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

Methodology

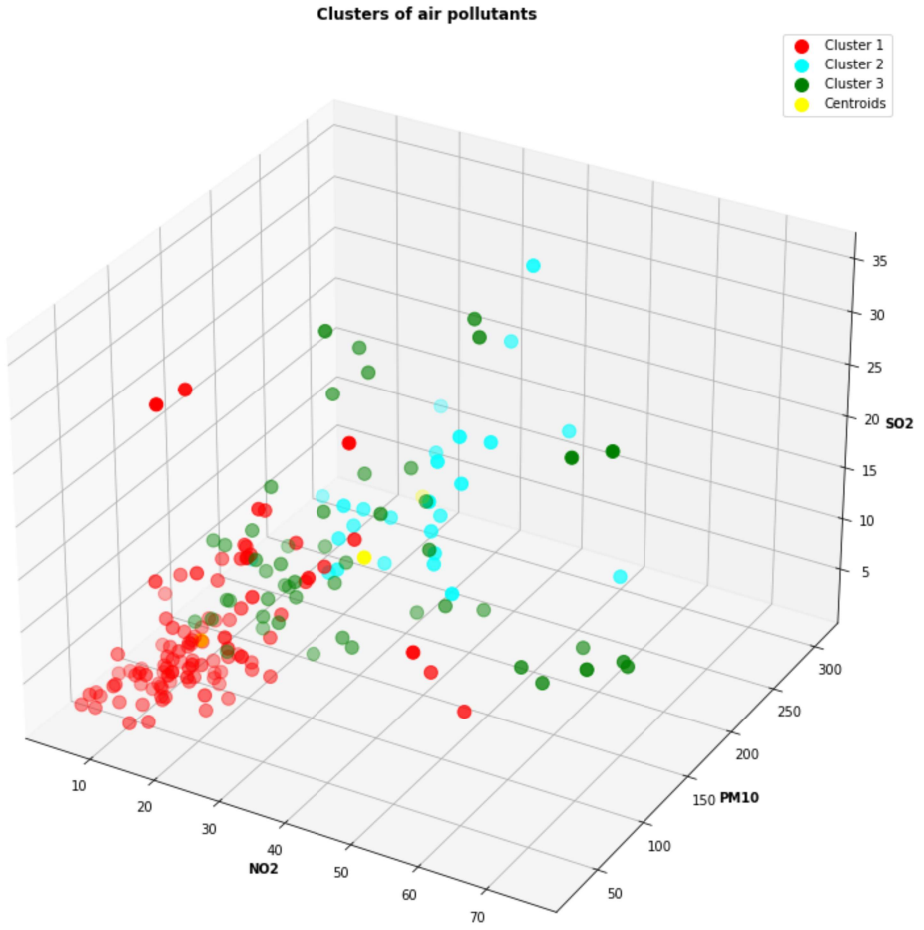
- The data set is checked for duplicates, null values and homogeneous features.
- Data distribution is checked with histogram, distribution and box plots for skewness and outliers.
- Numerous clustering models are engaged, based on co-location and density.
- K-mean model was chosen as the most suitable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

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



The end

Thank you for your attention!