Predicting Best Place to Build Café

In Delhi

1. Introduction

1.1. Background

Delhi is one of the largest city in India with geographic area of around 1484 Km^2. Being Capital of the country, it is also one of the most populous city in the country with total population of around 11 million people. Delhi has a large number of country's best colleges and also has the largest metro connectivity of the country.

1.2. The problem

All of the great Cafe are densely concentrated in one or two places in the city. So people usually have to move from one corner of the city to another just to hangout in a good cafe. Also opening a new cafe in the already existing competitive area will only increase the dense competition and decrease the profit. This project concentrates on what parameters all the Cafe exist in a certain area and then the goal is to find similar areas with similar parameters but with lesser Cafe or competition to solve the need of both Customers as well as Owners.

2. Data acquisition and cleaning

2.1 Data Sources

The population and all the neighborhood or sub division (as called in India) Data was taken from the government census website https://censusindia.gov.in

The location for the each sub division was taken with the help of geocoder library of python.

The data for the rest of the parameters was taken from Foursquare API

The parameters are as follows:

1) Nearby market- One of the most common reasons for people to hangout is shopping. So Nearby market can attract a large number of customers.

2) Nearby metro station- Opening a Cafe close to Metro station is always beneficial. As mostly students and younger generation travel by metro in the city, therefore the nearest cafe to a metro station will always have an upper hand in attracting more customers than the rest.

3) Population- The area which has higher population will always offer more customers. There is no point in building a cafe in a secluded corner of a city where no one might ever know that a Cafe exist.

4) Nearby Universities- Students always want to ditch their hostel food and hangout with their friends while bunking classes. Making a Cafe near Universities will offer a perfect place for them to hangout.

5) Nearby Cafe- There is no point of building a cafe that satisfy the above parameters but have countless competition in neighborhood. Therefore the number of existing Cafe will have a huge impact on our decision.

2.2. Data Cleaning

In data cleaning several tables were merged to present the data in the required form.

First from the census excel sheet, the data regarding only total population and total household of each sub district was extracted to form a new data frame.

Then with the help of geocoders library, Latitude and Longitude of each sub district was added to the data frame.

Then from the foursquare API, for each sub district, 100 nearby venues in a radius of 4 km were imported and a new data frame with sub district, nearby venues name, location of sib district and nearby venue category name was formed.

2.3. Feature Selection

There are total 146 venue categories and 1,767 venues around Delhi. For selecting the location to build a Café, the most crucial nearby venues are Nearby market, Nearby metro station, Nearby universities and other Nearby Café. Therefore these four categories were extracted from 146 venue categories and a data frame was formed.

3. Data Analysis

The data after the feature selection is taken for analysis. As analysis can be done on numbers therefore one hot encoding was done on data.

If a venue category is present in a sub division then it was given a value of 1 and if the venue is not present in a sub division then it was given a value of 0. Then base on this encoding, the mean of all venues of a venue category was taken for each sub division, the final data frame looked like this:

(Presenting only 5 sub divisions for example)

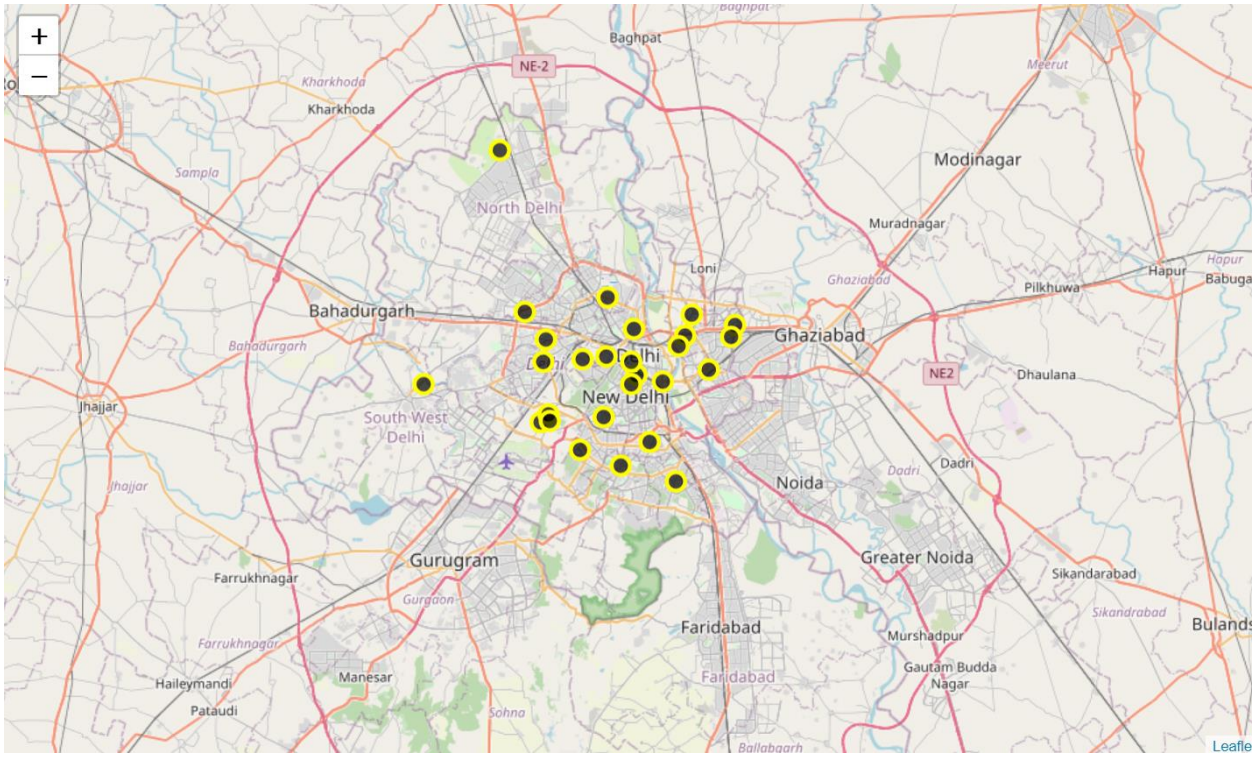
	sub division	Café	Metro Station	University	Market
0	Chanakya Puri	0.070000	0.0	0.01	0.02
1	Civil Lines	0.078431	0.0	0.00	0.00
2	Connaught Place	0.080000	0.0	0.00	0.01
3	Darya Ganj	0.090000	0.0	0.00	0.01
4	Defence Colony	0.080000	0.0	0.00	0.06

Then normalized population was added to this feature set to complete all the required parameters for building café.

The final data frame look like this

	sub division	Café	Metro Station	University	Market	Total population
0	Chanakya Puri	0.070000	0.0	0.01	0.02	0.000273
1	Civil Lines	0.078431	0.0	0.00	0.00	0.003059
2	Connaught Place	0.080000	0.0	0.00	0.01	0.000125
3	Darya Ganj	0.090000	0.0	0.00	0.01	0.001204
4	Defence Colony	0.080000	0.0	0.00	0.06	0.002834

To locate each sub division, a map with marker of each sub division was formed which is shown below:



4. Modeling

In this project K-Means Clustering algorithm is used to cluster the data set of sub divisions on the basis of 5 parameters that are, population, nearby market, nearby universities, nearby metro station and nearby café.

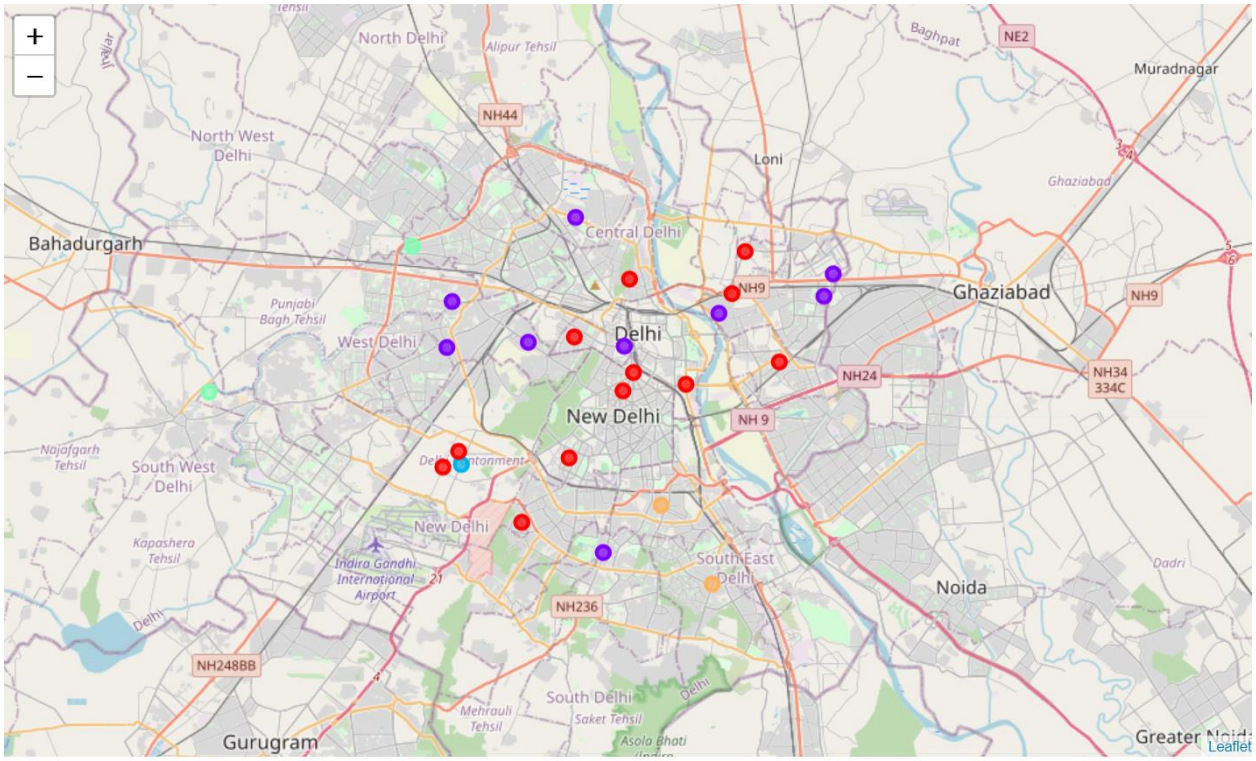
On the basis of above, the data was distributed into 5 clusters named as cluster 0, cluster 1, cluster 2, cluster 3 and cluster 4.

Then this clustering data was added to the data frame.

An example of the data frame formed is given below:

	Clusters	sub division	Café	Metro Station	University	Market	Total population
0	0	Chanakya Puri	0.070000	0.000000	0.010000	0.020000	0.000273
1	0	Civil Lines	0.078431	0.000000	0.000000	0.000000	0.003059
2	0	Connaught Place	0.080000	0.000000	0.000000	0.010000	0.000125
3	0	Darya Ganj	0.090000	0.000000	0.000000	0.010000	0.001204
4	4	Defence Colony	0.080000	0.000000	0.000000	0.060000	0.002834
5	2	Delhi Cantonment	0.137931	0.000000	0.034483	0.000000	0.001271

The final map with all sub districts represented by different clusters in different color is shown below:



All the data frames grouped by Cluster wise are shown below:

Cluster 0

	sub division	Café	Metro Station	University	Market	Total population
0	Chanakya Puri	0.070000	0.0	0.010000	0.020000	0.000273
1	Civil Lines	0.078431	0.0	0.000000	0.000000	0.003059
2	Connaught Place	0.080000	0.0	0.000000	0.010000	0.000125
3	Darya Ganj	0.090000	0.0	0.000000	0.010000	0.001204
9	Karol Bagh	0.081633	0.0	0.000000	0.010204	0.000607
10	Kotwali	0.076923	0.0	0.038462	0.000000	0.000307
15	Parliament Street	0.100000	0.0	0.000000	0.010000	0.000233
17	Preet Vihar	0.093750	0.0	0.000000	0.015625	0.004736
20	Sadar Bazar	0.100000	0.0	0.000000	0.000000	0.000578
22	Seelam Pur	0.066667	0.0	0.000000	0.000000	0.006126
24	Shahdara	0.083333	0.0	0.000000	0.000000	0.001435
25	Vasant Vihar	0.070000	0.0	0.010000	0.010000	0.002851

Cluster 1

	sub division	Café	Metro Station	University	Market	Total population
6	Gandhi Nagar	0.046154	0.000000	0.0	0.030769	0.001756
7	Hauz Khas	0.060000	0.000000	0.0	0.030000	0.005470
11	Model Town	0.053333	0.013333	0.0	0.000000	0.002647
14	Pahar Ganj	0.060000	0.000000	0.0	0.000000	0.000776
16	Patel Nagar	0.051948	0.000000	0.0	0.012987	0.005608
18	Punjabi Bagh	0.032967	0.000000	0.0	0.010989	0.003552
19	Rajouri Garden	0.045455	0.011364	0.0	0.022727	0.002140
23	Seema Puri	0.057143	0.000000	0.0	0.000000	0.002399
26	Vivek Vihar	0.039216	0.000000	0.0	0.000000	0.001101

Cluster 2

	sub division	Café	Metro Station	University	Market	Total population
5	Delhi Cantonment	0.137931	0.0	0.034483	0.0	0.001271

Cluster 3

	sub division	Café	Metro Station	University	Market	Total population
12	Najafgarh	0.000000	0.0	0.0	0.0	0.006065
13	Narela	0.000000	0.0	0.0	0.0	0.003598
21	Saraswati Vihar	0.022222	0.0	0.0	0.0	0.010000

Cluster 4

	sub division	Café	Metro Station	University	Market	Total population
4	Defence Colony	0.08	0.0	0.0	0.06	0.002834
8	Kalkaji	0.07	0.0	0.0	0.06	0.003834

5. Observations

We observed that Cluster 2 has least number of existing café and has some of the most densely populated areas making Sub Divisions in Cluster 2 Highly Recommended to Open a New Cafe.

Cluster 0 has most number of existing café because of High amount of Nearby Market. Opening a Cafe in sub divisions of this cluster will give fierce competition.

Cluster 4 has highest amount of nearby market and therefore has high amount of existing Cafe. Opening a Cafe here will also give tough competition.

Cluster 1 is an exception with no nearby café, good amount of population around and many nearby metro stations.

Cluster 3 is an exception too. Cluster 3 has Highest Number of Cafe with surprisingly No nearby market. This happened because Cluster 3 has Highest Number of nearby universities making it a Honey Comb for Students to hangout. We observed a similar trend in kotwali sub district of cluster 0.

6. Result

The analysis shows the amount of cafe, nearby market, nearby metro stations, nearby universities and population of each sub division and then merge the similar sub divisions into clusters whose locations are represented on the map. we also made it easy to choose a sub district to open a Cafe with maximum location advantage and showed the sub districts that you should avoid for opening a cafe.

7) Conclusion

From the above observations we conclude that Cluster 2 and cluster 1 are best suited for opening a new cafe. Opening a cafe in a suitable sub district of these clusters will give less competition and has capability to attract good amount of customers too.

Cluster 0 has the highest competition but if you are willing to ignore competition then Cluster 4 is better than Cluster 0 for opening a cafe. Not only Cluster 4 has equal amount of average competition as cluster 0 but Cluster 4 has way more amount of nearby market than cluster 0 which will attract lot more of customers.