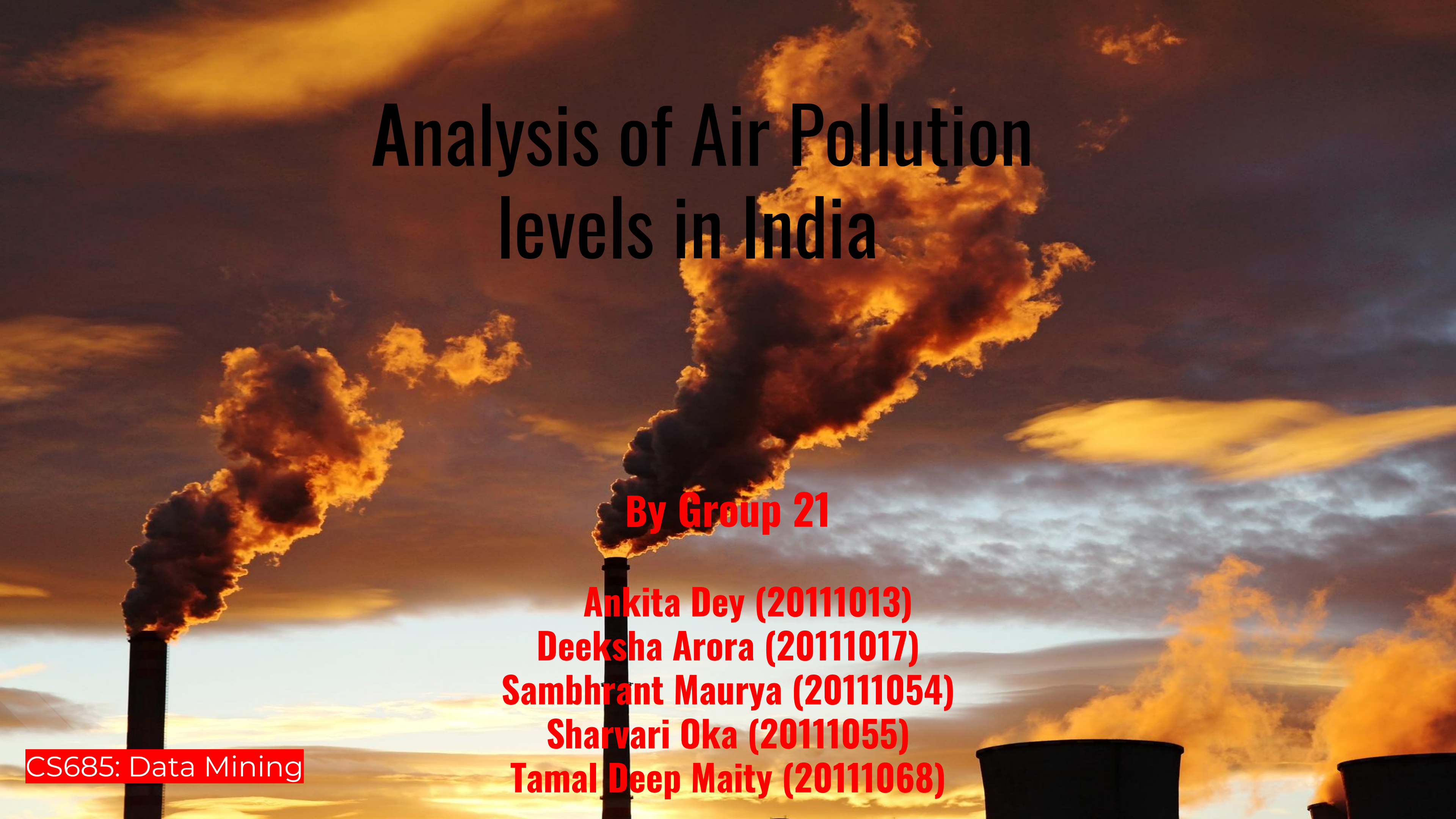


Analysis of Air Pollution levels in India

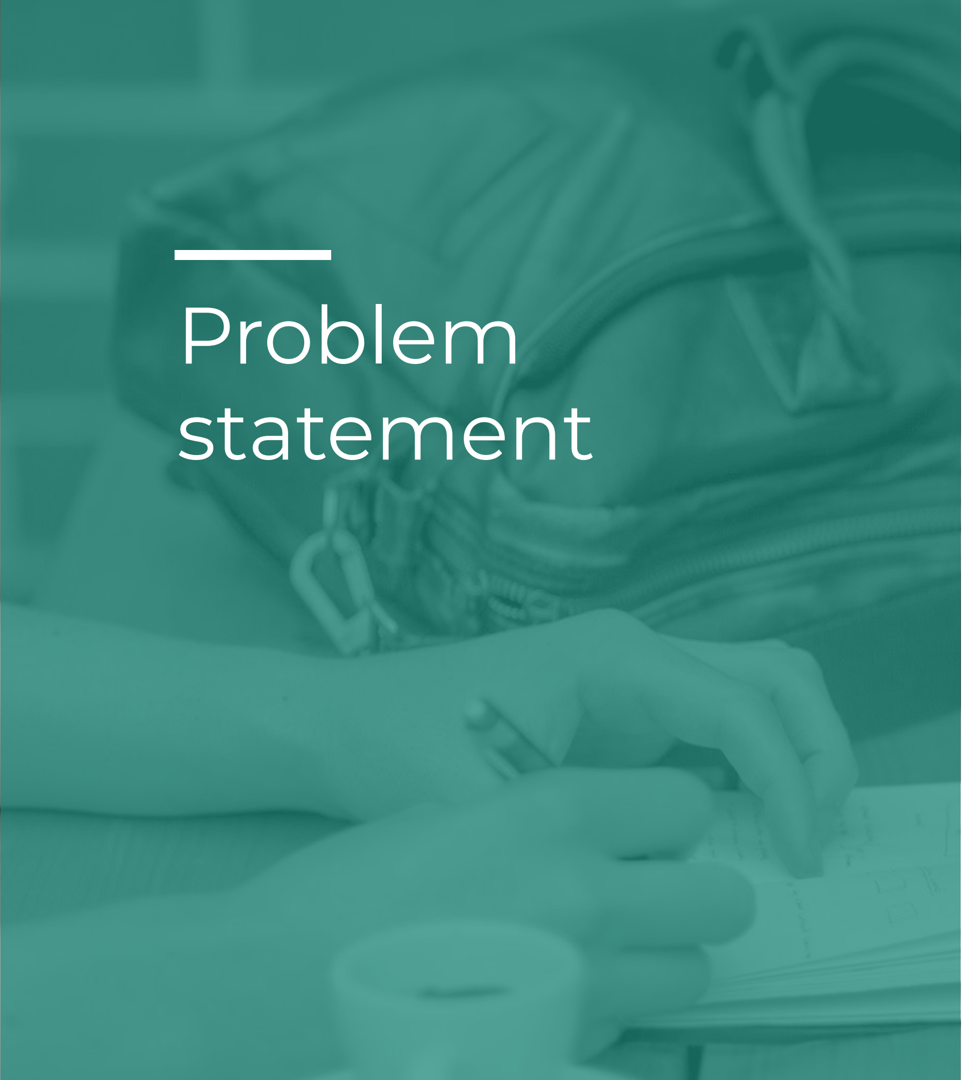
By Group 21

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

CS685: Data Mining

Outline

- **Problem Statement**
- **Datasets**
- **Analysis tools and Observations**
    - **Correlation**
    - **Heatmaps for air pollutant concentration**
    - **Z-Score Analysis for Hotspot Detection**
    - **Chi-Squared Test**
    - **Choropleth Map for Most polluted Cities using MPC**
    - **Clustering**
- **Results**



Problem statement

- ❑ To analyze air pollution trends in various states/UTs of India from 2005-2014.

- ❑ To find the most polluted states/UTs in India with respect to SO_2, NO_2 and RSPM concentrations.

- ❑ To understand the correlation between concentration of SO_2, NO_2 and RSPM with number of motor vehicles, industries and population density of the states

- ❑ To visualize the states/UTs as hotspots and coldspots on the basis of chi-score and z-score

- ❑ To cluster the states/UTs based on their pollution levels

Data Sets

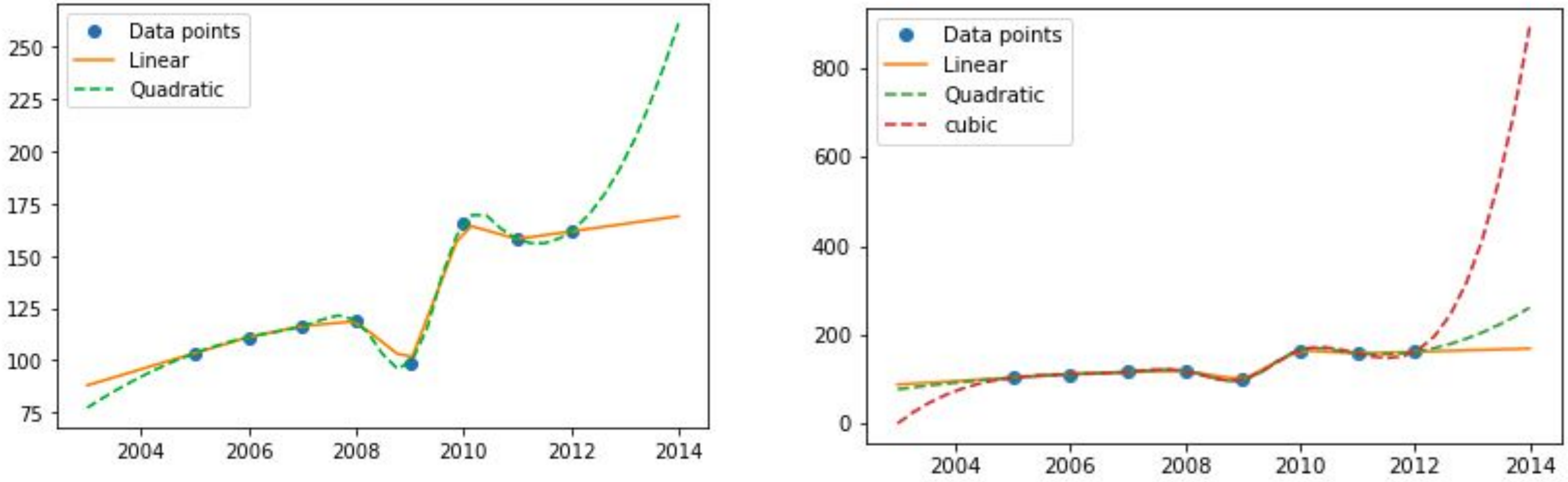
SO_2, NO_2 and RSPM statistics for all states and cities of India	**Link:** https://www.kaggle.com/shrutibhargava94/india-air-quality-data **Contents:** SO_2, NO_2, RSPM, SPM and PM 2.5 for all the states of India from 1990 to 2015 (csv)
State wise Motor Vehicle statistics	**Link:** http://mospi.nic.in/sites/default/files/statistical_year_book_india_2015/Table-20.4_0. **Contents:** Total registered motor vehicles for each state from 2001 to 2015 (xlsx)
State Wise number of Industries	**Links:** 2008-2014 data (xlsx)- https://www.mospi.gov.in/sites/default/files/statistical_year_book_india_2015/Table%2014.1_1.xlsx 2001-2006 data (docx)- http://labourbureau.gov.in/ASI_V2_2005_06_TAB27F.docx 2007-08 data (pdf)- http://labourbureaunew.gov.in/UserContent/ASI_Vol_1_2007_08.pdf
Census data of india	**Links:** 2001 census (html)- https://censusindia.gov.in/Census_Data_2001/Census_data_finder/A_Series/Total_population.htm 2011 census (xls)- http://mospi.nic.in/sites/default/files/statistical_year_book_india_2015/Table%20

Data Sets (contd.)

Geographical coordinates of borders of all States/UTs of India	**Link:** http://download.oracle.com/otn/samplecode/data-visualization/sample-MapLayers_Maps/INDIA_STATES_051120.zip **Contents:** A zipped json file containing the coordinates of all state borders of India
All Indian states and their neighbors	Built manually as a csv

Extrapolation of missing data

Only linear extrapolation gave values within the range of the dataset!



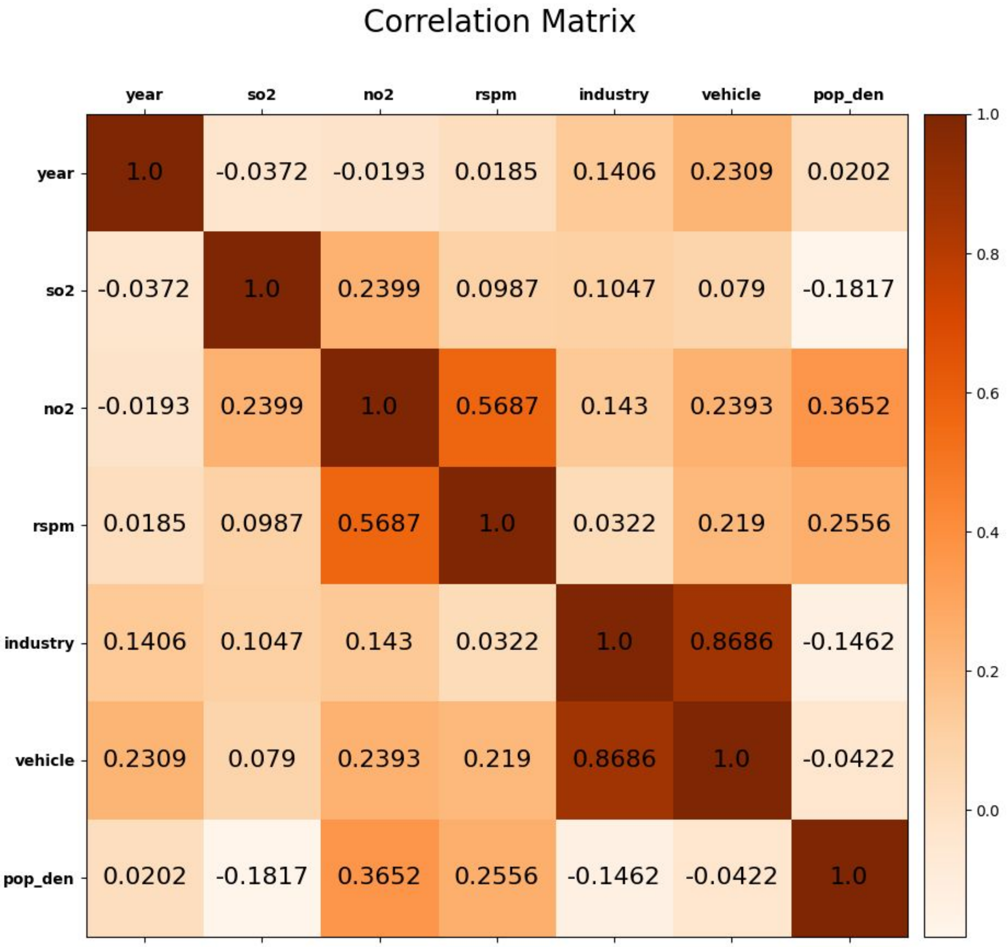
Plots of RSPM values for Bihar

Pearson’s Correlation

- Only works when features have linear relationship

- Very strong correlation between number of industries and vehicles.

- Correlation of [SO_2, NO_2, RSPM] vs [Industry, Vehicle] surprisingly low.

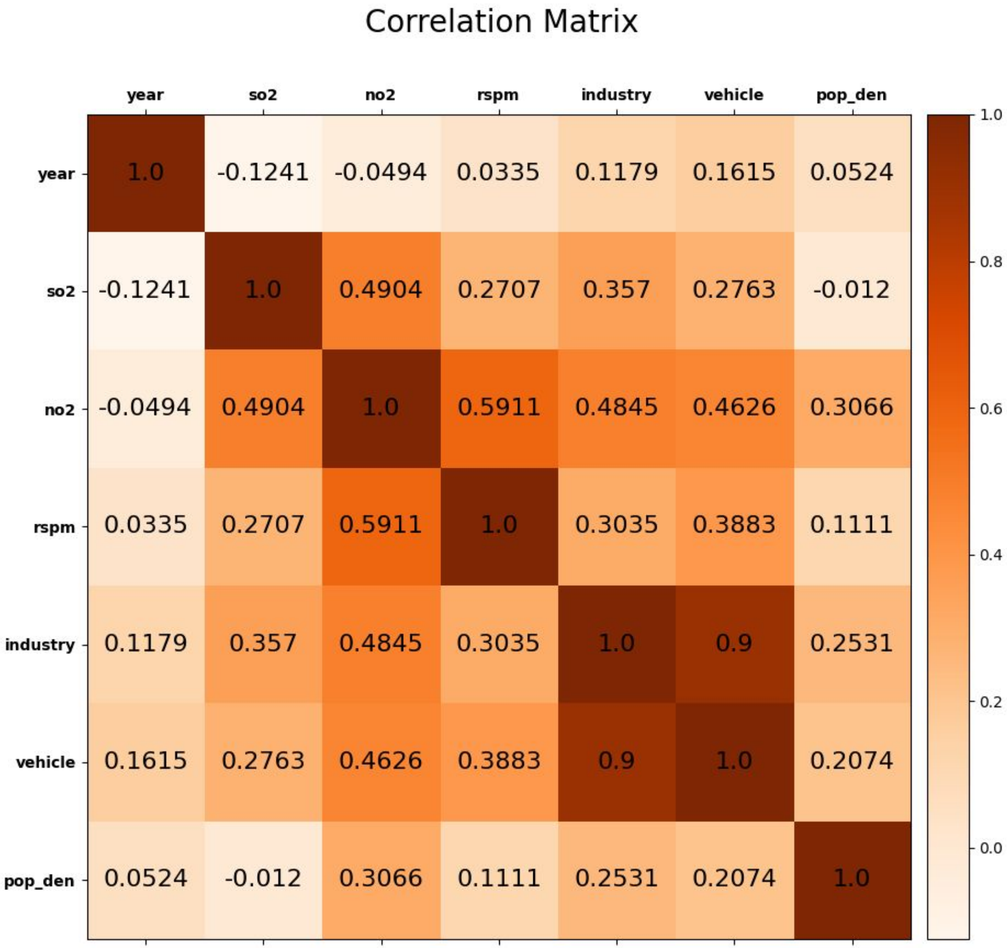


Correlation Matrix

	year	so2	no2	rspm	industry	vehicle	pop_den
year	1.0	-0.0372	-0.0193	0.0185	0.1406	0.2309	0.0202
so2	-0.0372	1.0	0.2399	0.0987	0.1047	0.079	-0.1817
no2	-0.0193	0.2399	1.0	0.5687	0.143	0.2393	0.3652
rspm	0.0185	0.0987	0.5687	1.0	0.0322	0.219	0.2556
industry	0.1406	0.1047	0.143	0.0322	1.0	0.8686	-0.1462
vehicle	0.2309	0.079	0.2393	0.219	0.8686	1.0	-0.0422
pop_den	0.0202	-0.1817	0.3652	0.2556	-0.1462	-0.0422	1.0

Spearman’s Correlation

- More generic than Pearson.
- Correlation of [SO_2, NO_2, RSPM] vs [Industry, Vehicle] also seems natural.
- Very high correlation(0.9) between no. of vehicles and no. of industries.
- Strong correlation observed between :
  - RSPM and NO_2 levels
  - SO_2 and NO_2 levels
  - NO_2 level and number of vehicles



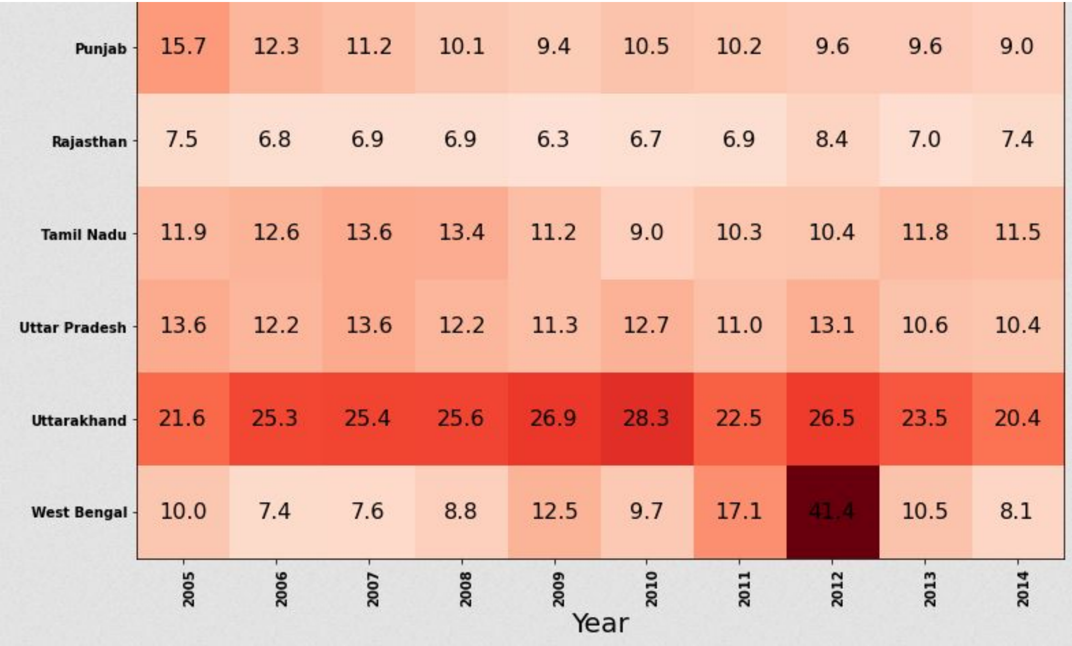
Correlation Matrix

	year	so2	no2	rspm	industry	vehicle	pop_den
year	1.0	-0.1241	-0.0494	0.0335	0.1179	0.1615	0.0524
so2	-0.1241	1.0	0.4904	0.2707	0.357	0.2763	-0.012
no2	-0.0494	0.4904	1.0	0.5911	0.4845	0.4626	0.3066
rspm	0.0335	0.2707	0.5911	1.0	0.3035	0.3883	0.1111
industry	0.1179	0.357	0.4845	0.3035	1.0	0.9	0.2531
vehicle	0.1615	0.2763	0.4626	0.3883	0.9	1.0	0.2074
pop_den	0.0524	-0.012	0.3066	0.1111	0.2531	0.2074	1.0

Heatmap of SO_2 concentration for some states

- For most states SO_2 levels haven't showed increasing trend other than UK.

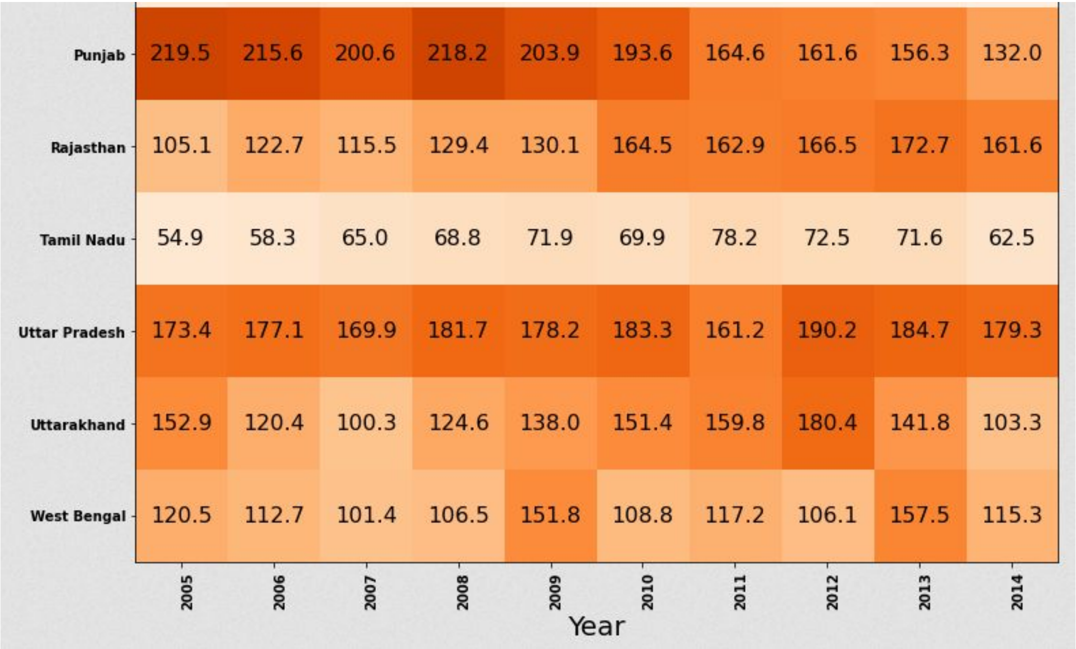
- Most prominent reason for SO_2 emission : power generation by burning fossil fuel, followed by vehicular emission.



	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Punjab	15.7	12.3	11.2	10.1	9.4	10.5	10.2	9.6	9.6	9.0
Rajasthan	7.5	6.8	6.9	6.9	6.3	6.7	6.9	8.4	7.0	7.4
Tamil Nadu	11.9	12.6	13.6	13.4	11.2	9.0	10.3	10.4	11.8	11.5
Uttar Pradesh	13.6	12.2	13.6	12.2	11.3	12.7	11.0	13.1	10.6	10.4
Uttarakhand	21.6	25.3	25.4	25.6	26.9	28.3	22.5	26.5	23.5	20.4
West Bengal	10.0	7.4	7.6	8.8	12.5	9.7	17.1	41.4	10.5	8.1

Heatmap of RSPM concentration for some states

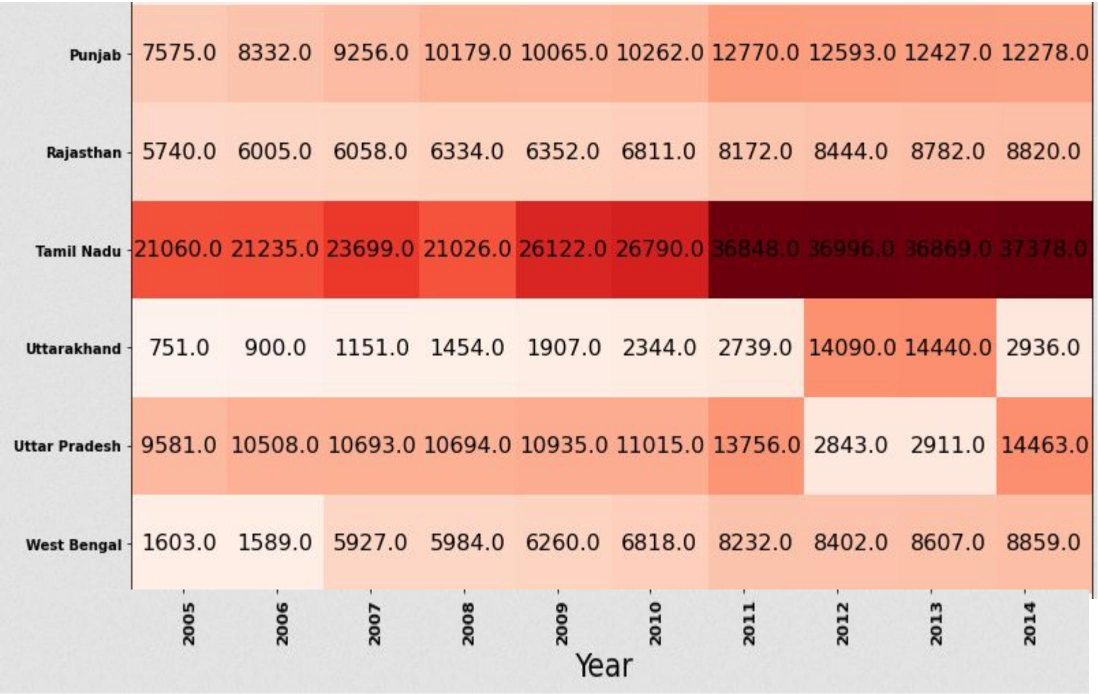
- Punjab's RSPM levels have seen a constant decline. (some state laws?)
- Levels have increased for states like Rajasthan.
- More or less constant throughout period of analysis in UP.



	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Punjab	219.5	215.6	200.6	218.2	203.9	193.6	164.6	161.6	156.3	132.0
Rajasthan	105.1	122.7	115.5	129.4	130.1	164.5	162.9	166.5	172.7	161.6
Tamil Nadu	54.9	58.3	65.0	68.8	71.9	69.9	78.2	72.5	71.6	62.5
Uttar Pradesh	173.4	177.1	169.9	181.7	178.2	183.3	161.2	190.2	184.7	179.3
Uttarakhand	152.9	120.4	100.3	124.6	138.0	151.4	159.8	180.4	141.8	103.3
West Bengal	120.5	112.7	101.4	106.5	151.8	108.8	117.2	106.1	157.5	115.3

Heatmap of industries for some states

- More and more industries setting up in Tamil nadu. Same for UP, Punjab, WB barring a few years.
- Most of these states affected (in terms of air quality).
- Expected as well by looking at correlation matrix.



	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Punjab	7575.0	8332.0	9256.0	10179.0	10065.0	10262.0	12770.0	12593.0	12427.0	12278.0
Rajasthan	5740.0	6005.0	6058.0	6334.0	6352.0	6811.0	8172.0	8444.0	8782.0	8820.0
Tamil Nadu	21060.0	21235.0	23699.0	21026.0	26122.0	26790.0	36848.0	36996.0	36869.0	37378.0
Uttarakhand	751.0	900.0	1151.0	1454.0	1907.0	2344.0	2739.0	14090.0	14440.0	2936.0
Uttar Pradesh	9581.0	10508.0	10693.0	10694.0	10935.0	11015.0	13756.0	2843.0	2911.0	14463.0
West Bengal	1603.0	1589.0	5927.0	5984.0	6260.0	6818.0	8232.0	8402.0	8607.0	8859.0

Hotspots using Z-Score

- To find Z-Score, the Mean Pollutant Concentration (MPC) for each state is computed, given by:

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

Since it's possible that air pollution of a state is also affected by the increasing air pollution level of it's neighboring states, we have defined a state as hotspot or coldspot taking into consideration the pollution level of its neighbors.

- A state is Hotspot if : $\text{MPC}_{\text{state}} > \text{Mean}_{\text{neighbor}} + ½ \text{ std}_{\text{neighbor}}$
- A state is Coldspot if : $\text{MPC}_{\text{state}} < \text{Mean}_{\text{neighbor}} - ½ \text{ std}_{\text{neighbor}}$

Choropleth Map to visualize the results of Z-Score

https://maurya-bitlegacy.github.io/codename-caeli/Maps/zscore-map.html

Analysis of Results obtained using Z-Score

- Major Hotspots: Delhi, Maharashtra, Jharkhand, Manipur, Rajasthan

- Nagaland and Delhi, both are hotspots, but reasons may vary. Delhi's pollution levels can be attributed to heavy vehicular emissions, dust from construction sites, etc. Nagaland's pollution levels are much lower, still hotspot as level of pollutant concentration is more than neighbors.

- Jharkhand, a hotspot : Possible reasons: Virginity of rural areas lost to industrialization. This state is home to most of India's coal mines which emit pollutants due to coal burning.

- Similarly, problems of Maharashtra (a hotspot) possibly revolve around having huge industry-vehicle concentration.

- Major Coldspots: Kerala, Arunachal Pradesh, Himachal Pradesh, Sikkim, Odisha. These state not only have low pollutant concentration as compared to their neighbors but in general as well.

Pollutant Concentration Analysis using Chi-Score

- Used Chi-Squared Test to detect outliers.

- Formula used:

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

Where, o : object is to be tested
o_i : value of o in ith dimension
E_i : mean value on ith dimension among all objects

- H_0 : State is not an outlier
  H_1 : State is an outlier

- A state is considered as an outlier if it's p-value is less than level of significance (1%).

Choropleth Map to visualize the results of Chi-Squared Test for the year 2014

https://maurya-bitlegacy.github.io/codename-caeli/Maps/chiscore-map.html

Limitation of Chi-Squared Test for outlier detection

- Tells whether a state is an outlier (high or low pollutant concentration) or not but doesn’t specify the nature of outliers.

- Therefore, we categorised the states as more polluted or less polluted based on Mean Pollutant Concentration levels.

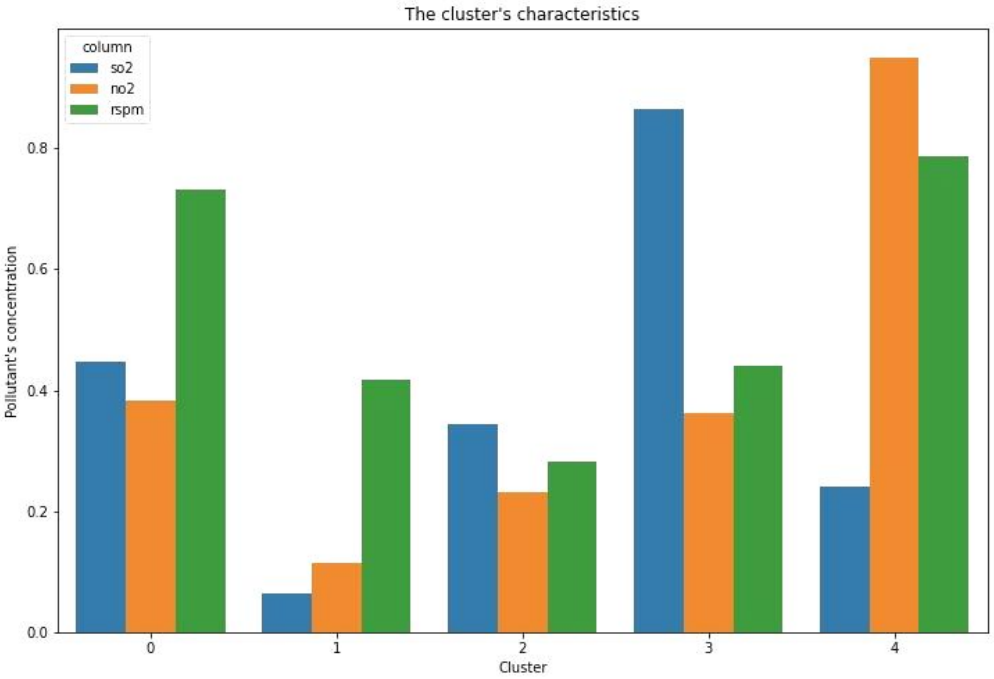
Pollution levels in 2014 and top 5 most polluted cities for every state

- Severely Polluted : MPC >=65
- Moderately Polluted : 45<= MPC <65
- Less Polluted: MPC<45

Choropleth Map for Visualization:
https://maurya-bitlegacy.github.io/codename-caeli/Maps/2014mpc-map.html

Clustering

- To group similar states based on concentrations of SO_2 , NO_2 and RSPM
- K-means Algorithm is used with 5 clusters (Got elbow point at K=5)



Interpretation of results obtained from Clustering

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

- High pollution level in Central part of India, especially RSPM.

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

- Low SO_2 and NO_2 levels and moderate level of RSPM in the hilly regions of India.

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

- Low levels of SO_2 , NO_2 and RSPM. States in Southern part of India are generally less polluted than Northern states.

Cluster 3: Gujarat, Maharashtra, Sikkim, Uttarakhand

- States with high SO_2 concentration.

Cluster 4: Delhi, West Bengal

- States with very high NO_2 and RSPM concentration.

Results: Summarized

- Our analysis shows that Delhi has the highest concentration of RSPM in the country, which again is not surprising as it can be read from any news article related to pollution in India over the past few years. RSPM levels in Delhi have almost increased continuously, and this increasing levels of RSPM in Delhi has been responsible for the deaths of thousands.

- Pollutant levels in states is affected by the number of vehicles and number of Industries. The seven sister states which are among the least polluted states in the country also have the least number of vehicles and factories in the country.

- Pollution levels are surprisingly not correlated to the population density of states. For example, West Bengal and Kerala have similar population densities, but Kerala came off as one of the least polluted states in the country while West Bengal came off as one of the most polluted states in the country.

- The analysis also shows that pollutant levels do not follow a continuously increasing or decreasing trend in many states. For instance, 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. In the initial years of analysis, Punjab has always made it to the top in terms of RSPM but later it's RSPM levels decreased and in 2014, Punjab has been categorized among the less polluted states.

Thank You