IBM Applied Data Science Capstone Project

Recommending Location to Open a Hotel at Chennai, India

by

Gregory Okechukwu Offodum

February, 2021

Contents

1 Introduction

Have you ever wondered about how to take the best decision in siting a business venture? - on a daily basis, entrepreneurs have to analyze before making this decision. This project investigates and studies the tourism industry; the considerations to take when going into the hotel business so as to maximize profit and satisfy customers. The case study deals with Chennai area, India.

Problem statement

As an entrepreneur in this area, if you could choose where to open a new hotel, where would it be?

This project aims to analyze, infer and recommend the best suited location to open a new hotel in Chennai, India.

2. Executive Summary

Chennai neighborhoods in India have been analyzed and clustered in 4 categories for suitability for a new hotel location. The findings have shown that the **Third Cluster (2)** is the best suited location. Second Cluster (1) proved to be the worst choice while First and Fourth clusters were moderately suitable.

3. Data description

The project utilizes dataset which we scrapped from Wikipedia, with a total of 67 neighborhoods, comprising of latitude and longitude, zip codes.

Link to the Data to be used: https://en.wikipedia.org/wiki/Category:Suburbs_of_Chennai

API

Foursquare API was used as the data source. Possessing a database of millions of location, the API is undoubtedly a top-notch source for this kind of project. Using credentials of Foursquare API features of near-by places of the neighborhoods would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

Neighbourhood: Name of the Neighbourhood

Neighbourhood Latitude: Latitude of the Neighbourhood

Neighbourhood Longitude: Longitude of the Neighbourhood

Venue: Name of the Venue

Venue Latitude: Latitude of Venue

Venue Longitude: Longitude of Venue

Venue Category: Category of Venue

4. Methodology

Summarily, exploratory data analysis was done to understand dataset. Preprocessing, cleaning of dataset and best feature selection followed suite. Machine learning and model development ensued, then K-Means Clustering technique was used. After Cluster Analysis, the best suited location was clear for selection.

Data collection was done from https://en.wikipedia.org/wiki/Category:Suburbs_of_Chennai, Beautiful Soup packages were used for web scrapping and data was extracted accordingly. Exploratory techniques were done, several statistical measures and visualization methods have been done to extensively explore the dataset. Geopy API facilitated accessing latitude and longitude values, these data were then transposed to pandas dataframes for analyses.

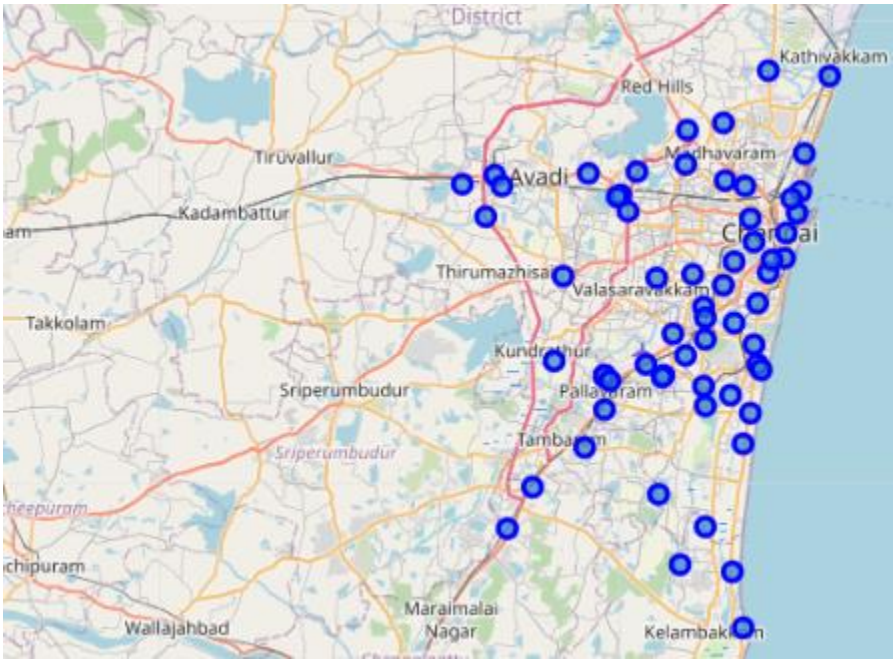
Data Scrapping

```
[6]: # send the GET request
data = requests.get("https://commons.wikimedia.org/wiki/Category:Suburbs_of_Chennai")
if data.status_code == 200:
    print('Page download successful')
else:
    print('Page download error. Error code: {}'.format(page.status_code))
data=data.text
```

```
Page download successful
```

	Neighborhood
0	► Adambakkam (13 F)
1	► Adyar (5 C, 17 F)
2	► Alwarthirunagar (9 F)
3	► Ambattur (1 C, 10 F)
4	► Anna Nagar (2 C, 6 F)

Follium library was used for map visualization by virtue of Chennai location coordinates:



Foursquare API was utilized to geberate locations within a 500-meter radius and the data was populated in a pandas dataframe:

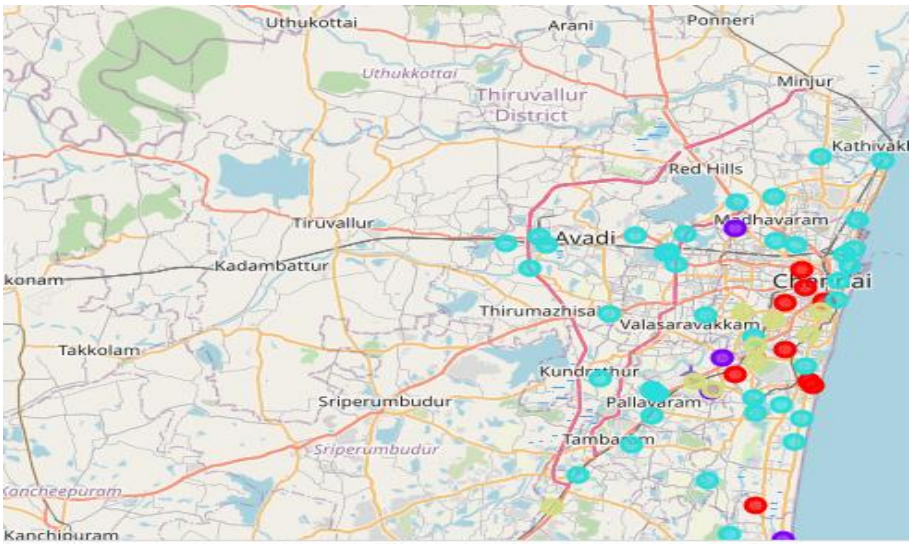
	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	► Adambakkam (13 F)	12.99192	80.20603	Luxe Cinemas	12.991041	80.216962	Multiplex
1	► Adambakkam (13 F)	12.99192	80.20603	Phoenix Market City	12.991710	80.217297	Shopping Mall
2	► Adambakkam (13 F)	12.99192	80.20603	Rajdhani	12.991081	80.217003	Rajasthani Restaurant
3	► Adambakkam (13 F)	12.99192	80.20603	IMAX®	12.990639	80.216310	Multiplex
4	► Adambakkam (13 F)	12.99192	80.20603	Zara	12.991411	80.216854	Clothing Store

K-Means was used to cluster neighborhoods into four based on hotel availability/frequency – identifying 4 centroids as specified, data points were assigned to their nearest center.

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
0	► Adambakkam (13 F)	0.051282	0	12.99192	80.20603
1	► Adyar (5 C, 17 F)	0.136364	3	12.97815	80.18883
2	► Alwarthirunagar (9 F)	0.000000	2	13.05055	80.18397
3	► Ambattur (1 C, 10 F)	0.000000	2	13.12908	80.16889
4	► Anna Nagar (2 C, 6 F)	0.000000	2	12.97673	80.14400

5. Result

The map was visualized after clustering giving an insight to the presence of hotels within the four clusters.



First Cluster, 0: Red

Second Cluster, 1: Purple

Third Cluster, 2: Light Blue

Fourth Cluster, 3: Yellow

6. Discussion

A more detailed look into the clusters was done to make objective inferences regarding their suitability for location of a new hotel.

First Cluster, 0

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
0	► Adambakkam (13 F)	0.051282	0	12.99192	80.20603
57	► Valmiki Nagar (4 F)	0.045455	0	12.98139	80.26377
54	► Triplicane (3 C, 26 F)	0.050000	0	13.06289	80.27146
52	► Thiruvanmiyur (1 C, 12 F)	0.034483	0	12.98619	80.26051
44	► Semmencherry (1 C, 1 F)	0.047619	0	12.86557	80.22051
21	► Kotturpuram (2 C, 4 F)	0.051546	0	13.01696	80.24276
20	► Kosapet (8 F)	0.063830	0	13.09453	80.25482
13	► Egmore (5 C, 5 F)	0.050000	0	13.07642	80.25756
31	► Nungambakkam (3 C, 7 F)	0.040000	0	13.06160	80.24315

Second Cluster, 1

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
60	► Villivakkam (1 C)	0.250000	1	13.13433	80.20618
29	► Nanganallur (3 F)	0.187500	1	12.97640	80.18760
27	► Muttukadu, Chennai (3 C, 1 F)	0.187500	1	12.83165	80.24207
47	► St. Thomas Mount (2 C, 41 F)	0.166667	1	13.00799	80.19599

Third Cluster, 2

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
9	► Chepauk (3 C, 6 F)	0.014706	2	13.064400	80.280650
36	► Perambur (3 C, 15 F)	0.000000	2	13.122470	80.235690
37	► Perungudi (8 F)	0.000000	2	12.963560	80.240010
38	► Poonamallee (2 C, 4 F)	0.000000	2	13.051370	80.112670
39	► Puzhal (2 C)	0.000000	2	13.159460	80.207180
40	► Ramapuram (1 C, 8 F)	0.000000	2	13.109980	80.152860
42	► Royapuram (1 C, 6 F)	0.000000	2	13.113940	80.294200
4	► Anna Nagar (2 C, 6 F)	0.000000	2	12.976730	80.144000
46	► Sithalapakkam (7 F)	0.000000	2	12.890090	80.184590
8	► Besant Nagar (3 C, 6 F)	0.011111	2	13.000170	80.257660
49	► Tambaram (2 C, 8 F)	0.000000	2	12.924890	80.128180
50	► Thirumullaivoyal (3 C, 1 F)	0.000000	2	13.127500	80.131640
51	► Thiruninravur (3 C, 1 P, 14 F)	0.000000	2	13.119850	80.036010
3	► Ambattur (1 C, 10 F)	0.000000	2	13.129080	80.168890
53	► Tiruvottiyur (1 C, 1 F)	0.000000	2	13.141780	80.296920
2	► Alwarthirunagar (9 F)	0.000000	2	13.050550	80.183970
58	► Vandalur (1 C, 8 F)	0.000000	2	12.895300	80.088800

Fourth Cluster, 3

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
12	► Covelong (23 F)	0.076923	3	12.78984	80.24914
1	► Adyar (5 C, 17 F)	0.136364	3	12.97815	80.18883
56	► Vadapalani (2 C, 10 F)	0.072464	3	13.05226	80.21120
28	► Mylapore (3 C, 16 F)	0.110000	3	13.03155	80.26022
25	► Meenambakkam (1 C, 1 F)	0.078947	3	12.98646	80.17600
16	► Guindy (4 C, 1 P, 17 F)	0.086957	3	13.00408	80.22012
41	► Royapettah (2 C, 1 F)	0.080000	3	13.05352	80.26826
43	► Saidapet (20 F)	0.081967	3	13.02027	80.22131
55	► Urapakkam (3 F)	0.071429	3	12.86342	80.06916
48	► T. Nagar (1 C, 1 P, 57 F)	0.074468	3	13.04536	80.23390

Inferences from the above cluster tables prove that the second and third clusters have the highest and lowest number of hotels respectively. A more descriptive table on suitability is shown below:

Cluster no	Cluster	Suitability	Description
0	First	Maybe	Moderate availability of hotels
1	Second	No	Highest number of hotels around
2	Third	Yes	Lowest dominance of hotels
3	Fourth	Maybe	Moderate availability of hotels

7. Recommendations for a better project

Further and extensive research would be possible with a larger dataset. Multiple model deployment and interpretation should be undertaken for better accuracy.

More so, consulting expertise from real estate professionals would facilitate better results.

8. Conclusion

From the results, it can be inferred that the **Third Cluster (2)** is the best suited location to open a new hotel by virtue of the relatively low number of hotels in these areas.

9. References

- Suburbs in Chennai, Wikipedia

  https://en.wikipedia.org/wiki/Category:Suburbs_of_Chennai

- Foursquare Developers Documentation.