaining factors in the environment. It contains a large number of different particles that, while not visible, aid in the maintenance of life. However, as science and technology advanced, particularly in the manufacturing and automobile industries, the greatest collateral effect was a decline in air quality or pollution.
Below are some of the most common causes of air pollution [1].
1.Exhaust from industry Emissions of hazardous gases such as SO2 and NOx from thermal power plants at Rajghat, Badarpur, Indraprastha, and other industrial areas contribute to the principal pollutants in Delhi's air.
2.Emissions from vehicles is the biggest contributors to the deterioration of Delhi air quality are traffic congestion and automobile emissions. According to data obtained from the Delhi government's transport department up to December 31, 2016, the total number of registered vehicles was 1, 01,06,791. Motorbikes and scooters account for the most registered vehicles in the city, with 63,40,136. These are regarded as significant contributors to air pollution.
The following contaminants are found in high concentration in Delhi's air[2]: - 1. Particulate Matter (PM2.5 and PM10), RSPM and SPM: Vehicle emissions, notably from heavy motor diesel vehicles, kerb-side dust, thermal power plants, industrial, and residential combustion processes, are the main sources of particulate matter in Delhi. NOx (nitrogen oxide): Nitrogen oxides are created in industrial combustion processes and are mostly found in vehicle exhaust. Because of transportation, NOx levels are highest in metropolitan areas. It is a key component in the formation of photochemical smog, which blankets the metropolitan environment in a haze. Benzene, toluene, and carbon monoxide are also included.
We've chosen Delhi as our research location. According to the World Health Organization (WHO) in 2014, Delhi is the most polluted city on the planet. High industrial and traffic emissions, construction work, and crop burning in neighbouring states all contributed to a 44 percent increase in fine particulate matter (PM) levels in Delhi. In Delhi, the level of airborne particulate matter (PM2.5) is extremely high. It is one of the most dangerous contaminants to human health. We're looking at statistics from Delhi (air pollution) between 2015 and 2020. We'll compare the statistics (air pollution) before and after the lockout, as well as the trends and seasonality.

- Related Work:

C. Srivastava et al. [3] presented research on air pollution estimation in Delhi utilizing IoT and Big Data analytics combined with machine learning based on meteorological parameters such as wind speed (WS), vertical wind speed (VWS), wind direction (WD), temperature (Temp) and relative humidity (RH). The dataset included data from three Delhi air pollution monitoring sites on hazardous particles such as PM2.5, PM10, and gases (O3, NO2, SO2, CO). Each station had three cases: the first case was for PM2.5 AQI, the second case was for PM10 AQI, and the third case was for gas AQI. Various machine learning methods were used to produce an accurate prediction model, which was then evaluated using the Mean

Square Error (MSE), Mean Absolute Error (MAE), and R2 metrics. In terms of overall performance, Neural Networks(MLP) and SVR were the best choices. Though the models predicted air quality with high accuracy, the model's potential is limited by the data set's short length. More meteorological features should've been addressed as well.

A. K. Sharma et al. [4] presented a study describing the extent of air pollution in Delhi and the magnitude of health problems inferred by it. The data was collected for 28 years on ambient air quality from a variety of locations all over Delhi. The parameters studied are varied in aerodynamic diameter (PM10) and those ≤2.5 μm in aerodynamic diameter (PM2.5) and others on suspended particulate matter (SPM) and respirable suspended particulate matter (RSPM). The existing knowledge gaps are highlighted. The exposure assessment, monitoring of physiological functional parameters, and onset and progress of air pollution-borne chronic illnesses can be captured. Continuous real-time monitoring of pollution parameters in all areas where cohorts are established. The key question of the extent of pollution and its distribution across various parts of the city could not be inferred as the locations of data collection have varied widely. The limitations faced in the paper are that none of the studies or a combination of them could present a complete picture of the burden of diseases like COPD, bronchial asthma, and other allergic conditions attributable to pollution in Delhi.

In this [5] the authors Shweta Taneja1, Dr. Nidhi Sharma2, Kettun Oberoi3, and Yash Navoria4 have offered a brief study on the trends of various air pollutants like sulphur dioxide(SO2), nitrogen dioxide (NO2), particulate matter (PM), carbon monoxide (CO), ozone (O3) seen in Delhi. The data is collected from 2011 to 2015 from Central Pollution Control Board (CPCB) consists of six attributes that are time (in months), air pollutants like SO2, NO2, CO, PM10and Ozone (03). The methodology used is the data mining techniques : linear regression and multilayer perceptron to interpret data and predict the trends and the tool used is WEKA. The agenda of the presented work is to minimize pollution through proper measures and ensure that the emitting of pollutants is observed within the range of regular pollution checks.

Anchal Garg et al. [6] presented research on a comprehensive study on the impact assessment of lockdown on overall ambient air quality amid COVID-19 in Delhi and its NCR, India. The dataset included data from 5 cities and 36 monitoring stations, twenty from Delhi and four each from Gurugram, Faridabad, Ghaziabad, and Noida. The data of air pollutants (PM10, PM2.5, NOx, NO, NO2, SO2, NH3, SO2, CO, and C6H6) from 36 locations in the study area were analyzed from 1st March to 1st May 2020. There was little chance of influx and outflux of air pollutants since the meteorological condition was almost the same for a week before lockdown and during the lockdown period. Since, during the lockdown period, all industrial, vehicular, and commercial activities were restricted, it provided ideal conditions for investigating their direct effect on the extent of pollution reduction and their role in adding the air pollution loads. Since there was little chance of influx or outflux, the variation in meteorological conditions was not significant as confirmed by ANOVA hypothesis testing. It was one of the most important outcomes of this study to identify the source more convincingly. This ideal condition (complete lockdown) in the real world is rarely found to pinpoint the exact cause of air pollution.

The authors of [7] focus on forecasting air pollution identifying air quality levels and recognizing the associated health impacts. The data were obtained from the official website of the Indian government where this research analyzed time-series data from 2005-to 2015 consisting of PM10, SO2, and NO2 with time variables day, month, and year. The comparative analysis of the approaches used for prediction: the Naïve Bayesian, Auto Regressive Integrated Moving Average (ARIMA), Exponential Smoothing, and TBATS is done. The purpose of this study is to forecast ambient air pollution to anticipate unwanted events in near future. Also, it is observed that each dataset has some unique characteristic so the forecasting model cannot guarantee to perform potentially for all datasets.

In this [8] the authors Sunil Gulia, Abhishek Mittal, and Mukesh Khare gave offered a novel study- Quantitative evaluation of source interventions for urban air quality improvement -a case study of Delhi city. The study has attempted the quantitative evaluation of selected management strategies for the reduction of air pollution in Delhi city. Air quality monitoring data of five differently located monitoring stations has been considered to evaluate the air quality status of Delhi.The evaluation has been carried out using updated emission inventory and advanced air quality dispersion model, i.e., AERMOD. The AERMOD validation has been done by comparing predicted and monitored concentrations at five different locations in Delhi and found satisfactory performance. A total of eight scenarios consisting of various control measures have been evaluated for the reduction of air pollution in Delhi. The limitation of this study was that it only evaluated these strategies based on their effectiveness in the reduction of pollution levels, however social and economic impact analysis must also be carried out before their implementation.

The authors in [9] performed time series analysis and prediction on air quality. In the months July to September 2018, data was collected using an IoT system from three locations in Delhi and the National Capital Region (NCR). AQI components and ppm concentrations of various gases such as Carbon Monoxide(CO), Carbon Dioxide(CO2), Ammonia(NH3), and Acetone ((CH3)2CO) are all included in the dataset. Linear regression was used to predict air quality, and the model was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Percentage Error (MAPE) (MAPE). In a polluted environment, the mean of CO2, CO, and NH3 indicates a mild drop, whereas the mean of (CH3)2CO) falls to a good level. The proposed model calculates the mean absolute error (7.57), root means square error (10.29), and mean absolute percentage error for various locations (0.07). As a result, the analysis can assist in maintaining AQI control in a specific location. Many factors that the authors overlooked, such as weather fluctuations, traffic density, and so on, could impact the model's accuracy.

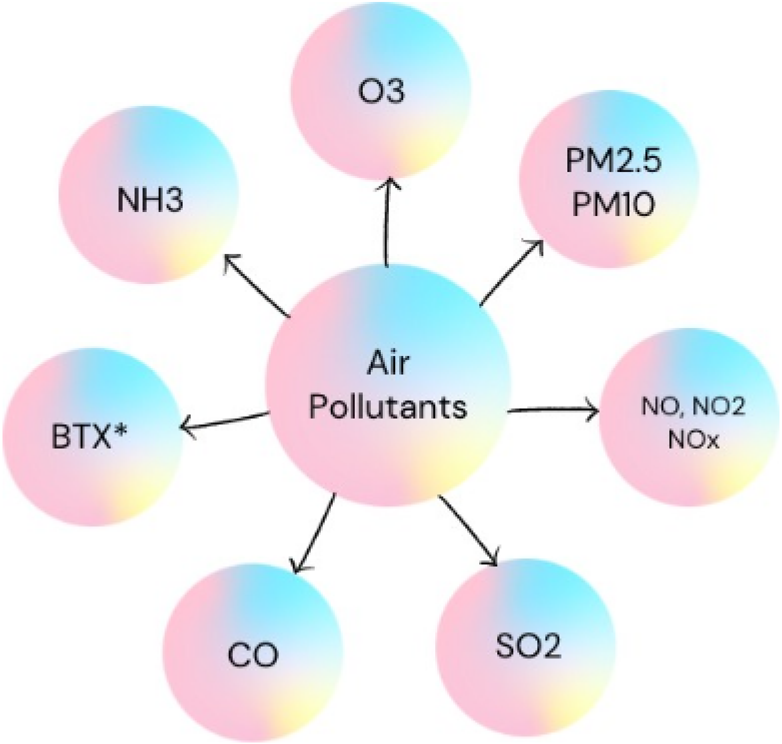
Miriam E. Marlier et al. [10] presented a study on Extreme Air Pollution in Global Megacities. In this study air pollution concentrations in the most populous global megacities were

examined and recent findings on the impacts of extreme air pollution, which were defined as concentrations exceeding
\' international guidelines were assessed. Station and satellite data of megacities were used. Important new methods for monitoring air pollution exposure, such as satellite-based estimates, including a more comprehensive understanding of the health and economic impact were also emphasized. There is a need to address the full cost of pollution from megacities, including techniques to value mortality and morbidity costs, integrating these costs with valuations from non-health impacts. The limitation of this study was that Monitoring data is not readily available for several megacities, making comparisons difficult.

- Dataset:

Data was collected from **CPCB: Central Pollution and Control Board.** CPCB collects air quality data which is updated every week. At present, the Real-Time Ambient Air Quality Monitoring Network consists of 261 Continuous Ambient Air Quality Monitoring Stations in 134 Cities covering 23 States & Union Territories which are connected to the web-based system.

The data keeps track of major air pollutants such as CO, NO2, NO, O3, SO2, PM2.5 etc**.** which are shown below:



❖ **Particulate matter (PM2.5 and PM10)**: Particulate matter is a mix of solids and liquids, including carbon, complex organic chemicals, sulphates, nitrates, mineral dust, and water suspended in the air. PM varies in size. Some particles, such as dust, soot, dirt or smoke are large or dark enough to be seen with the

naked eye. But the most damaging particles are the smaller particles, known as PM10 and PM2.5.

- **Nitrogen Oxides (NO, NO2, NOx):** Nitrogen oxides are a group of seven gases and compounds composed of nitrogen and oxygen, sometimes collectively known as NOx gases. The two most common and hazardous oxides of nitrogen are nitric oxide(NO) and nitrogen dioxide(NO2).
- **Sulphur Dioxide (SO2):** Sulfur dioxide or SO2 is a colourless gas with a strong odour, similar to a just-struck match. It is formed when fuel containing sulfur, such as coal and oil, is burned, creating air pollution.
- **Carbon Monoxide (CO):** Carbon monoxide is a colourless, highly poisonous gas. Under pressure, it becomes a liquid. It is produced by burning gasoline, natural gas, charcoal, wood, and other fuels.
- **Benzene, Toluene and Xylene (BTX):** Benzene, toluene, xylene, and formaldehyde are well-known indoor air pollutants, especially after house decoration. They are also common pollutants in the working places of the plastic industry, chemical industry, and leather industry.
- **Ammonia (NH3):** Ammonia pollution is pollution by the chemical ammonia (NH3) – a compound of nitrogen and hydrogen which is a byproduct of agriculture and industry.
- **Ozone(O3):** Ground-level ozone is a colourless and highly irritating gas that forms just above the earth's surface. It is called a "secondary" pollutant because it is produced when two primary pollutants react in sunlight and stagnant air. These two primary pollutants are nitrogen oxides (NOx) and volatile organic compounds (VOCs).

Along with these pollutants, the dataset also covers the AQI levels of each city.

Air Quality Index. The air quality index (AQI) is an index for reporting air quality daily. It is a measure of how air pollution affects one's health within a short period. A web-based system is designed to provide AQI on a 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 6 am on a day will incorporate data from 6 am on the 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.

Calculation of AQI. The AQI calculation uses 7 measures: PM2.5(Particulate Matter 2.5-micrometer), PM10, SO2, NOx, NH3, CO and O3 (ozone). For PM2.5, PM10, SO2, NOx and NH3 the average value in the last 24-hrs is used with the condition of having at least 16 values. For CO and O3 the maximum value in the last 8-hrs is used. Each measure is converted into a Sub-Index based on pre-defined groups. 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.

1. The Sub-indices for individual pollutants at a monitoring location is calculated using its 24-hourly average concentration value (8-hourly in case of CO and O3) and health

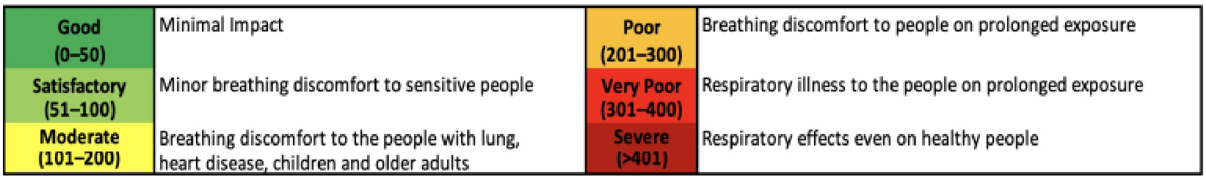
breakpoint concentration range. The worst sub-index is the AQI for that location.

2. All the eight pollutants may not be monitored at all the locations. Overall AQI is calculated only if data are available for a minimum of three pollutants out of which one should necessarily be either PM2.5 or PM10. Else, data are considered insufficient for calculating AQI. Similarly, a minimum of 16 hours' data is considered necessary for calculating the subindex.
3. The sub-indices for monitored pollutants are calculated and disseminated, even if data are inadequate for determining AQI. The Individual pollutant-wise sub-index will provide air quality status for that pollutant.

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:

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

There are 6 categories of the air 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

Daily and hourly city data, as well as daily and hourly Station data, is provided by CPCB. Station refers to the continuous pollution monitoring stations operated and maintained by the Central Pollution Control Board (CPCB) and the State Pollution Control Boards.

- Proposed Methodology:

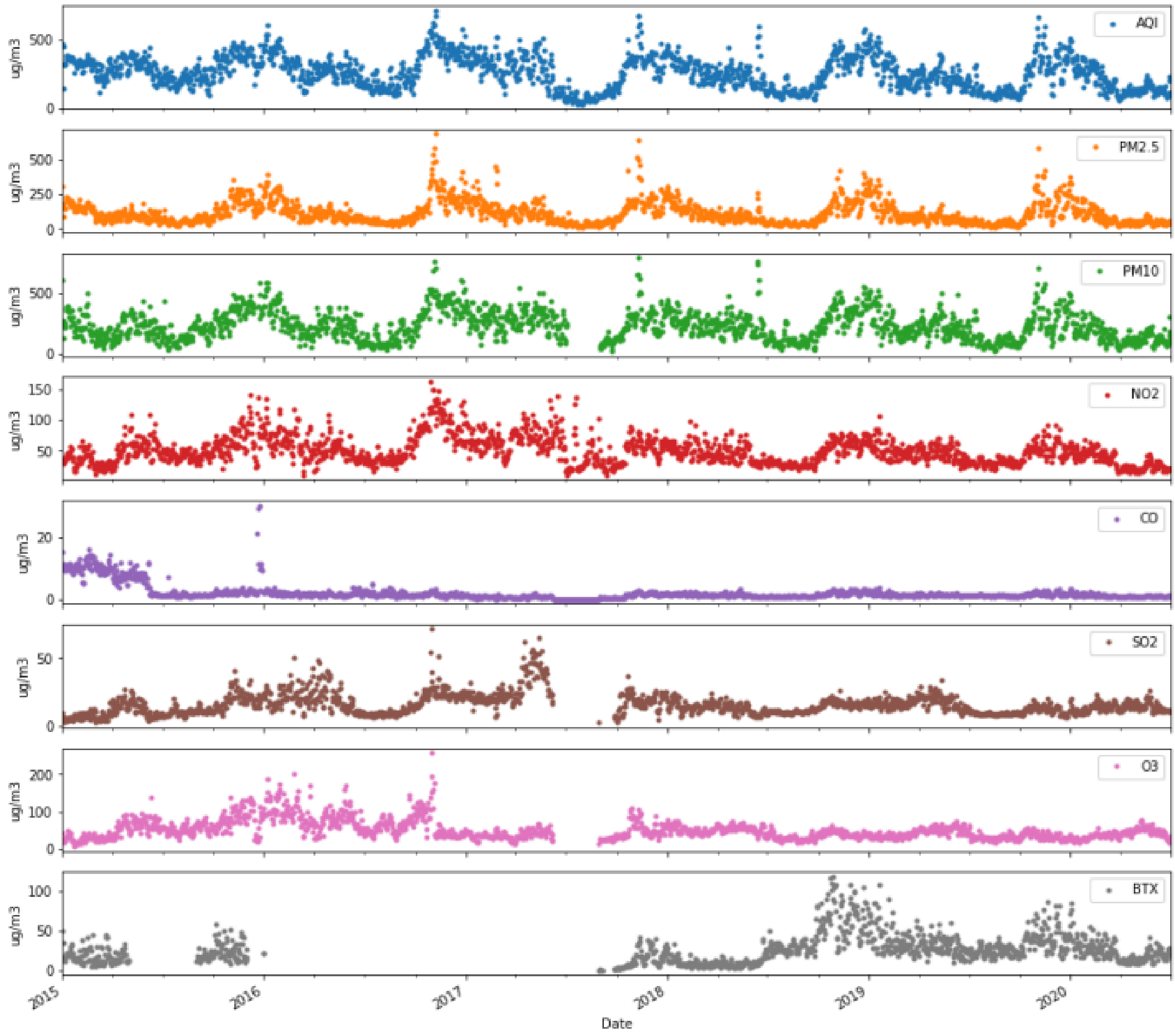
The initial step will be Data Preprocessing**.** The data is processed in several ways for many reasons which are shown and discussed below. The preprocessing is done in two stages:

Stage 1: To allow for better visualization for the analysis of the data and to focus on Delhi since the aim is to look at the AQI and different pollutant concentrations in Delhi.

Stage 2: To avoid encountering these NULL values while working with machine learning models.

Visualization of data after Stage 1 preprocessing. To understand the levels and concentrations of pollutants and the trends and seasonality they follow, we perform the following visualisations:

Visualising the concentration of pollutants in the air over the years. First, to get an overall bird's eye view of the entire Delhi dataset and pollutants we look at the concentration of pollutants in the air over the entire time period of the dataset i.e. 2015 - 2020. The concentrations are measured in ug/m3.

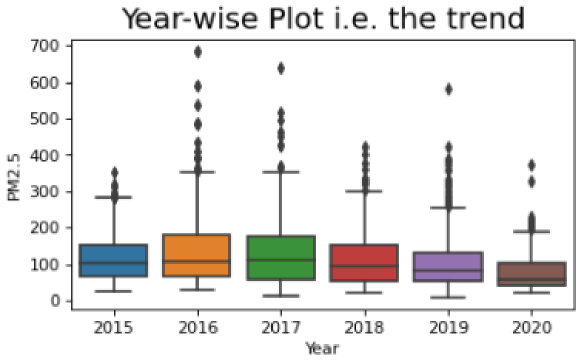


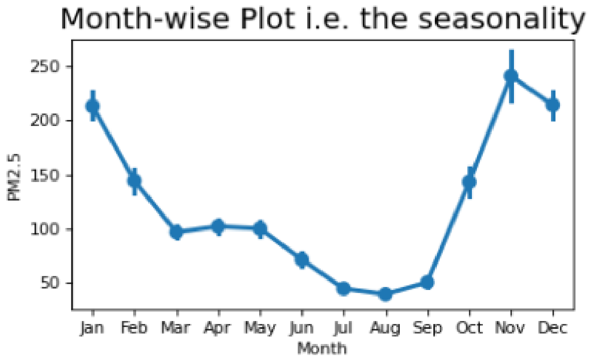
We notice a trend of increasing values over the time period in each pollutant from 2015 and the beginning of 2020, however, in the later part of 2020 i.e. after March 2020, we observe some decrease in the values of the pollutants while some like O_3 remain more or less the same. This observation confirms the effect of lockdown on the levels of these pollutants. We also observe the presence of missing plots implying missing values in the dataset.

Visualising the amount of pollutants in air over the years and months. Next, to get a more detailed view of the trend and seasonality seen in each pollutant we look at the concentration of pollutants in the air over the entire time period of the dataset i.e. 2015 - 2020 and well as the months. The concentrations are measured in ug/m3.

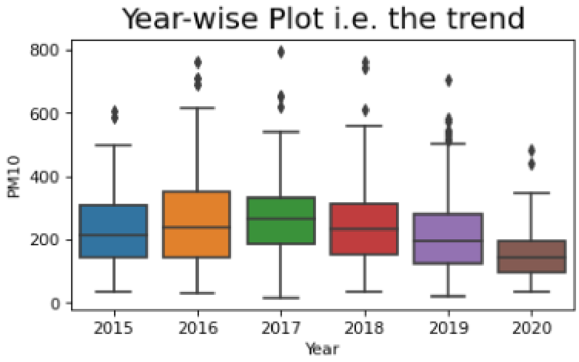
This process is carried out for all pollutants and AQI, we have mentioned the results below:

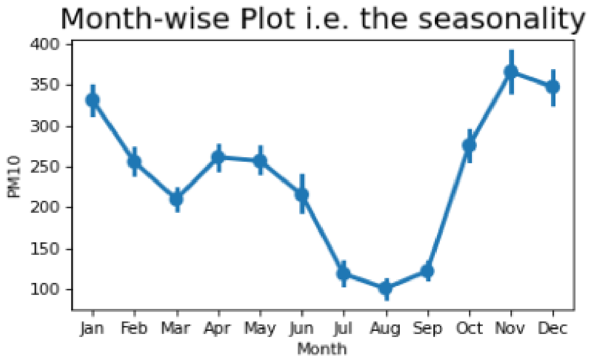
1. PM2.5


Year-wise Plot i.e. the trend

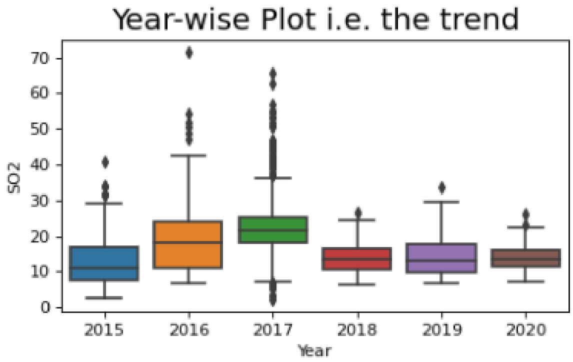

Month-wise Plot i.e. the seasonality

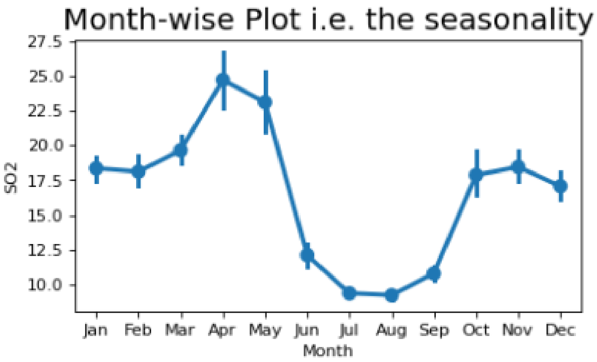
2. PM10


Year-wise Plot i.e. the trend

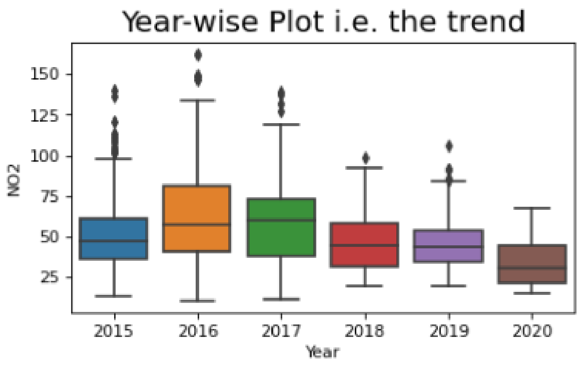

Month-wise Plot i.e. the seasonality

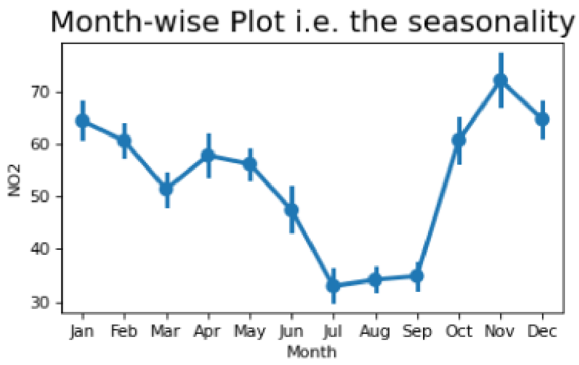
3. SO2


Year-wise Plot i.e. the trend

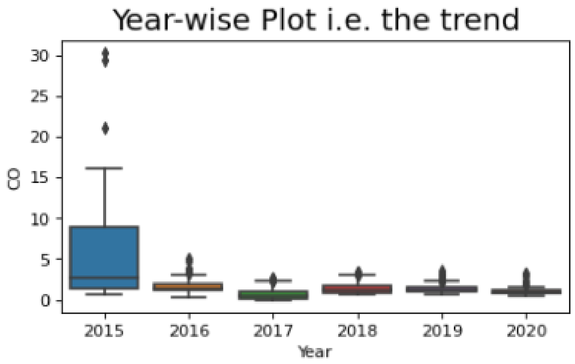

Month-wise Plot i.e. the seasonality

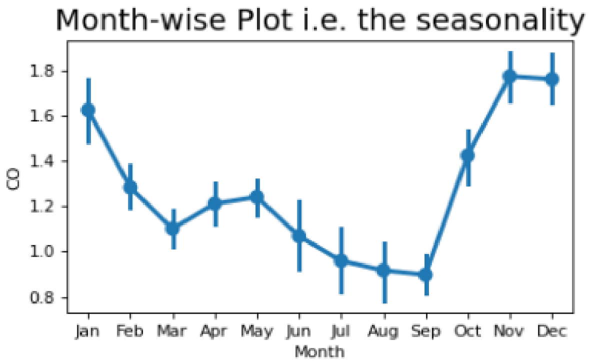
4. NO2


Year-wise Plot i.e. the trend
NO2
150
125
100
75
50
25
2015
2016
2017
2018
2019
2020
Year

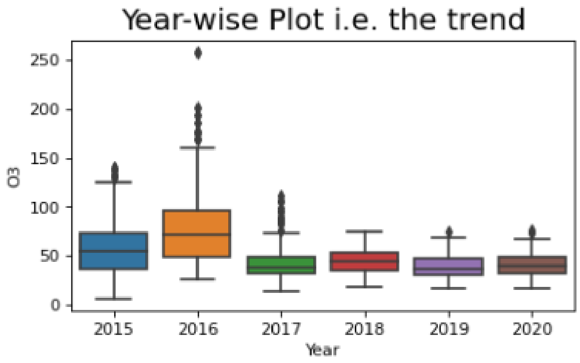

Month-wise Plot i.e. the seasonality
NO2
70
60
50
40
30
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

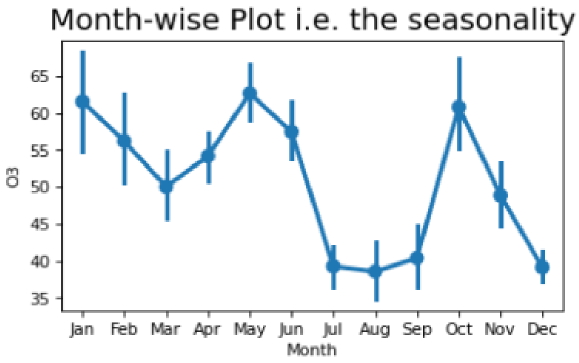
5. CO


Year-wise Plot i.e. the trend
CO
30
25
20
15
10
5
0
2015
2016
2017
2018
2019
2020
Year

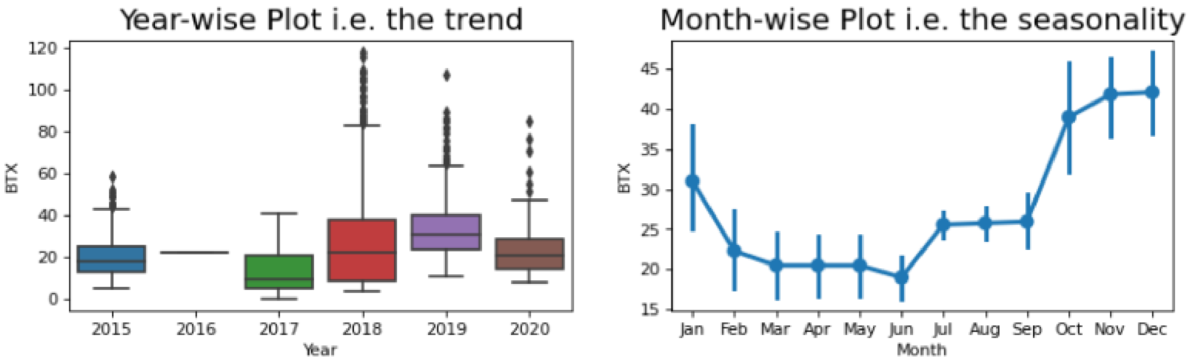

Month-wise Plot i.e. the seasonality
CO
1.8
1.6
1.4
1.2
1.0
0.8
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

6. O3

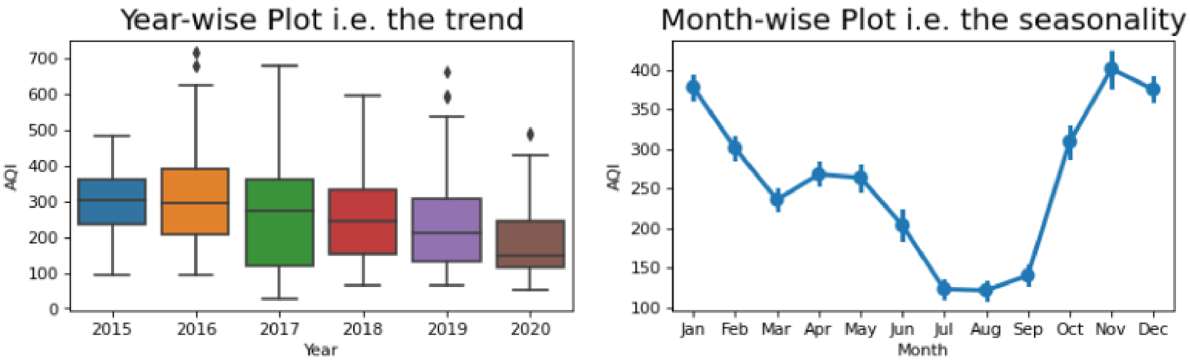

Year-wise Plot i.e. the trend
O3
250
200
150
100
50
0
2015
2016
2017
2018
2019
2020
Year


Month-wise Plot i.e. the seasonality
O3
65
60
55
50
45
40
35
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

7. BTX



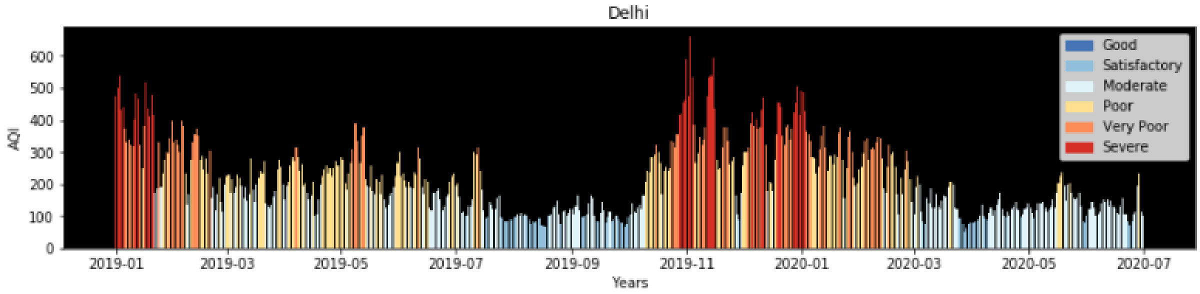
8. AQI



Delhi vs. pollutants. Next, for focused visualisation, we will look at Delhi’s AQI and pollutant concentrations over a fixed time period. This will also help us look at the changes during the lockdown.

Visualisation of concentrations over the last two years. To do so, we have focused only on 2019 and 2020. To look at the concentrations for the last two years, we have used Bar Graphs, where each bar represents the pollutant concentration or AQI value for a particular day and is shown in different colours depending on the category of the quality of air for that day.

AQI: We already know the grading of AQI levels from Good to Severe based upon the values of the air quality index. This grading or types of air was used to colour the bars to categorise them into Good, Satisfactory, Moderate, Poor, Very Poor and Severe for each day from January 2019 to July 2020. The missing values were taken care of. bfill() is used to backward fill the missing values in the dataset. It will backward fill the NaN values that are present in the data.



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 modelling of the seasonal component of the series is called SARIMA. SARIMAX is an extension of the SARIMA model that also includes the modelling 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).

Train the SARIMAX model on the Delhi_AQI data: First, the model is defined on Delhi City data and all Hyperparameters are initialised to zero with period interval(m) set as 12. auto_arima is used to find the Trend and Seasonal Element from which our model is prepared.

We will be training and looking at predictions in two phases to understand the effect of lockdown:

Phase 1. Without considering the effect of lockdown. For this phase, we trained the model on data from 2015-2018 and tested on 2019 data to test the accuracy and tune the parameters.

Next, after we have tuned our parameters depending on the predicted values, we go ahead to predict the future air quality i.e. the AQI for the year 2021 where we train the model on data from 2015-2020 and then predict 2020 July to 2021 July AQI values. The predictions of the model for pollutant concentrations and AQI of Delhi for the year 2019, its evaluation and predictions for 2020- 2021 will be discussed in the next section.

Phase 2. Considering the effect of lockdown. For this phase, we trained the model on data from 2015-2019 and predicted the unknown values for 2021. (Since the model parameters have already been tuned by testing the model on 2019 data in the previous phase, that part is not repeated.)

Model 2. LSTM. Long-short term memory is a common artificial recurrent neural network structure that is often used for time-series predictions. It stores past observations in memory, and while training it learns how to use this memory so as to not lose track of longer-term patterns. LSTMs are well regarded for a large variety of time-series tasks, although being vanilla neural network components some more configuration is required to set it up as compared to other models like ARIMA, SARIMAX, or Prophet.

Being as vanilla as they are, some measures need to be taken to account for inconsistencies in data and accounting for random variations. Unlike Prophet, this system cannot simply handle outliers like holidays and festivals. In this case, our primary concern is dealing with the missing data points - missing out on invalid Nans can cause our model to fail - and account for the inconsistencies brought in through lockdown.

Handling missing values as LSTM does not handle missing values :

A common issue encountered during training was when the LSTM kept giving Nan as the output for any input sequence where a single one of the entries was nan. So, in order to fix that we chose to simply eliminate the missing values altogether:

Training and testing: First, we trained and tested our models on 2015-18 and 2019 data respectively to get parameters right and see how the accuracy results are turning out. Due to long training times, not much tuning could be done beyond a couple of runs, but the best results at this phase were taken for the next bit.

Predictions for 2021. For this phase, we trained the model on data from 2015-2020 and predicted the unknown values for 2021. This requires a change in the split of the train and test, and the rest of the code remains unchanged.

Model 3. Prophet. Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well. It uses a simple, modular regression model that often works well with default parameters. That allows analysts to select the components relevant to their forecasting problem and easily make adjustments as needed.

Parameters in Prophet: There are many parameters that Prophet can consider while modelling a time series including national and regional holidays. The parameter we are most interested in is seasonality. Seasonalities are estimated in the Prophet using a partial Fourier sum.

Seasonal effects s(t) are approximated by the following function:

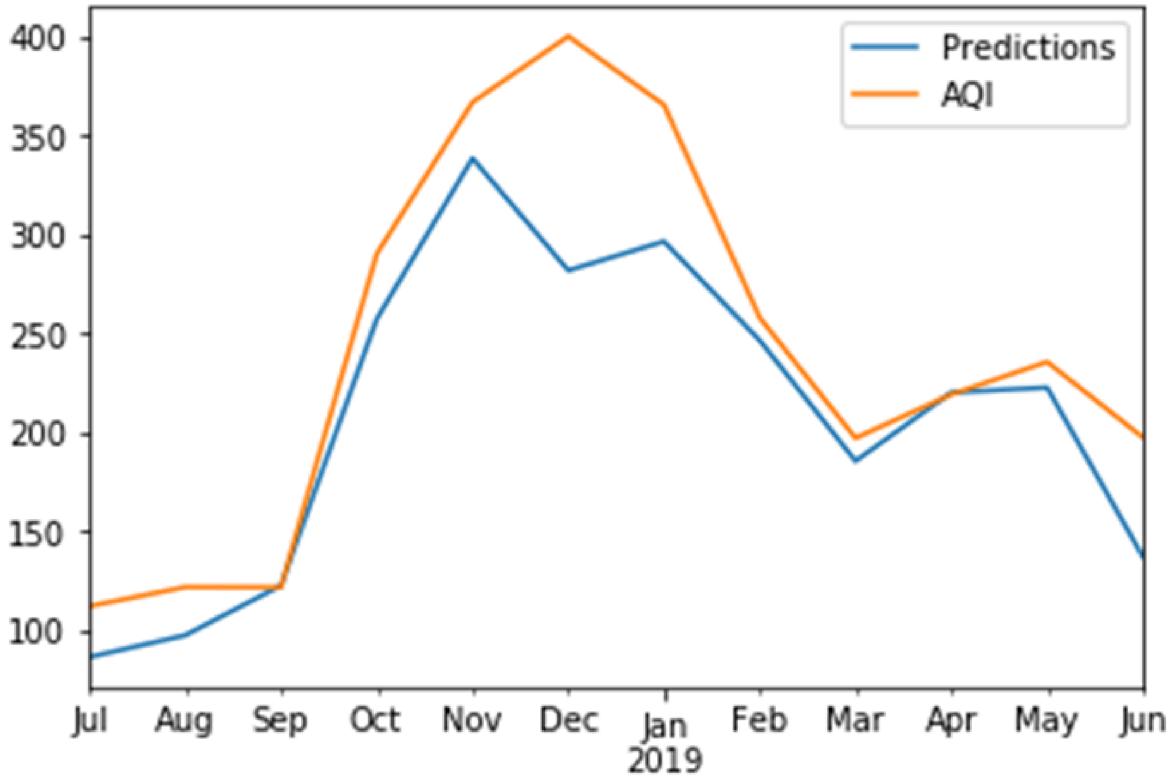
$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right)$$

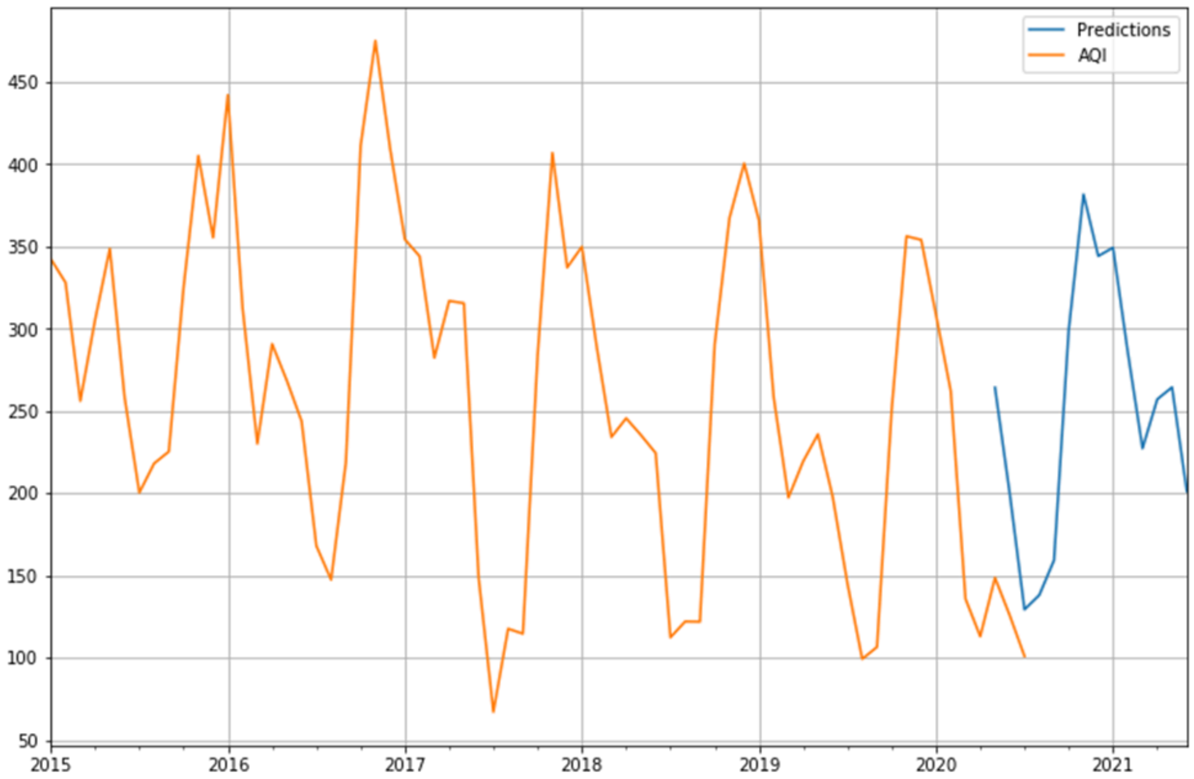
Results. Now we will look at the predictions made by the models for the two tasks - one of predicting the AQI values for a year already present in the dataset and another for 2020-2021.

Model 1. SARIMAX. Here are the results for both phases of this model:

Phase 1. Without considering the effect of lockdown. For this phase, we trained the model on data from 2015-2018 and tested on 2019 data to test the accuracy and tune the parameters. We haven't considered the 2019-2020 data because it is an outlier due to lockdown and may give incorrect results.

First, we look at the predictions for the 2019(year already available) AQI values.

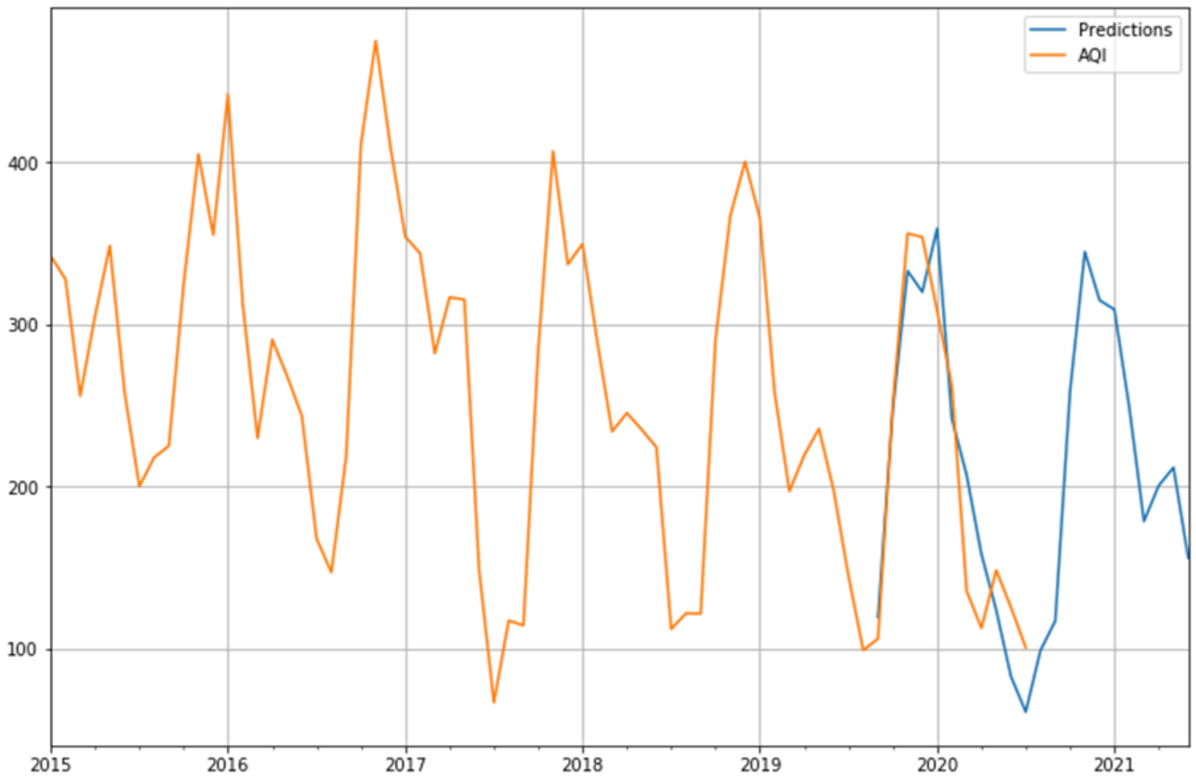




- ❖ We note that the 2021 prediction is that of very poor air quality i.e. approximately 380 AQI which is in the severe side of the Very Poor category.
- ❖ Even the lower bound on the AQI has increased from previous years.
- ❖ This plot represents the quality of air if there was no lockdown to begin with or if the lockdown conditions were to completely disappear.

Phase 2. Considering the effect of lockdown. For this phase, we trained the model on data from 2015-2020 and predicted the unknown values for 2021. (Since the model parameters have already been tuned by testing the model on 2019 data in the previous phase, that part is not repeated.)

Next, we look at the predictions for the unknown i.e. 2020-2021 AQI values considering the effect of lockdown.

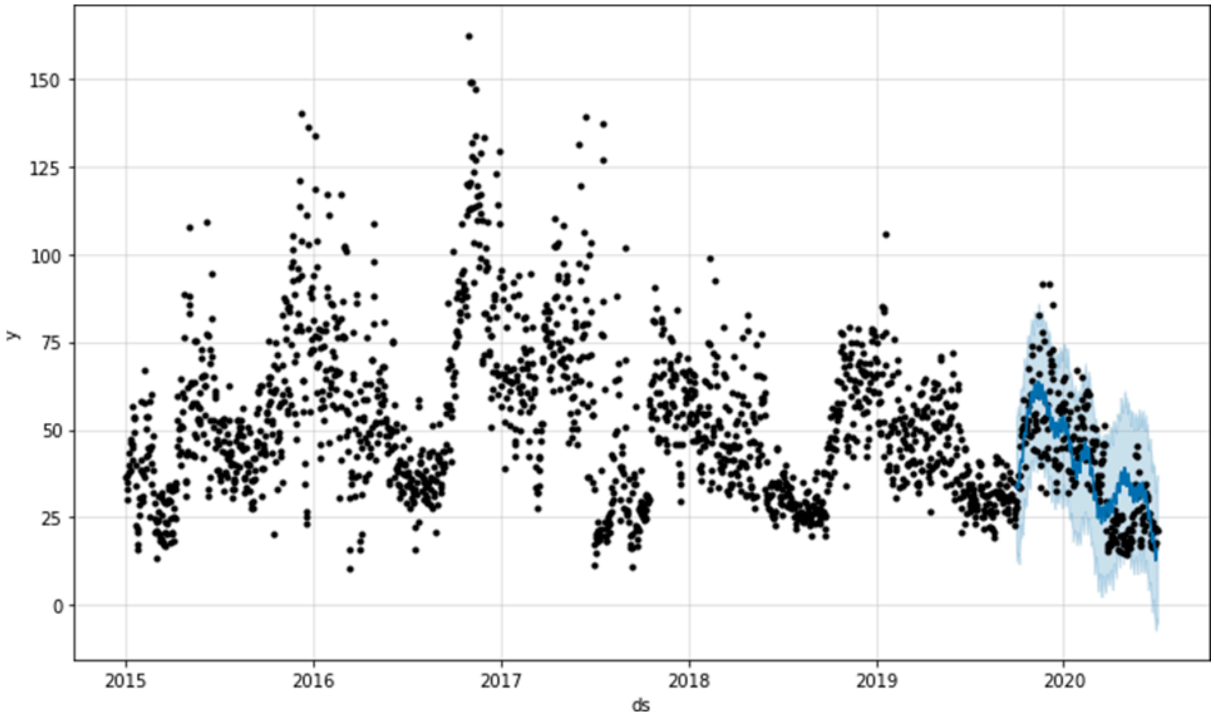


- ❖ We note that the 2021 prediction is not a highly optimistic one i.e. approximately 340 AQI which is in the Very Poor category. However, the peak has reduced a little.
- ❖ The decrease in the peak is purely due to the fact that 2020 is such an outlier. So, we can say that if lockdown were to continue, the air might have better quality.
- ❖ When the 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 and can be seen in phase 1.

Model 3. Prophet.

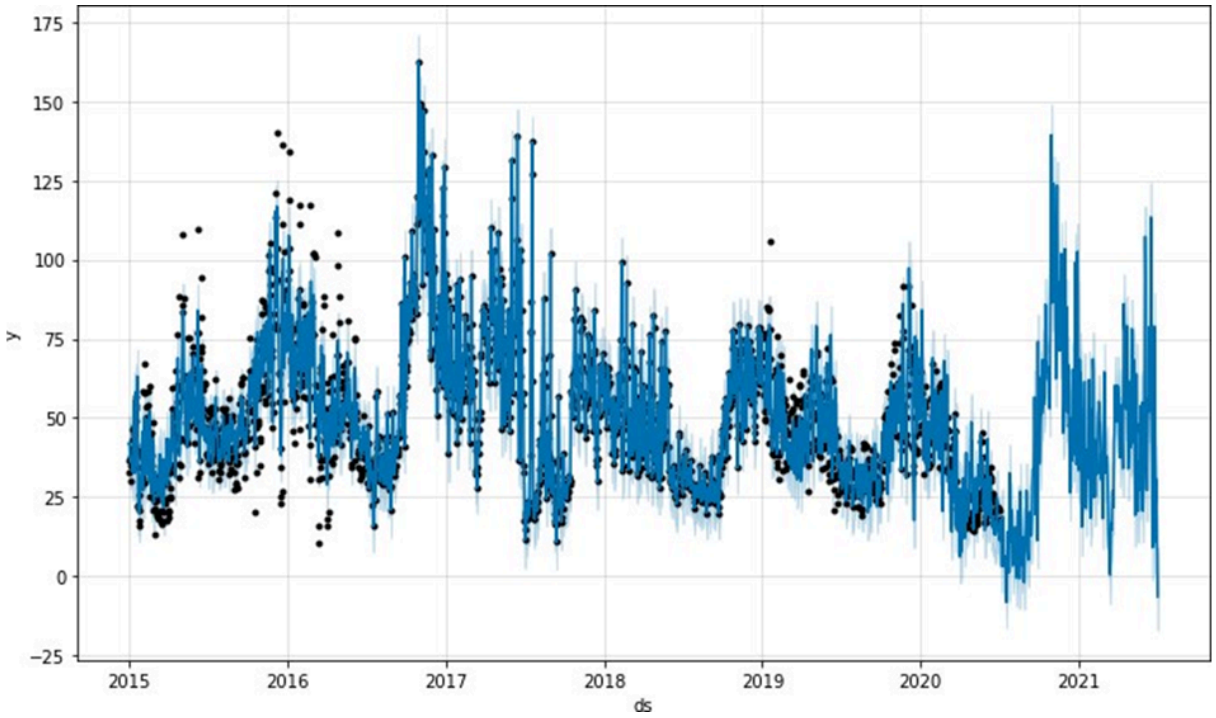
First, let’s see how well the model performed for the test data and then we will use it to predict future values.

The predict method will assign each row in future a predicted value which it names ‘yhat’. If you pass in historical dates, it will provide an in-sample fit. The forecast object here is a new dataframe that includes a column ‘yhat’ with the forecast, as well as columns for components and uncertainty intervals which are yhat_lower and yhat_upper which are for the lower and upper bounds respectively.



- The black points represent the actual values.
- The blue line represents the predicted values.
- The light blue shading represents the range of values.

We calculated the RSME for the final predictions which was **9.054781512949596** Next we make a dataset which is beyond the historical data available to us. This is known as an out-of-sample forecast. We can achieve this in the same way as an in-sample forecast and simply specify a different forecast period.



- References:

1. Nidhi Sharma, Shweta Taneja, Vaishali Sagar, Arshita Bhatt,Forecasting air pollution load in Delhi using data analysis tools,Procedia Computer Science,Volume 132,2018,Pages 1077-1085,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2018.05.023.

2. Ravindra, K., Singh, T., Biswal, A. *et al.* Impact of COVID-19 lockdown on ambient air quality in megacities of India and implication for air pollution control strategies. *Environ Sci Pollut Res* 28, 21621–21632 (2021). https://doi.org/10.1007/s11356-020-11808-7

3. C. Srivastava, S. Singh, and A. P. Singh, "Estimation of Air Pollution in Delhi Using Machine Learning Techniques," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 2018, pp. 304-309, doi: 10.1109/GUCON.2018.8675022.

4. Sharma AK, Baliyan P, Kumar P. “Air pollution and public health: the challenges for Delhi, India.” Reviews on environmental health. 2018 Mar 1;33(1):77-86.

5. Taneja S, Sharma N, Oberoi K, Navoria Y. Predicting trends in air pollution in Delhi using data mining. In 2016 1st India international conference on information processing (IICIP) 2016 Aug 12 (pp. 1-6). IEEE.

6. Garg A, Kumar A, Gupta NC. Comprehensive study on impact assessment of lockdown on overall ambient air quality amid COVID-19 in Delhi and its NCR, India. Journal of hazardous materials letters. 2021 Nov 1;2:100010.

7. Taufik MR, Rosanti E, Prasetya TA, Septiarini TW. Prediction algorithms to forecast air pollution in Delhi India on a decade. InJournal of Physics: Conference Series 2020 Mar 1 (Vol. 1511, No. 1, p. 012052). IOP Publishing.
8. Gulia S, Mittal A, Khare M. Quantitative evaluation of source interventions for urban air quality improvement-A case study of Delhi city. Atmospheric Pollution Research. 2018 May 1;9(3):577-83.

9. Raghavendra Kumar, Pardeep Kumar, Yugal Kumar,Time Series Data Prediction using IoT and Machine Learning Technique,Procedia Computer Science,Volume 167,2020,Pages 373-381,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2020.03.240.

10. Marlier ME, Jina AS, Kinney PL, DeFries RS. Extreme air pollution in global megacities. Current Climate Change Reports. 2016 Mar;2(1):15-27.