COURSERA

IBM Applied Data Science Capstone

The Battle of Neighborhoods

By: Baddam Charitha

Introduction:

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies, etc.. Shopping malls are like a one-stop destination for all types of shoppers. For retailers, the central

location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more shopping malls to cater to the demand. As a result, there are many shopping malls in the city of Hyderabad and many more are being built. Opening shopping malls allows property developers to earn consistent rental income. Opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Proposal:

The objective of this capstone project is to analyze and select the best locations in the city of Hyderabad, India to open a new shopping mall. 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 Hyderabad, India, if a property developer is looking to open a new shopping mall, where would you recommend that they open it?

Data Section:

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

- List of neighborhoods in Hyderabad. This defines the scope of this project which is confined to the city of Hyderabad, India.

- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.

- Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighborhoods.

Foursquare API:

We will need data about different venues in different neighborhoods of that specific city. In order to gain that information we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 2000 meter.

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:

1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Venue
5. Venue Latitude
6. Venue Longitude
7. Venue Category (Shopping mall)

Methodology Section:

Firstly, we need to get the list of neighborhoods in the city of Hyderabad. Fortunately, the list is available in the page (https://commons.wikimedia.org/wiki/Category:Suburbs_of_Hyderabad,_India). I will do web scraping using Python requests and beautiful-soup packages to extract the list of neighborhoods data. However, this is just a list of names. I 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 Geo-coder 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 Data Frame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical co-ordinates data returned by Geo-coder are correctly plotted in the city of Hyderabad.

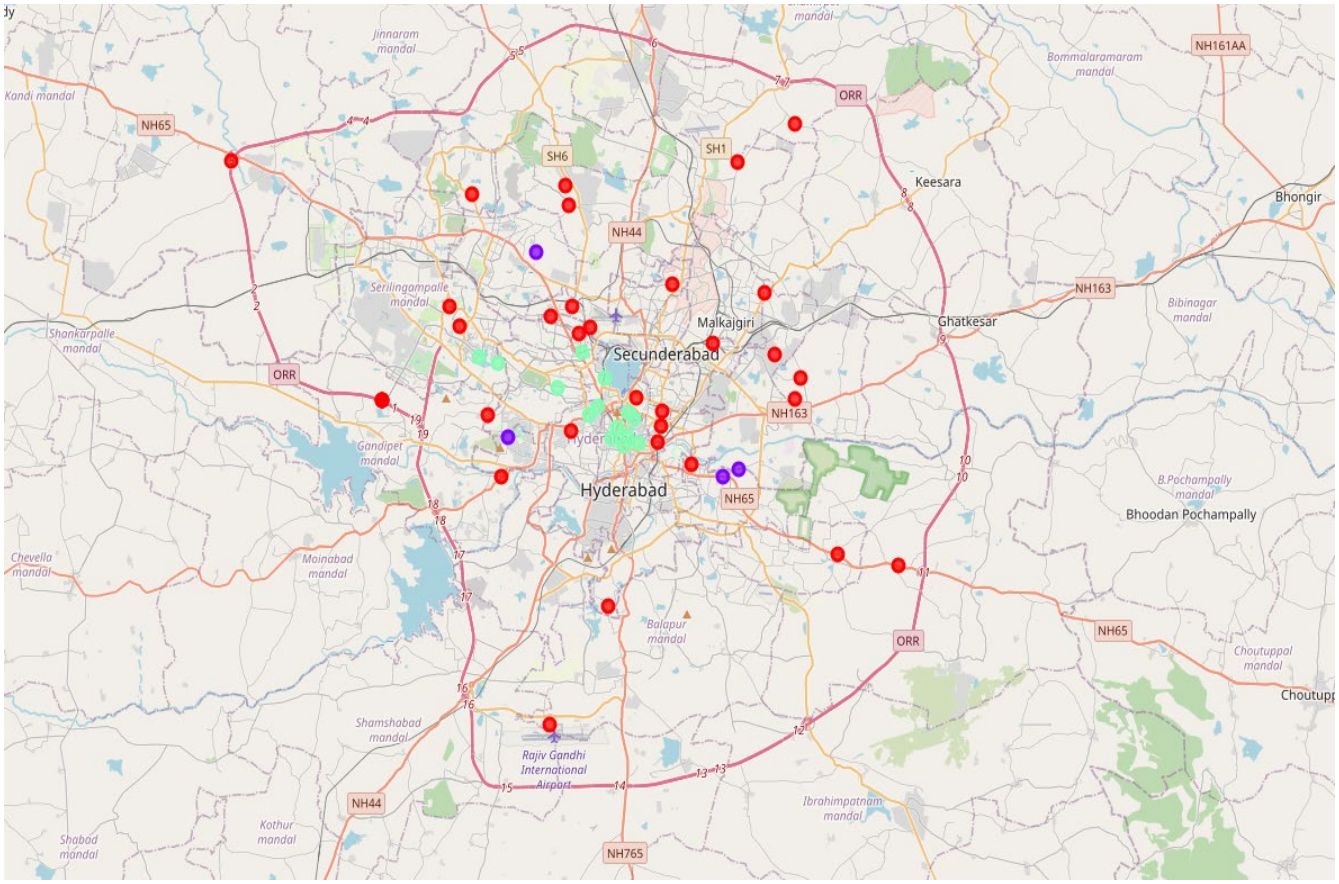
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods 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 neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood 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 analyzing the “Shopping Mall” data, we will filter the “Shopping Mall” as venue category for the neighborhoods.

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 neighborhoods into “3” clusters based on their frequency of occurrence for “Shopping Mall”. The results will allow us to identify which

neighborhoods have higher concentration of shopping malls while which neighborhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new shopping malls.

Results Section:

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Shopping Mall”:



- **Cluster 0:** Neighborhoods with low number to no existence of shopping malls.

:

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
51	▶ Trimulgherry (1 C, 3 F)	0.0	0	17.470723	78.504503
21	▶ Hydershakote (14 F)	0.0	0	17.368380	78.399990
44	▶ Sanathnagar (8 F)	0.0	0	17.458760	78.443100
23	▶ Kachiguda (1 C, 4 F)	0.0	0	17.386880	78.495530
43	▶ Pedda Amberpet (1 F)	0.0	0	17.321150	78.642370
50	▶ Tarnaka (1 C, 6 F)	0.0	0	17.408935	78.326740
27	▶ L. B. Nagar (16 F)	0.0	0	17.512650	78.441290
28	▶ Madhapur (1 C, 19 F)	0.0	0	17.459000	78.368100
29	▶ Malakpet (3 C, 2 F)	0.0	0	17.374930	78.515670
30	▶ Malkajgiri (3 C, 6 F)	0.0	0	17.439300	78.529200
31	▶ Manikonda (8 F)	0.0	0	17.401390	78.391630
33	▶ Mehdipatnam (1 C)	0.0	0	17.392630	78.442190
34	▶ Miyapur (5 F)	0.0	0	17.421020	78.582440
41	▶ Nizampet (2 C, 32 F)	0.0	0	17.518330	78.381860
36	▶ Moula-Ali (3 C, 5 F)	0.0	0	17.465770	78.560180
37	▶ Nacharam (1 C, 4 F)	0.0	0	17.433510	78.566730
45	▶ Sanjeeva Reddy Nagar (10 F)	0.0	0	17.444380	78.447240
18	▶ HITEC City (5 C, 29 F)	0.0	0	17.448230	78.374290
19	▶ Hayathnagar (1 C, 14 F)	0.0	0	17.327070	78.605330
8	▶ Bolarum (3 C, 1 F)	0.0	0	17.536219	78.235043
1	▶ Alwal (1 C, 1 F)	0.0	0	17.535430	78.544270
3	▶ Bandlaguda, Rangareddy (1 C, 2 F)	0.0	0	17.299820	78.464950
48	▶ Sitaphalmandi (1 C, 1 F)	0.0	0	17.408935	78.326740
47	▶ Shamshabad (1 C, 4 F)	0.0	0	17.236650	78.429370
6	▶ Begumpet (5 C, 9 F)	0.0	0	17.447290	78.453960

7	▶ Boduppal (2 F)	0.0	0	17.409540	78.578960
9	▶ Cavalry Barracks, Hyderabad (1 C)	0.0	0	17.408935	78.326740
40	▶ Narayanguda (1 C, 4 F)	0.0	0	17.395474	78.497594
11	▶ Dabirpura (1 C)	0.0	0	17.408935	78.326740
13	▶ Domalguda (3 C)	0.0	0	17.409950	78.482290
14	▶ Erragadda (3 F)	0.0	0	17.453330	78.430340
46	▶ Shamirpet (3 C, 5 F)	0.0	0	17.555611	78.578848
16	▶ Gajularamaram (2 F)	0.0	0	17.522760	78.438620
10	▶ Chikkadpally (7 F)	0.0	0	17.403010	78.497920

- **Cluster 1:** Neighborhoods with high concentration of shopping malls.

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
26	▶ Kukatpally (16 F)	0.100000	1	17.487350	78.420870
17	▶ Golconda (5 C, 4 F)	0.076923	1	17.389410	78.404060
12	▶ Dilsukhnagar (1 C, 2 F)	0.050000	1	17.368570	78.535150
38	▶ Nagole, Hyderabad (4 F)	0.066667	1	17.372426	78.544543

- **Cluster 2:** Neighborhoods with moderate no. of shopping malls.

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
42	► Old City (Hyderabad, India) (8 C, 26 F)	0.010870	2	17.39487	78.47076
49	► Somajiguda (5 F)	0.020000	2	17.42072	78.46300
0	► Abids (1 C, 13 F)	0.012658	2	17.38980	78.47658
35	► Moazzam Jahi Market (16 F)	0.019608	2	17.38448	78.47442
32	► Masab Tank (4 F)	0.010000	2	17.40093	78.45362
24	► Khairtabad (1 C, 2 F)	0.010000	2	17.40592	78.45856
22	► Jubilee Hills (3 C, 8 F)	0.010000	2	17.42865	78.39762
20	► Hyderguda (2 F)	0.011111	2	17.39923	78.48073
15	► Gachibowli (4 C, 17 F)	0.010000	2	17.43181	78.38636
5	► Basheerbagh (1 C, 7 F)	0.010417	2	17.40211	78.47770
4	► Banjara Hills (3 C, 25 F)	0.020000	2	17.41535	78.43435
2	► Ameerpet, Hyderabad (3 C, 21 F)	0.020000	2	17.43482	78.44949
39	► Nampally (2 C, 10 F)	0.017241	2	17.38897	78.46733
25	► Koti, Hyderabad (3 C, 7 F)	0.014706	2	17.38594	78.48338

Discussion Section:

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

Conclusion Section:

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 “3” clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 2 are the most preferred locations to open a new shopping mall. 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 shopping mall.

References:

Details of Suburbs in Hyderabad retrieved from (https://commons.wikimedia.org/wiki/Category:Suburbs_of_Hyderabad,_India).