Capstone Project Report

Air Quality

Analysis of AIR in India

Nema, Isha

Table of Contents

Problem Statement

Since industrialization, there has been an increasing concern about environmental pollution. As mentioned in the WHO report 7 million premature deaths annually linked to air pollution , air pollution is the world's largest single environmental risk. Moreover as reported in the NY Times article, India's Air Pollution Rivals China's as World's Deadliest it has been found that India's air pollution is deadlier than even China's.

Metrological department needs us to identify and rank the districts based on the air quality index. This air quality index is the formula which takes into account various parameters such as presence of Sulphur Dioxide, Nitrogen dioxide etc. The outcome of this will help in updating the Environment policy of the district.

Requestor – Metrological department

Audience - Metrological department, Policy Department India, Pollution Control India

Data

This data is combined(across the years and states) and largely clean version of the Historical Daily Ambient Air Quality Data released by the Ministry of Environment and Forests and Central Pollution Control Board of India under the National Data Sharing and Accessibility Policy (NDSAP).

```
<City id="Kolkata">
    <Station id="Rabindra Bharati University, Kolkata - WBPCB" lastupdate="18-02-2019 01:00:00">
     <Pollutant_Index Avg="294" Max="359" Min="212" id="PM2.5"/>
     <Pollutant_Index Avg="178" Max="236" Min="137" id="PM10"/>
     <Pollutant_Index Avg="100" Max="214" Min="49" id="NO2"/>
     <Pollutant_Index Avg="6" Max="8" Min="3" id="NH3"/>
     <Pollutant_Index Avg="11" Max="18" Min="6" id="SO2"/>
     <Pollutant_Index Avg="29" Max="78" Min="20" id="CO"/>
     <Pollutant_Index Avg="34" Max="53" Min="9" id="OZONE"/>
    </Station>
    <Station id="Victoria, Kolkata - WBPCB" lastupdate="18-02-2019 01:00:00">
     <Pollutant_Index Avg="156" Max="301" Min="81" id="PM2.5"/>
     <Pollutant_Index Avg="121" Max="182" Min="85" id="PM10"/>
     <Pollutant_Index Avg="110" Max="195" Min="42" id="NO2"/>
     <Pollutant_Index Avg="18" Max="38" Min="11" id="NH3"/>
     <Pollutant_Index Avg="16" Max="38" Min="5" id="SO2"/>
     <Pollutant_Index Avg="32" Max="106" Min="24" id="CO"/>
```

```
    <Pollutant_Index Avg="54" Max="111" Min="3" id="OZONE"/>
  </Station>
</City>
```

Above is the xml excerpt of the data.

Please find the detailed description of all the pollutants captured in this dataset.

Field Name	Field Description
PM2.5	PM2.5 are tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated.
PM10	Particles less than or equal to 10 micrometers in diameter are so small that they can get into the lungs, potentially causing serious health problems.
NO2	Its presence in aircontributes to the formation and modification of other air pollutants, such as ozone and particulate matter, and to acid rain.
NH3	The neutral, un- ionized form (NH3) is highly toxic to fish and other aquatic life.
SO2	It irritates the nose, throat, and airways to cause coughing, wheezing, shortness of breath, or a tight feeling around the chest.
CO	When too much carbon monoxide is in the air, your body replaces the oxygen in your red blood cells with carbon monoxide. This can lead to serious tissue damage, or even death.
Ozone	Ozone is an air pollutant that is harmful to breathe and it damages crops, trees and other vegetation. It is a main ingredient of urban smog.

Based on Indian government guideline below is the breakup.

AQI Category, Pollutants and Health Breakpoints

AQI Category (Range)	PM_{10} (24hr)	$PM_{2.5}$ (24hr)	NO_2 (24hr)	O_3 (8hr)	CO (8hr)	SO_2 (24hr)	NH_3 (24hr)	Pb (24hr)
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200	0–0.5
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400	0.5–1.0
Moderately polluted (101–200)	101–250	61–90	81–180	101–168	2.1–10	81–380	401–800	1.1–2.0
Poor (201–300)	251–350	91–120	181–280	169–208	10–17	381–800	801–1200	2.1–3.0
Very poor (301–400)	351–430	121–250	281–400	209–748	17–34	801–1600	1200–1800	3.1–3.5
Severe (401–500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

Data Cleaning

Step 1: The XML data was converted into csv format.

Step 2: Station Id was the address of station from which the values where calculated. Thus for working on the dataset the need was to convert the station id to city names.

Step 3: Calculating average by grouping on City name

Data Acquisition

Pollution data	https://data.gov.in/catalog/real-time-air-quality-index
City Data	http://censusindia.gov.in/Tables_Published/Admin_Units/Admin_links/Town_Codes_2001.xls
Location Details	geopy.geocoders

Data Analysis

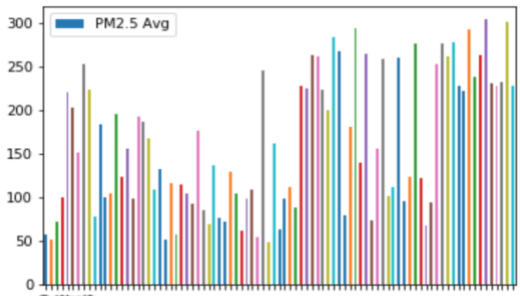
Correlation of variables using Kendall method:

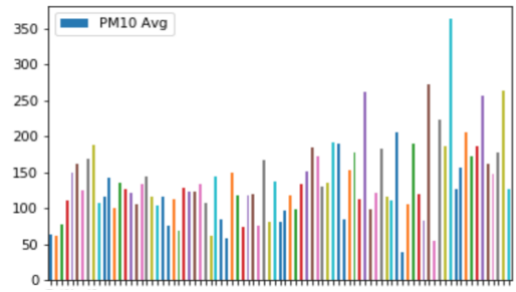
	PM2.5 Avg	PM10 Avg	NO2 Avg	NH3 Avg	SO2 Avg	CO Avg	Ozone Avg
PM2.5 Avg	1.000000	0.715840	0.373806	0.184749	0.222715	0.250984	-0.116536
PM10 Avg	0.715840	1.000000	0.427071	0.139657	0.270026	0.258382	-0.109821
NO2 Avg	0.373806	0.427071	1.000000	0.048914	0.162769	0.229331	0.045578
NH3 Avg	0.184749	0.139657	0.048914	1.000000	0.179315	0.140135	-0.111810
SO2 Avg	0.222715	0.270026	0.162769	0.179315	1.000000	0.154708	-0.119880
CO Avg	0.250984	0.258382	0.229331	0.140135	0.154708	1.000000	-0.169432
Ozone Avg	-0.116536	-0.109821	0.045578	-0.111810	-0.119880	-0.169432	1.000000

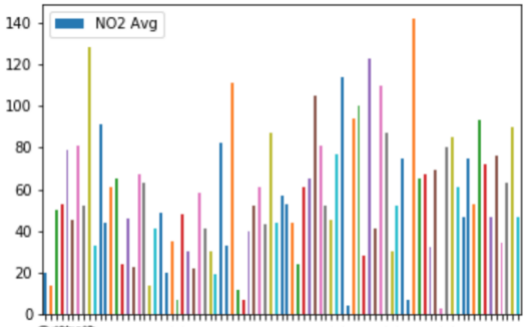
There is mild correlation between PM2.5 and PM 10. But none of the strong magnitude.

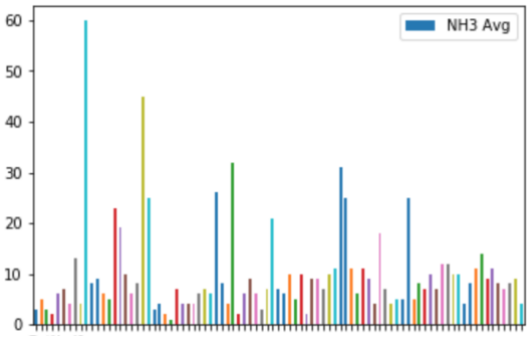
Data Analysis

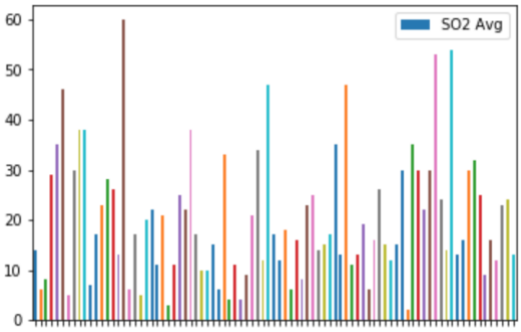
Bar graph are created to understand the value ranges of each pollutants.

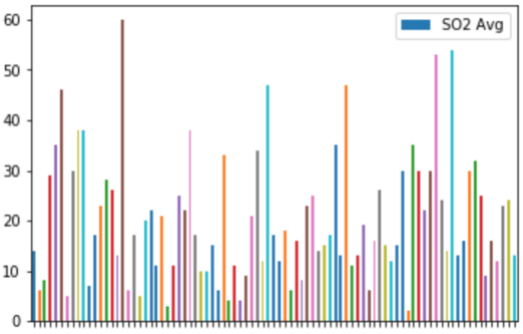

PM2.5 Avg
300
250
200
150
100
50
0

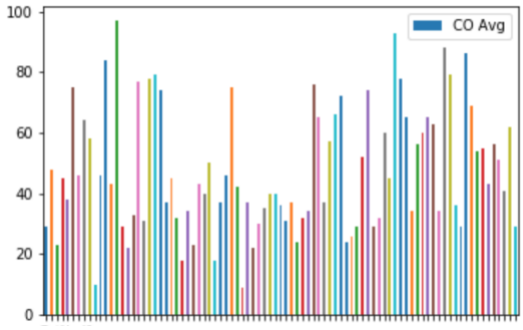

PM10 Avg
350
300
250
200
150
100
50
0

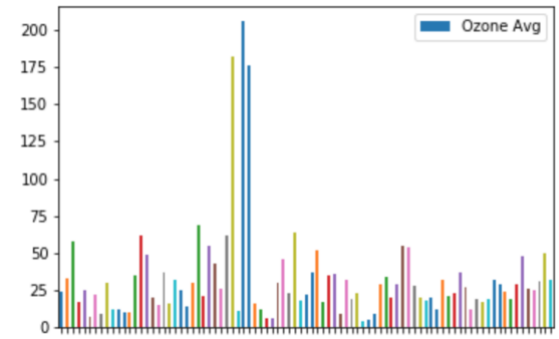

NO2 Avg
140
120
100
80
60
40
20
0


NH3 Avg
60
50
40
30
20
10
0


SO2 Avg
60
50
40
30
20
10
0


SO2 Avg
60
50
40
30
20
10
0


CO Avg
100
80
60
40
20
0


Ozone Avg
200
175
150
125
100
75
50
25
0

Modeling

Metrological department needs us to identify and rank the districts based on the air quality index. To do that k- means clustering algorithm is used.

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

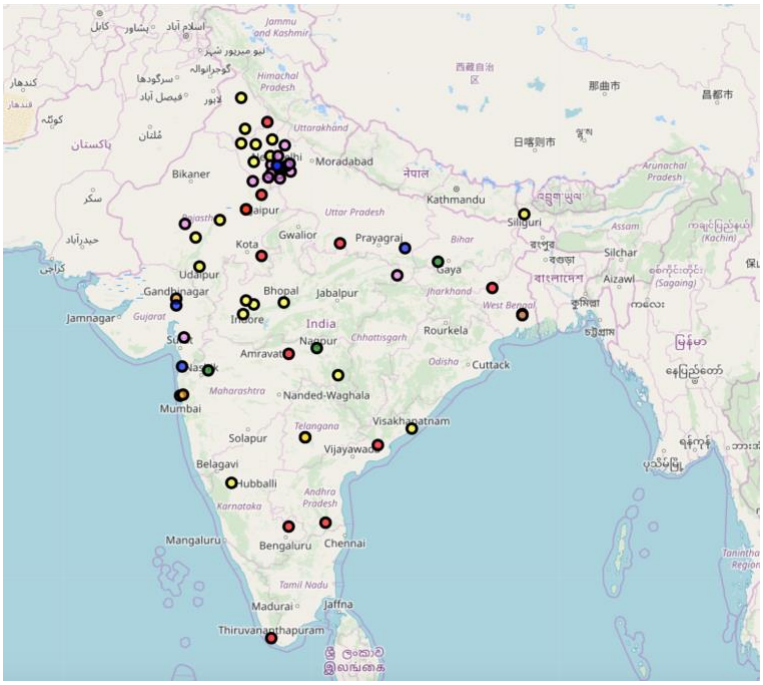
1. The centroids of the K clusters, which can be used to label new data

2. Labels for the training data (each data point is assigned to a single cluster)

3. Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS)

	Cluster	City_strip
0	0	Delhi, Vapi , Noida, Vatva, Varanasi, Ghaziabad
1	1	Alwar, Rajamahendravaram, Khanna, Tirupati, Th...
2	2	Ankleshwar, Hapur, Baghpat, Bhiwadi, Bulandsha...
3	3	Aurangabad, Nagpur, Nashik
4	4	Udaipur, Visakhapatnam, Hubballi, Pali, Ajmer,...
5	5	Yamuna Nagar, Siliguri, Jind, Jodhpur, Palwal,...

Conclusion



In this study, I analyzed the relationship between pollutants and the impact of those on various parts of India. I built k-means algorithm which created 6 clusters based on . pollutant data present in the dataset. This will give a much needed visual representation to metrological department and creation of this clusters will help them to identify the areas in which environment policies need to be altered.