Air Pollution in India - Clustering

Presented by
Artem Ramus

Introduction

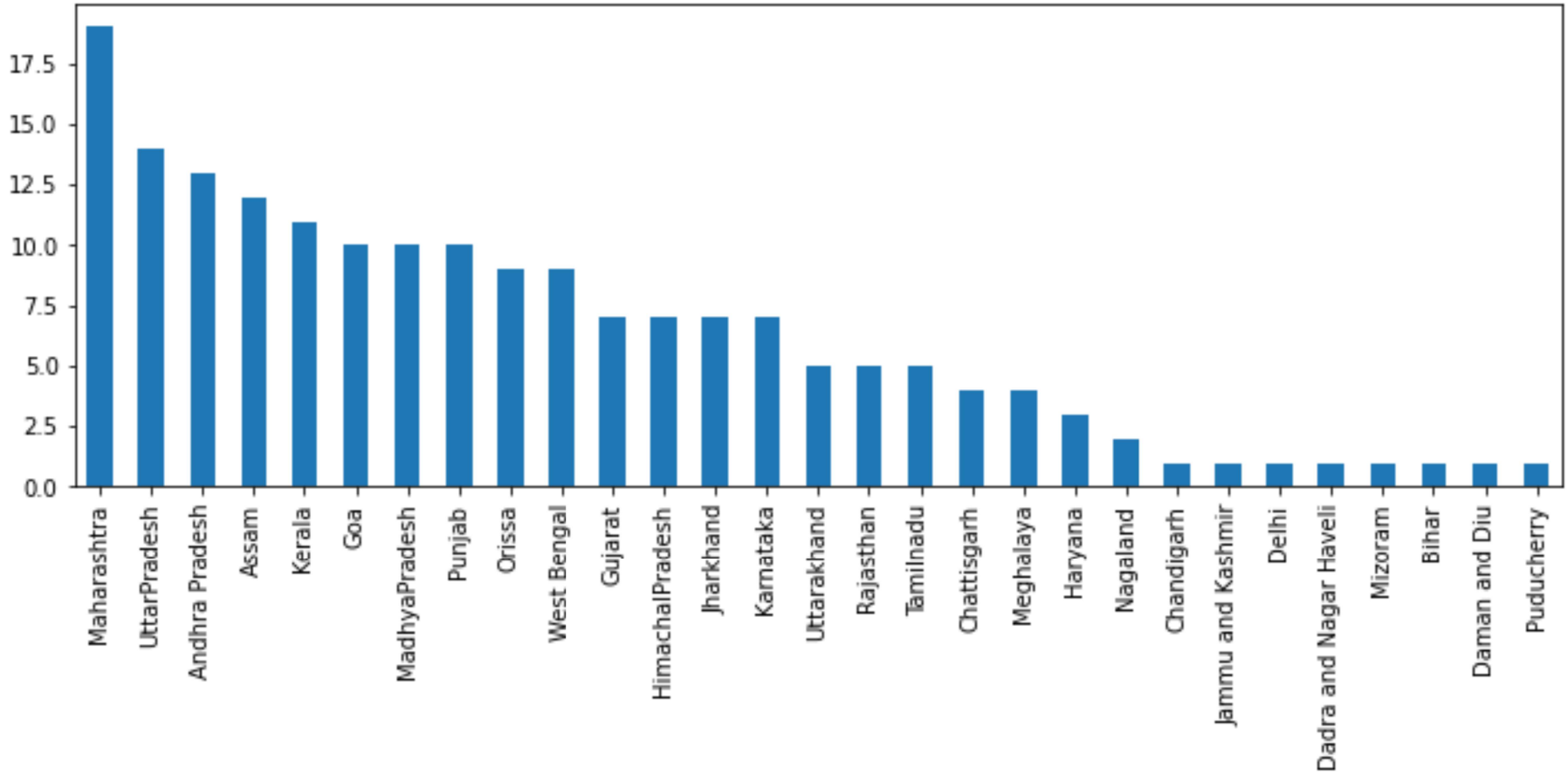
This data set comprises three types of air pollutant in India for specific cities. Additional column is "State".

Inspiration: derive meaningful insights about the air pollution in India by dividing the data to categories by common properties similarity.

Link to the dataset at Kaggle:
https://www.kaggle.com/adityadeshpande23/pollution-india-2010

Data Analysis

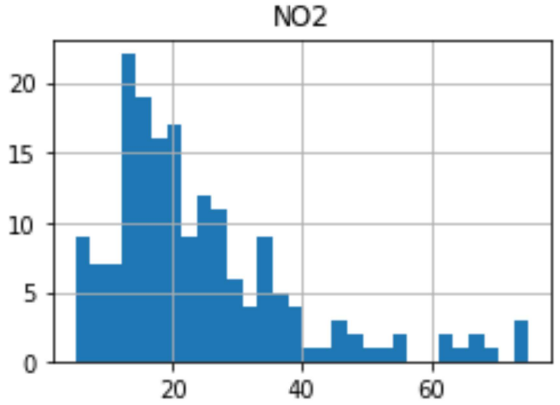
There is 181 cities as 29 states. Cities per state are shown below

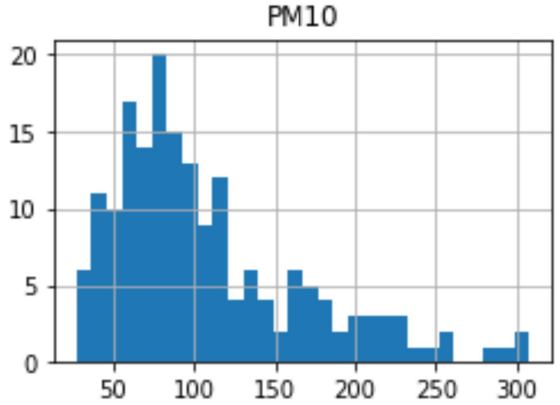


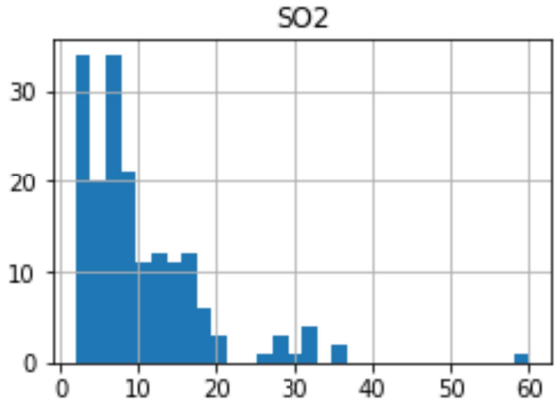
Data Analysis

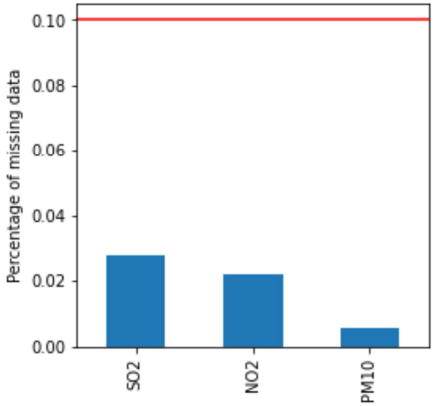
There is 3 pollutants - NO2, PM10, SO2. Statistical distribution is shown below:

	NO2	PM10	SO2
count	177.0	180.0	176.0
mean	24.1	108.1	10.0
std	14.7	60.9	8.3
min	5.0	27.0	2.0
25%	14.0	65.0	4.8
50%	20.0	89.5	7.5
75%	29.0	135.0	13.2
max	75.0	308.0	60.0









Data Analysis

The 5 most polluted states with NO_2 are West Bengal, Delhi, Bihar, Jharkhand and Maharashtra
The 5 least polluted states with NO_2 are Puducherry, Kerala, Meghalaya, Mizoram and Nagaland

The 5 most polluted states with PM10 are Delhi, Jharkhand, Bihar, Uttar Pradesh and Haryana
The 5 least polluted states with PM10 are Kerala, Mizoram, Dadra and Nagar Haveli, Puducherry and Daman and Diu

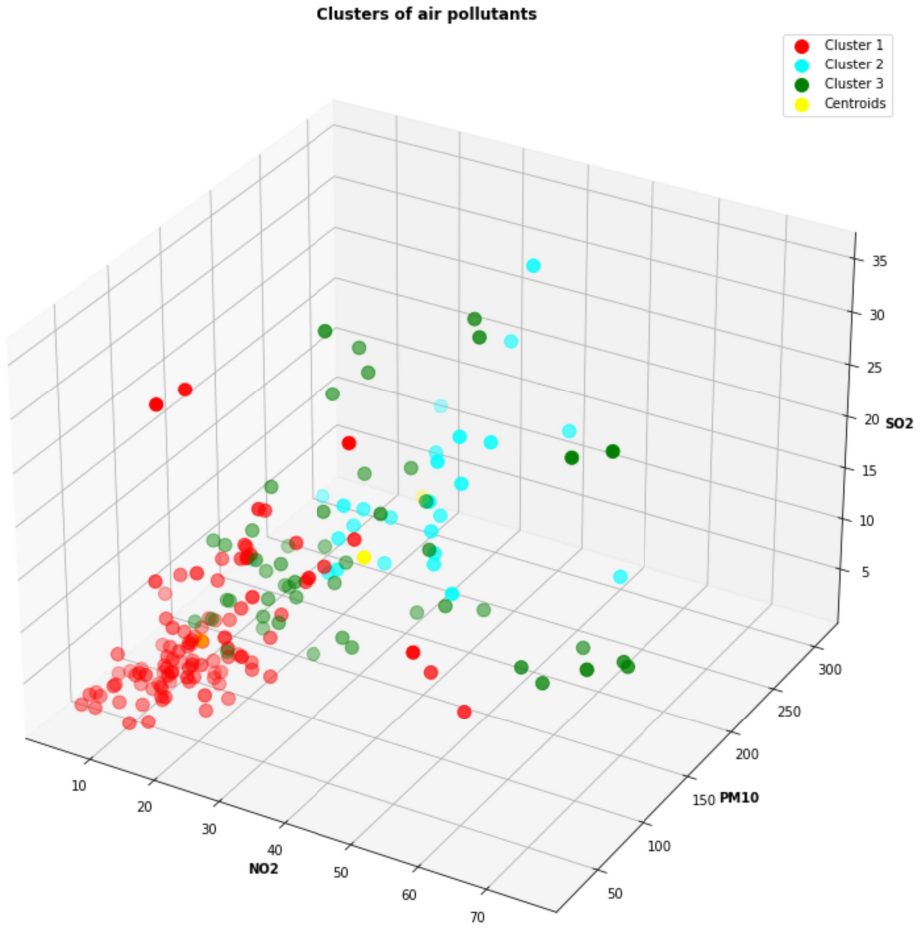
The 5 most polluted states with SO_2 are Uttarakhand, Jharkhand, Maharashtra, Gujarat and Uttar Pradesh
The 5 least polluted states with SO_2 are Himachal Pradesh, Meghalaya, Mizoram, Nagaland andS Chandigarh

Results

Cluster 1 'NO2' min and max are 5.0 and 56.0
Cluster 1 'PM10'min and max are 27.0 and 97.0
Cluster 1 'SO2'min and max are 2.0 and 32.0

Cluster 2 'NO2' min and max are 6.0 and 55.0
Cluster 2 'PM10'min and max are 181.0 and 308.0
Cluster 2 'SO2'min and max are 3.0 and 30.0

Cluster 3 'NO2' min and max are 11.0 and 75.0
Cluster 3 'PM10'min and max are 99.0 and 175.0
Cluster 3 'SO2'min and max are 2.0 and 35.0



Methodology

- The data set is checked for duplicates, null values and homogeneous features.
- Data distribution is checked with histogram, distribution and box plots for skewness and outliers.
- Numerous clustering models are engaged, based on co-location and density.
- K-mean model was chosen as the most suitable because of spatial characteristics. 3 clusters were chosen with the elbow method as the most representative.

The end

Thank you for your attention!