Capstone Project

Comparing the neighbourhoods of Hyderabad City, India.

1. Introduction:

Hyderabad is the capital city of Indian state, Telangana. It is located in the northern part of south India. The city comprises of an estimated population of 9.7 million as of 2021 and is sixth most populous metropolitan area in India. It is also the fifth largest urban economy in India. The amalgamation of local and migrated individuals led to a distinctive culture and the city is emerged as the foremost center of oriental culture. Many crafts such as painting, jewelry, literature, clothing… so on are still prominent in Hyderabad city.

The Telugu film industry in the city is the second largest film production industry in the country. The city, emerged as pharmaceuticals and biotechnology hub in India. The formation of HITEC city, dedicated to information technology has encouraged multinational companies like Google, Amazon, Apple, Facebook and Microsoft to set up their operations in Hyderabad city.

Just like India, the Hyderabad city is one of the best representations of its great history, diverse culture and food. The city has different cuisines, and is listed as UNESCO creative city of gastronomy. The city is famous for its popular food, “Biryani”. Many restaurants provide different cuisines of food. The city has rich food culture dating back to Nizam’s and Mughal empire. Some of the cuisines the city offers are Arabic, Turkish, Iranian and native Telugu cuisines.

1.1 The Business Problem:

The Hyderabad city is the one of the best locations in the country to open a restaurant. The city offers different cuisines and rich diverse food to the people. Clearly, setting up a restaurant in the city is most profitable for business. The diversity of people and their food preferences enable to compete in food business. Our main problem here is to choose a location or a neighborhood in the city to open a restaurant. It is the job of a data scientist to gather the information on all neighbourhoods in the city and present it to the stake holders. So that they will make the business decision on the location to open a restaurant.

2. Data sources and data cleaning:

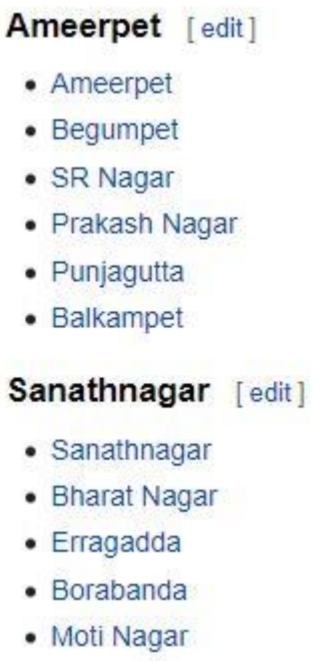
For this project, Hyderabad city neighborhood names, their respective latitude and longitude coordinates and the nearby venues for each neighborhood data are required. The following are the data sources for our project.

2.1 Get Neighborhood names:

After a quick google search, it was found that there was a Wikipedia page, that provides the information on neighbourhoods of Hyderabad city. Below is the link for the Wikipedia webpage.

Wikipedia-link:
https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Hyderabad

The data on the Wikipedia webpage was as shown below:



Ameerpet [edit]

- Ameerpet
- Begumpet
- SR Nagar
- Prakash Nagar
- Punjagutta
- Balkampet

Sanathnagar [edit]

- Sanathnagar
- Bharat Nagar
- Erragadda
- Borabanda
- Moti Nagar

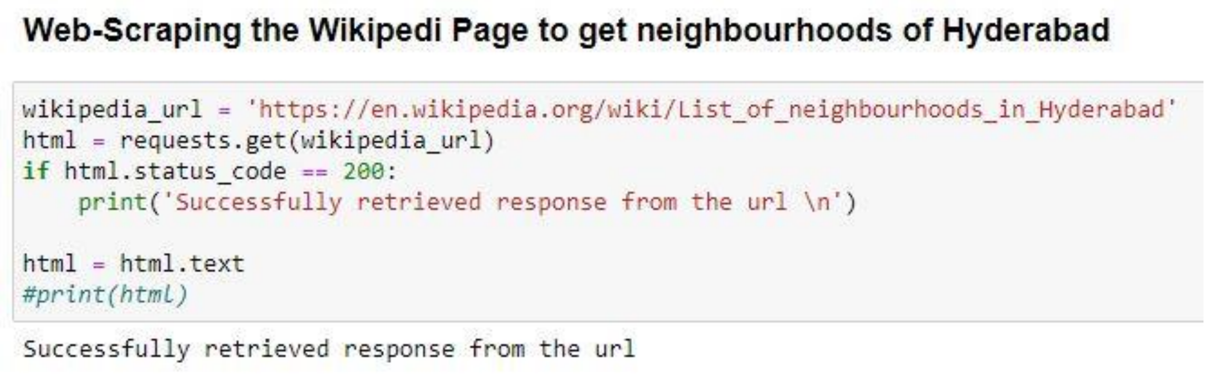
The neighborhood names are displayed row wise. Each neighborhood name is a link to its own Wikipedia webpage. We need to fetch the neighborhood names from the Wikipedia website.

To get data, let’s web-scrape this Wikipedia webpage and get the required data. For web-scraping, beautiful soup python package was used. On inspecting the webpage, our required data in the “<li>” tags.

Python requests library was used to fetch data from the URL. Beautiful soup object was used on the data from webpage. The fetched data is then filtered-out to obtain our required data. Finally, our required data is stored in a pandas data frame.

The web scraping was done as shown in the below figures.

Fig 1:



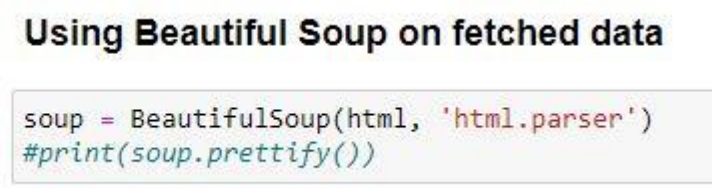
Web-Scraping the Wikipedi Page to get neighbourhoods of Hyderabad

```
wikipedia_url = 'https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Hyderabad'
html = requests.get(wikipedia_url)
if html.status_code == 200:
    print('Successfully retrieved response from the url \n')

html = html.text
#print(html)
```

```
Successfully retrieved response from the url
```

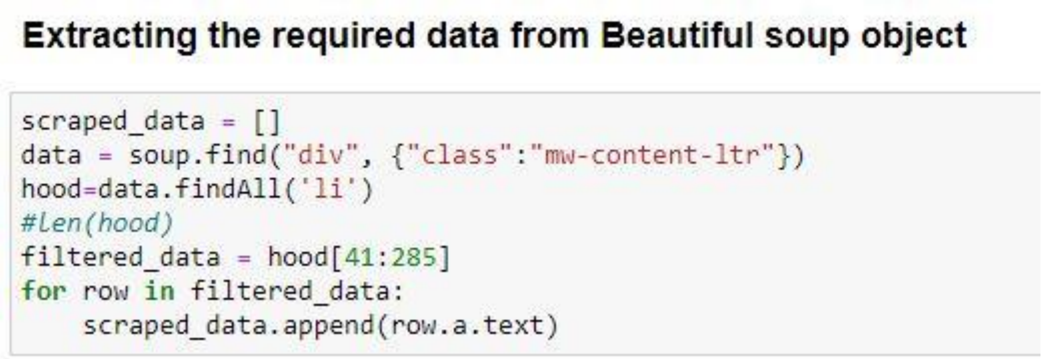
Fig 2:



Using Beautiful Soup on fetched data

```
soup = BeautifulSoup(html, 'html.parser')
#print(soup.prettify())
```

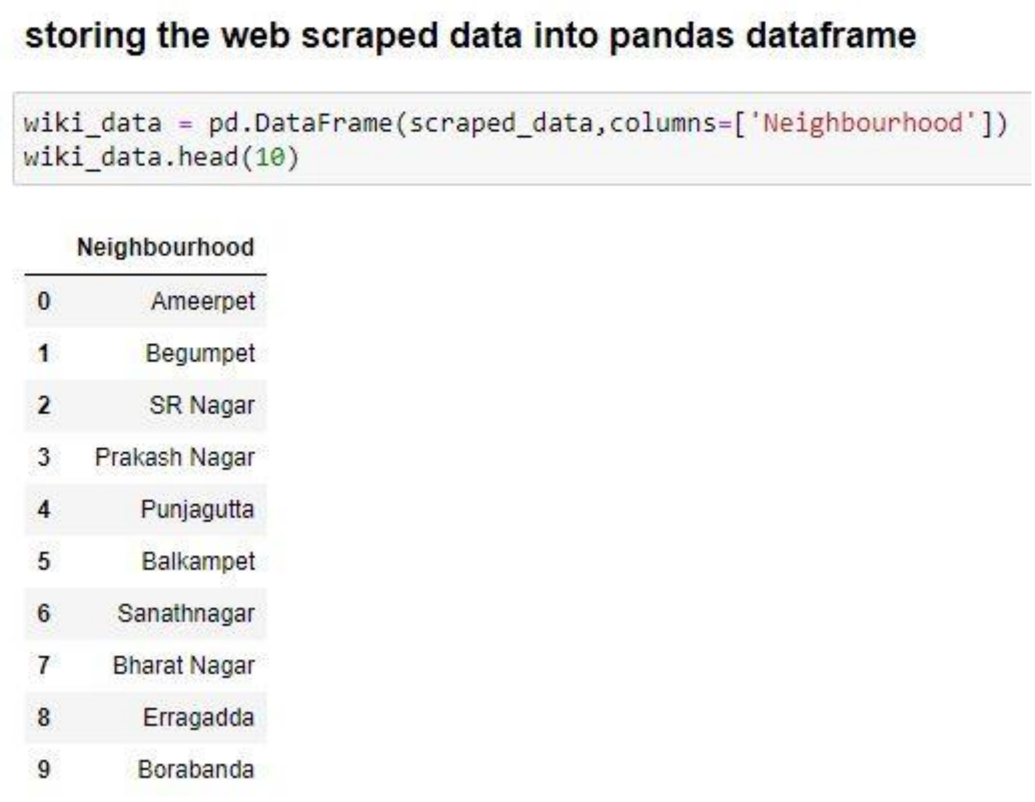
Fig 3:



Extracting the required data from Beautiful soup object

```
scraped_data = []
data = soup.find("div", {"class":"mw-content-ltr"})
hood=data.findAll('li')
#len(hood)
filtered_data = hood[41:285]
for row in filtered_data:
    scraped_data.append(row.a.text)
```

Fig 4:



storing the web scraped data into pandas dataframe

```
wiki_data = pd.DataFrame(scraped_data,columns=['Neighbourhood'])
wiki_data.head(10)
```

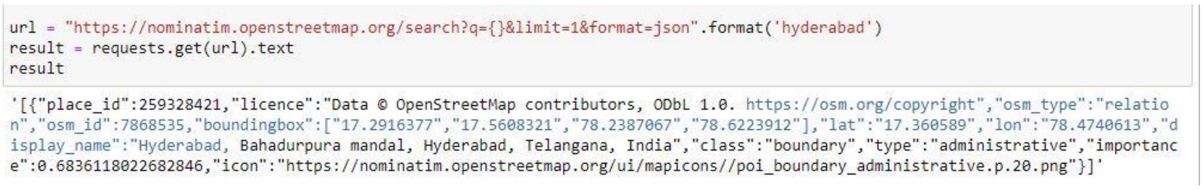
	Neighbourhood
0	Ameerpet
1	Begumpet
2	SR Nagar
3	Prakash Nagar
4	Punjagutta
5	Balkampet
6	Sanathnagar
7	Bharat Nagar
8	Erragadda
9	Borabanda

2.2 Get coordinates for neighborhoods:

We got the neighborhood names of Hyderabad city by web scraping the Wikipedia webpage. The latitude and longitude coordinates of the neighborhoods were not available in the Wikipedia page. To obtain the coordinates, the openstreetmap.org website's nominatim API was used. An API request should be sent to the API to get data. The following is the URL format.

URL:
"https://nominatim.openstreetmap.org/search?q={}&limit=1&format=json"

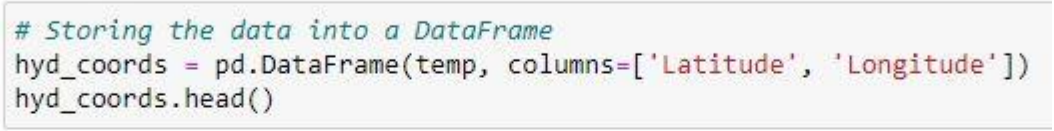
The response we get, contains the coordinates for a requested address in json format. By using this API, coordinates for all the neighborhoods of Hyderabad city are obtained. This is shown below.



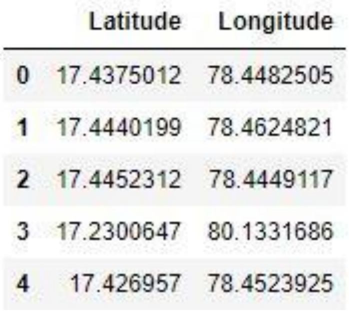
```
url = "https://nominatim.openstreetmap.org/search?q={}&limit=1&format=json".format('hyderabad')
result = requests.get(url).text
result
```

```
'[{"place_id":259328421,"licence":"Data © OpenStreetMap contributors, ODbL 1.0. https://osm.org/copyright","osm_type":"relatio
n","osm_id":7868535,"boundingbox":["17.2916377","17.5608321","78.2387067","78.6223912"],"lat":"17.360589","lon":"78.4740613","d
isplay_name":"Hyderabad, Bahadurpura mandal, Hyderabad, Telangana, India","class":"boundary","type":"administrative","importanc
e":0.683611802268284б,"icon":"https://nominatim.openstreetmap.org/ui/mapicons//poi_boundary_administrative.p.20.png"}]'
```

The API response is in json format, here our required data is latitudes and longitude coordinates. The coordinates are filtered-out from the response and stored in a pandas data frame as shown below.



```
# Storing the data into a DataFrame
hyd_coords = pd.DataFrame(temp, columns=['Latitude', 'Longitude'])
hyd_coords.head()
```



	Latitude	Longitude
0	17.4375012	78.4482505
1	17.4440199	78.4624821
2	17.4452312	78.4449117
3	17.2300647	80.1331686
4	17.426957	78.4523925

2.3 Data Cleaning: The challenge with coordinates data

By using the nominatim API, coordinates for 14 neighborhoods are not obtained. This is mainly because,

1. For some neighbourhoods, the API doesn’t provide coordinates due to unknown reason.
2. Few neighborhood names in the Wikipedia webpage are misspelt. This was found by manually searching for coordinates on google website.

This challenge has overcome by searching coordinates for remaining neighborhoods, using the below mentioned website.

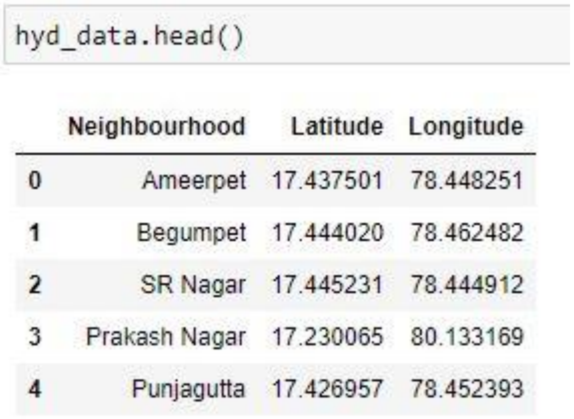
Website: https://www.latlong.net/

This website provides latitude and longitude coordinates for a requested address or a place. This is a free website; it allows only a limited number of searches in a day.

Finally, after getting coordinates for all the neighborhoods, the data is stored in a pandas data frame. Now our required data is stored in two data frames. One for neighborhood names and the other for coordinates. These two data frames are merged into a single data frame which is required data for our project.

After merging the neighborhoods data and the coordinates data,

The final data frame obtained was as shown below.



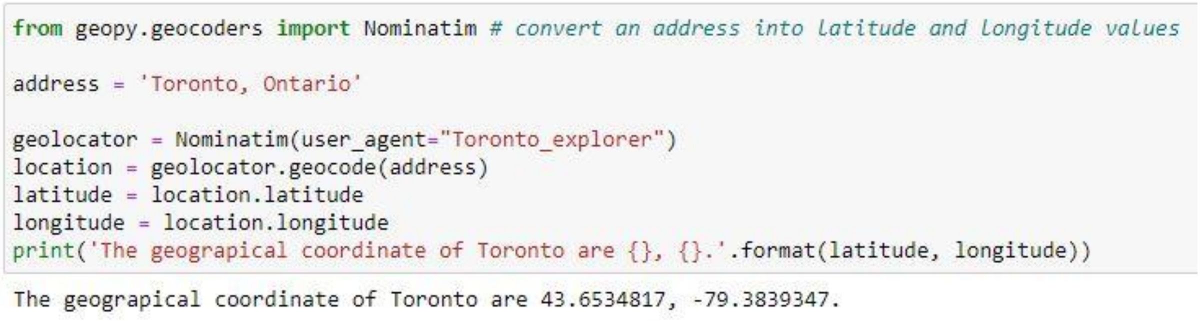
```
hyd_data.head()
```

	Neighbourhood	Latitude	Longitude
0	Ameerpet	17.437501	78.448251
1	Begumpet	17.444020	78.462482
2	SR Nagar	17.445231	78.444912
3	Prakash Nagar	17.230065	80.133169
4	Punjagutta	17.426957	78.452393

2.4 Geopy library to get coordinates of Hyderabad city:

Geopy is a python library that can be used to fetch coordinates of an address. This library was used to get the coordinates of Hyderabad city. The Hyderabad city coordinates will used to plot the map of Hyderabad city using python's folium visualization library. Below is an example.

Fig 5:



```
from geopy.geocoders import Nominatim # convert an address into latitude and longitude values

address = 'Toronto, Ontario'

geolocator = Nominatim(user_agent="Toronto_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Toronto are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of Toronto are 43.6534817, -79.3839347.
```

2.5 Neighborhood location data using Foursquare API:

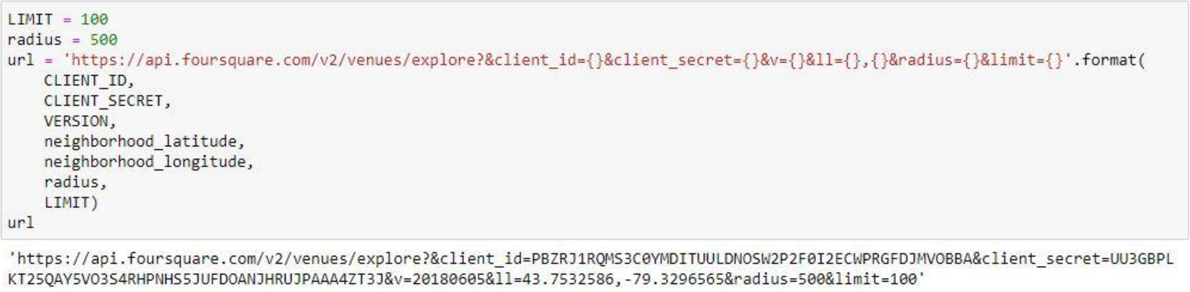
To get the nearby venues of a neighborhood, Foursquare API was used. Foursquare API provides location data of an address. It provides diverse information about venues, users, photos, check-in's, geo-tagging…etc.

This API was used in this project to get nearby venue details of a neighborhood. To get data from the API, a search query is to be sent to the foursquare API.

The response from the API contains the requested data in json format. The required data from Foursquare API was venues of a neighborhood.

The API requires credentials like CLIENT_ID, CLIENT_SECRET and VERSION to get data. These can be obtained by creating an account on Foursquare website. The response from API is in json format. The response is filtered-out to get our required data. Finally, the data is stored in a data frame.

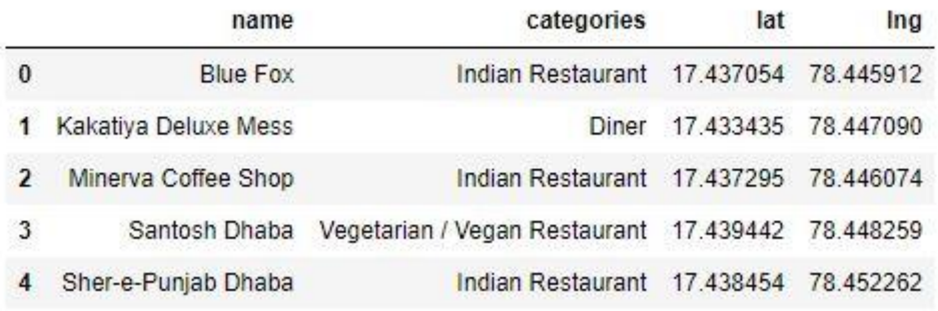
Fig 6: Request to Foursquare API:



```
LIMIT = 100
radius = 500
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    neighborhood_latitude,
    neighborhood_longitude,
    radius,
    LIMIT)
url
```

```
'https://api.foursquare.com/v2/venues/explore?&client_id=PBZRJ1RQMS3C0YMDITUULDNOSW2P2F0I2ECWPRGFDJMVOBBA&client_secret=UU3GBPL
KT25QAY5VO3S4RHPNHS5JUFDOANJHRUJPAAA4ZT3J&v=20180605&ll=43.7532586,-79.3296565&radius=500&limit=100'
```

Fig 7: The filtered json response stored in a data frame:



	name	categories	lat	lng
0	Blue Fox	Indian Restaurant	17.437054	78.445912
1	Kakatiya Deluxe Mess	Diner	17.433435	78.447090
2	Minerva Coffee Shop	Indian Restaurant	17.437295	78.446074
3	Santosh Dhaba	Vegetarian / Vegan Restaurant	17.439442	78.448259
4	Sher-e-Punjab Dhaba	Indian Restaurant	17.438454	78.452262

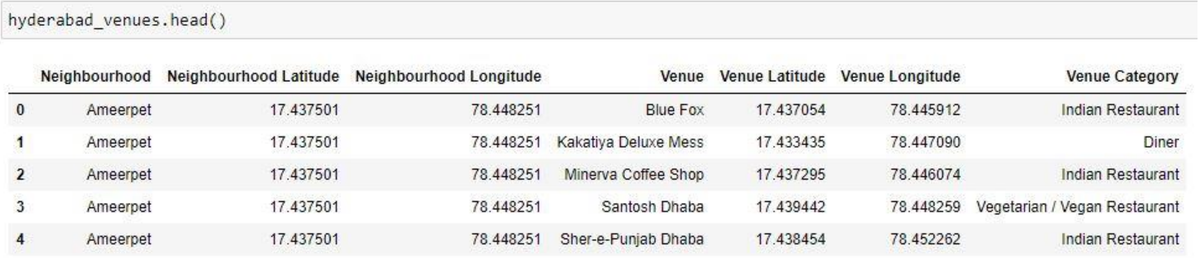
2.6 Exploratory Data Analysis:

Exploratory data analysis is not required for our project. As our project requires detailed analysis of location data such as venues of neighborhoods and neighborhood coordinates. This data was obtained from the data sources as explained above. The location data will suffice for making a decision on where to open the restaurant.

3. Methodology:

We have the data required for our project. Let’s see the statistical analysis of data. There are a total of 244 records in our final data frame named hyd_data. Therefore, we have 244 neighborhoods and their coordinates in this data frame.

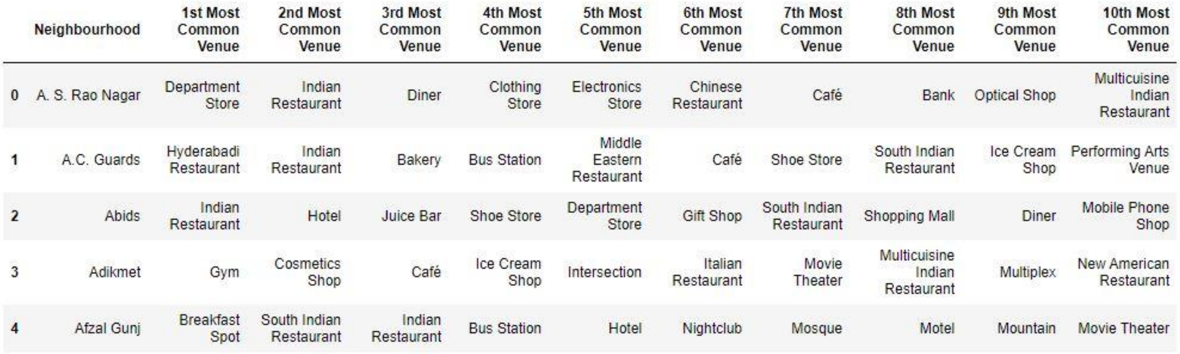
Now, we have another data frame that contains the data of neighborhood venues. The second data frame was named hyderabad_venues. This data frame contains the venues of neighborhoods. Below is screenshot of the data frame.



```
hyderabad_venues.head()
```

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ameerpet	17.437501	78.448251	Blue Fox	17.437054	78.445912	Indian Restaurant
1	Ameerpet	17.437501	78.448251	Kakatiya Deluxe Mess	17.433435	78.447090	Diner
2	Ameerpet	17.437501	78.448251	Minerva Coffee Shop	17.437295	78.446074	Indian Restaurant
3	Ameerpet	17.437501	78.448251	Santosh Dhaba	17.439442	78.448259	Vegetarian / Vegan Restaurant
4	Ameerpet	17.437501	78.448251	Sher-e-Punjab Dhaba	17.438454	78.452262	Indian Restaurant

There are 1154 records in the venues data frame. On checking the data frame, there are 186 unique venue categories. We are interested in the neighborhoods with restaurants. Let’s sort the data frame and check the top 10 most common venues in each neighborhood. So that we get an idea about neighborhoods that have restaurants. This is achieved by grouping the neighborhoods, and by taking the mean of the frequency of occurrence of each venue category. Finally, sorted data frame looks as shown below.



	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	A. S. Rao Nagar	Department Store	Indian Restaurant	Diner	Clothing Store	Electronics Store	Chinese Restaurant	Café	Bank	Optical Shop	Multicuisine Indian Restaurant
1	A.C. Guards	Hyderabadi Restaurant	Indian Restaurant	Bakery	Bus Station	Middle Eastern Restaurant	Café	Shoe Store	South Indian Restaurant	Ice Cream Shop	Performing Arts Venue
2	Abids	Indian Restaurant	Hotel	Juice Bar	Shoe Store	Department Store	Gift Shop	South Indian Restaurant	Shopping Mall	Diner	Mobile Phone Shop
3	Adikmet	Gym	Cosmetics Shop	Café	Ice Cream Shop	Intersection	Italian Restaurant	Movie Theater	Multicuisine Indian Restaurant	Multiplex	New American Restaurant
4	Afzal Gunj	Breakfast Spot	South Indian Restaurant	Indian Restaurant	Bus Station	Hotel	Nightclub	Mosque	Motel	Mountain	Movie Theater

3.1 Machine Learning Algorithm:

Now that we have the sorted data frame. We will use the unsupervised machine learning algorithm “K Means Clustering” to cluster the neighborhoods. This clustering algorithm, forms clusters of neighborhoods with similar most common venues. This helps us get an idea of most common venue in a cluster.

3.2 K Means Algorithm:

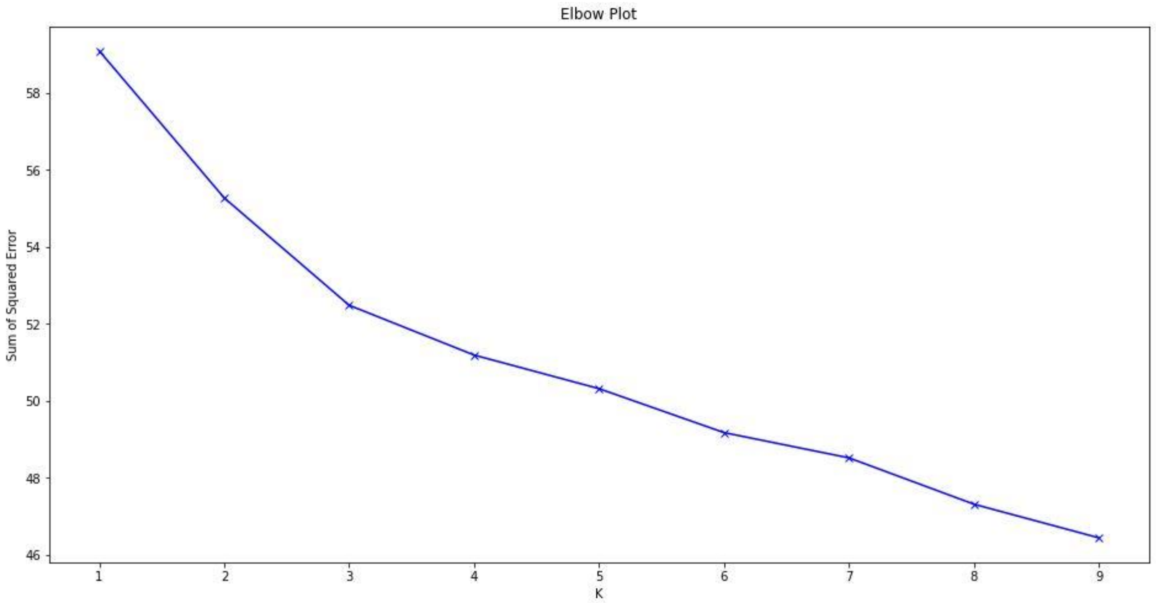
K Means algorithm is a clustering algorithm. It is the most popular and widely used clustering algorithm. The algorithm takes a dataset of items as input and categorize those items into groups called clusters. The algorithm is as shown below.

1. Randomly initialize n number of data points from the dataset called cluster centroids. Here n is the number of cluster centroids.
2. Now the distance between remaining data points and cluster centroids are calculated. The data points with least distance to cluster centroids are assigned to them to form clusters.
3. The mean of distances between the data points and the centroid of that particular cluster is calculated. The centroid is moved to this calculated mean position.

The above process is repeated in an iterative manner, until the centroids can no longer move to a different mean position. At the end of this iterative process, we will have our clusters.

3.3 Choosing number of clusters:

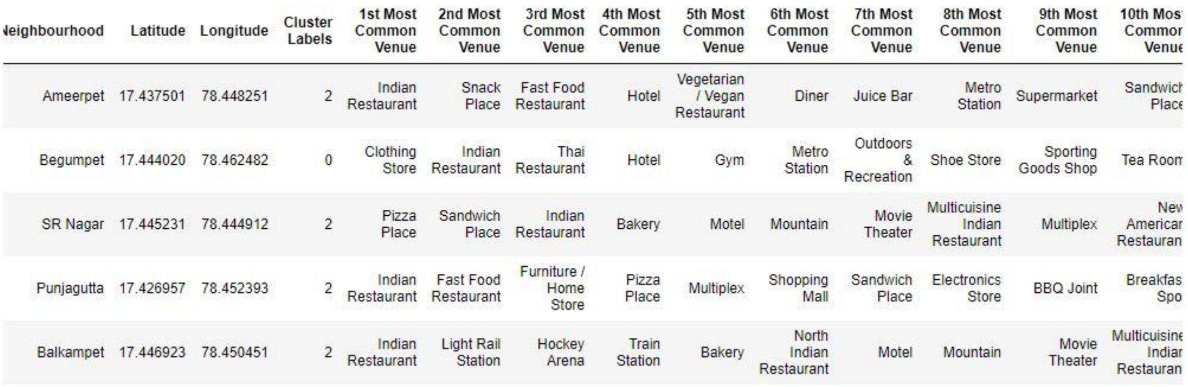
In some problems, we may have no idea on choosing 'k', the number of clusters. So, to find the optimum value for choosing number of clusters (k), we test the algorithm using elbow method. In elbow method, we select a predefined range for number of clusters (k). We run the K Means algorithm on this range of n and calculate the sum of squared distances of data points to their cluster centroid. If we plot sum of squared distances Vs number of clusters (k), we got the plot as shown below figure.



The shape of the plot is just like the shape of an elbow, hence the name elbow plot. In elbow plot, we choose k, where the plot line deviates at an angle. In above plot the line deviates at k = 3. Therefore, the optimum value for k is 3.

4. Results:

Now we use the K Means algorithm with k =3 on the sorted neighborhoods data. We add a new column to the data frame that labels the neighborhood to its cluster. The final data frame contains the neighborhood name, its coordinates, cluster label and its top 10 most common venues. The final data frame looks as shown below.

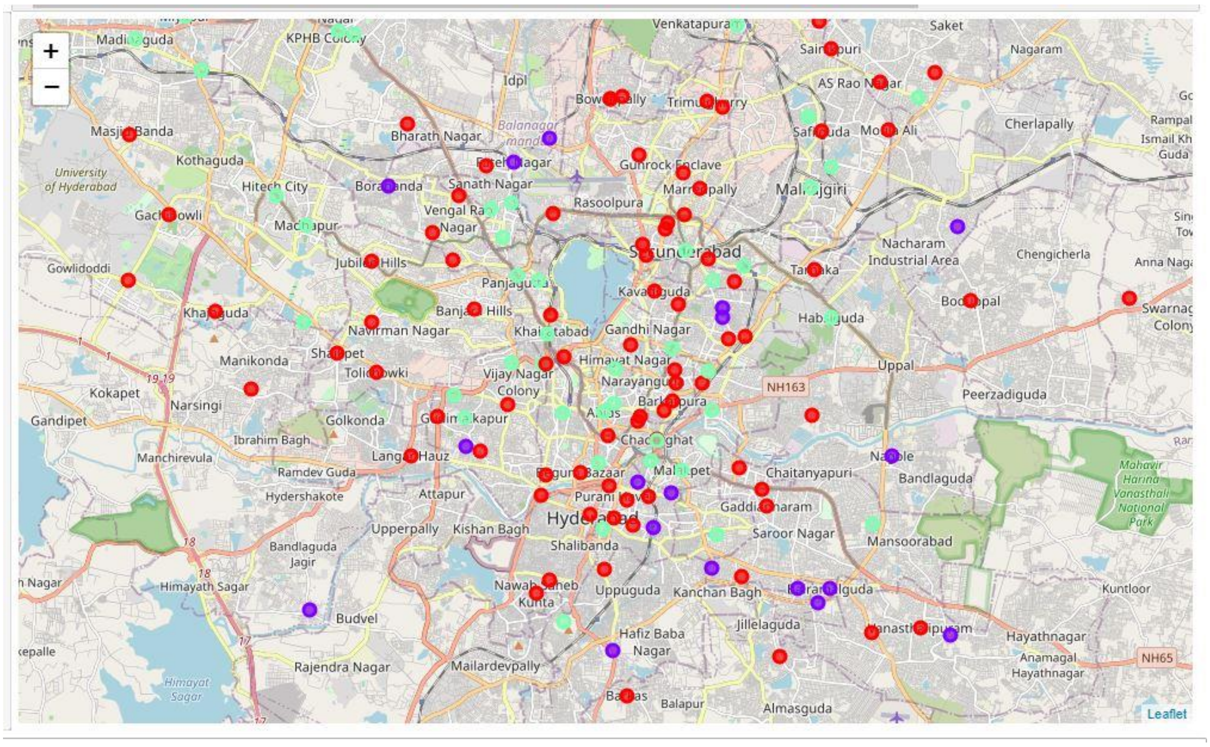


leighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Mos Commor Venue
Ameerpet	17.437501	78.448251	2	Indian Restaurant	Snack Place	Fast Food Restaurant	Hotel	Vegetarian / Vegan Restaurant	Diner	Juice Bar	Metro Station	Supermarket	Sandwich Place
Begumpet	17.444020	78.462482	0	Clothing Store	Indian Restaurant	Thai Restaurant	Hotel	Gym	Metro Station	Outdoors & Recreation	Shoe Store	Sporting Goods Shop	Tea Roon
SR Nagar	17.445231	78.444912	2	Pizza Place	Sandwich Place	Indian Restaurant	Bakery	Motel	Mountain	Movie Theater	Multicuisine Indian Restaurant	Multiplex	New Americar Restauran
Punjagutta	17.426957	78.452393	2	Indian Restaurant	Fast Food Restaurant	Furniture / Home Store	Pizza Place	Multiplex	Shopping Mall	Sandwich Place	Electronics Store	BBQ Joint	Breakfas Spo
Balkampet	17.446923	78.450451	2	Indian Restaurant	Light Rail Station	Hockey Arena	Train Station	Bakery	North Indian Restaurant	Motel	Mountain	Movie Theater	Multicuisine Indiar Restauran

4.1 Plotting the Clusters:

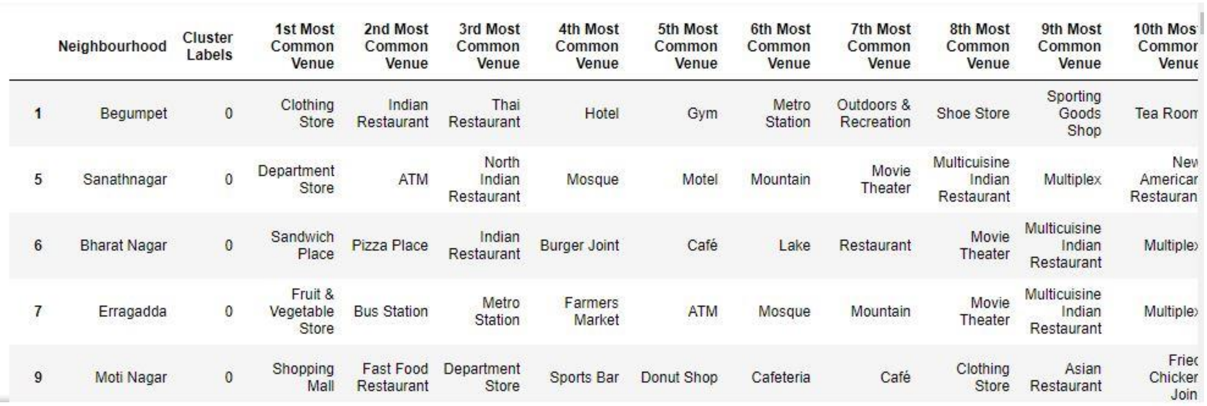
By visually plotting the results, we get better idea of the clusters and the neighborhoods. For plotting, we use the folium library. This library is very useful for plotting the location data. The folium library is used for creating beautiful map visualizations. It also has zoom functionality, that enables to zoom in the map and explore the areas in the map. The clusters in the plot are denoted with circular color markers. The three clusters are marked red, blue and green circles. The final plot result is as shown below.

Fig 8:



In the map, the red circular mark denotes cluster 1, the blue circular mark denotes cluster 2 and the green color mark denotes cluster 3. These clusters are created based on the mean of, the frequency of occurrence of similar venue categories in neighborhoods. We can verify the results of the map by taking a look at the clustered final data frame. The clustered final data frame is nothing but the final data frame split cluster wise. As we have 3 clusters in our result, the final data frame is split into 3 data frames cluster-wise. Let's take a look at them.

4.2 Cluster 1 data frame:

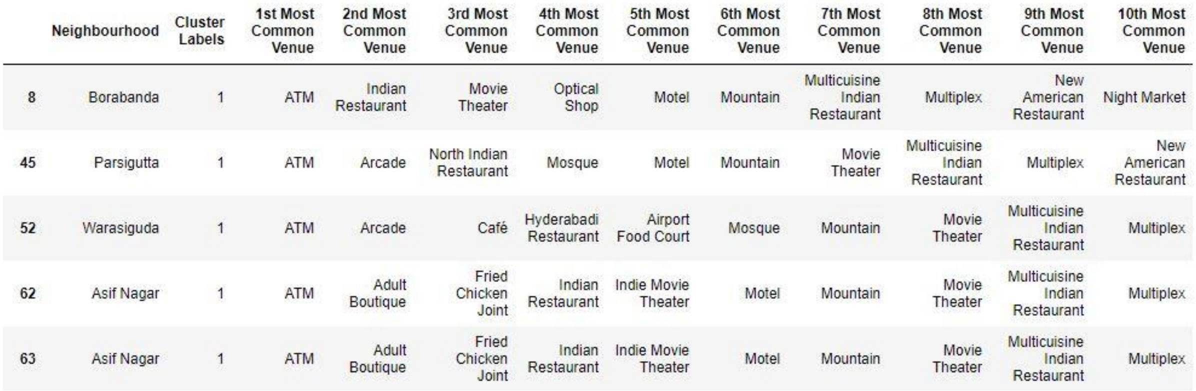


	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Mos Commor Venue
1	Begumpet	0	Clothing Store	Indian Restaurant	Thai Restaurant	Hotel	Gym	Metro Station	Outdoors & Recreation	Shoe Store	Sporting Goods Shop	Tea Room
5	Sanathnagar	0	Department Store	ATM	North Indian Restaurant	Mosque	Motel	Mountain	Movie Theater	Multicuisine Indian Restaurant	Multiplex	New Americar Restauran
6	Bharat Nagar	0	Sandwich Place	Pizza Place	Indian Restaurant	Burger Joint	Café	Lake	Restaurant	Movie Theater	Multicuisine Indian Restaurant	Multiple›
7	Erragadda	0	Fruit & Vegetable Store	Bus Station	Metro Station	Farmers Market	ATM	Mosque	Mountain	Movie Theater	Multicuisine Indian Restaurant	Multiple›
9	Moti Nagar	0	Shopping Mall	Fast Food Restaurant	Department Store	Sports Bar	Donut Shop	Cafeteria	Café	Clothing Store	Asian Restaurant	Friec Chicker Join

The above picture is the cluster 1 data frame. It contains the neighborhood name, cluster label and top 10 most common venues in the neighborhood. Let's check the second cluster data frame.

4.3 Cluster 2 data frame:

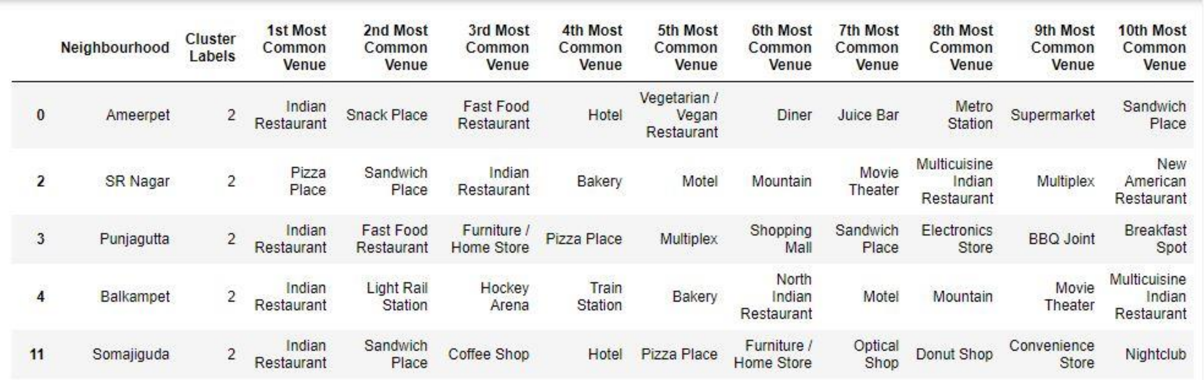
The below picture is the second cluster data frame.



	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
8	Borabanda	1	ATM	Indian Restaurant	Movie Theater	Optical Shop	Motel	Mountain	Multicuisine Indian Restaurant	Multiplex	New American Restaurant	Night Market
45	Parsigutta	1	ATM	Arcade	North Indian Restaurant	Mosque	Motel	Mountain	Movie Theater	Multicuisine Indian Restaurant	Multiplex	New American Restaurant
52	Warasiguda	1	ATM	Arcade	Café	Hyderabadi Restaurant	Airport Food Court	Mosque	Mountain	Movie Theater	Multicuisine Indian Restaurant	Multiplex
62	Asif Nagar	1	ATM	Adult Boutique	Fried Chicken Joint	Indian Restaurant	Indie Movie Theater	Motel	Mountain	Movie Theater	Multicuisine Indian Restaurant	Multiplex
63	Asif Nagar	1	ATM	Adult Boutique	Fried Chicken Joint	Indian Restaurant	Indie Movie Theater	Motel	Mountain	Movie Theater	Multicuisine Indian Restaurant	Multiplex

4.4 Cluster 3 data frame:

The below picture is cluster 3 data frame.



	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Ameerpet	2	Indian Restaurant	Snack Place	Fast Food Restaurant	Hotel	Vegetarian / Vegan Restaurant	Diner	Juice Bar	Metro Station	Supermarket	Sandwich Place
2	SR Nagar	2	Pizza Place	Sandwich Place	Indian Restaurant	Bakery	Motel	Mountain	Movie Theater	Multicuisine Indian Restaurant	Multiplex	New American Restaurant
3	Punjagutta	2	Indian Restaurant	Fast Food Restaurant	Furniture / Home Store	Pizza Place	Multiplex	Shopping Mall	Sandwich Place	Electronics Store	BBQ Joint	Breakfast Spot
4	Balkampet	2	Indian Restaurant	Light Rail Station	Hockey Arena	Train Station	Bakery	North Indian Restaurant	Motel	Mountain	Movie Theater	Multicuisine Indian Restaurant
11	Somajiguda	2	Indian Restaurant	Sandwich Place	Coffee Shop	Hotel	Pizza Place	Furniture / Home Store	Optical Shop	Donut Shop	Convenience Store	Nightclub

5. Discussion:

From the pictures of all three cluster data frames, we can make some analysis and discuss them. In first cluster data frame, we can say that restaurants are not the top 1st most common venues. But they are in 3rd and 9th most common venues. From this observation, we can say that either the competition or demand for restaurants is less in that particular cluster. If we further dive into this, and find the answer to "why the restaurants are not the top most common venues in first cluster?" We may get a clear idea whether or not to open a restaurant in this cluster.

In second cluster, restaurants are 4th and 9th most common venues. But the top most common venue are ATM's. Here, the situation is similar to the first cluster. Therefore, the same type of analysis can be applied here.

The third cluster is really interesting. Restaurants are top 1st most common venues in this cluster. This means there is a lot of demand and competition for restaurants in cluster 3. So, opening a restaurant in this cluster is good for business. But it also depends on other factors such as competition, cost of opening a restaurant in the cluster and so on. The final decision will be taken by the stake holders.

6. Conclusion:

In this project, our business problem is choosing a neighborhood to open a restaurant in Hyderabad city, India. For this problem, we first need the neighborhoods of Hyderabad and their location data. The required data was acquired from Wikipedia site and by using nominatim API. The venues data of each neighborhood was obtained using the Foursquare API. The K Means clustering algorithm was used on the data and grouped the data into 3 clusters. Each cluster was created by grouping the neighborhoods by taking the mean of the frequency of occurrence of each venue category. The final result was labelled and segmented into 3 clusters. The data was plotted using the folium library. Analysis was made on the final data sets and observations were made. The observations and the comments made on the data help the stock holders to make a decision on where to open the restaurant.