Air Pollution in India - Clustering

Presented by
Artem Ramus

Introduction

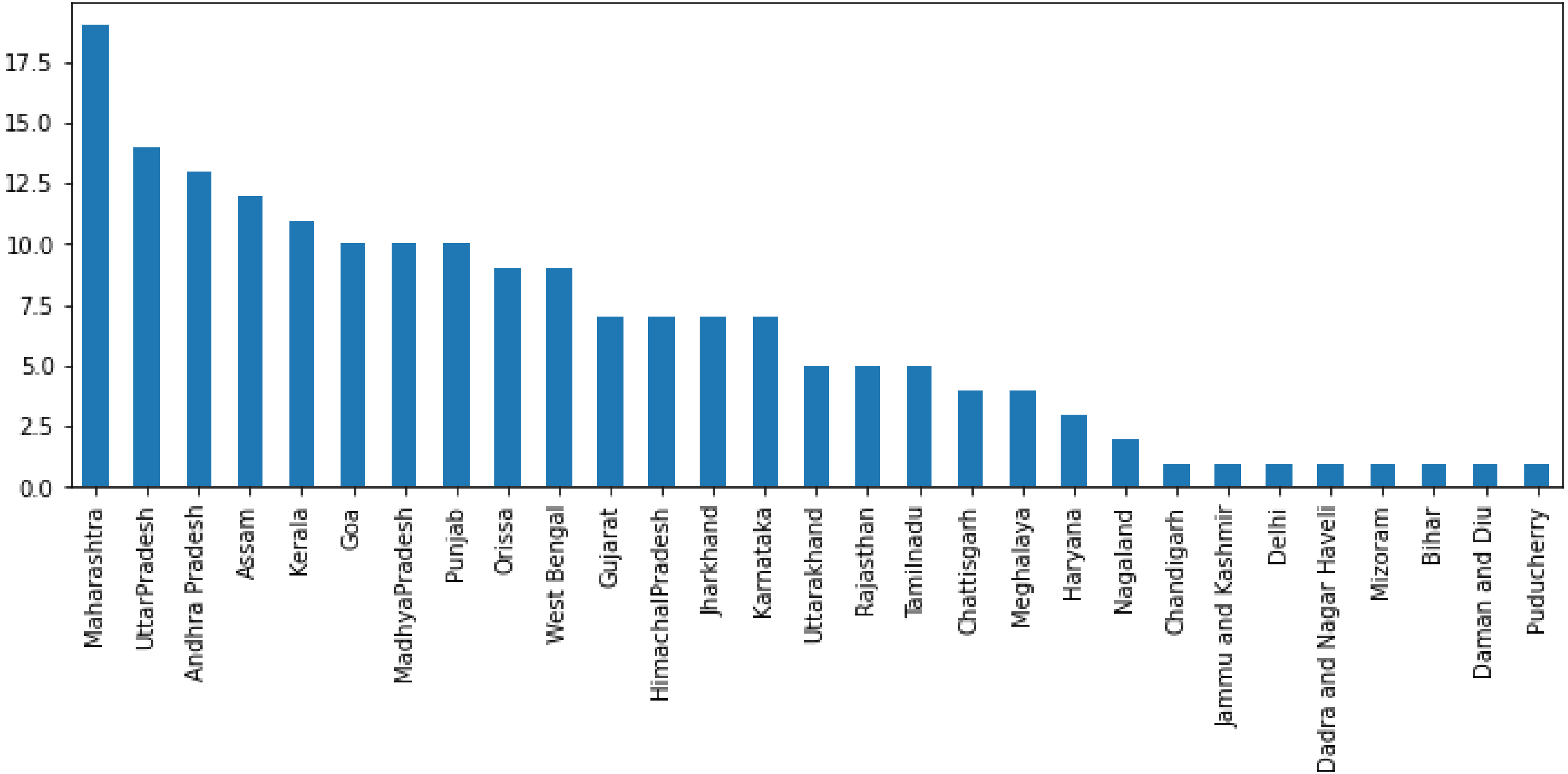
The goal of this project is to find areas in India with similar air pollution characteristics.

Data set contains levels of 3 common air pollutant: NO2 - nitrogen dioxide from NOX family, SO2 - sulfur dioxide, and MP10 - particle matters that have a diameter of 10 micrometers (0.01 mm) or smaller.

The data was taken from Kaggle: https://www.kaggle.com/adityadeshpande23/pollution-india-2010.

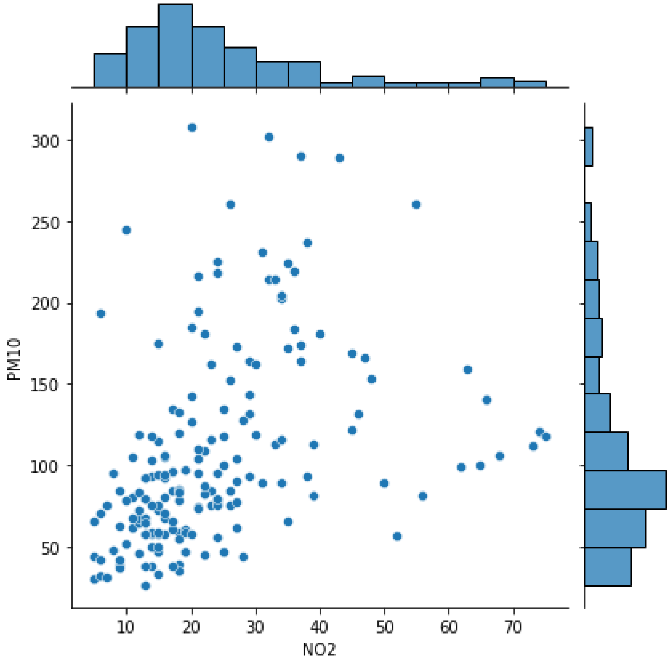
Data Analysis

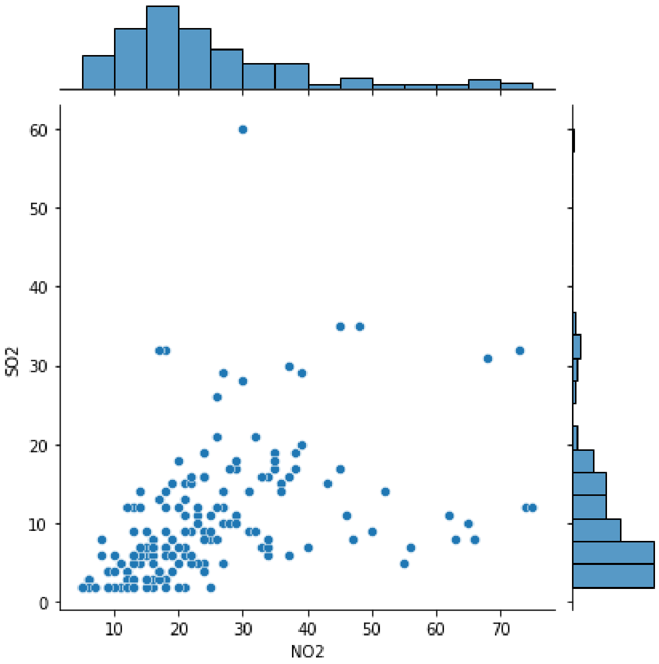
Measurements were taken in 181 cities in 29 states. Cities counts are shown below.

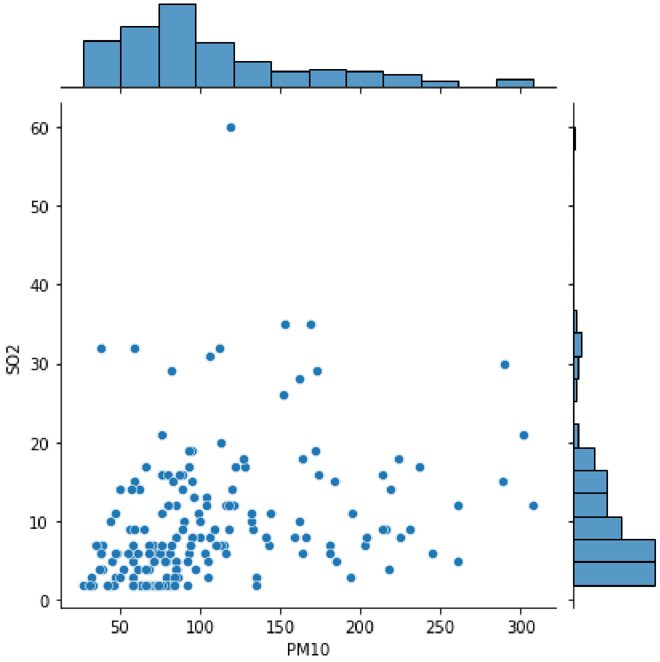


Data Analysis

There are 3 pollutants - NO2, PM10, SO2. Scatter plots are shown below:







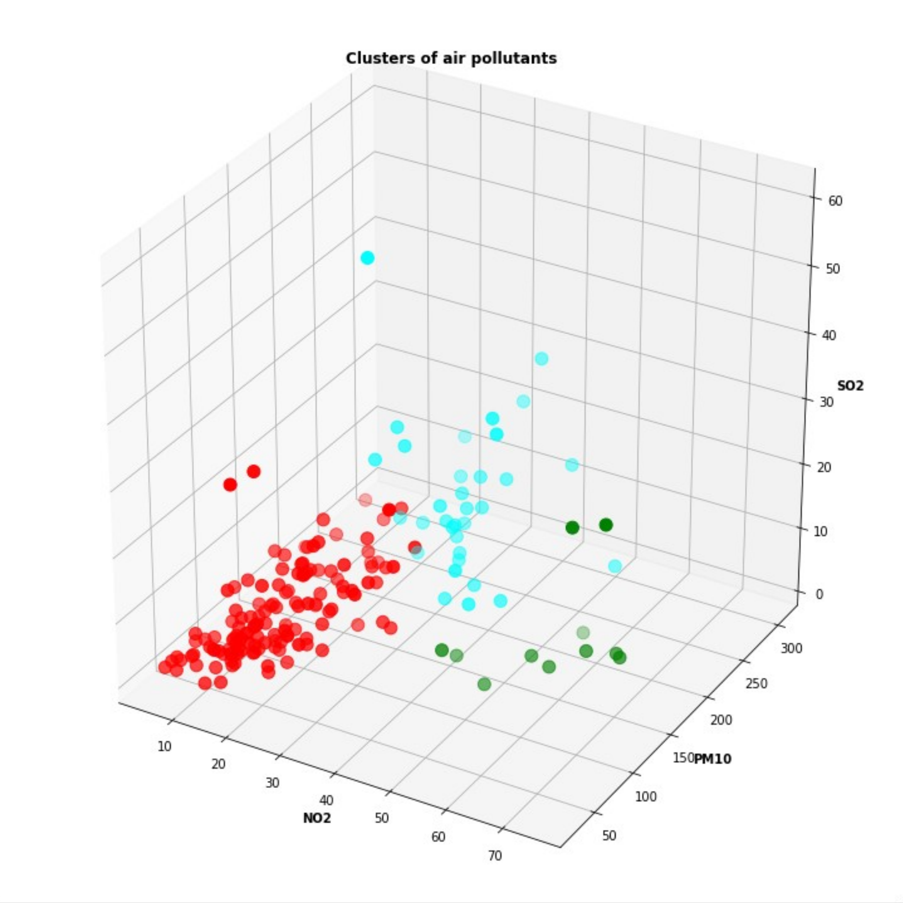
Results

Agglomerative clustering with 3 clusters was chosen.

Red cluster 1 comprises the least polluted areas by all means.

Light blue cluster 2 comprises the most polluted areas.

Green cluster 3 low in 'SO2' and high with 'NO2' and 'PM10'.



Methodology and Assumptions

- Models of choice:
  - Agglomerative clustering with 2 to 5 clusters
  - DBSCAN clustering resulted in 2 clusters and one area of low density
- Missing values ~2% were imputed with MICE approach under assumption of ‘data missing at random’.
- MIN-MAX normalization was chosen.

The end

Thank you for your attention!