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

Outline

- Problem Statement

Outline

- Problem Statement
- Datasets

Outline

- Problem Statement
- Datasets
- Analysis tools and Observations

Outline

- Problem Statement
- Datasets
- Analysis tools and Observations
  - Chi-Squared Test

Outline

- Problem Statement
- Datasets
- Analysis tools and Observations
  - Chi-Squared Test
  - Z-Score Analysis for Hotspot Detection

Outline

- Problem Statement
- Datasets
- Analysis tools and Observations
  - Chi-Squared Test
  - Z-Score Analysis for Hotspot Detection
  - Choropleth Map for Most polluted Cities using MPC

Outline

- Problem Statement
- Datasets
- Analysis tools and Observations
  - Chi-Squared Test
  - Z-Score Analysis for Hotspot Detection
  - Choropleth Map for Most polluted Cities using MPC
  - Clustering

Outline

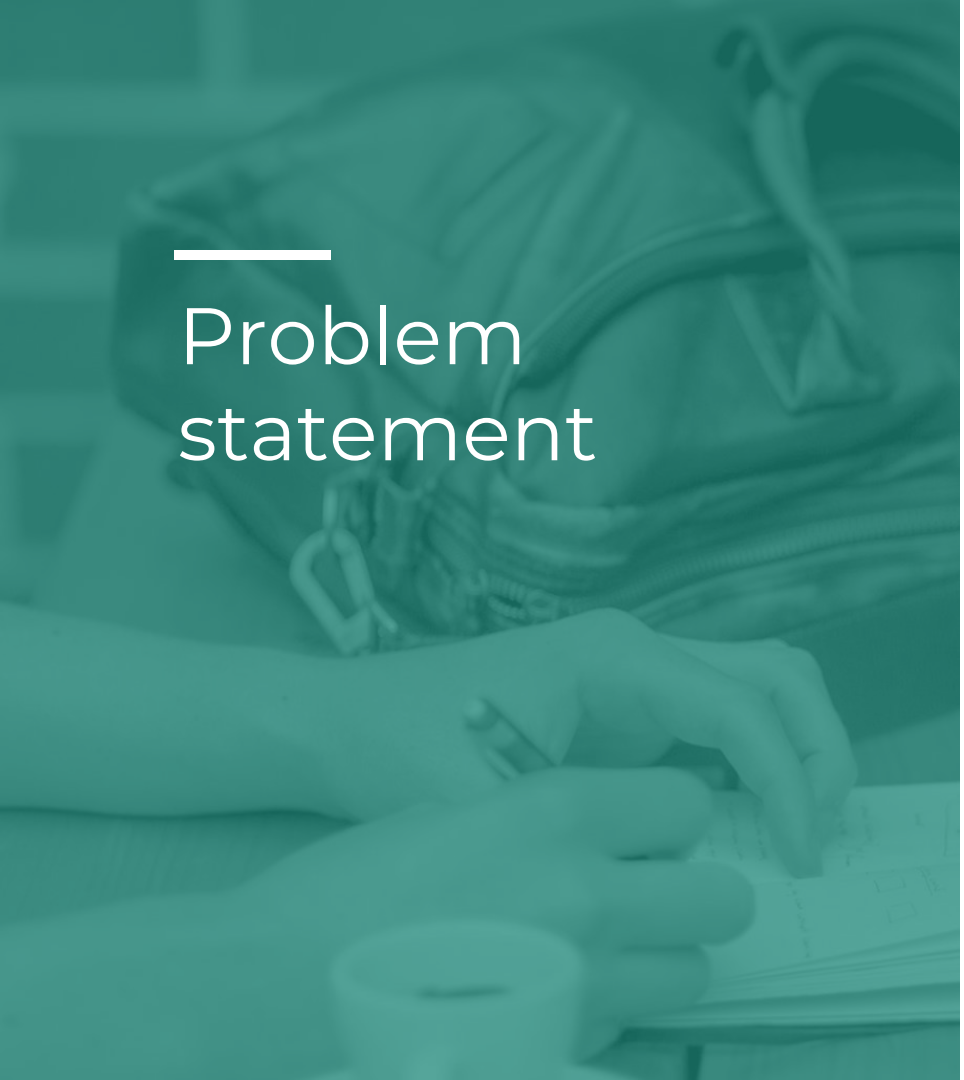
- Problem Statement
- Datasets
- Analysis tools and Observations
  - Chi-Squared Test
  - Z-Score Analysis for Hotspot Detection
  - Choropleth Map for Most polluted Cities using MPC
  - Clustering
  - Correlation

Outline

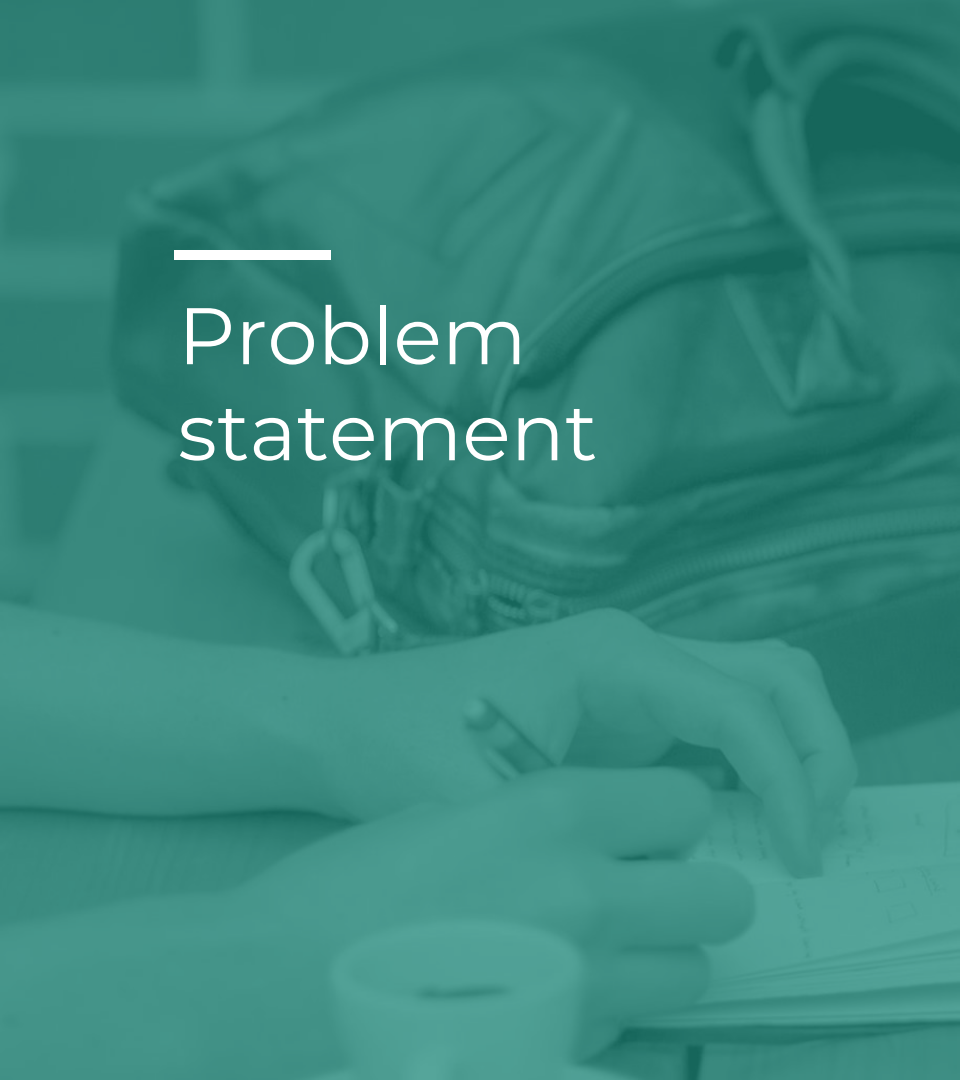
- Problem Statement
- Datasets
- Analysis tools and Observations
  - Chi-Squared Test
  - Z-Score Analysis for Hotspot Detection
  - Choropleth Map for Most polluted Cities using MPC
  - Clustering
  - Correlation
  - Heatmaps for air pollutant concentrationResults

Outline

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

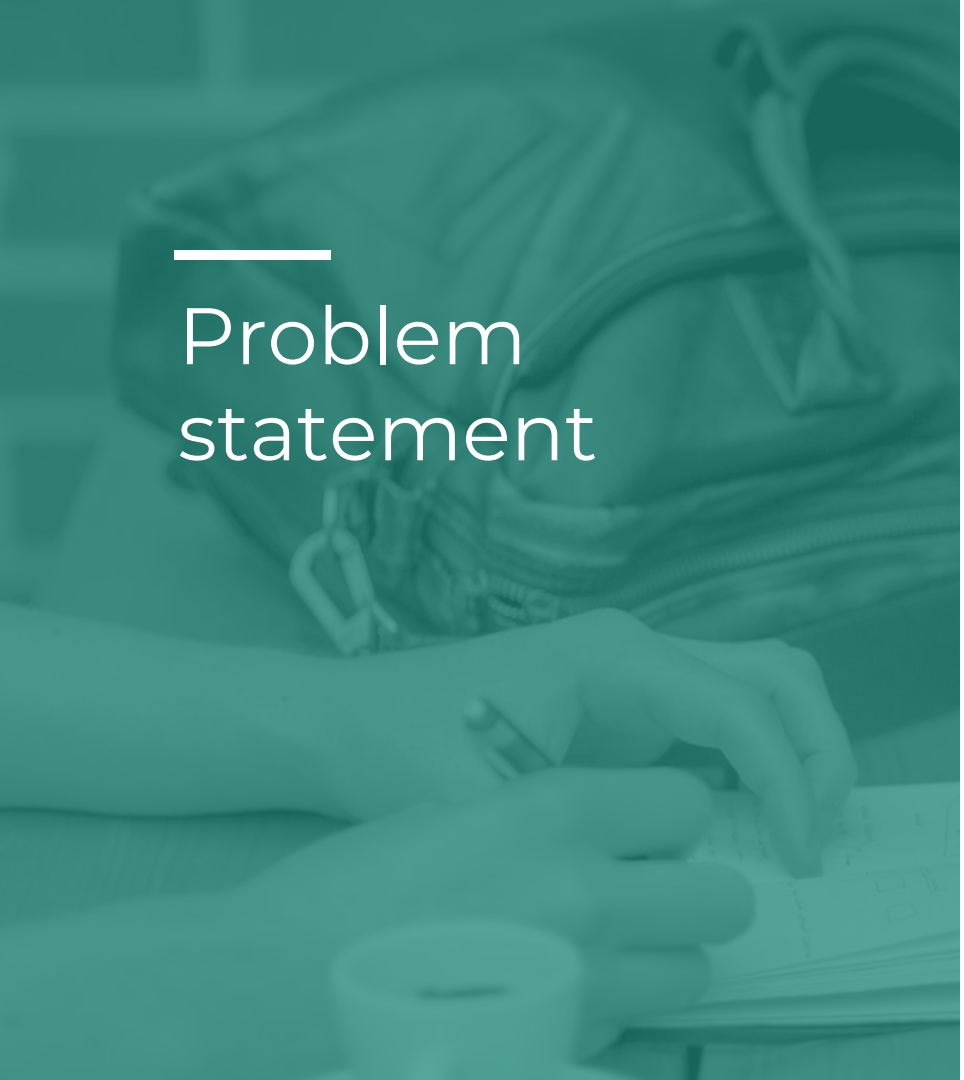


Problem statement



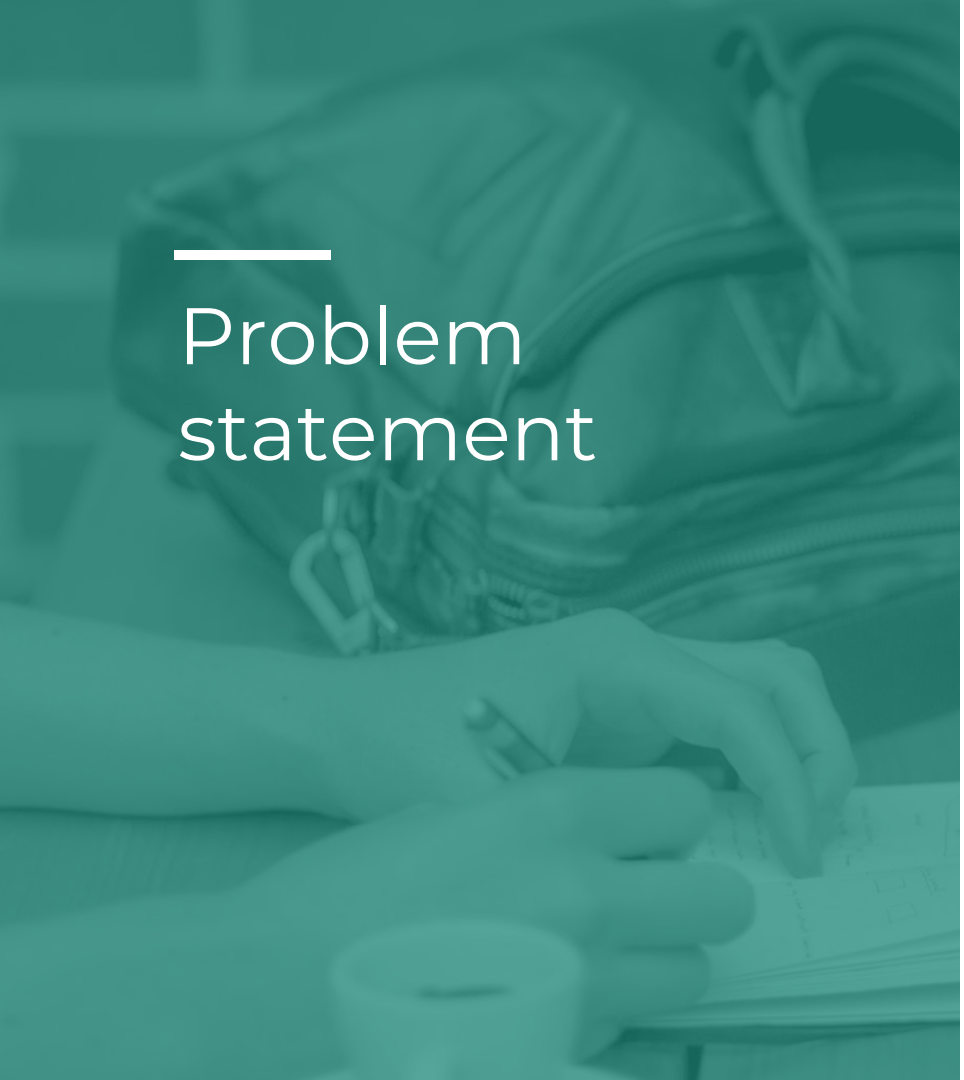
Problem statement

- To analyze air pollution trends in various states in India from 2005-2014.



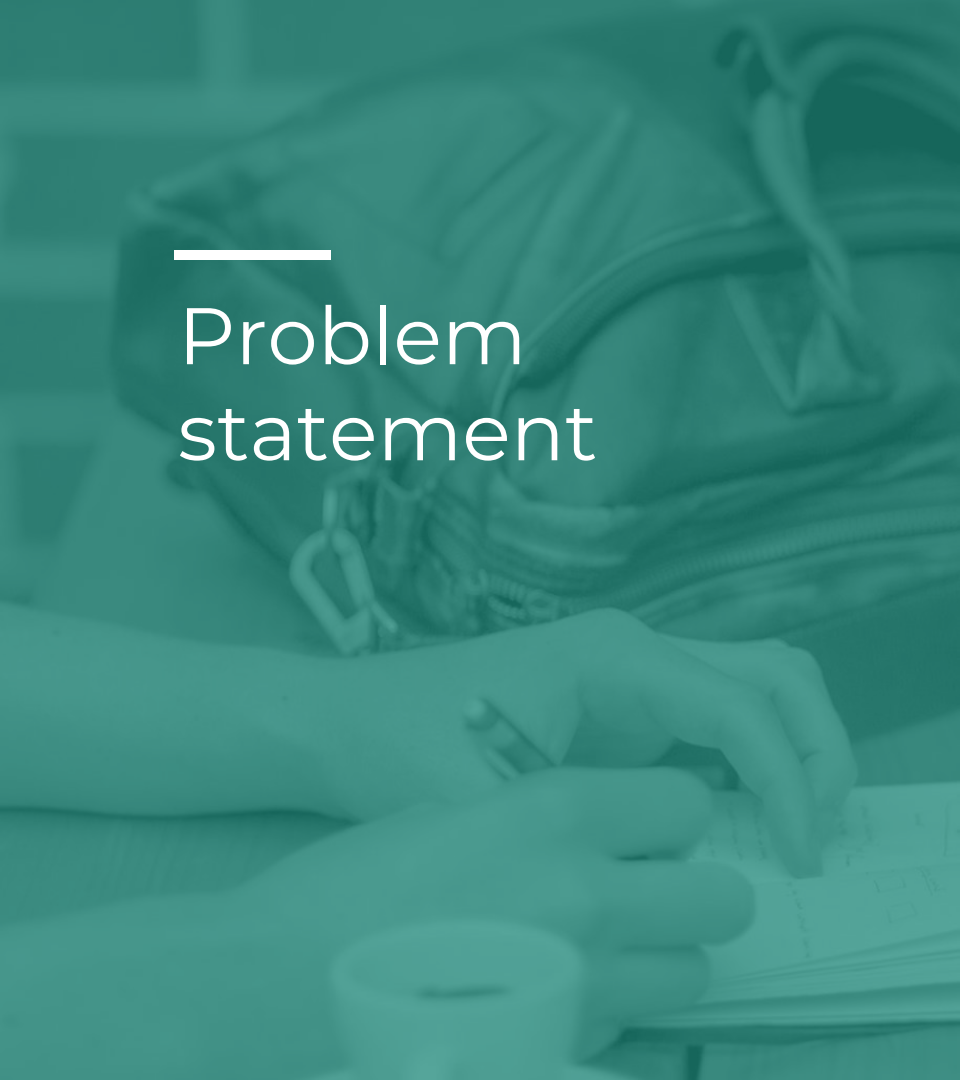
Problem statement

- To analyze air pollution trends in various states in India from 2005-2014.
- To find the most polluted states in India with respect to SO_2, NO_2 and RSPM concentrations.



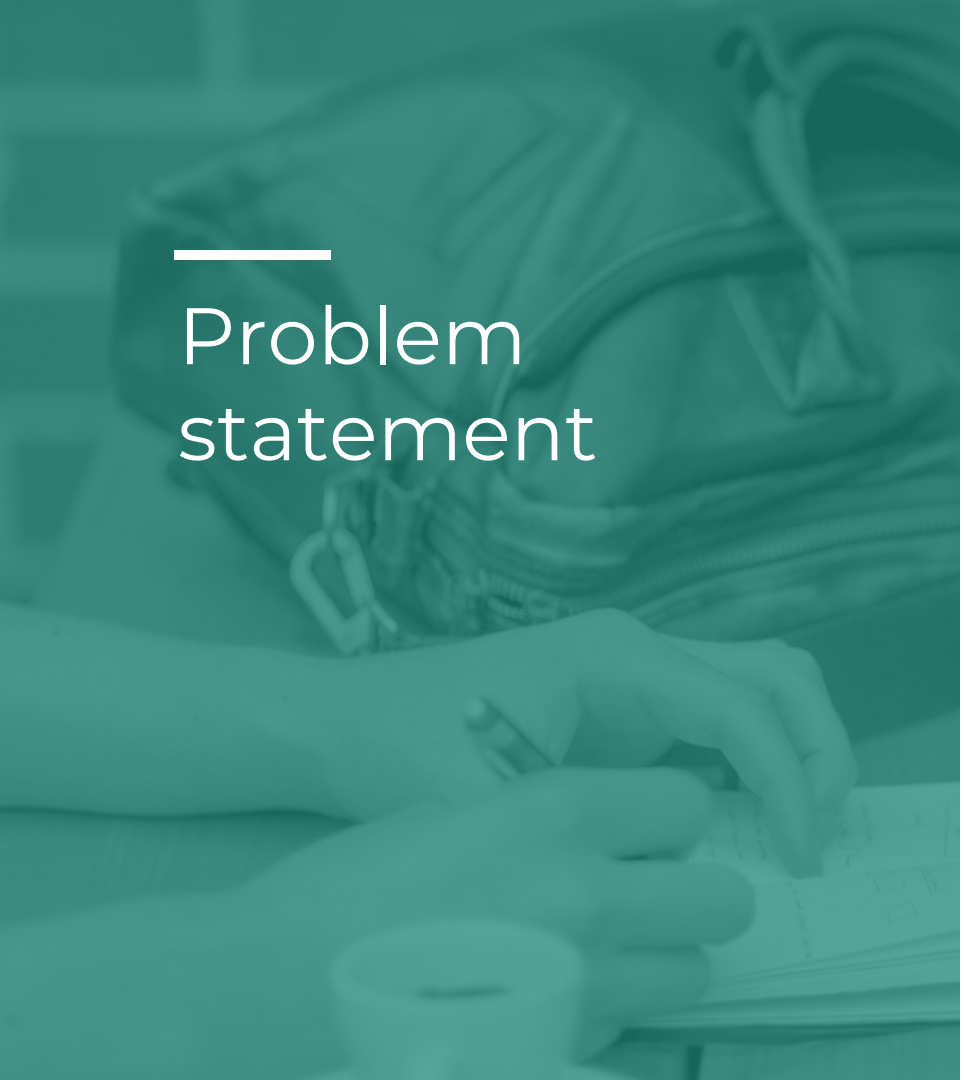
Problem statement

- To analyze air pollution trends in various states in India from 2005-2014.
- To find the most polluted states in India with respect to SO_2, NO_2 and RSPM concentrations.



Problem statement

- To analyze air pollution trends in various states in India from 2005-2014.
- To find the most polluted states 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



Problem statement

- To analyze air pollution trends in various states in India from 2005-2014.
- To find the most polluted states 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 project the states as hotspots and coldspots on the basis of chi-score and z-score

Data Sets

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)- http://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

Pollutant Concentration Analysis using Chi-Score

Pollutant Concentration Analysis using Chi-Score

- Used modified version of Chi-Squared Test to detect outliers.

Pollutant Concentration Analysis using Chi-Score

- Used modified version of Chi-Squared Test to detect outliers.
- Formula used:

Pollutant Concentration Analysis using Chi-Score

- Used modified version of 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

- A state is considered as 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

Limitation of Chi-Squared Test for outlier detection

- Tells whether a state is outlier (hotspot or coldspot) or not but doesn’t specify the nature of outlier.

Limitation of Chi-Squared Test for outlier detection

- Tells whether a state is outlier (hotspot or coldspot) or not but doesn't specify the nature of outlier.

- To overcome this limitation we used Z-score to find hotspots.

Hotspots using Z-Score

Hotspots using Z-Score

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

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

Hotspots using Z-Score

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

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

Since it's very rare that air pollution of a state isn't affected by the increasing air pollution level of it's neighboring state, we have defined a state as hotspot or coldspot taking into consideration the pollution level of it's neighbor.

- A state is Hotspot if : $MPC_{state} > Mean_{neighbor} + \frac{1}{2} std_{neighbor}$
- A state is Coldspot if : $MPC_{state} < Mean_{neighbor} - \frac{1}{2} std_{neighbor}$

Choropleth Map to visualize the results of Z-Score for the year 2014

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

Analysis of Results obtained using Z-Score

Analysis of Results obtained using Z-Score

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

Analysis of Results obtained using Z-Score

- Major Hotspots: Delhi, Maharashtra, Jharkhand, Nagaland, Rajasthan
- 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.

Analysis of Results obtained using Z-Score

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

- 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.

- Nagaland and Delhi, both are hotspots, but reasons for that 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 levels more than neighbors. Dust from poorly maintained roads, burning of wastes, some possible reasons.

Analysis of Results obtained using Z-Score

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

- 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.

- Nagaland and Delhi, both are hotspots, but reasons for that 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 levels more than neighbors. Dust from poorly maintained roads, burning of wastes, some possible reasons.

- Jharkhand, a hotspot : Possible reasons: Virginity of rural areas lost to industrialization. Most of the industries spit up pollutants like manufacturing of cement, concrete and mud bricks.

Analysis of Results obtained using Z-Score

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

- 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.

- Nagaland and Delhi, both are hotspots, but reasons for that 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 levels more than neighbors. Dust from poorly maintained roads, burning of wastes, some possible reasons.

- Jharkhand, a hotspot : Possible reasons: Virginity of rural areas lost to industrialization. Most of the industries spit up pollutants like manufacturing of cement, concrete and mud bricks.

- Similarly, Maharashtra's problems revolve around having both huge industry-vehicle concentration as well as every other metro city problem.

Analysis of Results obtained using Z-Score

Analysis of Results obtained using Z-Score

- Using Chi-Squared Test, we got Punjab and Bihar as outliers but Z-Score doesn’t mark them as hotspot or coldspot.

Analysis of Results obtained using Z-Score

- Using Chi-Squared Test, we got Punjab and Bihar as outliers but Z-Score doesn't mark them as hotspot or coldspot.

- Possible Reason: Although the pollution levels of Punjab and Bihar are high and possibly amongst the worst in the country, but since their neighbors are hotspots and have much higher pollution levels, they don't show up in the Z-score hotspot detection.

Analysis of Results obtained using Z-Score

- Using Chi-Squared Test, we got Punjab and Bihar as outliers but Z-Score doesn't mark them as hotspot or coldspot.

- Possible Reason: Although the pollution levels of Punjab and Bihar are high and possibly amongst the worst in the country, but since their neighbors are hotspots and have much higher pollution levels, they don't show up in the Z-score hotspot detection.

- Chi-Squared Test doesn't use the 'neighbor' statistic unlike Z-score computation and such differences are expected to arise because of that.

Most Polluted Cities in a State

Most Polluted Cities in a State

- Found top 5 most polluted cities in a state.
- Severely Polluted : MPC >=65
- Moderately Polluted : 45<= MPC <65
- Less Polluted: MPC<45

Most Polluted Cities in a State

- Found top 5 most polluted cities in a 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

Clustering

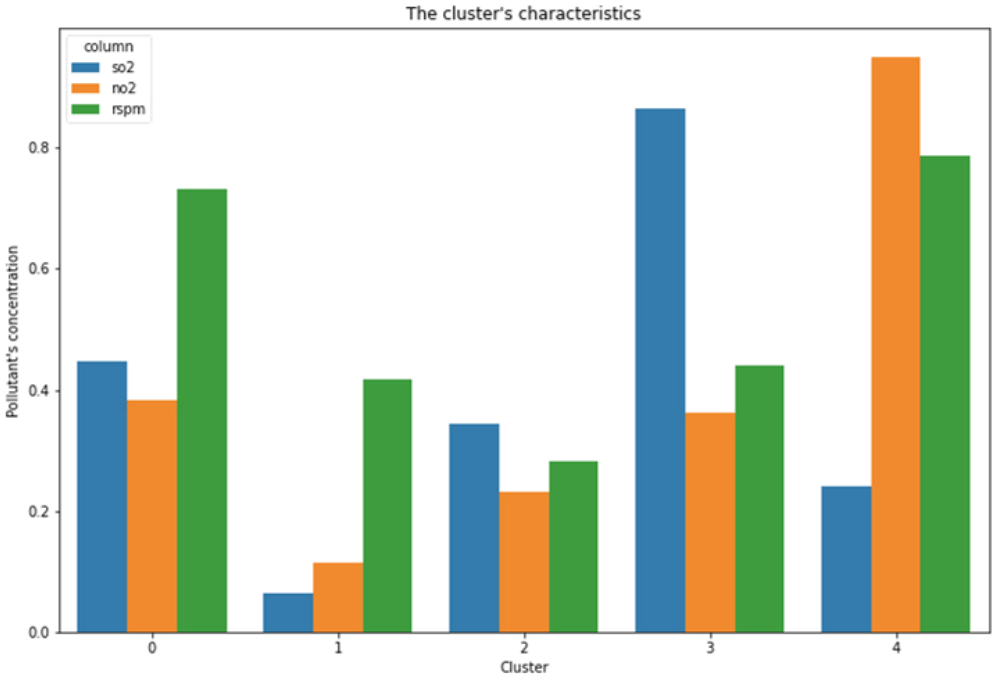
- To group similar states based on concentrations of SO_2 , NO_2 and RSPM

Clustering

- To group similar states based on concentrations of SO_2 , NO_2 and RSPM
- K-means Algorithm is used with 5 clusters

Clustering

- To group similar states based on concentrations of SO_2 , NO_2 and RSPM
- K-means Algorithm is used with 5 clusters



Interpretation of results obtained from Clustering

Interpretation of results obtained from Clustering

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

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.

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

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.

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

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.

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

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.

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

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.

Correlation

Correlation

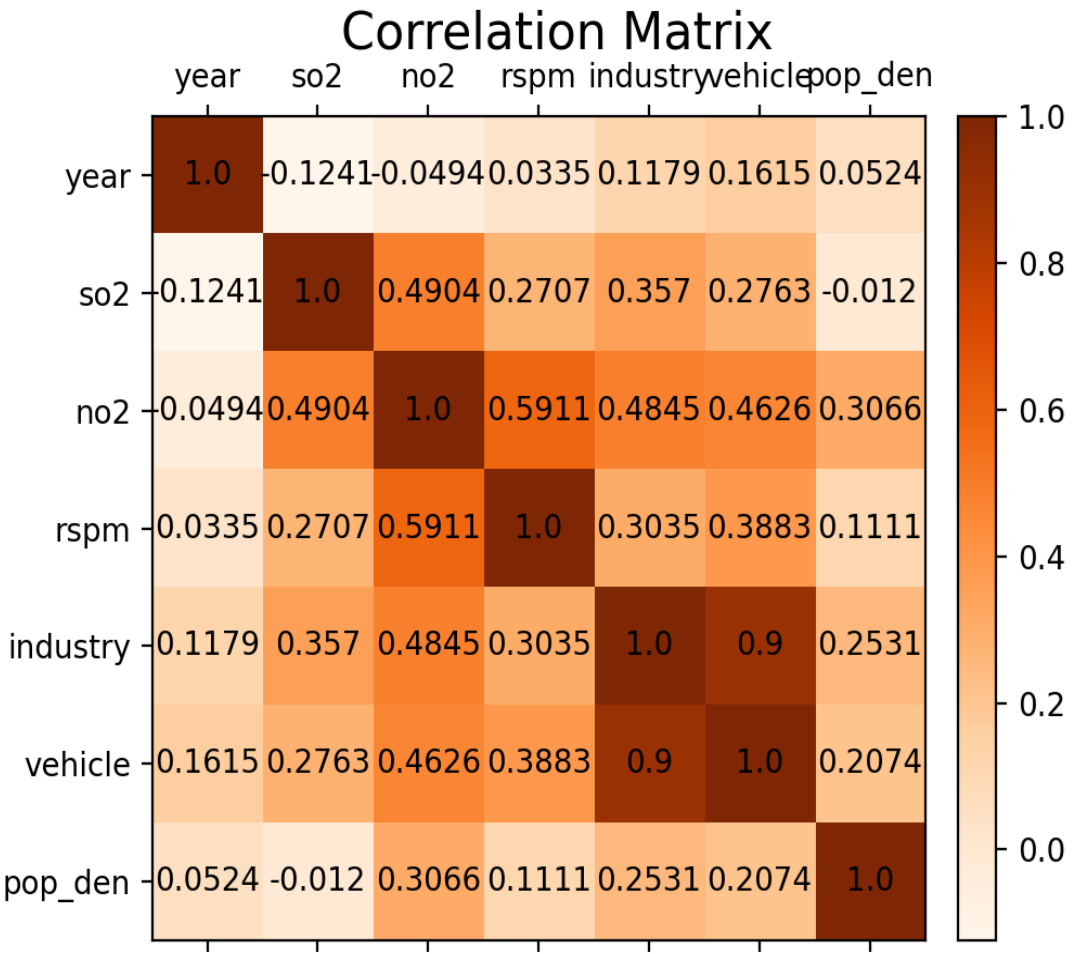
Features:

- SO_2
- NO_2
- RSPM
- Number of Industries
- Number of vehicles
- Population Density

Correlation

Features:

- SO_2
- NO_2
- RSPM
- Number of Industries
- Number of vehicles
- Population Density



Observations from Correlation Matrix

Observations from Correlation Matrix

- Correlation does NOT imply causation.

Observations from Correlation Matrix

- Correlation does NOT imply causation.
- Very high correlation(0.9) between number of vehicles and number of industries.

Observations from Correlation Matrix

- Correlation does NOT imply causation.
- Very high correlation(0.9) between number of vehicles and number of industries.
  - Possible reason : places with greater industries also require more vehicles to transport goods

Observations from Correlation Matrix

- Correlation does NOT imply causation.
- Very high correlation(0.9) between number of vehicles and number of industries.
  - Possible reason : places with greater industries also require more vehicles to transport goods
- Strong correlation is observed between :

Observations from Correlation Matrix

- Correlation does NOT imply causation.
- Very high correlation(0.9) between number of vehicles and number of industries.
  - Possible reason : places with greater industries also require more vehicles to transport goods
- Strong correlation is observed between :
  - RSPM and NO_2 levels

Observations from Correlation Matrix

- Correlation does NOT imply causation.
- Very high correlation(0.9) between number of vehicles and number of industries.
  - Possible reason : places with greater industries also require more vehicles to transport goods
- Strong correlation is observed between :
  - RSPM and NO_2 levels
  - SO_2 and NO_2 levels

Observations from Correlation Matrix

- Correlation does NOT imply causation.
- Very high correlation(0.9) between number of vehicles and number of industries.
  - Possible reason : places with greater industries also require more vehicles to transport goods
- Strong correlation is observed between :
  - RSPM and NO_2 levels
  - SO_2 and NO_2 levels
  - NO_2 level and number of vehicles

Observations from Correlation Matrix

- Correlation does NOT imply causation.
- Very high correlation(0.9) between number of vehicles and number of industries.
  - Possible reason : places with greater industries also require more vehicles to transport goods
- Strong correlation is observed between :
  - RSPM and NO_2 levels
  - SO_2 and NO_2 levels
  - NO_2 level and number of vehicles

- Subtle interdependencies possible for the above observations.

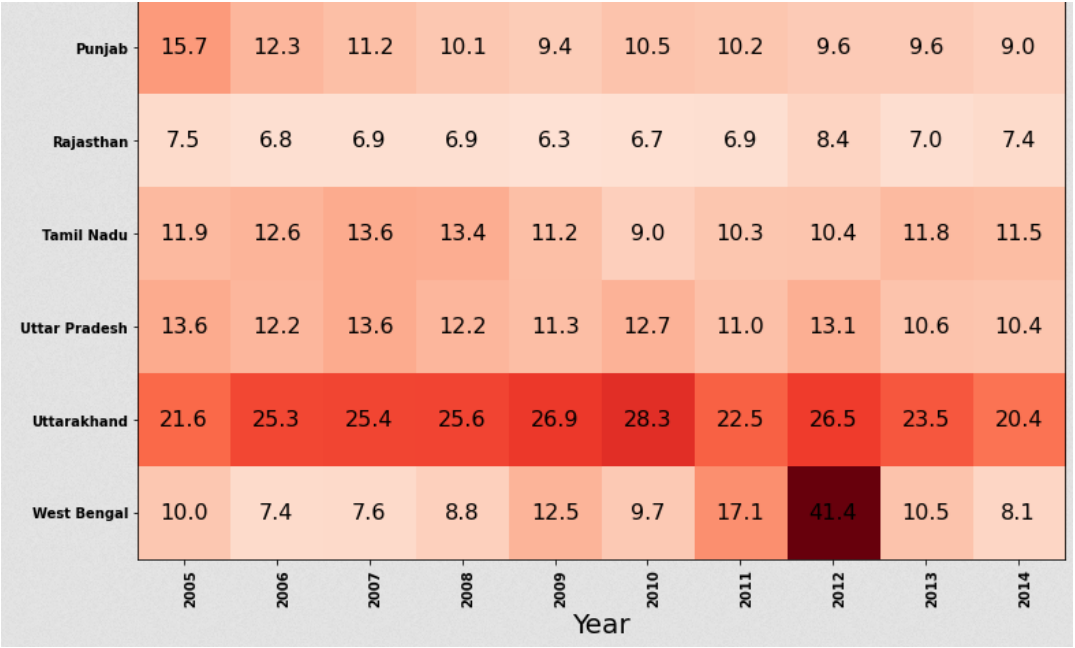
Observations from Correlation Matrix

- Correlation does NOT imply causation.
- Very high correlation(0.9) between number of vehicles and number of industries.
  - Possible reason : places with greater industries also require more vehicles to transport goods
- Strong correlation is observed between :
  - RSPM and NO_2 levels
  - SO_2 and NO_2 levels
  - NO_2 level and number of vehicles

- Subtle interdependencies possible for the above observations.
- For eg., the strong correlation between NO_2 and SO_2 is probably because of some reason which isn't a part of the feature set.

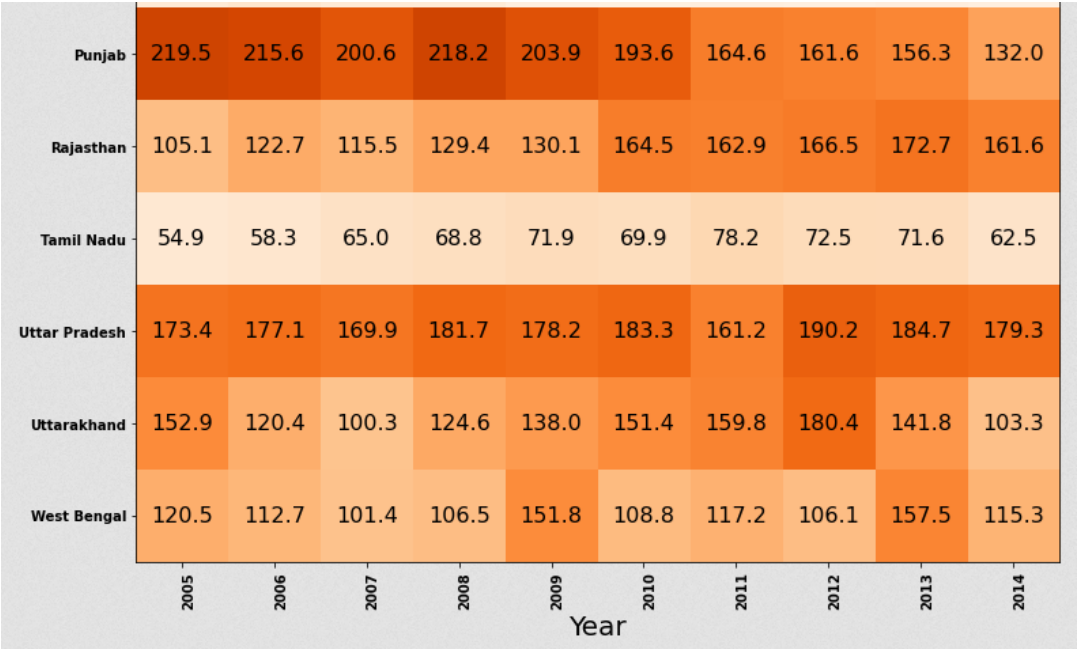
Heatmap of SO_2 concentration for some states

Heatmap of SO_2 concentration for some states



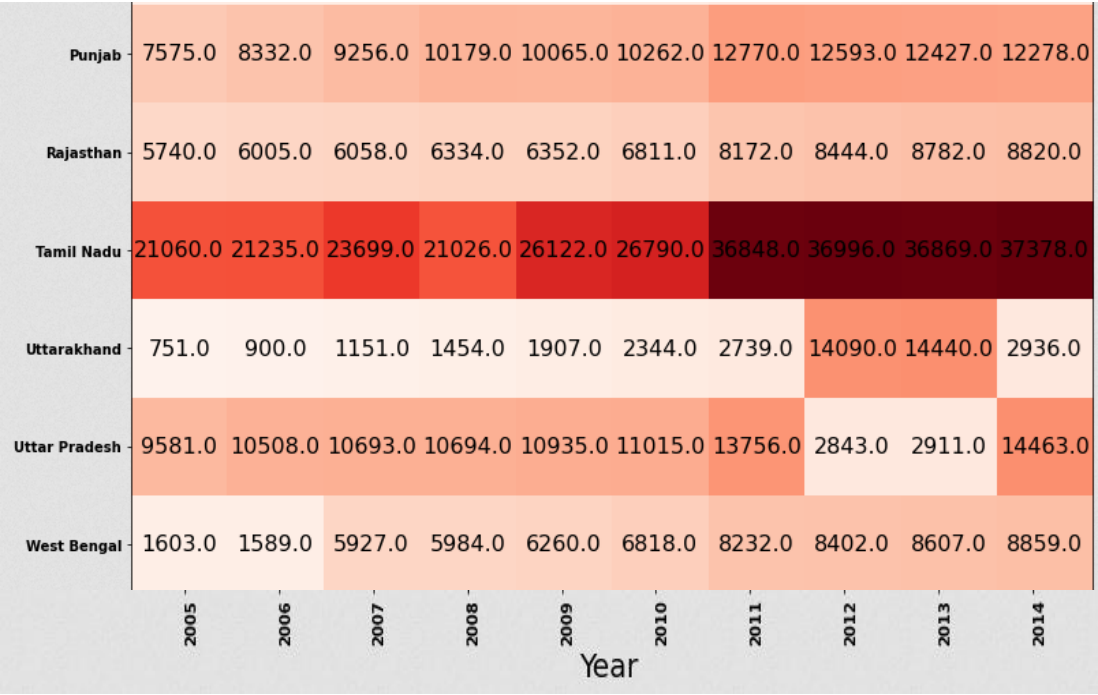
Heatmap of RSPM concentration for some states

Heatmap of RSPM concentration for some states



Heatmap of industries for some states

Heatmap of industries for some states



Results

Results

- 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. The increasing levels of pollutant concentration in Delhi has made suffer a lot and has been responsible for the deaths of thousands.

Results

- 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. The increasing levels of pollutant concentration in Delhi has made suffer a lot and has been responsible for the deaths of thousands.
- It is observed that Uttar Pradesh 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.

Results

- 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. The increasing levels of pollutant concentration in Delhi has made suffer a lot and has been responsible for the deaths of thousands.
- It is observed that Uttar Pradesh 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.
- 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 the Punjab's is categorized in less polluted states. The primary reason for air pollution in Punjab has been the burning of stubble by the farmers. The government has released many policies and programs to prevent this and these actions played a significant role in controlling the pollution.

Results

- 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. The increasing levels of pollutant concentration in Delhi has made suffer a lot and has been responsible for the deaths of thousands.
- It is observed that Uttar Pradesh 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.
- 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 the Punjab's is categorized in less polluted states. The primary reason for air pollution in Punjab has been the burning of stubble by the farmers. The government has released many policies and programs to prevent this and these actions 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.

Thank You