A PROJECT REPORT ON
PREDICTION OF AQI WITH TIME SERIES FORECASTING
BY

Vivek Bari **Payas Sonkusare**

Samruddhi Bhor **Chetan Palve**

Nikhil Shinde **Satish Wagh**

PREPARED IN PARTIAL FULLFILMENT
OF
CAPSTONE PROJECT UNDER THE SUPERVISION
OF
Mr. KONETI NAVEEN KUMAR YADAV

DSE Capstone Project

Project Summary

Batch details	PGPDSE-FT Pune June23
Team members	Vivek Bari Samruddhi Bhor Chetan Palve Payas Sonkusare Satish Wagh Nikhil Shinde
Domain of Project	Environment Analysis
Proposed Project title	Air Quality Index in India
Group Number	01
Team Leader	Vivek Bari
Mentor Name	Mr. Koneti Naveen Kumar Yadav

Date:

Signature of the Mentor

Signature of the Team Leader

Project Details

Introduction

- Air pollution is a complex mixture of different gases particles perceived as a modern-day curse, due to the increased amount of urbanization and industrialization across the world.
- Several countries have taken some serious measures to maintain the air quality. India, which holds largest amount of human population, also took various measures to improve the air quality. A report by WHO shows, about 43% of all lung disease and lung cancer are attributable to Air Pollution. As per World Bank study released in 2016 revealed that India lost more than 8.5% of its GDP in 2013 due to the cost of increased welfare and lost labor due to air pollution. Various studies performed previously on the Air quality shows the Particulate Matters (PM2.5 and PM10) as the most dangerous and life-threatening pollutant among the group of pollutants. Particulate matter contributes to approximately 800,000 premature deaths each year.
- This analysis explores the factors which are influencing the Air pollution, this will help us analyze the trend in air quality across the years which help us to derive some business solutions. Also, we believe that forecasting the air quality of cities helps us to prevent before it impacts our environment.

Business Problem

- Air pollution levels in most of the urban areas have been a matter of serious concern. It is the right of the people to know the quality of air they breathe. In view of this, we took initiative for developing a national Air Quality Index (AQI) for Indian cities. AQI is a tool to disseminate information on air quality in qualitative terms (e.g., good, satisfactory, and poor) as well as its associated likely health impacts. There are six AQI categories, namely Good, Satisfactory, Moderately polluted, Poor, Very Poor, and Severe. The AQI considers eight pollutants for which short-term (up to 24-hourly averaging period) standards are prescribed, however, AQI can be calculated if monitoring data are available for minimum three pollutants of which one should necessarily be PM2.5 or PM10. Based on the measured ambient concentrations, corresponding standards and likely health impact, a sub-index is calculated for each of these pollutants.
- WHO reports, particulate matter (PM) contributes to approximately 800,000 premature deaths each year. If the situation continues, this might have some serious not only on country's GDP but also on the health of our future generations. Several studies were previously done to study the Air quality of Delhi, which is the most polluted city India. But for a large metropolitan city, where the growth of corporate

companies is in exponential order, a proper study has to be done in order to increase the quality of lives. So, in the above context we planned to analyze the Air quality of India. This report explores the factors influencing the Air pollution; help us analyze the trend in air quality across the years which help us to derive some business solutions. Also, we believe that forecasting the air quality helps us to prevent before it impacts our environment, which in turn increase the production level of the industries and other companies, which in turn increase the country's GDP.

Problem Statement

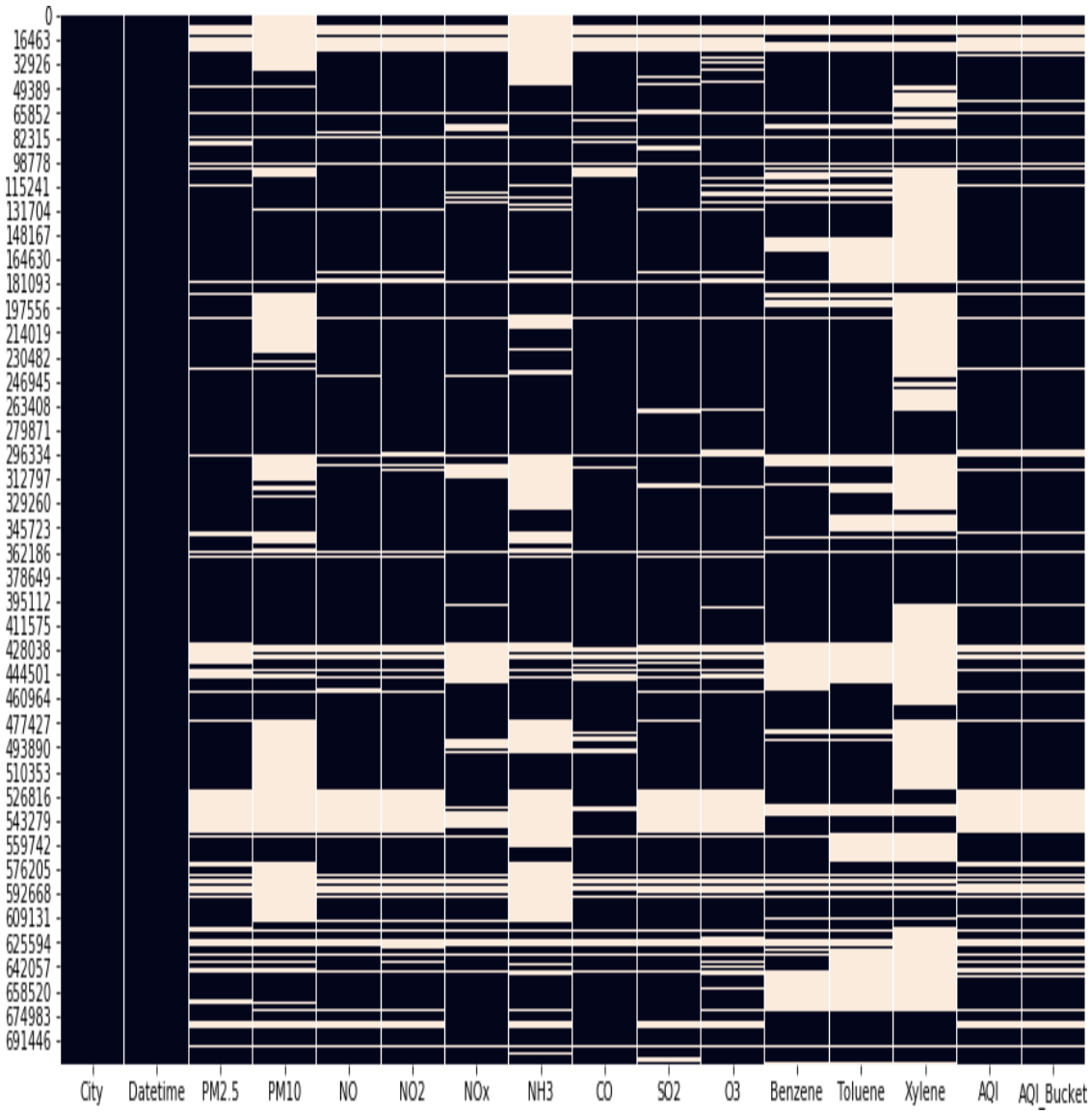
- As India is developing economy from Industrialization perspective. Development may be good for the nation in terms of GDP and economic gain, but it also has a flipside as more and more industries are being established, they are emitting more smoke, pollutants and Particulate Matter (PM) into the air which is increasing Air pollution. Air Pollution is causing diseases such as Asthma, Lung Cancer and other cardiovascular diseases. Air pollution is being measured by AQI (Air Quality Index) and higher the AQI higher the air pollution. In this project we will analyze the air quality data of major developed/ developing cities in India to find some underlying principles or patterns which will help us in predicting the AQI with the given data. Since this data has been captured at specific intervals over a period of time, we will be conducting a Time Series Analysis to predict the AQI of a city at a point in time.

Methodology

- **Data Collection**
  The data has been made publicly available by the Central Pollution Control Board: https://cpcb.nic.in/ which is the official portal of Government of India.

- **Data Understanding**
  The dataset contains air quality data and AQI (Air Quality Index) at hourly and daily level of various stations across multiple cities in India. We are focusing on monthly data of cities for forecasting the AQI on all over India. It contains different types of pollutants like PM2.5, PM10, NO, CO, Benzene, Toluene, and Xylene etc.

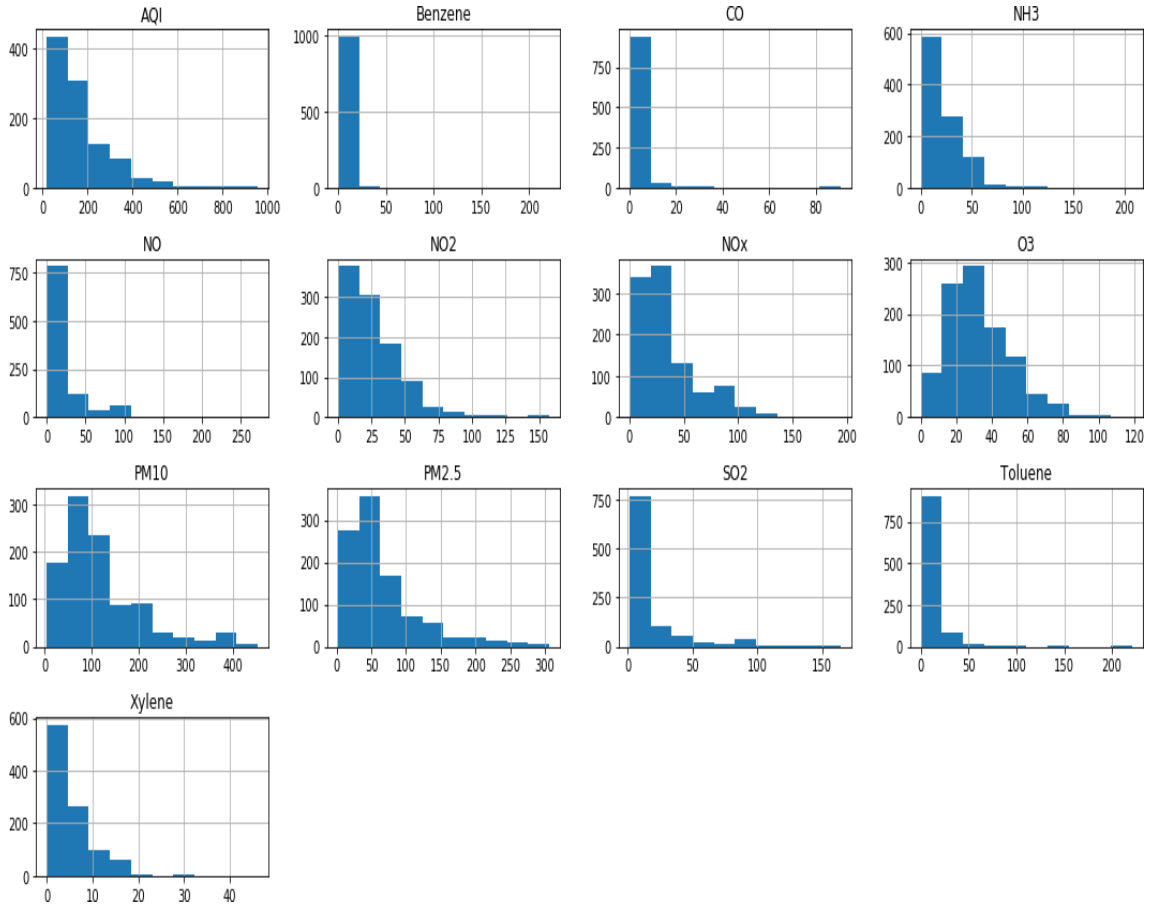
- **Data Cleaning**
  The dataset having missing values presents in all variable except City and Date time. Below heatmap shows missing values in variables.



We have to impute these missing values in data. As per the data description, missing values impute by city and date time. First it imputed by daily average, then monthly average and remaining by backward and forward fill. After this, check the distribution, it shows after imputing missing values, it's not much changed.

- **Univariate Analysis**

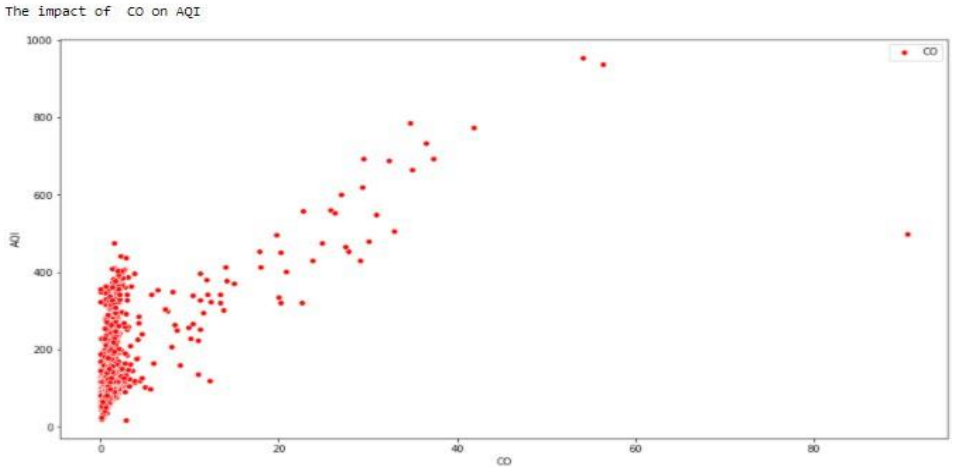
  All the variables in dataset are numerical variables. So, first check the distribution of all the variables and skewness of the data.

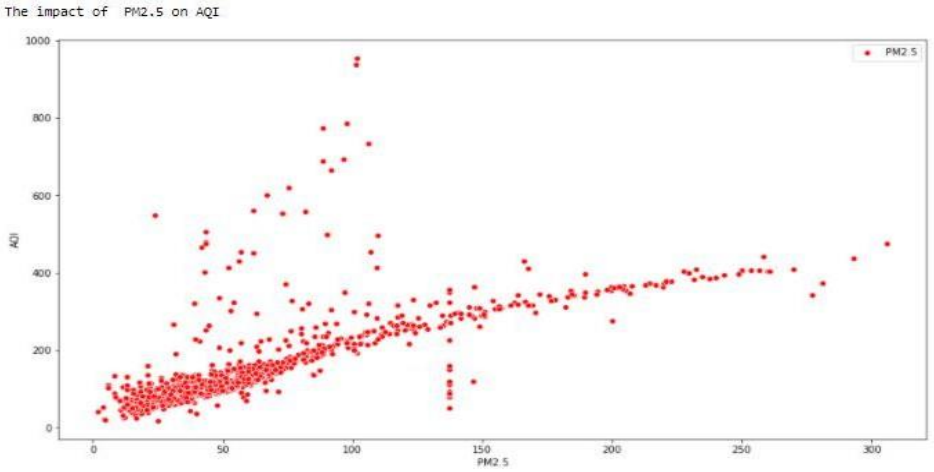


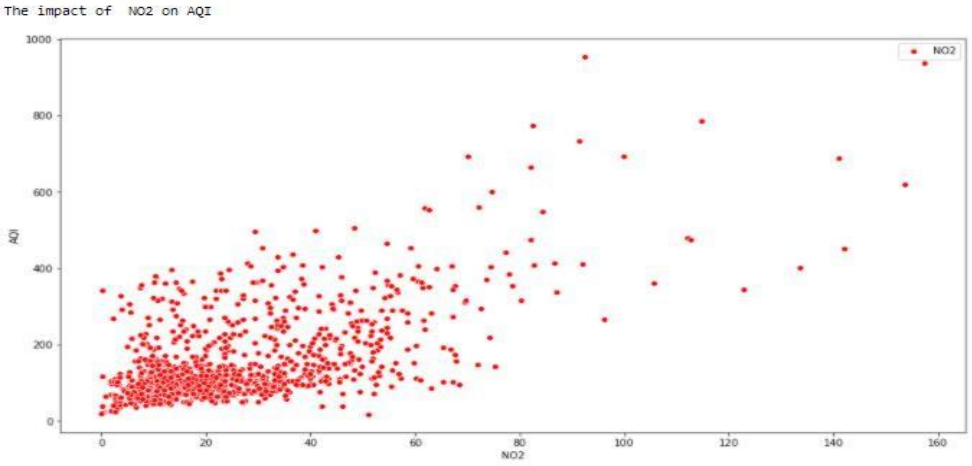
- From the plot, can say that most of the variables are skewed. All the variables are right skewed. Skewness of the Benzene, Toluene and CO is on higher side than other variables.

- **Multivariate Analysis**

  For multivariate analysis, plot graphs of all variables with AQI variable. It shows that the PM2.5, PM10, NO2 and CO are positive correlated with AQI, means these three are increases AQI also increase. It also visualizes by heatmap shown below.
  It shows that the PM2.5, PM10, NO2 and CO are having strong positive correlation, which means, higher these three AQI also higher. All other variables are also positive correlated with AQI but the correlation between them is weak, which means they are create less impact on AQI.



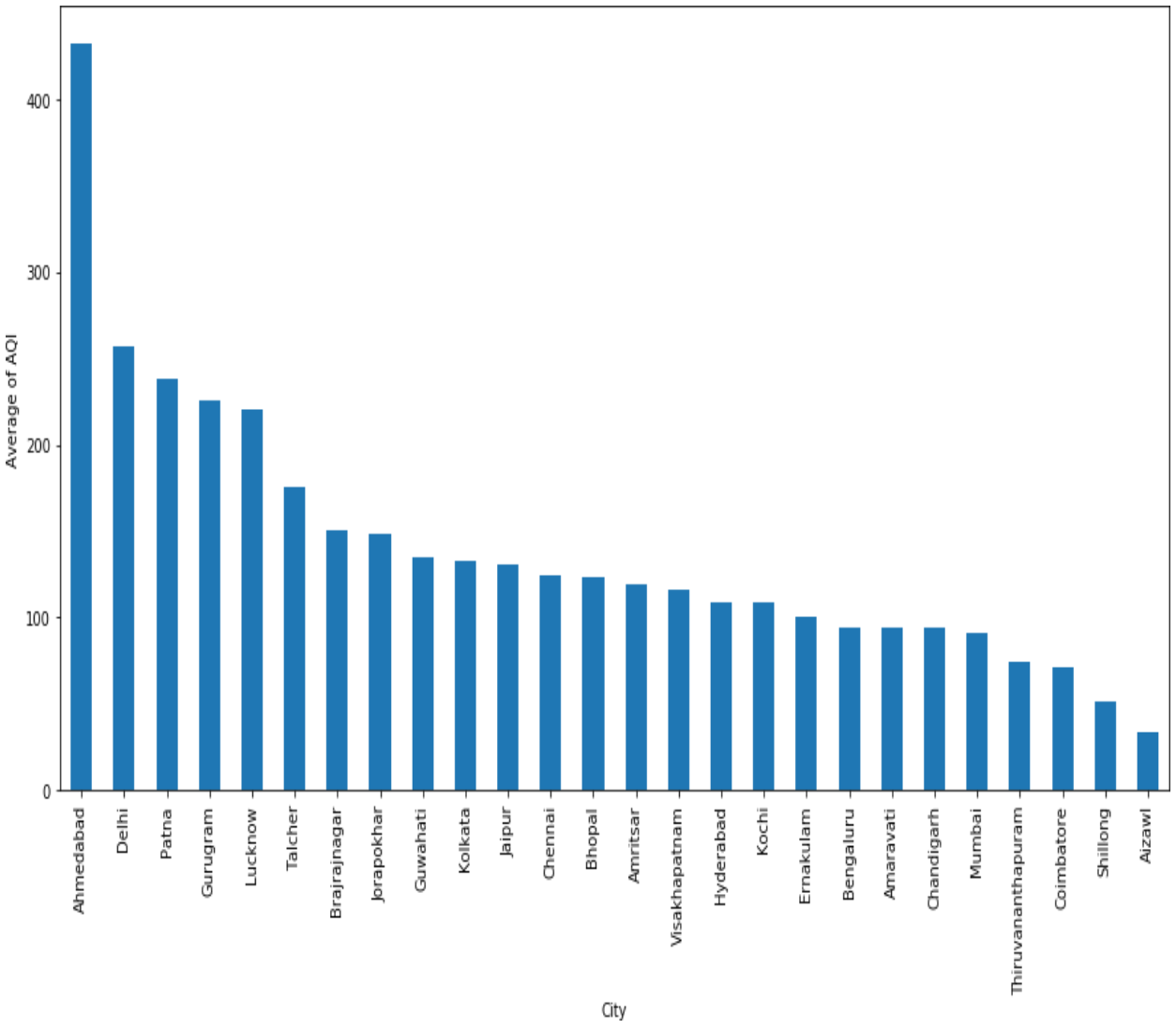




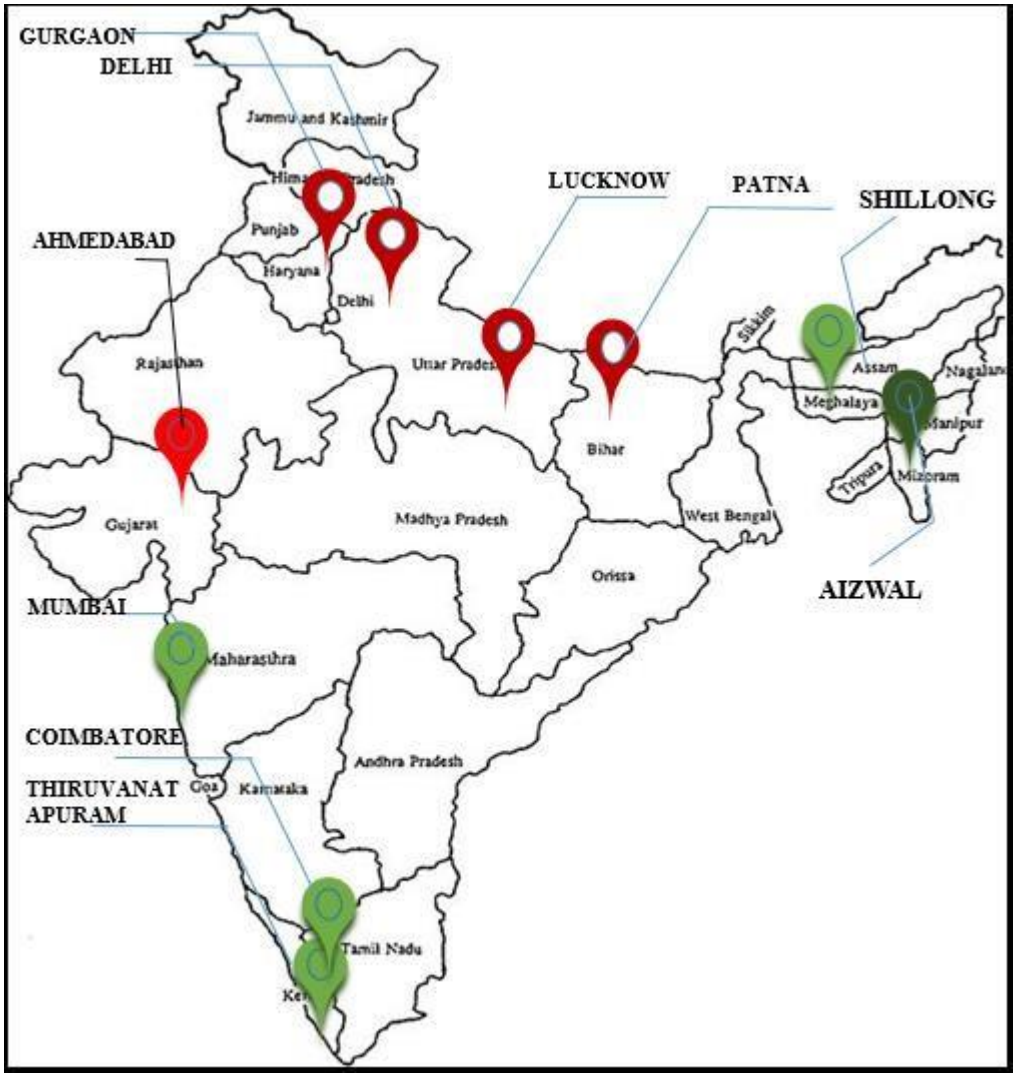
- We checked relationship of all the pollutants with AQI and plotted scattered plot. And we see that CO, PM2.5, NO2 has the strong visual relation with AQI.

Also we checked for the city wise average AQI, which plot is shown in below.

- It shows that, Ahmedabad city have a higher average of AQI in last five years. And Aizawl have very low average of AQI, may be the reason for these is mountainous area.



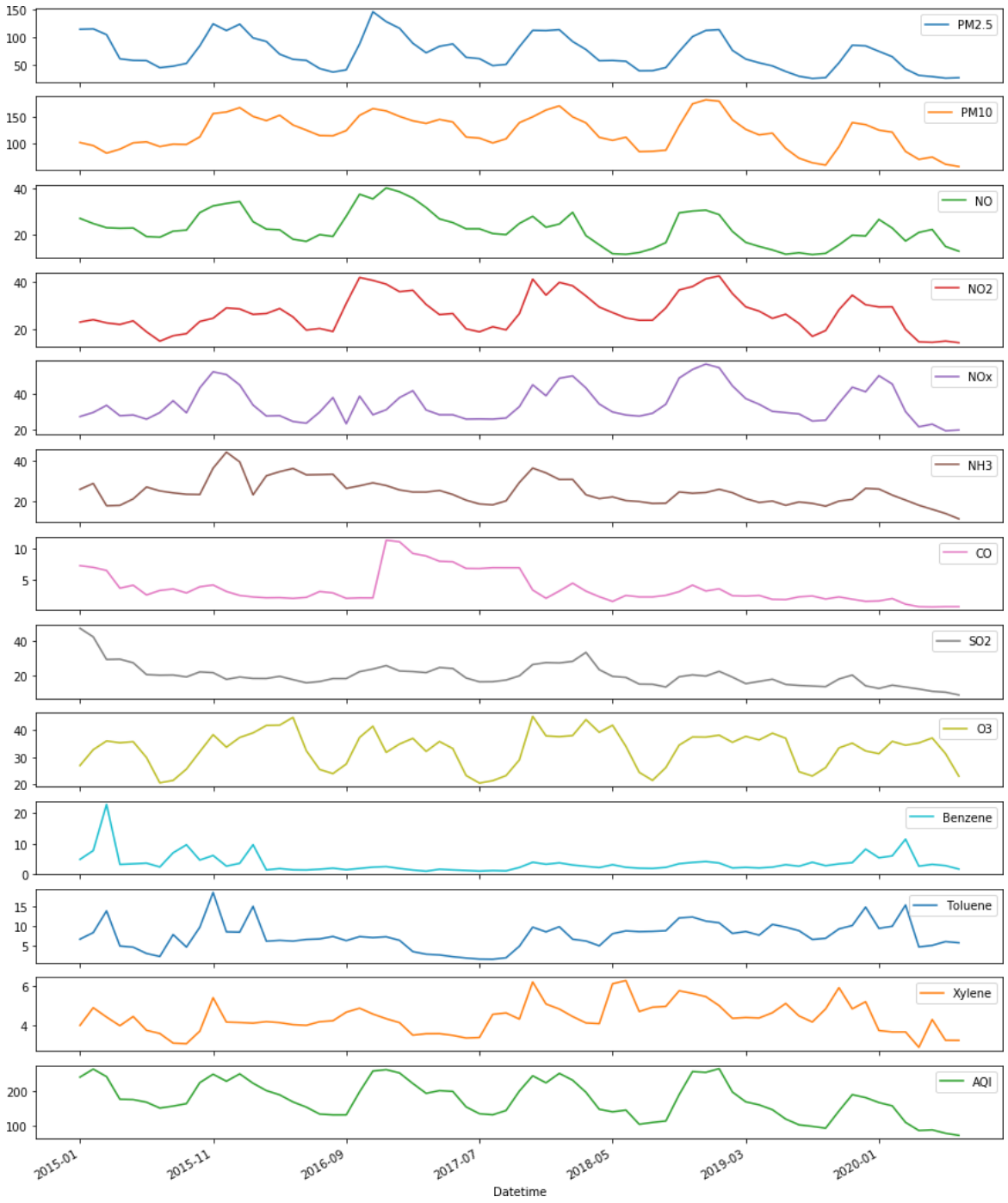
- **City wise representation of AQI on map of India**



- According to the map we can see the top 5 cities with the highest AQI and top 5 cities with the lowest AQI
- Cities with the highest AQI are represented by the red marker and the cities with the lowest AQI are represented by the green marker. We can infer that places towards the north region have high AQI whereas the AQI falls as we come down south.
- Ahmedabad is the city with the highest AQI and Aizwal has the lowest AQI

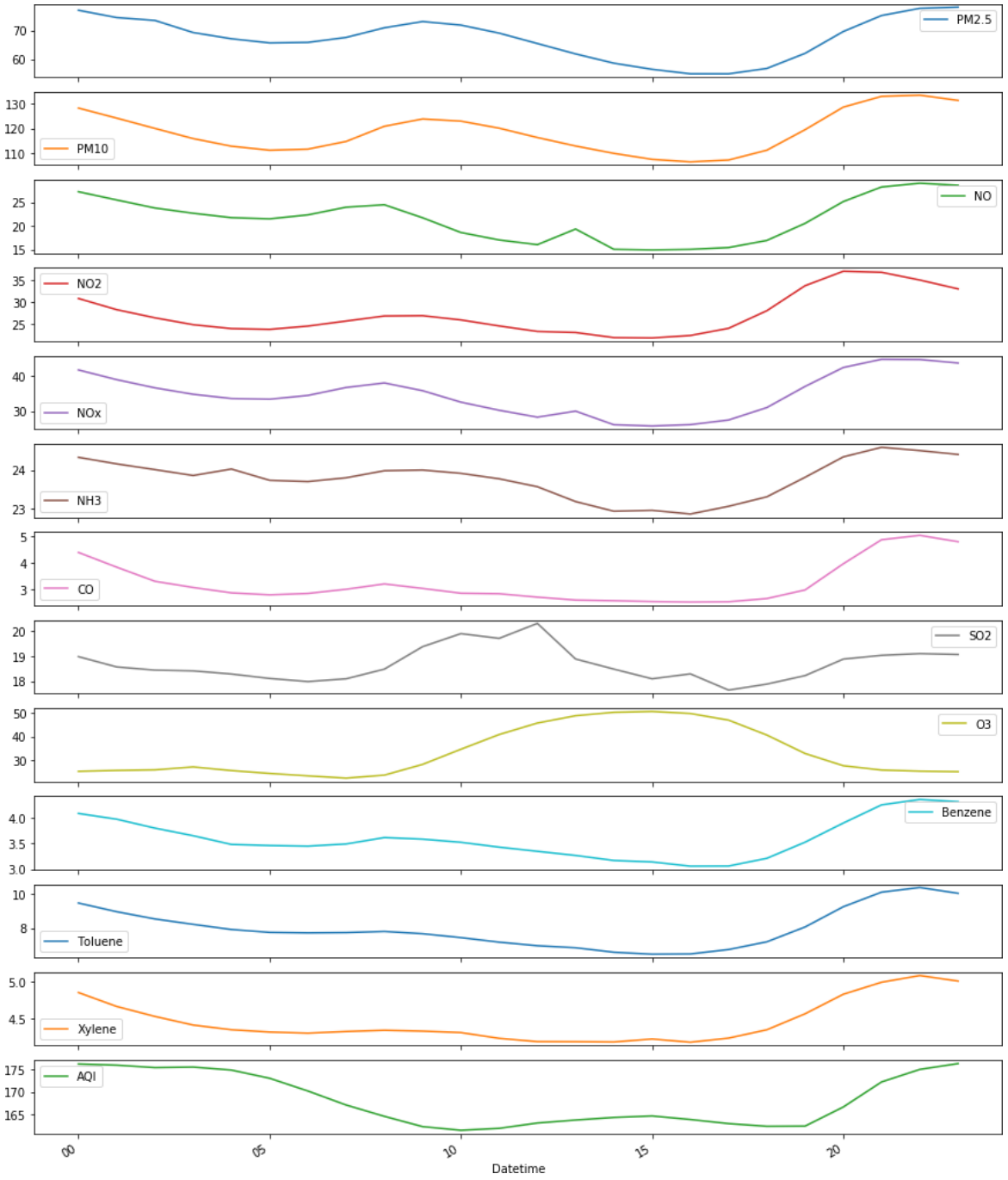
- **Time Series Visualization**

For visualize the time series data, first check the year and month wise visualization which shown in below.



From the above graph, can see that the AQI is decreases after January 2020, this is because of the COVID Pandemic. Industries were completed closed; Transport was halted for 5-6 months. So due to that we can see the Decline in AQI During 2020

Also check the visualization by hour considering all over India.



From the plot of pollutants hour wise, can say that the AQI and all other pollutants are having low average during night and higher at mid of the day.

As per the scientific analysis, higher the temperature, lower the AQI. During night temperature is low so AQI is high and during day it is vice versa.

- **Statistical Tests**

Here we perform the statistical test i.e. Anova test with all the columns and checked which pollutants are affecting more on AQI. From the statistical test we found that the CO, PM2.5 and NO2 are highly affected on AQI.

- **Stationarity**

Stationarity is an important concept in time series analysis. Stationarity means that the statistical properties of a time series do not change over time. Stationarity is important because many useful analytical tools and statistical tests and models rely on it.

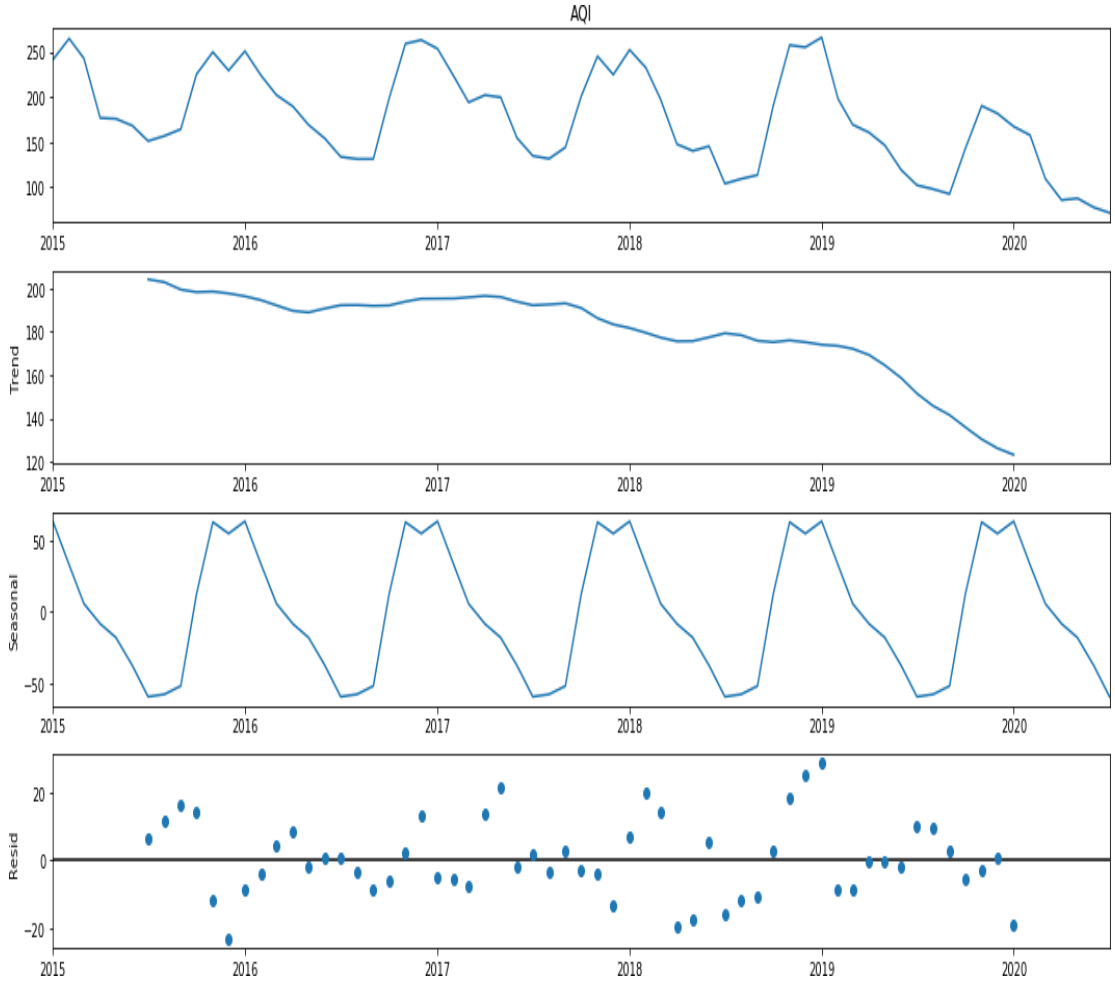
We checked for stationarity with adfuller (Dickey Fuller) test for AQI.

We found that from the test the pvalue is < 0.05, hence we reject the H0. This means that data is stationary.

- **Time Series Decomposition**

We can also visualize our data using a method called time-series decomposition that allows us to decompose our time series into three distinct components like trend, seasonality, and noise.

The below graph gives us Observed values in the data. Next three graphs are Trend, Seasonality and Residuals. By looking at the trend in the data, we can see that the trend is decreasing gradually year by year and there is sudden decrease after 2019. The seasonal graph shows the cyclic changes in the data. The data points which doesn't follow trend as well as seasonality, are plotted in the residual graph.

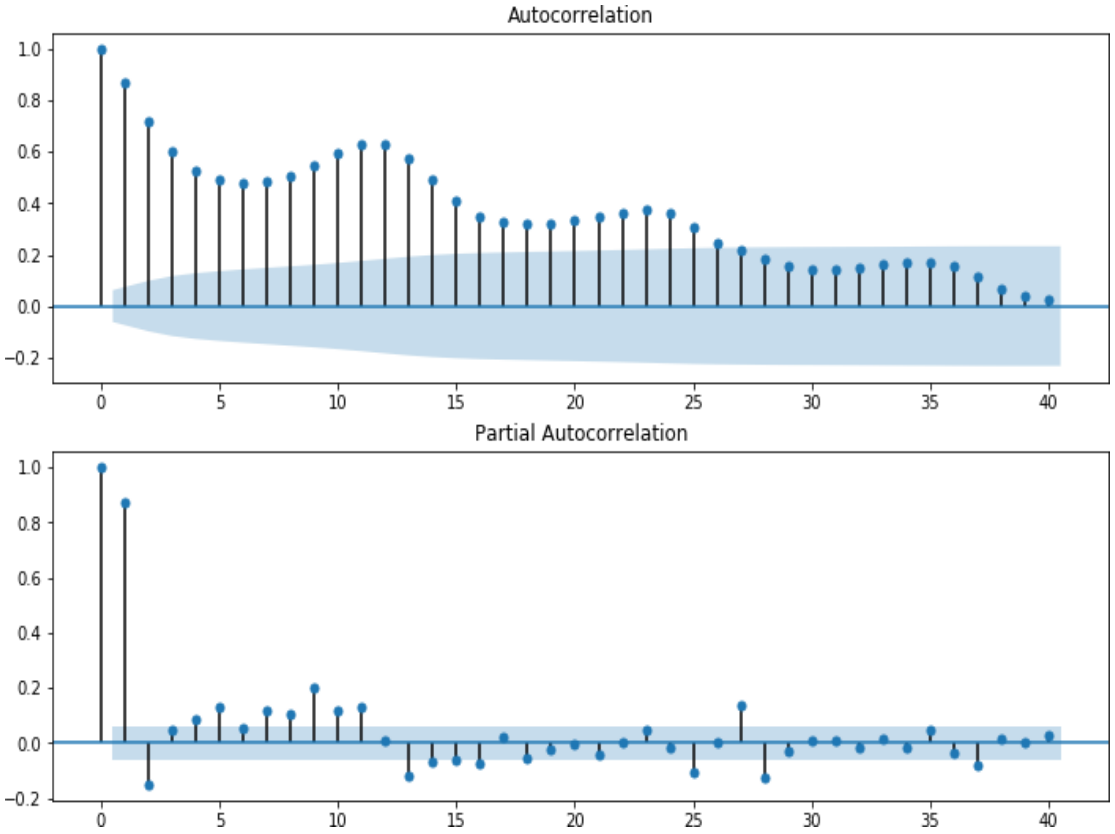


- **ACF and PACF plot**

After checking Stationarity, plot ACF and PACF which shown in below.
Here ACF is stands for Autocorrelation Function and PACF for Partial Autocorrelation Function. From the ACF and PACF plot we can determine the value for lag of the AR and MA model.

Partial autocorrelation computes the "pure" correlation between xt and xt−2 by removing the "transitive" correlation, that is, the amount of correlation explained by the first lag, and recomputing. For the partial autocorrelation between xt and xt−3, we will remove the correlation with both xt−1 and xt−2 and recomputed, and so on. A partial autocorrelation is the amount of correlation between a variable and a lag of itself that is not explained by correlations at all lower-order-lags

ACF is an (complete) auto-correlation function which gives us values of auto-correlation of any series with its lagged values.

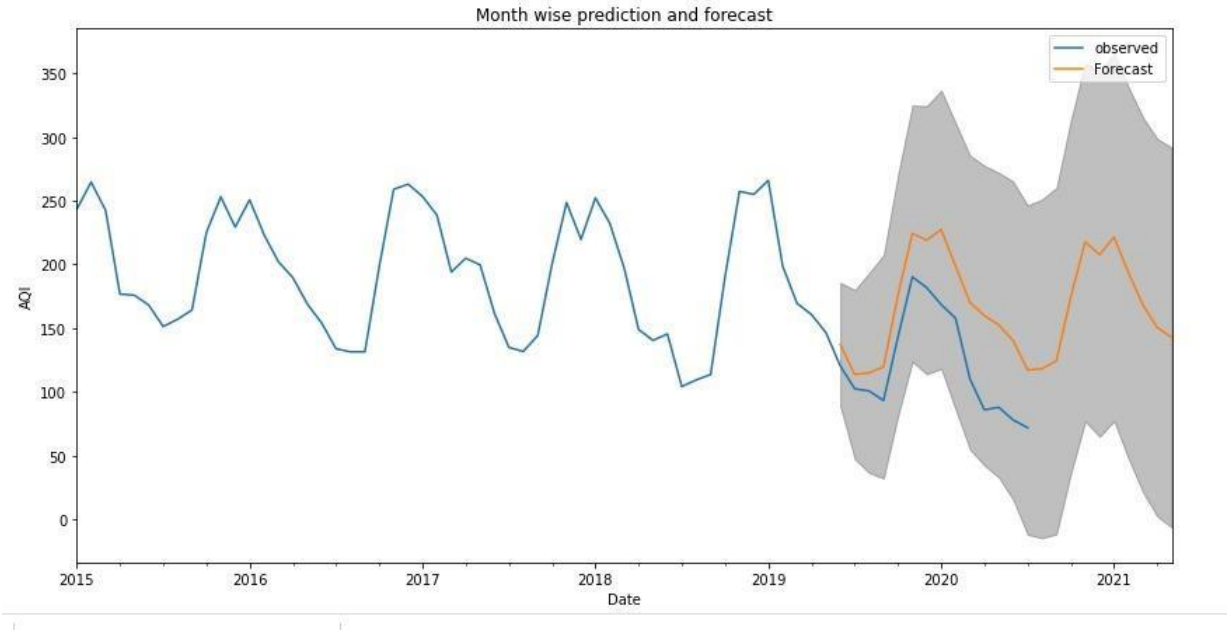


- **Base Model**

Here we choose ARIMA model for building a base model. We are taking order (p,d,q) from ACF and PACF plot and built ARIMA model which gives good result. Auto ARIMA model is also used for finding the best values for order (p,d,q). It's performs the step wise and search the best pair which gives low AIC value and built a model for that. Auto ARIMA gives the best order (3,0,2) and build model for that. Its also gives the Seasonality, but for ARIMA we didn't require Seasonality, so considering Seasonal = False. ARIMA (3,1,3) gives good result than ARIMA (1,1,1), Its AIC value is also low than the ARIMA (1,1,1) and log-likelihood is higher than the ARIMA (1,1,1).

- **Month wise Univariate analysis of AQI**

After the auto ARIMA we used the order (3,0,2) and we add the seasonality order to build the SARIMAX model. SARIMAX model is the same as an ARIMA model but it also takes into account the seasonality factor. Seasonality is the presence of variations that occur at specific regular intervals less than a year. Seasonality may be caused by weather and consists of periodic, repetitive, and generally regular and predictable patterns in the levels of a time series.



From the above graph we have observed and forecasted value. The predicted value of 2020 is very high on the basis of previous data. But the actual value of AQI seems less the reason for decrease might be pandemic.

- **Month Wise Multivariate analysis with VAR**

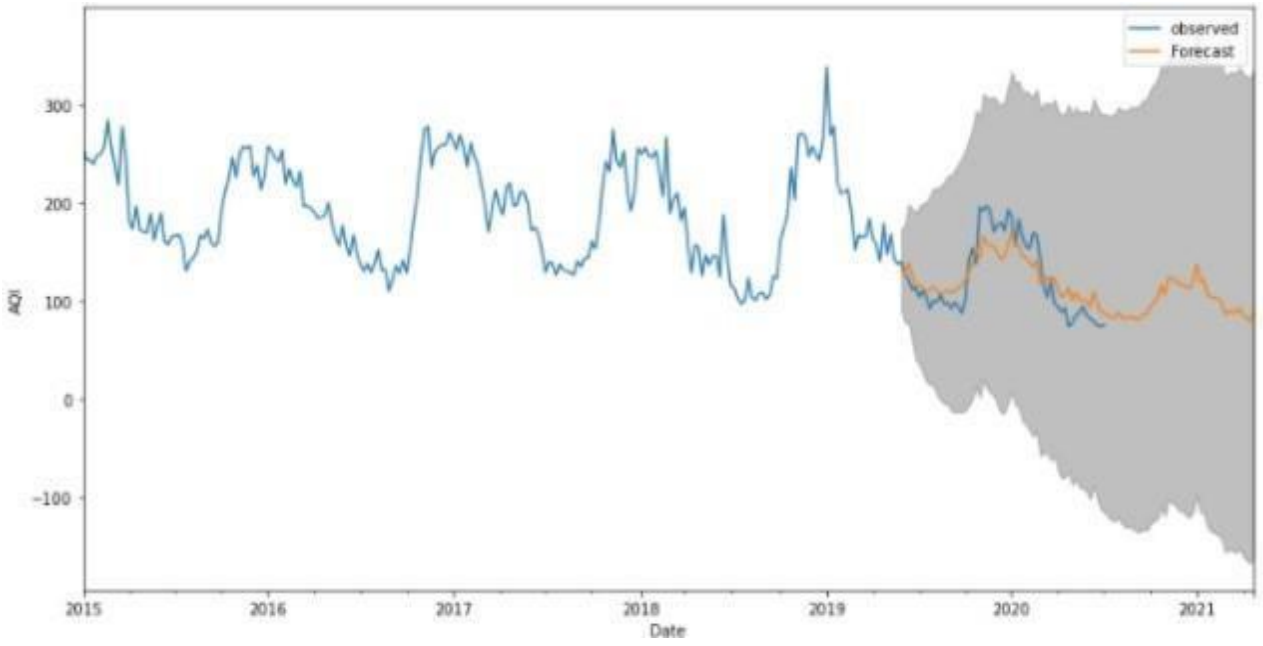
VAR is similar to the AR and MA models the only difference that AR and MA are used for univariate analysis whereas VAR is a multivariate analysis. The AR and MA models are unidirectional and VAR model is a bidirectional model. The algorithm calculates the vectors of the coefficients and multiplies it to the vectors of (t-1) value and adds the vectors of errors to it. It gives us the accuracy for all the variables.

- **Week wise Univariate Analysis with AQI**

In week wise analysis, first we create the data frame for week by aggregating method and split into train and test. After this we have built the auto ARIMA model and find the best order for this. By using this order, we add seasonality in this and built SARIMAX model.

SARIMAX model is the same as an ARIMA model but it also takes into account the seasonality factor. Seasonality is the presence of variations that occur at specific regular intervals less than a year. Seasonality may be caused by weather and consists of periodic, repetitive, and generally regular and predictable patterns in the levels of a time series.

With the output of SARIMAX and ARIMA model we forecast the AQI value for year 2020/21. From the observed value we can see that the forecasted value of 2020 was very high and actual value is pretty low. The reason behind this might be the covid pandemic.



From the above graph we have observed and forecasted value. And forecasted the value of AQI for 2021.

- **Week Wise Multivariate analysis with VAR**

We have performed the week wise analysis using VAR which is similar to AR and MA models. We created the data frame for week by aggregating method and split into train and test. After this we have built the auto ARIMA model and find the best order for this. We got rmse and mape values for each feature. But we here only considering AQI value. So the RMSE value of AQI is 62.86 and MAPE value is 55.32. Hence VAR is giving good result.

- **Comparative analysis**

We have compared Univariate SARIMAX and multivariate VAR for predicting AQI. We have built the auto ARIMA model and find the best order for this. By using this order, we add seasonality in this and built SARIMAX model.
Month wise MAPE value of SARIMAX is 39.10 and RMSE value is 45.74.
Month wise VAR, MAPE value is 55.32 and RMSE value is 20.91
So value of MAPE value of SARIMAX is less than VAR. So, we can conclude that SARIMAX is better than VAR.
Week wise MAPE value of SARIMAX is 15.32 and RMSE value is 20.91.
Week wise VAR, MAPE value is 52.76 and RMSE value is 61.93
So, we can say that the SARIMAX is better than VAR.

- **Conclusion**

- We have done a Univariate and Multivariate analysis on AQI.
- Univariate Analysis is done by using ARIMA, Auto-ARIMA and SARIMAX models.
- Multivariate Analysis is done by using VAR model.
- We have received the best results for predicting AQI through SARIMAX, therefore we chose SARIMAX as our final model and the forecasted values are visualized through graphical representation.
- From the predicted values we could observe that the AQI for 2020 is predicted to be very high on the basis of the previous data, but the actual values of AQI are very less in 2020. The reason for the decrease might be the pandemic which lead to lockdown and hence decrease in pollutants.
- From the forecasted graph we can conclude that AQI will be high for coming year 2021. The focus should be on how the pollutants affecting AQI can be reduced.
- According to our analysis the pollutants affecting the most on AQI are PM2.5, NO2 and CO on the basis of Correlation and Statistical Analysis.

Business problem Solution

From the above analysis we found that the AQI will increase in upcoming year. And we know that the higher AQI is harmful for human as well as environment. We also know that PM 2.5, PM10, CO, NO2 are highly affected on AQI so we need to control it. We should take appropriate measures to reduce the AQI by more tree plantation and taking precautions pollutants produced by industries.

References

https://paginas.fe.up.pt/~ec/files_0405/slides/02%20CRISP.pdf
https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Other References:

1. Anand Kumar, Dr. Ashish Grag and Prof. Upender Pandel, 2011, A Study of Ambient Air Quality Status in Jaipur City (Rajasthan, India), Using Air Quality Index, Nature and Science, 9(6), 38 – 43.

2. Sarella G and Khambete A K 2015. Ambient Air Quality Analysis using Air Quality Index – A. Case Study of Vapi. International Journal for Innovation Research in Science & Technology. 1(10)