Capstone Project – The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

Analysis and clustering of neighborhoods in Delhi, India for opening a Chinese restaurant

Rohan Karthikeyan

August, 26, 2020

1. Introduction: Business Problem

Delhi, officially known as the National Capital Territory of Delhi (NCT) is a city and union territory of India containing New Delhi, the capital of India. According to the 2011 Indian census, the population of just the city of Delhi was over 11 million, the second-highest in India after Mumbai. The city is of great historical significance as an important commercial, transport, and cultural hub, as well as the political center of India.

As a resident of Delhi for some time now, it would be an understatement to say that the people of Delhi love street food and enjoy other cuisines. The dearth of food habits among the city's residents have created a unique style of cooking which has become popular world over, such as Kebab, biryani and tandoori. But, talking about international cuisines, there may be no doubt that Chinese cuisine is the favorite among so many Delhiites.

There is also no overlooking the fact that Delhi is home to so many educational institutions of repute such as the Indian Institute of Technology, Delhi University, among many others. This attracts a lot of students from Indian cities as well as from other countries, including students from China. Consequently, Chinese restaurants are of huge demand especially in areas which has a larger concentration of Chinese populations.

We will try to analyze and select the best locations in Delhi to open a new Chinese restaurant. Using data science methodology and tools such as data analysis and visualization, we will try to provide the answer to these business questions:

- What is / are the best locations for opening a Chinese restaurant in Delhi?
- In what neighborhood should we open our restaurant to have better chances of success?

2. Data acquisition and cleaning

Based on the definition of our problem, we will need the following data:

- All the neighborhoods of Delhi
- Venue data, particularly related to restaurants. This data will be used to perform further analysis of the neighborhoods.
- Latitude and longitude of the neighborhoods

2.1 Data sources

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

- Geo coordinates of Delhi are obtained using Geopy Nominatim
- Number of restaurants and their type and location in every neighborhood is obtained using the Foursquare API

GET https://api/foursquare/com/v2/venues/explore

Following parameters are passed to the Foursquare API in addition to the Client ID and Client Secret:

ll – Latitude and longitude of the location

radius – 1000

limit - 100

- List of all neighborhoods of Delhi is obtained by web scraping the Wikipedia page: Neighborhoods of Delhi

2.2 Data cleaning

The list of neighborhoods of Delhi scraped from the Wikipedia page was converted into a data-frame. However, we observe that there are a lot of redundancies, because of the webpage's HTML layout. So, this required some cleaning.

On adding the latitude and longitude values to create a new data-frame, we observe that some areas had 0 latitude and 0 longitude. Since there were only few entries (around 6) like that, we can just remove them.

2.3 Data selection

The Get Venue Explore endpoint of Foursquare API is used to retrieve the list of all venues along with the location and other details. This helps us to identify the areas with more footfall and then explore if there are any Chinese restaurants nearby.

One could also first collect the list of all Chinese restaurants (however, this uses the Get Venue Search endpoint of Foursquare API) in the neighborhoods of Delhi and then analyze the locations, cluster them and find out the areas where there is a lesser concentration of restaurants.

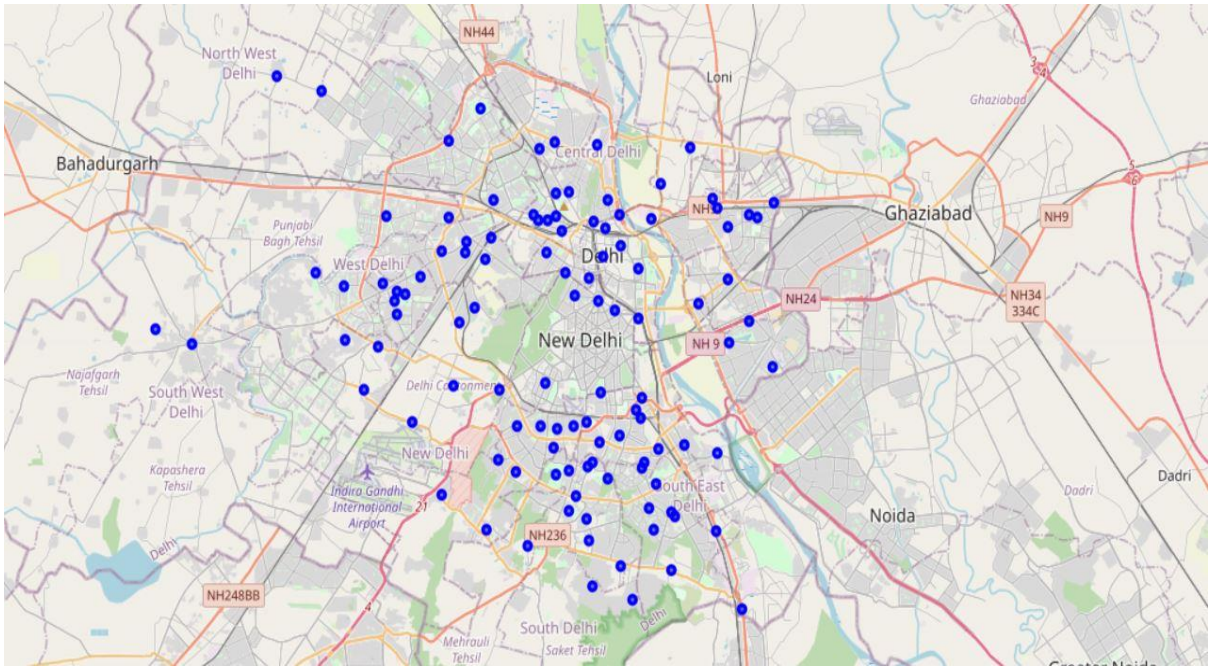
3. Methodology

a. We first collect data on the neighborhoods of Delhi from the Web. We need to scrape data from the above Wikipedia page, since there are no relevant datasets available for this. The location coordinates of each neighborhood will be obtained from the GeoPy Nominatim geolocator and appended to the neighborhood data.
b. Foursquare API will be used to explore each of the neighborhoods and their venues. The venues of the neighborhoods are analyzed in detail and patterns will be discovered. The discovery of patterns will be carried out by grouping the neighborhoods using k-means clustering.
c. Each cluster will be examined and a decision will be made regarding which cluster fits the shareholder's requirements.

d. Finally, if there are multiple neighborhoods that fit these conditions, the data on rent of properties in the neighborhoods of Delhi can be used to influence the shareholder's decision. The results of the analysis will highlight potential neighborhoods where a Chinese restaurant may be opened based on geographical location and proximity to competitors.

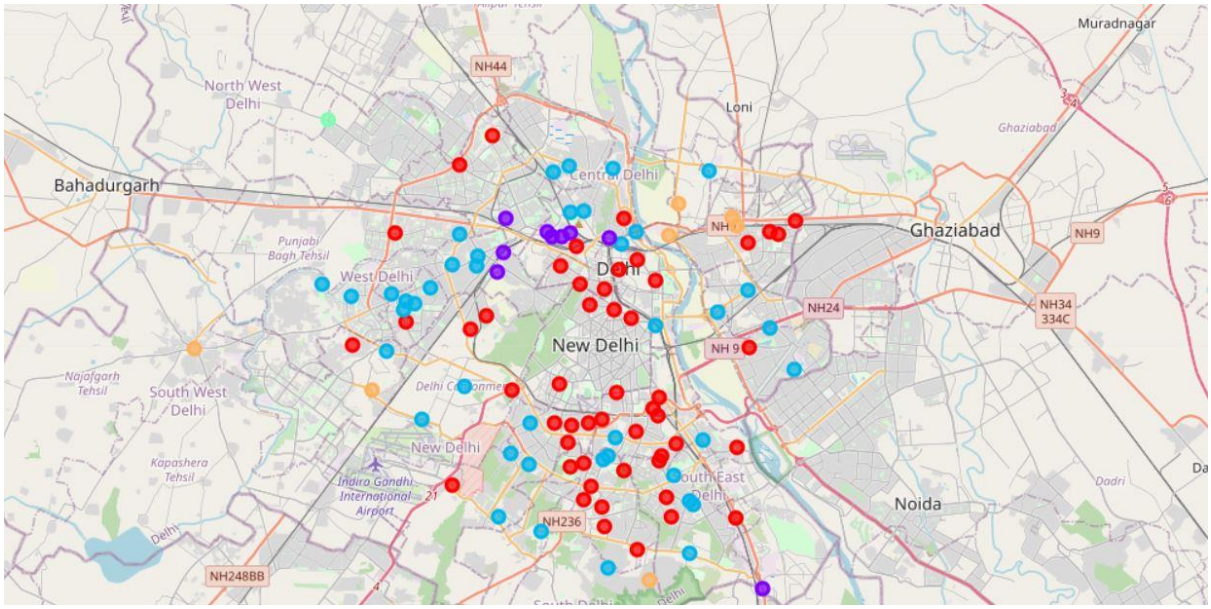
3.1 Exploratory data analysis

The merged data-frame is analyzed and plotted on a map using folium function. This helps us to get a fair idea of the location for the Chinese restaurants in the neighborhoods of Delhi.



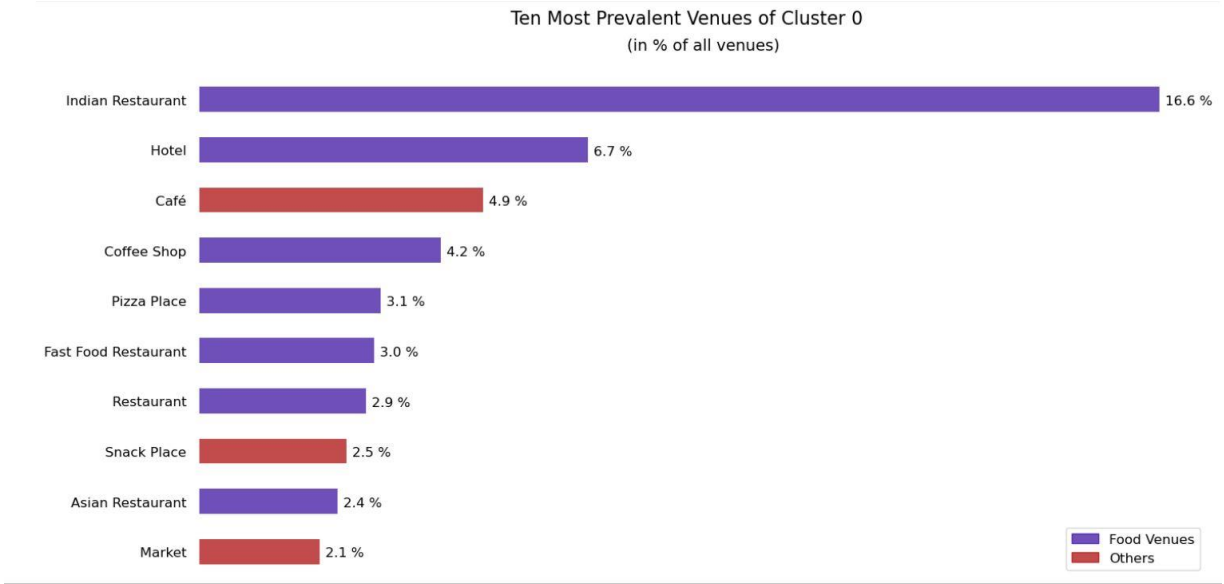
3.2 Machine Learning Algorithms

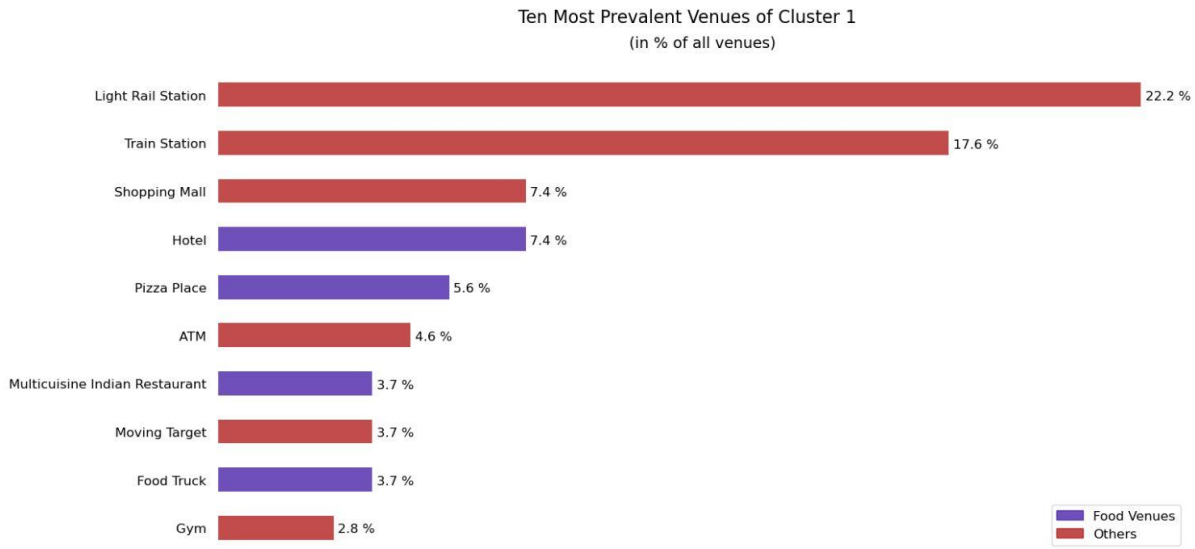
Now that we have got all the neighborhoods plotted on the map of Delhi, we cluster them using k-means clustering. Using the elbow method, we choose the optimum value of 5. The k-means clusters were plotted on a map using folium for easy visualization.

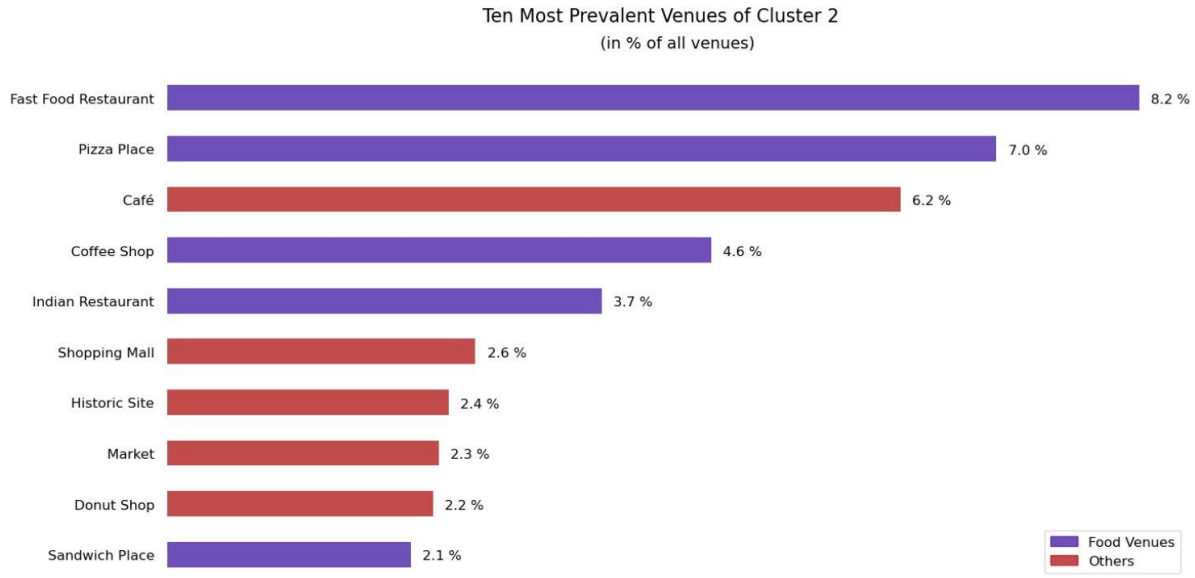


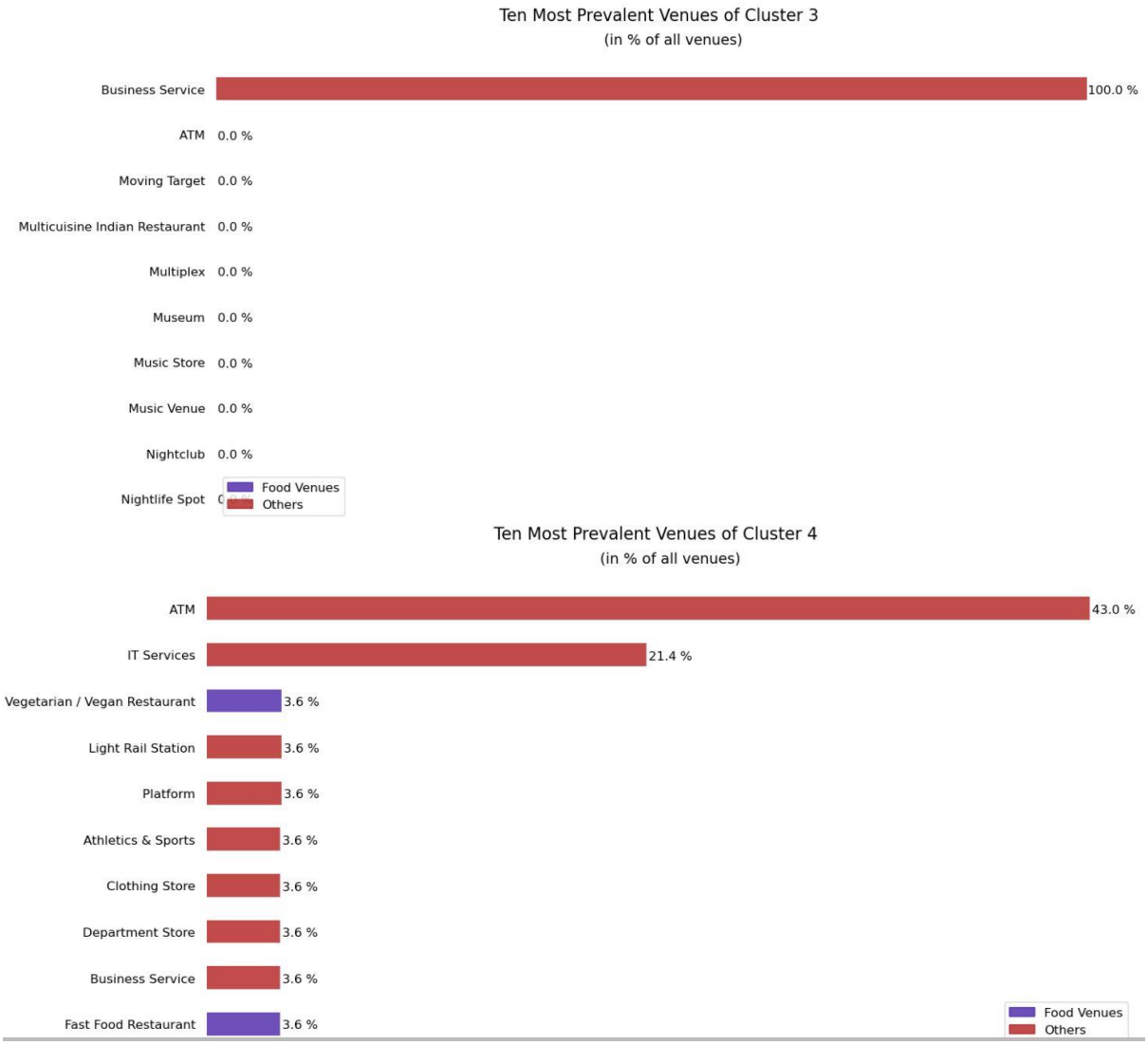
4. Results

We analyze which among the 5 clusters are suited for opening our restaurant. The charts for the top 10 most prevalent venues of each cluster is given below:




Ten Most Prevalent Venues of Cluster 1
(in % of all venues)
Light Rail Station
22.2 %
Train Station
17.6 %
Shopping Mall
7.4 %
Hotel
7.4 %
Pizza Place
5.6 %
ATM
4.6 %
Multicuisine Indian Restaurant
3.7 %
Moving Target
3.7 %
Food Truck
3.7 %
Gym
2.8 %
Food Venues
Others


Ten Most Prevalent Venues of Cluster 2
(in % of all venues)
Fast Food Restaurant
8.2 %
Pizza Place
7.0 %
Café
6.2 %
Coffee Shop
4.6 %
Indian Restaurant
3.7 %
Shopping Mall
2.6 %
Historic Site
2.4 %
Market
2.3 %
Donut Shop
2.2 %
Sandwich Place
2.1 %
Food Venues
Others



5. Discussion

We observe that there are many restaurants of different cuisines in Cluster 0. So, let's check if Cluster 1 is OK. Well, it seems that we might have our task cut out here, but let's keep exploring. Just as in Cluster 0, we have heavy competition in Cluster 2, so let's skip this. Cluster 3 doesn't seem a good choice, obviously. Yes, even setting up a restaurant in Cluster 4 seems good.

So, the competition is between cluster 1 and cluster 4. How can we make a decision? If we had some data on the rent of properties in Delhi by neighborhoods, we could utilize them here, however, it's not available. If we analyze the neighborhoods in the two clusters, we observe that places in cluster 1 are closest to the city center than those in cluster 4.

Consequently, we have greater chances of more footfall here. So, we are now going to explore the neighborhoods in cluster 1 which are:

- Gulabi Bagh
- Sarai Rohilla

- Shastri Nagar
- Tis Hazari
- Pandav Nagar
- Jor Bagh
- Badarpur
- Moti Bagh
- Patel Nagar

We then check if there are any Chinese restaurants in these places. It really comes as a sur prise to us that there are no Chinese restaurants in the neighbourhood. One could really lo ok for neighbourhoods among Cluster 1 that offer a combination of popularity among tour ists, closeness to city centre, strong socio-economic dynamics.

The final result of this is 9 zones (in Cluster 1) have largest number of potential new restau rant locations. This, of course, does not imply that those zones are optimal locations for a new restaurant! That's because it's entirely possible that there's a good reason for no Chin ese restaurants in any of those areas, reasons which would not make them suitable for a n ew restaurant regardless of lack of competition in the area. Recommended zones should t herefore be considered only as a starting point for more detailed analysis which could eve ntually result in locations which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

6. Conclusion

The purpose of this project was to identify areas in Delhi (one might add, closer to the city centre) in order to aid stakeholders in narrowing down the search for optimal location for a new Chinese restaurant. By calculating restaurant density distribution from Foursquare, we first identified that neighbourhoods in Cluster 1, some of which include: Badarpur, Gul abi Bagh, Jor Bagh, Patel Nagar, Shastri Nagar and Tis Hazari, would be the ideal location f or starting a Chinese restaurant.

The final decisions on our optimal restaurant location will be made by stakeholders based on specific characteristics of neighbourhoods and locations in every recommended zone, t aking into consideration additional factors like: attractiveness of each location (proximity t o park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighbourhood etc.