COURSERA

IBM Applied Data Science Capstone

THE BATTLE OF NEIGHBOURHOODS

Introduction :

Restaraunts are a great place to spend a quality time with family and friends while having delicious food together.There are different restaraunts having various cuisines.But the cuisine selection and what to offer to the customers depends on the locality of where the investor want to place the restaurant.Business investors contribute in the growth of the economy and also the restaurant business is a great way for their growth."Who dosen't invest in good food,right?".Here we suggest business investors to establish a restaurant at the desired location by investigating the in-sights of restaraunts in the city of Hyderabad,India

The objective of this capstone project is to analyze and select the best locations in the city of Hyderabad, India to open a new restaraunt. 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 business investor is to open a restaurant, where would you recommend that they open it?

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 Restaraunt. 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 business investor is looking to open a new Restaurant, 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 Restaurants. 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 (Restaurant)

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.

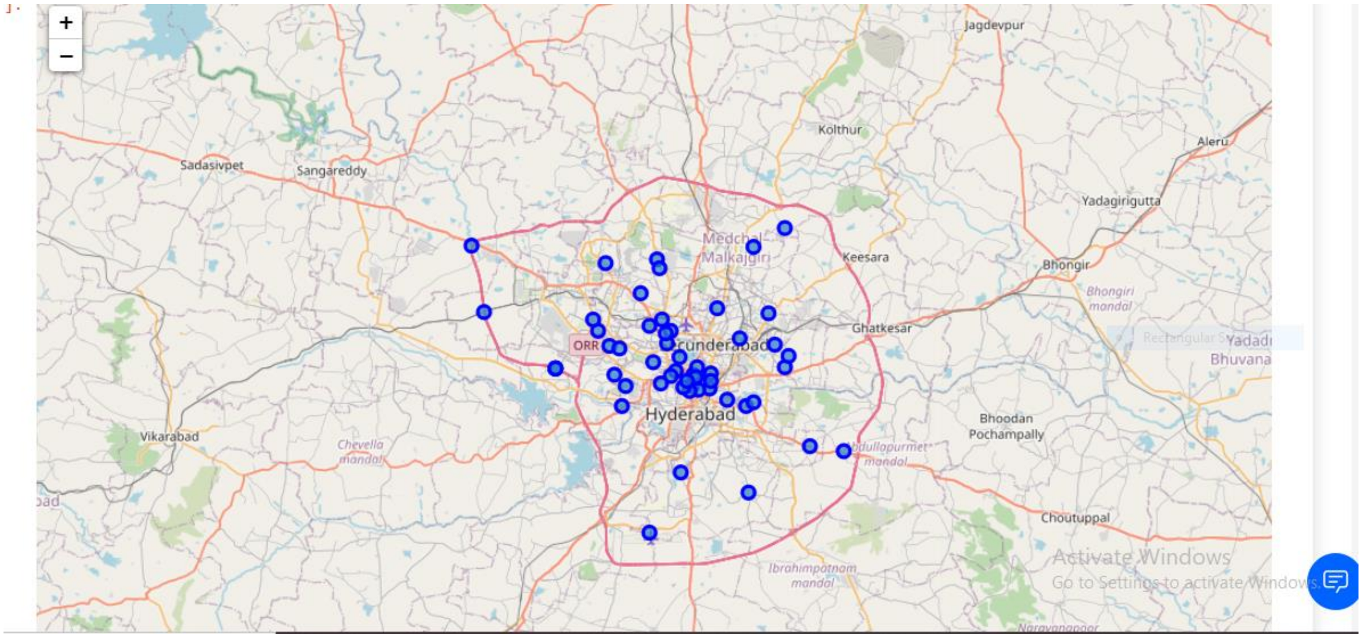
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods 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 "Restaurant" data, we will filter the "Restaurant" as venue category for the neighborhoods.

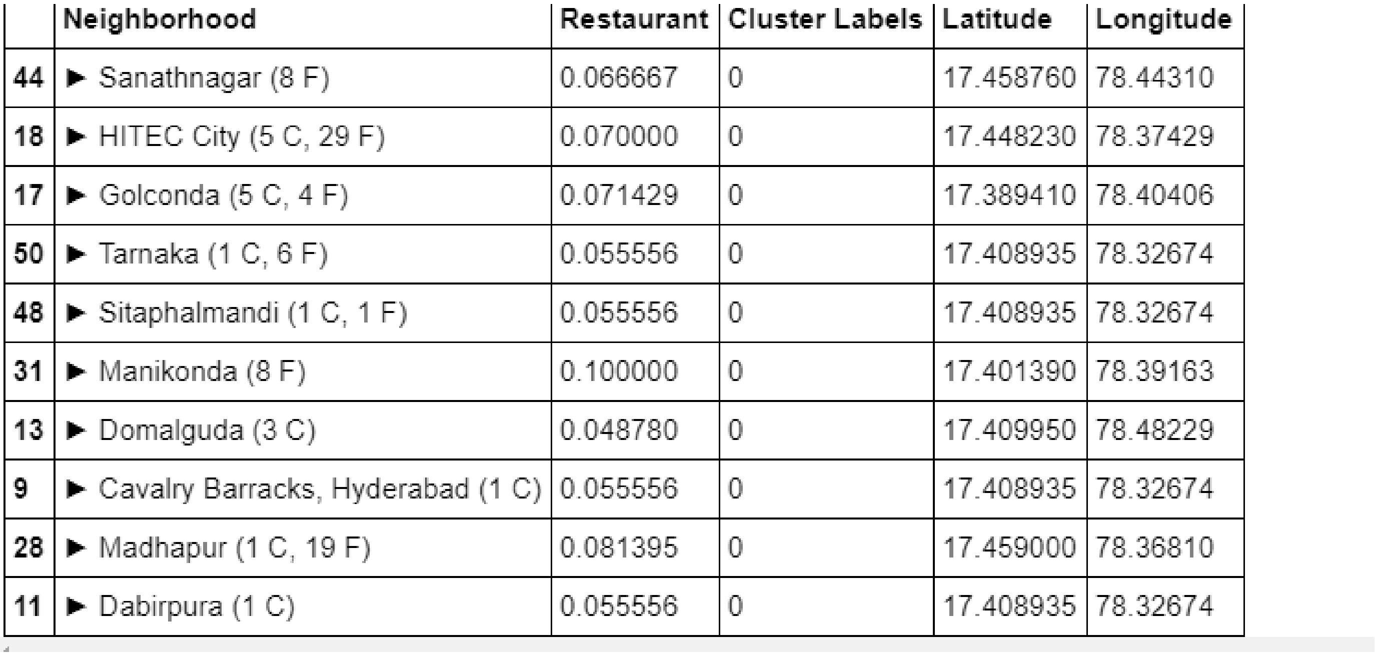
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine 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 "Restaurants". The results will allow us to identify which neighborhoods have higher concentration of Restaurants while which neighborhoods have fewer number of Restaurants. Based on the occurrence of Restaurants in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Restaurants.

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 "RESTAURANTS":

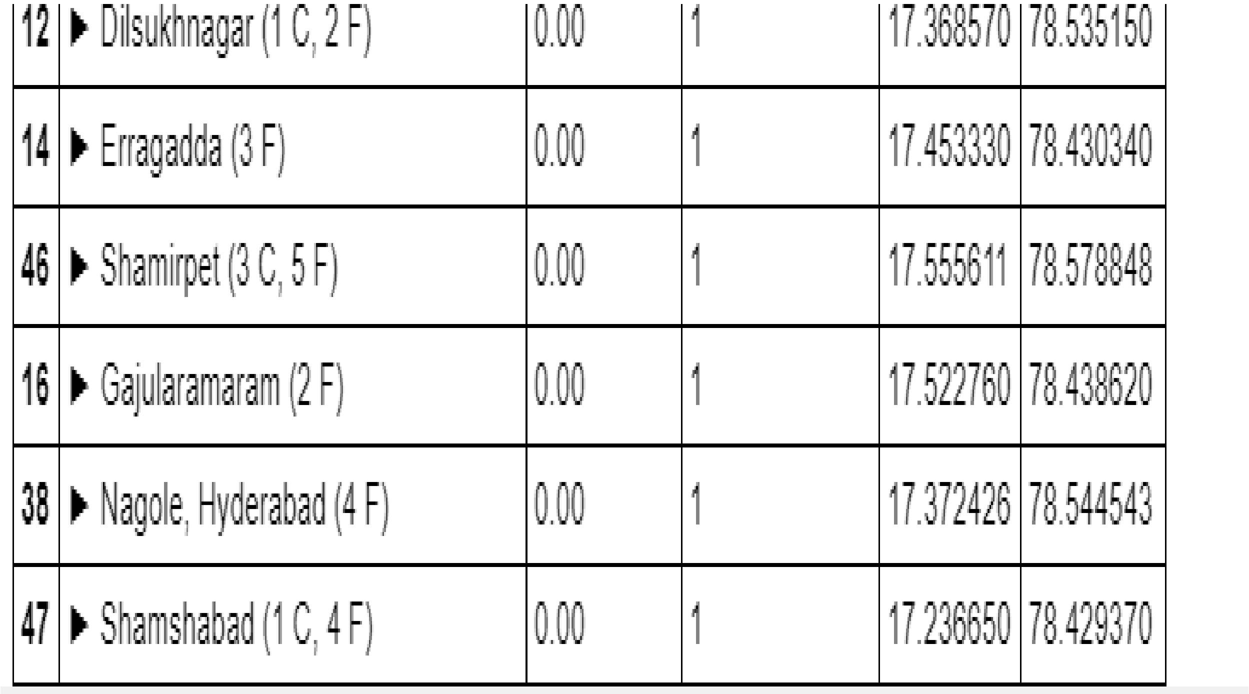


- **Cluster 0:** Neighborhoods with moderate number of existence of Restaurants.

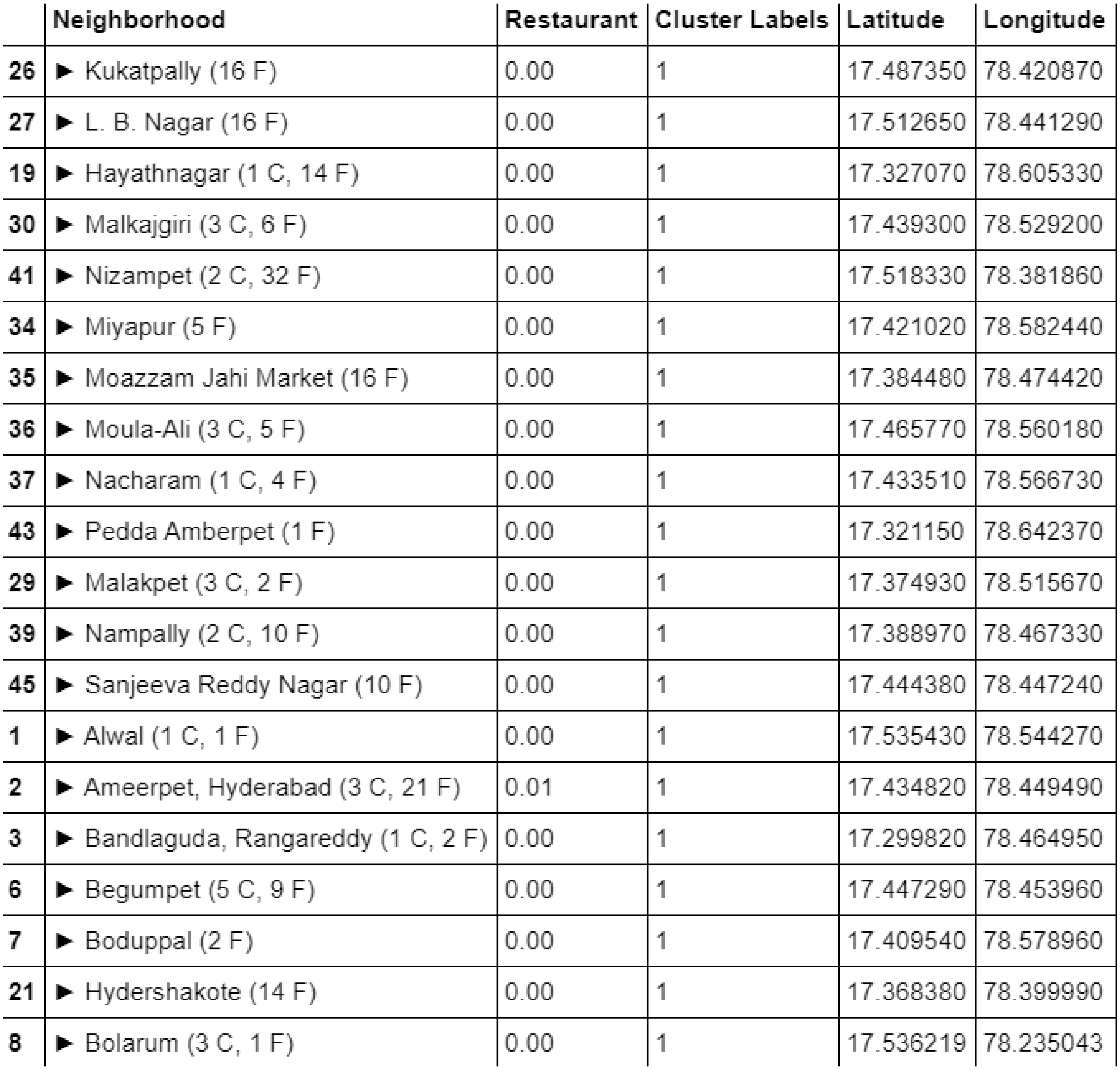


	Neighborhood	Restaurant	Cluster Labels	Latitude	Longitude
44	► Sanathnagar (8 F)	0.066667	0	17.458760	78.44310
18	► HITEC City (5 C, 29 F)	0.070000	0	17.448230	78.37429
17	► Golconda (5 C, 4 F)	0.071429	0	17.389410	78.40406
50	► Tarnaka (1 C, 6 F)	0.055556	0	17.408935	78.32674
48	► Sitaphalmandi (1 C, 1 F)	0.055556	0	17.408935	78.32674
31	► Manikonda (8 F)	0.100000	0	17.401390	78.39163
13	► Domalguda (3 C)	0.048780	0	17.409950	78.48229
9	► Cavalry Barracks, Hyderabad (1 C)	0.055556	0	17.408935	78.32674
28	► Madhapur (1 C, 19 F)	0.081395	0	17.459000	78.36810
11	► Dabirpura (1 C)	0.055556	0	17.408935	78.32674

- **Cluster 1:** Neighborhoods with low or no concentration of Restaurants

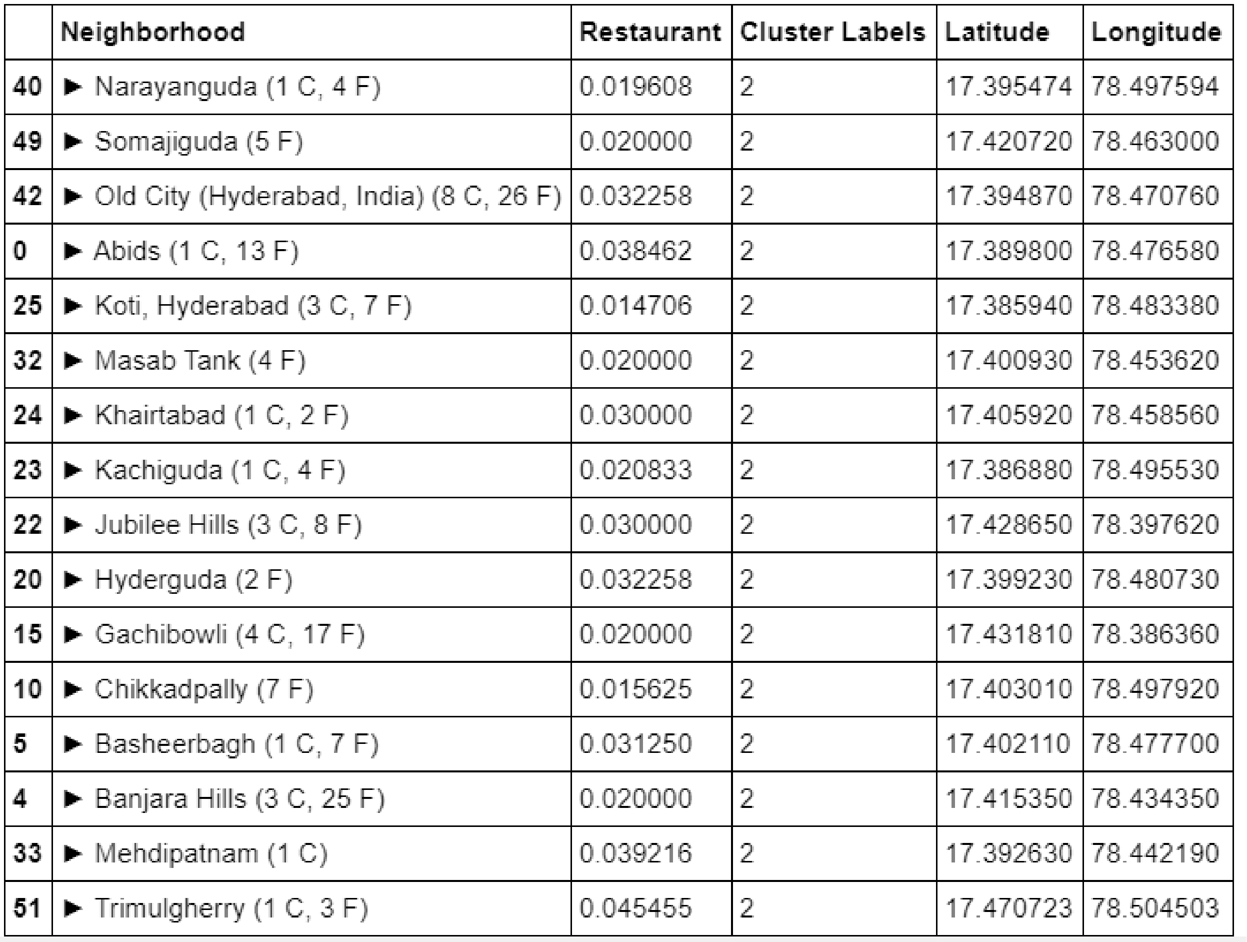


12	▶ Dilsukhnagar (1 C, 2 F)	0.00	1	17.368570	78.535150
14	▶ Erragadda (3 F)	0.00	1	17.453330	78.430340
46	▶ Shamirpet (3 C, 5 F)	0.00	1	17.555611	78.578848
16	▶ Gajularamaram (2 F)	0.00	1	17.522760	78.438620
38	▶ Nagole, Hyderabad (4 F)	0.00	1	17.372426	78.544543
47	▶ Shamshabad (1 C, 4 F)	0.00	1	17.236650	78.429370



	Neighborhood	Restaurant	Cluster Labels	Latitude	Longitude
26	► Kukatpally (16 F)	0.00	1	17.487350	78.420870
27	► L. B. Nagar (16 F)	0.00	1	17.512650	78.441290
19	► Hayathnagar (1 C, 14 F)	0.00	1	17.327070	78.605330
30	► Malkajgiri (3 C, 6 F)	0.00	1	17.439300	78.529200
41	► Nizampet (2 C, 32 F)	0.00	1	17.518330	78.381860
34	► Miyapur (5 F)	0.00	1	17.421020	78.582440
35	► Moazzam Jahi Market (16 F)	0.00	1	17.384480	78.474420
36	► Moula-Ali (3 C, 5 F)	0.00	1	17.465770	78.560180
37	► Nacharam (1 C, 4 F)	0.00	1	17.433510	78.566730
43	► Pedda Amberpet (1 F)	0.00	1	17.321150	78.642370
29	► Malakpet (3 C, 2 F)	0.00	1	17.374930	78.515670
39	► Nampally (2 C, 10 F)	0.00	1	17.388970	78.467330
45	► Sanjeeva Reddy Nagar (10 F)	0.00	1	17.444380	78.447240
1	► Alwal (1 C, 1 F)	0.00	1	17.535430	78.544270
2	► Ameerpet, Hyderabad (3 C, 21 F)	0.01	1	17.434820	78.449490
3	► Bandlaguda, Rangareddy (1 C, 2 F)	0.00	1	17.299820	78.464950
6	► Begumpet (5 C, 9 F)	0.00	1	17.447290	78.453960
7	► Boduppal (2 F)	0.00	1	17.409540	78.578960
21	► Hydershakote (14 F)	0.00	1	17.368380	78.399990
8	► Bolarum (3 C, 1 F)	0.00	1	17.536219	78.235043

- **Cluster 2:** Neighborhoods with high concentration of Restaurants.



	Neighborhood	Restaurant	Cluster Labels	Latitude	Longitude
40	► Narayanguda (1 C, 4 F)	0.019608	2	17.395474	78.497594
49	► Somajiguda (5 F)	0.020000	2	17.420720	78.463000
42	► Old City (Hyderabad, India) (8 C, 26 F)	0.032258	2	17.394870	78.470760
0	► Abids (1 C, 13 F)	0.038462	2	17.389800	78.476580
25	► Koti, Hyderabad (3 C, 7 F)	0.014706	2	17.385940	78.483380
32	► Masab Tank (4 F)	0.020000	2	17.400930	78.453620
24	► Khairtabad (1 C, 2 F)	0.030000	2	17.405920	78.458560
23	► Kachiguda (1 C, 4 F)	0.020833	2	17.386880	78.495530
22	► Jubilee Hills (3 C, 8 F)	0.030000	2	17.428650	78.397620
20	► Hyderguda (2 F)	0.032258	2	17.399230	78.480730
15	► Gachibowli (4 C, 17 F)	0.020000	2	17.431810	78.386360
10	► Chikkadpally (7 F)	0.015625	2	17.403010	78.497920
5	► Basheerbagh (1 C, 7 F)	0.031250	2	17.402110	78.477700
4	► Banjara Hills (3 C, 25 F)	0.020000	2	17.415350	78.434350
33	► Mehdipatnam (1 C)	0.039216	2	17.392630	78.442190
51	► Trimulgherry (1 C, 3 F)	0.045455	2	17.470723	78.504503

Discussion Section:

As observations noted from map in the result section, most of Restaurants are concentrated in the central area of Hyderabad, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to no Restaurants in the neighborhoods. This represents a great opportunity and high potential areas to open new Restaurants as there is very little to no competition from existing Restaurants. From another perspective, the results also show that the oversupply of Restaurants mostly happened in the central area of the city, with the suburb area still have very few Restaurants Therefore, this project recommends investors to capitalize on these findings to open new Restaurants in neighborhoods in cluster 2 with little to no competition. Business investors 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 0 and 1 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 investors regarding the best locations to open a new Restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 0 are the most preferred locations to open a new Restaurant. The findings of this project will help the relevant investors to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Restaurant.

References:

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