IBM Applied Data Science Capstone Project

Opening New Bar in Delhi, India

By

Kumar Sanu

August'20

Introduction

For many people, going at a place with their family and friends to have a good time and have good food and drinks are a good way to relax during weekends and holidays. These drinks also help people in reducing their depression and frustration occurred due to both professional and personal life. With this said, bars are becoming a common place for all sorts of people across the city. Whether you stay in your home or you are visiting a place, you should consider visiting a bar at least once a week or more. Many stressful things are happening in the world around us, so sometimes all one needs is to go to a place that serves good food and drinks. Hence the availability of bars became the requirement in the life of people.

And where there is demand of services, there is business. Hence business people can think to start their business. But to open any business there is many things to be taken in consideration.

Chances are we've come across the term 'location, location, location' and when it comes to opening a new bar, nothing is more important. Sure, we'll need a great concept, a first-rate product, and excellent customer service, but our bar's location will be the foundation of your business and have a major impact on its success. So for a project developer, location is the key factor to consider in case of opening the bar so that they can target the desired number of peoples.

Business Problem

The aim of this project is to help the project developers in selecting the best locations in Delhi, India to open bar. We will use Data Science Methodology and Machine Learning Techniques to solve this problem.

Basically this project will provide answer to this Business Problem:-

If someone wants to open a new bar in Delhi, the capital of India, where they should open?

Data

We need following data to solve this problem:-

1. List of neighborhoods in Delhi. The scope of the project will be confined to the city of Delhi.
2. Latitude and longitude of each neighborhoods. This is required in plotting the map and getting venues information.
3. Venue data especially related to bar, which is the area of interest for us. This will be used in clustering those neighborhoods.

Source Information and Methods for extraction

We will use the Wikipedia page for getting data which consists of 9 Boroughs comprising of 117 Neighborhoods of Delhi. (https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi)

1. We will use web scraping techniques to extract the list of neighborhoods from wiki page with help of python packages.
2. We will use geopy package to find latitude and longitude of each neighborhoods.
3. We will use Foursquare API to get all venues related information of neighborhoods using their latitude and longitude values.

Foursquare API will provide many categories of data in those neighborhoods. Out of those categories, we are highly interested in data related to bar in order to solve this problem.

This project will make use of many data science skills from data extracting (web scraping), Working with API(Foursquare),data cleaning, data wrangling, data visualization(Folium) to Machine Learning(K-means Clustering). We will discuss these all methodologies and the steps taken in detail while implementing it.

Target Audience of this Project

Basically this project is for those project developers and investors who want to new open bar in Delhi, the capital of India.

For 2020-21, the city-state projected a total revenue of around Rs 6,279 crore through the sale of liquor, compared to the revised estimate of around Rs 5,480 crore for the previous fiscal. In 2018-19, Delhi earned around Rs 5,007 crore through the sale of liquor, according to budget documents.

In terms of income, alcohol accounts for around 14.1% of the city government's total revenue projections for 2020-21.

Hence, there is great opportunity for project developers to do investment in opening bars in Delhi.

Methodology and Steps

First of all, we need list of neighborhoods in Delhi. And this information is readily available at Wiki page (https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi).

Neighbourhoods of Delhi

From Wikipedia, the free encyclopedia

Delhi is a vast city and is home to a population of more than 16 million people. It is a microcosm of India and its residents belong to varied ethnic, religious and linguistic groups. As the second-largest city, and the capital of the nation, its nine districts comprise multiple neighbourhoods. The large expanse of the city comprises residential districts that range from poor to affluent, and small and large commercial districts, across its municipal extent. This is a list of major neighbourhoods in the city and only pertains to the National Capital Territory of Delhi. It is not complete, and outlines the various neighbourhoods based on the different districts of the metropolis.

Contents [hide]

So in order to get this list, we will use web scraping techniques using python requests and BeautifulSoup packages. This list contains only names of neighborhoods. After using web scraping we will have this dataframe:-

	Borough	Neighbourhood
0	North West Delhi	Adarsh Nagar
1	North West Delhi	Ashok Vihar
2	North West Delhi	Begum Pur
3	North West Delhi	Karala
4	North West Delhi	Model Town

But Foursquare API require geographical coordinates of all those neighborhoods in order to get venues information in those neighborhoods. We will use geopy package of python to convert those neighborhood addresses in terms of latitude and longitudes. After gathering this data, we will put this data in pandas dataframe.

	Borough	Neighbourhood	Latitude	Longitude
0	North West Delhi	Adarsh Nagar	28.614193	77.071541
1	North West Delhi	Ashok Vihar	28.699453	77.184826
2	North West Delhi	BegumPur	28.536978	77.209941
3	North West Delhi	Karala	28.735140	77.032511
4	North West Delhi	Model Town	28.702714	77.193991

Now we are good to go as we have all data handy which is required to get the venue information in each neighborhood of Delhi.

Then we will visualize this data in a map using Folium package. So here we will plot map of Delhi and superimpose those neighborhoods on it. By this way, we can also verify that those geographical coordinates obtained through geopy packages are plotted correctly. Below map is the map of Delhi on which blue circle markers are representing all the neighborhoods of it.

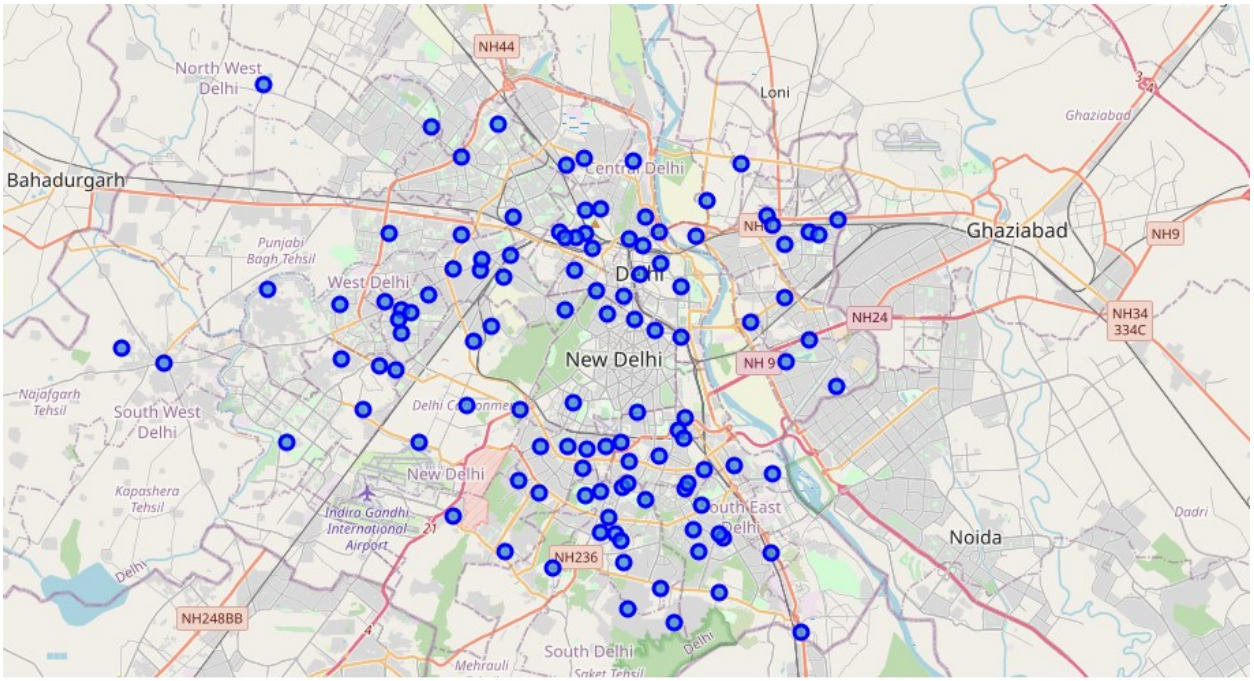

Fig. Map of Delhi and its neighborhoods (blue circle marker)

Next we will use Foursquare API to get top 100 venues in all neighborhoods within radius of 2000 m. For using Foursquare API, we need its ClientId and ClientSecret. So we will sign up a developer account on Foursquare. We then make API calls passing latitude and longitude of the neighborhoods along with the account ClientId and ClientSecret to the Foursquare. Foursquare will return the venue data in json format. We will extract Name, Latitude, Longitude, Category of each venue of each neighborhood. And finally adding these venue information and dropping Borough column to the previously created dataframe. The structure of this dataframe will become as follows:-

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adarsh Nagar	28.614193	77.071541	Shree Maakhan	28.625825	77.068608	Indian Restaurant
1	Adarsh Nagar	28.614193	77.071541	Cafe Coffee Day	28.622388	77.088353	Café
2	Adarsh Nagar	28.614193	77.071541	Domino's Pizza	28.630000	77.080000	Pizza Place
3	Adarsh Nagar	28.614193	77.071541	Cafe Coffee Day	28.629794	77.079771	Coffee Shop
4	Adarsh Nagar	28.614193	77.071541	Subway	28.630026	77.079541	Sandwich Place

With this data, we can calculate the total number of venues in each neighborhood. Also we can get the number of unique categories of venues. Then we will analyze each neighborhood by grouping rows by neighborhood and by taking the mean of the frequency of occurrence of each category. By doing so, we are also preparing the data for clustering. Since we are interested in analyzing the "Bar" data so we will filter our data related to bar category of neighborhoods. Final dataframe looks like this:-

	Neighbourhood	Bar
0	Adarsh Nagar	0.0
1	Alaknanda	0.0
2	Anand Vihar	0.0
3	Ashok Nagar	0.0
4	Ashok Vihar	0.0

Now we are left with the dataframe consisting of two columns i.e. neighborhoods name and Frequency of occurrence of Bar in the respective neighborhoods. Then we will apply K-means clustering algorithm on this data to cluster those neighborhoods into 3 clusters based on their frequency of occurrence for "Bar".

The result will help us in identifying which cluster of neighborhood has high concentration of Bar and which has the least. Based on the occurrence of bar in different clusters, it will help us to answer the problem statement. That means it will tell us that in which neighborhood one should open the bar?

Results

The result from K-Means Clustering shows that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for "Bar".

`Cluster 0`: Neighborhoods with moderate concentration of Bar

`Cluster 1`: Neighborhoods with non-existence of Bar

`Cluster 2`: Neighborhoods with higher concentration of Bar

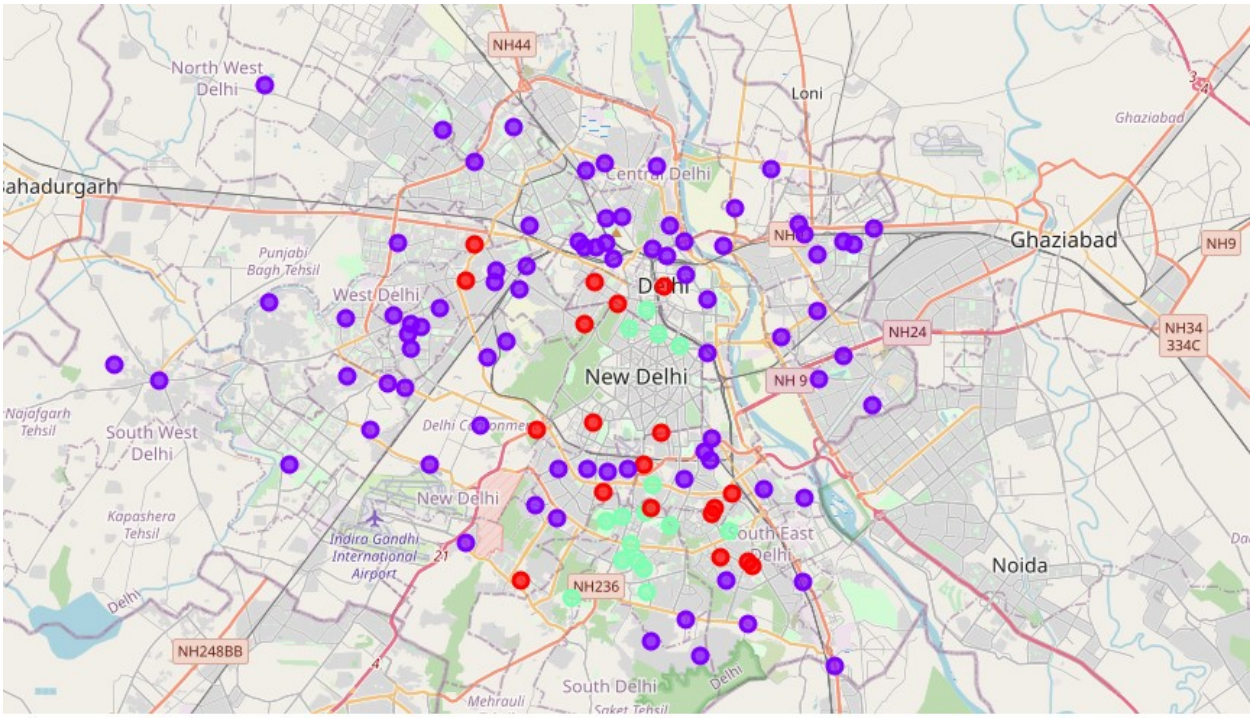

Fig. Map of Delhi and its neighborhoods(circle marker with different color)

Note: - In the above map, cluster 0 represented in red color, cluster 1 in violet color and cluster 2 in aquamarine color.

Discussion

As per the cluster map above in the result section, we can say that most of the bar are concentrated in South Delhi and New Delhi with highest number in **`Cluster 2`** and moderate in **`Cluster 0.`**

Descriptive Analysis

It's clearly visible from the above map that there is great opportunity and high potential areas of opening bar in `Cluster 1` neighborhoods which covers mainly West Delhi, East Delhi and North Delhi as there is no competition due to non-existence of bar there. If we talk about `Cluster 2` neighborhoods which covers mainly South Delhi and New Delhi which have high concentration of bar, bars are more likely to suffer from high competition due to oversupply of products. Meanwhile one can open bar in `Cluster 0` which covers mainly some parts of South Delhi and Central Delhi who are able to stand out in moderate competitions by proving better services. Below are the screenshots of examining 3 clusters.

Analysing Cluster 0

```
In [34]: dl_merged[dl_merged['Cluster_Label']==0]
```

Out[34]:

	Borough	Neighbourhood	Latitude	Longitude	Bar	Cluster_Label
13	North Delhi	Kotwali	28.651718	77.221939	0.023256	0
21	North Delhi	Wazirabad	28.433762	77.087757	0.019231	0
30	Central Delhi	Jhandewalan	28.644319	77.199917	0.014706	0
31	Central Delhi	Karol Bagh	28.652998	77.189023	0.024390	0
34	Central Delhi	Rajinder Nagar	28.635786	77.184140	0.019608	0
36	New Delhi	Chanakyapuri	28.594677	77.188521	0.013889	0
39	New Delhi	INA Colony	28.577281	77.212649	0.014706	0
53	South Delhi	Chittaranjan Park	28.538752	77.249249	0.033708	0
55	South Delhi	East of Kailash	28.557032	77.244614	0.030000	0
56	South Delhi	Govindpuri	28.535156	77.263794	0.029412	0
57	South Delhi	Greater Kailash	28.559157	77.246018	0.030000	0
64	South Delhi	Kalkaji	28.537070	77.261805	0.028571	0
67	South Delhi	Lodi Colony	28.590702	77.220921	0.027778	0
70	South Delhi	Neeti Bagh	28.559251	77.216166	0.034884	0

Analysing Cluster 1

```
In [35]: dl_merged[dl_merged['Cluster_Label']==1]
```

Out[35]:

	Borough	Neighbourhood	Latitude	Longitude	Bar	Cluster_Label
0	North West Delhi	Adarsh Nagar	28.614193	77.071541	0.000000	1
1	North West Delhi	Ashok Vihar	28.699453	77.184826	0.000000	1
3	North West Delhi	Karala	28.735140	77.032511	0.000000	1
4	North West Delhi	Model Town	28.702714	77.193991	0.000000	1
5	North West Delhi	Narela	28.842610	77.091835	0.000000	1
6	North West Delhi	Pitam Pura	28.703268	77.132250	0.000000	1
7	North West Delhi	Rohini	28.716209	77.117074	0.000000	1
8	North West Delhi	Shalimar Bagh	28.717453	77.150867	0.000000	1
9	North Delhi	Civil Lines	28.676851	77.225030	0.000000	1
10	North Delhi	Gulabi Bagh	28.669649	77.194726	0.000000	1
11	North Delhi	Kamla Nagar	28.680344	77.202129	0.000000	1

Analysing Cluster 2

```
In [36]: dl_merged[dl_merged['Cluster_Label']==2]
```

Out[36]:

	Borough	Neighbourhood	Latitude	Longitude	Bar	Cluster_Label
2	North West Delhi	BegumPur	28.536978	77.209941	0.050000	2
33	Central Delhi	Paharganj	28.641499	77.214061	0.060000	2
35	New Delhi	Barakhamba Road	28.626952	77.229950	0.050000	2
37	New Delhi	Connaught Place	28.631383	77.219792	0.050000	2
38	New Delhi	Gole Market	28.633719	77.205627	0.060000	2
58	South Delhi	Green Park	28.555537	77.202497	0.050000	2
59	South Delhi	Gulmohar Park	28.557101	77.213005	0.040000	2
60	South Delhi	Hauz Khas	28.544256	77.206707	0.040000	2
61	South Delhi	Hauz Khas Village	28.553855	77.194713	0.050000	2
68	South Delhi	Malviya Nagar	28.533920	77.212447	0.050000	2
69	South Delhi	Mehrauli	28.521826	77.178323	0.090909	2
71	South Delhi	Nehru Place	28.549257	77.252952	0.040816	2
78	South Delhi	Saket	28.524411	77.213725	0.050000	2
82	South Delhi	Sarvodaya Enclave	28.537478	77.202089	0.040000	2
83	South Delhi	Siri Fort	28.552146	77.224698	0.060000	2
84	South Delhi	South Extension	28.568715	77.216896	0.041096	2

Limitations and Future Scope of Analysis

As we have limited our scope of analysis to only one parameter i.e. frequency of occurrence for bar in neighborhoods but there are other factors also which should be taken into consideration for selecting location to open a bar like:

- Our Style- Are we formal? Elegant? Casual? This can determine the type of customers you're likely to appeal to.
- Demographics – Different areas of our city appeal to different demographics. If we're looking to appeal to college students, opening near a university makes sense. If we're looking to attract higher-class customers, set up shop in the more affluent part of town.
- Accessibility and Parking – If we're attracting tourists, parking is less of an issue since they're likely to call a cab or use Ola or Uber. However, tourists are less likely to be repeat customers. Keep this in mind when choosing location.

- Zoning Restrictions – Can we open a bar here?
- Rent and Utilities Costs – Will we be able to make this up in sales with the type of customers you're going to draw in?

Unfortunately, we don't have these data available to neighborhood level that are required for the project. Future Research can devise a methodology to use all these factors in the analysis to find the best location. As we have used Foursquare free tier developer account, there are some limitations in number of API calls and result returned. So Future research can also use paid account in order to overcome these limitations and get better results.

Conclusion

In this Project, we have gone through the process of identifying Business Problem, specifying the data required, extracting and preparing data and finally applying K-means machine learning algorithm to help project developers in determining the best place in Delhi to open a new bar.

Hence this project recommends that if business people are thinking to open bar in Neighborhoods of Delhi, they should focus more on `Cluster 1` neighborhoods with no competition. If business people have good sales and marketing strategy, they can open bar in `Cluster 0` neighborhoods with moderate competition. Last but not the least, they should avoid considering `Cluster 2` which already have high concentration of Bar.

So answer to the business problem 'If someone wants to open a new bar in Delhi, the capital of India, where they should open?' is:-

Neighborhoods of Delhi which is in cluster 1 are the best places to open new bar.