Impact of Air Pollution on our Lives

by-DINESH Y

DATA

Dataset Taken from Kaggle:https://www.kaggle.com/rohanrao/air-quality-data-in-india

Description

- Air is what keeps humans alive. Monitoring it and understanding its quality is of immense importance to our well-being.
- The dataset contains air quality data and AQI (Air Quality Index) at hourly and daily level of various stations across multiple cities in India.

List of Cities there in Data

Ahmedabad, Aizawl, Amaravati ,Amritsar ,Bengaluru ,Bhopal ,Brajrajnagar ,Chandigarh ,Chennai ,Delhi, Gurugram ,Guwahati, Hyderabad ,Jaipur ,Jorapokhar ,Kochi ,Kolkata ,Lucknow ,Mumbai ,Patna ,Shillong, Talcher Thiruvananthapuram

Acknowledgements

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. Similar to air monitoring data, a dataset on noise decibel levels in India is available here: https://www.kaggle.com/rohanrao/noise-monitoring-data-in-india

More About the DATA

List of Datasets from the data are:

1.City_day 2.city_hour 3.Station_day 4.station_hour 5.Stations

1.city_day : Consists per day reading of pollutants across all cities
2.city_hour: Consists per hour reading of pollutants across all cities
3.station_day: Consists per day reading of pollutants across all AQI stations
4.Station_hour: Consists per hour reading of pollutants across all AQI stations
5.Stations: Consists of Stations IDs and Station Details (like: Place)

- For Analysis purpose city_day is taken as city_hourly data is more volatile and fluctating
- For city_day.csv
- COLUMNS are ['City', 'Date', 'PM2.5', 'PM10', 'NO', 'NO2', 'NOx', 'NH3', 'CO', 'SO2', 'O3', 'Benzene', 'Toluene', 'Xylene', 'AQI']
- INDEX :Time Series(DD-MM-YY)

Assumptions

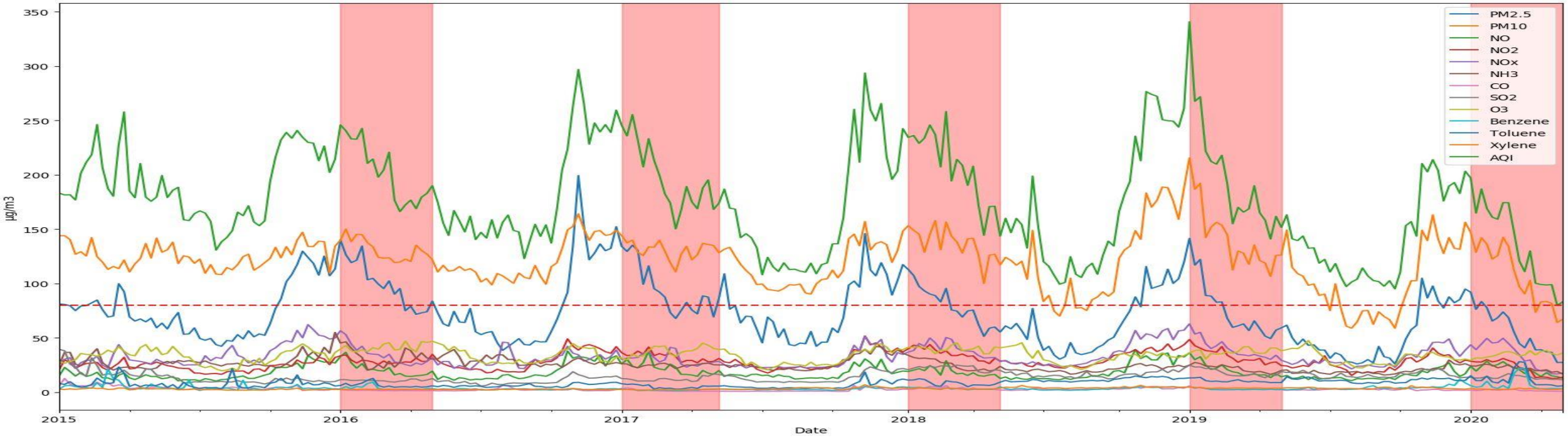
The data represents only the Metropolitan Cities in India and does not include rural regions due to there not enough Air Monitoring stations in India however the main pollution problem lies in cities only not in rural areas. What ever interpretation done on this presentation purely for top metropolitan cities.

To differentiate between the Vehicular and Industrial pollution I have taken statistical info from the website yourarticlelibrary.

Throughout the analysis AQI and pollutants depends upon major factors like Vehicle Pollution ,Industrial Pollution and I have ignored Minor factors like forest fires,cropping burnings,Diwali fireworks etc .

I have taken City_day data rather than city_hour data because city_hour has more volatile and fluctuating data point which fails to find General trend. Still this problem of Volatility persists in city_day dataset to overcome with this problem resampling of data based on WEEK is done which means instead of per day pollution data points, pollution level per week is taken into account.

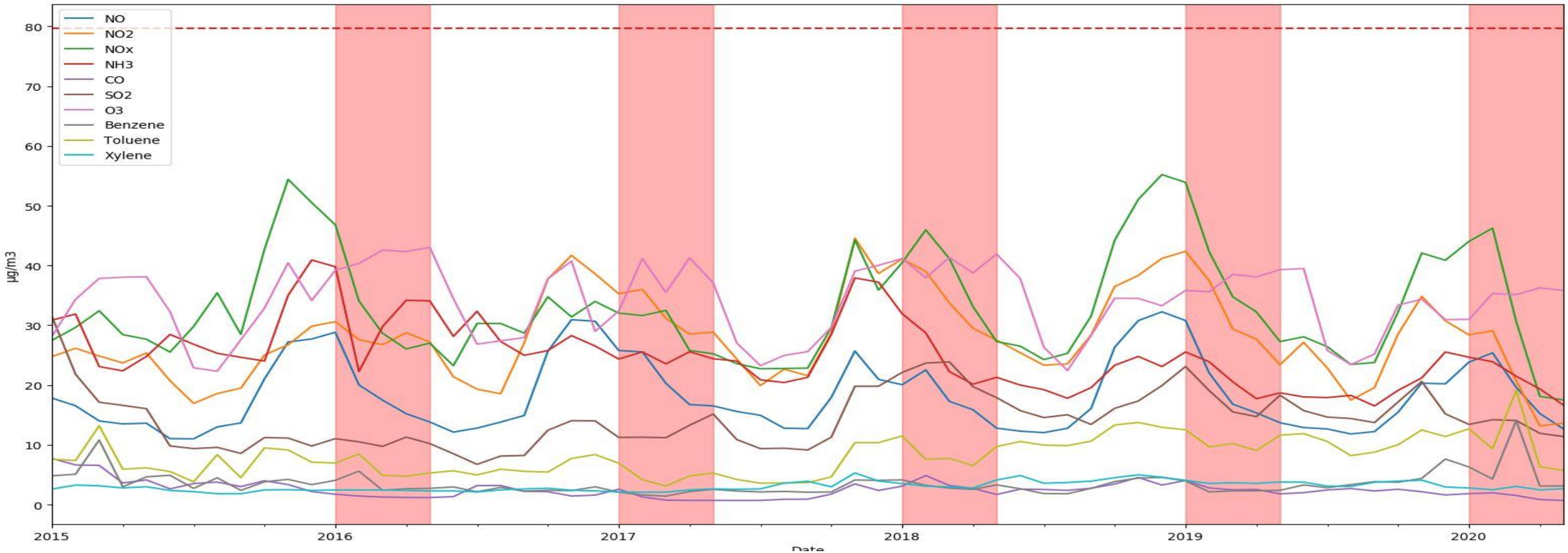
Average per Week of all Pollutants from all Cities vs Time



OBSERVATION:

1.All Pollutants shows the Seasons Effect.
2..Every year India has its AQI at its Peak at Beginning of Year .Strangely for year 2019 there hasn't been a Perfect Peak
3..The horizontal Red dotted line indicates the AQI level at 2020-05-01,which is All Time Low compared to Previous Years
4..The Concentration of P10 and PM 2.5 are higher than all other Pollutants.

Average per Week of all Pollutants from all Cities vs Time without AQI



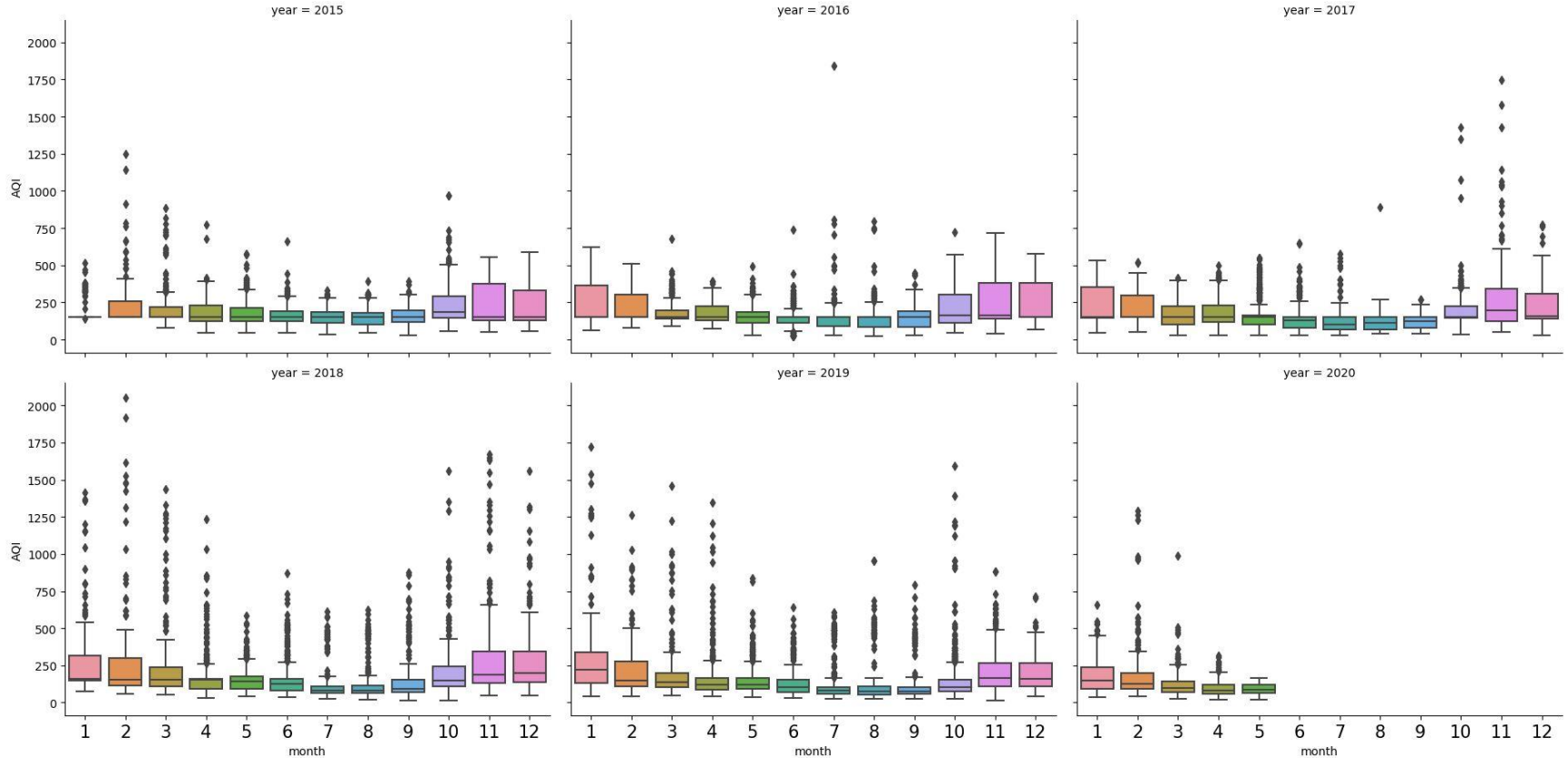
OBSERVATIONS:

1.Comparing to other pollutants NOx conc. Is very higher.
2.Lower ones is Xylene and Caron Monoxide.
3.It would better if we have correlation plot across all pollutants.

This plot shows how the spread of AQI levels across each month for each year.

We can also make interpretation about the outliers of AQI levels for each month.

Outliers in the sense how AQI level makes difference with common region



At glance the outliers are increasing from year to year and there is less spread of data in the middle of every year. Beginning and Ending of year has more AQI levels.

Before co-relation plot it would be better if we analyze each pollutant sources.

More About Pollutants

PM2.5, PM10 :

Road dust and tiny bits of, well, stuff sent into the air by stone processing and other crushing operations are common PM 10 pollutants. PM 2.5 comes primarily from combustion. Fireplaces, car engines, and coal- or natural gas–fired power plants are all major PM 2.5 sources.

NO, NO2, NOx, NH3, NOx :

These emissions from electricity generation. NO and NO2 are collectively known as NO_x because they are rapidly inter-converted during the day. Nitrogen oxides are produced from the reaction of nitrogen and oxygen gases in the air during combustion, especially at high temperatures. In large cities, nitrogen oxides are produced from fuel combustion in mobile and stationary sources.

The combustion of gasoline in automobiles emit nitrogen oxides into the atmosphere (mobile source). Stationary emissions come from coal fired power plants, electric power plant boilers.

Natural sources: Mainly thunderstorms due to the extreme heat of lightning. Forest fire is another natural source.
Biogenic: Agricultural fertilization and the use of nitrogen fixing plants (through nitrogen fixation by microorganisms).

OZONE:

Automobile exhaust and industrial emissions release a family of **nitrogen oxide gases** (NOx) and **volatile organic compounds** (VOC) by-products of burning gasoline and coal. NOx and VOC combine chemically with oxygen to form ozone during sunny, high-temperature conditions of late spring, summer and early fall. High levels of ozone are usually formed in the heat of the afternoon and early evening, dissipating during the cooler nights.

SULPHUR DIOXIDE:

Industrial activity that processes materials that contain Sulphur.
Example: The generation of electricity from coal, oil or gas that contains **Sulphur.** Some mineral ores also contain **sulfur**, and **sulfur dioxide** is released when they are processed.

India is the largest emitter of sulphur dioxide (SO2) in the world, contributing more than 15 per cent of global anthropogenic emissions. The primary reason for India's high emission output is the expansion of coal-based electricity generation over the past decade. The greatest source of SO2 in the atmosphere is the burning of fossil fuels in power plants and other industrial facilities.

BENZENE:

The major **sources of benzene** exposure are tobacco smoke, automobile service stations, exhaust from motor vehicles, and industrial emissions. Vapors (or gases) from products that contain **benzene**, such as glues, paints, furniture wax, and detergents

TOULENE:

The primary sources of toluene are the industries that manufacture it or use it in production. Some of the industries that manufacture it or use it in production are oil refiners, chemical industry, rubber manufacturers, pharmaceutical industry, metal degreasing, printing, manufacturers of paints, varnishes and lacquers. These emissions mainly are to the air, but are also to the soil and water.

Other possible emitters of toluene are vapors and spilling of petrol, commercial and household painting and paint, varnish and lacquer removal, tobacco smoke, and consumer products containing toluene. These emissions are to the air unless there is a spill.

References:

http://www.npi.gov.au/resource/toluene-methylbenzene
https://www.downtoearth.org.in/news/air/india-emits-the-most-sulphur-dioxide-in-the-world-66230
https://science.howstuffworks.com/environmental/green-science/gasoline.htm
https://www.nrdc.org/onearth/particulars-pm-25
https://www.aeroqual.com/meet-the-nitrogen-oxide-family

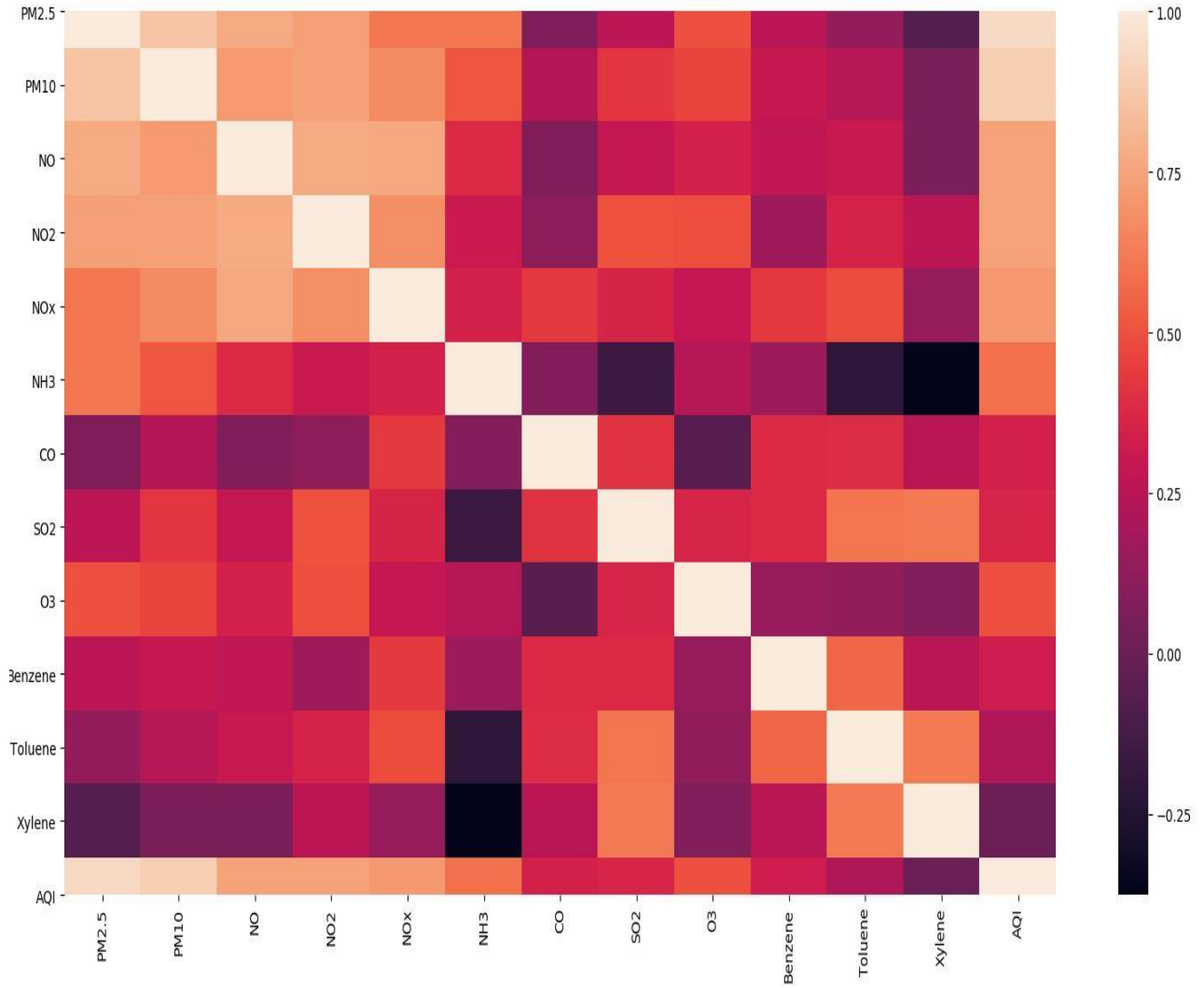
CORELATION AMONG ALL POLLUTANTS

Find Relationships Between Multiple Time Series Using Spearman

Kendall Tau or Spearman rank can be used to compute the correlation coefficient between variables for which the relationship is thought to be non-linear.

Range: [-1, 1]
0: no relationship
1: strong positive relationship
-1: strong negative relationship

- Xylene Corelated with Toluene,SO2
- Toluene Corelated with SO2,Benzene and Xylene
- Benzene Corelated with Toluene
- O3 Corelated with PM2.5
- SO2 Corelated with NO2,Xylene and Toluene
- NH3 Corelated with PM2.5
- PM 2.5 Corelated with PM10,NO,NO2,Nox
- CO is correlated with NOx.
- NO2 highly corelated with PM202,PM10,Nox,NO
  moderately corelated with SO2 and O3



Out of all Pollutants CO and NO2 are Very Highly Toxic

So Analysis of Cities Prone to CO and NO2 Would be Helpful

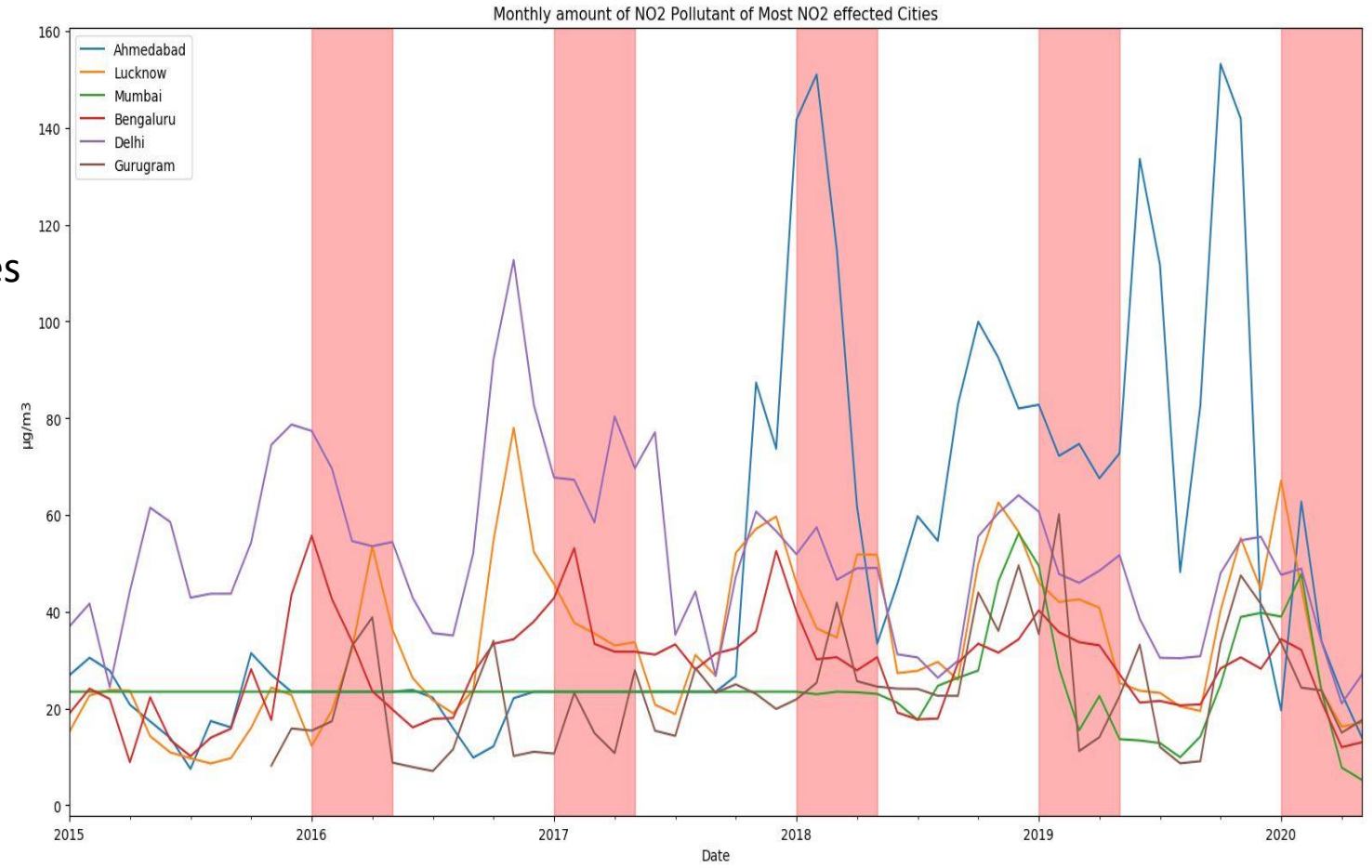
Analyzing Most Prone Cities to NO2 Pollution

1.Delhi used to have high amount of NO2 content Compared to others until 2017 ending,
but Ahmedabad Has higher than Delhi in recent times

2.Remaining Cities have Similar Trend.

3.NO2 content depend upon the NOx and NO due to Rapid interconversion among them in day time.

4.From Correlation Plot it depends upon O3, SO2,PM2.5,PM 10.



RECOMMNEDATIONS:

Encourage and provide public transport systems like BRTS and MetroRails
Identify toxic pollutants emitted by industries manufacturing Industries and put high restrictions on them

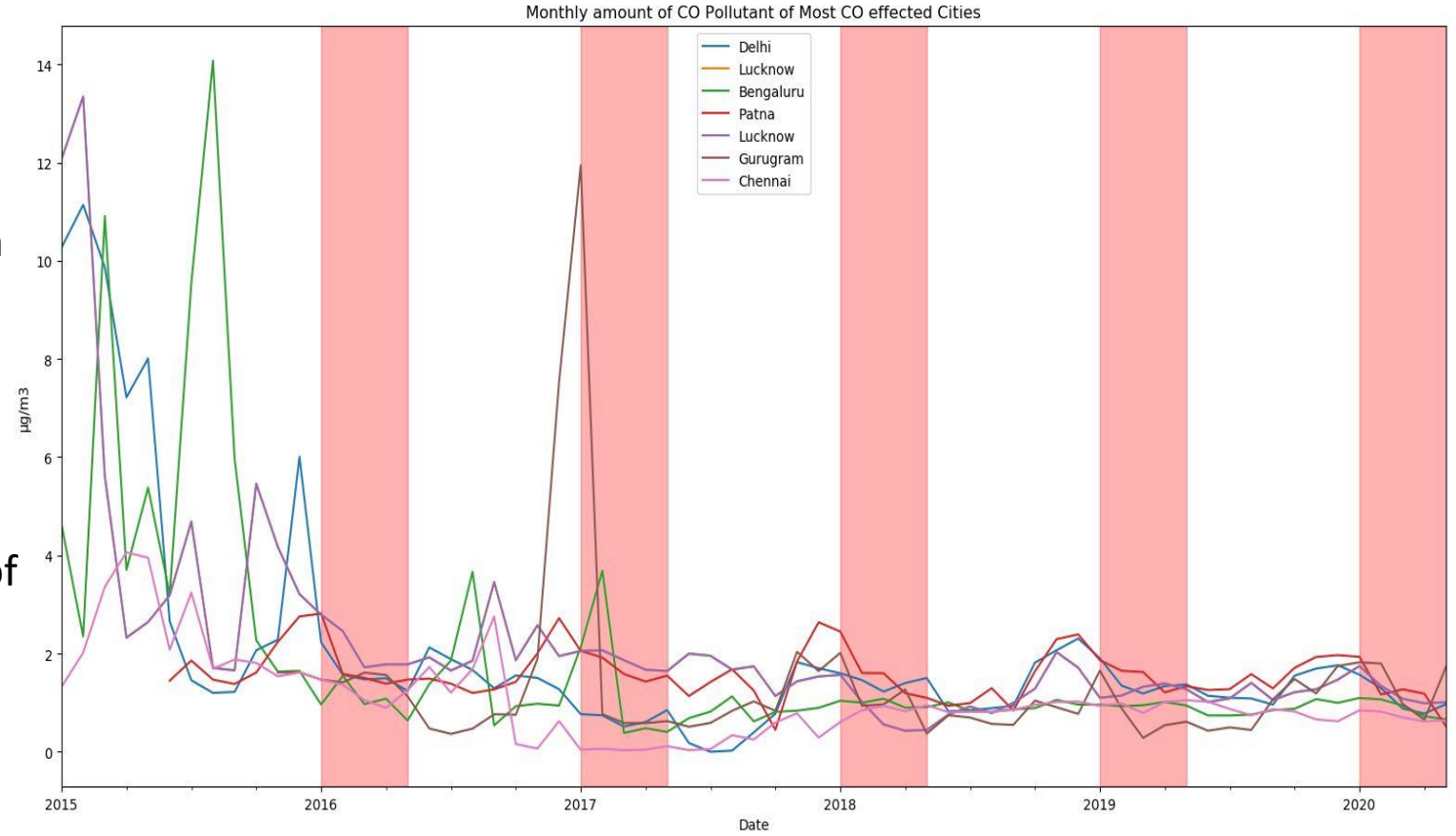
Analyzing Most Prone Cities to CO Pollution

1.Previously the CO pollutant is higher in Gurugram But declined after 2017

2.After 2017 the Most effected Cities by CO have Similar Trend.

3. The main source of CO is complete combustion of fossil fuels and the latter by automobile exhausts

4.From Correlation Plot CO is moderately corelated with NOx

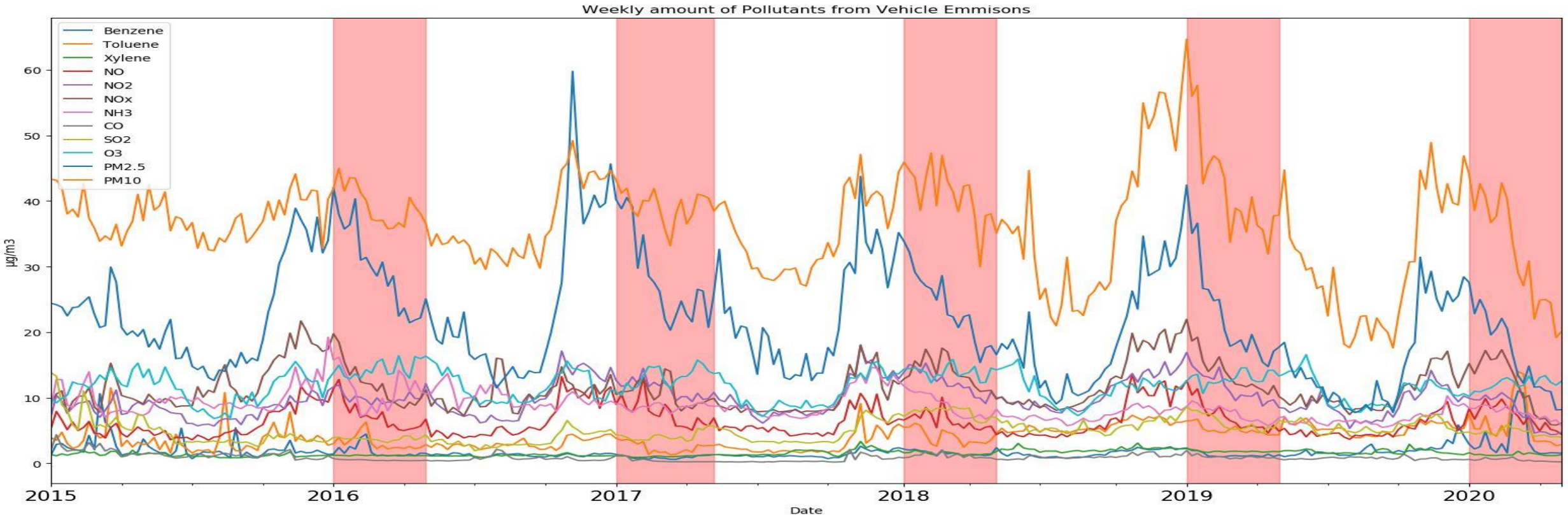


Vehicular vs Industrial Pollution

In the major Metropolitan cities,
vehicular exhaust accounts for 70% of all CO,
50% of all hydrocarbons,
30-40% of all oxides and
30% of all SPM.
So rest of pollution is caused by Industries ,Fire Forests ,Burning Crops etc. in Cities.

Most air contaminants do not have an associated AQI. Many countries monitor ground-level ozone, particulates, sulfur dioxide, carbon monoxide and nitrogen dioxide, and calculate air quality indices for these pollutants.

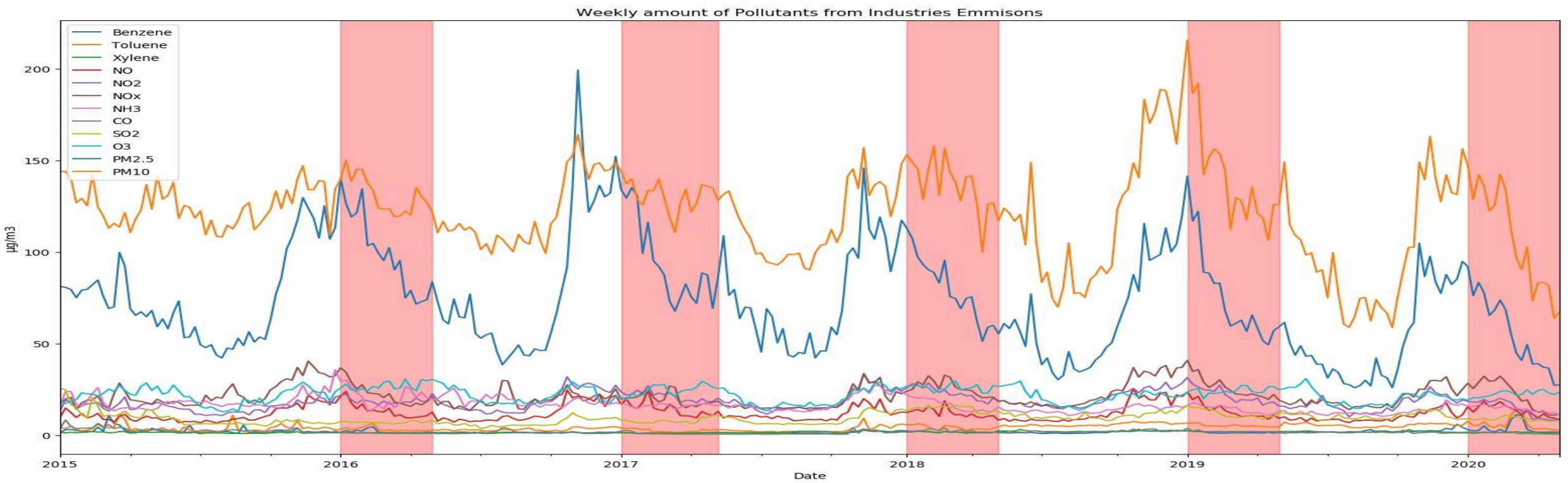
FROM WIKIPEDIA
The 51% of pollution is caused by the industrial pollution, 27% by vehicles, 17% by crop burning and 5% by diwali fireworks in India.The data availability only on Urban Places makes analysis only on Metropoltian Cities.



1.Surprisingly, Benzene shows all time high record in February Ending 2020 year although Benzene, Xylene and carbon monoxid have nearly same trend across the years.

2.(NO,SO2,Toluene), (NOx , O3, NH3), have similar trends.

3.Every pollutant has dropped drastically at the ending of April month, except Xylene and SO2

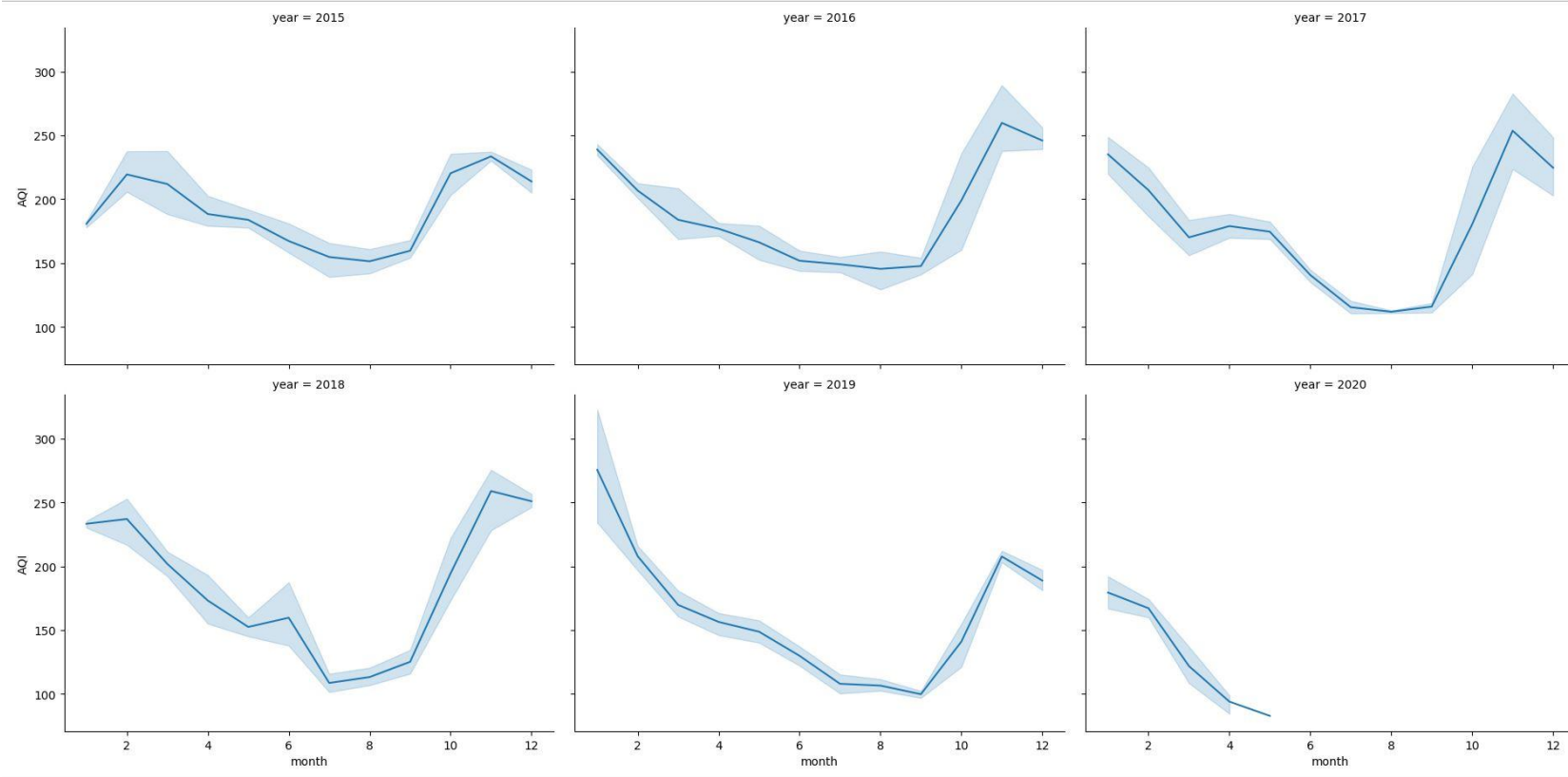


PM10,PM2.5 have higher Emission’s.

PM10 records all time high in the beginning of 2019.But at 2020 there isn’t all time high Record value.

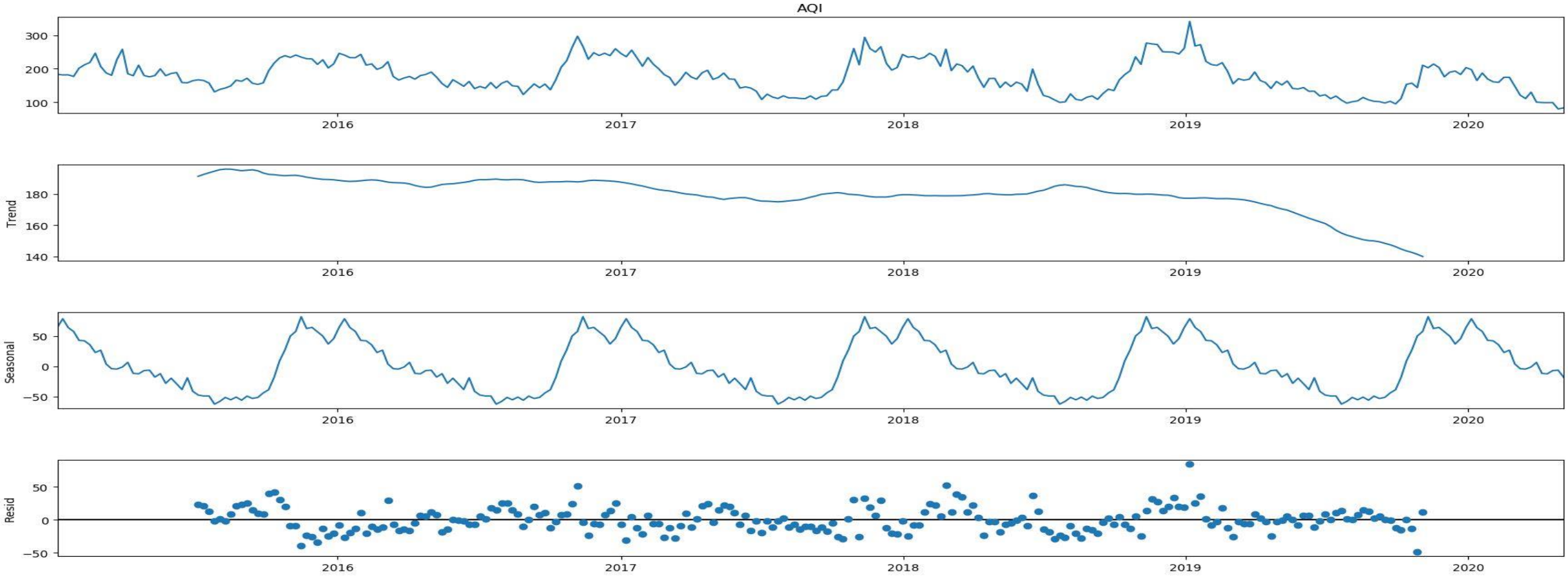
In the beginning of Recent years 2018 and 2019 the AQI Level is 230 and 280 nearly But Beginning of 2020 the AQI is nearly 174 Which is all time low. WHY?

This can be answered if we observe between the subplots.
NOTICE, the AQI level in the beginning of the Particular year is depends on how high the AQI has raised in its ending of predecessor year.



From plots the AQI level is raising steadily from 9th month to 11th month but ,this raise hasn't happened for year 2019 that reflects low AQI in in the beginning of 2020 year.

Components of AQI Time Series Data



Trend : The trend of pollution is almost same from 2017 to mid 2018 and it has decreased due to COVID 19 Impact.

Seasonality : From Seasonal sub-plot its clear that the AQI is seasonal

FORCASTING MODEL MAKING

List of Algorithms available for Forecasting's:

ARMA and ARIMA

Autoregressive Integrated Moving Average, or ARIMA, is one of the most widely used forecasting methods for univariate time series data forecasting. It supports both an autoregressive and moving average elements. Although the method can handle data with a trend, it does not support time series with a seasonal component.Thn extension to ARIMA that supports the direct modeling of the seasonal component of the series is called SARIMA.

SARIMAX

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.
It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

RNN’s – LSTM

LSTM is a special kind of RNN composed of a set of cells with features to memorize the sequence of data. The cell captures and stores the data streams. Further the cells inter-connect one module of past to another module of present one to convey information from several past time instants to the present one. Due to the use of gates in each cell, data in each cell can be disposed, filtered, or added for the next cells.Still there hasn’t been a good machine learning algorithms which used for Time Series Analysis, even SARIMAX algorithms outperforms Machine Learning LSTM’s Neural nets.

Modelling Seasonal Auto-Regressive Integrated Moving Average (SARIMAX)

For statsmodel library SARIMAX Documentation clickhere.

Seasonal **A**uto**R**egressive **I**ntegrated **M**oving **A**verage with eXogenous regressors model
Simply Exogenous helps to tell our AQI data that PANDEMIC IS OCCURRED IN 2020,which makes 2020 year special for ur model.

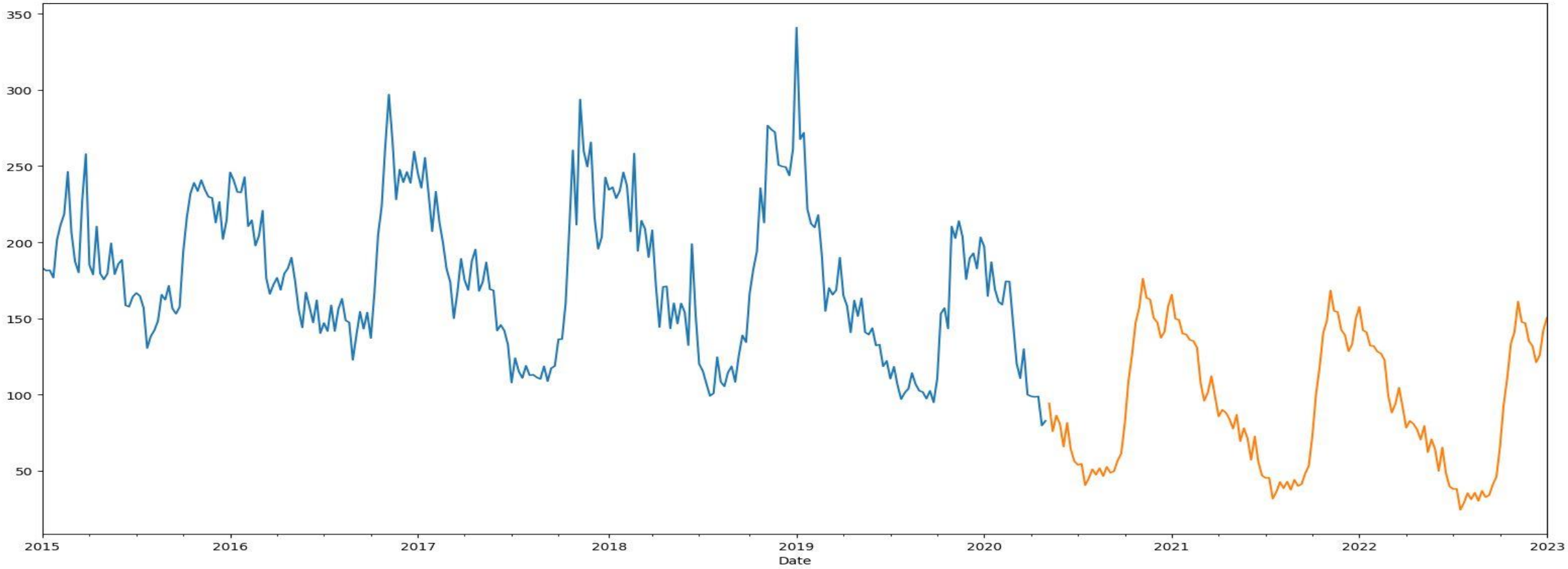
Exog Dataset:

From 2015-2019(ending) covid19 is not there so we create a column with time series 2015-2019(ending) and put 0 value to it which implies there are no covid19 cases.

From 2020(JANUARY-MAY) we impute number of cases per week to the same exog dataset column
Assuming covid 19 cases will fall by this year ending i.e December

From 2020(DECEMBER) – 2023 we impute exog dataset with 0,implying that there is no effect of covid19.

This way we can tell our model about COVID19 Impqct



The SARIMAX model depicts that the all time very low AQI will come in coming weeks in mid of this year. And the same Pattern will for next 2-3 years which might go wrong as same pattern can not occur.
This error occurred may be due to not creating appropriate exog dataset and model has been more biased on seasonal Effect and doesn't learned from exog.

The Main takeaway point from above plot is AQI is going to be ALL TIME LOW AQI recorded in past 8 year in next coming months.

SUMMARY

It has been observed that every pollutant concentration is low in COVID19 Pandemic. It has been Observed in past pandemics That the cities which are effected by Air Pollutions are more likely corelated with Pandemic Cases which means Air Pollutions Makes immune system weak.

It has been observed that all pollutants and AQI levels are low at mid of the year this may due to Monsoons. As raindrop falls through the atmosphere, it can attract tens to hundreds of tiny aerosol particles to its surface before hitting the ground. Sunshine makes some pollutants undergo chemical reactions, producing smog and various gases.

Definitely there will be correlation among the air pollutants as they react with each other with respects to atmospheric conditions like wind,humidity,pressure etc.
The 51% of pollution is caused by the industrial pollution, 27% by vehicles, 17% by crop burning and 5% by Diwali fireworks for country including all other Places that this data doesn't has. The strong variation between CITIES pollution and OVERALL INDIA pollution can be explained as there are very Large scale industries which will we situated far away from cities has no contribution to CITIES data which I have analyzed.

The AQI levels in 2019 at cities are not raised as much as previous year at corresponding peak time. This might due to less Vehicular Pollution in Cities .Although in upcoming years the AQI levels contributed by Vehicles may drop significantly in cities if India correctly adopt the Electric Vehicle Policies.As of now 53 percent Air Pollution is from industries. But the industrial pollution will raise rapidly post covid19 pandemic as many MNC manufacturing companies trying to migrate their Factories from china to india.