Data Science Capstone Project

Battle of Neighbourhoods

A report by
Smriti Chaudhury

Table of Contents

Introduction

In this project, the chosen city is New Delhi. One of the country's largest urban agglomerations, Delhi sits alongside the *Yamuna River*, the tributary of *Ganges River,* about 100 miles south of the *Great Himalayas.* Delhi is the administrative, legislative and judicial capital of India and surrounds the metropolitan region as well as surrounding rural regions.

It is of great historical significance as an important commercial,transport and cultural hub as well as the political centre of India. The whole delhi is divided into Old Delhi and New Delhi. New Delhi is characterized by wide straight avenues, with trees in double rows on either side which connects various points of interest.

This is the main east-west axis; it divides New Delhi into two parts, with a large shopping and business district, Connaught Place, in the north and extensive residential areas in the south.

I have lived in Delhi since my birth and have travelled to different places in the city. It comprises people coming from various other states having different religions who live in harmony which is one of the great things about the city. Many people have migrated from rural areas and urban areas in search of livelihood and job opportunities. It has many landmarks from historical monuments like *Lal Quila (Red Fort), Qutub Minar, India Gate, Purana Quila, Humayun's Tomb* to famous temples like *Lotus Temple, Akshardham Temple, ISKCON Temple* and various famous attractions like *Agrasen Ki Baoli, National Museum, National Zoological Park, National Rail Museum etc,.*

Delhi cuisine comprises all types of dishes. Centuries of global trade, colonization and conquest have made the city one of the most multicultural cities in the world. Connaught Place, popularly called CP, is the mecca of multicultural cuisine, from momos to samosas, fruit chats to bhelpuri, biryanis to pizzas and fine dining restaurants, CP has it all.

The city has absorbed, over the centuries, settlers, and visitors from across the globe. The emperors, the nobles, the viceroys and the sahibs all provided generous patronage to the cuisine of Delhi and contributed the cultivation of fine taste. Exploring the melting pot of the city can be a fascinating and rewarding experience.

Objective

The objective of this project is to choose the best neighbourhood where a Restaurant can be set up. Delhi is a hotspot for Continental, Thai, Mexican and Chinese food as well. Keeping all this in mind, the focus of this capstone project will be to:

- Choose a neighbourhood among the busiest ones popularity wise and distance wise.
- The stakeholders who want to open a different cuisine restaurant such as Italian, Thai or Mexican around famous attractions of the city would be interested in the results of this project.

Data

As in any data science project, collection and managing data is a huge task. There should be a vast amount of data to apply adequate machine learning algorithms and extract valuable information from the data. As we cannot find all data sets, we'll need to extract data from web pages, i.e. *Web Scraping* and convert it into a dataset.

Based on the criteria above, factors which will influence the decision of choosing a neighborhood are:

- Number of existing multi cuisine restaurants (if any present)
- Distance of neighborhood from the centre of the city
- Popular attractions in the neighborhood

To compute the dataset, the following data will be needed :

1. List of neighbourhoods in Delhi.
   Source : https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi
2. Coordinates of all neighbourhoods and venues
   Source : GeoPy Nominatim Geocoding
3. Number of restaurants and their type and location in every neighbourhood.
   Source: Foursquare API https://developer.foursquare.com/

Methodology

In this project, the first step will be to collect the data. Gathering data from webpages to build the dataset of neighborhoods in the city. Then, using a geopy nominatim library, extract the longitudes and latitudes of the neighborhoods of that dataset.

	index	Neighborhood	Latitude	Longitude
0	0	Adarsh Nagar	28.714401	77.167288
1	1	Ashok Vihar	28.699453	77.184826
2	2	Azadpur	28.707657	77.175547
3	3	Bawana	28.799660	77.032885
4	5	Dhaka	28.708698	77.205749
5	6	Jahangirpuri	28.725972	77.162658
6	7	Karala	28.735140	77.032511
7	8	Keshav Puram	28.688926	77.161683
8	9	Kingsway Camp	28.698778	77.204626
9	10	Kohat Enclave	28.698041	77.140539
10	11	Model Town	28.702714	77.193991
11	12	Narela	28.842610	77.091835
12	13	Pitam Pura	28.703268	77.132250
13	14	Rani Bagh	28.685982	77.132524
14	15	Rithala	28.720806	77.107181
15	17	Shalimar Bagh	28.717453	77.150867
16	18	Shakti Nagar	28.679790	77.194914
17	19	Bara Hindu Rao	28.659518	77.205010
18	20	Chandni Chowk	28.655983	77.232194

Fig 4.1 Dataset of Neighborhoods with latitudes and longitudes alongside.

Using this data, a folium map of the Delhi neighbourhoods marked on it will be created.

The second step will be to explore each of neighbourhoods and their venues using Foursquare location data. The venues of the neighbourhoods will be analyzed in detail and patterns will be discovered. Observing the data, it is a segmentation and clustering problem of Machine Learning. This discovery of patterns will be carried out by grouping the neighbourhoods using k-means clustering.

Following this, each cluster will be examined and a decision will be made regarding which cluster fits the shareholder's requirements. The factor that will determine this is the frequency of occurrence of restaurants and other food venues within the cluster.

We will use a for loop to see what is the adequate number of clusters that should be formed by silhouette coefficient.The technique provides a succinct graphical representation of how well each object has been classified. The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters.

We get ‘2’ as the value to make clusters in our project.

Once a cluster is picked, the neighbourhoods in that cluster will be investigated with regards to the number of Italian restaurants in its vicinity. The ones that fit the requirements will be further explored and shortlisted based on how small their respective distances to the center or Delhi are. The results of the analysis will highlight potential neighbourhoods where a multi cuisine restaurant may be opened based on geographical location and proximity to competitors. This will only serve as a starting point since there are a lot of other factors that influence such a decision.

Results

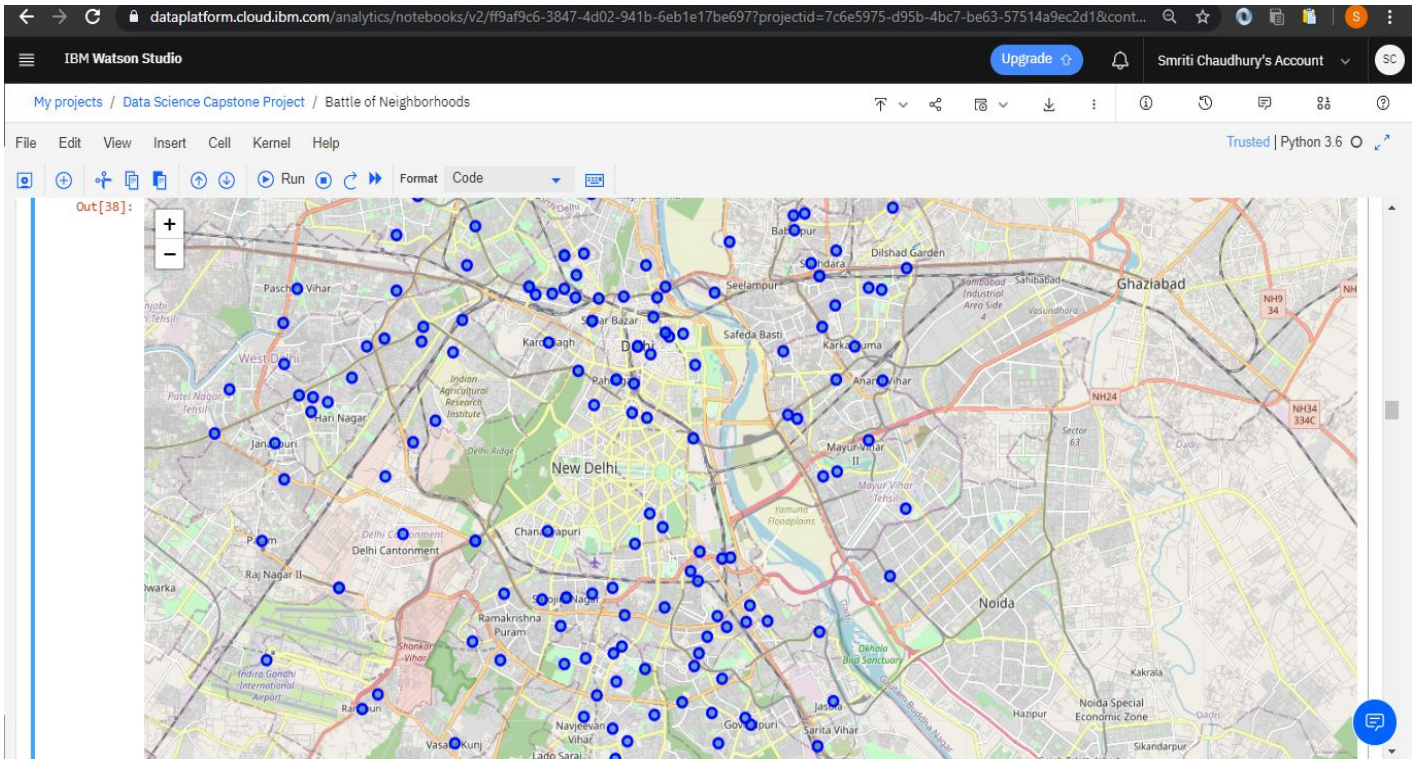

Fig 5.1 Map of Delhi,India showing all the neighborhoods marked on it

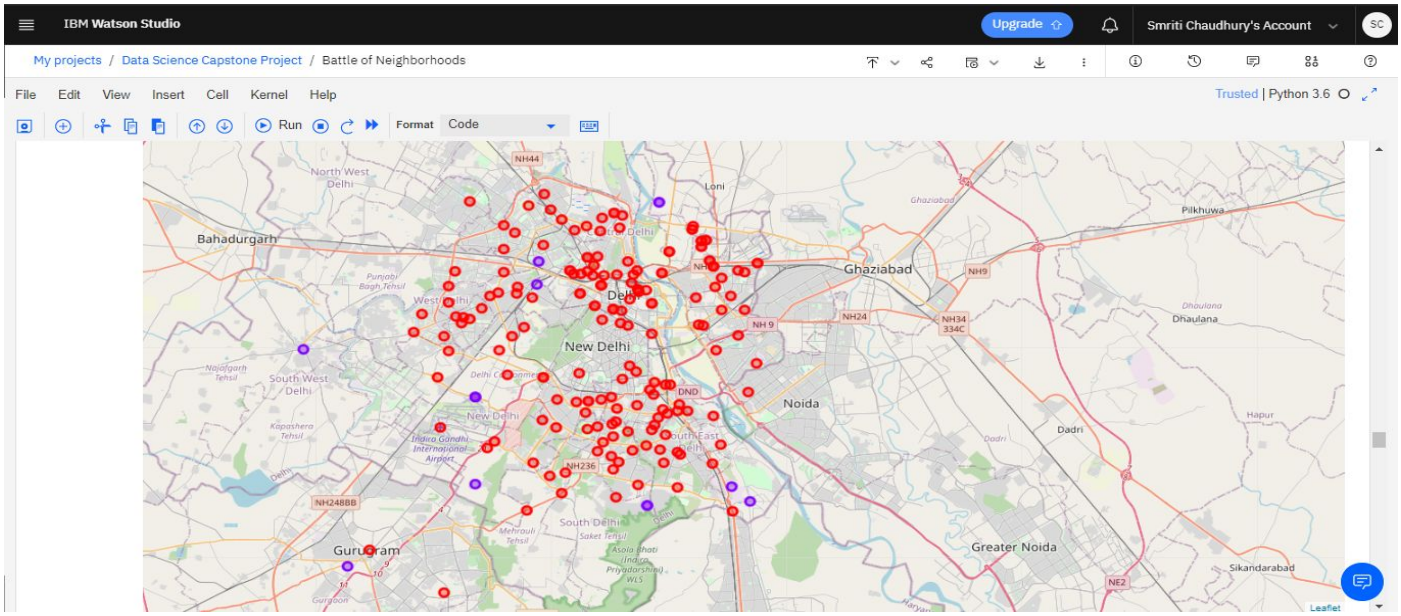

Fig 5.2 Clusters distributed all around Delhi

Two clusters formed around the city as per Silhouette Coefficient and neighborhoods are color coded accordingly.

---------- Adarsh Nagar ----------

	name	categories	distance	lat	lng
0	Chai Point	Tea Room	641	28.709372	77.170490
1	Giani's	Ice Cream Shop	754	28.717900	77.173907
2	Pahalwan Dhaba	Indian Restaurant	475	28.714594	77.172155
3	Standard Ice Cream Falooda	Dessert Shop	805	28.710360	77.160440
4	Moti Dhaba	Indian Restaurant	650	28.710008	77.171688
5	Republic of Chicken	Fried Chicken Joint	728	28.711205	77.160776
6	Giani's, Shalimar Bagh, Delhi	Ice Cream Shop	621	28.709142	77.169425

Fig 5.3 List of sorted venues with their distances to the city and their latitudes and longitudes.

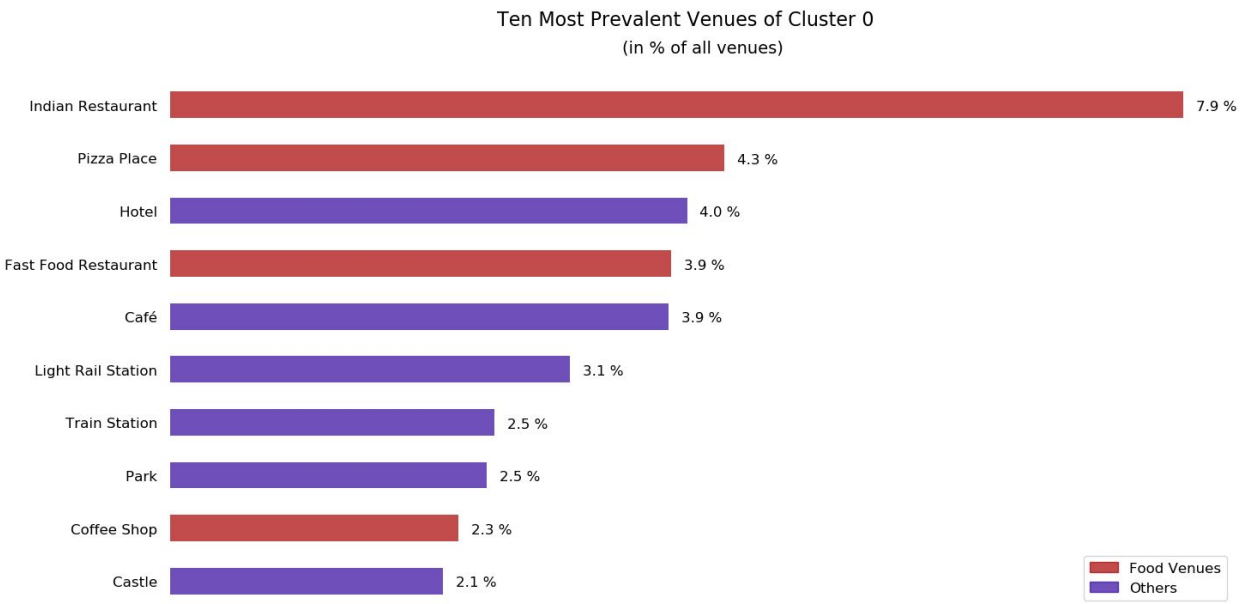

Fig 5.3 Ten Most Prevalent Venues of Cluster 0

Neighborhoods in Cluster 0 have many famous venues, landmarks and attraction spots as compared to Cluster 1 neighborhoods.

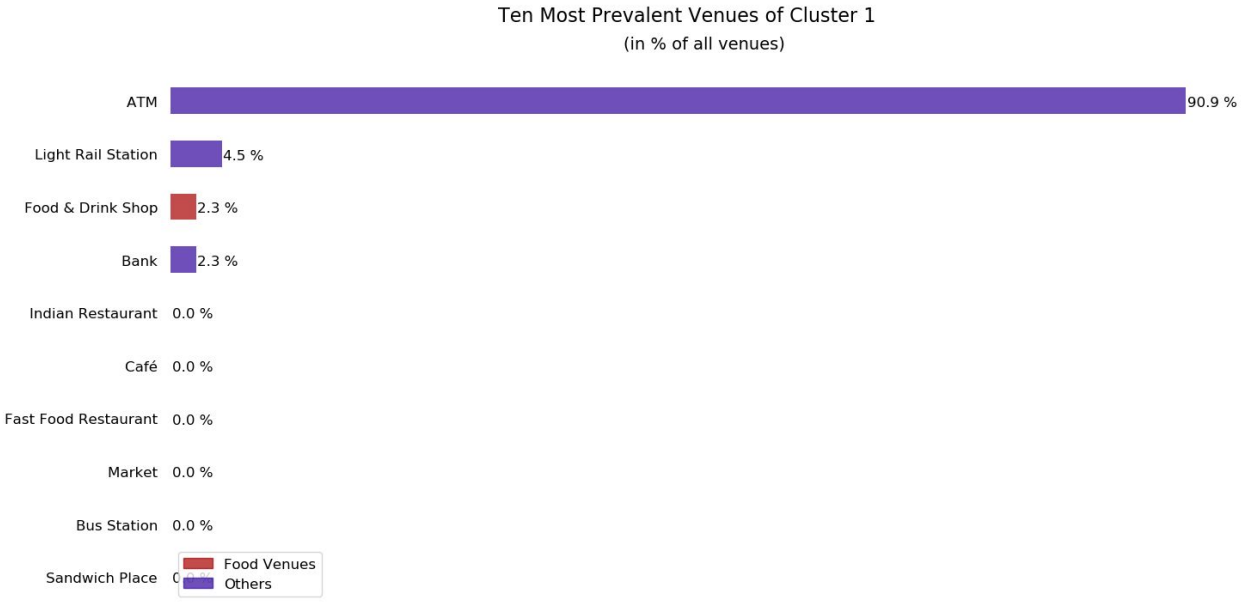

Fig 5.4 Ten Most Prevalent Venues of Cluster 1

Discussion

Looking at the map of Delhi, Cluster 0 has a proximal number of neighborhoods and an efficient number of centres are there to attract many customers for setting up a multi cuisine restaurant. Even though Cluster 0 has restaurants and cafes but the main idea behind this is to set up a different cuisine restaurant like Italian, Mexican, Thai, or Mughlai. Top 5 venues are listed in Cluster 0 for each neighborhood and the neighborhoods which have many restaurants already are excluded.

Out[98]:

	Neighborhood	Latitude	Longitude	Distance from Delhi center (in km)
0	Sadatpur	28.651718	77.221939	0.000000
1	Kotwali	28.651718	77.221939	0.000000
2	Chawri Bazaar	28.649265	77.226515	0.523282
3	Chandni Chowk	28.655983	77.232194	1.107369
4	Chandni Chowk	28.655983	77.232194	1.107369
5	Dariba Kalan	28.654602	77.233379	1.161517
6	Old Delhi Railway Station	28.660905	77.227715	1.166768
7	New Delhi Railway Station	28.640282	77.220410	1.280367
8	Paharganj	28.641499	77.214061	1.371925
9	Paharganj	28.641499	77.214061	1.371925
10	Lahori Gate	28.655787	77.238720	1.698815
11	Tis Hazari	28.667163	77.216631	1.793786
12	Kashmere Gate	28.666814	77.229055	1.816565
13	Bara Hindu Rao	28.659518	77.205010	1.865711
14	Bara Hindu Rao	28.659518	77.205010	1.865711
15	Daryaganj	28.646090	77.243048	2.152835
16	Pul Bangash	28.666407	77.207416	2.162380
17	Kashmiri Gate	28.669977	77.232059	2.257658
18	Connaught Place	28.631383	77.219792	2.270845
19	Jhandewalan	28.644319	77.199917	2.301103

Fig 6.1 Dataset of Cluster 0 neighborhoods with their distances,latitudes and longitudes.

The above figure shows all the neighborhoods in Cluster 0 and enables the stakeholders and businessmen to choose a location as per their convenience and requirements by looking at each neighborhood selected in Cluster 0 which have greater number of famous attractions, landmarks and most visited places to attend so the restaurant incomes can go up.

Conclusion

The objective of this project was to identify the best potential neighbourhoods in Delhi, India where a multi cuisine restaurant can be set up. All the required neighbourhood data was either scraped of the internet or obtained using a geolocator. After the neighbourhoods were visualized on a folium map, their venues were explored using Foursquare location data. Based on the frequency of occurrences of different venue types, the neighbourhoods were divided into two groups with the help of k-means clustering. The clusters were examined and the best one in which a restaurant could be set up was chosen. The neighbourhoods were filtered further based on proximity to existing restaurants and distance from the center of the city. As touched upon earlier, the results of the analysis highlight potential neighbourhoods where a multi cuisine restaurant may be opened solely based on geographical location and proximity to competitors. This will only serve as a starting point in the overall investigation since there are a lot of other factors - availability of commercial spaces, appeal of each location, proximity to major roads, access through public transport, etc. - that influence such a decision.