IBM Applied Data Science Capstone

Recommending a Hotel at Chennai, India

Jesly Serin Jose

1. Introduction

1.1 Background

For this Capstone project, I am creating a hypothetical scenario regarding opening a hotel in the best location in the city of Chennai area, India. The main booming sector across the globe is Tourism. Primarily focussing on the places where tourists are attracted. So it is mandatory to have a good hotel for the tourist's comfort. Particularly, the location of the hotel is one of the most important decisions that will determine whether the hotel will be a success or a failure. Thereby they can enjoy the nearby popular places as much as needed. The solution is to recommend the best neighbourhood for opening the hotel in Chennai area.

1.2 Business Problem

The objective of this capstone is to analyze and observe the most suitable location for the entrepreneur to open a new Hotel in Chennai, India, where they can make use of best opportunity. In Chennai, if an entrepreneur wants to open a hotel, which place should be considering?

1.3 Target Audience

This project is particularly useful for the entrepreneur who wants to find the location to open Hotel in Chennai, India. This project also gives benefit for the customers and the property developers.

2. Data acquisition and cleaning

To tackle the problem, we need to have the dataset that contains:

- Complete list of neighbourhoods in Chennai, India
- Latitude and longitudes (i.e. Geographical coordinates) of the listed neighbourhoods. Source of data from FourSquare site
- Venue data related to hotels, using Machine Learning methods called Clustering to solve the problem

Data Source

- The Wikipedia page https://en.wikipedia.org/wiki/Category:Suburbs_of_Chennai is the major source of data that is being used to obtain all the districts of Chennai, with a total of 67 neighbourhoods. With the help of beautifulsoup package, Web scrapping techniques are used to extract data from Wikipedia page and convert it into a pandas dataframe.
- By using the geocoder package, we will able to get the geological coordinates of the neighbourhoods i.e. the latitude and longitude coordinates of all neighbourhoods present in the dataframe.
- The foursquare location platform will be used as the sole data source for venue data of neighbouring places can be obtained through the API.

3.Methodology

The best location to open a new hotel in Chennai, India is based on the neighbourhood's cluster of potential locations. On the basis of the overall number of venues for different categories of locations, all of the venues were grouped into the same category. Machine learning techniques called Clustering can perform this analysis.

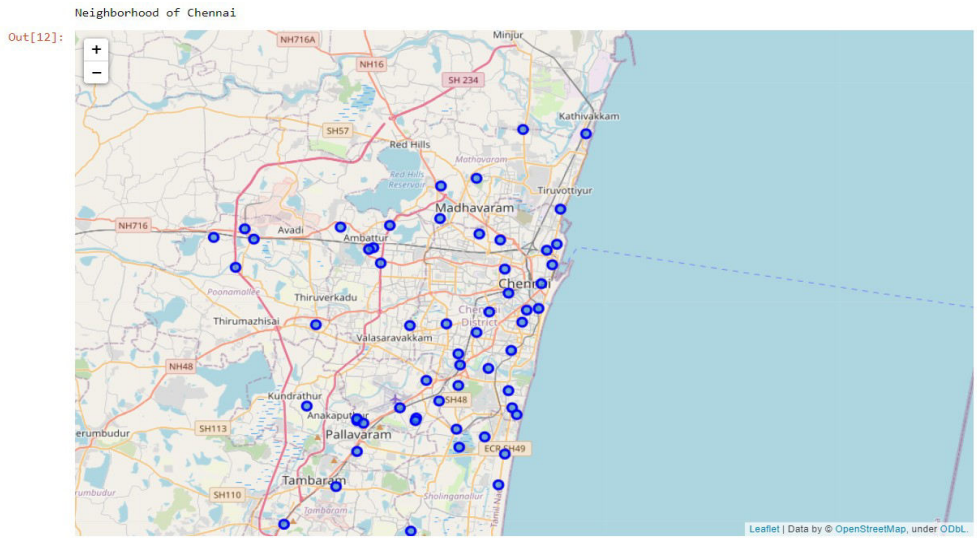
The first major step should be to collect the relevant data. On the dataset, available on the page (https://commons.wikimedia.org/wiki/Category:Suburbs_of_Chennai), web scraping is performed by using Python requests and beautiful-soup packages to extract the neighbourhood data list. The data from the Wikipedia were in a JSON format, which requires JSON parsing to extract the different data parameters.

The second phase deals with Exploratory Data Analysis. Statistical measures and data visualization techniques have been used to understand the internal structures of data and the relationships between the various data parameters. Then, to get the latitude and longitude of all the districts of the State, Geopy API is used, which contain Gecoder package. After collecting the data, it will be populated in a pandas DataFrame and then use the Folium package to visualize the neighbourhoods in a map. And the missing values which were removed in the dataset.

The final dataset of initial five of 63 neighbourhoods has shown the following:

	Neighborhood	Latitude	Longitude
0	▶ Adambakkam (13 F)	12.991920	80.206030
1	▶ Adyar (5 C, 17 F)	12.978150	80.188830
2	▶ Alwarthirunagar (9 F)	13.050550	80.183970
3	▶ Ambattur (1 C, 10 F)	13.129080	80.168890
4	▶ Anna Nagar (2 C, 6 F)	12.976730	80.144000

By using gecoder and follium libraries, the map is plotted and visualized with the coordinates of all locations in Chennai neighborhood as shown the following:



Next, the Foursquare API is used to generate the complete data to get the locations within a 500-meter radius. With the help of Foursquare Account, by making API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the JSON format of the venue data and create the table.

```
(1974, 7)
```

ut[16]:

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	▶ Adambakkam (13 F)	12.99192	80.20603	Luxe Cinemas	12.991041	80.216962	Multiplex
1	▶ Adambakkam (13 F)	12.99192	80.20603	Phoenix Market City	12.991710	80.217297	Shopping Mall
2	▶ Adambakkam (13 F)	12.99192	80.20603	Mainland China	12.991028	80.217084	Chinese Restaurant
3	▶ Adambakkam (13 F)	12.99192	80.20603	IMAX®	12.990639	80.216310	Multiplex
4	▶ Adambakkam (13 F)	12.99192	80.20603	Rajdhani	12.991081	80.217003	Rajasthani Restaurant

Finally, using k-means clustering, the data performs clustering, which is the simple and most popular unsupervised machine learning algorithms. K-means clustering identifies k number of centroids, and then assigns each data point to the nearest center, while keeping centroids as small as possible. Based on the frequency of occurrence for "Hotel", the neighbourhoods are grouped into "4" cluster category. The findings will able to identify which neighbourhoods are most appropriate for the opening of new hotel.

```
(63, 5)
```

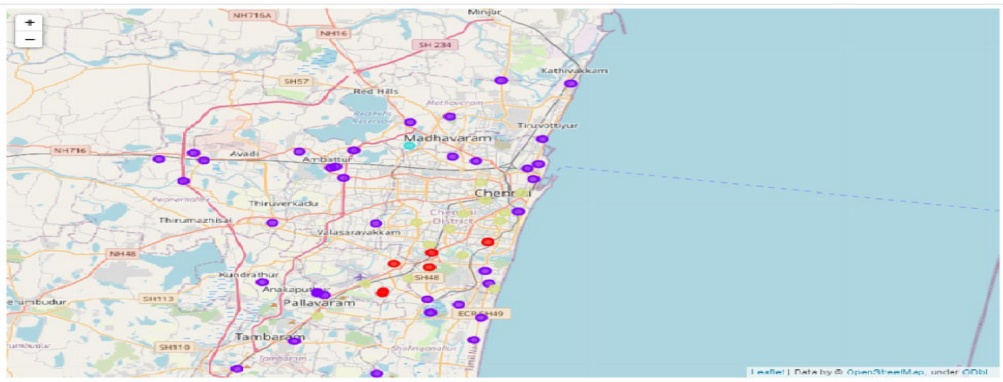
Out[29]:

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
1	▶ Adyar (5 C, 17 F)	0.090909	0	12.978150	80.188830
47	▶ St. Thomas Mount (2 C, 41 F)	0.142857	0	13.007990	80.195990
29	▶ Nanganallur (3 F)	0.133333	0	12.976400	80.187600
55	▶ Urapakkam (3 F)	0.083333	0	12.863420	80.069160
43	▶ Saidapet (20 F)	0.101695	0	13.020270	80.221310
16	▶ Guindy (4 C, 1 P, 17 F)	0.085714	0	13.004080	80.220120
28	▶ Mylapore (3 C, 16 F)	0.110000	0	13.031550	80.260220
62	▶ Washermanpet (1 C, 1 F)	0.000000	1	13.109500	80.287010
30	▶ Neelankarai (2 F)	0.000000	1	12.950140	80.255050
61	▶ Vyasarpadi (1 C)	0.000000	1	13.117780	80.251680

4. Result

The results show that the clusters can be categorized into four based on the frequency of occurrence for “Hotel”:

Out[30]:



The colors red, purple, green and yellow represents cluster 0, 1, 2 and 3 respectively.

5. Observations

After looking at the different locations that can be used for a new hotel, it can be observed that some are more suitable for hotel use. The table shows details of the availability of hotels nearby, as well as the venue and place.

First Cluster:

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
1	► Adyar (5 C, 17 F)	0.090909	0	12.97815	80.18883
47	► St. Thomas Mount (2 C, 41 F)	0.142857	0	13.00799	80.19599
29	► Nanganallur (3 F)	0.133333	0	12.97640	80.18760
55	► Urapakkam (3 F)	0.083333	0	12.86342	80.06916
43	► Saidapet (20 F)	0.101695	0	13.02027	80.22131
16	► Guindy (4 C, 1 P, 17 F)	0.085714	0	13.00408	80.22012
28	► Mylapore (3 C, 16 F)	0.110000	0	13.03155	80.26022

There are fewer locations in this cluster, but hotel availability is good, if not the best.

Second Cluster:

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
62	► Washermanpet (1 C, 1 F)	0.000000	1	13.109500	80.287010
30	► Neelankarai (2 F)	0.000000	1	12.950140	80.255050
61	► Vyasarpadi (1 C)	0.000000	1	13.117780	80.251680
32	► Padi, Chennai (6 F)	0.000000	1	13.099710	80.161680
33	► Pallavaram (2 C, 19 F)	0.000000	1	12.974440	80.148520
34	► Pallikaranai (1 C, 1 F)	0.000000	1	12.955670	80.220800
35	► Pattaravakkam (1 C)	0.000000	1	13.111644	80.156137
37	► Perungudi (8 F)	0.000000	1	12.963560	80.240010
38	► Poonamallee (2 C, 4 F)	0.000000	1	13.051370	80.112670
39	► Puzhal (2 C)	0.000000	1	13.159460	80.207180
40	► Ramapuram (1 C, 8 F)	0.000000	1	13.109980	80.152860
52	► Thiruvanmiyur (1 C, 12 F)	0.022472	1	12.986190	80.260510
42	► Royapuram (1 C, 6 F)	0.000000	1	13.113940	80.294200
51	► Thiruninravur (3 C, 1 P, 14 F)	0.000000	1	13.119850	80.036010
45	► Siruseri (11 F)	0.000000	1	12.837580	80.201850
46	► Sithalapakkam (7 F)	0.000000	1	12.890090	80.184590
36	► Perambur (3 C, 15 F)	0.000000	1	13.122470	80.235690
26	► Mundagakanniamman Koil railway station (3 F)	0.000000	1	13.118530	80.066030
24	► Manali New Town (3 F)	0.000000	1	13.203710	80.268540
50	► Thirumullaivoyal (3 C, 1 F)	0.000000	1	13.127500	80.131640
2	► Alwarthirunagar (9 F)	0.000000	1	13.050550	80.183970
3	► Ambattur (1 C, 10 F)	0.000000	1	13.129080	80.168890
4	► Anna Nagar (2 C, 6 F)	0.000000	1	12.976730	80.144000
5	► Anna Salai (4 C, 32 F)	0.000000	1	13.125990	80.059450

This cluster has the Very Low/None Hotels accessibility as seen in the Hotel column of the data.

Third Cluster:

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
60	► Villivakkam (1 C)	0.250000	2	13.13433	80.20618
27	► Muttukadu, Chennai (3 C, 1 F)	0.214286	2	12.83165	80.24207

There are only two places in this cluster, but the hotel's accessibility score is very high.

Fourth Cluster:

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
56	▶ Vadapalani (2 C, 10 F)	0.059701	3	13.05226	80.21120
54	▶ Triplicane (3 C, 26 F)	0.050000	3	13.06289	80.27146
57	▶ Valmiki Nagar (4 F)	0.030303	3	12.98139	80.26377
0	▶ Adambakkam (13 F)	0.042254	3	12.99192	80.20603
44	▶ Semmencherry (1 C, 1 F)	0.047619	3	12.86557	80.22051
41	▶ Royapettah (2 C, 1 F)	0.070000	3	13.05352	80.26826
25	▶ Meenambakkam (1 C, 1 F)	0.058824	3	12.98646	80.17600
21	▶ Kotturpuram (2 C, 4 F)	0.050000	3	13.01696	80.24276
20	▶ Kosapet (8 F)	0.062500	3	13.09453	80.25482
18	▶ Kodambakkam (8 F)	0.028571	3	13.02883	80.21999
13	▶ Egmore (5 C, 5 F)	0.050000	3	13.07642	80.25756
12	▶ Covelong (23 F)	0.076923	3	12.78984	80.24914
11	▶ Chromepet (3 F)	0.055556	3	12.95237	80.14413
10	▶ Chetput (6 F)	0.040000	3	13.08362	80.28252
48	▶ T. Nagar (1 C, 1 P, 57 F)	0.068966	3	13.04536	80.23390
31	▶ Nungambakkam (3 C, 7 F)	0.060000	3	13.06160	80.24315

The fourth cluster includes locations with lower to moderate hotel availability.

6. Limitations and Suggestions for Future Research

In this study, the effects of population and income of residents are studied on the location decision of a new hotel. However, the relevant data are not readily available at a neighbourhood level required by this study. Future research could devise a method to estimate these data to be used in the clustering algorithm to determine when to locate a new Hotel. In addition, the software was limited in the type of API calls it could make, and what kind of data it could obtain. Future research would require paid accounts to get better results than those available through free accounts.

7. Conclusion

According to the results, it seems that the places that are part of cluster 1 (label – 1) are most suitable for opening a new hotel. The second cluster has the least number of existing hotels, but it has excellent connections with other popular public places.

This analysis suggests the opening of a new hotel in the **Second Cluster (Cluster Label – 1)**

8. References

- Category:Suburbs in Chennai. Wikipedia. https://commons.wikimedia.org/wiki/Category:Suburbs_of_Chennai
- Foursquare Developers Documentation. Foursquare. https://developer.foursquare.com/docs