Indian Institute of Technology, Kanpur
Department of Computer Science and Engineering

Analysis of Air Pollution Levels in India

Project Report - GROUP 21
CS685: Data Mining
Under the guidance of
Prof. Arnab Bhattacharya
Academic Year 2020 - 2021

Ankita Dey (20111013)
ankitadey20@iitk.ac.in

Deeksha Arora(20111017)
deeksha20@iitk.ac.in

Sambhrant Maurya(20111054)
samaurya20@iitk.ac.in

Sharvari Oka(20111055)
okasharvar20@iitk.ac.in

Tamal Deep Maity(20111068)
tamalmaity20@iitk.ac.in

Acknowledgement

We would like to extend our sincere gratitude to our Project Guide, Prof. Arnab Bhattacharya for his advice, guidance, patience and timely help during the project. The freedom he gave us, to work with anything at any time, encouraged us to try out new things and helped to work more efficiently. It is a great honor for us to have been a part of his course. The successful completion of the work has been only possible due to his excellent guidance, meticulous observation and critical analysis.

Ankita Dey (20111013)
Deeksha Arora (20111017)
Sambhrant Maurya (20111054)
Sharvari Oka (20111055)
Tamal Deep Maity (20111068)

Contents

Abstract

Air pollution is a complex term to define, but any release of toxic gases or harmful particles into the atmosphere that are detrimental to human health can be termed as Air Pollution.There exist about 200 known pollutants, the major pollutants found in air include Carbon Dioxide, Sulphur Dioxide, Nitrogen Oxides, Soot, Smog, Dioxins, Polycyclic aromatic hydrocarbons, Chlorofluorocarbons and Methane.

The average human being inhales $10000 - 15000$ litres of air everyday. When polluted air is inhaled, pollutants enter our lungs; they can enter our bloodstream and be carried to our internal organs such as the brain. This can cause severe health problems such as lung diseases, cardiovascular diseases and even cancer. New studies have also found that air pollution affects every organ in the body and reduces the quality as well as number of years of life. The well known annual smog of India's capital-Delhi has been known to irritate the eyes, throat and lungs. Long exposure to severe pollutants like smog can affect an individual's IQ and the ability to learn. Smog aggravates heart problems, bronchitis, asthma, and other lung diseases. Sulphur Dioxide- a major pollutant, can cause respiratory problems like bronchitis and asthma attacks and has been linked to cardiovascular diseases as well.

Air pollution costs the world economy $5 trillion per year as a result of productivity losses and degraded quality of life, according to a study conducted by the World Bank. The World Bank states that additional economic losses caused by air pollution, including health costs and the adverse effect on agricultural and other productivity were not calculated in their report, and thus the actual costs to the world economy are far higher than $5 trillion.

WHO estimates that 9 out of 10 people in India are exposed to polluted air everyday. Air Pollution was among the top 5 risk factors for deaths in India for 2019. A study has found that India suffers most pollution related deaths in the world, where about 2 million deaths are linked to Air Pollution annually. Hence, being aware about air pollution, it's causes and it's effects is of utmost importance.

Chapter 1

Introduction and Broad Aims of the Project

1.1 Introduction and Motivation

In today's era, there are many potential sources of Air Pollution. Air Pollution mostly comes from energy use and production. In India, air pollution is believed to have grown more than the birth rate in the country. India has watched the most elevated increment in the pollution level through the most recent decade, 10 out of 20 most polluted cities are in India. As per a gauge given by WHO, 1 out of 8 deaths were ascribed to introduction to air contamination making it the biggest ecological hazard factor for the sick well being. The number of vehicles running on the roads is on the rise in India and has possibly contributed to the much of Air Pollution in India. With booming industrialization in the country, pollution is seemingly unavoidable because people are interested in making more money, but are not equally concerned about the environment. The number of factories is on the rise, and factories are known to emit tremendous amount of pollutants into the atmosphere. Deforestation is also on the rise, because the need for rapid development demands clearing of forests for more available land. Rapid deforestation has possibly fueled the rise in air pollution, because trees inherently prevent pollutants from settling around. The ever increasing population of India is also possibly correlated with the rise in the Air Pollution levels in the country.

This project uses 3 input features indicative of possible causes of Air Pollution- *Number of Vehicles per state*, *Number of Factories per state* and *Population Density*.

1.2 Broad aims of the project

This project aims to use data mining techniques to accomplish the following:

1. To find trends in the Air Pollution levels in various states of India over the years $2005 - 2014$.
2. To find the most polluted states in India with respect to SO_2, NO_2 and RSPM concentrations.
3. To find the hotpots and coldspots of Air Pollution by comparing the Pollution levels in each state with it's neighboring states using z-score and to classify the states as outliers and non-outliers using chi-score.
4. To find a correlation between the trends in Air Pollution of each state with these possible factors- Number of Vehicles in the state, Number of Factories operating in the state and Population Density of the state.
5. To cluster the states based on their pollution levels.

Chapter 2

Datasets Required

The following datasets are required for the analysis:

1. Air Quality Data of India
2. Statewise Total Registered Motor Vehicles in India
3. Statewise Total Number of Industries in India
4. Statewise Population Enumeration Data
5. Coordinates of the state boundaries of India
6. Neighboring states of each state of India

Chapter 3

Data Preparation

This chapter includes the source of datasets and how to clean these datasets for our analysis.

3.1 Data Collection

- **Air Quality Data of India**: The air quality data of India has been obtained from *Kaggle*, available at https://www.kaggle.com/shrutibhargava94/india-air-quality-data. This dataset is available as a **csv** file. It contains SO_2, NO_2, RSPM, SPM and PM 2.5 for all the states of India from 1990 to 2015.
- **Statewise Total Registered Motor Vehicles in India**: The dataset of statewise total registered motor vehicles has been taken from *www.mospi.nic.in* available at http://mospi.nic.in/sites/default/files/statistical_year_book_india_2015/Table-20.4_0.xlsx as **.xls** file.
- **Statewise Total Number of Industries in India**: The dataset for total number of industries in each state has been taken from:
  - 2009-2014 data is obtained from *www.mospi.gov.in* available at http://mospi.nic.in/sites/default/files/statistical_year_book_india_2015/Table%202.1.xls as **.xlxs** file.
  - 2001-2006 data is obtained from *labourbureau.gov.in* available at http://www.labourbureau.gov.in/ASI_V2_2005_06_TAB27F.docx as **docx** file.
  - 2007-08 data is obtained from *labourbureaunew.gov.in* available at http://www.labourbureaunew.gov.in/UserContent/ASI_Vol_1_2007_08.pdf as **pdf** file.
- **Statewise Population Enumeration Data**: The population data of India has been taken from Indian census 2001 and census 2011. The 2001 census data is available at https://censusindia.gov.in/Census_Data_2001/Census_data_finder/A_Series/Total_population.htm webpage and 2011 census data is available at http://mospi.nic.in/sites/default/files/statistical_year_book_india_2015/Table%202.1.xls as **.xls** file.
- **Coordinates of state boundaries of India**: The coordinates of the boundaries of all the states of India has been taken from Oracle, available as a zipped GeoJson file at http://download.oracle.com/otn/samplecode/data-visualization/sample-MapLayers_Maps/INDIA_STATES_051120.zip.

3.2 Data Cleaning

The data collected needs to be cleaned to make it ready for any further processing. Few things that we made sure during the cleaning stage were that:

- All missing values were either removed or interpolated.
- Non numeric values were made numeric.
- Names of states were made to be in sync with their most recent names.
- Telangana and Andhra Pradesh data were merged.
- The data of Andaman & Nicobar Islands, Lakshadweep and Tripura was removed as these states are not present in the Air Quality Dataset.
- Final cleaned datasets were stored in csv format with headers as per requirement.

The various steps taken to clean each dataset are:

- **Air Quality Data of India**: This dataset contains the headers: 'stn_code', 'sampling_date', 'state', 'location', 'agency', 'type', 'so2', 'no2', 'rspm', 'spm', 'location_monitoring_station', 'pm2_5', 'date'. SO_2, NO_2, *RSPM*, *SPM* and *PM* 2.5 values recorded for different cities from 1990 to 2015. Some insights about the data — the number of entries in each column, the type of entry in each column, etc are shown in the below figure, where we see that we have 435742 entries in our dataset. There are very few non-null values present for PM 2.5 and SPM. So, only SO_2, NO_2 and RSPM levels has been used in our analysis from 2005 to 2014.

```
In [4]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435742 entries, 0 to 435741
Data columns (total 13 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   stn_code                     291665 non-null  object
 1   sampling_date                435739 non-null  object
 2   state                        435742 non-null  object
 3   location                     435739 non-null  object
 4   agency                       286261 non-null  object
 5   type                         430349 non-null  object
 6   so2                          401096 non-null  float64
 7   no2                          419509 non-null  float64
 8   rspm                         395520 non-null  float64
 9   spm                          198355 non-null  float64
 10  location_monitoring_station  408251 non-null  object
 11  pm2_5                        9314 non-null    float64
 12  date                         435735 non-null  object
dtypes: float64(5), object(8)
memory usage: 43.2+ MB
```

Figure 3.1: Insights of Air Quality Data of India

To get year-wise and state-wise values of SO_2, NO_2 and RSPM for each year from 2005 to 2014, firstly 'date' has been modified to contain just the year and secondly, the mean of data available for cities of the corresponding state for that particular year has been calculated. The missing values have been replaced by the mean value of that particular state for that particular year.

```
In [16]:
1 # Get rows that have null values for no2
2 df[df['no2'].isnull()]
```

Out[16]:

		so2	no2	rspm
state	date			
Mizoram	2005	NaN	NaN	27.856287

```
In [17]:
1 # Get rows that have null values for so2
2 df[df['so2'].isnull()]
```

Out[17]:

		so2	no2	rspm
state	date			
Mizoram	2005	NaN	NaN	27.856287
	2008	NaN	7.246286	37.304207
Nagaland	2005	NaN	14.760000	79.764706

```
In [18]:
1 # Get rows that have null values for rspm
2 df[df['rspm'].isnull()]
```

Out[18]:

		so2	no2	rspm
state	date			
Dadra & Nagar Haveli	2012	7.618667	20.181333	NaN
Daman & Diu	2012	7.628947	20.147368	NaN

Figure 3.2: Missing Values in Air Quality data after merging location wise and year wise information for each State/Union Territory

We experimented linear and quadratic interpolation for the values of SO_2,NO_2 and $RSPM$ and we encountered that the data variations were almost similar for few states and after plotting the linear and quadratic plots for the interpolated values, we inferred that linear interpolation provided the best approach in this scenario with non-negative values. Quadratic interpolation did not yield a promising number of positive values. It can be observed that the values of SO_2,NO_2 and $RSPM$ concentrations available in the data set are in a particular range and do not show much variance. Hence when we apply linear interpolation, the calculated values appear to fit in the acceptable range. In Figure 3.3, we can observe that during the year 2009-2011, there was a spike in the obtained values. Here the quadratic interpolation suggests that the curve should be fitted upwards, but doing so would be incorrect as this surge is an exception when compared to the trend of other data points which show almost

linear variation. The quadratic interpolation gives more negative values in this scenario because it tries to smoothen the curve and is particularly more accurate for a high variance data set. These reasons justify the use of linear interpolation to fill the missing values of SO_2,NO_2 and $RSPM$.

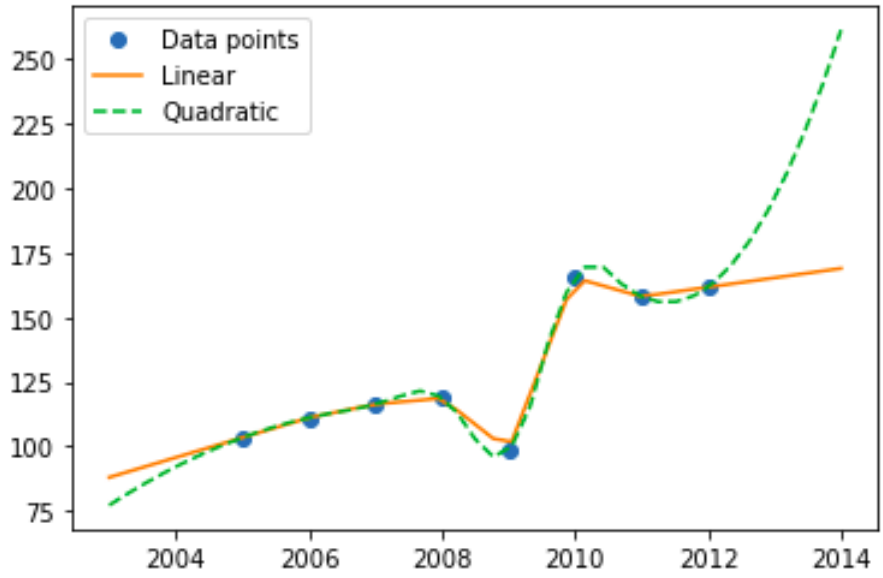

Figure 3.3: Linear vs quadratic interpolation for RSPM values of Bihar

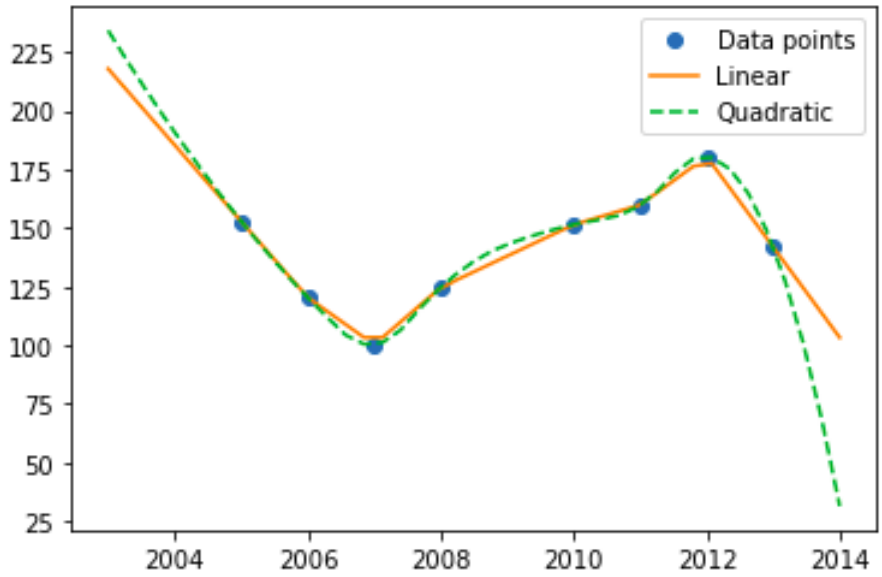

Figure 3.4: Linear vs quadratic interpolation for RSPM values of Uttarakhand

As shown in above figures, for year 2014 quadratic interpolation gives either very high or low values whereas linear interpolation gives values in the range of the data set.

For Sikkim, Arunachal Pradesh and Manipur, the data was available only for the years 2007, 2008 and (2007 and 2008) respectively. This data was not

sufficient to perform temporal interpolation for finding missing values, therefore we resorted to spatial interpolation where we fill the data of the missing years of a state using the data of it's neighbors . To fill missing data for each of these 3 states, we found the ratio of pollutant concentrations for the state and the mean of its neighbors using the data available for a particular year, and then we fill the missing data of the state in the same ratio using the data of the neighboring states for each missing year.

- **Total Registered Motor Vehicles in India**: This dataset contains state-wise total registered motor vehicles from 2001 to 2015. The data for the years 2005 to 2014 has been used for this project. This dataset had no missing values. However, some fields had non numeric values appended with numeric values which had to be corrected to contain numeric values only. Odisha, Chhattisgarh and Dadra & Nagar Haveli had discrepancies in their names which were corrected.

- **Total Number of Industries in India**: This dataset contains state-wise distribution of factories, fixed capital, working capital, productive capital, invested capital, number of workers, total persons engaged, wages to workers, total emoluments, fuel consumed, materials consumed, total input, products and by products, total output, depreciation, net value added, rent paid for fixed assets, interest paid, gross fixed capital formation and value of addition in stock for 2008 to 2014. The information about number of factories in a state from this dataset has been used. Also, for Mizoram there is lack of information about number of factories. This datset had missing values for Arunachal Pradesh for the years 2009 to 2014 and for Sikkim for the year 2009. Uttarakhand and Dadra & Nagar Haveli had discrepancies in their names which were corrected.

- **Statewise Population Enumertion Data**: Total population of a state for 2001 was obtained from the 2001 census data. The data was scraped from the mentioned webpage using python's BeautifulSoup library. The census 2011 data contains area of state(in sq. km), male population, female population, total population, rural population and urban population for each state. The total population and area of a state(in sq. km) has been used from this dataset. Using the data obtained from 2001 and 2011 census of India, the population for intervening years and years after 2011 has been estimated for each state. For estimation, we used a constant rate of growth r per year given by $(1 + r)^{time\,duration} = \frac{population\,in\,2011}{population\,in\,2001}$. Using this rate of growth, the population for any year can be estimated using $(population\ in\ 2001) \times (1+r)^{year\,-\,2001}$. There were some discrepancies in the names of states such as Uttarakhand, Odisha, Puducherry, Andaman and Nicobar Islands, Dadra and Nagar Haveli, Manipur, Jammu and Kashmir which were corrected.

Chapter 4

Methodology and Analysis

4.1 Chi-Score for finding the Hotspots

We have used a modified version of Chi-Squared Test to detect outliers in multivariate data. To determine whether a state is an outlier i.e. hotpsot or coldspot, the following formula is used:

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

where o is the object to be tested (a state) and o_i is the value of o on the i^{th} dimension. E_i is the mean value on the i^{th} dimension among all objects. The object may be identified as an outlier if the Chi-value is larger than a threshold value.

4.1.1 Methodology

If the *p-value* of a state is less than *level of significance* (1%) then the state is considered as *outlier i.e. the state is either a hotspot or coldspot.* The results obtained from this *modified version of Chi-Squared Test* have been plotted on an interactive choropleth map of India for the year 2014 which is available at https://deekshaarora05.github.io/Analysis-of-Air-Pollution-levels-in-India/Maps/chiscore-map.html

4.1.2 Observations

The Choropleth map of India in Figure 4.1 shows that Arunachal Pradesh, Bihar, Dadra and Nagar Haveli, Daman and Diu, Delhi, Goa, Haryana, Jammu and Kashmir, Jharkhand, Kerala, Madhya Pradesh, Maharashtra, Meghalaya, Mizoram, Nagaland, Puducherry, Punjab, Rajasthan, Sikkim, Tamil Nadu, Uttar Pradesh, Uttarakhand and West Bengal are outliers but using *Chi-Squared Test for outlier detection* we can't determine whether the outlier state is a hotspot or coldspot.

A lot of states in Northern India fall under the outlier category. Though these can be either hotspots or coldspots, looking at the annual pollution problem in states like Uttar Pradesh, Punjab and Delhi, it can be argued that most of those states would possibly come out as hotspots. A similar argument can be put forth for states like West Bengal and Maharashtra where the pollution problem is significant in reality. On the other hand, some of the North Eastern Seven Sister states turning out to be outliers gives us the impression that there is a larger possibility of those states turning out as coldspots, especially so when they are neighboured by Assam, which is more developed than its neighbors. This leads us to believe that the pollution levels in the neighboring states of Assam would be much lesser than the pollution levels in Assam itself. To get a more accurate picture of the hotspots and coldspots, we next employ the method of $Z - score$ analysis.

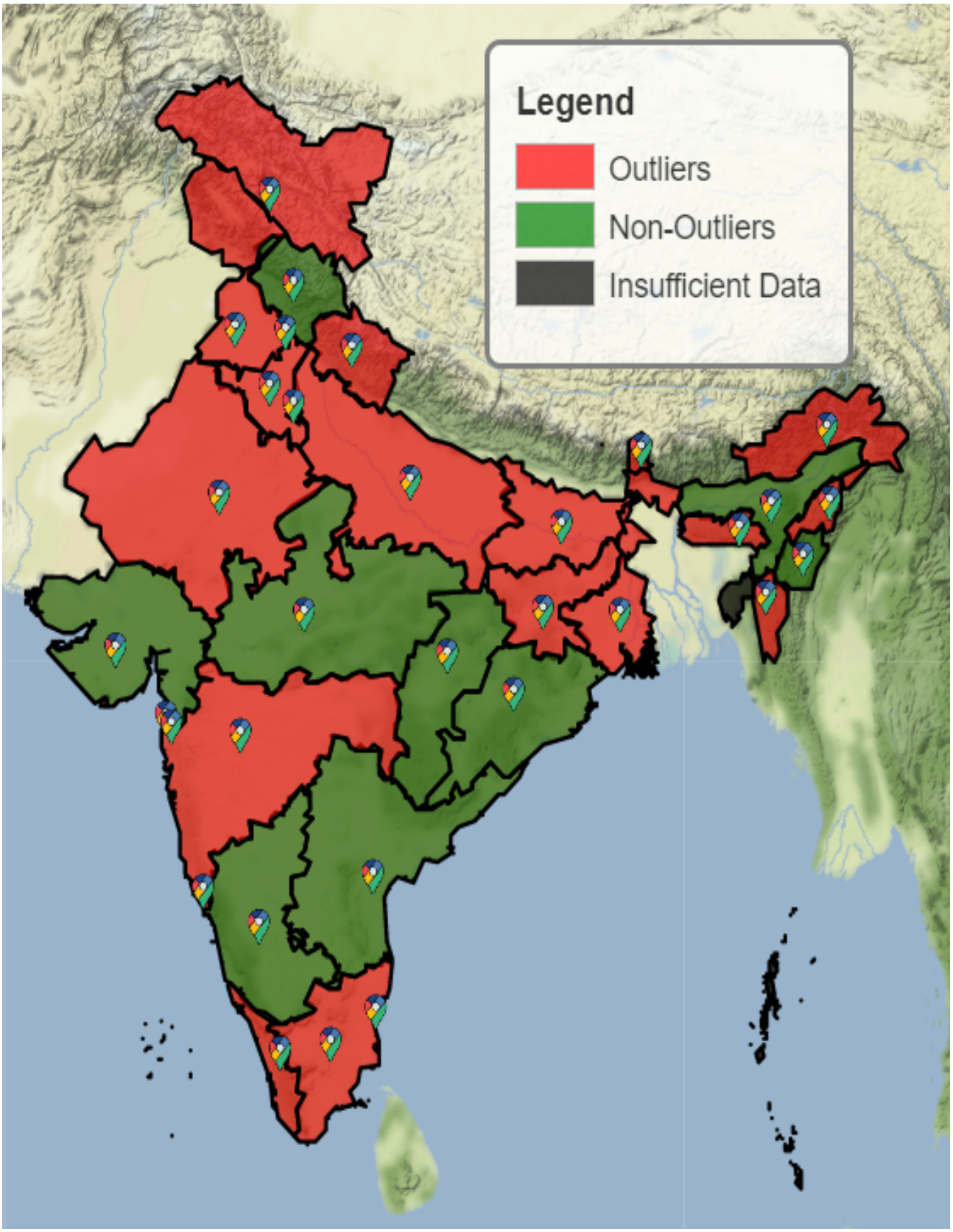

Figure 4.1: Hotspots in 2014 Using Chi-Score as seen on Choropleth maps

4.2 Z-Score for finding the Hotspots

The limitation of *Chi-Squared Test for outlier detection* is that it only tells whether the state is an outlier or not but it doesn't give any information about the nature of outliers i.e. whether the outlier state is a hotspot or coldspot. To determine this we have used *Z-Score analysis*.

4.2.1 Input data

To determine whether a state is hotpsot or not, we calculated the Mean Pollutant Concentration for each state from year 2005 to 2014 using the formula:

$$Mean\ Pollutant\ Concentration = \frac{SO_2\ conct. + NO_2\ conct. + RSPM\ conct.}{3}$$

4.2.2 Methodology

A state is considered as hotspot (respectively, coldspot) if the Mean Pollutant Concentration (MPC) of the state is greater than (respectively, less than) the mean plus (respectively, minus) half of the standard deviation of its neighbors. The results obtained from *Z-Score* are plotted on an interactive Choropleth map of India for the year 2014 which is available at https://deekshaarora05.github.io/Analysis-of-Air-Pollution-levels-in-India/Maps/zscore-map.html.

4.2.3 Observations

The Choropleth map of India in Figure 4.2 below shows that in the year 2014, Delhi, Haryana, Jharkhand, Maharashtra, Manipur, Nagaland, Rajasthan and West Beangal were the hotspot states whereas Andhra Pradesh, Arunachal Pradesh, Chandigarh, Dadra Nagar Haveli, Daman Diu, Goa, Himachal Pradesh, Kerala, Meghalaya, Mizoram, Odisha, Puducherry and Sikkim were coldspot states and the remaining states are neither hotspots nor coldspots.

We start by first analysing the hotspots. As expected from the chi score analysis, the states in Northern India do show up as hotspots. These results seem to tally with reality as Rajasthan, Uttar Pradesh, Delhi and Haryana are proximal to each other and these states witness extreme crop burning every year. Combined with this, the abundant presence of factories and vehicles in these states is also expected to drop the air quality severely. Interestingly, for a state to show up as a hotspot, loosely it needs that the neighbouring states be somewhat less polluted. These bunch of states are surrounded by states that are indeed less polluted. Various studies regarding the AQI (Air Quality Index) feature several cities in Uttar Pradesh like Moradabad and Greater Noida time and again as the dirtiest when it comes to air quality. On the other hand cities from states like Uttarakhand, Madhya Pradesh and Bihar seldom find place in such studies. A steep variation in AQI from their neighbouring states thus makes these set of states as National hotspots. States like West Bengal and Maharashtra find themselves densely populated while also being home to some of the most factory/industry oriented cities like Durgapur, Haldia, Raniganj in West Bengal and Nashik, Pune, Nagpur in Maharashtra. Both these states have another thing in common: Ports. Haldia Port in West Bengal and Jawaharlal Nehru Port in Maharashtra give ideal locational opportunities to factories that depend on raw

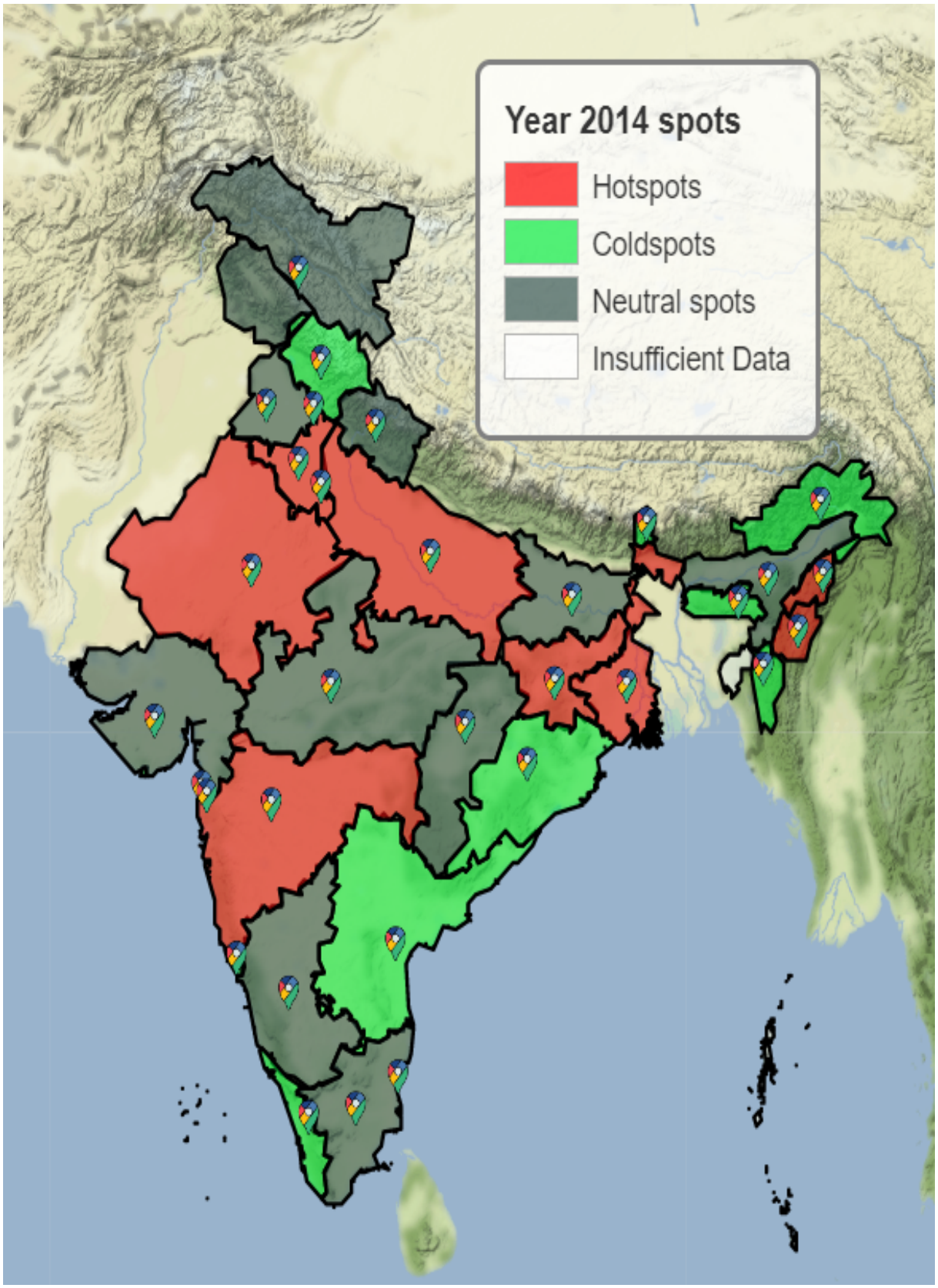

Figure 4.2: Hotspots in 2014 using Z-Score as seen on the interactive Choropleth maps

materials that are water transported. Settling in of these industries also mean more pollutant deposition in and around the air of these cities. Both these states also have to take the brunt of hosting metropolitan cities which mean more vehicles, more garbage burning and more construction activities which can only make the air pollution scenario graver.

The coldspot states in India are generally not representatives of superb air quality. Rather, they indicate that their neighbouring states are extremely poor when it comes to air quality. Take the example of Odisha that borders two of the most polluted states in the country - West Bengal and Jharkhand. This however might not be the case for the North Eastern states of Arunachal Pradesh, Meghalaya and Mizoram. Although they do border the more polluted state of Assam, yet these states have very little scope of being heavily polluted. There haven't been too many industries set up. Personal vehicle ownership isn't very common. And as such air in these states is much cleaner than in any other state of India.

The top-5 hotspot and top-5 coldspot states for each year using the z-score values are shown below:

Year	spot	State1	State2	State3	State4	State5
2005	hot	Punjab	Gujarat	Jharkhand	Uttar Pradesh	West Bengal
2005	cold	Goa	Odisha	Dadra & Nagar Haveli	Bihar	Mizoram
2006	hot	Punjab	Gujarat	Jharkhand	Uttar Pradesh	Delhi
2006	cold	Chandigarh	Bihar	Odisha	Goa	Mizoram
2007	hot	Punjab	Jharkhand	Uttar Pradesh	Delhi	Maharashtra
2007	cold	Odisha	Goa	Kerala	Mizoram	Dadra & Nagar Haveli
2008	hot	Punjab	Delhi	Jharkhand	Maharashtra	Uttar Pradesh
2008	cold	Arunachal Pradesh	Bihar	Chandigarh	Kerala	Mizoram
2009	hot	Jharkhand	Haryana	West Bengal	Punjab	Maharashtra
2009	cold	Chandigarh	Kerala	Bihar	Dadra & Nagar Haveli	Mizoram
2010	hot	Delhi	Punjab	Jharkhand	Tamil Nadu	Maharashtra
2010	cold	Himachal Pradesh	Chandigarh	Dadra & Nagar Haveli	Kerala	Mizoram
2011	hot	Delhi	Jharkhand	Punjab	Tamil Nadu	Maharashtra
2011	cold	Arunachal Pradesh	Dadra & Nagar Haveli	Chandigarh	Kerala	Mizoram
2012	hot	Delhi	Jharkhand	Maharashtra	Manipur	Punjab
2012	cold	Kerala	Arunachal Pradesh	Dadra & Nagar Haveli	Mizoram	Chandigarh
2013	hot	Delhi	West Bengal	Maharashtra	Tamil Nadu	Jharkhand
2013	cold	Goa	Himachal Pradesh	Chandigarh	Mizoram	Kerala
2014	hot	Delhi	Maharashtra	Jharkhand	Nagaland	Rajasthan
2014	cold	Chandigarh	Goa	Kerala	Arunachal Pradesh	Mizoram

Figure 4.3: Top five Hotspot and Coldspot states using Z-Score

4.3 Correlation

4.3.1 Input data

Air Pollutant Concentration data and number of industries, Motor Vehicles and Population Density of states from year 2005 to 2014 that has been used for Correlation.

4.3.2 Methodology

Correlation refers to the relation between the observed values of two variables. A positive association can be interpreted as : 'increase in value of one variable increases the value of other variable' whereas a negative association can be described as the instance : 'increase in value of one variable decreases the value of other variable'. Variables can also not affect one another at all in which case the correlation will be neither postive nor negative. The statistical phrase, "Correlation does not imply causation" holds true and we have to be careful in our analysis of the correlation coefficient so as to not draw false conclusions.

Correlation is typically measured by a number called correlation coefficient that lies in the range of -1.0 to 1.0 and signifies the strength of correlation. A correlation coefficient of 1.0 means that the variables are perfectly positively correlated while a value of -1.0 signifies perfect negative correlation. 0 correlation coefficient implies that there is no relationship between the variables. The book"Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach" further tries to interpret the correlation coefficient and mentions that values of +-0.1, +-0.3 and +-0.5 signify weak, moderate and strong relationship strength of the variables respectively. In this project, we find out two different correlation coefficients, namely **Pearson Correlation coefficient** and **Spearman Correlation Coefficient**.

Pearson Correlation Coefficient measures the linear correlation between two variables where each variable has a Gaussian distribution. An important feature of Pearson correlation is that it can find out only the linear relationship between variables x and y, i.e., change in x should yield proportional change in y. Spearman Correlation Coefficient on the other hand finds out a monotonic (and not necessarily linear) relationship between variables. A monotonic relationship means that as the value of one variable increases, the value of the other variable either keeps on increasing or keeps on decreasing and not necessarily at some fixed rate. Spearman Correlation Coefficient also differs from Pearson Correlation Coefficient in the sense that while Pearson's correlation, which is a parametric measure of correlation is a calculation of the covariance between two variables normalized by the variance of both variables, Spearman's correlation is a non parametric rank correlation, i.e., it computes the correlation using rank values and not real values.

We first find out both the Pearson Correlation Coefficient and the Spearman Correlation Coefficient. We do this by using the functions 'pearsonr' and 'spearmanr' provided by the scipy library in Python. Both these functions return several computations, out of which we have captured the correlation coefficient and the p-value for our analysis. The correlation coefficient is the same coefficient we discussed above and the p-value roughly indicates the probability of observing data generated by an uncorrelated system with equal or more correlation as obtained. For eg., a p-value of 0 indicates that the probability of observing the data given the samples are uncorre-

lated is higly unlikely and thus the samples are indeed correlated. (p-value reliable 500).

A look at the tables of the Pearson Correlation Coefficient and the Spearman Correlation Coefficient tells us that the latter seems more reliable. Some of the reasons for that are theoretical such as Spearman can capture more generality of relations than Pearson. If there is some linear relationship then it is monotonic too and thus Spearman correlation coefficient can be relied upon. However, the opposite doesn't hold. Pearson Correlation Coefficient can't be relied upon in case of monotonic non-linear relationships. This can be further justified by looking at the scatterplots. Barring the plots where there seems to be no relation between the variables, the rest aren't just linear - some of them seem to have a monotonic non linear relationship too. Another reason for choosing Spearman over Pearson for analysis is that the correlation values for Industry and Vehicle with $SO_2/NO_2/RSPM$ seems to be lower for Pearson and it falls under the range of weak correlation. However, from our practical knowledge we can safely estimate that the relationship between the number of vehicles and SO_2 is fairly high, which is exactly what the Spearman Correlation Coefficient gives. Further the p-values for those relations are higher for Pearson than Spearman suggesting that the chances for the Spearman Correlation Coefficient's value being accurate are more. Thus, Spearman Correlation Coefficient is used for any further analysis. A point to note here is that if a scatterplot indicates that a relationship isn't linear or monotonic then neither of the two Correlation Coefficients should be analysed for that particular pair of features/variables.

4.3.3 Observation

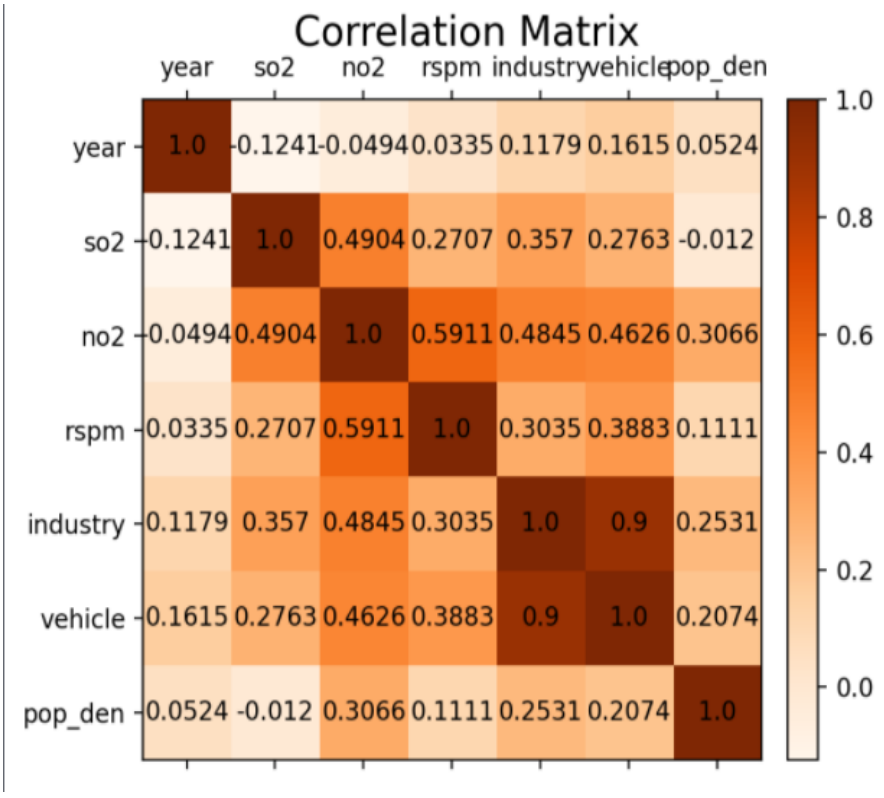

Figure 4.4: Correlation Matrix

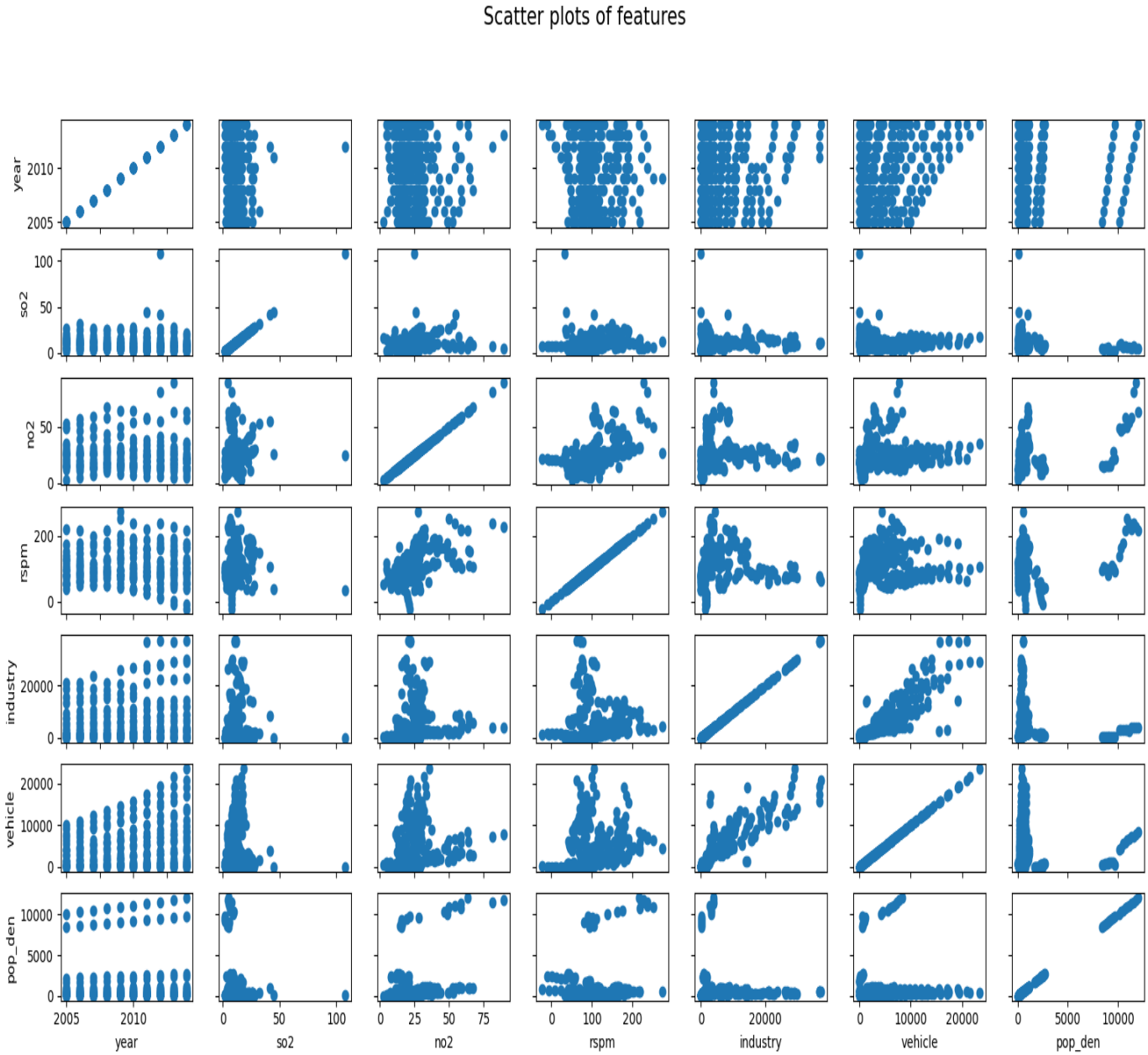

Figure 4.5: Correlation ScatterPlots

Now, coming to the correlation value analysis between the features, we can see from Figure 4.9 that the diagonals are 1, as expected. It signifies that a feature will always be strongly correlated with itself and doesn't give us much insight. Looking at the other values, we see a strong correlation (values greater than or close to 0.5) between $RSPM$ and NO_2 levels (value 0.59), SO_2 and NO_2 levels (value 0.49), NO_2 and industry (0.48), and NO_2 and vehicles (value 0.46). The correlation between number of vehicles and number of industries is very high with a value of 0.9 and this can be attributed to the fact that places with greater industries also require more vehicles to transport goods manufactured or utilised by the factories. There also seems to be a moderate correlation (values greater than 0.3) between vehicle and $RSPM$ (value 0.39), industry and SO_2 (value 0.36), and $RSPM$ and industry (value 0.30). All these correlations might not imply causation and there may be subtle interdependencies at play here. For eg., the strong correlation between NO_2 and SO_2 is most probably because of some feature which isn't a part of the feature set. Had the responsible feature been a part of our feature set, we would also have seen a much stronger correlation between SO_2 or NO_2 and that parameter.

A very interesting case develops here which shows how the correlation values actually make practical sense. We see that the correlation value between population density and all other features that we think impact air quality isn't very high and thus we can think of it as not being one of the primary reasons of AQI. This infact also turns out to be the case for West Bengal and Kerala for the year 2014. Though these states have similar population density, West Bengal comes as a hotspot while Kerala comes as a coldspot in our $Z - Score$ analysis indicating that population density isn't a key player in the air pollution scenario.

4.4 Clustering

We have used Air Pollutant Concentration data for year 2014 to group the states based on levels of SO_2, NO_2 and $RSPM$.

4.4.1 Methodology

We have grouped similar states based on 2014 mean pollutant concentration using clustering. We have used K-means algorithm that clusters data with similar features together with the help of euclidean distance.

The Euclidean distance is a distance measure between two points or vectors in a two - or multidimensional (Euclidean) space based on Pythagoras' theorem. The distance is calculated by taking the square root of the sum of the squared pairwise distances of every dimension. Considering n dimensional space, formula for Euclidean distance would be,

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

In our analysis, the number of dimensions is 3: SO_2, NO_2 and $RSPM$ concentration.

4.4.2 Observation

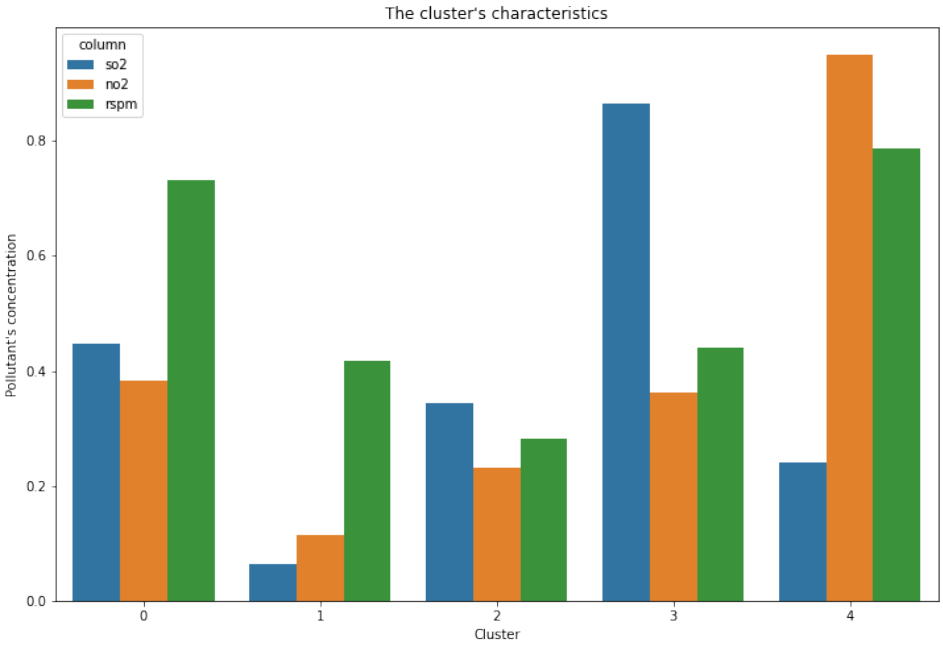

Figure 4.6: State Cluster vs Pollutant's concentration

Cluster 0:
Bihar, Chhattisgarh, Haryana, Jharkhand, Madhya Pradesh, Punjab, Rajasthan, Uttar Pradesh
Total Members: 8

Cluster 1:
Arunachal Pradesh, Chandigarh, Goa, Himachal Pradesh, Jammu & Kashmir, Kerala, Manipur, Meghalaya, Mizoram, Nagaland, Odisha
Total Members: 11

Cluster 2:
Andhra Pradesh, Assam, Dadra & Nagar Haveli, Daman & Diu, Karnataka, Puducherry, Tamil Nadu
Total Members: 7

Cluster 3:
Gujarat, Maharashtra, Sikkim, Uttarakhand
Total Members: 4

Cluster 4:
Delhi, West Bengal
Total Members: 2

Cluster 1 contains most of the hilly states of India. These states have the least concentration of SO_2 and NO_2 but still moderately high $RSPM$ concentration. In

reality, this is expected to cause severe weather in the hilly regions. In winters, as temperatures come down, the moisture content is condensed into fog. The high concentration of RSPM would expedite the process of fog formation, leading to drop in day temperatures and chilly weather.

Cluster 2 consists of some South Indian states and Union Territories. These state also have less pollutant concentration. This cluster shows that Southern part of India is in general less polluted than Northern part of India.

Cluster 4, consisting of Delhi and West Bengal, leads in NO_2 and $RSPM$ concentrations. Cluster 3 on the other hand leads in SO_2 concentration. Cluster 0 also has high pollutant levels. Notably all of the clusters which have high pollutant concentration consists of states from Central and Northern part of India.

4.5 HeatMaps and Bar Plots

4.5.1 Plots for SO_2 concentration of states

:

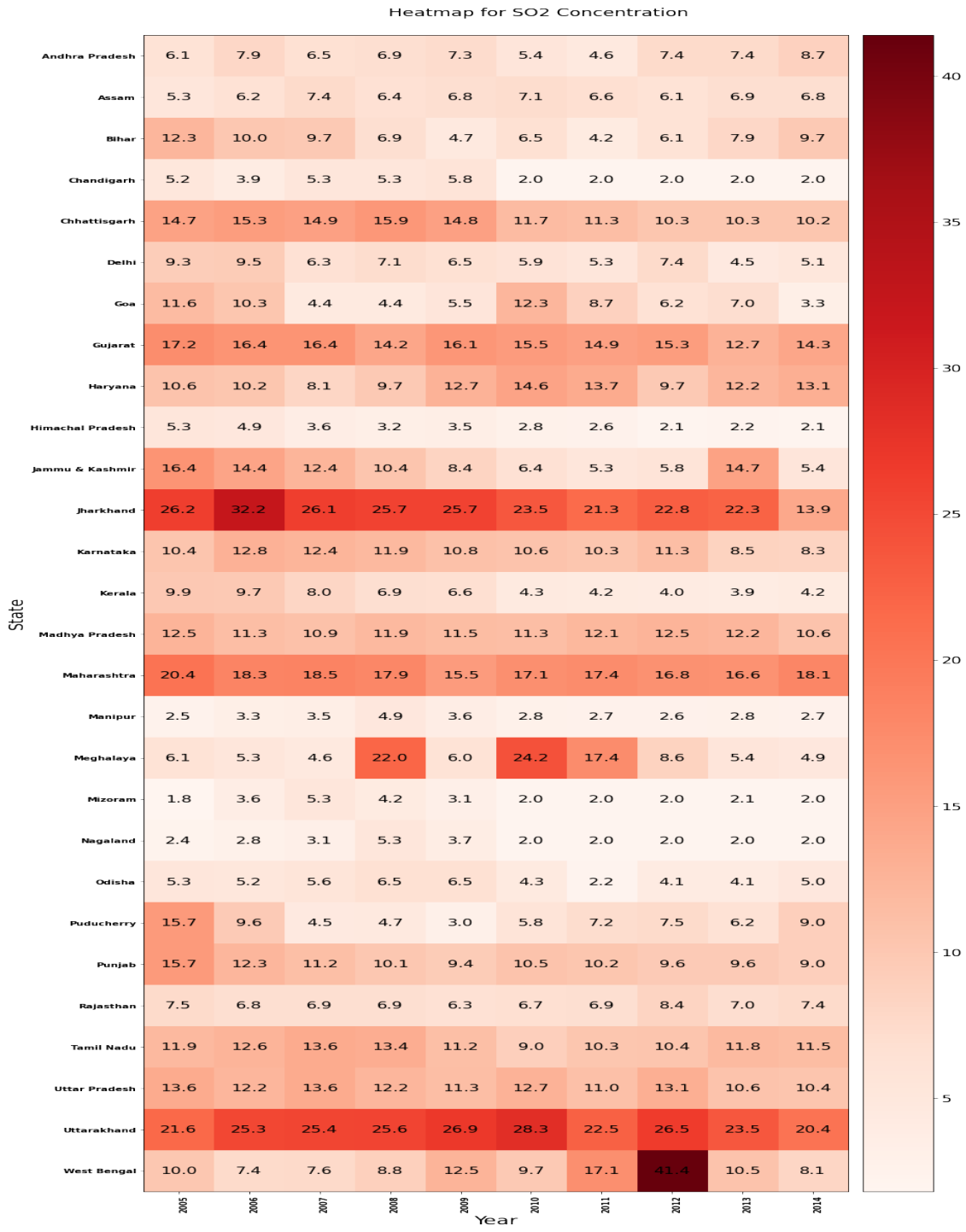

Figure 4.7: Heatmap for SO2 Concentration

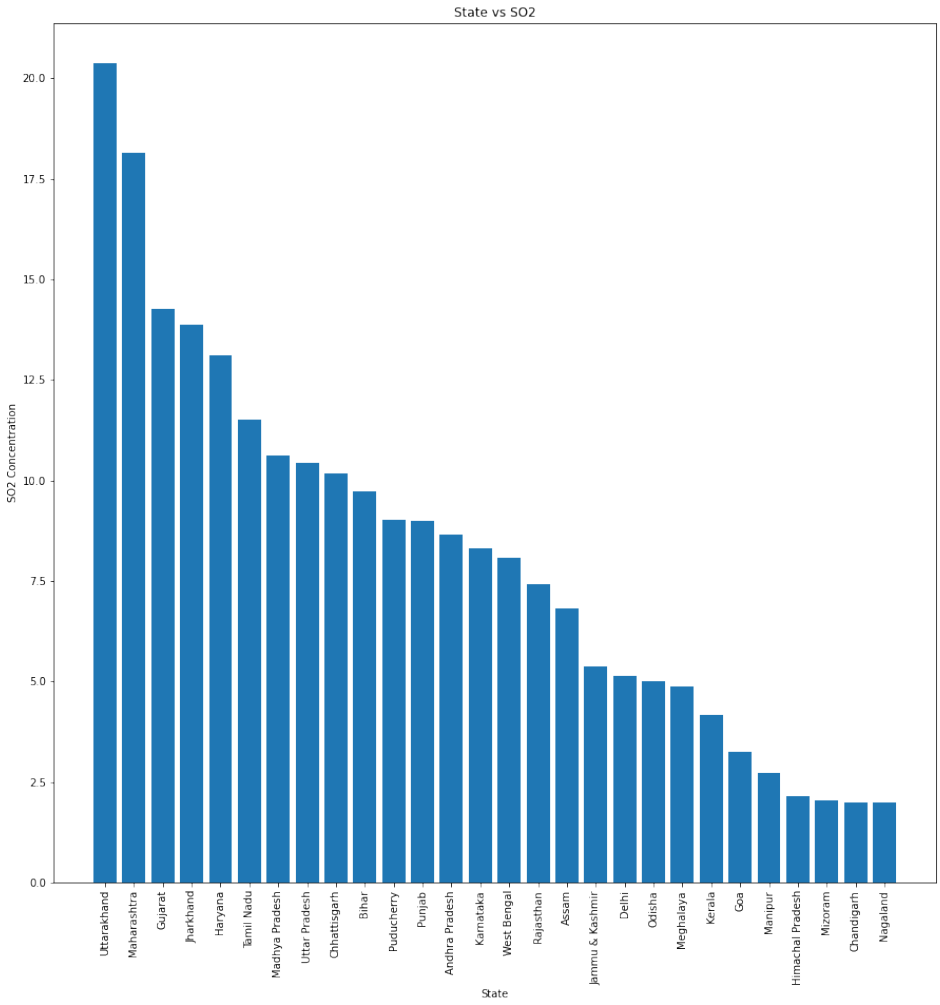

Figure 4.8: States vs their SO_2 Concentration for the year 2014

From the heatmap we can say that , West Bengal has witnessed highest SO_2 concentration level in year 2012. We can observe high leap in the SO_2 concentration in year 2008, 2010 and considerably high concentration in year 2011 for Meghalaya. Himachal Pradesh, Manipur, Mizoram, Nagaland show less concentration for SO_2 for the ten years span.
As shown in above barplot, Uttarakhand has the highest SO_2 concentration whereas Nagaland has the least SO_2 concentration for the year 2014. We can observe that there is not much difference between the levels of SO_2 concentration for Uttarakhand and Maharashtra which places Maharashtra at second rank. The SO_2 levels in Nagaland, Mizoram and Chandigarh are nearly same.

4.5.2 Plots for NO_2 concentration of states

:

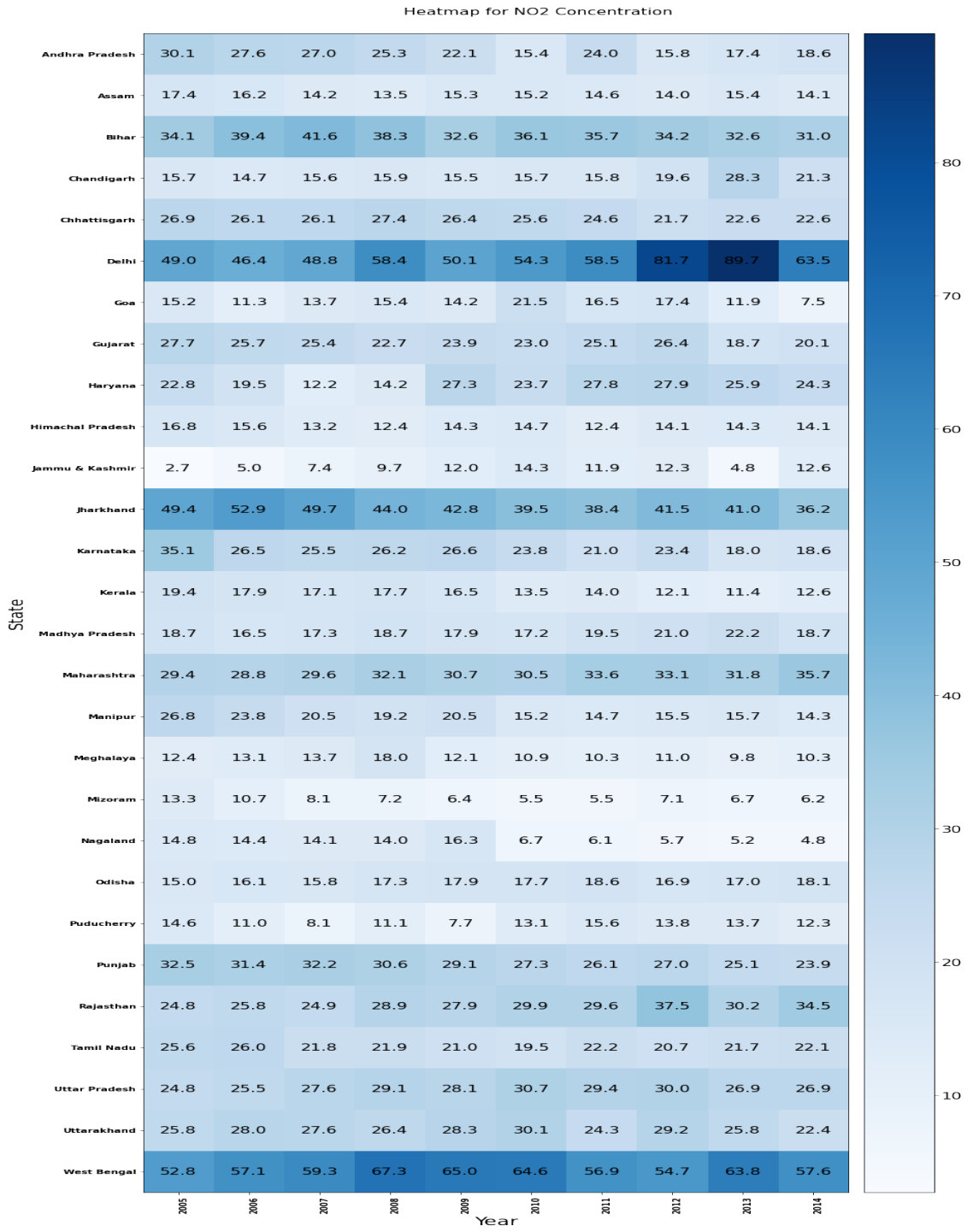

Figure 4.9: Heatmap for NO_2 Concentration

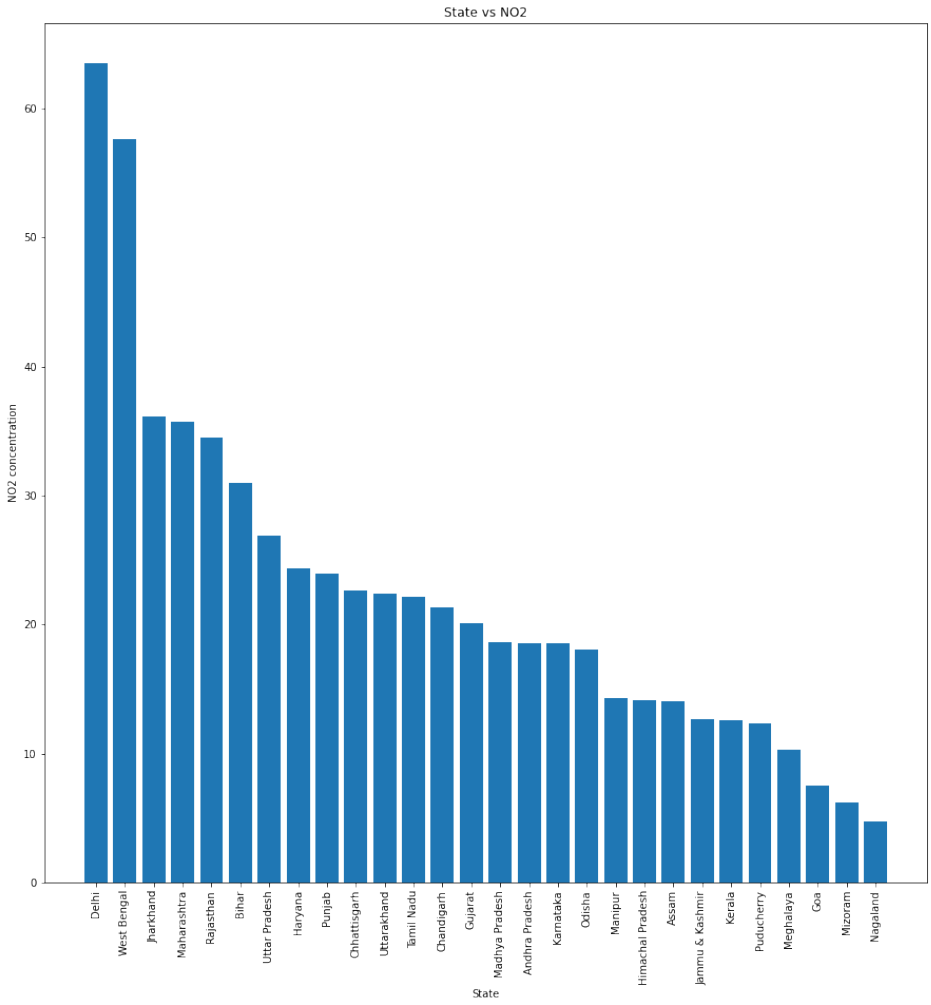

Figure 4.10: States vs their NO_2 Concentration for the year 2014

It can be seen from the above NO_2 concentration barplot, Delhi has very high NO_2 concentration. West Bengal comes at the second position in the NO_2 concentration levels. Jharkhand, Maharashtra, Rajasthan show moderate range of NO_2 concentration.
From the above heatmap as well, we can see that Delhi shows quite high concentration of NO_2 for the years 2005-2014. Nagaland, Mizoram have very less NO_2 concentration. For Rajasthan, there's sudden increase in NO_2 levels for the year 2012.

4.5.3 Plots for *RSPM* concentration of states

:

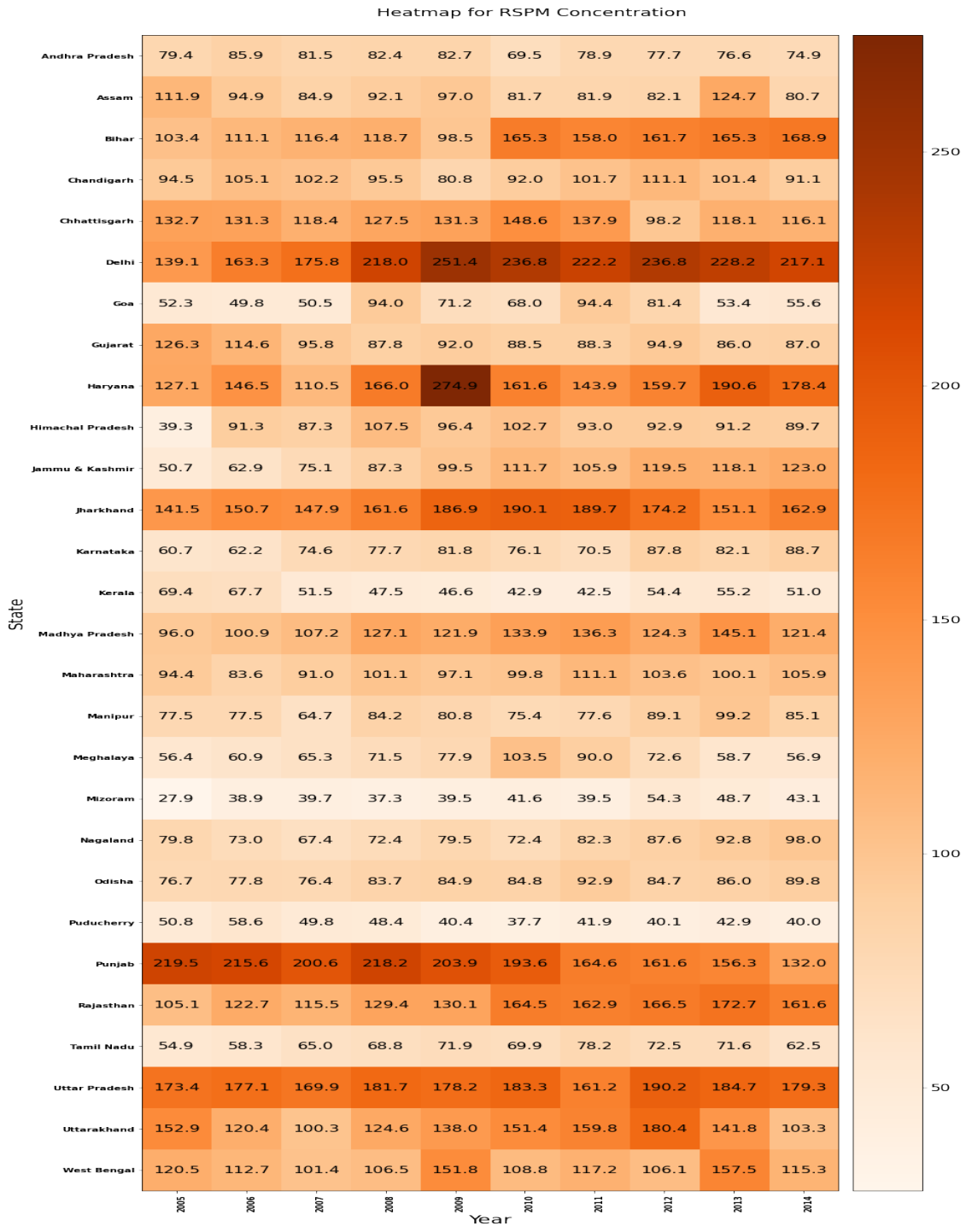

Figure 4.11: Heatmap for *RSPM* Concentration

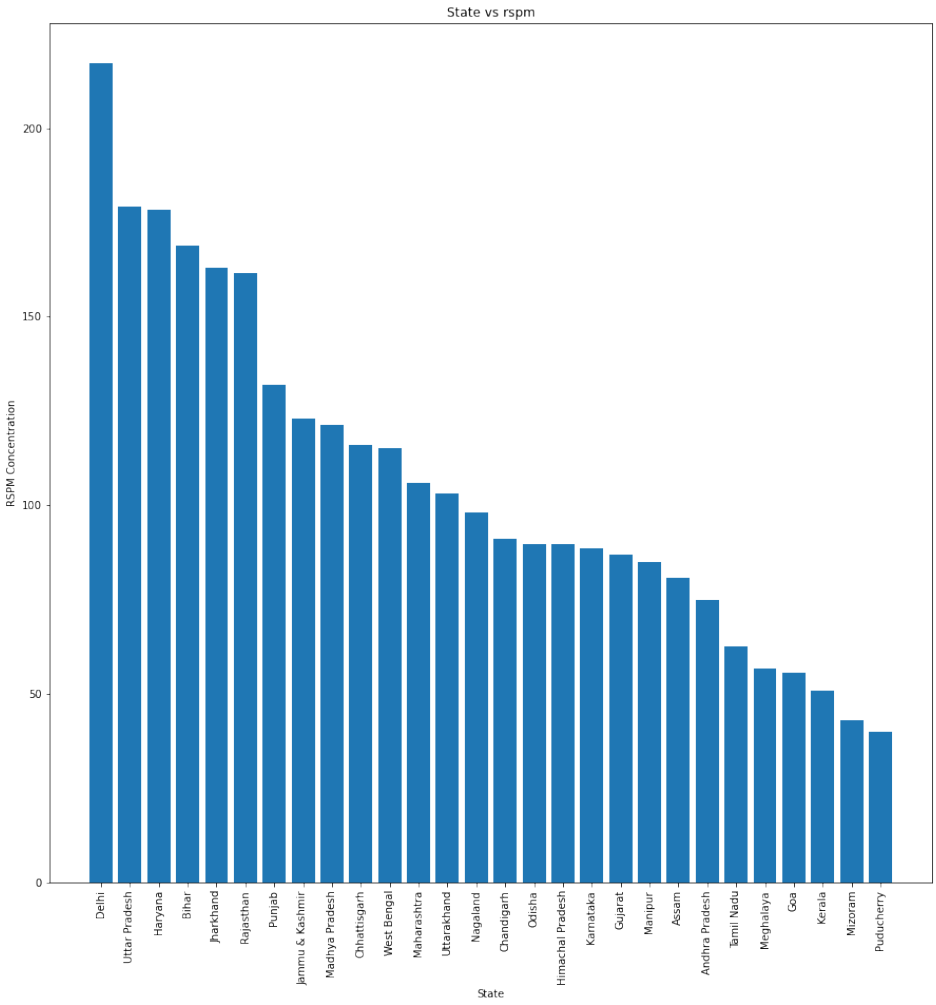

Figure 4.12: States vs their $RSPM$ Concentration for the year 2014

From the heatmap, it is evident that Delhi has high $RSPM$ concentration for all the 10 years. Bihar has witnessed sudden increase in $RSPM$ concentration after 2009. In Haryana, there was a sudden spike in $RSPM$ concentration in 2009 and for rest of the years it witnessed moderate to high concentration of $RSPM$. Also, Punjab, Uttarakhand and Puducherry have shown gradual decrease in $RSPM$ concentration.
From the bar plot for the year 2014, we can see that Delhi has the highest $RSPM$ levels thus proving Delhi to be the most polluted city when it comes to air pollution. Uttar Pradesh and Haryana also show nearly the same $RSPM$ concentration levels and are ranked at second and third position respectively. These three states are the most densely populated states, therefore there is an urgent need to take actions to bring the pollution level of these states in control. It can also be observed that Puducherry has the least $RSPM$ concentration.

4.5.4 Plots for Number of industries in states

:

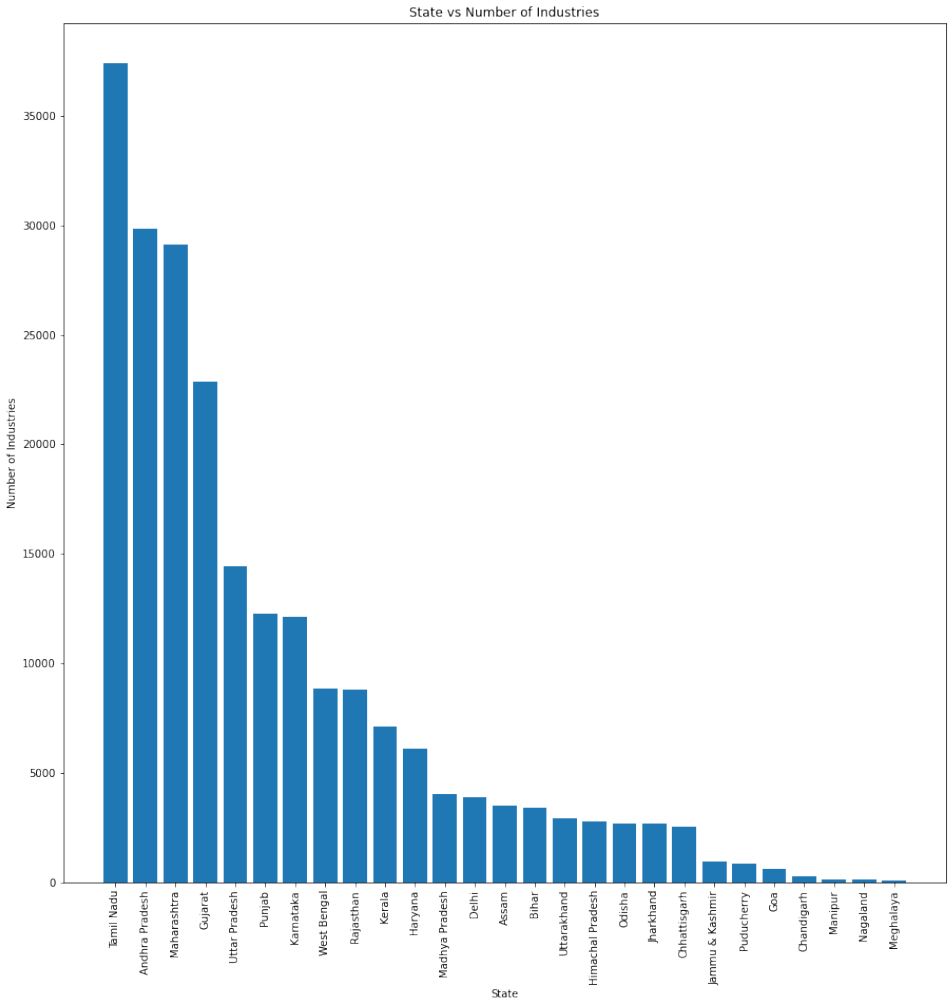

Figure 4.13: State vs number of Industries for the year 2014

4.5.5 Plots for number of vehicles in states

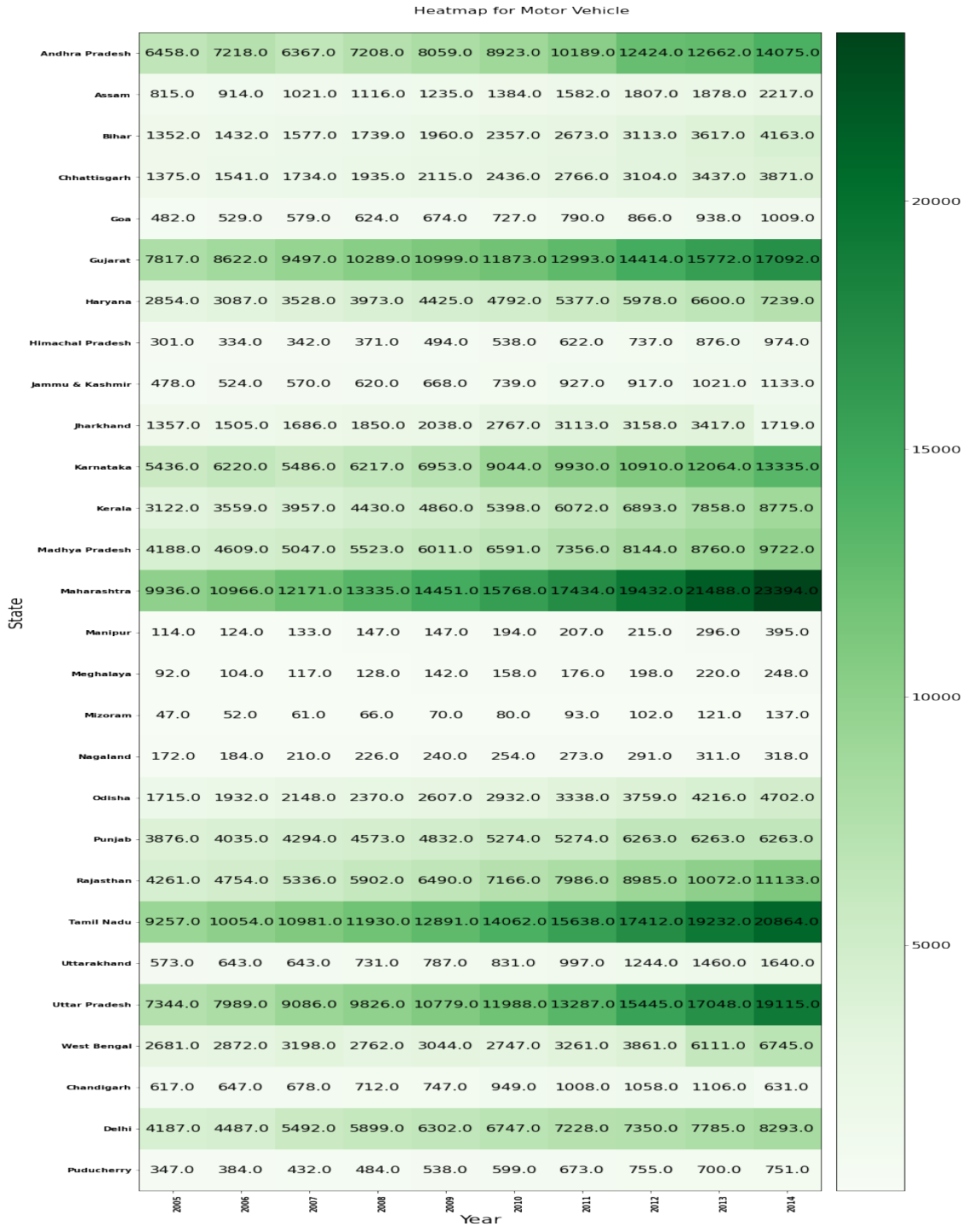

Figure 4.14: Heatmap for number of Motor Vehicle

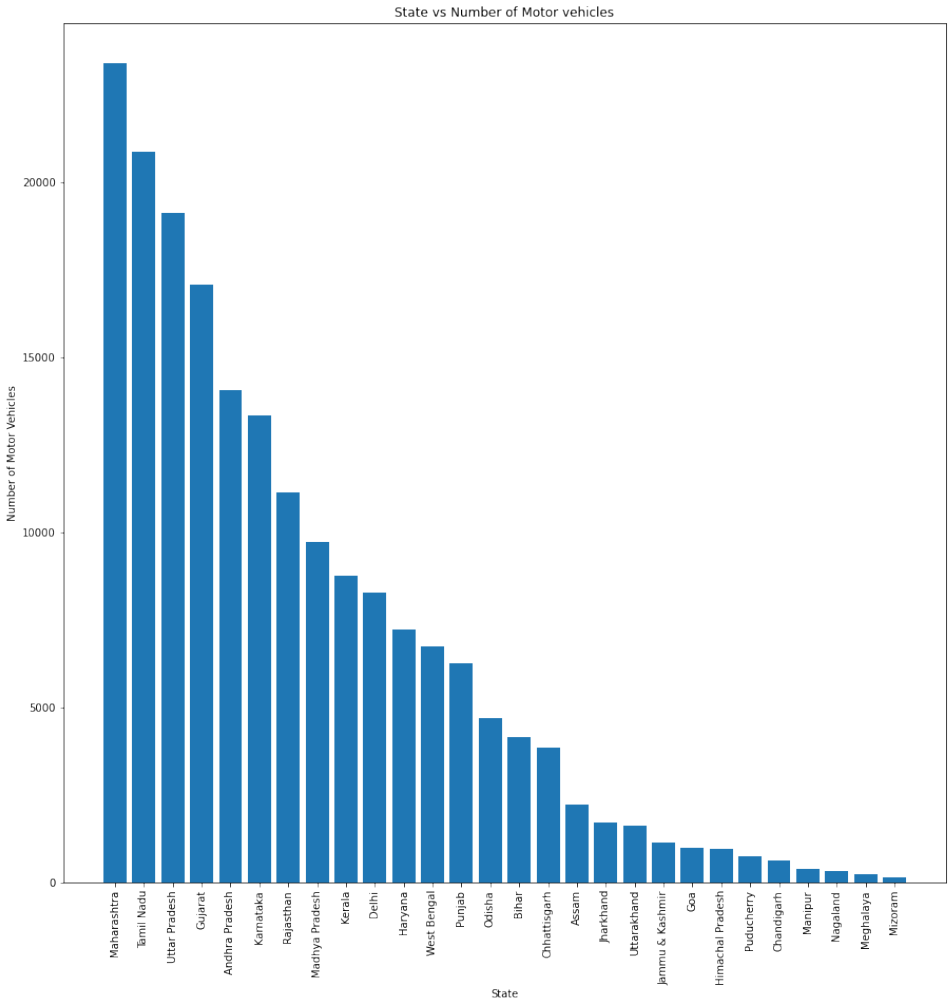

Figure 4.15: State vs number of Motor Vehicles for the year 2014

From the heatmap, we can say that over the span of ten years, the number of motor vehicles has increased by 3-4 times which certainly affects the air pollution.

4.5.6 Plots for population density of a state

:

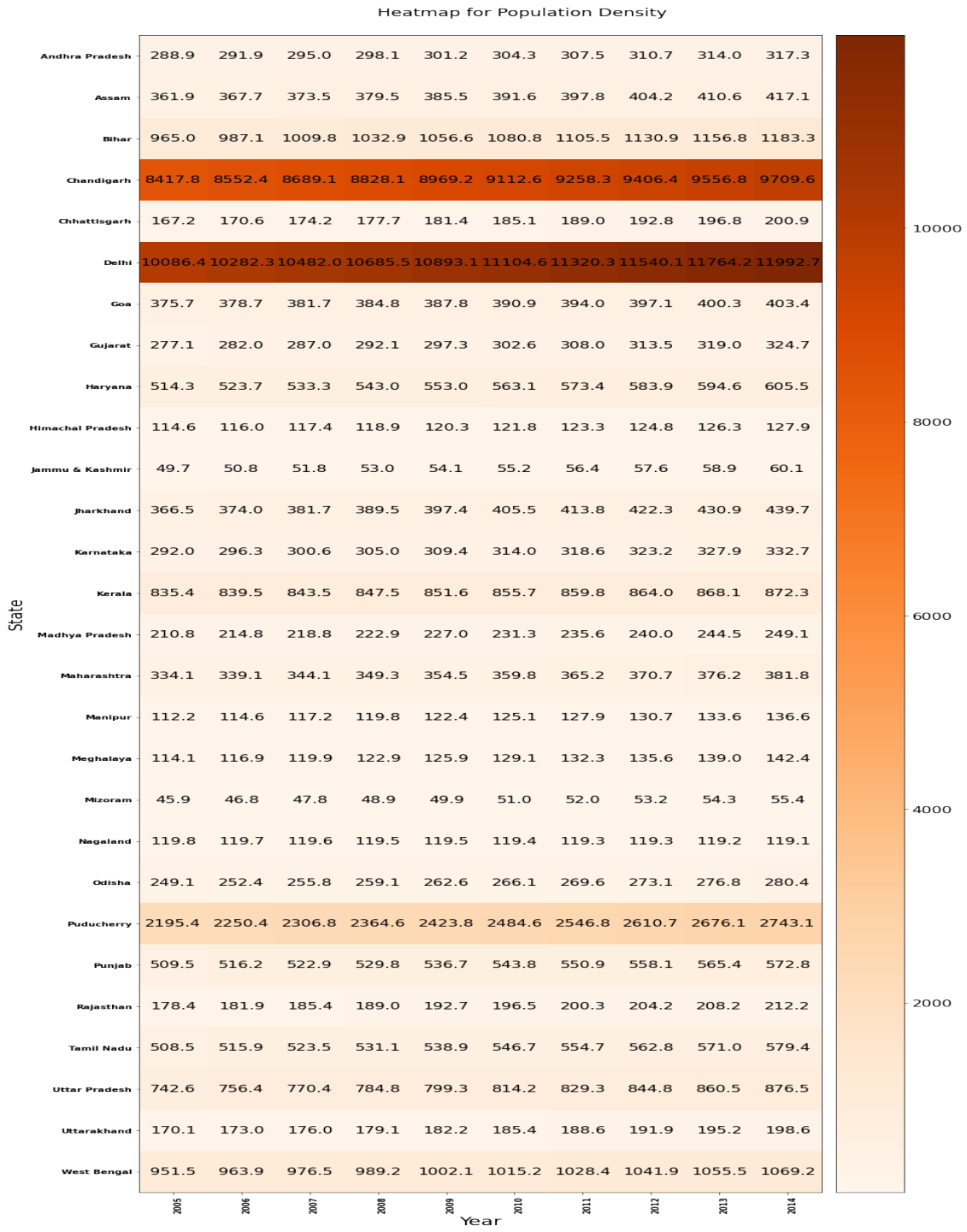

Figure 4.16: Heatmap for Population Density

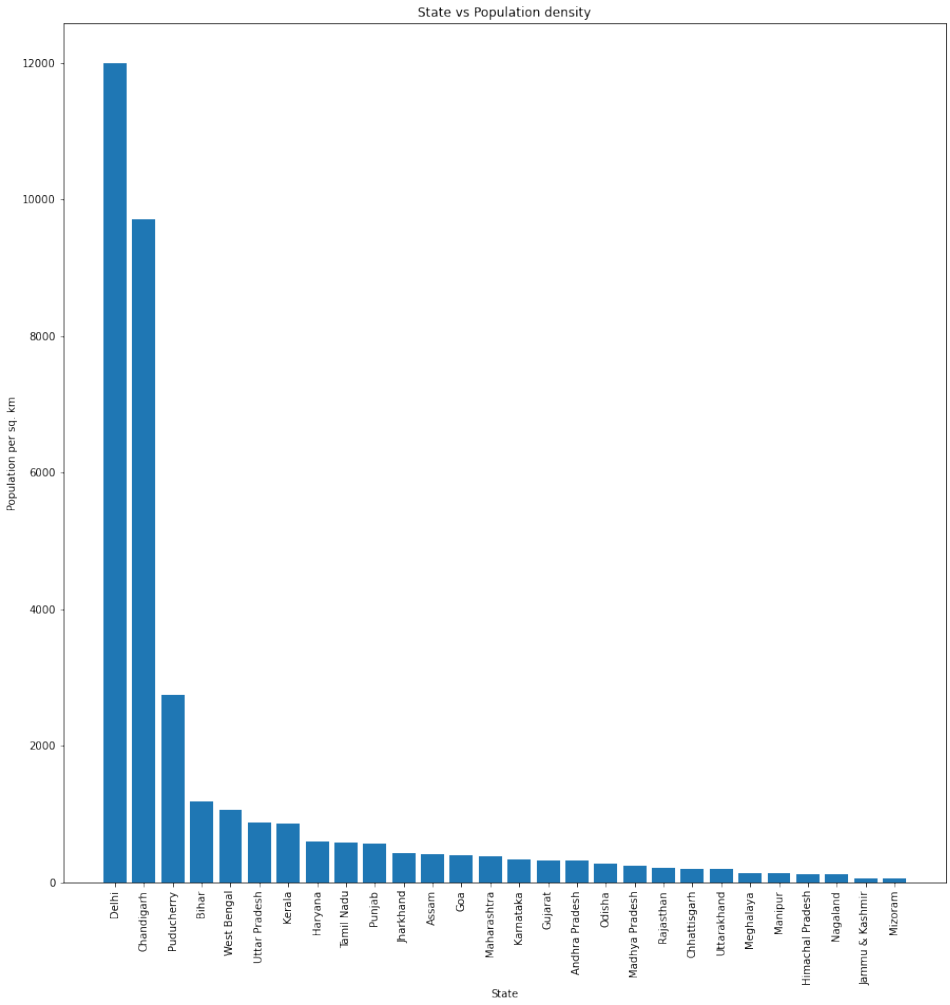

Figure 4.17: State vs Population Density for the year 2014

From the heatmap, it can be observed that the Population density shows gradual increase for all the states over the span of 10 years. Heatmap also depicts that Delhi has highest population density from 2005-2014.

4.6 Bonus section: Interactive Choropleth map for Pollutant concentrations for the year 2014

4.6.1 About the map

The map is available at https://deekshaarora05.github.io/Analysis-of-Air-Pollution-levels-in-India/Maps/2014mpc-map.html. It classifies the states as red (highly polluted), yellow (moderately polluted) and green (less polluted) for the year 2014. The classification is based on the following metric, where MPC is the Mean Pollutant Concentration that has been defined before:

- $MPC > 65$: A highly polluted state
- $45 >= MPC > 65$: A moderately polluted state
- $MPC < 45$: A less polluted state

The map is built using the Folium library in python, which is built on top of Javascript and HTML and uses Leaflet based maps. The map has a popup feature for each state which shows the top 5 most polluted cities of a state for the year 2014. A snapshot of the map is shown in Figure 4.18.

4.6.2 Observations

- Northern states of India - Uttar Pradesh, Delhi, Haryana, Rajasthan and Jharkhand were the most polluted states in 2014.
- The moderately polluted category mostly contains the Central Indian states - Madhya Pradesh, Chattisgarh along with West Bengal, Maharashtra and the two states in the extreme north - Punjab and Jammu & Kashmir.
- The least polluted category contains all the Southern Indian states and the Seven Sister states, along with the states - Gujarat and Himachal Pradesh.

This further adds to the fact that in general South India is less polluted than the north. The northern states around the Union Territory of Delhi are the most polluted states of India and the states in Central India lie somewhere in between them - in the moderately polluted category.

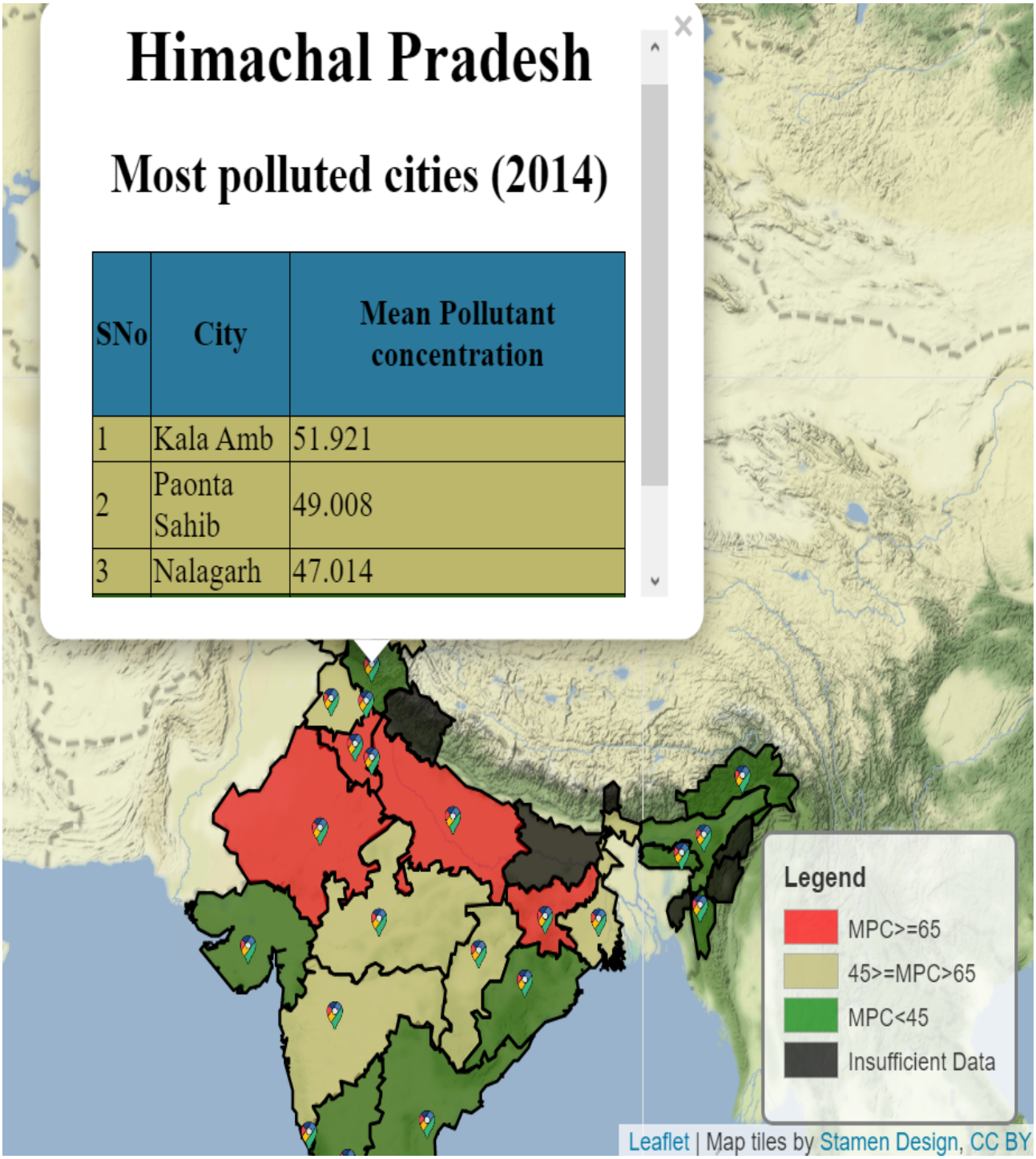

Figure 4.18: A snapshot of the MPC map for year 2014

Chapter 5

Results

After doing a thorough analysis of the pollutant concentrations and the factors affecting these concentrations, the following results can be derived:

- Delhi has the highest concentration of rspm, which again is not surprising as it can be read from any news article related to pollution in India over the past few years. When it comes to air pollution Delhi has been an important target. The increasing levels of pollutant concentration in Delhi has made a lot suffer and has been responsible for the deaths of thousands.

- It is observed that Uttar Pradesh(UP) is also not far away from Delhi in terms of pollutant concentrations. Being the most populated state in the country, it becomes more than necessary to deal with the rising level of pollution. Also, UP and Delhi are neighbouring states.

- In the initial years of analysis, Punjab has always made it to the top in terms of pollution but later it's pollution level decreased and in 2014 Punjab is categorized in less polluted states. The primary reason for air pollution in Punjab has been the burning of stubble by the farmers. This has been such a significant concern that the government has released so many policies and programs to prevent rice farmers from clearing their fields by burning the stubble that remains once paddy is harvested. It seems that the actions taken by the government played a significant role in controlling the pollution.

- The analysis also shows that the presence of the pollutant sulphur dioxide has been high from 2005 to 2008 in some states but has decreased later. Chandigarh, Daman Diu, Dadra Nagar Haveli are example of such states. To support our analysis we found a news article which goes as follows:
  Data released by NASA's Aura satellite calls into question the veracity of Central Pollution Control Board's (CPCB) claim made in 2012 that the mean sulphur dioxide (SO2) emissions in India decreased in 2010 as compared to 2001 level.
  However, some states like Uttarakhand, Maharashtra, Gujarat, etc. still experience considerably high levels of SO_2 concentration.

- Heatmaps indicate the alarming levels of $RSPM$ in West Bengal and the government needs to take urgent actions to control the increasing level of RSPM.

Chapter 6

Conclusion and Future Work

Air pollution is turning out to be 'the challenge' of the century. Not just in India, but countries all over the world are scrambling for laws and reforms that would help keep air pollution in check. In a November 2019 ruling, the Supreme Court of India criticised the federal and state governments over the sorry state of pollution in the capital. The same saga has continued every year in the Northern part of the country, especially during the winters, when the heavy air settles down and along with it, the pollutants. States like West Bengal have seen more deaths attributed to poor air quality than any other.

According to a WHO report 8 out of 10 most polluted cities in the world are from India. The situation in the country demands stringent laws to be introduced and implemented with the utmost priority. This isn't an impossible task by itself but it requires contribution from each and every corner of the society. From the smallest acts like preferring the use of public transport over private ones to conserving energy by switching off air conditioners and lights when not in use, **we can make a difference**. The government must also hold industries accountable for the damage they are doing to the air quality. Crop burning is another practice that is ancient and needs to be replaced with more nuanced approaches. The vehicular emissions need to be capped. Adherence to protocols can't be sidelined anymore and the government has to ensure this. Programmmes like NCAP (National Clean Air Programme) are a step in the right direction. It is a National level programme meant to set up and achieve reduction goals with respect to particulate matters every few years. Our fate, as a country, as far as air pollution is concerned is still in our hands and we need to act quick or else prepare to suffer the consequences.

The project can be extended by analysing the air pollution trends for 2015 to 2020 and for future years as well. We can predict the future levels of air pollutant concentrations by incorporating Machine learning models to our analysis.