School of Business and Economics
Loughborough University

22BSP417

Process and Programming for Analytics

"Data Science Report"

Word Count: 1468

Data Science Report

Abstract:

In this analysis report, a huge amount of data set was taken and analyzed in order to find if the air quality improved in India. The available data is between 2015-01-01 00:00:00 and 2020-05-01 00:00:00.

This includes the period when the Indian government decided to lockdown 1.3 billion population on March 25th, 2020. This was done in order to reduce the rapid incline of covid cases. A major difference was seen in the air quality for some cities due to the lockdown.

A Year wise and Monthly comparison for different groups of pollutants are done suggesting that the levels actually decreased steeply while the lockdown. Using the data, vehicular and industrial pollution levels in different cities are also analyzed.

Data collection:

The data was uploaded on Kaggle as a dataset but the original source is the Central Pollution Control Board: https://cpcb.nic.in/ which is the official portal of the Government of India. The number of rows present in the dataset is 26,220 which consists of the name of the cities that are monitored with the air quality data and AQI (Air Quality Index).

There were a total of 24 cities that are analyzed throughout the dataset. However for some parts of the analyses only the major cities like Ahmedabad, Mumbai, Hyderabad, Delhi, Bengaluru, and Chennai. in India are considered.

Fig. 1: Cities being Monitored

Missing Values and Percentage

It was found that a large number of values were missing from the data. This can be caused by a number of reasons such as:

1. The station in the city did not have the device to capture reading
2. Technical fault in the meter

After going through the missing values data, it was known that 14 out of 16 columns had null values. A table was made to keep a check on the number and percentage of missing values for each pollutant.

	Missing Values	% of Total Values
CO	1961	7.500000
NO2	3217	12.300000
NO	3233	12.300000
SO2	3544	13.500000
O3	3660	14.000000
NOx	4043	15.400000
AQI	4282	16.300000
AQI_Bucket	4282	16.300000
PM2.5	4289	16.400000
Benzene	5287	20.200000
Toluene	7555	28.800000
NH3	9847	37.600000
PM10	10766	41.100000
Xylene	16807	64.100000

Table1: Missing values and % of total values

Air Pollutants combined into groups

The Air pollutants were divided into groups based on the effects they have on the environment and human body such as:

1. Nitrogen family: NO, NO2 and NOx

   NO2 and NO are the most dangerous oxides in the group nitrogen. Studies have shown that symptoms of bronchitis in asthmatic children have increased long-term exposure to NO_2.

2. BTX : Benzene, Toluene, Xylene

   BTX belongs to an important group of aromatic volatile organic compounds (VOCs) that are usually emitted from various sources especially working places like plastic, chemical, and leather industry. BTX plays a vital role in tropospheric chemistry as well as poses a health hazard to human beings.

3. Particulate_Matter: PM2.5 and PM10

    Particles such as dust, dirt, or smoke are to be seen with the naked eye. But the most hazardous particles are the smaller ones like Particulate matter. Some of the diseases caused by these are acute lower respiratory infections and lung cancer.

Apart from the missing values and combining air pollutants into groups, the ***Date format was not in the Date-time format*** which could have led to issues in the analysis. Therefore the date format was changed to the needed format for further analysis.

Data Analysis:

The data has been analyzed using various libraries such as pandas, numpy, plotly, matplotlib.pyplot,seaborn, pycountry, plotly.express, cufflinks etc.

Data was firstly cleaned, secondly grouped according to similarities, and then analyzed using various interactive graphs such as pie charts, histograms, bar graphs, and Scatter plots.

Methodology

Since the data is vast, it has been used in its best capacity to come to different aspects for conclusions.

1. Vehicle vs Industrial Pollution caused in different cities
2. Pollution trends over the years and months
3. Effects of lockdown (if any) on air quality

Air quality vs Nitrogen Oxides vs BTX levels - City wise

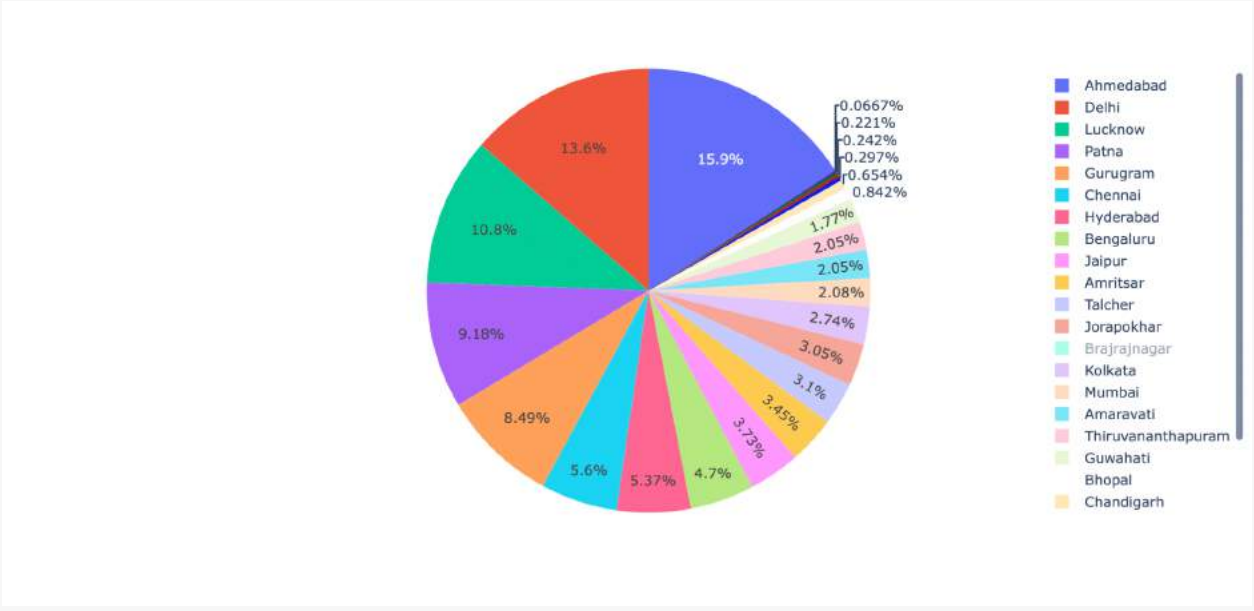


Fig2:Interactive Pie Chart comparing the Air Quality Levels city wise

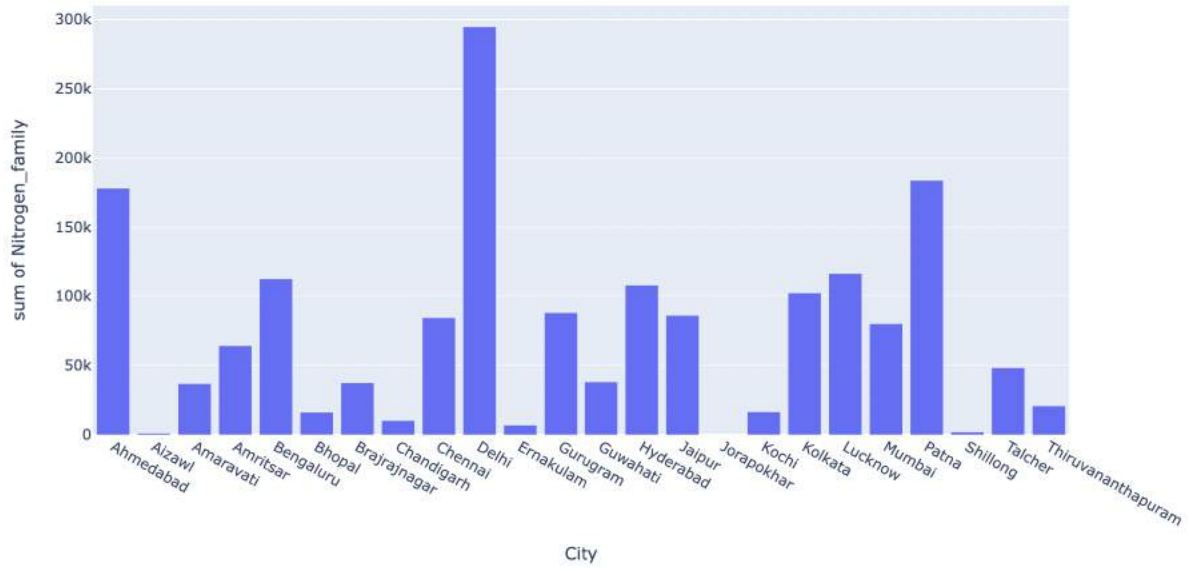


Fig3: Bar graph comparing the Nitrogen Oxides level city wise

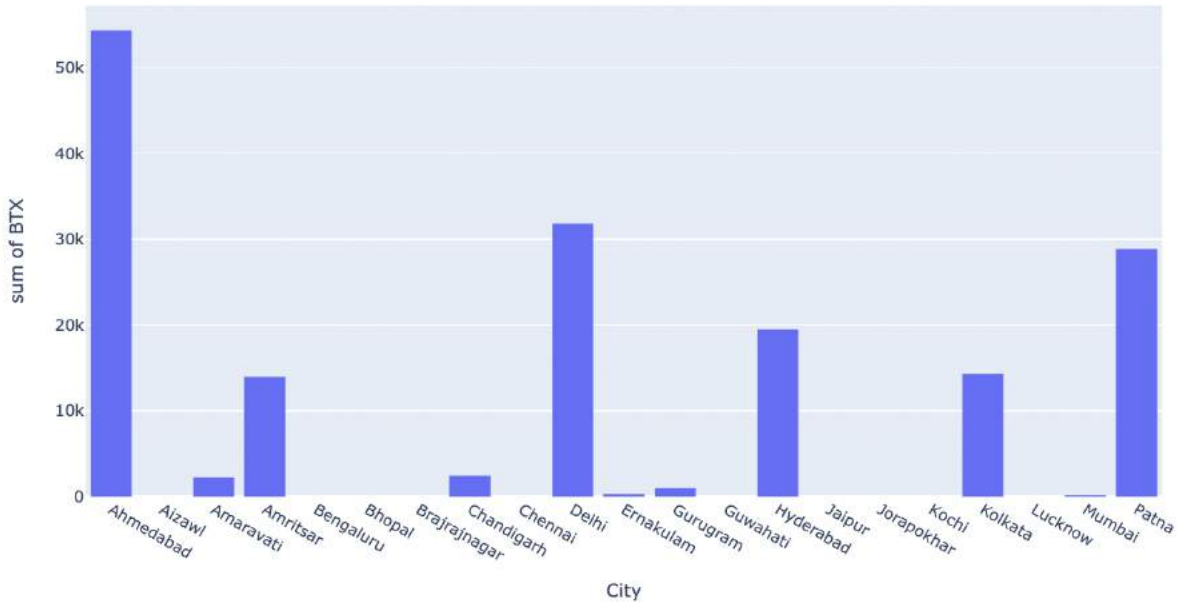


Fig4: Bar graph comparing the BTX level city wise

A pie chart has been used to make a sense of part to whole data of the cities' AQI levels. This helps to signify the levels of Air quality levels in respect to the city. It is an interactive plot, the major cities can also be selected with the others being unselected. This will help to analyze which cities are taking the most part in air quality within the whole. From the above Fig2. It is prominently shown that Ahemdabad has the worst Air quality levels followed by Delhi, Patna, and Hyderabad.

To compare the highest or most common, and how other cities compare against the others in terms of nitrogen oxides, a bar chart has been used. As seen in Fig3, It is topped by Delhi, followed by Patna and Ahemdabad. Similarly in Fig4. Ahmedabad tops the list for BTX components in the air followed by Delhi.

Vehicular vs Industrial Pollution

To further know the reasons for the pollution levels, analyses have been done on the basis of Industrial and Vehicle pollution. This analysis has been done using bar charts to get an effective visualization of which are the top 10 cities for both categories.

For this analysis, Air pollutants were divided on the basis of their emissions. Particulate matter, Nitrogen Oxides, and Carbon were categorized as Vehicular Pollutants and the others as Industrial Pollutants.

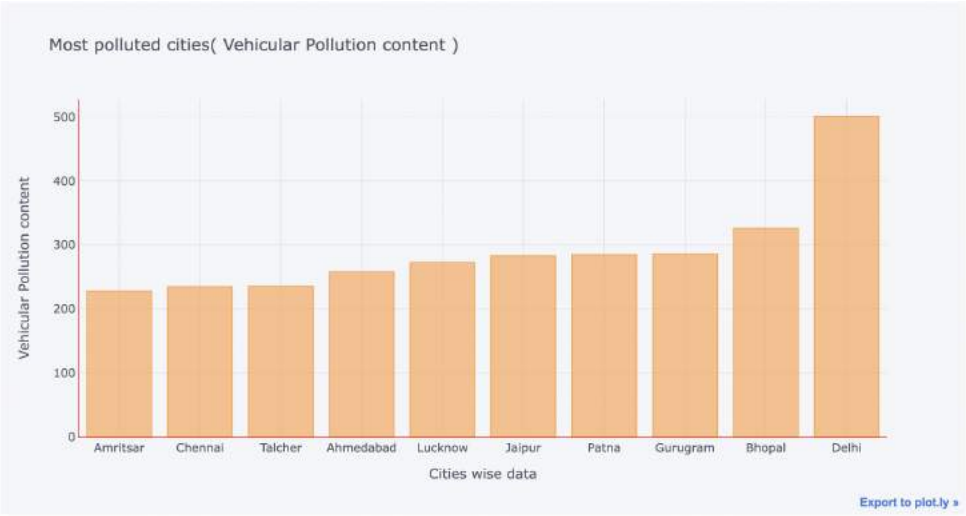


Fig5: Bar graph comparing the Vehicular Pollution level city wise

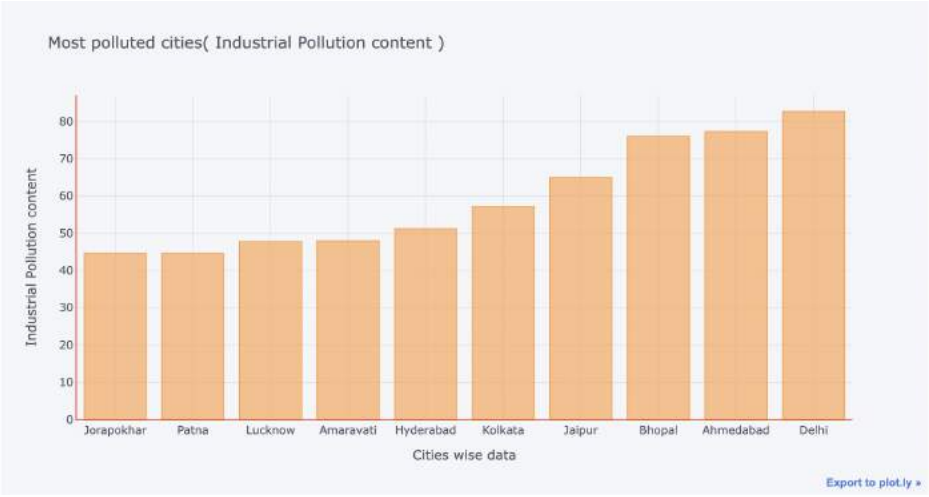


Fig6: Bar graph comparing the Industrial Pollution level city wise

Fig5 and Fig6 represent that Delhi is the most polluted city both Vehicular and Industrial pollution wise. Ahemdabad is more polluted than Bhopal in terms of Industrial pollution and vice versa for Vehicular pollution.

Yearly analysis of various pollutants

Due to the vast data, analyzing pollution levels pollutant-wise over the years is a little difficult. Therefore, a scatter plot is used instead of a line graph for the same. This plot has the levels for “air pollutants” which were identified as the most hazardous pollutants causing pollution.

Air pollutants categorized:

1. particulate matter
2. Nitrogen family
3. CO
4. SO2
5. O3
6. BTX'

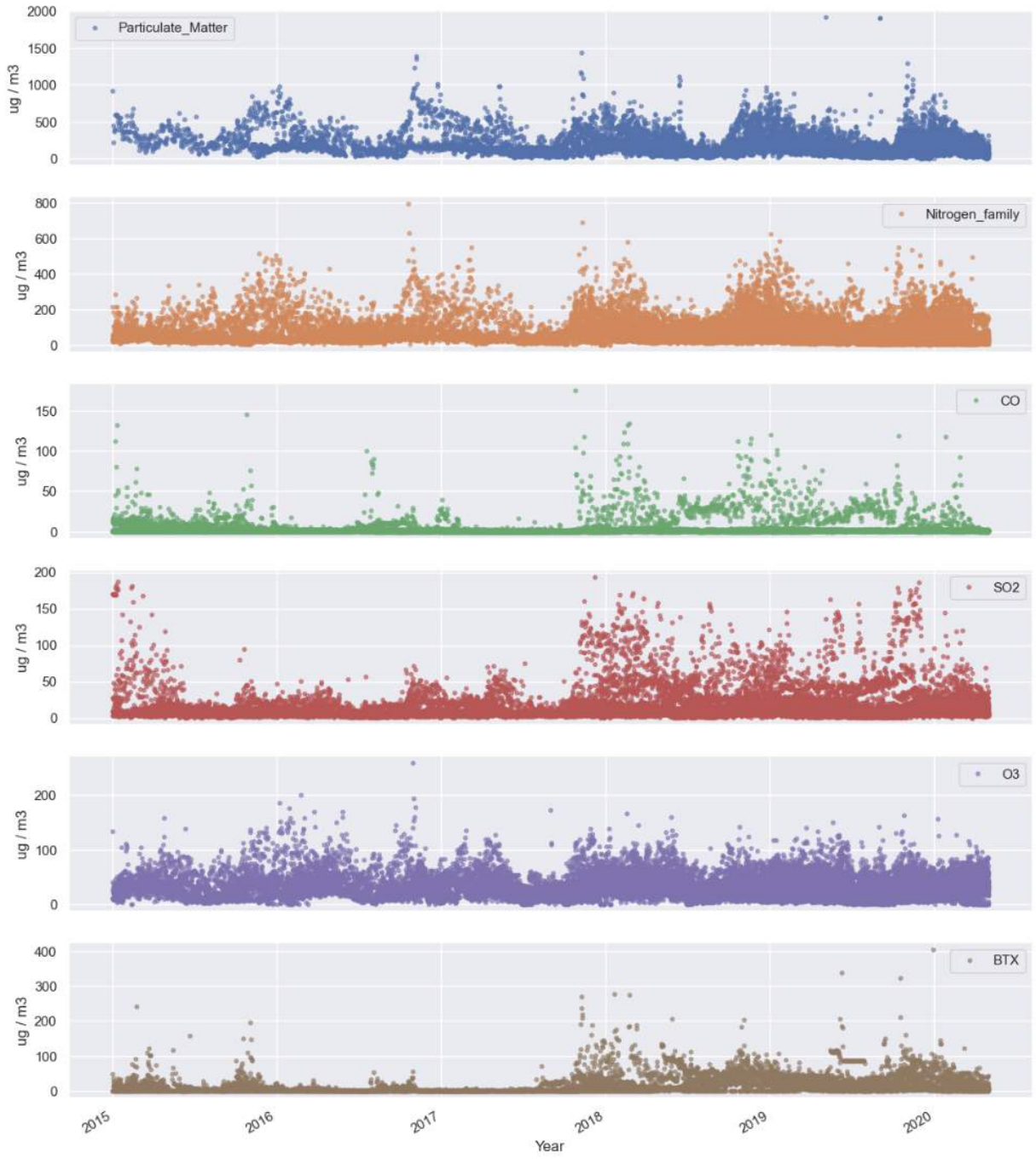
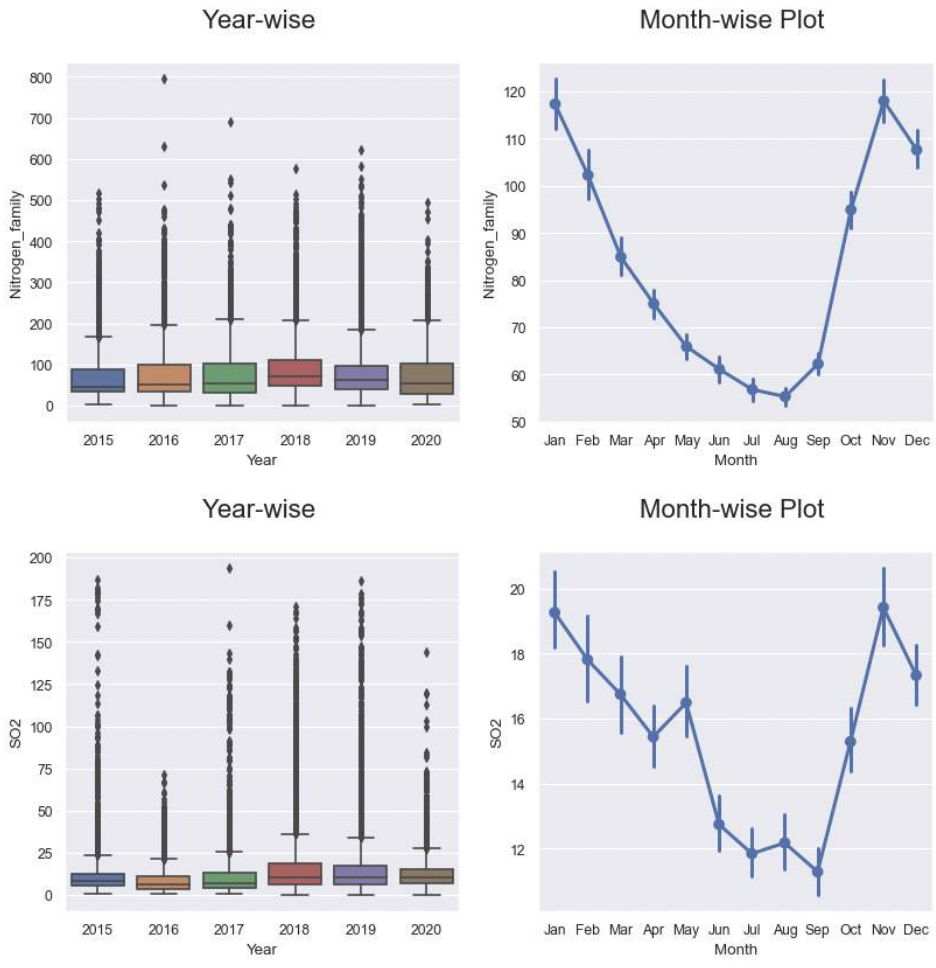


Fig7: Scatter plot comparing the Pollution level over years

Year and Month wise Comparison of Air pollutants

Box plots have been used to provide high-level information at a glance, and offer general information about the air pollutants categories by showing where the majority of the data lies in the “box”. It also gives a more in-depth analysis by including the outliers, the median, and the mode. BTX, Nitrogen family (Nitrogen oxides), Particulate matter and SO2 have been compared using the box plots.

The trend for these box plots suggests that there has been a decline in levels since 2019 which can be an effect of the lockdown. Also, According to the monthly plot, the levels start rising again in winter for all the categories.



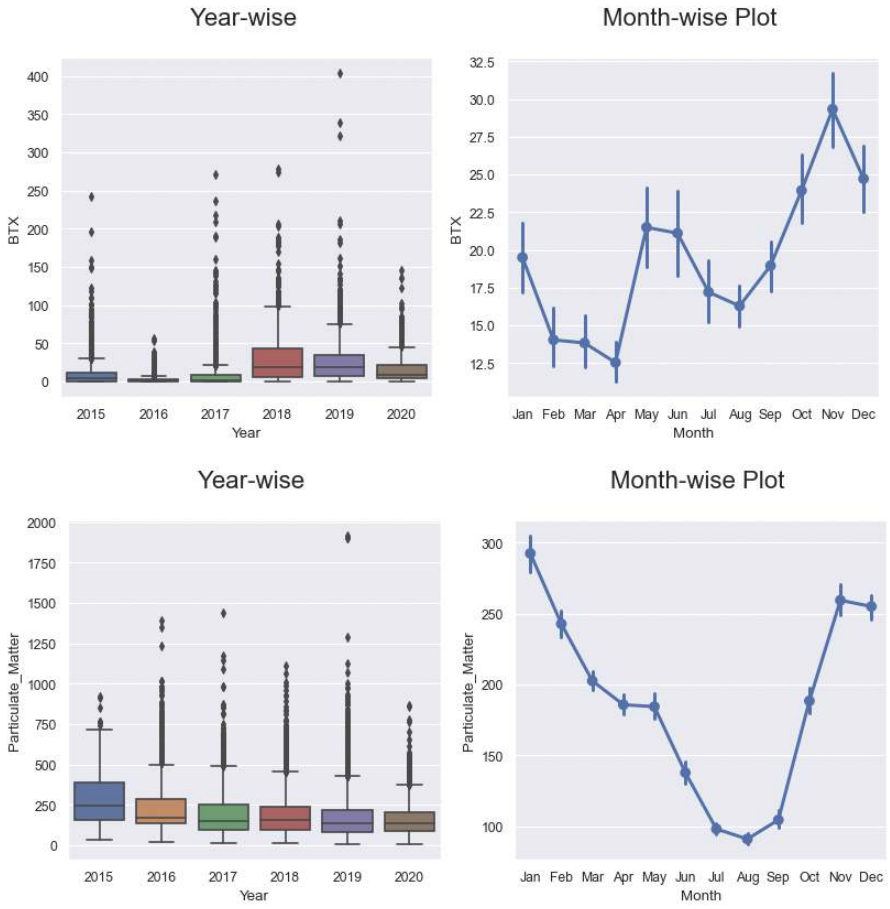


Fig8: Box Plot comparing the Pollution level over years and months

Comparing City wise data vs Air Quality over the years

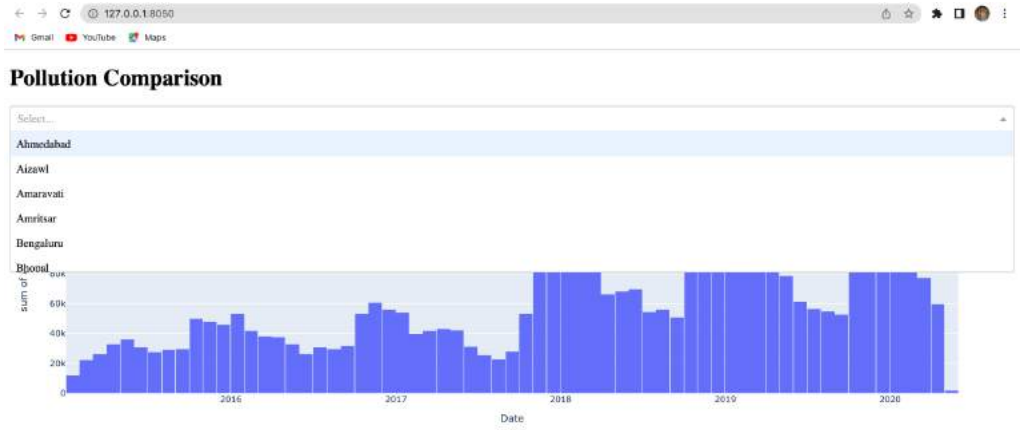


Fig9: Comparing the City wise data vs Air quality over years

In fig9. There is a comparison of city-wise data over the years. To analyze the data for the major cities. The trend shows that the major cities such as Delhi and Ahemdabad showed a better air quality situation after the lockdown compared to the before lockdown situation. Whereas there was no significant change in the situation for Mumbai and some other cities.

Conclusions and Recommendations:

Air Pollution over months

For Particulate matter (PM2.5 and PM10) there is pollution higher in the winter months rather than in the summer ones. The same is true for SO2 and Nitrogen Oxides. The pollution generated by winter crops can also be the reason for the same. Especially in the northern region of India. Pollution levels fall in the summer months specifically from July onwards. The sudden decline in the level of pollution can have a direct relation to the rainy season. Pollution levels in summers are not too high or low compared to monsoons and winters. A clear trend can be seen between pollution levels and the seasons.

Recommendation: To lower the pollution levels in winter which are majorly caused by the burning crops, Precautionary steps should be taken by the concerned authority to dispose off the stubble rather than burning it.

Air Pollution over the lockdown period

After the lockdown, the air quality increased which can be inferred as the cause of the lockdown period. Around March 2020, a steep decline can be noticed in pollution. After going through all the data, it can be suggested that there was an improvement in the cities in terms of air quality. Also, Delhi, Ahmedabad, Bhopal, and Patna are one of the most polluted cities in India.

Recommendation: Collecting data till 2022 will be a strong way to check if the air quality is back to its pre-lockdown time or not.