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 analyse the Air quality of India. This report explores the factors influencing the Air pollution; help us analyse 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

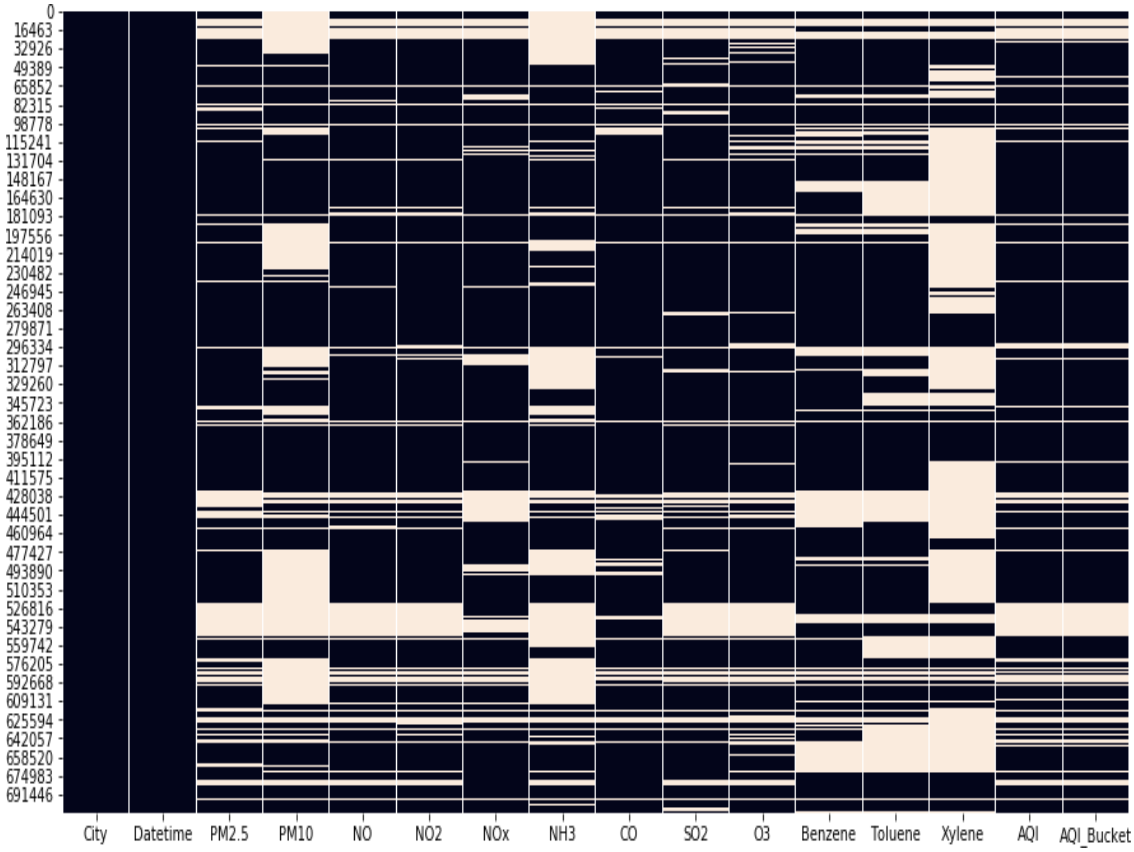
- As India is developing economy from Industrialization perspective. Development may be good for the nation in terms of GDP and economic gain, but it also has a flipside as more and more industries are being established, they are emitting more smoke, pollutants and Particulate Matter (PM) into the air 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, Xylene etc.

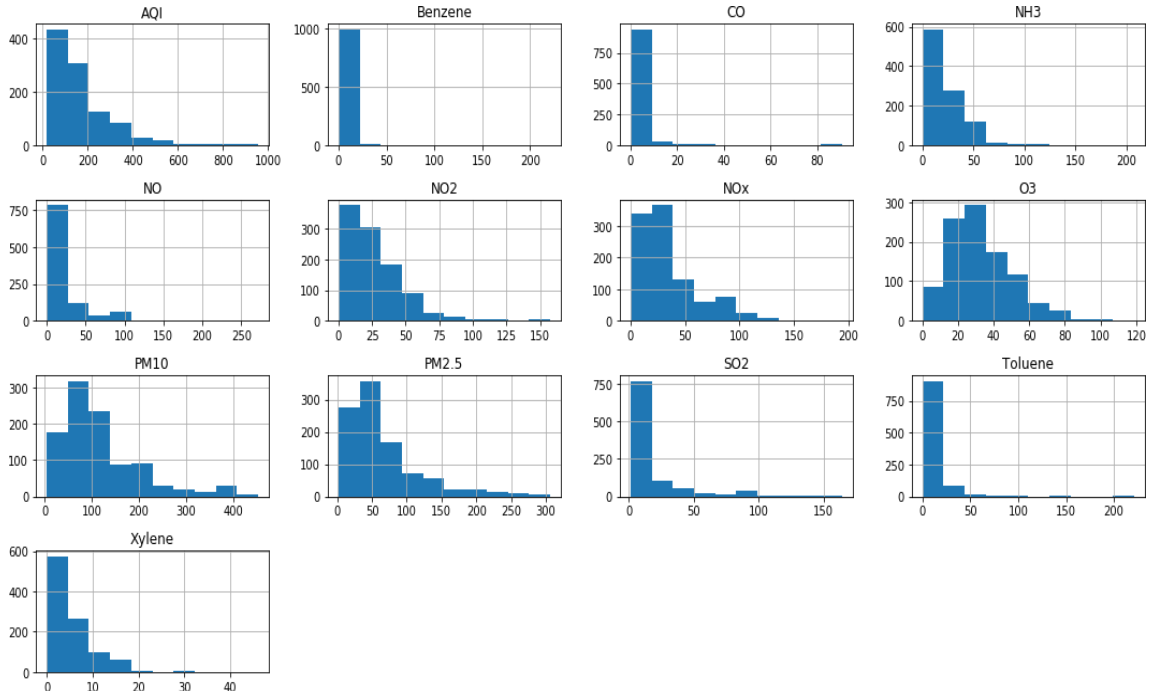
- Data Cleaning
  The dataset having missing values presents in all variable except City and Datetime. Below heatmap shows missing values in variables.



We have to impute these missing values in data. As per the data description, missing values impute by city and datetime. 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

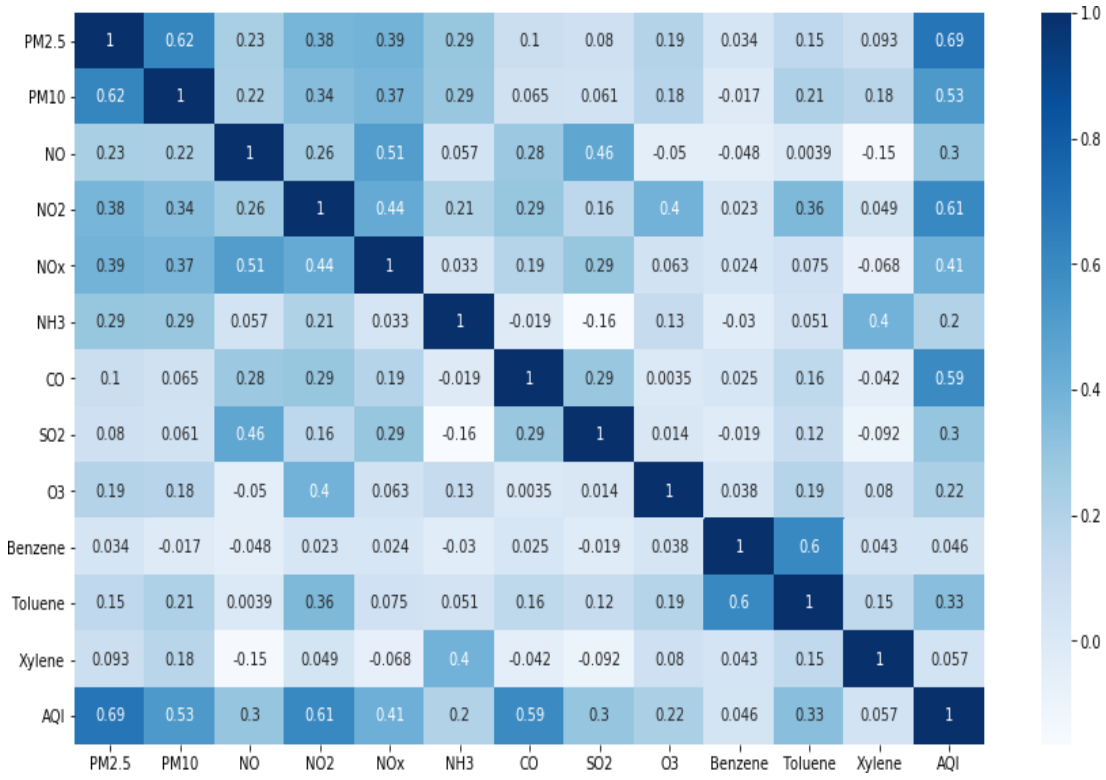
  All the variables in dataset are numerical variables. So, first check the distribution of all the variables and skewness of the data.



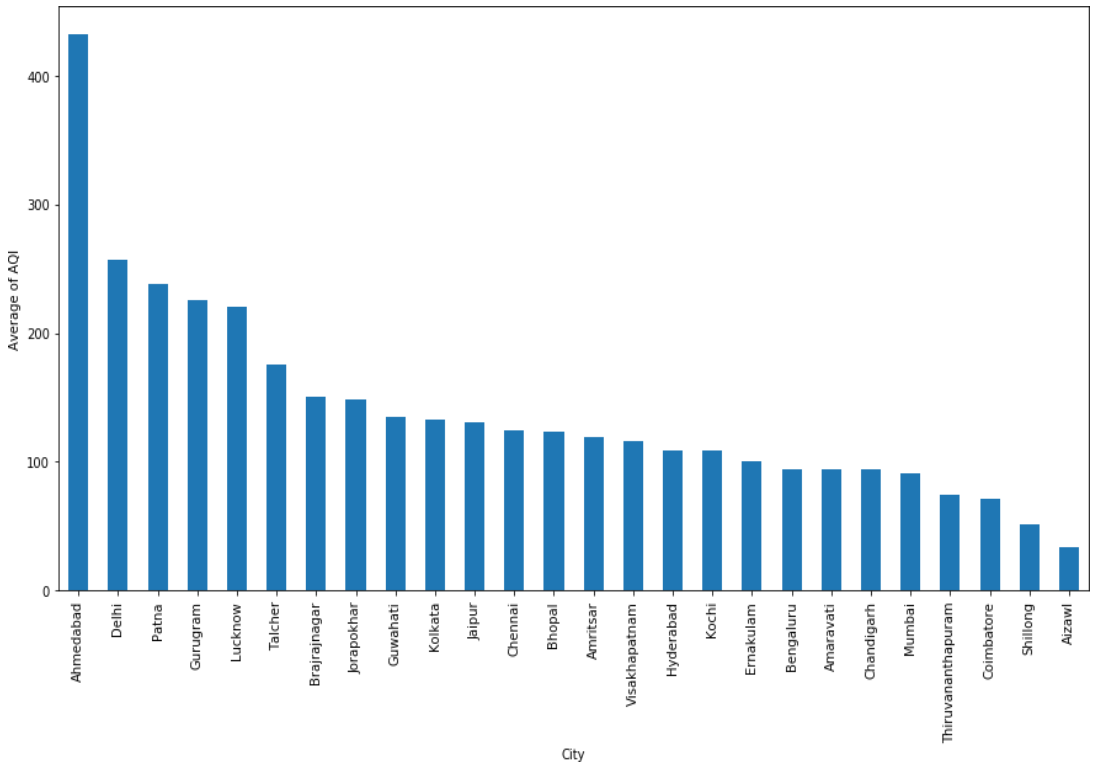
From the plot, can say that most of the variables are skewed. All the variables are right skewed. Skewness of the Benzene, Toluene and CO is on higher side than other variables.

- Multivariate Analysis

  For multivariate analysis, plot graphs of all variables with AQI variable. It shows that the PM2.5, PM10 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 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.



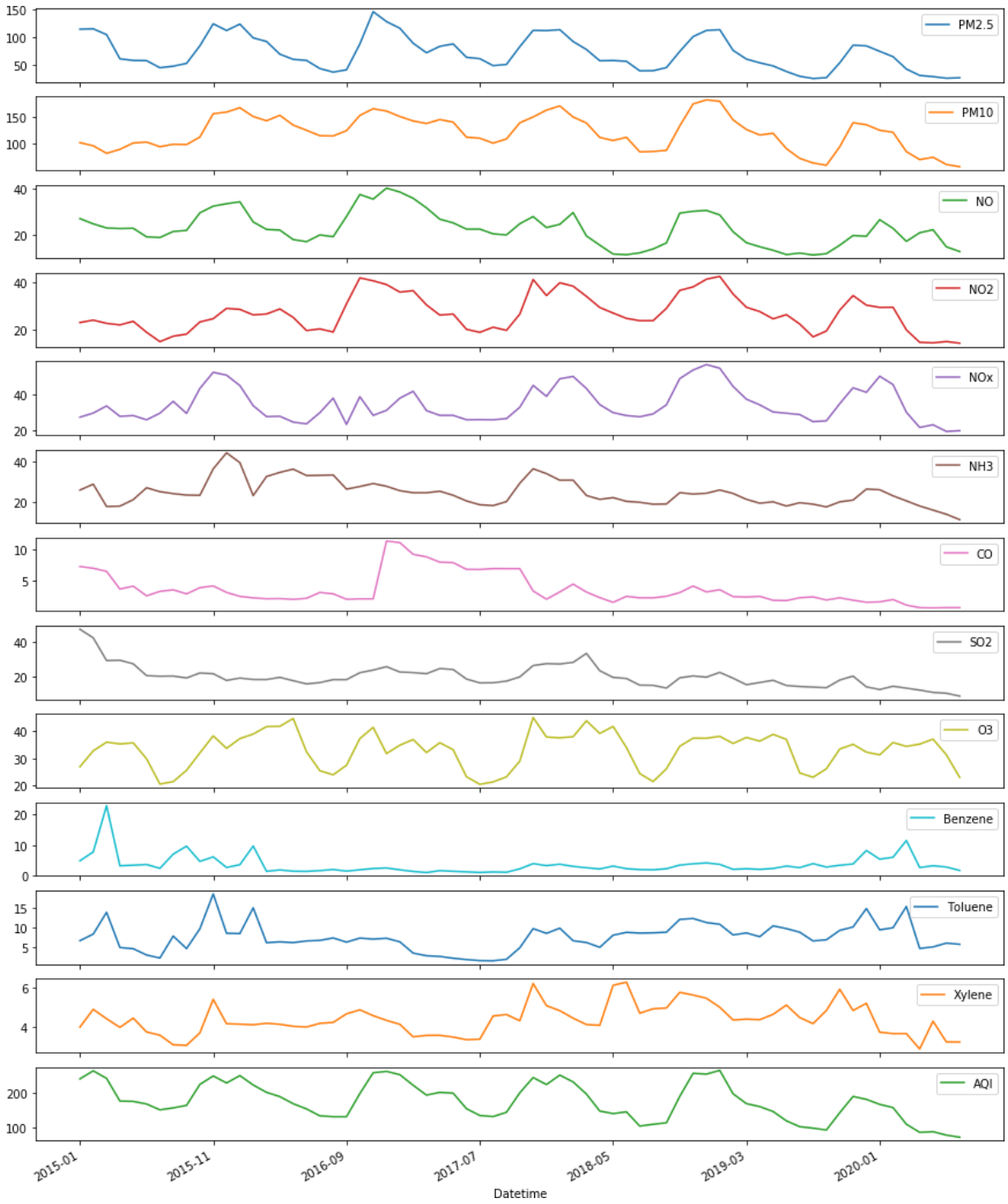
Also check for the city wise average AQI, which plot is shown in below.



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.

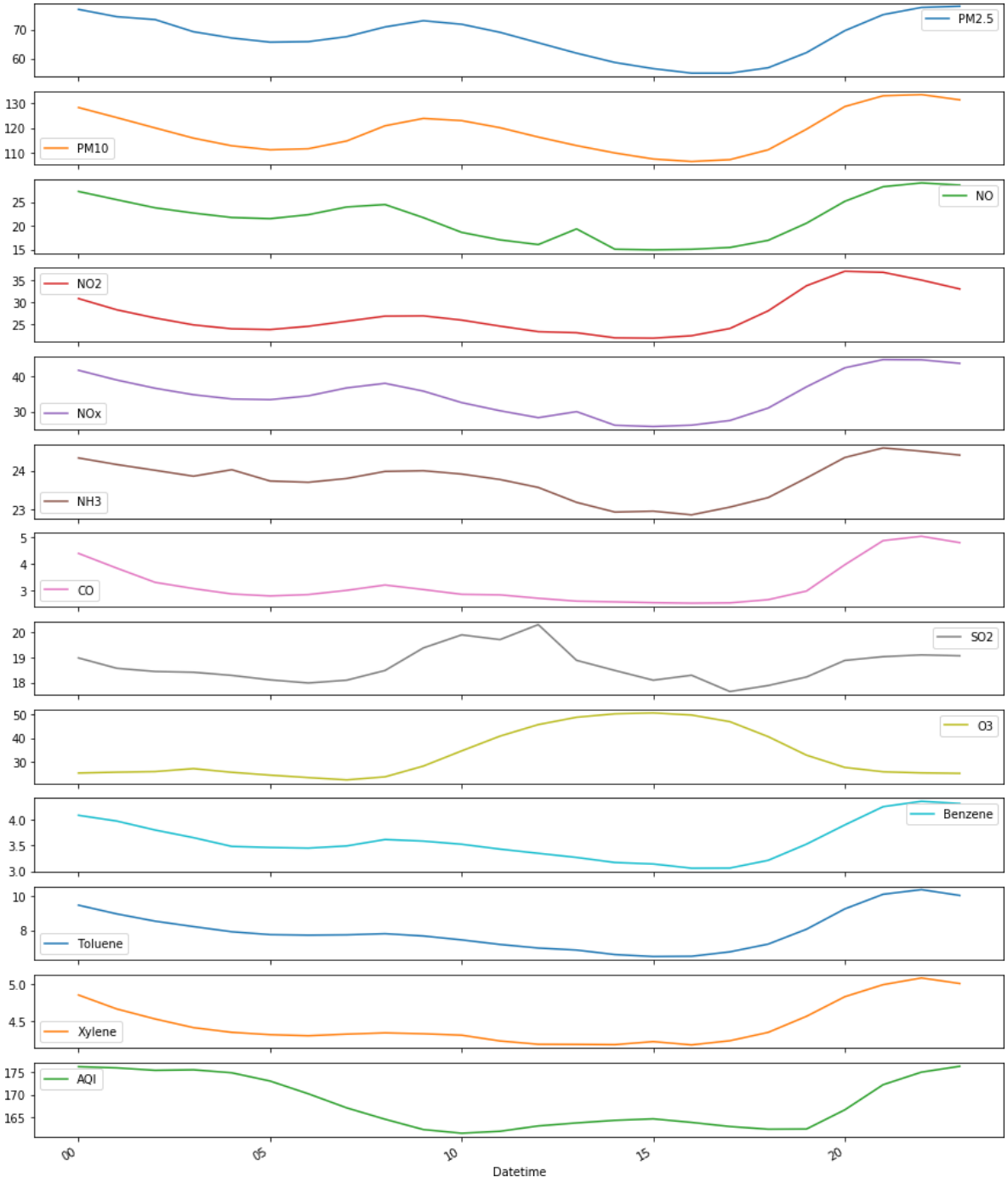
- Time Series Visualization

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



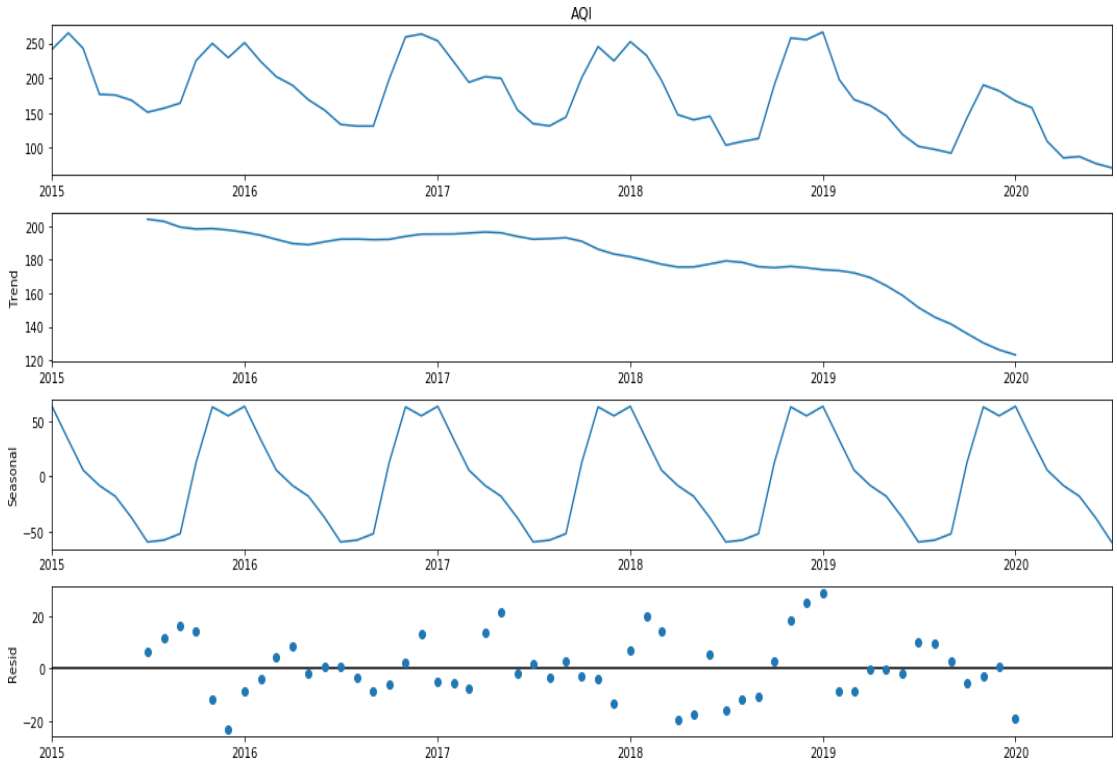
From the above graph, can see that the AQI is decreases after January 2020, this is because of the COVID Pandemic. Industries were completed closed; Transport was halted for 5-6 months. So due to that we can see the Decline in AQI During 2020

Also check the visualization by hour considering all over India.

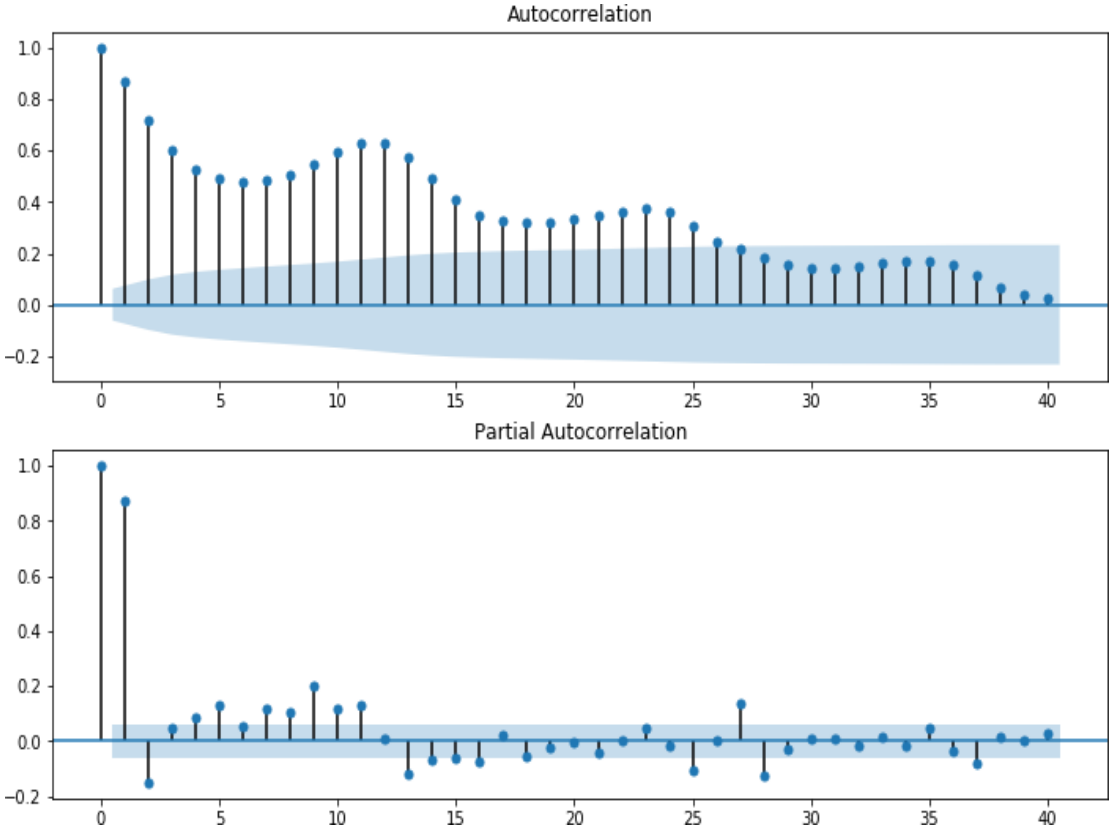


From the plot of pollutants hour wise, can say that the AQI and all other pollutants are having low average during night and higher at mid of the day. As per the scientific analysis, higher the temperature, lower the AQI. During night temperature is low so AQI is high and during day it is vice versa.

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.



And after checking decomposition, check for the stationarity with adfuller test. From the pvalue of adfuller test determines the data is stationary or not. From the test statistics and pvalue, pvalue is less than the 0.05, which means reject the null hypothesis and our data is stationary.
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.



- 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 (3,1,3) 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 (1,1,2), Its AIC value is also low than the ARIMA (1,1,2) and log-likelihood is higher than the ARIMA (1,1,2).

- 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)