DECADAL POLLUTION ANALYSIS

AND

IMPACT OF COVID-19

Project report submitted in partial fulfilment of the requirements for the course CS685A: Data Mining in the academic year 2021-2022.

In the Department of Computer Science and Engineering

Indian Institute of Technology, Kanpur

Under the guidance of

Prof. Arnab Bhattacharya

Submitted by

Abhishek Sahu (21111002)
abhisheks21@iitk.ac.in

Archit Gupta(21111015)
garchit21@iitk.ac.in

Alok Kumar Trivedi (21111008)
alokt21@iitk.ac.in

Sai Kiran Tanikella (21111061)
tskiran21@iitk.ac.in

Amar Raja Dibbu (21111009)
amard21@iitk.ac.in

Acknowledgement

At the very outset of this report, we would like to extend our sincere & heartfelt gratitude to our project guide, Prof. Arnab Bhattacharya, for his active guidance, help, cooperation & encouragement, without which we would not have made headway in the project.

The liberty he gave, right from the commencement of the course to choosing the project, encouraged us to explore diverse areas of interest and work efficiently. It is an absolute honour for us to be a part of this course. The successful completion of this project has been possible only because of his constant guidance, motivation, and critical analysis.

Abhishek Sahu (21111002)

Alok Kumar Trivedi (21111008)

Amar Raja Dibbu (21111009)

Archit Gupta (21111015)

Sai Kiran Tanikella (21111061)

ABSTRACT

The motive of this project is twofold, Performing a Decadal analysis of air pollution in India for 2011-2020 and understanding the impact of the covid-19 on the pollution levels across the country. For this purpose, various methods are used at both state and city levels. Chi-square test, Z-score are used to find Outliers, Hotspots and coldspots of Cities and states. Various plots and heatmaps are used to understand the changing pollutant concentration levels in the country. Correlation between different pollutants with Number of industries, Number of motor vehicles and Population of the state is found. Later five-year average monthly trend for pollutants is plotted to understand the variation in their concentrations in different months. In the later part, the Impact of covid on pollution in different cities and states is analysed using both AQI values and pollutant levels before and after lockdown.

TABLE OF CONTENTS

Chapter 1
Introduction and Objectives of the project

1.1 Introduction

Pollution, in general, is defined as the addition of those substances that adulterate the original characteristics of the natural resources like air, water, land on this planet. These substances, accumulated in the environment, take a longer time to disperse, decompose naturally. Thus, causing a series of detrimental effects to the health and quality of life of all the living beings on the planet, most importantly, humans.

It is caused by various natural events like forest fires, volcanoes and anthropogenic sources, including various human activities from agriculture, animal husbandry, transportation, burning fossil fuels for energy generation, manufacturing industries, construction.

Significant forms of pollution with utmost concern to humans are Air pollution, Water pollution, Land pollution and Noise pollution. Air pollution is the contamination of the atmosphere by any chemical substances like Carbon Monoxide, Ozone, Sulphur Dioxide, Oxides of Nitrogen, Particulate Matter, Fly ash, Benzene and its derivatives or biological substances like animal dander and pollen.

The primary reason for the existence and the survival of life on earth is the abundant availability of oxygen and the non-existence or negligible amounts of other harmful chemicals in the air we breathe. Humans inhale 7-8 litres of air every minute. Pollutants, when present in quantities more than the prescribed limits can cause adverse effects in our body, ranging from respiratory diseases to heart stroke, cancer and even death. Modern diseases and the ongoing pandemic like COVID-19 have all been aggravated by the rising air pollution levels. It is now a fact that pollution in metropolitan cities has put the covid-19 patients at higher risk and led to mass hospitalisations. Besides these, the well known annual smog in the Northern plains of India and around the NCR region creates visual problems for the commuters. It leads to the shutting down of schools and offices, resulting in valuable time for education and economic activity.

According to the World Health Organization(WHO), 99% of the global population breathes air that exceeds the WHO guideline limits for the pollutants, with low and middle-income countries like India suffering the most. Also, estimates show that, annually, air pollution kills 7 million people worldwide. India is at the juncture of exposing ninety per cent of its population to air pollution every day. Global Burden of Disease (GBD) 2019 survey has put air pollution in the top 5 risk factors for deaths in India. Hence it is essential to study air pollution and its impact on global sustenance.

1.2 Objectives

The following are the objectives in this project with the support of data mining techniques.

1. **Performing Decadal analysis of air pollution in India using various pollutant levels.**

2. **Finding outlier states which are Hotspots and Coldspots during 2011-2020.**

3. **Correlation between different pollutants with Number of industries, Number of motor vehicles and Population of the state.**

4. **Finding the most polluted states in India with respect to SO2, NO2 and RSPM levels.**

5. **Clustering States based on their pollution levels.**

6. **Monthly analysis of Pollutants(NO2, SO2, PM) over last 5 years**

7. **Cities with Highest Pollutant Concentrations**

8. **Impact of Covid: Analyzing AQI of different cities over 2020.**

9. **Comparison of AQI before and after lockdown**

10. **Effect of Lockdown on individual pollutant levels**

Chapter 2

Datasets

In order to achieve the objectives mentioned in the previous chapter and make a proper analysis of the pollution in the country, certain data is required.
For the period 2011-2020 in India, the datasets required are listed below:

1. **Pollutant Concentration across India**
2. **Population Enumeration Data**
3. **State-wise total number of Industries**
4. **State-wise Number of motor vehicle registrations.**
5. **Coal Production Data**
6. **Industrial Coal Consumption**
7. **Coordinates of state boundaries**
8. **Neighbouring states of each state in India.**

Chapter 3

Data Preprocessing

The importance of data is now evident by the phrase**, "Data is the new oil",** initially coined by the British data science entrepreneur Clive Humby. Even though it is crucial, its genuine worth can be acknowledged just when it is refined.
This chapter directs the cycle from collecting, cleaning, integrating and transforming the data into a suitable format.

3.1 Data Collection

1. **Pollutant concentrations across India:**
    a. Air Quality data for the period 1990-2015 for all stations in India has been obtained from kaggle, available at https://www.kaggle.com/shrutibhargava94/india-air-quality-data
    b. Air Quality data for 2015-2020 for all stations in India has been obtained from kaggle, available at https://www.kaggle.com/rohanrao/air-quality-data-in-india.

    Description: These are the unprocessed raw datasets obtained by web scraping the official CPCB website https://app.cpcbccr.com/AQI_India/.
    These kaggle files are CSV files containing concentrations of various pollutants like SO2, NO2, PM2.5 etc., which were recorded at different stations across the country.

2. **Population Enumeration Data**
    The population data of India has been taken from the decadal census 2001 and census 2011 websites.
    a. The 2001 census data is available at https://censusindia.gov.in/Census_Data_2001/Census_data_finder/A_Series/Total_population.htm webpage
    b. 2011 census data is available at http://censusindia.gov.in/pca/DDW_PCA0000_2011_Indiastatedist.xlsx
    The state-wise population is during 2001 and 2011 are recorded in the above datasets.

3. **State-wise total number of Industries**
    Dataset for the total number of industries in each state has been obtained from:
    a. 2001-2006 data is sourced from labourbureau.gov.in available at http://www.labourbureau.gov.in/ASI_V2_2005_06_TAB27F.docx
    b. 2007-08 data is obtained from labourbureaunew.gov.in available at http://www.labourbureaunew.gov.in/UserContent/ASI_Vol_1_2007_08.pdf
    c. 2009-2015 is taken from http://mospi.nic.in/sites/default/files/statistical_year_book_india_2015/Table%2014.1_2.xlsx

4. **State-wise Number of motor vehicle registrations**
   Dataset for the state-wise annual motor vehicle registrations is taken from the MOSPI official website.
   http://mospi.nic.in/sites/default/files/statistical_year_book_india_2015/Table-20.4_1.xlsx for the period 2001-2016
5. **Coal Production Data**
   Annual coal production data for 2001-2016 is collected from the official Statistical data provided by MOSPI
   http://mospi.nic.in/sites/default/files/statistical_year_book_india_2015/Table%2016.3_1.xlsx.
6. **Industrial Coal Consumption**
   Annual Coal consumption of Industries is provided at
   http://mospi.nic.in/sites/default/files/statistical_year_book_india_2015/Table%2016.4_1.xlsx.
7. **Coordinates of state boundaries**
   Coordinates of a State including State border and Capital city is provided at
   https://github.com/geohacker

3.2 Data Cleaning

The data gathered from various sources comprises discrepancies such as missing, non-numeric, incorrectly formatted values. Such data has to be cleaned to remove those discrepancies.
Actions taken in the specific datasets to clean the data are as follows:

- **Pollutant Concentration Data:**
  a. The information about the dataset for **1990-2015** is shown in the figure. Those columns which are considered in the analysis are considered and of the total 435742 entries, we can see the non-null values of various attributes. 'stn_code' is the Station from which the data is collected, the corresponding city name is given by the 'location'.
  'so2', 'no2', 'rspm' are the values measured by the station at the date given by the 'data' column.

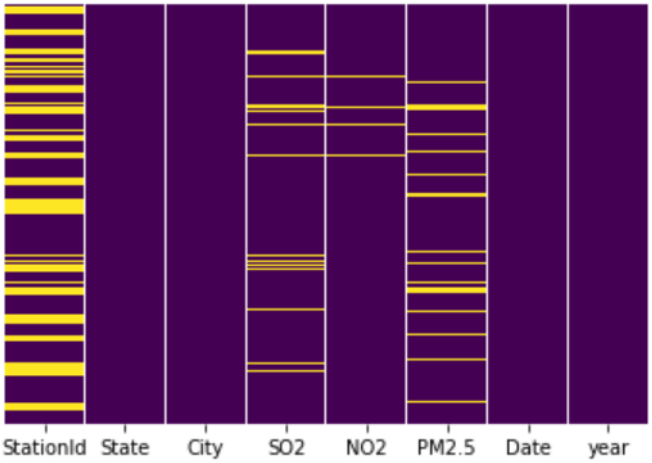
```
df_2011.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435742 entries, 0 to 435741
Data columns (total 7 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   stn_code  291665 non-null  object
 1   state     435742 non-null  object
 2   location  435739 non-null  object
 3   so2       401096 non-null  float64
 4   no2       419509 non-null  float64
 5   rspm      395520 non-null  float64
 6   date      435735 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(3), object(3)
memory usage: 23.3+ MB
```

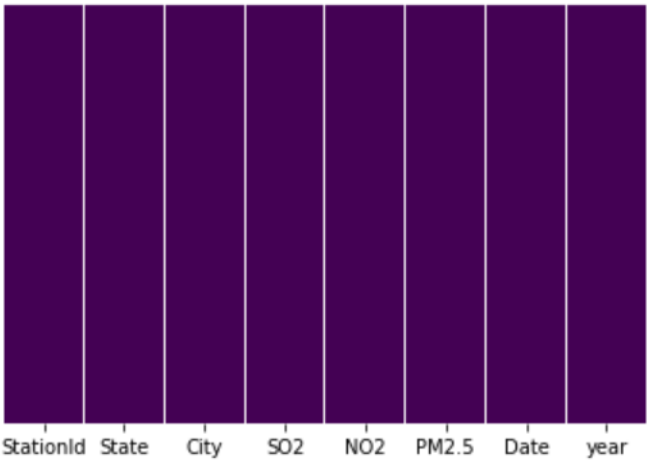
➢ Tuples containing the null values in the date column are dropped,, and those with non-null values in the date column are considered. A new column with the title ‘year’ is created using the ‘date’ column. Column names are renamed as shown in the figure.

```
#take data values not null
df_2011=df_2011[~df_2011["date"].isnull()]

df_2011["year"]=df_2011.date.dt.year

#rename columns
df_2011.columns=['StationId','State','City','SO2','NO2',"PM2.5",'Date','year']
df_2011
```

➢ Null values present in different columns of the data file are displayed by the horizontal yellow bars of the heatmap provided by the seaborn library. The heatmap shows that there are Null values present in SO2, NO2, PM2.5 and also in the StationId column.



➢ **Interpolation** is applied by grouping the station id to find the missing pollutant concentration values for each station. The index method is chosen for interpolation because the values at each station more or less remain close to the value at the nearest index. Stations with only one or two entries in the data are removed to facilitate interpolation. Next, the data is grouped according to the City name and interpolated by using the index method to get the approximate values for the missing data. Thus, the obtained heatmap shows no null values present in the data.



➢ Annual concentrations of pollutants at each station is calculated by grouping the data on StationId, City and by taking the mean of corresponding columns.

➢ Data of each station from the year 2011 is explicitly extracted and extrapolated for the missing years in 2011-2015, using the scipy library. The previous interpolation covers only the missing values in the data frame. This extrapolation provides the values for the missing data outside the range of the data frame. Later, the city-wise annual mean is taken to get the annual concentrations of pollutants for respective cities.

```
df_2011_final = df_2011_final.groupby(["State",'City',"year"]).mean()
df_2011_final
```

State	City	year	SO2	NO2	PM2.5
Andhra Pradesh	Ananthapur	2011	4.000000	12.564815	70.009259
		2012	4.000000	12.564815	70.009259
		2013	4.000000	12.564815	70.009259
		2014	4.787037	10.314815	75.990741
		2015	5.166667	10.685185	87.907407
...	...	...	...	...	...
West Bengal	South Suburban	2012	118.959360	8.278941	59.496305
		2013	8.514492	59.551415	179.981492
		2014	4.213592	37.762136	96.373786
		2015	3.461632	36.880695	90.215366

➢ Data of Delhi and Noida is merged to represent as they make a larger capital territory region.

➢ State name for cities within Telangana but in erstwhile Andhra Pradesh was changed to Telangana using the 2015 data.

➢ Bangalore is replaced with its new name Bengaluru and in other states, city names were replaced with their new or modified names.

b. For the pollutant concentration dataset from **2015-2020**, similar procedures employed for the previous data are applied to clean the data. As this dataset contains only StationId and not state or city names, another dataset(stations.csv) containing the mapping for StationId and city, the state name is merged to map the data with city and state names. Interpolation for the missing data using the Index method is done and extrapolation for the missing data in the range 2015-2020.

- **Population Enumeration Data:** Census data for 2021 is unavailable as the decadal census could not take place due to the Covid-19 pandemic. The data is available only for the 2001 and 2011 years. For the yearly data, the growth rate between 2001 to 2011 is calculated. It is then used for extrapolating the data for the years after 2011. For 2001, data was scraped from the corresponding webpage using python's BeautifulSoup library. The census 2011 data is directly fetched from the web link provided. Using the data obtained from the 2001 and 2011 census of India, the population for the years after 2011 has been estimated for each state. To achieve this, we used a constant growth rate 'R' per year given by $(1 + R)^{duration} = Population\ in\ 2011 \div Population\ in\ 2001$. Using this rate of growth, the population for any year can be estimated using $Population(year) = (1 + R)^{year-2001} \times Population(2001)$.

- **State-wise total number of Industries:** The official website of MOSPI provides information about Industries. This dataset contains the various columns ranging from state-wise distribution of factories, fixed capital, working capital to productivity information from 2008 to 2015. Telangana and Lakshadweep were removed as there is a lack of information. The information about the number of factories in a state from this dataset has been used. For some states like Mizoram, Arunachal Pradesh, Sikkim, there is a discrepancy in the data. Names of states like Uttarakhand, Andaman and Nicobar Islands, Dadra and Nagar Haveli have been corrected.

- **State-wise Number of motor vehicle registrations**: Motor Vehicles Registration data is collected from the official website of MOSPI. This data set has a state-wise total number of registrations from 2001 to 2016. Although there are no missing values in the dataset, many of them are non-numeric, with special characters appended. The names of Odisha, Chhattisgarh, Andaman and Nicobar Islands, Dadra and Nagar Haveli have been corrected.

- **Coal Production Data:** Production of coal, coal derivatives & by-products is a dataset available on the MOSPI website. It contains the production of cooking,non-cooking, lignite, and coal derivatives like hard coal, washed coal, and Total coal produced annually from 2001-2016. For this project, only the year and the total coal produced columns were considered.

- **Industrial Coal Consumption:** Data for raw coal consumption by different industries is taken from the MOSPI website. It contains consumption data for industries like Electricity, Steel, Paper, Cement, Textile, Iron, Fertilizers, Brick and Total coal consumption by all those industries annually for 2001-2016. The year and the total consumption were considered for this project.

3.3 Data Integration

- ➢ The cleaned data of the pollutant concentrations for 2011-2015 and 2015-2020 is merged. The mean of the data is calculated by grouping the city and year to remove any overlapping years. Now, this data is extrapolated to fill the city-wise missing years' data. State-wise mean values are calculated by grouping the state and year. Cities common to both the datasets (2011-2015 and 2015-2020) were only considered for the city-wise pollution analysis. There are 21 common cities between the two datasets.
- ➢ Year Wise population is calculated using the growth rate calculated earlier and stored in a CSV file.

```
#extrapolated population of India in 2020
df_pop_test["2020"].sum()
```

```
1399441209
```

- ➢ State-wise total number of industries data is extrapolated for the years 2016-2020.
- ➢ State-wise total number of motor vehicles data is extrapolated for the years 2017-2020.
- ➢ Coal Production and Industrial Coal Consumption data are extrapolated for 2017-2020.

Chapter 4

Decadal analysis of Pollution

In this chapter, using chi-score, Z-score, states which are outliers, hotspots and coldspots are found. Later, analysis about different pollutant levels across various states and cities across the country. Cities and states where pollutant levels have rapidly increased over the decade are found using plots and heatmaps.

4.1 Chi-Squared for finding the outliers

Our main task is to find whether a state is an outlier, i.e. either Hotspot or Coldspot. To determine this, we have used a modified version of the Chi-Square Test to detect Outliers. The formula gives it

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - e_i)^2}{e_i}$$

Where x is the State to be tested, here x_i represents the i^{th} measurement. And $\mathbf{e_i}$ is the expected value of the i^{th} measurement of all States. The state with Chi-value more significant than a threshold value is considered as an outlier.

4.1.1 Methodology

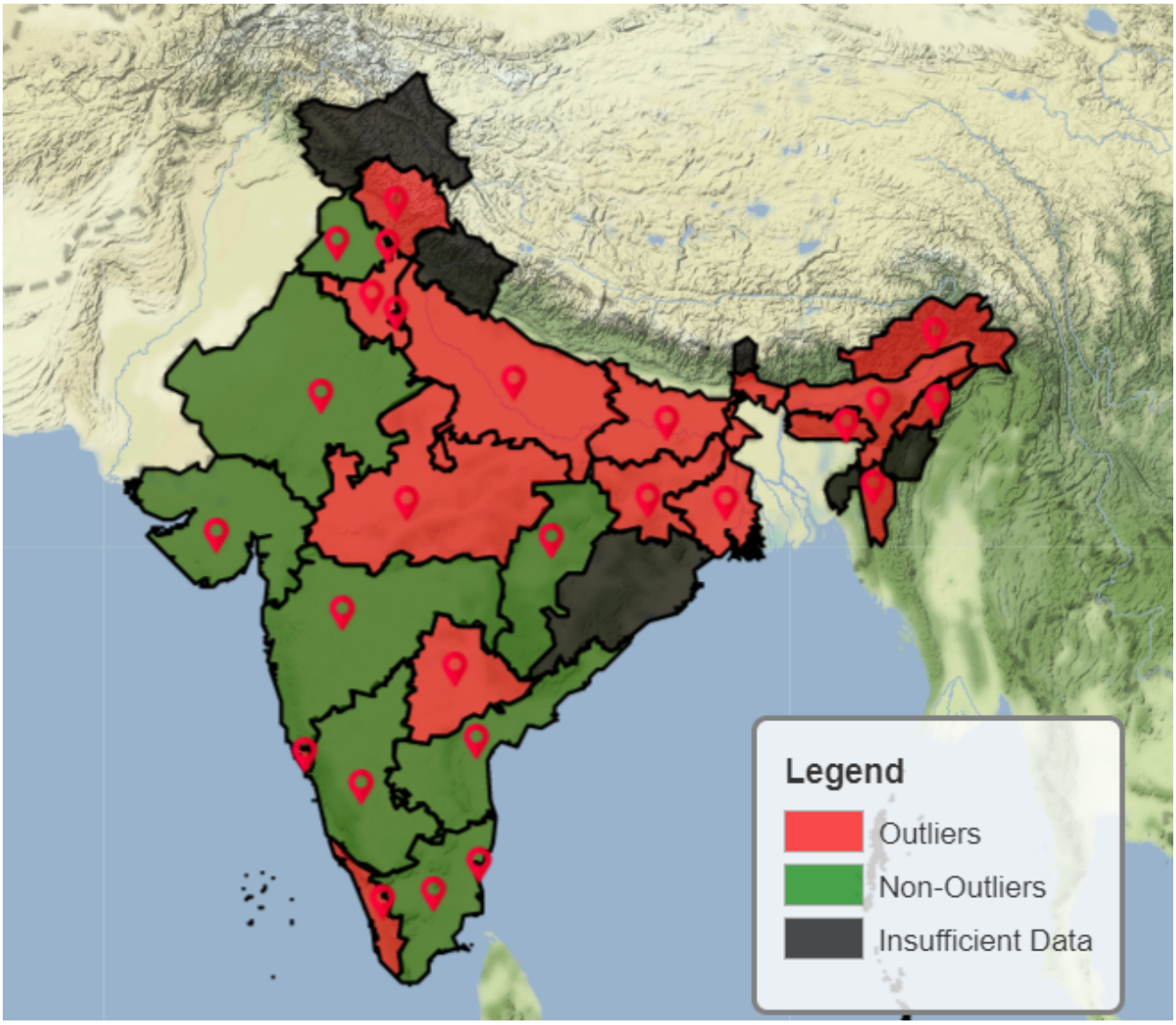
For this test, we have set a significant value as 1% (or 0.01). From the above formula, we have calculated the Chi-square value for all States. Then from the Chi-Square distribution, we got the p-value corresponding to a State. And a State is considered as an outlier if the p-value is less than the significant value. The Interactive map with the Chi-Square test is available in the HTML folder of our project.

4.1.2 Observation

The Map of India shown in the Figure shows that Himachal Pradesh, Haryana, Uttar Pradesh, Madhya Pradesh, Bihar, Jharkhand, West Bengal, Assam, Meghalaya, Mizoram, Nagaland, Arunachal Pradesh, Telangana, Goa, Kerala are outliers. But using the Chi-Square test for outlier detection, we can't determine whether the outlier state is Hotspot or Coldspot.

A lot of the Northern part fall under this category. States like Uttar Pradesh, Haryana, Madhya Pradesh may be hotspots because they are highly polluted States of India, and the average temperature of the States are on higher sides. On the other hand, states like Himachal Pradesh and Northeastern states like Assam, Meghalaya, Arunachal Pradesh, etc.,

can be coldspots as they are the Himalayas and Mountain regions. The states like West Bengal and Kerala might be an outlier as recently they have been hit by multiple natural calamities like Cyclone, Flood, etc. To get an accurate picture of the Hotspot and Coldspot, we will test with a Z-score.



4.2 Z-score test for finding Hotspot and Coldspot

To overcome the limitation of the Chi-Square Test, we have used the Z-score test. Z-score will help us determine whether a State is a Hotspot/Coldspot/Neural spot.

4.2.1 Methodology

For this process, we need to find the mean pollutant concentration of each State from 2011 to 2020. To calculate this, we have used the following formula

Mean Pollutant Concentration (MPC)= $\frac{(SO_2+NO_2+PM)}{3}$

Any state is considered Hotspot (respectively Coldspot) if the MPC of a State is greater than the Mean MPC and Half the Standard Deviation of MPC of all its neighbouring States. The result is available on an interactive map which is available in the HTML folder of our project .

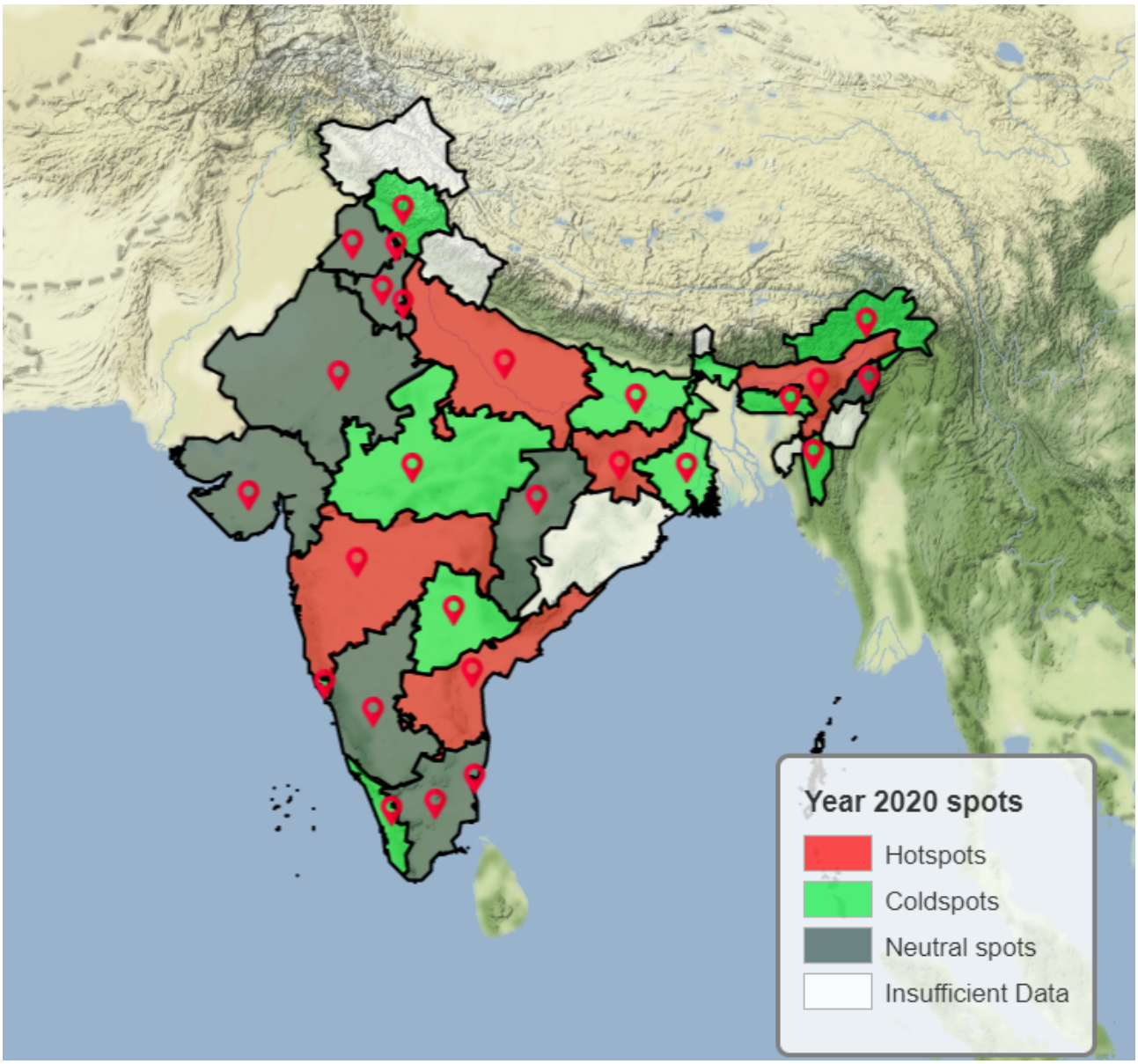
4.2.2 Observation

The Map of India shown in the Figure shows that Uttar Pradesh, Maharashtra, Jharkhand, Assam, Andhra Pradesh are the Hotspots. A possible reason could be the large population and the ongoing Industrial works which might increase the temperature. Whereas in Assam, it might happen that due to a low-pressure system over the Bay of Bengal, there is less moisture which results in dry weather.

Also, States like Himachal Pradesh, Bihar, West Bengal, Arunachal Pradesh, Meghalaya, Mizoram, Nagaland, Madhya Pradesh, Telangana, Goa and Kerala are some of the cold spots. States having mountains and Himalayas are known to be cold spots, whereas, in Goa, Kerala, the overall temperature becomes low due to heavy rainfall.

Other states like Rajasthan, Gujrat, Karnataka, Tamil Nadu, etc., are in neutral spots as the average temperature of these States remains almost constant because they are hotter in the day and colder at night.

Again we also have some States like Odisha, Sikkim, etc., where we don’t have enough pieces of evidence and data to support our hypothesis.


Year 2020 spots
Hotspots
Coldspots
Neutral spots
Insufficient Data

4.3 Correlation

The relationship between the observed values of two variables is referred to as correlation. A positive association is defined as: “an increase in the value of one variable improves the value of another variable”, whereas a negative association is defined as: “an increase in the value of one variable decreases the value of another variable”.

Variables can also not affect one another, resulting in a correlation that is neither positive nor negative. The statistical adage “correlation does not imply causation” holds, and we must use caution while analyzing the correlation coefficient to avoid drawing incorrect conclusions.

4.3.1 Input data

Data on air pollutant concentrations, number of industries, motor vehicles, and population of states were utilized for correlation from 2011 to 2020.

4.3.2 Methodology

Correlation is usually expressed as the association coefficient, which ranges from -1.0 to 1.0 and indicates the degree of correlation. A correlation coefficient of 1.0 means the perfect positive connection between the variables, while a value of -1.0 indicates a perfect negative correlation. There is no association between the variables if the correlation coefficient is zero. The book "Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach" goes on to explain the correlation coefficient, stating that values of +/-0.1, +/-0.3, and +/-0.5 indicate weak, moderate, and muscular association strength, respectively. We discovered two different correlation coefficients in this project: Pearson Correlation Coefficient and Spearman Correlation Coefficient.

Pearson Correlation Calculator The coefficient is a metric that evaluates the linear correlation between two variables with Gaussian distributions. Pearson correlation has the advantage of determining solely the linear relationship between variables x and y, i.e., a proportionate change in x should result in a corresponding change in y. On the other hand, Spearman Correlation Coefficient produces a monotonous (and not always necessary) result.

The relationship between variables is almost always linear. A monotonic connection is one in which, as time passes, the relationship becomes monotonous. When the value of one variable rises, the value of the other variable either falls or stays the same.

Rising or continues to decrease, but not necessarily at a fixed rate While Pearson's correlation is a parametric measure of correlation that calculates the covariance between two variables normalized by the variance of both variables, Spearman's correlation is a non-parametric rank correlation that computes the correlation using rank values rather than fundamental values.

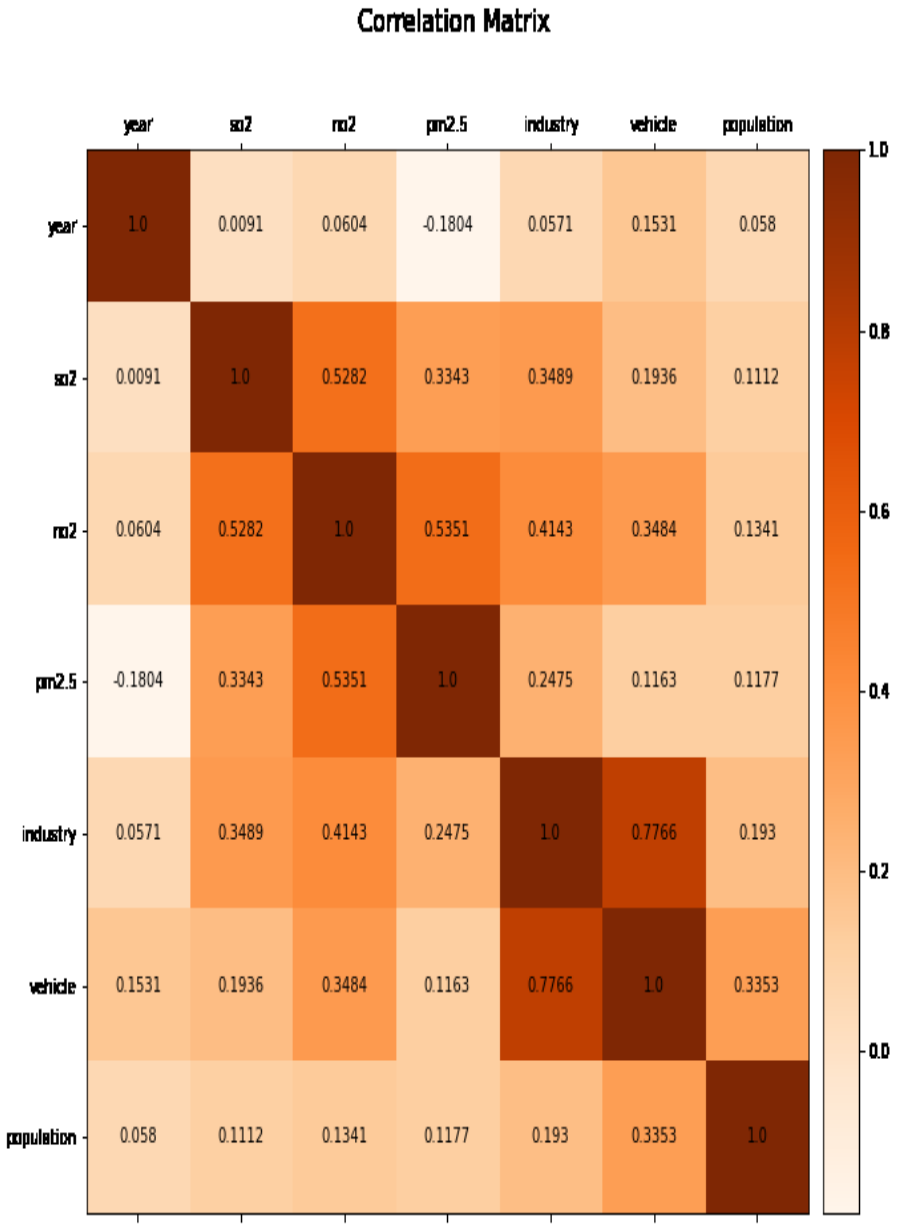
The Pearson Correlation Coefficient and the Spearman Correlation Coefficient must first be determined. We accomplish this by utilizing the scipy library's functions 'Pearson' and spearman.' Both of these functions yield several calculations, from which we extracted the correlation coefficient and the p-value for our analysis. The p-value generally represents the likelihood of witnessing data created by an uncorrelated system with equal or more correlation as obtained, and the correlation coefficient is the same coefficient we mentioned earlier. A p-value of 0 indicates that the probability of observing the data given the uncorrelated samples is exceptionally implausible, implying that the models are linked. (Reliable p-value 500)

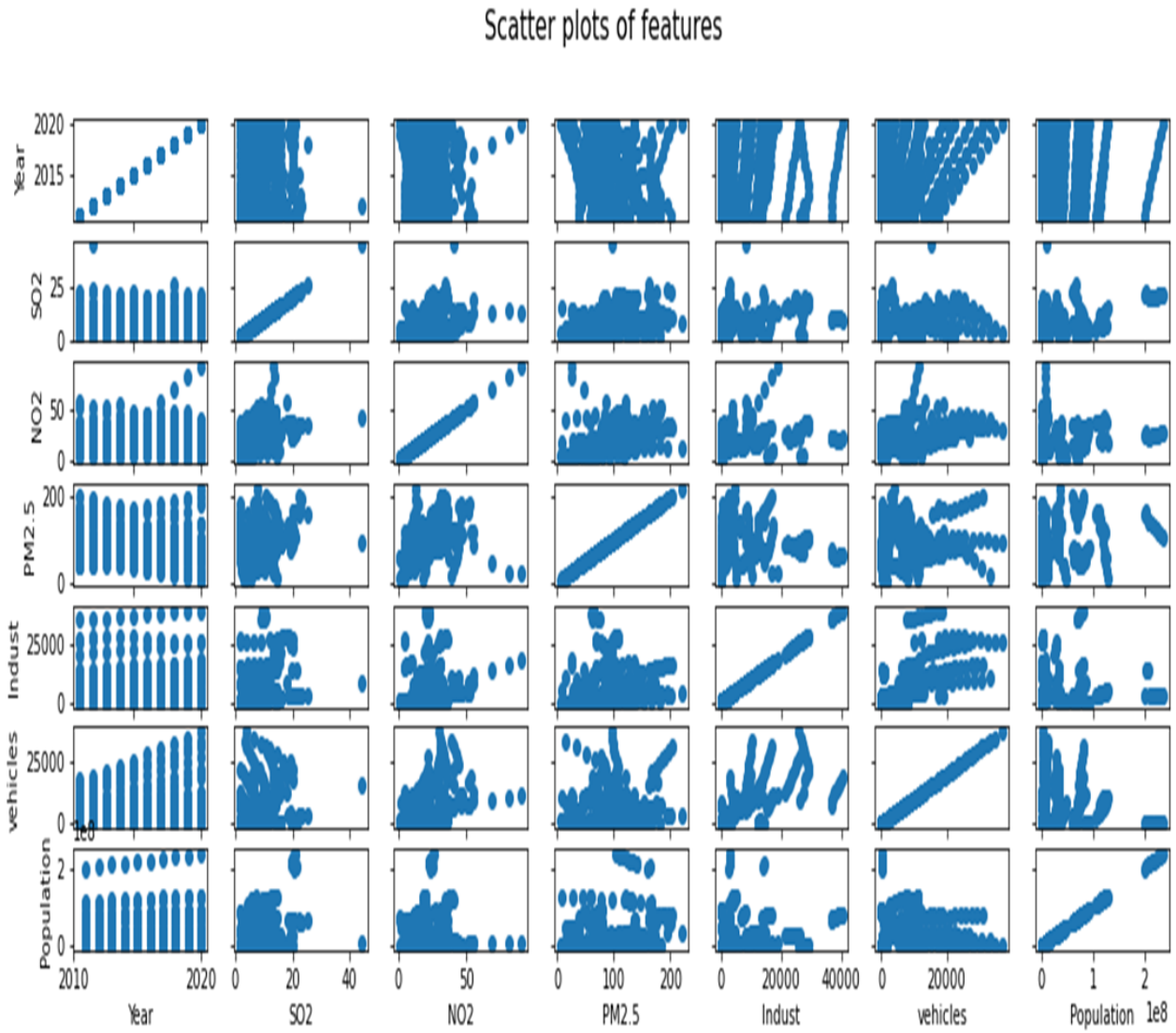
The Pearson Correlation Coefficient and the Spearman Correlation Coefficient tables show that the latter appears to be more dependable. Some of the explanations are theoretical, such as the fact that Spearman can capture a broader range of relationships than Pearson. It is also monotonic if a linear relationship exists; hence, the Spearman correlation coefficient can be trusted. The opposite, however, is not valid. In the event of monotonic non-linear correlations, the Pearson Correlation Coefficient cannot be determined. By looking at the scatterplots, this may be further justified. Except for the plots in which there appears to be no relationship between the variables, the rest aren't merely linear; some seem to have a monotonic non-linear relationship as well.

Another rationale for using Spearman instead of Pearson for analysis is that Pearson's correlation values for Industry and Vehicle with SO2=NO2=PM2.5 appear to be lower. It falls into the weak correlation region. However, based on our experience, we can confidently assume that the relationship between the number of vehicles and SO2 is moderate, precisely what the Spearman Correlation Coefficient indicates.

Furthermore, the p-values for such relationships are more significant for Pearson than for Spearman, implying that the Spearman Correlation Coecient's value is more likely to be correct. For any further study, the Spearman Correlation Coefficient is utilized. It's worth noting that if a scatterplot suggests that a relationship isn't linear or monotonic, neither of the two Correlation Coefficients for that pair of features/variables should be examined.

4.3.3 Observation





It denotes that a component will always be highly associated with itself and provides little information. We detect a substantial connection (values greater than or close to 0.5) between PM2.5 and NO2 levels (value 0.53), SO2 and NO2 levels (value 0.52), NO2 and industry (0.41), and NO2 and cars (0.41). (value 0.46). With a value of 0.7, the correlation between the number of vehicles and the number of industries is powerful. This can be attributed to areas with more substantial industries, which necessitate more automobiles to move goods made or used by manufacturers.

There is also a moderate association (values larger than 0.1) between vehicle and PM2.5 (value 0.33), industry and SO2 (value 0.34), and PM2.5 and Industry (value 0.35). (value 0.30).There may be subtle interdependencies at work here, and all of these correlations may not imply causation. For example, the significant association between NO2 and SO2 is most likely due to the absence of some feature from the feature set. We would have noticed a more significant association between SO2 or NO2 and that parameter if the responsible quality had been included in our feature set.

4.4 Clustering

We grouped the states based on SO2, NO2, and PM2.5 levels using Air Pollutant Concentration data from 2014.

4.4.1 Methodology

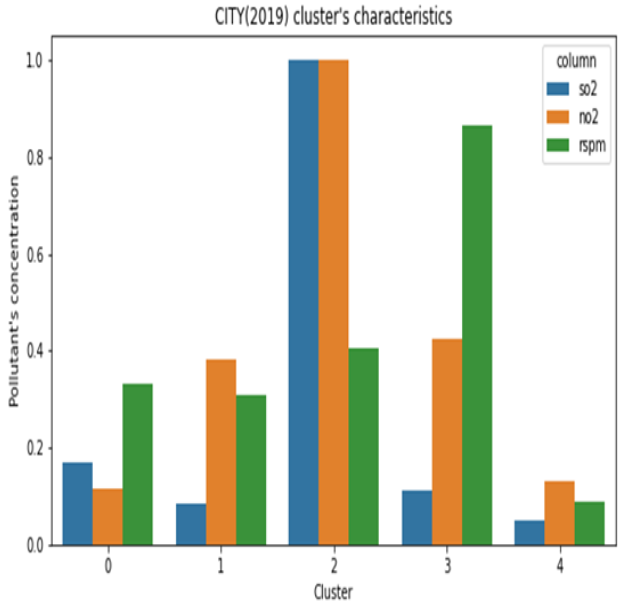
Using clustering, we grouped states with comparable pollution concentrations in 2014. We employed the K-means technique, which uses Euclidean distance to group data with identical features together. The Euclidean distance is a measurement of the distance between two points or vectors in a coordinate system. Pythagoras' theorem defines a two-dimensional or multidimensional (Euclidean) space.

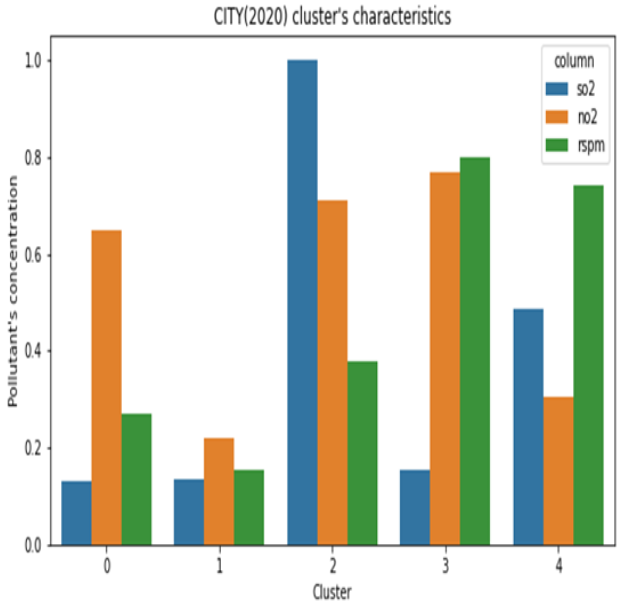
The square root of the sum of the squared pairwise distances is used to determine the length. Every dimension has wise distances. In n-dimensional space, The formula for the Euclidean distance is,

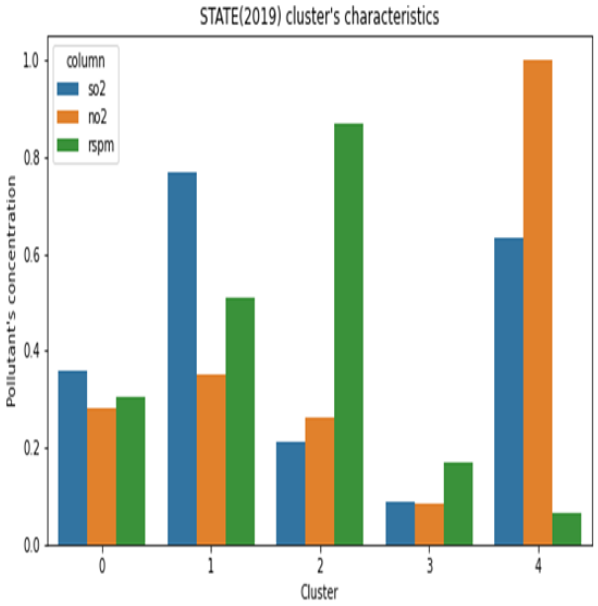
$$\sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

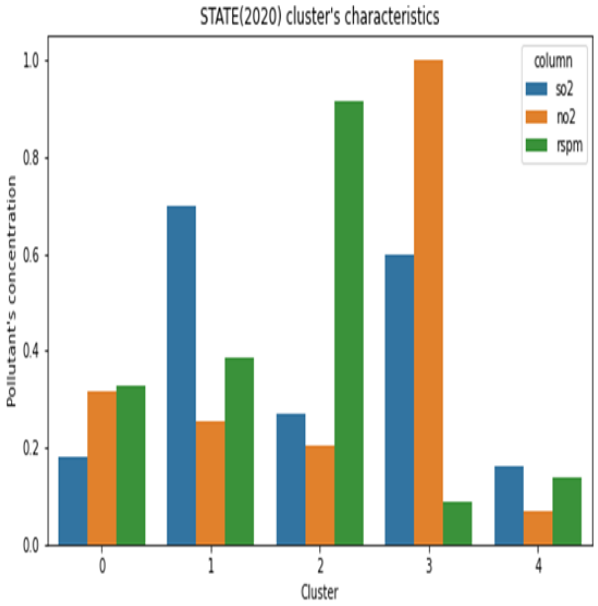
In our analysis, the number of dimensions is 3: SO2, NO2, and PM2.5 concentration.

4.4.2 Observation





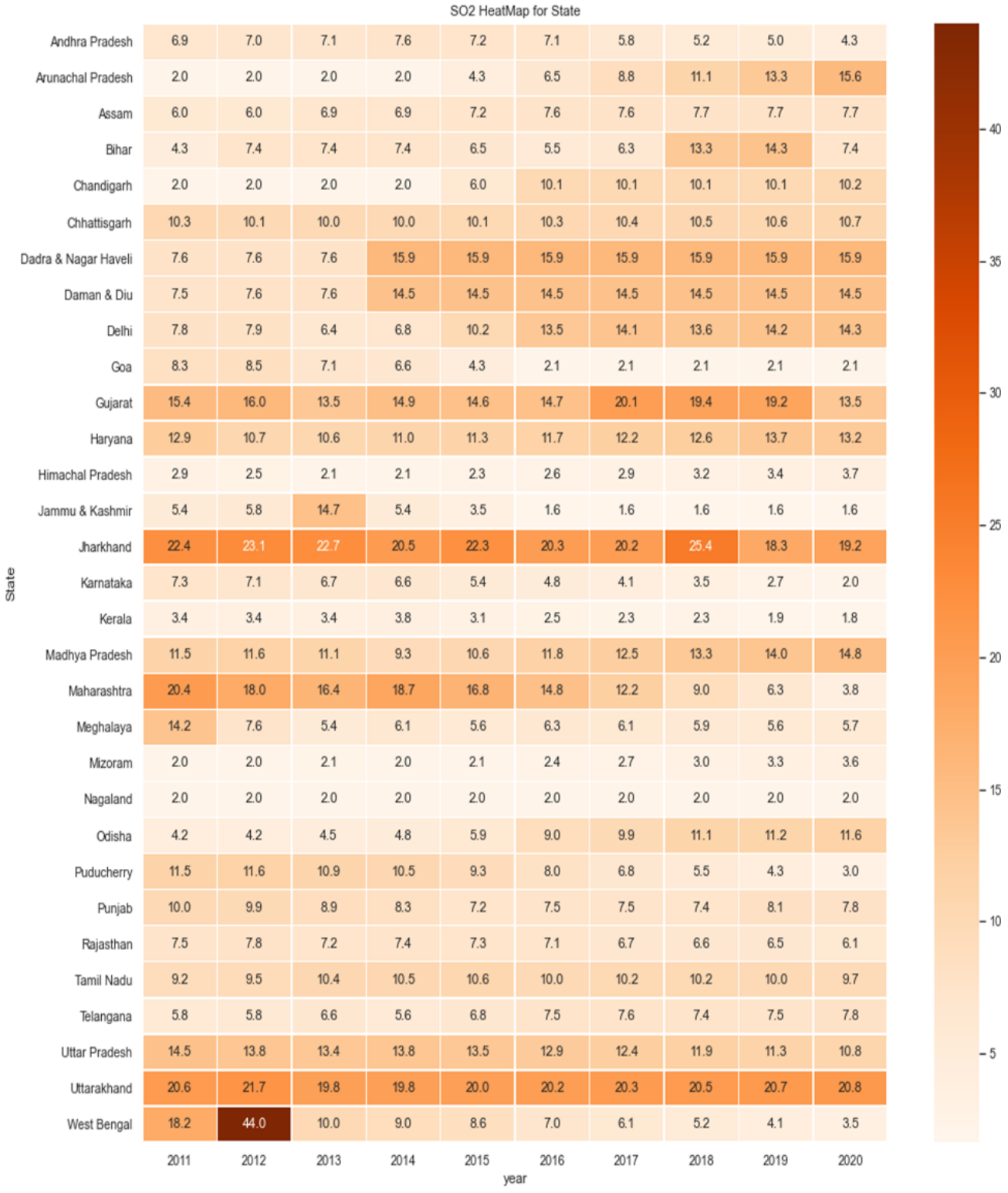




Cluster	States-2019	States-2020	Cities-2019	Cities-2020
0	Andhra Pradesh Chandigarh, Chhattisgarh Madhya Pradesh Maharashtra, Odisha Punjab, Rajasthan Tamil Nadu, Telangana West Bengal	Andhra Pradesh, Goa Himachal Pradesh Karnataka, Kerala Maharashtra, Meghalaya, Mizoram Puducherry, Rajasthan West Bengal	Visakhapatnam Bhopal, Jaipur Hyderabad, Kolkata	Chandigarh, Kochi ThiruvananthapuramShillong, Aizawl Amritsar, Chennai
1	Arunachal Pradesh Bihar Dadra & Nagar Haveli Daman & Diu Delhi Gujarat Jharkhand Uttarakhand	Arunachal Pradesh,Dadra & NagarHaveli Daman & Diu,Delhi| Gujarat, Jharkhand Uttarakhand	Bengaluru Kochi Thiruvananthapuram, Mumbai, Shillong Aizawl Chennai Coimbatore	Guwahati, Talcher
2	Haryana	Bihar, Chandigarh, Chhattisgarh, Madhya Pradesh, Odisha, Punjab, Tamil Nadu, Telangana	Ahmedabad	Visakhapatnam Bengaluru, Bhopal Mumbai, Jaipur Coimbatore Hyderabad, Kolkata
3	Assam, Jammu & Kashmir, Nagaland, Uttar Pradesh	Haryana	Guwahati, Chandigarh, Talcher, Amritsar	Ahmedabad
4	Goa, Himachal Pradesh, Karnataka Kerala Meghalaya Mizoram, Puducherry	Assam, Jammu & Kashmir, Nagaland, Uttar Pradesh	Patna, Delhi, Lucknow	Patna, Delhi, Lucknow

4.5 HeatMaps and Bar Plots

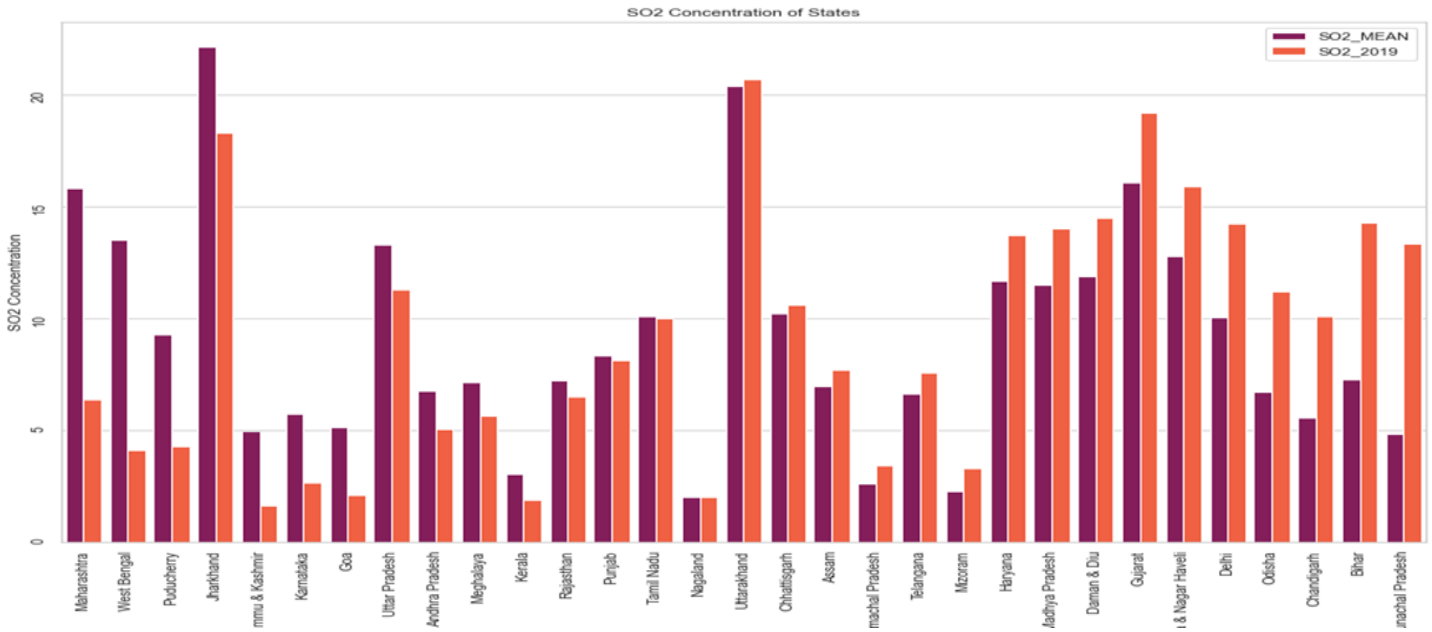
4.5.1 State and City wise SO2 concentration




SO2 HeatMap for City
Ahmedabad 13.3 12.1 12.6 13.4 22.0 22.0 70.6 70.2 73.7 33.7
Aizawl 2.0 2.0 2.1 2.1 2.0 3.1 4.2 5.2 6.3 7.4
Amritsar 14.3 14.6 12.8 13.4 8.8 6.4 6.4 5.2 13.1 10.3
Bengaluru 15.6 14.0 13.4 12.6 6.4 6.0 5.8 8.6 5.4 6.7
Bhopal 3.4 2.8 2.5 2.1 7.9 12.5 12.5 12.5 12.5 13.4
Chandigarh 2.0 2.0 2.0 2.0 6.0 10.1 10.1 10.1 10.1 10.2
Chennai 13.1 12.3 12.2 12.6 12.0 5.3 7.9 9.0 8.3 8.8
Coimbatore 4.3 3.5 4.2 4.6 6.8 9.5 9.5 9.5 9.5 7.5
Delhi 7.8 7.9 6.4 6.8 10.2 13.5 14.1 13.6 14.2 14.3
Guwahati 6.1 6.1 7.0 6.8 10.9 14.7 14.7 14.7 14.7 14.6
Hyderabad 4.9 4.9 5.2 5.0 9.5 14.9 12.7 8.9 6.9 6.6
Jaipur 6.8 9.1 7.5 7.1 8.2 9.3 9.3 10.9 12.1 12.1
Kochi 3.3 3.2 2.3 2.4 2.5 5.5 8.5 11.5 14.6 17.6
Kolkata 12.9 37.4 10.5 8.3 8.0 9.0 9.0 9.0 7.9 10.6
Lucknow 8.1 8.0 7.5 7.7 12.4 11.4 11.1 10.8 8.6 8.8
Mumbai 5.4 4.9 3.0 4.0 16.8 29.4 28.5 13.4 11.5 10.8
Patna 4.3 7.4 7.4 7.4 6.5 5.5 6.3 13.3 14.3 7.4
Shillong 2.9 2.3 2.4 2.0 3.8 5.5 5.5 5.5 5.5 7.3
Talcher 7.8 7.8 8.4 9.3 24.2 38.3 38.3 33.1 27.2 24.1
Thiruvananthapuram 10.3 9.1 9.4 9.1 7.0 3.8 3.8 7.0 4.7 5.8
Visakhapatnam 12.8 12.8 13.3 13.1 15.2 21.6 9.5 11.1 12.9 8.2
2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
year
70
60
50
40
30
20
10

From the heatmap of states, from 2011-2020, Uttarakhand, Maharashtra, and Jharkhand paid the most attention to SO2. For Meghalaya, we can see a lot of bounce within the SO2 attention in 2013, 2018, and 2011, and a lot of attention in 2011. In the last ten years, Himachal Pradesh, Manipur, Mizoram, and Nagaland have paid substantially less attention to SO2. For ten years, Himachal Pradesh, Manipur, Mizoram, and Nagaland have displayed much less attention for SO2. Himachal Pradesh, Manipur, Mizoram, and Nagaland have paid far less attention to SO2 over the last ten years.

In the same way from the heatmap of cities, we can say Ahmedabad has the highest SO2 concentration in the year 2019 Talcher of Odisha has higher in 2016. In contrast, Shillong of Meghalaya and Aizawl have the most negligible SO2 concentration.



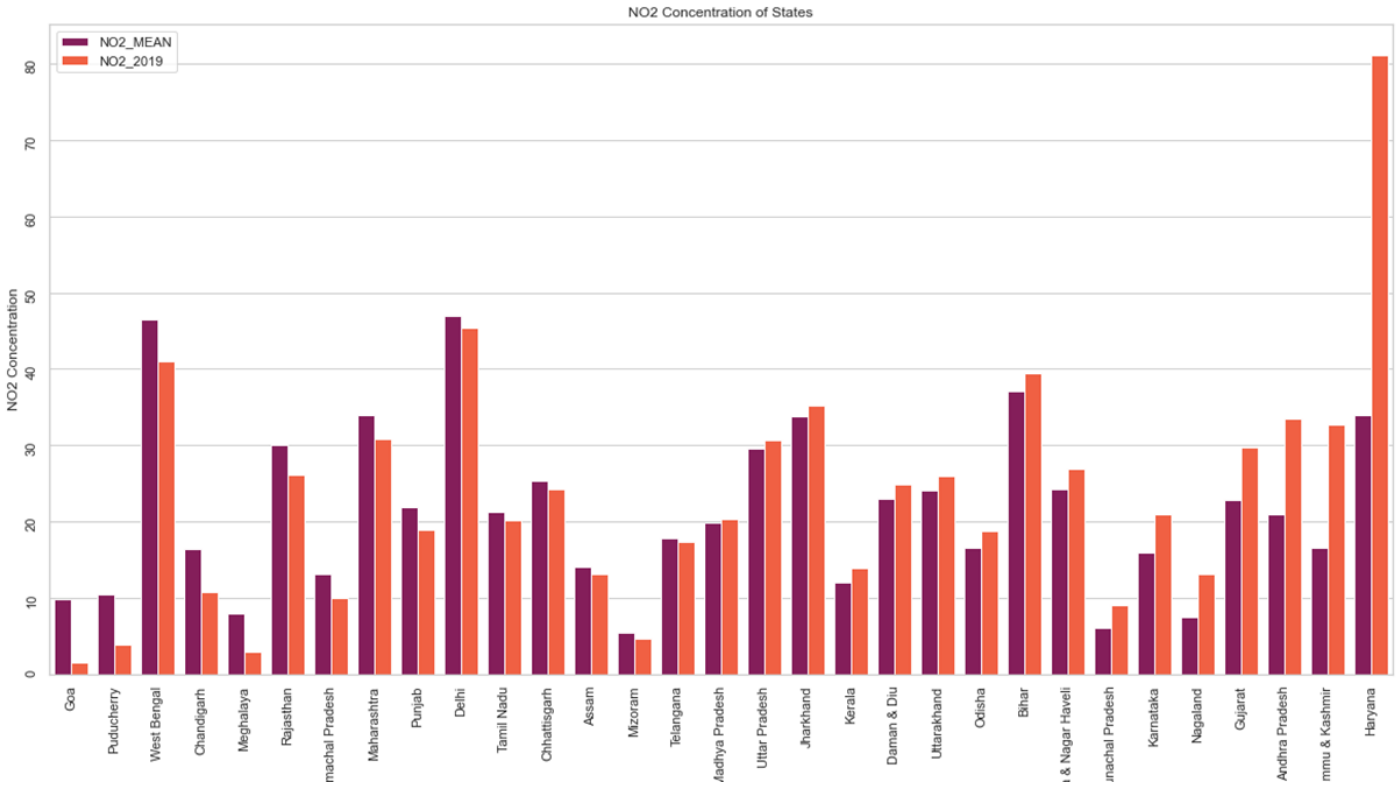
As shown in the above barplot, Jharkhand has the best SO2 attention for 2019, while Nagaland has the least SO2 attention for 2019. We can see that the degrees of SO2 attention for Jharkhand and Uttarakhand are nearly identical, putting Uttarakhand in second place. Nagaland, Mizoram, and Kerala have SO2 levels that are almost identical.

4.5.2 State and City wise NO2 concentration


NO2 HeatMap for City
City
year
Ahmedabad 25.2 23.8 17.7 19.7 21.8 22.1 32.1 85.0 90.8 31.1
Aizawl 5.2 6.1 9.3 7.6 8.4 6.8 5.2 3.6 2.0 0.4
Amritsar 36.3 37.5 39.6 40.4 25.4 21.0 21.0 22.4 15.8 14.5
Bengaluru 29.5 28.5 27.7 29.6 26.8 32.4 33.1 32.6 28.8 22.2
Bhopal 16.5 20.9 27.1 21.2 33.7 44.7 44.7 44.7 44.7 22.6
Chandigarh 15.9 19.9 23.2 21.3 18.1 10.9 10.9 10.9 10.9 12.3
Chennai 22.1 22.9 21.3 22.1 18.6 14.9 17.4 17.0 16.0 11.3
Coimbatore 32.9 27.1 24.5 25.4 20.2 15.5 15.5 15.5 15.5 43.7
Delhi 50.9 52.3 46.0 44.6 45.8 43.3 46.5 46.7 45.4 33.7
Guwahati 14.1 14.1 15.7 13.8 13.7 13.1 13.1 13.1 13.1 14.4
Hyderabad 28.6 27.7 22.3 23.6 23.5 24.0 23.6 37.5 29.6 27.7
Jaipur 36.7 51.7 40.3 42.7 37.0 38.0 38.0 34.6 33.6 28.4
Kochi 12.8 9.9 6.1 8.6 11.4 12.0 12.6 13.3 13.9 14.5
Kolkata 65.8 56.6 67.6 66.1 46.7 37.9 37.9 37.9 34.5 29.1
Lucknow 32.9 31.7 29.4 28.3 31.5 43.1 43.4 37.9 34.8 29.8
Mumbai 34.5 20.8 13.1 19.6 21.2 20.2 20.2 21.7 21.9 22.7
Patna 36.5 36.6 36.6 36.6 35.8 34.8 39.8 40.4 39.5 37.6
Shillong 10.1 10.7 11.0 10.7 8.0 2.9 2.9 2.9 2.9 2.7
Talcher 18.9 18.9 21.4 23.7 12.3 0.1 0.1 16.9 10.1 12.7
Thiruvananthapuram 23.2 21.5 22.4 22.5 16.2 6.9 6.9 10.1 7.0 14.1
Visakhapatnam 13.9 13.9 18.0 20.2 30.7 42.6 19.2 38.7 37.4 30.8
2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
80
60
40
20

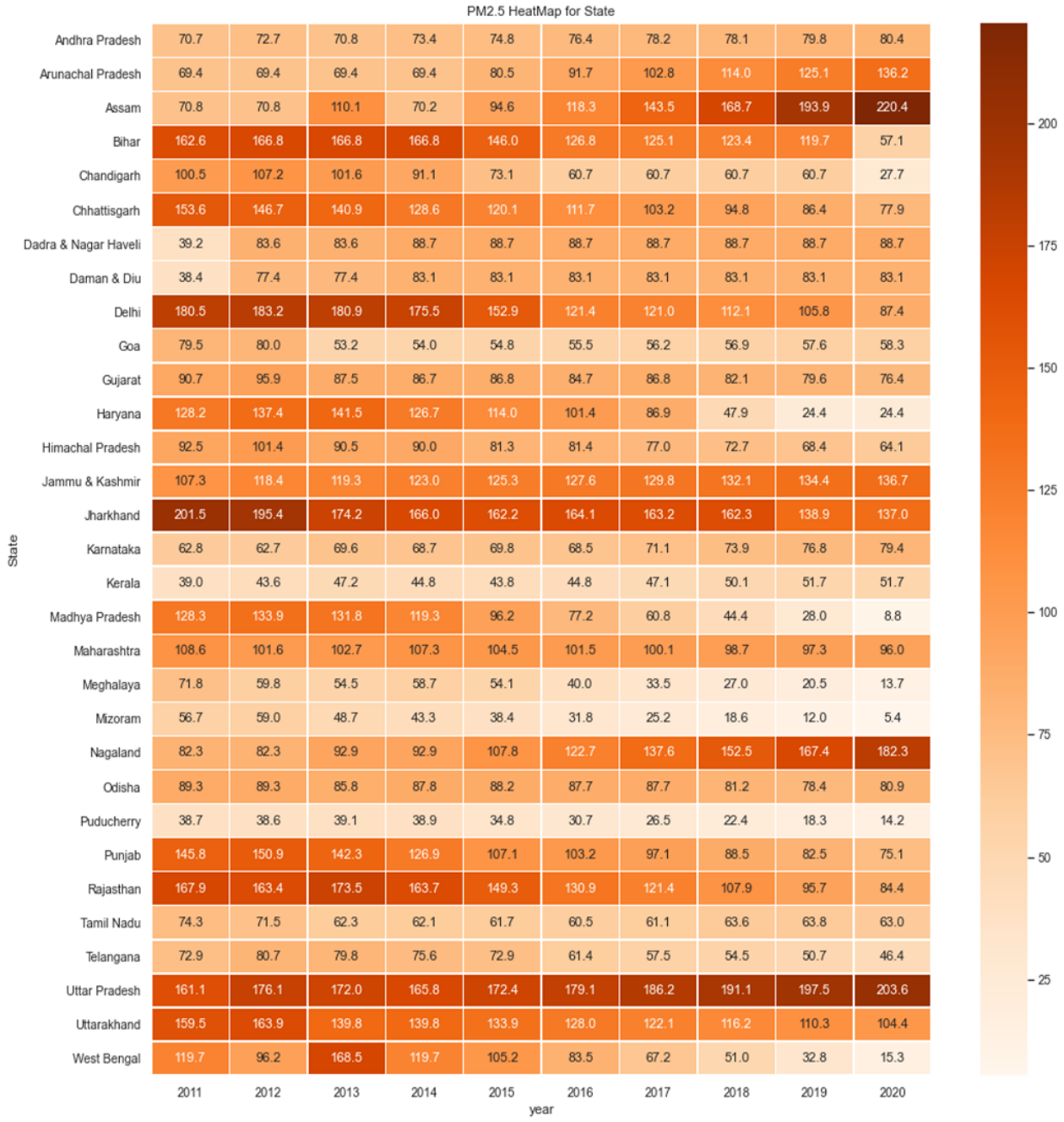
We can observe from the above heatmap that Haryana has a somewhat high NO2 concentration from 2015-2020. We can see that the NO2 concentrations in Jammu and Jharkhand are very similar to those in Delhi. For the years 2016-2020, Nagaland, Mizoram, and Meghalaya have the lowest NO2 concentrations. For the year 2019, NO2 levels in Jharkhand have risen dramatically.

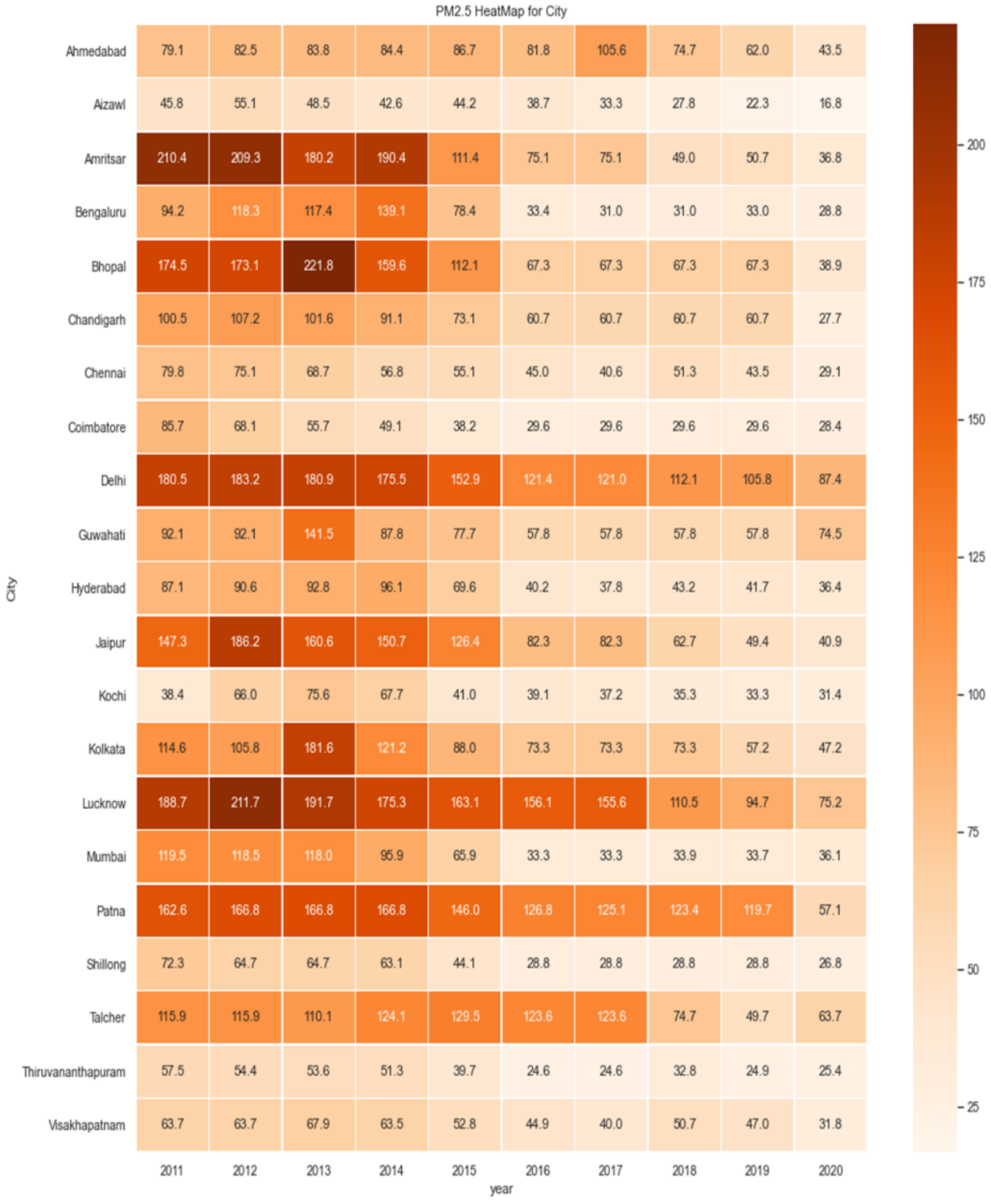
Similarly, from the heatmap of cities, we can say Ahmedabad has the highest NO2 concentration, and Aizawal has the least.



As observed from the NO2 concentration barplot above, Haryana has initiated NO2 concentration in 2019 more significant than the mean. In terms of NO2 concentrations, Delhi and West Bengal are in second place. NO2 concentrations are modest in Jharkhand, Maharashtra, and Bihar. In the same way that Goa has the lowest SO2 concentration, it also has the lowest NO2 concentration.

4.5.3 State and City wise PM2.5 concentration

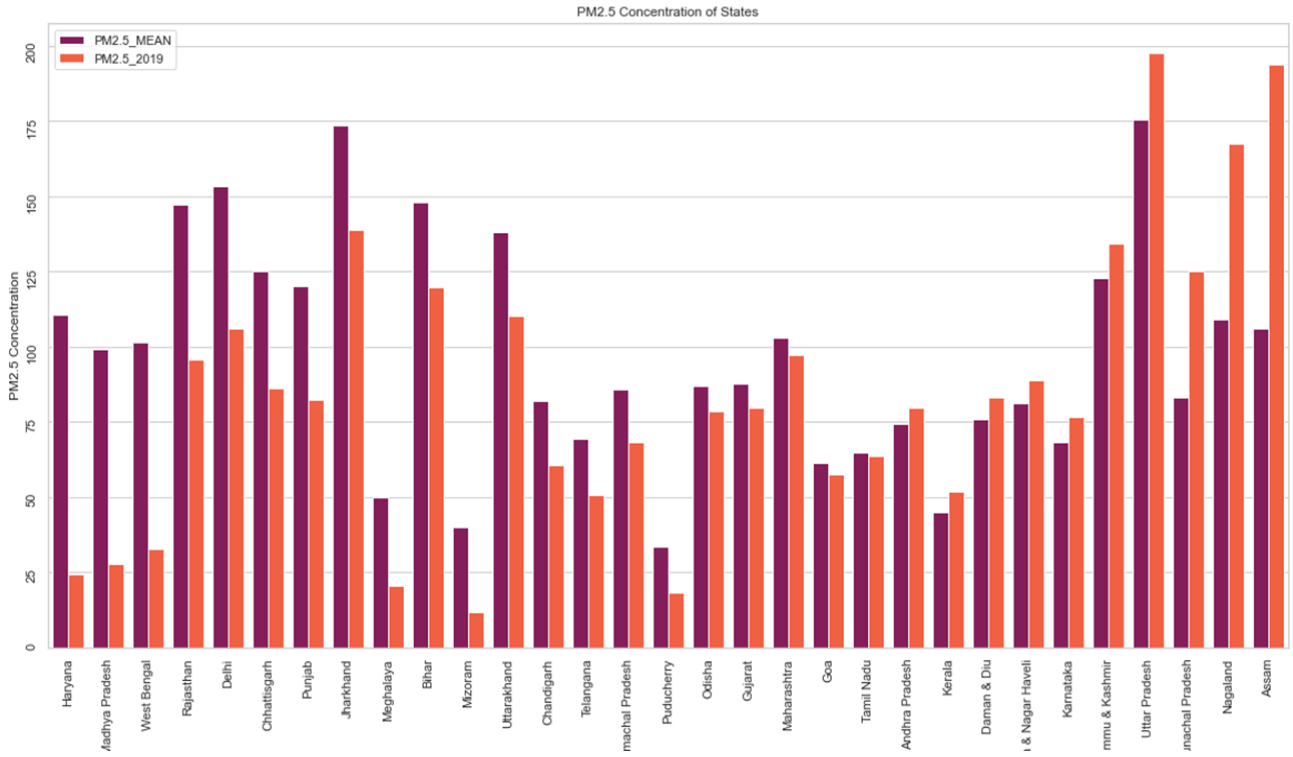




According to the heatmap, Assam and Uttar Pradesh have had high PM2.5 concentrations for the past ten years. After 2015, the concentration of PM2.5 in Nagaland increased dramatically. In Jharkhand, there was a sharp increase in PM2.5 concentration in 2012,

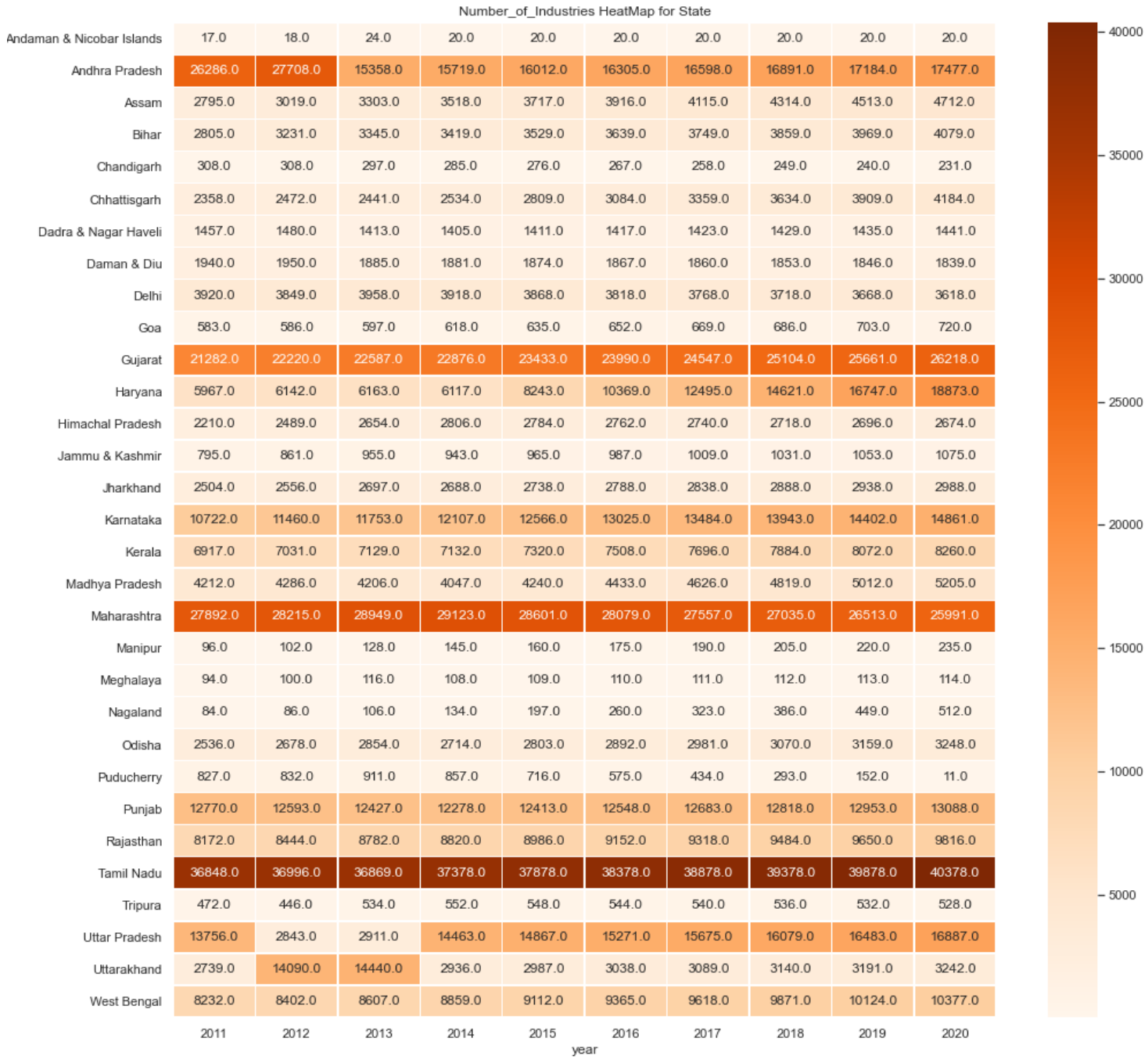
followed by moderate to high PM2.5 concentration. In addition, PM2.5 concentrations in Punjab, Uttarakhand, and Puducherry have gradually decreased.

Similarly, According to the city heat maps, Bhopal had the highest PM2.5 concentration in 2013 while Aizawal had the lowest.

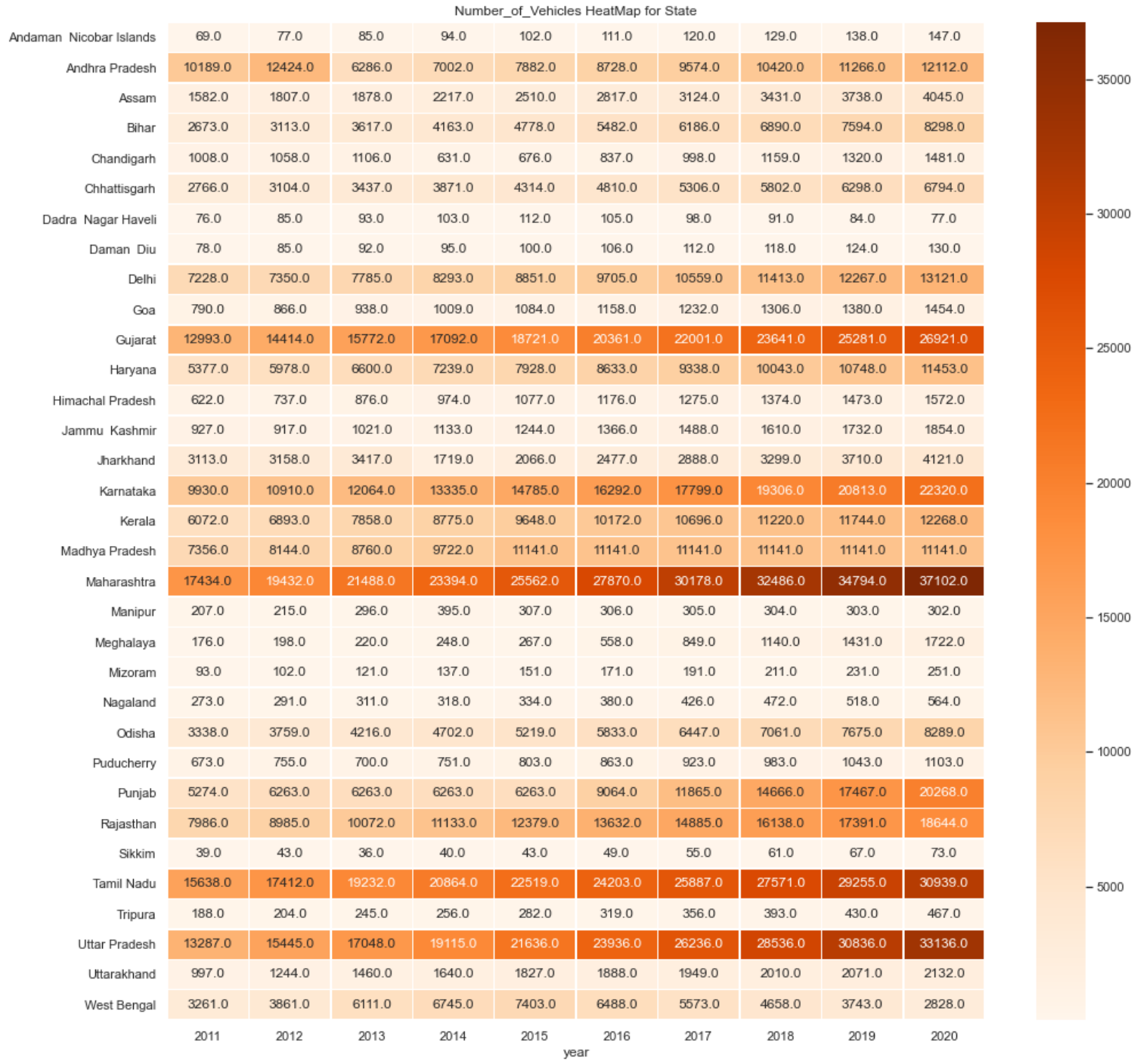


The bar plot for 2019 shows that Uttar Pradesh has the highest PM2.5 levels. Uttar Pradesh and Jharkhand, which are placed second and third, have practically identical PM2.5 mean concentration values but they are in opposite directions through the years while Uttar Pradesh kept on increasing till 2019 while Jharkhand gradually decreased to 2019. Uttar Pradesh is the most densely inhabited, immediate action is required to control pollution levels in these states. It's also worth noting that Puducherry has the lowest PM2.5 levels.

4.5.4 Plots for Number of industries in states



4.5.5 Plots for number of vehicles in states



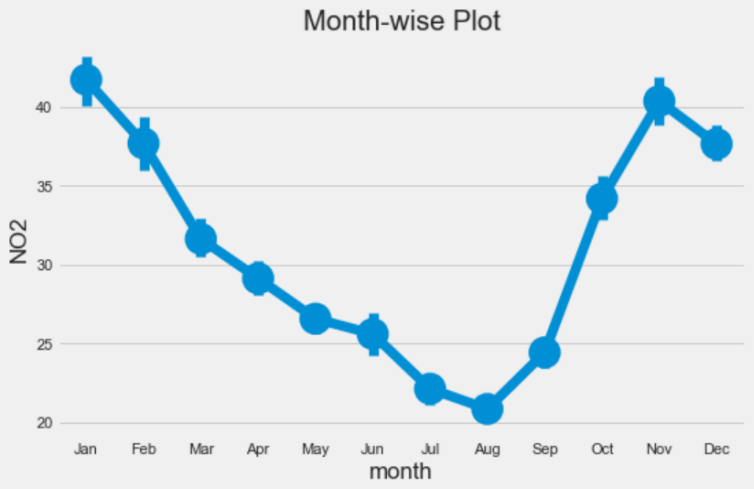
Chapter 5

Impact of COVID-19 on Pollution

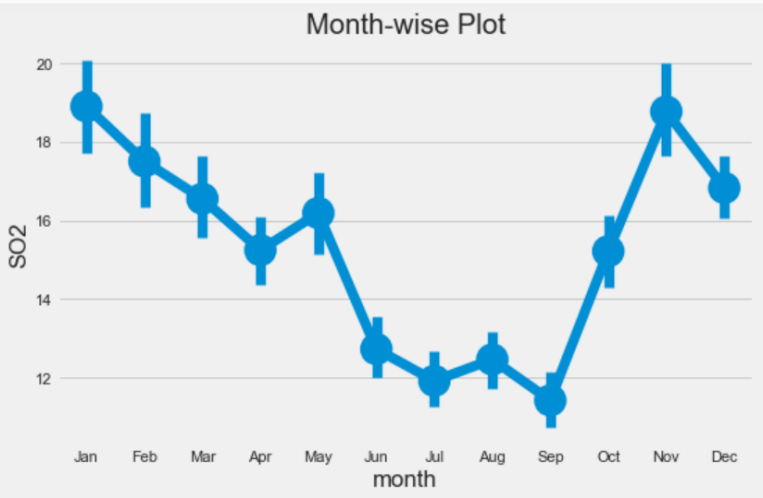
Lockdown due to the Covid-19 pandemic was announced on the 22nd of March 2020. For the next four months, India has imposed the strictest lockdown in the world, in order to curb the spread of covid-19. This was the time when the movement of vehicles and industrial activity was minimal. The emission levels were the lowest. So here we are finding the Pollutant levels, AQI in different places and comparing them with their previous levels.

5.1 Monthly analysis of Pollutants(NO2, SO2, PM) over last 5 years

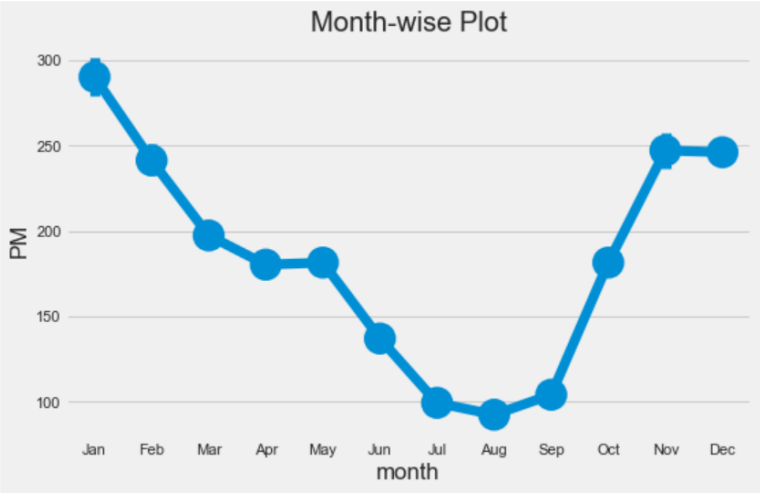
a. **NO2:** The concentration of NO2 in the atmosphere increases during the winter season starting September and peaks during November and January. As the Monsoon arrives in the months after June, the concentration falls to the lowest in the month of August due to heavy rains. Vehicular and Industrial consumption of fossil fuels increases the concentration levels after September.



b. **SO2:** The concentration of SO2 in the atmosphere increases during the winter season starting September and peaks during November. As the Monsoon arrives in the months after June, the concentration falls to the lowest in the month of September due to heavy rains. Vehicular and Industrial consumption of fossil fuels increases the concentration levels after September. There is an increase in SO2 levels during the month of May which is a result of increased generation of Thermal power due to high demand in the Summer season.



c. **Particulate Matter(PM):**



It is the combination of organic and inorganic particles which are suspended in the air. These are microscopic in nature and can enter bloodstreams if their size is less than 2.5microns. Just like NO2 and SO2 levels, this is also low in August as the monsoon rainfall settles the particulate matter and washes it onto the ground. Later due to vehicular and industrial combustion and also because of the stubble burning in the Northern Western parts of India, PM levels increase reaching the peak in January before falling down. The slight increase during the month of May is due to the dry summer season and loo winds which bring dust and sand from the desert onto the Northern plains. Roadside dust is also stirred up which contributes to the increase during summer.

5.2 Cities with Highest Pollutant Concentrations

Five year average of pollutant concentrations in different cities is plotted to understand which places are experiencing the maximum effect of rising pollutant levels.

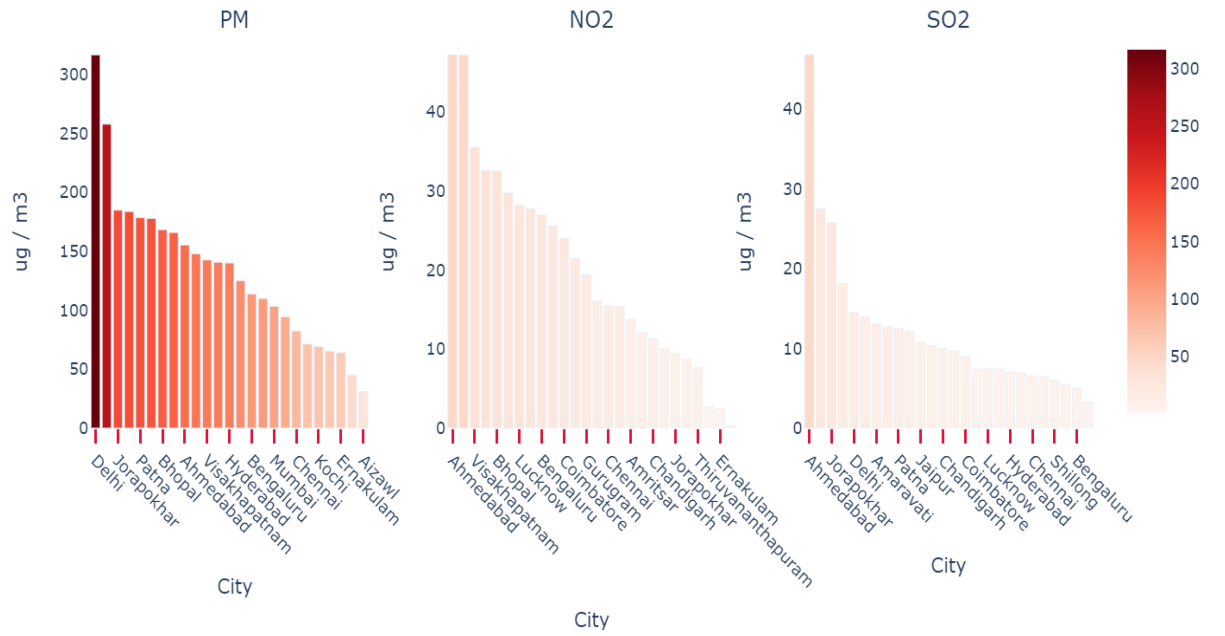
➢ From the above plot, it is evident that Delhi has the maximum Particulate Matter average over the past 5 years whereas Ahmedabad topped both NO2 and SO2 levels over 5 years. There are nearly 11 Thermal power plants around the NCR firing coal continuously round the clock to generate electricity for NCR. Fly ash produced from these thermal power plants and stubble burning in the surrounding regions of NCR contributes to the higher values of particulate matter in Delhi.

5.3 Effect of lockdown due to Covid-19

In this section, various trends in the AQI, pollutant levels during the months of 2019 and 2020, especially in the lockdown are analysed and presented.

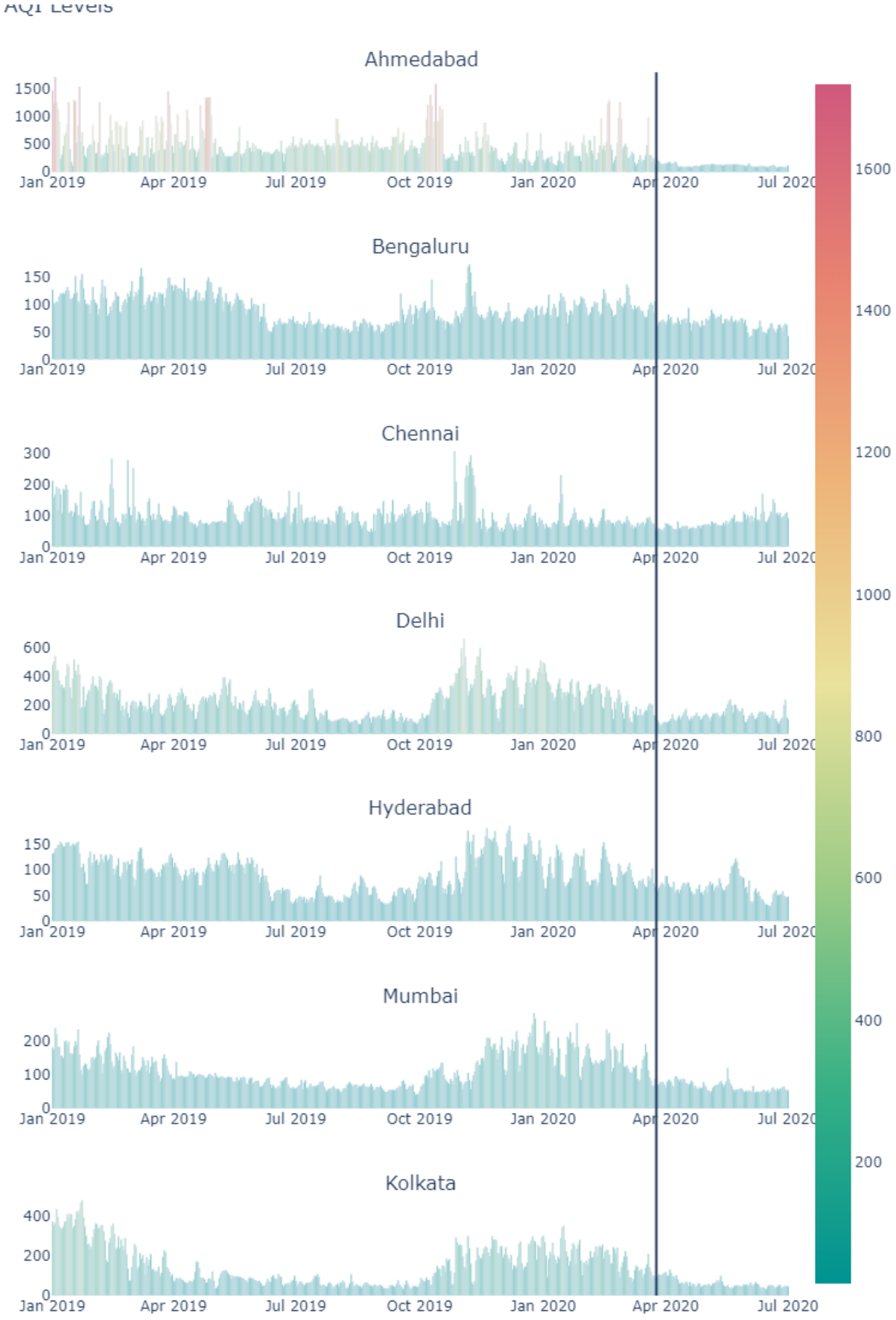
5.3.1 Analyzing AQI of different cities over 2020

For this analysis 7 cities are chosen based on population, industrialization and location.

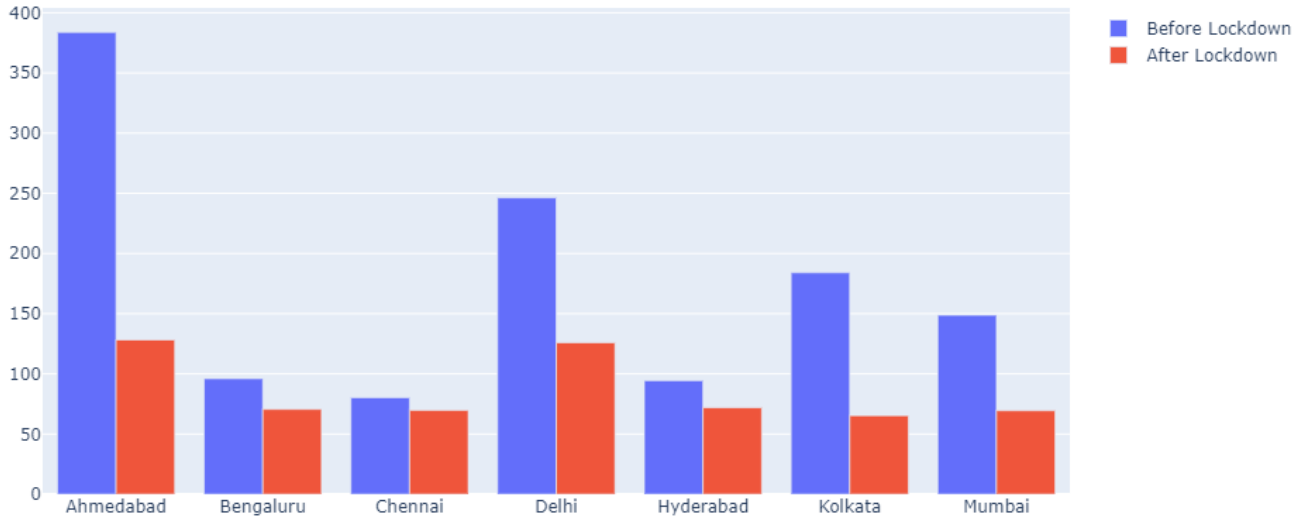


Ahmedabad, Delhi, Bengaluru, Mumbai, Hyderabad, Chennai, Kolkata are preferred. These cities will cover all regions of the country.

- The figure shows the variation of various pollutant levels, from Jan 2019 onwards till July 2020.
- The black vertical line shows the date on which the first phase of lockdown came into effect in India.
- Apparently, all the above Indian cities seem to have a dangerously high level of pollution before the lockdown.
- We can observe a general downward trend in AQI during the monsoon season starting June-July till September-October across all parts of the country.
- Clearly, there is a rapid downfall after 25th March 2020(after lockdown) in all the cities under consideration.
- Of all the cities, Ahmedabad had the steepest decline followed by Mumbai and Delhi. Whereas, Bengaluru hasn't recorded such steeper falls compared to other cities.


AQI Levels
Ahmedabad
1500
1000
500
0
Jan 2019
Apr 2019
Jul 2019
Oct 2019
Jan 2020
Apr 2020
Jul 2020
Bengaluru
150
100
50
0
Chennai
300
200
100
Delhi
600
400
200
Hyderabad
Mumbai
Kolkata
1600
1400
1200
1000
800
600
400
200

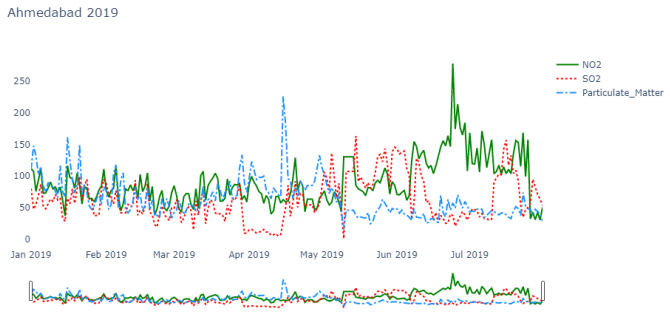
5.3.2 Comparison of AQI before and after lockdown

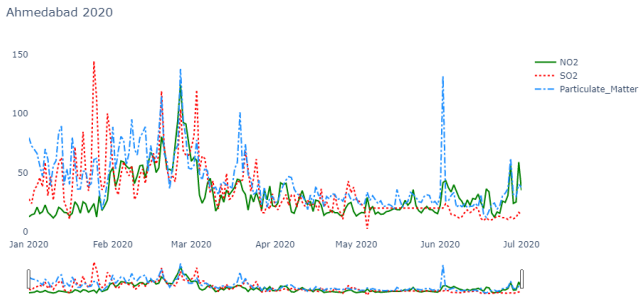


By this comparison chart, it is clearly evident that the AQI level after Lockdown has drastically fallen down compared to before Lockdown levels. Ahmedabad and Delhi experienced a greater decrease in AQI. Also, it is clear from this comparison that the lockdown has a significant effect on the pollution levels in India.

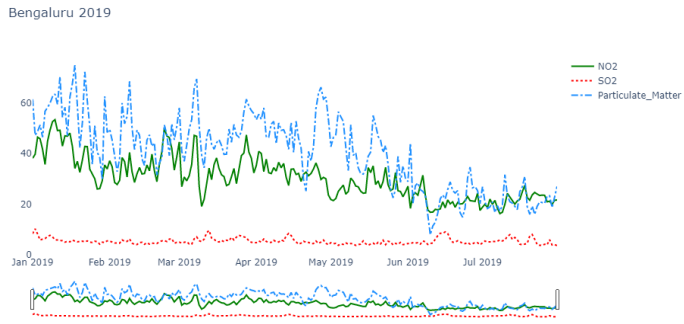
5.3.3 Effect of Lockdown on individual pollutant levels

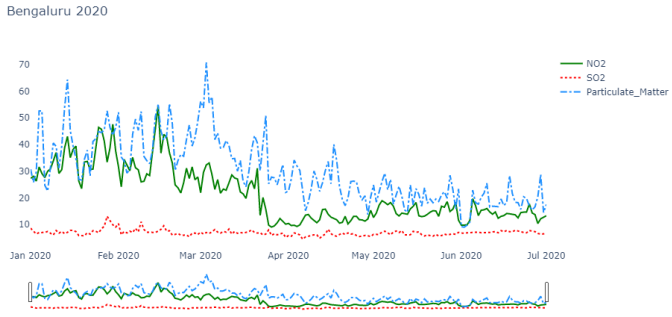
Ahmedabad



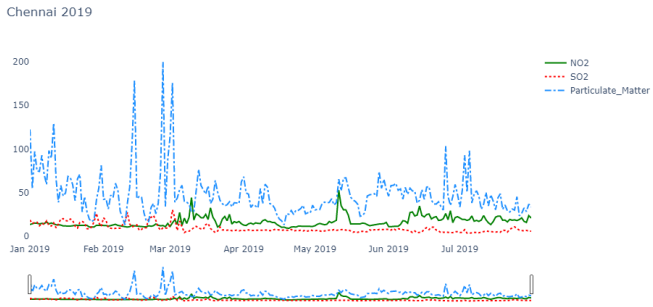


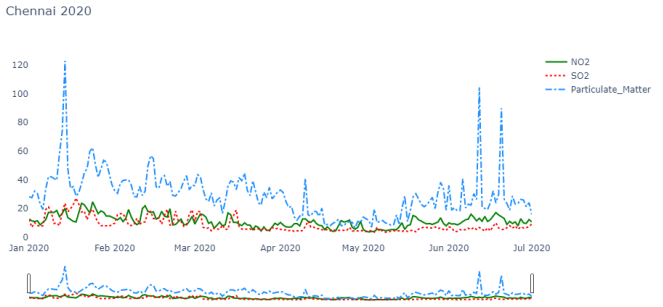
Bengaluru



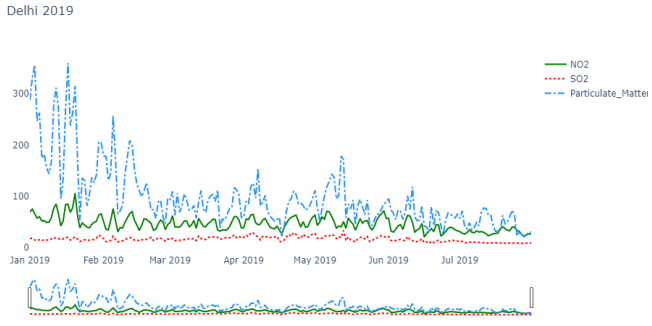


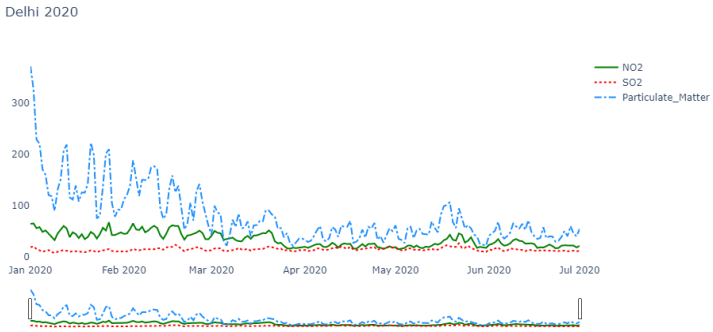
Chennai



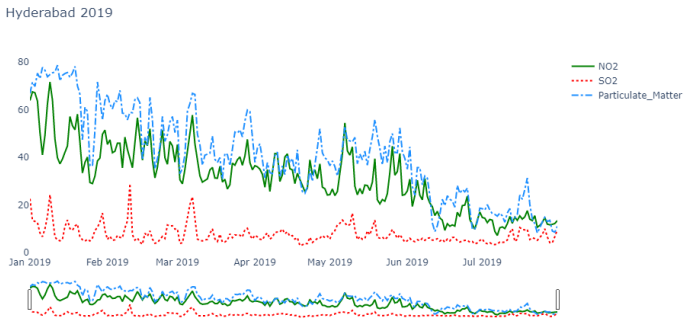


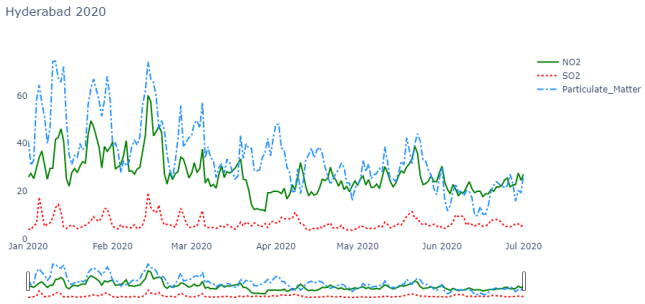
Delhi



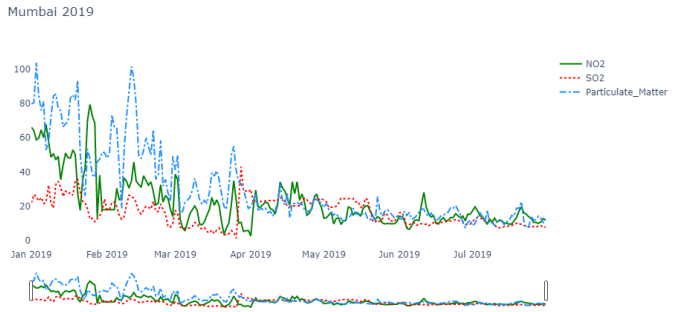


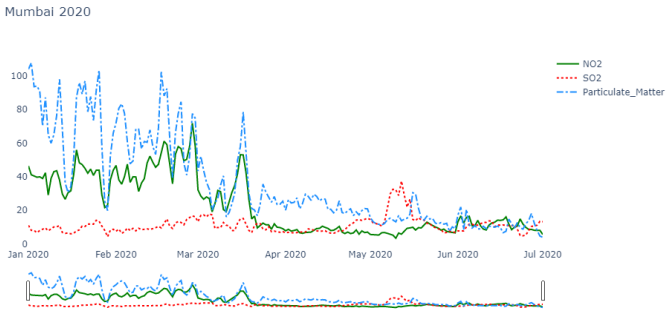
Hyderabad



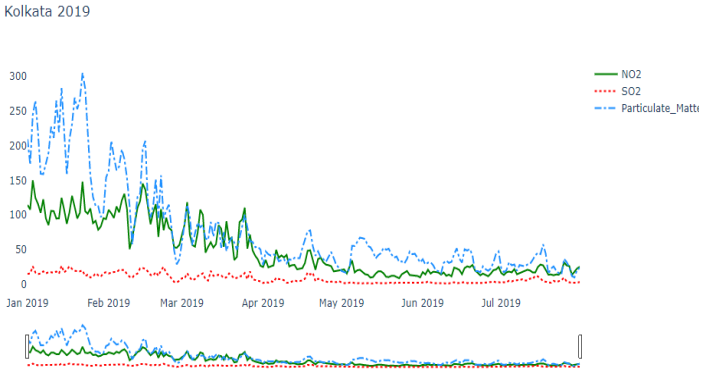


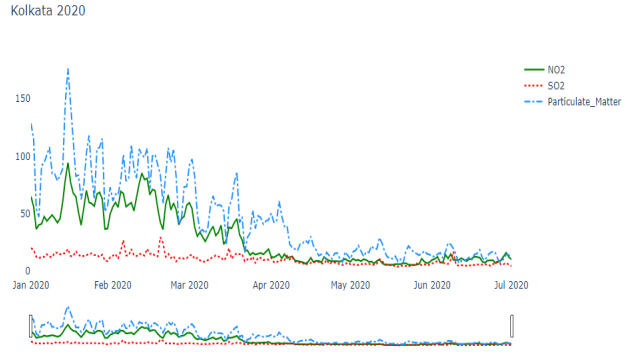
Mumbai





Kolkata

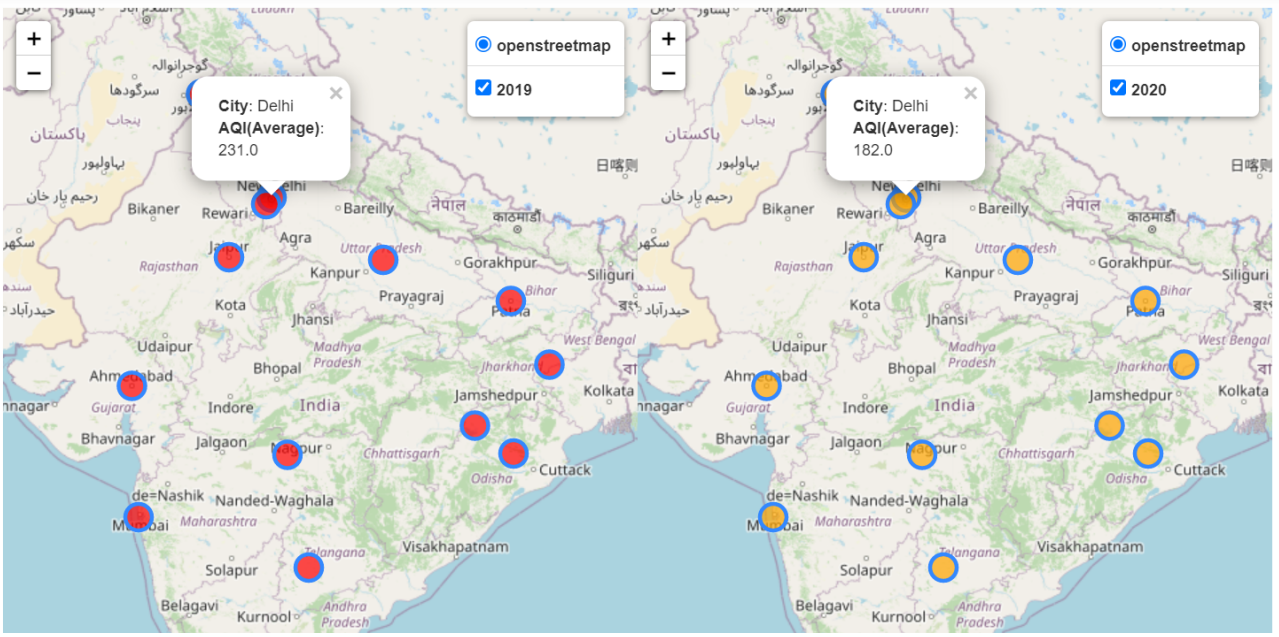




- It is interesting to note that the various pollutant levels in India generally drop down as summer approaches. This can also be affirmed by the graphs above.
- However, the reduction in march 2020 is more pronounced as compared to March 2019.

5.4 Dual map to visualize the AQI change during 2019 and 2020

This map shows the comparison between AQI levels in different cities in India for the year 2019 and 2020.



Chapter 6

Results and Conclusion

6.1 Decadal Pollution Analysis

- States with large populations and high industrial activity are hotspots. Such states include Uttar Pradesh, Maharashtra, Andhra Pradesh etc.
- Other states like Himachal Pradesh, Goa, Kerala and almost all the seven sister states are coldspots.
- NO2 levels are highly correlated with the number of vehicles and the number of industries.
- The number of vehicles is also highly correlated with the number of industries. This shows that vehicular registration is more in industrialized places.
- A significant correlation between PM2.5 and Industry/vehicles is observed.
- Uttarakhand, Maharashtra, and Jharkhand paid the most attention to SO2. For Meghalaya, we can see a lot of bounce within the SO2 in 2013, 2018, and 2011. For ten years, Himachal Pradesh, Manipur, Mizoram, and Nagaland have displayed much less change in SO2.
- Ahmedabad has the highest SO2 concentration in the year 2019 Talcher of Odisha has higher in 2016. In contrast, Shillong of Meghalaya and Alwal have the most negligible SO2 concentration.
- Haryana has a somewhat high NO2 concentration from 2015-2020. We can see that the NO2 concentrations in Jammu and Jharkhand are very similar to those in Delhi. For the years 2016-2020, Nagaland, Mizoram, and Meghalaya have the lowest NO2 concentrations. For the year 2019, NO2 levels in Jharkhand have risen dramatically.
- Ahmedabad has the highest NO2 concentration, and Aizawal has the least.
- Haryana has NO2 concentration in 2019 more significant than the mean. Delhi and West Bengal are in second place. NO2 concentrations are modest in Jharkhand, Maharashtra, and Bihar. Whereas Goa has the lowest SO2 and NO2 concentrations.
- Assam and Uttar Pradesh have had high PM2.5 concentrations for the past ten years. After 2015, the concentration of PM2.5 in Nagaland increased dramatically. In Jharkhand, there was a sharp increase in PM2.5 concentration in 2012, followed by moderate to high PM2.5 concentration. In addition, PM2.5 concentrations in Punjab, Uttarakhand, and Puducherry have gradually decreased.
- Bhopal had the highest PM2.5 concentration in 2013 while Aizawal had the lowest.
- Uttar Pradesh has the highest PM2.5 levels in 2019. Uttar Pradesh and Jharkhand, which are placed second and third, have practically identical PM2.5 mean concentration values.
- PM2.5 in Uttar Pradesh kept on increasing till 2019 while in Jharkhand, it gradually decreased.

- Uttar Pradesh is the most densely inhabited, immediate action is required to control pollution levels in these states.
- It's also worth noting that Puducherry has the lowest PM2.5 levels.

6.2 Impact of Covid-19

- Delhi has the maximum Particulate Matter average over the past 5 years whereas Ahmedabad topped both NO2 and SO2 levels over 5 years.
- Pollution in all the Indian cities is more than the safer limits before the lockdown.
- Starting June-July, till October, AQI during the monsoon generally falls due to monsoonal rains washing the pollutants from the air.
- Clearly, there is a rapid downfall in AQI levels after 25th March 2020(after lockdown) in all the cities under consideration. Of all the cities, Ahmedabad had the steepest decline followed by Mumbai and Delhi. Whereas, Bengaluru hasn’t recorded such steeper falls compared to other cities.
- Ahmedabad and Delhi experienced a greater decrease in AQI.

Chapter 7

Future work

The 21st Century pandemic, Covid-19 has created a huge impact on the pollution levels by decreasing the AQI and pollutant concentrations in India and across the world. With much better awareness about climate change, the Greenhouse effect and global warming, the global population and leadership is very much concerned about the wellness of this planet and is determined to make the earth a better place to live. The recently concluded 26th Conference of Parties in the Glasgow United Kingdom reiterated this commitment towards the environment. So it is important and also interesting to monitor the change in pollution after the steep fall during the covid pandemic. This study can be extended in the upcoming years to track the progress of the commitments made on such big stages. More pollutants can also be added to the study and find the effect of pollution on respiratory illnesses.