Kolkata Neighborhood Clustering and it's relationship with Pollution

Aniketo Ghosh

6th May, 2020.

Introduction: Business Problem

Background: Global warming and climate change are pertinent issues of the current world and the conditions are getting worse by the day. A growing concern among citizens is the ever degrading air quality, especially so in cities. Thanks to accurate sensors of air pollution, we have data readily available for us to understand the real time situation ourselves.

Intended audience: In this project we will try to find the air quality index of Kolkata, India using PM 2.5 readings from across the city and its relationship with the neighbourhoods of the city. Specifically, this report will be targeted to stakeholders interested in climate change and ever increasing concern of degrading air quality ie. Government, NGOs or independent organizations.

Data: Acquisition and Characteristics

Acquisition: The data for this project has been acquired from https://cleair.io/. Cleair is a startup which specializes in low cost high accuracy network sensors which gather data about various aspects of air and noise pollution from their numerous sensors across the city.

The neighbourhood data of Kolkata will be collected using the Foursquare API. The venues in the neighbourhood, their types, categories and hence nature will be leveraged to make the study.

Characteristics: The data comes in csv format with the coordinates (latitude and longitude) of the location of the sensors and their average readings over the course of a week.

Feature selection: The features required for this study are Latitude, Longitude and PM 2.5 readings from the Cleair data and the headers we would get from the Foursquare API. Namely, Venue name, location, category, type and their frequency.

Data

Based on definition of the problem, factors that will influence the decission are:

- PM 2.5 readings from a particular neighbourhood
- intrinsic characteristic of neighbourhood influencing air quality
- location data of the neighbourhood

I decided to use coordinates to define our neighborhoods.

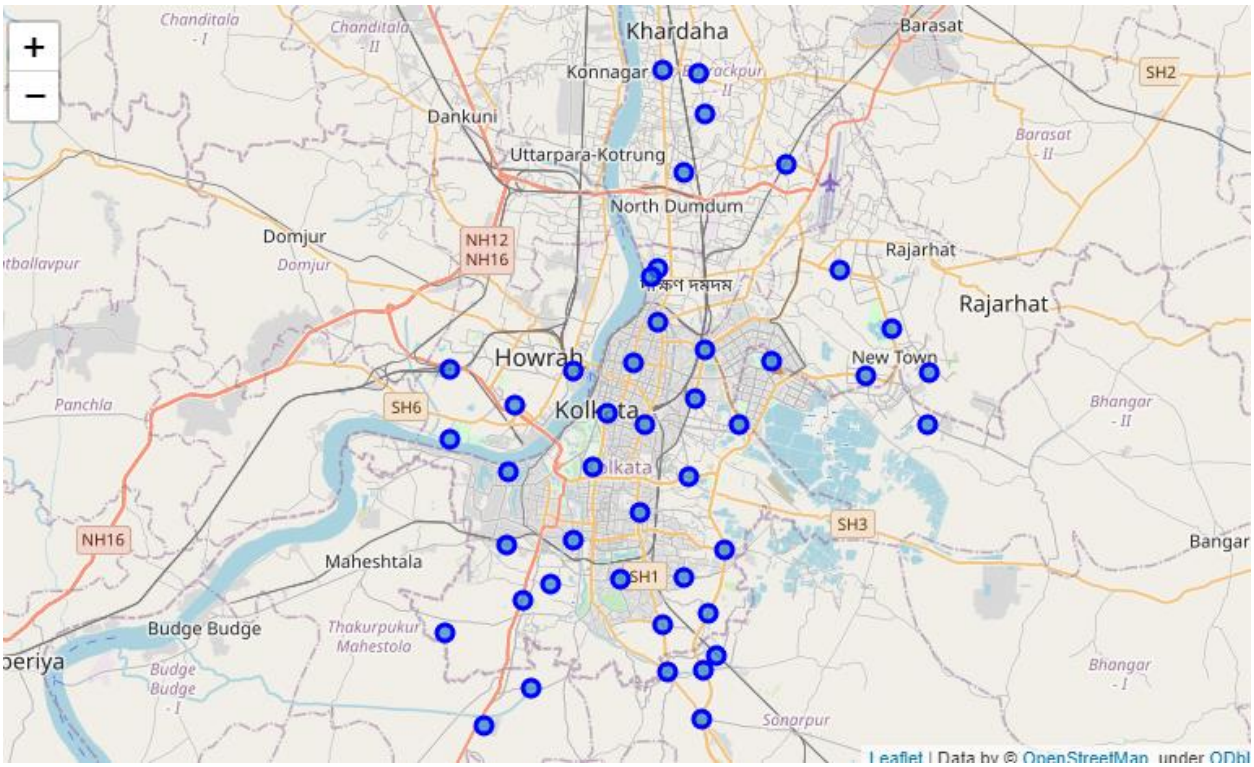
Following data sources will be needed to extract/generate the required information:

- coordinates of Kolkata **geolocator**
- different venues and their type and location in every neighborhood will be obtained using **Foursquare API**
- coordinate of Kolkata neighbourhoods and their respective PM 2.5 emission data from the a csv I found online from https://cleair.io/

The data looks like the following after preparation:

	Neighborhood	longitude	latitude	PM2.5
0	Victoria Memorial	88.345560	22.545673	92.56
1	Howrah Station	88.337300	22.583000	156.78
2	Taratala Road (Marine Engineering & Research I...	88.309656	22.515289	45.65
3	Chetla (Desher Khabar)	88.337631	22.517270	115.95
4	Lords More	88.357841	22.502047	84.96
5	Adarsha Palli (Ray Bahadur Road, Lions Club)	88.327682	22.499827	65.96
6	City Centre 2	88.450100	22.622300	88.65
7	Karunamoyee Crossing	88.421400	22.586500	101.27
8	Pallisree (Nabarun Club)	88.375265	22.483984	73.69
9	Garia (Depot)	88.377689	22.465832	91.68
10	Ajoynagar	88.394577	22.488743	80.14
11	Ruby More	88.401792	22.513483	124.49
12	Safui Para	88.384202	22.502452	78.21
13	Ballygunge Phari	88.366061	22.527527	98.42
14	Topsia more	88.386660	22.541758	114.51
15	Moulali (Kolkata Youth Center)	88.367681	22.561898	121.30
16	Beleghata (Building more)	88.407753	22.561732	95.32
17	Esplanade (park in front of Victoria House)	88.352086	22.566212	154.71
18	Phoolbagan	88.389428	22.572236	75.21
19	Ultadanga (below foot bridge)	88.393156	22.591409	131.99
20	Girish Park	88.362815	22.586040	99.69
21	Shyambazar (Five Points)	88.373673	22.601787	88.26

The data look like the following after plotting on the map:



Methodology

In the first step, we have **collected the required data**: location and names of every venue in our given neighbourhoods.

In the second step, we have **found the types, categories of all venues** (according to Foursquare categorization).

In the third step, we will **explore the neighbourhoods and their venue characteristics**, so as to give us an idea as to what kind of a neighbourhood it is

In the fourth step, we will **explore each neighbourhood separately** and find the top 10 venues of each neighbourhood

In the fifth step, we will perform **k-means clustering algorithm** on the data so as to **find clusters of similar neighbourhoods** together

In the sixth step, we will find **the average PM 2.5 ratings** of these neighbourhoods, whether they are **similar for a given cluster and how they differ from cluster to cluster.**

In the seventh step, we will find what the **venue data says about the neighbourhood** and how it is **related to the PM 2.5 emissions of the same.**

Analysis

The raw data was analyzed to find some insights into the data.

We found a total of 217 venues from all our neighbourhoods from the Foursquare data.

For evaluationary purposes, we went over the first neighbourhood in the dataset, Victoria Memorial which had 11 corresponding venues in the neighbourhood

	name	categories	lat	lng
0	Victoria Memorial	History Museum	22.545844	88.342890
1	Maidan	Field	22.549906	88.344219
2	Kenilworth Hotel	Hotel	22.546211	88.350133
3	Academy of Fine Arts	Art Gallery	22.543275	88.345138
4	Nandan	Indie Theater	22.542034	88.345440

Here we are showing only 5. So we get the names and categories of the venues, thus getting a fair idea from the categories, what kind of a neighbourhood it might be, whether it's a residential area or a commercial area.

To get a better idea of this, we analyze each neighbourhood on an individual level and take the mean of each venue category frequency and create a table to show the top 10 venues for each neighborhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9 M C V
0	Adarsha Palli (Ray Bahadur Road, Lions Club)	Dance Studio	Vegetarian / Vegan Restaurant	Clothing Store	Coffee Shop	Convenience Store	Deli / Bodega	Department Store	Dhaba	D S
1	Agarpara	ATM	Vegetarian / Vegan Restaurant	Dumpling Restaurant	Coffee Shop	Convenience Store	Dance Studio	Deli / Bodega	Department Store	D
2	Ajoynagar	Bus Station	Fast Food Restaurant	Bakery	Mughlai Restaurant	Vegetarian / Vegan Restaurant	Dhaba	Coffee Shop	Convenience Store	D S
3	Asoka Cinema Hall	ATM	Indian Restaurant	Indie Movie Theater	Mughlai Restaurant	Vegetarian / Vegan Restaurant	Discount Store	Coffee Shop	Convenience Store	D S
4	BNR (Engine Gate)	ATM	Vegetarian / Vegan Restaurant	Dumpling Restaurant	Coffee Shop	Convenience Store	Dance Studio	Deli / Bodega	Department Store	D

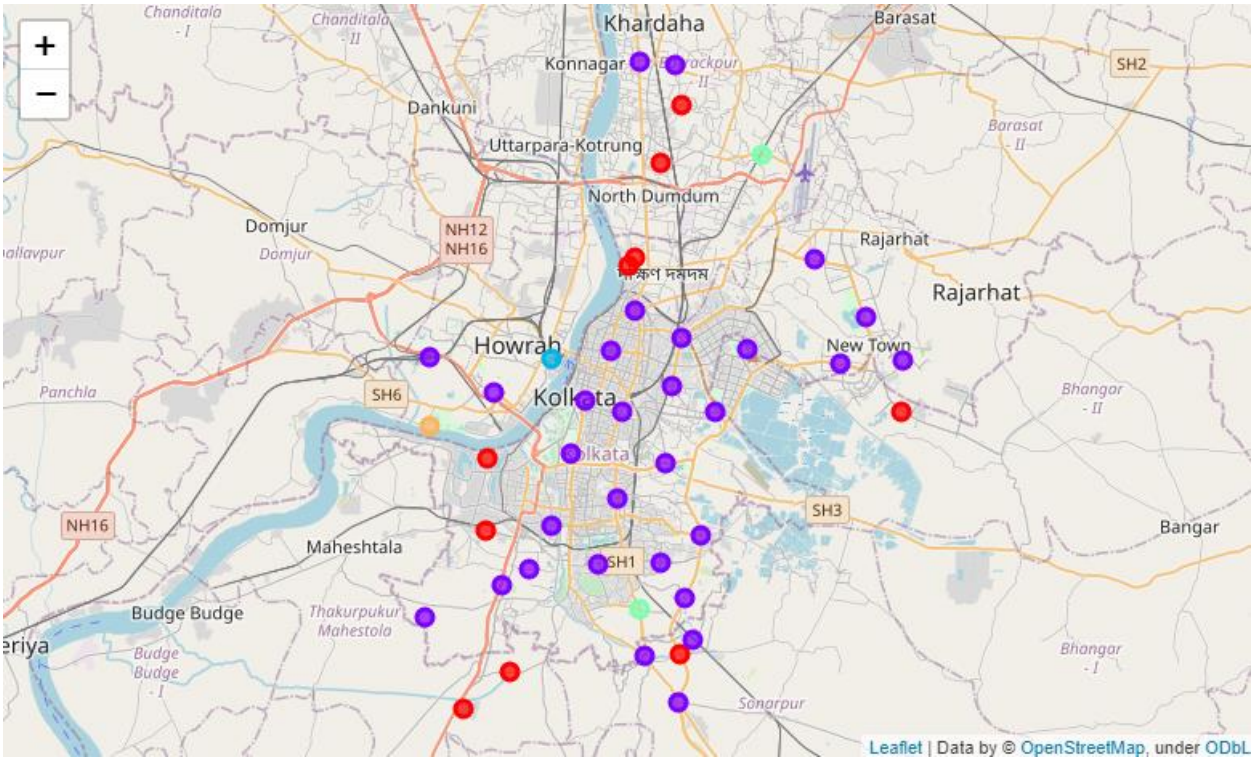
Thereafter, we perform k-means clustering algorithm on the dataset. It is a rather simple classification method, yet a very powerful one on unlabeled data such as this, in which we are trying to find a pattern among the neighbourhoods based on the category of venues present.

Results

We get the resultant dataframe after passing through the k-means clustering algorithm. We specified the number of clusters to be 5 and now every neighbourhood has been specified with a particular cluster label.

	Neighborhood	longitude	latitude	PM2.5	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Mos Commo Venue
0	Victoria Memorial	88.345560	22.545673	92.56	1	Shopping Mall	Food Court	Performing Arts Venue	History Museum	Art Galle
1	Howrah Station	88.337300	22.583000	156.78	2	Platform	Vegetarian / Vegan Restaurant	Discount Store	Coffee Shop	Conveni Store
2	Taratala Road (Marine Engineering & Research I...	88.309656	22.515289	45.65	0	ATM	Restaurant	Vegetarian / Vegan Restaurant	Dumpling Restaurant	Coffee S
3	Chetla (Desher Khabar)	88.337631	22.517270	115.95	1	Indian Sweet Shop	Bengali Restaurant	Park	Pharmacy	Jewelry Store
4	Lords More	88.357841	22.502047	84.96	1	Café	Clothing Store	Chinese Restaurant	Dumpling Restaurant	Multiple

Let us visualize the result on a map



Discussion

When we analyze the resultant dataset we find that there are mainly 2 clusters.

Cluster 1 and Cluster 2 are the **main clusters** that have been formed, each with very distinguishable characteristics.

Cluster 1 consists of neighbourhoods where most common venues are ATMS, convenience stores and restaurants (see dataframe c2) - clearly indicating that these are **residential areas**. Now here also, the air pollution characteristic is interesting to note, as the PM 2.5 emissions in these regions are **significantly lower** than other neighbourhoods and the average reading of PM 2.5 over all the neighbourhoods in Cluster 3 is **44.46 µg/m³**

	Hospital						Restaurant			
32	Agarpara	88.393100	22.683400	57.19	0	ATM	Vegetarian / Vegan Restaurant	Dumpling Restaurant	Coffee Shop	Convenie Store
33	Dhalai Bridge	88.392700	22.466000	37.64	0	ATM	Metro Station	Vegetarian / Vegan Restaurant	Dumpling Restaurant	Convenie Store
39	Cossipore Gun Shell Factory	88.370900	22.619400	43.67	0	ATM	Vegetarian / Vegan Restaurant	Dumpling Restaurant	Coffee Shop	Convenie Store
41	Belgharia Head Post Office	88.384000	22.660400	47.73	0	ATM	Pharmacy	Vegetarian / Vegan Restaurant	Discount Store	Coffee Sh
43	St. Xavier's University, Kolkata	88.487500	22.561900	40.36	0	ATM	Discount Store	Vegetarian / Vegan Restaurant	Dumpling Restaurant	Coffee Sh

```
c0['PM2.5'].mean()
```

```
44.458000000000006
```

Cluster 2 consists of neighbourhoods where most common venues are shopping malls, cafes and office buildings (see dataframe c1) - clearly indicating that these are **commercial areas**. Now the air pollution characteristic is interesting to note, as the PM 2.5 emissions in these regions are **significantly higher** than other neighbourhoods and the average reading of PM 2.5 over all the neighbourhoods in Cluster 2 is **92.98 µg/m³**

						Restaurant		Phone Shop	Restaurant	Store
34	Kavi Subhash metro station	88.397900	22.472200	30.87	1	ATM	Train Station	Hotel	Ice Cream Shop	Metr Stati
35	Mother's Wax Museum	88.472100	22.599400	86.87	1	Motorcycle Shop	Park	Art Museum	Clothing Store	Vege Vega Rest
36	TCS Gitanjali Park	88.487900	22.582000	41.70	1	IT Services	ATM	Vegetarian / Vegan Restaurant	Dumpling Restaurant	Conv Store
37	Sarsuna	88.283500	22.481000	78.67	1	American Restaurant	Jewelry Store	Vegetarian / Vegan Restaurant	Dumpling Restaurant	Conv Store
38	One Rajarhat	88.461200	22.580900	86.41	1	Bakery	Hotel	Restaurant	Furniture / Home Store	Multi
40	Khardaha	88.375300	22.700300	91.43	1	Clothing Store	Men's Store	Discount Store	Coffee Shop	Conv Store

```
c1['PM2.5'].mean()
```

```
92.98066666666664
```

Results

This result goes on to show that the air quality index of residential areas is much lesser than that of commercial areas in Kolkata, India.