**IBM DATA SCIENCE CAPSTONE PROJECT**

**OPENING A NEW RESTAURANT IN HYDERABAD, INDIA**

**By Varija Reddy Sankapalli**

**INTRODUCTION**

Hyderabad city is the sixth-most populous urban cluster in India. Western parts of the city have rapidly grown in the recent times, owing to IT and Service Sector while the eastern parts have become residential hubs while the central part forms the commercial, economic and cultural core of the city. I was born and grew up in this city and I have huge bonding to this place. The multi-cultural nature and increase in multinational companies in the city of Hyderabad has brought along with-it numerous cuisines from all over the world. The people of India generally love food and I personally love to try different cuisines and experience different flavors. Thus, the aim of this project is to study the neighborhoods in Hyderabad to determine possible locations for opening a restaurant.



**BUSINESS PROBLEM**

The objective of this Capstone project is to analyze and select the best locations in the city of Hyderabad, India to open a new restaurant. Using data science methodology and machine learning techniques like clustering, this project can be useful for business owners and entrepreneurs who are looking to invest in a restaurant in Hyderabad. The main objective of this project is to carefully analyze appropriate data and find recommendations for the stakeholders.

**DATA**

For this project, we need the following data:

1. Neighborhood Data: The data for the neighborhoods in Hyderabad was extracted from Wikipedia website. The hyperlink of the website is found [here](https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Hyderabad). The data is read into a pandas data frame using the read\_html() method. Wikipedia page provides a comprehensive and detailed table of the data which can easily be extracted using the read\_html() method of pandas.
2. Geographical coordinates: The geographical coordinates of Hyderabad can be obtained from the GeoPy library in python. This data is relevant for plotting the map of Hyderabad using the Folium library in python. The geocoder library in python can be used to obtain latitude and longitude data for various neighborhoods in Hyderabad and the coordinates of all neighborhoods in Hyderabad are used to check the accuracy of coordinates given on Wikipedia and replace them in our data frame if the absolute difference is more than 0.001. These coordinates are then further used for plotting using the Folium library in python.
3. Venue Data: The venue can be extracted using the Foursquare API. This data contains venue recommendations for all neighborhoods in Hyderabad and is used to study the popular venues of different neighborhoods.

**HYDERABAD MAP**

This map represents all the 200 nearby neighborhoods of Hyderabad, India. We can use K-Means clustering machine learning technique to group all the neighborhoods into different clusters.

Map

Description automatically generated

**METHODOLOGY**

Firstly, data is retrieved from Wikipedia website, to get the list of neighborhoods in Hyderabad, India. The hyperlink for the website is given [here](https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Hyderabad). Web scraping is done using Python requests and Beautiful Soup packages to extract the list of neighborhoods of Hyderabad. Here, we get a list of names. Next, Foursquare API is used to get the geographical coordinates in the form of latitudes and longitudes. So, wonderful Geocoder package is used to convert the address into geographical coordinates in the form of latitude and longitude. After gathering the data, the data id populated into pandas Data Frame and the neighborhoods are visualized using Folium package. This allows us to check whether the geographical coordinates data returned by Geocoder are plotted in the city of Hyderabad. Next, Foursquare API is used to get the top venues that are within a radius of 1000 meters. In order to use Foursquare API, we need to register a Foursquare Account in order to obtain the Foursquare ID and Foursquare secret key.

API calls are made 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 extract the venue name, venue category, venue latitude and longitude. With this 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. We analyze each neighborhood by grouping the rows by neighborhood and take the mean of the frequency of occurrence of each venue category. By doing so, we can also use the data in clustering. Since we are analyzing the best place to start a new restaurant, most common top ten venues are sorted for the neighborhood.

Now, Clustering neighborhoods of Hyderabad, India is done using K-Means Clustering. Here, we used Silhouette Score to determine how many clusters are needed to use for the required data. A function is defined to plot the Silhouette Score that will be calculated using different number of clusters. A plot is generated where we can see that the silhouette scores are not very high even as we increase the number of clusters. This means that the inter-cluster distance between different clusters is not very high over the range of k-values. However, we will try to cluster our data as best as we can. For this, we will use 2 clusters for our clustering model since it provides the highest silhouette score. A new data frame is created that includes cluster labels and top 10 venues.

**RESULTS**

The results from K-Means clustering show that the neighborhood is categorized into two clusters.

Cluster 0: Neighborhoods with less numbers of restaurants.

Cluster 1: Neighborhoods with a greater number of restaurants.

The results of Clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color.

Map

Description automatically generated

**DISCUSSIONS**

By analyzing the two clusters obtained we can see that some of the clusters are more suited for restaurants and hotels, whereas other clusters are less suited. These clusters contain a higher degree of other venues like hotels, pizza place and clothing store. Thus, they are not well suited for opening a new restaurant. Thus, the neighborhoods in these clusters would be well suited for opening a new restaurant.

Comparing clusters 1 and 2, neighborhoods in cluster 1 seem to be more suited for starting a restaurant since they contain a larger percentage of food joints in the top 10 most common venues than cluster 2. The neighborhoods in cluster 1 contain a variety of food joints like restaurants, tea rooms, bakery, cafe and also contain very diverse cuisines like Indian, Chinese and seafood restaurants. Most neighborhoods in cluster 2 seem to have Indian Restaurant as their topmost common venue; however, on careful analysis we can see that neighborhoods in cluster 2 also contain other venues like Dessert Shop, Pizza place and so on. Thus, it is recommended that the new restaurant can be opened in the neighborhoods belonging to cluster 1.

**LIMITATIONS AND SUGGESTIONS**

This project was done only considering best place to open a new restaurant in Hyderabad, India. There are also other factors such as population and income of residents that could influence the decision of selecting a place to open a new restaurant. However, such data is not available to the neighborhood level required for this project. Future research may include a methodology to estimate such data to be used in clustering algorithm to determine the preferred locations to open a new restaurant in Hyderabad. In addition, Sandbox Tier Account of Foursquare API is used here that has limitations to the number of API calls and results returned. Future research may use paid account to overcome these limitations to get results.

**CONCLUSION**

We have successfully analyzed the neighborhoods in Hyderabad, India for determining which would be the best neighborhoods for opening a new restaurant. Based on our analysis, neighborhoods in cluster 1 are recommended as locations for the new restaurant. The stakeholders and investors can further tune this by considering various other factors like transport, legal requirements, and costs associated. These were out of the scope for this project and thus were not considered.