Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Bakery in Delhi, India

By: Anurag Kandalkar

May 2020

Introduction

People of all ages are affectionate of different bakery products, because of their taste, color and easy to digest nature. They eat and serve different bakery products in their parties and festivals. Celebrating any moment of happiness is incomplete with bakery products.Bakery products are becoming prominent day by day. They are very popular because of its taste and simple to digest. Bakery items are usually loved by all. Nowadays individuals have virtually no time to invest much on making breakfast it is the bread and bun or biscuits which had occurred instead of other sorts of stuff.. As a result, there are many Bakery in the city of Delhi and many more are being built. Opening Bakery allows Bakers to earn consistent income. Particularly, the location of the Bakery is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Delhi, India to open a new Bakery. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Delhi, India, if a Baker is looking to open a new Bakery, where would you recommend that they open it?

Data

To solve the problem, we will need the following data:

- List of neighbourhoods in Delhi. This defines the scope of this project which is confined to the city of Delhi, the capital city of the country of India in Asia.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to Bakery. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi) contains a list of neighbourhoods in Delhi, with a total of 174 neighbourhoods. We will use Json file mentioning Borough and neighbourhood. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Bakery category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Delhi. Fortunately, the list is available in theJSON file. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Delhi.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Bakery” data, we will filter the “Bakery” as venue category for the neighbourhoods.

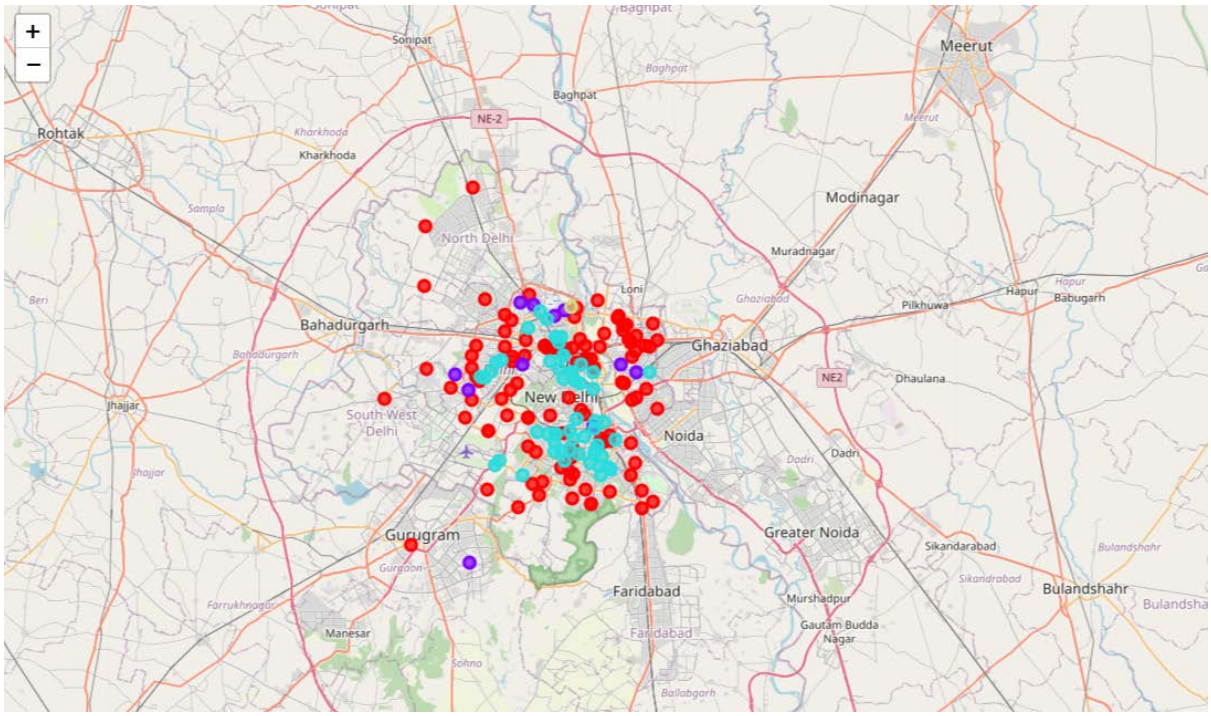
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 4 clusters based on their frequency of occurrence for “Bakery”. The results will allow us to identify which neighbourhoods have higher concentration of Bakery while which neighbourhoods have fewer number of Bakery. Based on the occurrence of Bakery in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Bakery.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 4 clusters based on the frequency of occurrence for “Bakery”:

- Cluster 0: Neighbourhoods with low number to no existence of Bakery
- Cluster 1: Neighbourhoods with moderate number of Bakery
- Cluster 2: Neighbourhoods with high concentration of Bakery
- Cluster 3: Neighbourhoods with very high concentration of Bakery
-

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, cluster 2 in mint blue colour and cluster 3 in light yellow colour



Discussion

As observations noted from the map in the Results section, most of the Bakery are concentrated in the central area of Delhi city, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to no Bakery in the neighbourhoods. This represents a great opportunity and high potential areas to open new Bakery as there is very little to no competition from existing malls. Meanwhile, Bakery in cluster 3 are likely suffering from intense competition due to oversupply and high concentration of Bakery. From another perspective, the results also show that the oversupply of Bakery mostly happened in the central area of the city, with the suburb area still have very few Bakery. Therefore, this project recommends Baker to capitalize on these findings to open new Bakery in neighbourhoods in cluster 0 with little to no competition. Baker with unique selling propositions to stand out from the competition can also open new Bakery in neighbourhoods in cluster 1 with moderate competition. Lastly, Baker are advised to avoid neighbourhoods in cluster 2 andcluster 3 which already have high concentration of Bakery and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of Bakery, there are other factors such as population and income of residents that could influence the location decision of a new Bakery. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new Bakery. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 4 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e.Bakers regarding the best locations to open a new Bakery. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 0 are the most preferred locations to open a new Bakery. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Bakery.

References

Category:Neighbourhoods of Delhi. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi

Foursquare Developers Documentation. *Foursquare*. Retrieved from https://developer.foursquare.com/docs

Neighbourhoods under Cluster 0

```
[30]: Delhi_merged.loc[Delhi_merged['Cluster Labels'] == 0]
```

[30]:

	Neighborhood	Bakery	Cluster Labels	Borough	Latitude	Longitude
83	Maujpur	0.000000	0	North East Delhi	28.692078	77.279697
112	Pitam Pura	0.000000	0	North West Delhi	28.703268	77.132250
111	Patparganj	0.000000	0	East Delhi	28.611592	77.290564
110	Patel Nagar	0.000000	0	West Delhi	28.659809	77.156957
109	Paschim Vihar	0.000000	0	West Delhi	28.669578	77.095956
107	Palam	0.000000	0	South West Delhi	28.591893	77.082824
105	Okhla	0.000000	0	South Delhi	28.563662	77.289055
103	New Usmanpur	0.000000	0	North East Delhi	28.683855	77.256005
100	Nehru Vihar	0.000000	0	North Delhi	28.710745	77.221174
98	Nehru Nagar	0.010000	0	West Delhi	28.568511	77.251385
97	Neeti Bagh	0.011765	0	South Delhi	28.559251	77.216166
96	Naveen Shahdara	0.000000	0	North East Delhi	28.677339	77.286016
95	Narela	0.000000	0	North West Delhi	28.842610	77.091835
94	Naraina	0.000000	0	South West Delhi	28.622055	77.138644
93	Nand Nagri	0.000000	0	North East Delhi	28.694458	77.316144
92	Najafgarh	0.000000	0	South West Delhi	28.612304	76.982391
91	Munirka	0.000000	0	South West Delhi	28.554886	77.171084
89	Moti Nagar	0.000000	0	West Delhi	28.657859	77.142429
69	Kirti Nagar	0.000000	0	West Delhi	28.653281	77.141773
70	Kishangarh Village	0.000000	0	South West Delhi	28.519428	77.166535
71	Kohat Enclave	0.000000	0	North West Delhi	28.698041	77.140539
73	Krishna Nagar	0.000000	0	East Delhi	28.657846	77.290185
74	Lahori Gate	0.000000	0	North Delhi	28.655787	77.238720
76	Laxmi Nagar	0.000000	0	East Delhi	28.630553	77.277575
114	Pratap Nagar	0.000000	0	North Delhi	28.666718	77.198897
77	Laxmibai Nagar	0.000000	0	New Delhi	28.575419	77.205109
81	Malviya Nagar	0.010000	0	South Delhi	28.533920	77.212447
82	Mandoli	0.000000	0	North East Delhi	28.681336	77.295243
138	Shahpur Jat	0.010204	0	South Delhi	28.548330	77.214104
129	Saket	0.010000	0	South Delhi	28.524411	77.213725
130	Sangam Vihar	0.000000	0	South Delhi	28.497702	77.239174
131	Sant Nagar	0.000000	0	North Delhi	28.646418	77.091188
133	Sarai Rohilla	0.000000	0	North Delhi	28.667873	77.190267
134	Sarita Vihar	0.000000	0	South Delhi	28.528574	77.288331
136	Sarvodaya Enclave	0.010000	0	South Delhi	28.537478	77.202089
137	Shahdara	0.000000	0	North East Delhi	28.673333	77.289025
130	Sangam Vihar	0.000000	0	North Delhi	28.497702	77.239174
67	Kingsway Camp	0.000000	0	North West Delhi	28.613749	77.212133
163	Yamuna Vihar	0.000000	0	North East Delhi	28.700372	77.272773
65	Khanpur	0.000000	0	South Delhi	28.512798	77.232395
22	Daryaganj	0.000000	0	Central Delhi	28.646090	77.243048
25	Delhi Cantonment	0.000000	0	South West Delhi	28.593833	77.134979
49	Jaitpur	0.000000	0	South Delhi	28.500477	77.316192
48	Jahangirpuri	0.000000	0	North West Delhi	28.725972	77.162658
47	Inderpuri	0.000000	0	South West Delhi	28.628684	77.147098
27	Dhaula Kuan	0.000000	0	South West Delhi	28.591891	77.161703
27	Dhaula Kuan	0.000000	0	West Delhi	28.591891	77.161703
28	Dilshad Garden	0.000000	0	North East Delhi	28.675826	77.321516
42	Gulabi Bagh	0.000000	0	North Delhi	28.669649	77.194726
29	East Vinod Nagar	0.000000	0	East Delhi	28.622766	77.307503
66	Khirki Village	0.010000	0	South Delhi	28.529885	77.218077
31	Fateh Nagar	0.000000	0	West Delhi	28.631615	77.100878
32	Friends Colony	0.000000	0	South Delhi	28.566751	77.261918
33	Gandhi Nagar	0.000000	0	East Delhi	28.453175	77.015329
34	Gautampuri	0.000000	0	New Delhi	28.511570	77.302623
36	Ghitorni	0.000000	0	South West Delhi	28.493751	77.149187
20	Dabri, New Delhi	0.000000	0	South West Delhi	28.610770	77.091106
21	Dariba Kalan	0.000000	0	North Delhi	28.654602	77.233379
18	Civil Lines	0.000000	0	North Delhi	28.676851	77.225030
64	Khan Market	0.010000	0	New Delhi	28.600135	77.226491
62	Kashmiri Gate	0.000000	0	North Delhi	28.669977	77.232059
84	Mayur Vihar	0.000000	0	East Delhi	28.613107	77.295722
85	Meera Bagh	0.000000	0	West Delhi	28.658844	77.090848
86	Mehrauli	0.000000	0	South Delhi	28.521826	77.178323
88	Moti Bagh	0.000000	0	South West Delhi	28.667619	77.184227
79	Maharani Bagh	0.000000	0	South Delhi	28.572001	77.263372
116	Pul Bangash	0.000000	0	North Delhi	28.666407	77.207416
117	Punjabi Bagh	0.000000	0	West Delhi	28.668945	77.132461
118	Rajokri	0.000000	0	South West Delhi	28.513170	77.110766
142	Shastri Nagar	0.000000	0	East Delhi	28.670088	77.181859
143	Shastri Park	0.000000	0	North East Delhi	28.668434	77.250530
146	Sonia Vihar	0.000000	0	North East Delhi	28.719926	77.248182
148	Sriniwaspuri	0.010204	0	South Delhi	28.565220	77.254704
150	Tilak Nagar	0.000000	0	West Delhi	28.636548	77.096496
151	Timarpur	0.000000	0	North Delhi	28.701263	77.218680
142	Shastri Nagar	0.000000	0	North Delhi	28.670088	77.181859
152	Tis Hazari	0.000000	0	North Delhi	28.667163	77.216631
154	Uttam Nagar	0.000000	0	West Delhi	28.624847	77.065286
156	Vasant Vihar	0.000000	0	South West Delhi	28.560691	77.160791
157	Vasundhara Enclave	0.000000	0	East Delhi	28.601726	77.321122
158	Vikas Nagar	0.000000	0	West Delhi	28.645015	77.035028
160	Vishwas Nagar	0.000000	0	East Delhi	28.664436	77.294960
161	Vivek Vihar	0.000000	0	East Delhi	28.669164	77.312267
153	Tughlaqabad	0.000000	0	South Delhi	28.511192	77.262327
68	Kirby Place	0.000000	0	South West Delhi	28.611741	77.128527
139	Shakarpur	0.000000	0	East Delhi	28.629489	77.281061
137	Shahdara	0.000000	0	East Delhi	28.673333	77.289025
122	Rani Bagh	0.000000	0	North West Delhi	28.685982	77.132524
123	Rithala	0.000000	0	North West Delhi	28.720806	77.107181
125	Sadar Bazaar	0.000000	0	North Delhi	28.577151	77.111153
125	Sadar Bazaar	0.000000	0	Central Delhi	28.577151	77.111153
126	Sadatpur	0.011364	0	North East Delhi	28.651718	77.221939
128	Sainik Farm	0.000000	0	South Delhi	28.503746	77.216073
6	Babarpur	0.000000	0	North East Delhi	28.687431	77.279755
7	Badarpur	0.000000	0	South Delhi	28.493170	77.303024
60	Karala	0.000000	0	North West Delhi	28.735140	77.032511
11	Bawana	0.000000	0	North West Delhi	28.799660	77.032885
52	Jasola	0.000000	0	South Delhi	28.542233	77.294386
13	Chanakyapuri	0.000000	0	New Delhi	28.594677	77.188521
12	Brij Puri	0.000000	0	East Delhi	28.702510	77.273777
14	Chandni Chowk	0.000000	0	Central Delhi	28.655983	77.232194
56	Kabir Nagar	0.000000	0	North East Delhi	28.692670	77.283544
55	Jor Bagh	0.000000	0	South Delhi	28.676608	77.158692
16	Chhattarpur	0.000000	0	South Delhi	28.507007	77.175417
54	Jhilmil Colony	0.000000	0	East Delhi	28.669814	77.307283
14	Chandni Chowk	0.000000	0	North Delhi	28.655983	77.232194
38	Golf Links	0.010000	0	New Delhi	28.595970	77.231163

Neighbourhoods in cluster 1

Cluster 1

```
[31]: Delhi_merged.loc[Delhi_merged['Cluster Labels'] == 1]
```

[31]:

	Neighborhood	Bakery	Cluster Labels	Borough	Latitude	Longitude
26	Dhaka	0.052632	1	North West Delhi	28.708698	77.205749
141	Shalimar Bagh	0.052632	1	North West Delhi	28.717453	77.150867
115	Preet Vihar	0.058824	1	East Delhi	28.641441	77.295259
159	Vikaspuri	0.041667	1	West Delhi	28.638419	77.070836
35	Geeta Colony	0.055556	1	East Delhi	28.650101	77.275921
0	Adarsh Nagar	0.062500	1	North West Delhi	28.714401	77.167288
162	Wazirabad	0.037736	1	North Delhi	28.433762	77.087757
87	Model Town	0.038462	1	North West Delhi	28.702714	77.193991
51	Jangpura	0.044944	1	South Delhi	28.582457	77.241500
50	Janakpuri	0.041667	1	West Delhi	28.621927	77.087476
108	Pandav Nagar	0.050000	1	East Delhi	28.650024	77.153676

Neighbourhood in cluster 3

Cluster3

```
[33]: Delhi_merged.loc[Delhi_merged['Cluster Labels'] == 3]
```

[33]:

	Neighborhood	Bakery	Cluster Labels	Borough	Latitude	Longitude
90	Mukherjee Nagar	0.1	3	North East Delhi	28.712557	77.214404

Neighbourhood in cluster 3

[32]:

	Neighborhood	Bakery	Cluster Labels	Borough	Latitude	Longitude
15	Chawri Bazaar	0.020000	2	North Delhi	28.649265	77.226515
80	Mahipalpur	0.022222	2	South West Delhi	28.544485	77.125691
140	Shakti Nagar	0.027778	2	North Delhi	28.679790	77.194914
140	Shakti Nagar	0.027778	2	North Delhi	28.679790	77.194914
57	Kailash Colony	0.030000	2	South Delhi	28.553052	77.242969
78	Lodi Colony	0.014085	2	South Delhi	28.590702	77.220921
58	Kalkaji	0.015152	2	South Delhi	28.537070	77.261805
59	Kamla Nagar	0.027027	2	North Delhi	28.680344	77.202129
144	Shivaji Place	0.016949	2	West Delhi	28.651657	77.121703
145	Siri Fort	0.030000	2	South Delhi	28.552146	77.224698
75	Lajpat Nagar	0.031250	2	South Delhi	28.579262	77.244033
37	Gole Market	0.020000	2	New Delhi	28.633719	77.205627
149	Tihar Village	0.022727	2	West Delhi	28.634636	77.107112
9	Bara Hindu Rao	0.022222	2	Central Delhi	28.659518	77.205010
9	Bara Hindu Rao	0.022222	2	North Delhi	28.659518	77.205010
8	Bali Nagar	0.020833	2	West Delhi	28.654138	77.128178
61	Karol Bagh	0.022727	2	Central Delhi	28.652998	77.189023
72	Kotwali	0.026316	2	North Delhi	28.641022	77.241605
155	Vasant Kunj	0.017241	2	South West Delhi	28.529249	77.154134
5	Azadpur	0.034483	2	North West Delhi	28.707657	77.175547
4	Ashok Vihar	0.033333	2	North West Delhi	28.699453	77.184826
3	Ashok Nagar	0.021739	2	West Delhi	28.636021	77.101822
63	Keshav Puram	0.030303	2	North West Delhi	28.688926	77.161683
2	Anand Vihar	0.032258	2	East Delhi	28.641115	77.312502
1	Alaknanda	0.018519	2	South Delhi	28.529336	77.251632
10	Barakhamba Road	0.030000	2	New Delhi	28.629589	77.225138
147	South Extension	0.027027	2	South Delhi	28.568715	77.216896
53	Jhandewalan	0.028169	2	Central Delhi	28.644319	77.199917
135	Sarojini Nagar	0.017241	2	South West Delhi	28.574157	77.195370
106	Paharganj	0.020000	2	Central Delhi	28.641499	77.214061
106	Paharganj	0.020000	2	North Delhi	28.641499	77.214061
39	Govindpuri	0.030303	2	South Delhi	28.535156	77.263794
113	Pragati Maidan	0.026316	2	New Delhi	28.623475	77.242528
104	Nizamuddin West	0.020833	2	South Delhi	28.588365	77.244955
40	Greater Kailash	0.030000	2	South Delhi	28.541878	77.238455
102	New Friends Colony	0.025641	2	South Delhi	28.567101	77.269764
101	Netaji Nagar	0.017544	2	South Delhi	28.573534	77.186359
41	Green Park	0.020000	2	South Delhi	28.555537	77.202497
119	Rajouri Garden	0.018868	2	West Delhi	28.642152	77.116060
120	Rama Krishna Puram	0.018868	2	South West Delhi	28.575561	77.172303
121	Rangpuri	0.019608	2	South West Delhi	28.540007	77.119775
99	Nehru Place	0.030612	2	South Delhi	28.549257	77.252952
43	Gulmohar Park	0.020000	2	South Delhi	28.557101	77.213005
124	Roshanara Bagh	0.034483	2	North Delhi	28.673718	77.199320
44	Hauz Khas	0.020000	2	South Delhi	28.544256	77.206707
45	Hauz Khas Village	0.030000	2	South Delhi	28.553855	77.194713
46	INA Colony	0.028986	2	New Delhi	28.577281	77.212649
127	Safdarjung Enclave	0.030000	2	South Delhi	28.565642	77.193438
24	Defence Colony	0.020408	2	South Delhi	28.571222	77.231776
23	Dayanand Colony	0.020000	2	South Delhi	28.562200	77.247613
132	Sarai Kale Khan	0.023810	2	North Delhi	28.586626	77.256228
19	Connaught Place	0.030000	2	New Delhi	28.631383	77.219792
135	Sarojini Nagar	0.017241	2	South Delhi	28.574157	77.195370
17	Chittaranjan Park	0.021978	2	South Delhi	28.538752	77.249249
30	East of Kailash	0.030000	2	South Delhi	28.557032	77.244614