Applied DS Capstone Project - The Battle of Neighbourhoods

-Shambo Basu

August,2019

1. Business Problem

1.1 Background

Delhi is the capital and one of the best cities in India. It has people from diverse backgrounds and cultures and even larger diversity of food and choice of cuisine.
There are already a lot of restaurants that are open in the city and hence for anyone trying to start a new restaurant in the city it will have to compete with a lot of others hence resulting in small profit margins

1.2 Problem

In this project I am trying to answer the question - **What will be the ideal location for starting a new restaurant of a particular cuisine so as to maximise profit?**

1.3 Interest

This is for restaurant owners who are looking to start a new restaurant or expand an existing chain of restaurant in the city.

2. Data

1. The Delhi geoJson File was provided by - https://github.com/datameet/Municipal_Spatial_Data/blob/master/Delhi/Delhi_Boundary.geojson

2. http://www.elections.in/delhi/mcd-elections/mcd-ward-list-2017.html - This will provide us with data of the neighbourhoods in Delhi but each neighbourhood is segmented into further sub-neighbourhoods hence it will have to cleaned to get only the major neighbourhoods of the city

3. The latitude and longitude data is obtained using geolocator library in python

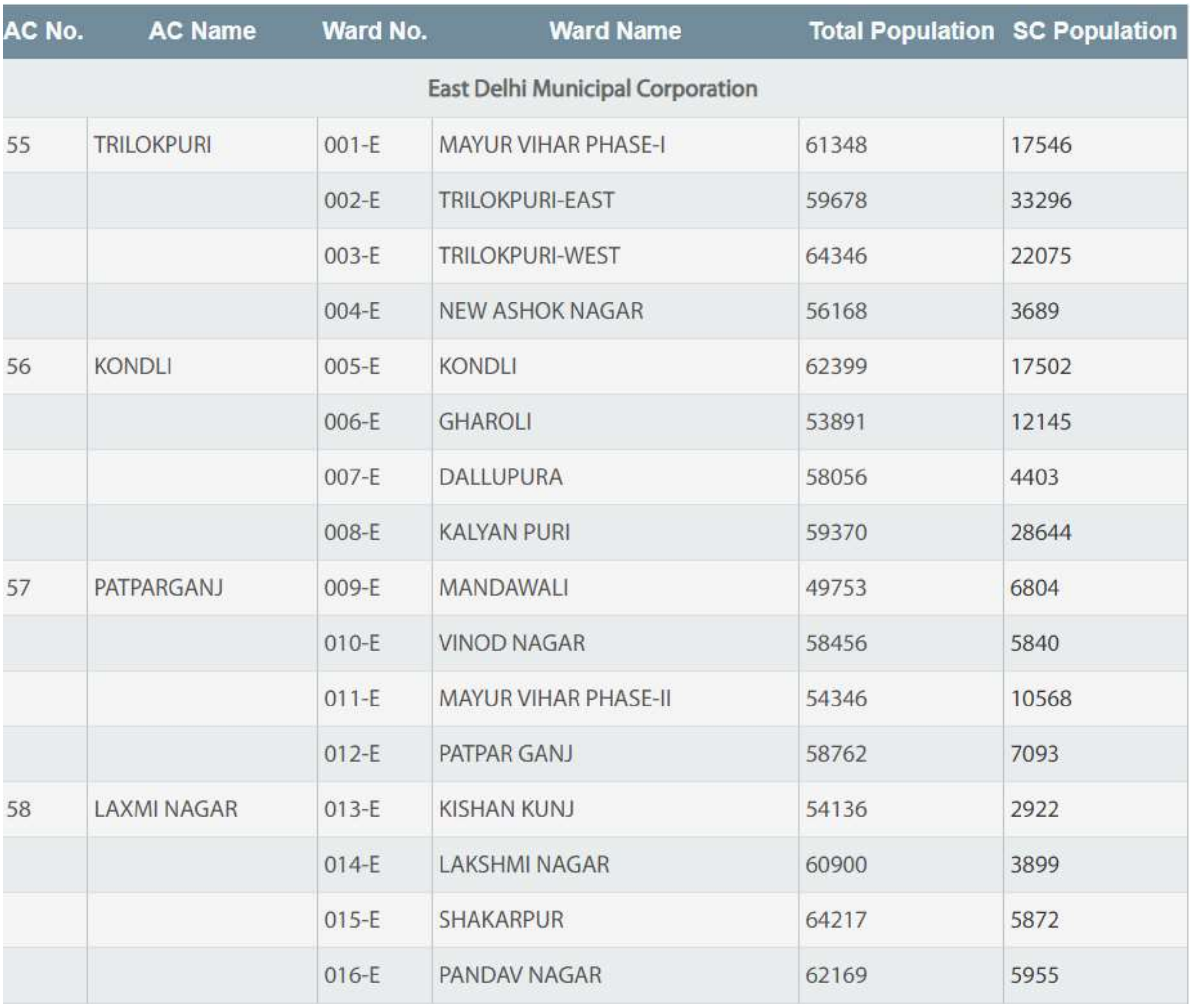
4. Places/Neighbourhoods which geolocator can't locate is obtained from "https://www.distancesto.com"

5. Foursquare API- This will provide us with the trending neighbourhoods in the area and the number of restaurants of a particular cuisine in the neighbourhood and number of restaurants in general

3. Methodology

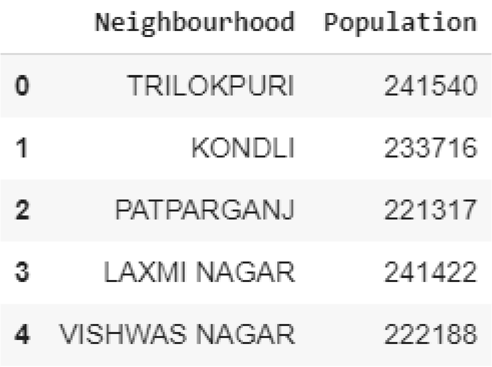
3.1 Data Cleaning

Once the dataset was imported and stored in a Dataframe using pandas it looked like the following:



AC No.	AC Name	Ward No.	Ward Name	Total Population	SC Population
East Delhi Municipal Corporation					
55	TRILOKPURI	001-E	MAYUR VIHAR PHASE-I	61348	17546
		002-E	TRILOKPURI-EAST	59678	33296
		003-E	TRILOKPURI-WEST	64346	22075
		004-E	NEW ASHOK NAGAR	56168	3689
56	KONDLI	005-E	KONDLI	62399	17502
		006-E	GHAROLI	53891	12145
		007-E	DALLUPURA	58056	4403
		008-E	KALYAN PURI	59370	28644
57	PATPARGANJ	009-E	MANDAWALI	49753	6804
		010-E	VINOD NAGAR	58456	5840
		011-E	MAYUR VIHAR PHASE-II	54346	10568
		012-E	PATPAR GANJ	58762	7093
58	LAXMI NAGAR	013-E	KISHAN KUNJ	54136	2922
		014-E	LAKSHMI NAGAR	60900	3899
		015-E	SHAKARPUR	64217	5872
		016-E	PANDAV NAGAR	62169	5955

- AC. No., Ward No. and SC Population columns were dropped as they were of no use to us.
- Rows such as ‘East Delhi Municipal Corporation’ were removed.
- Ward Name rows were combined under the AC Name column and the total population of these Ward Names added up
- The AC Name and Population were copied into a new DataFrame under the columns Neighbourhood and Population.
  At the end the new DataFrame looked like this:

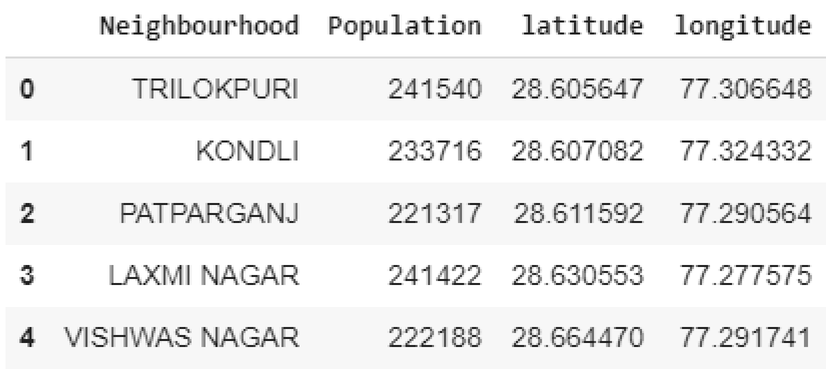


	Neighbourhood	Population
0	TRILOKPURI	241540
1	KONDLI	233716
2	PATPARGANJ	221317
3	LAXMI NAGAR	241422
4	VISHWAS NAGAR	222188

3.2 Longitude Latitude Data

Latitude and Longitude data of each neighbourhood was added using the **geopy library**

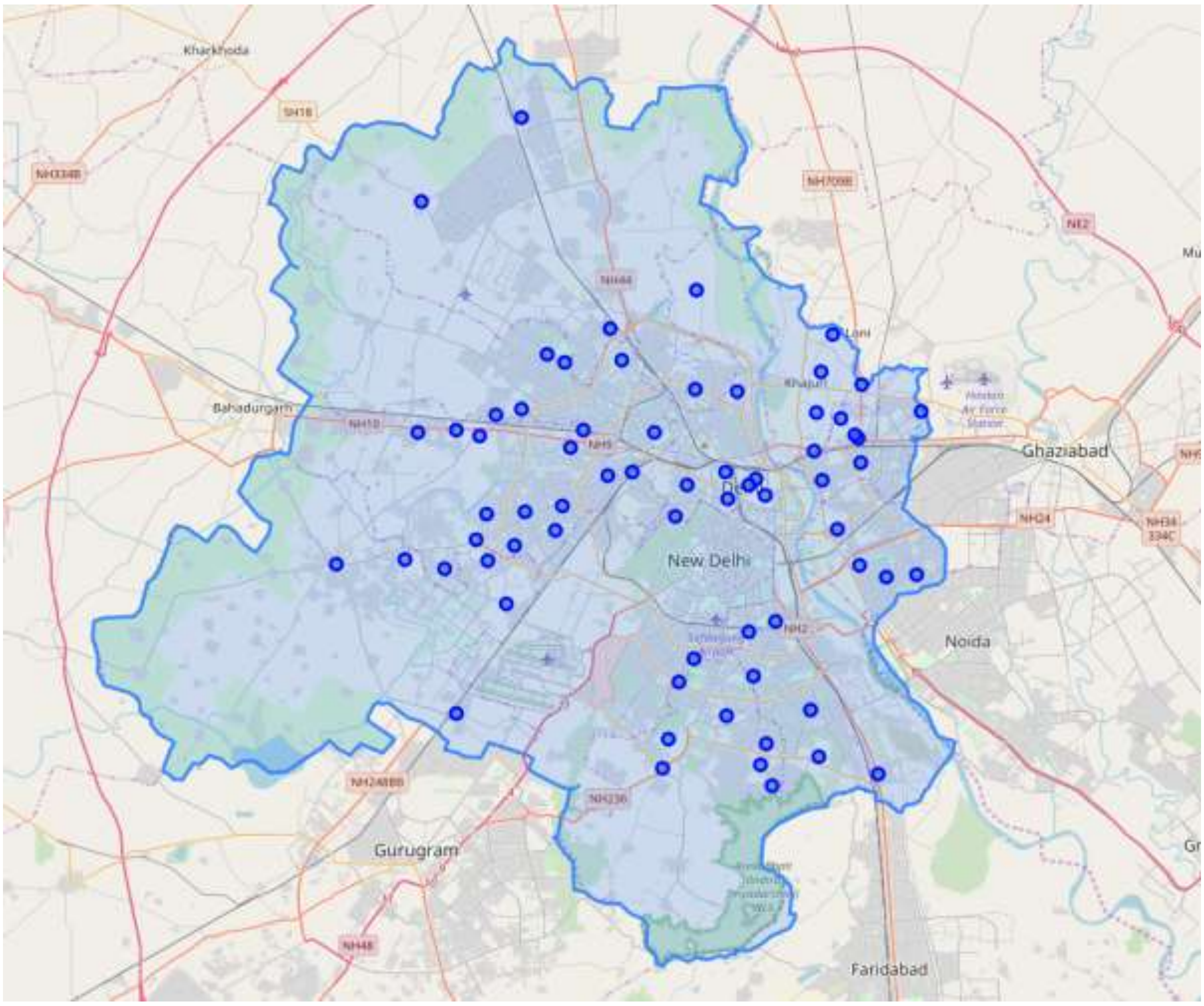
But geopy didn't return this data for a few neighbourhoods and those had to be manually entered from the website mentioned in 2(4) and these resulted in:



	Neighbourhood	Population	latitude	longitude
0	TRILOKPURI	241540	28.605647	77.306648
1	KONDLI	233716	28.607082	77.324332
2	PATPARGANJ	221317	28.611592	77.290564
3	LAXMI NAGAR	241422	28.630553	77.277575
4	VISHWAS NAGAR	222188	28.664470	77.291741

3.3 Exploring the locations

The locations were plotted using the geoJson file for the boundaries and the latitude, longitude data of the neighbourhoods using the folium mapping library:



3.4 Defining Neighbourhoods

It is clear from the map that the neighbourhoods in Delhi are quite randomly situated while neighbourhoods are far apart in some places and really clustered in others and some places seem they have no neighbourhoods at all, but these are actually divided in terms of population where the number of people in each Neighbourhood is roughly the same(as can be seen in the DataFrame in section 3.1 and 3.2) for election purposes.

For our usage we need to define the are around each location, I have defined it as:

- If the Neighbourhoods are very close then the radius of circle around the location is 750m
- But if they are far apart it is half the distance of the nearest neighbour from it.

The logic behind this decision is that If someone is at one location and travels 750m in any one direction, they may end up in an area that can be defined as the neighbourhood of an entirely different location, but if it is far apart then travelling half the distance of the nearest neighbour will end up in the neighbourhood of the nearest neighbour itself

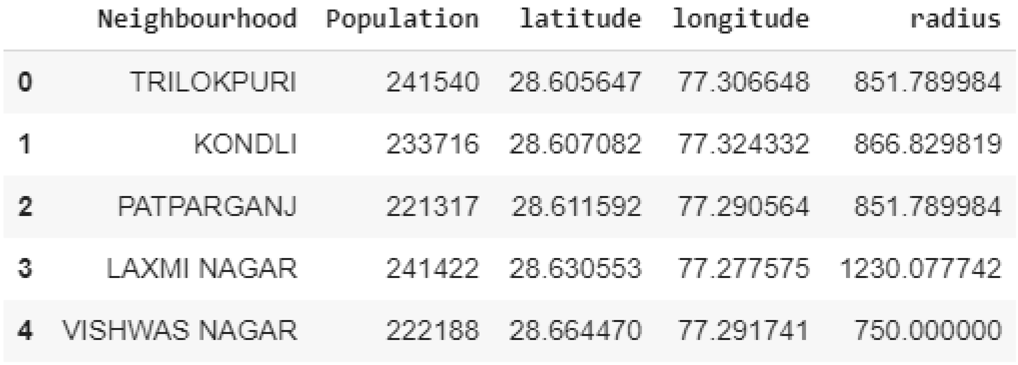
#NOTE: This definition of Neighbourhood creates some problems which will be discussed in 5

3.5 Finding Nearest Neighbours and their distances

The NearestNeighbour Algorithm from scikit learn is used to find the nearest neighbour and haversine formula is used to find the distances between them.

#NOTE: The NearestNeighbour is and unsupervised algorithm that uses Euclidean/Minkowski distance to measure between two points, which will result in wrong calculation of distances, but for our purposes it is fine since we are interested in knowing the nearest neighbour rather than the distance itself.

The resulting distances is:



	Neighbourhood	Population	latitude	longitude	radius
0	TRILOKPURI	241540	28.605647	77.306648	851.789984
1	KONDLI	233716	28.607082	77.324332	866.829819
2	PATPARGANJ	221317	28.611592	77.290564	851.789984
3	LAXMI NAGAR	241422	28.630553	77.277575	1230.077742
4	VISHWAS NAGAR	222188	28.664470	77.291741	750.000000

3.5 Venue Data

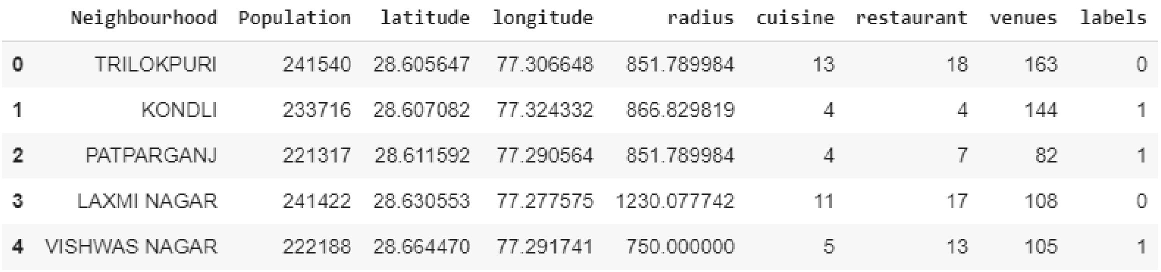
First an input was taken about the cuisine of the restaurant to be opened

Foursquare API was used to find the venue information around the neighbourhood in the radius as defined in section 3.5

Then the number of restaurants, restaurant of that particular cuisine (Indian in this case) and the number of venues people visited other than restaurants was stored in a DataFrame

#NOTE: The intent in the API call was kept to checkin rather than browse to get the venues people visited and not just the venues that are there. Also limit was kept higher than the probable number of venues in that area.

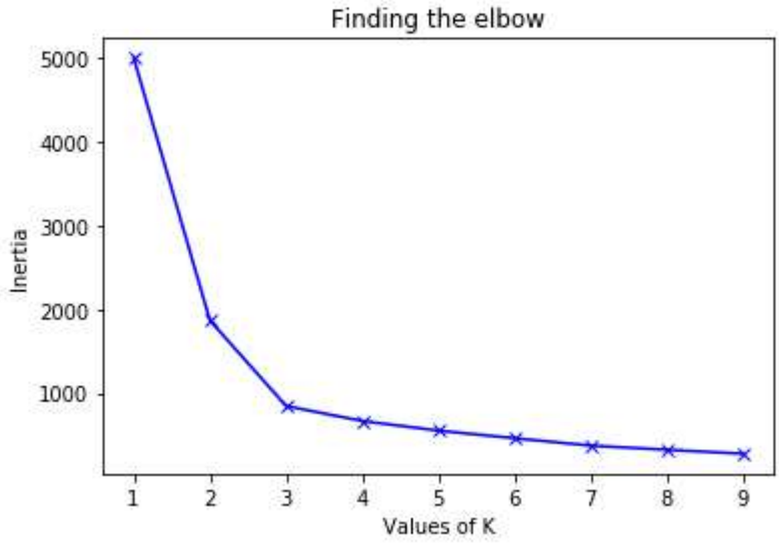
The DataFrame that resulted is:



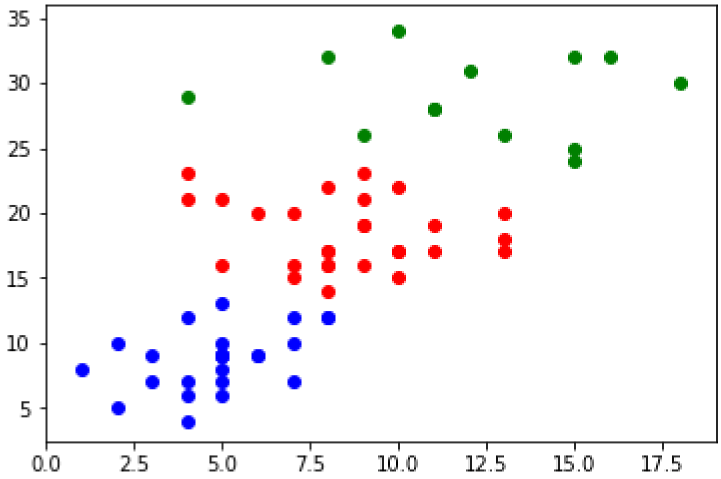
	Neighbourhood	Population	latitude	longitude	radius	cuisine	restaurant	venues	labels
0	TRILOKPURI	241540	28.605647	77.306648	851.789984	13	18	163	0
1	KONDLI	233716	28.607082	77.324332	866.829819	4	4	144	1
2	PATPARGANJ	221317	28.611592	77.290564	851.789984	4	7	82	1
3	LAXMI NAGAR	241422	28.630553	77.277575	1230.077742	11	17	108	0
4	VISHWAS NAGAR	222188	28.664470	77.291741	750.000000	5	13	105	1

3.6 Machine Learning

The areas were clustered using KMeans cluster algorithm from scikit learn according to number of restaurants and number of restaurants of that particular cuisine. Elbow method is used to fin optimum k(here it is 3) such that the error is less but the model is not overfitted.

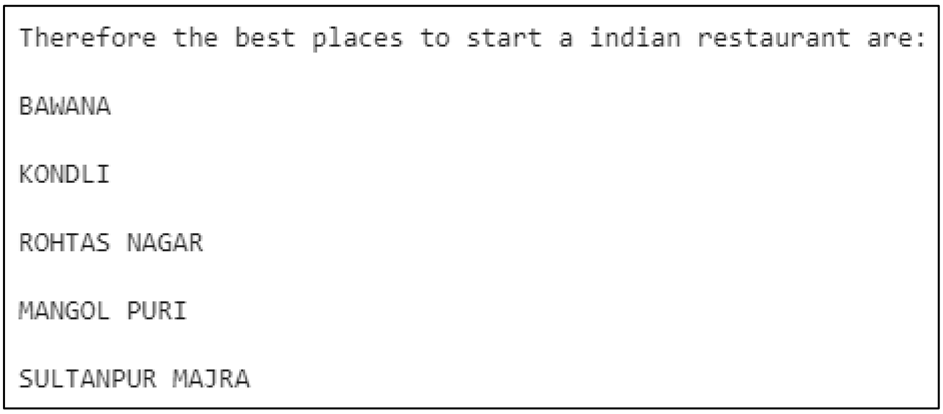


The cluster closest to the origin is selected . Then the items of this cluster were sorted according to the number of venues



4. Result

The result of the model was :



```
Therefore the best places to start a indian restaurant are:

BAWANA

KONDLI

ROHTAS NAGAR

MANGOL PURI

SULTANPUR MAJRA
```

5. Discussion

This Model has a lot of assumptions that can be sources of error:

- It doesn’t take into account other factors that make a restaurant successful/unsuccessful such as-price, demographic, renting , electricity etc.
- It also assumes that the demand for all type of cuisine is equally in demand at every neighbourhood of the city and hence less number of that already existing means supply is needed to fulfil that demand, however it may be that the reason for less number of restaurants of a cuisine is due to its low demand in that are.

- The definition of Neighbourhood as described in section 3.4 doesn’t include a lot of area of the city where there might be higher probability of profit.

6. Conclusion

I have tried to find out the best location to start a restaurant of a cuisine by comparing number of restaurant and other venues in that area, this could help restauranteurs in identifying places to open their businesses.