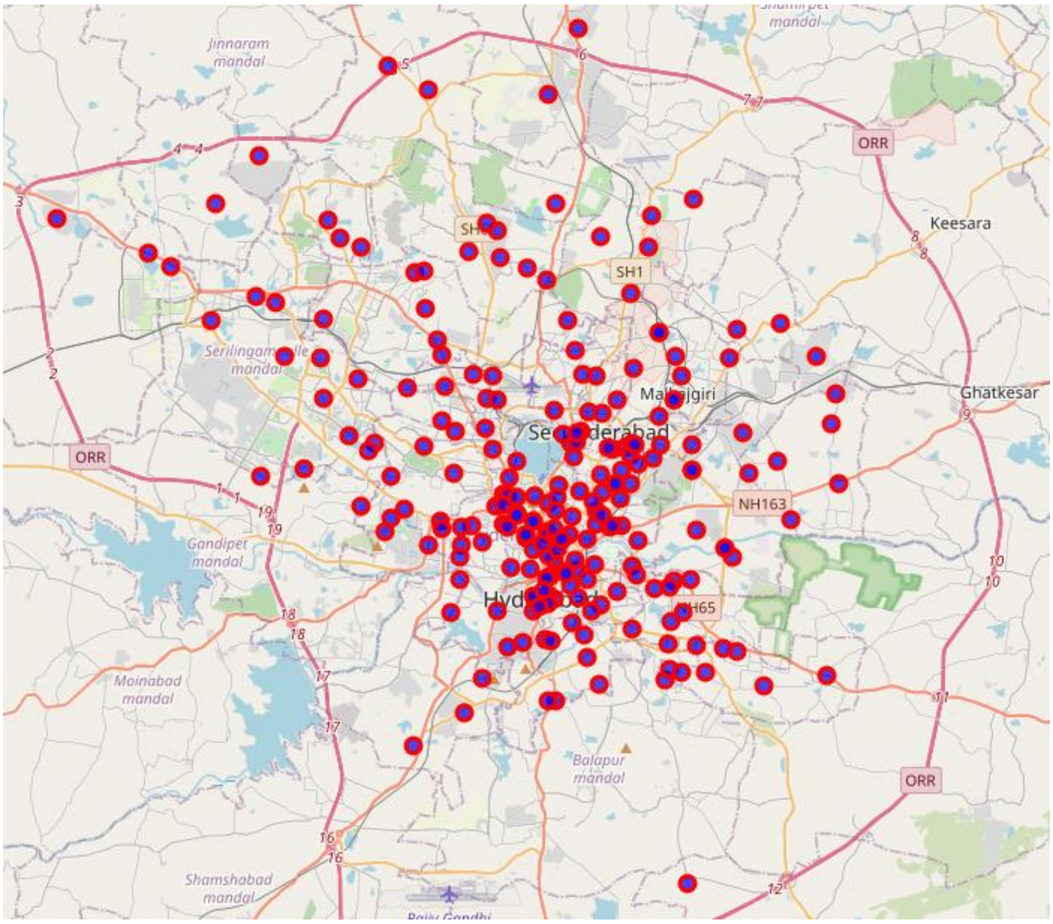
Applied Data Science Capstone Project by IBM on Coursera

Clustering of Neighbourhoods – Hyderabad, India



Sujan Kumar Kummara

Contents

List of Figures

Introduction

Hyderabad is one of the three popular cities in southern part of India. This is the city where global conglomerates such as Microsoft, Google, IBM, Facebook, Amazon, etc had set up their offices. There are more than 1000 IT firms established in this city. Apart from being an IT powerhouse, Hyderabad is also a manufacturing city with establishments such as BHEL, DRDO, NMDC, HAL, etc. Having such a huge number of firms and industries located in and around Hyderabad, this city attracts many people from various parts of India. The population of this city is ever-growing at a pace of around 2.9 % annually. Currently, the city's population is close to 10 million and is still increasing. Owing to this increasing population in the city, there is a large scope to set up businesses like, shopping malls, hotels, restaurants, coffee shops, departmental stores, etc.

Business Problem

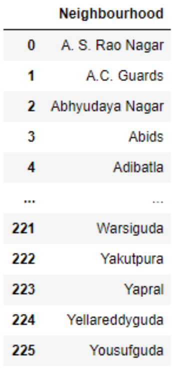
The objective of this project is to analyse the neighbourhoods in the Hyderabad city and segregate them into different clusters based on the popular venues at each neighbourhood, by using **data science methodology** and **machine learning techniques** like **clustering**. This project aims to help small business owners to select a suitable cluster to set up their businesses like hotels, restaurants, shopping malls, etc.

The **target audience** of this project are investors & developers who wish to construct shopping malls, hotels, etc., and small business owners who want to set up their businesses like restaurants, coffee shops, departmental stores, etc. It would help them to find suitable location to set up their business based on its category

Data Sources

Hyderabad city has many neighbourhoods. In order to solve the business problem of this project, the following data is required.

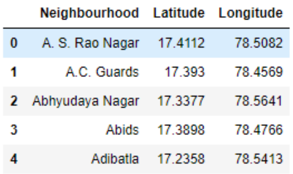
1. **List of neighbourhoods in Hyderabad:** The list of neighbourhoods in Hyderabad can be obtained from Wikipedia's page https://en.wikipedia.org/w/index.php?title=Category:Neighbourhoods_in_Hyderabad,_India&pageuntil=Sikh+Village#mw-pages. Web scraping techniques are used to scrape this Wikipedia page with the help of Python packages '**requests**' and '**beautiful soup**'



	Neighbourhood
0	A. S. Rao Nagar
1	A.C. Guards
2	Abhyudaya Nagar
3	Abids
4	Adibatla
...	...
221	Warsiguda
222	Yakutpura
223	Yapral
224	Yellareddyguda
225	Yousufguda

Figure 1 - Neighbourhoods of Hyderabad Dataset

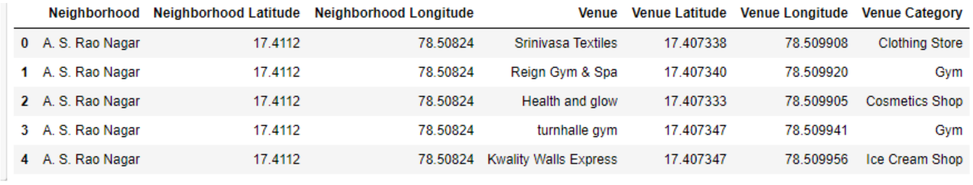
2. **Latitudes & Longitudes of Neighbourhoods:** The latitude and longitude coordinates of each neighbourhood can be obtained by using Python Geocoder package



	Neighbourhood	Latitude	Longitude
0	A. S. Rao Nagar	17.4112	78.5082
1	A.C. Guards	17.393	78.4569
2	Abhyudaya Nagar	17.3377	78.5641
3	Abids	17.3898	78.4766
4	Adibatla	17.2358	78.5413

Figure 2 - Neighbourhoods with latitude and longitude coordinates

3. **List of Venues:** The list of venues in each neighbourhood along with venue details like its latitude, longitude and category, by using FourSquare API. https://developer.foursquare.com/docs/places-api/. FourSquare has a very large dataset of venues across the globe and this data is being used by many developers.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	A. S. Rao Nagar	17.4112	78.50824	Srinivasa Textiles	17.407338	78.509908	Clothing Store
1	A. S. Rao Nagar	17.4112	78.50824	Reign Gym & Spa	17.407340	78.509920	Gym
2	A. S. Rao Nagar	17.4112	78.50824	Health and glow	17.407333	78.509905	Cosmetics Shop
3	A. S. Rao Nagar	17.4112	78.50824	turnhalle gym	17.407347	78.509941	Gym
4	A. S. Rao Nagar	17.4112	78.50824	Kwality Walls Express	17.407347	78.509956	Ice Cream Shop

Figure 3 - Venues in Hyderabad Dataset

Methodology

There are total of **225 neighbourhoods** in the city of Hyderabad. A total of **1170 venues** with **151 unique categories** are obtained in the whole city using the **FourSquare API**. As many neighbourhoods have very few venues, the results may not be accurate. So, the neighbourhoods having less than 10 venues can be filtered out from our analysis, and only the neighbourhoods having 10 or more venues will be used for our further analysis.

One hot encoding will be performed on the obtained data and get the ten most common venue categories in each neighbourhood, from that data.

Clustering technique is applied on the data containing 10 most common venue categories for each neighbourhood, to segregate the neighbourhoods of Hyderabad into separate clusters. **K-Means clustering** is used here to cluster the neighbourhoods. **Silhouette score** is used as a performance metric to obtain the optimal number of clusters.

Once the clusters are obtained, each cluster can be analysed for its existing most common venue categories. This analysis will be helpful for investors/developers to construct shopping malls, hotels etc., and small business owners to set up their restaurants, coffee shops, departmental stores etc.

Analysis

After looking into the venue data set, it is observed that there are many neighbourhoods whose count of venues is less than 10. So, the neighbourhoods having less than 10 venues are removed from the dataset to get better results. Below is the plot showing only the neighbourhoods those are having 10 or more than 10 venues.

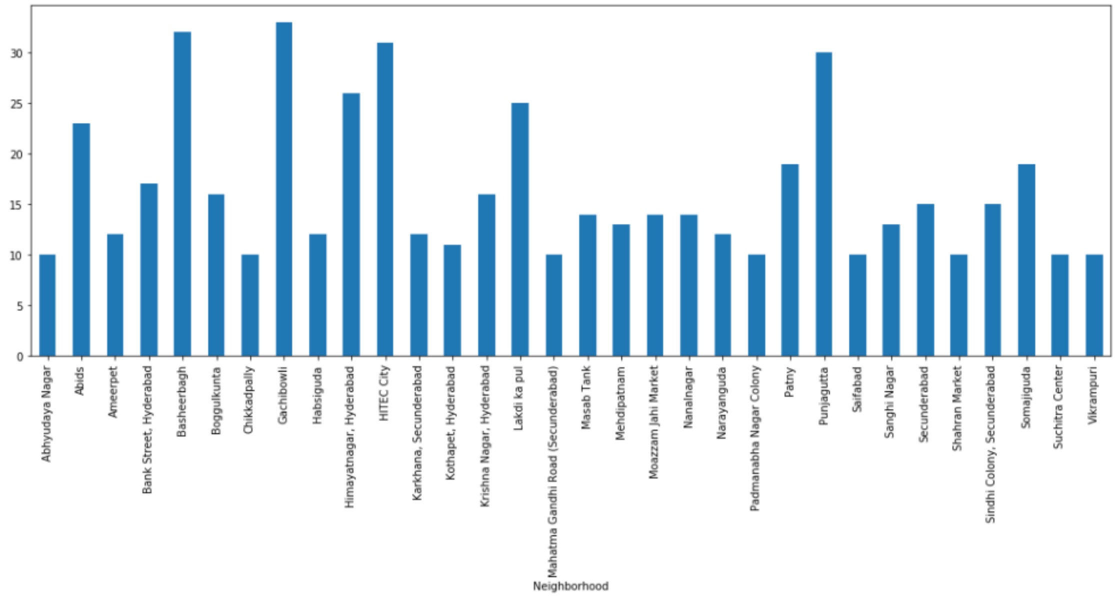
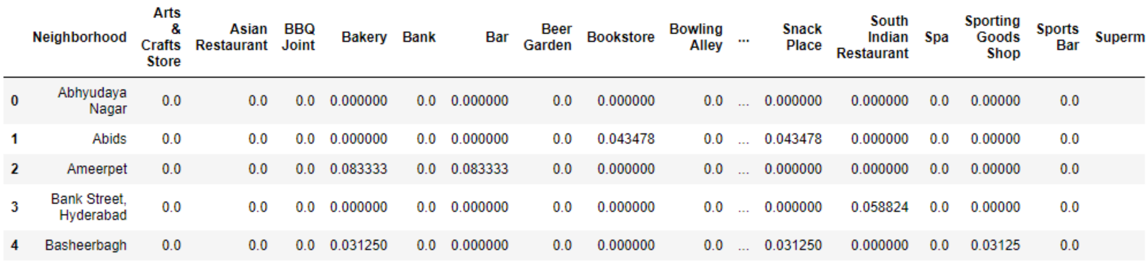


Figure 4 - Filtered Neighbourhoods

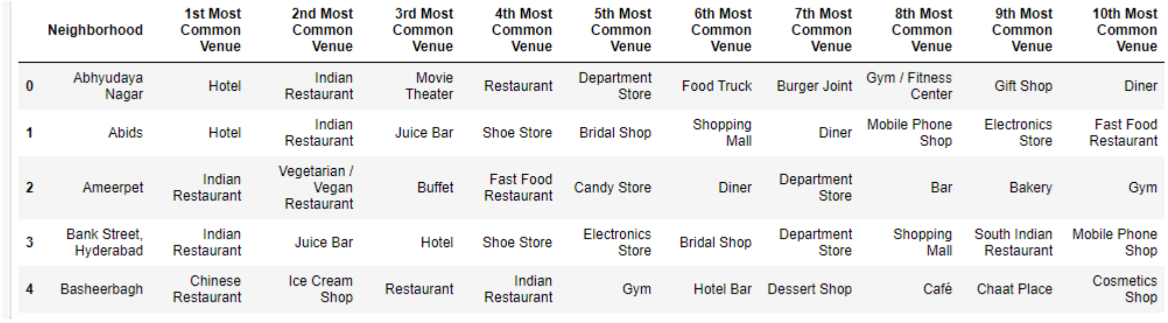
Now, **one hot encoding** is performed on this filtered dataset of venues to obtain the venue categories in each neighbourhood. Then the data is grouped by neighbourhood and average value of the frequency of occurrence of each category is obtained. A sample of this one hot encoded dataset is shown below.



	Neighborhood	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bakery	Bank	Bar	Beer Garden	Bookstore	Bowling Alley	...	Snack Place	South Indian Restaurant	Spa	Sporting Goods Shop	Sports Bar	Superm
0	Abhyudaya Nagar	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.000000	0.0	...	0.000000	0.000000	0.0	0.00000	0.0	
1	Abids	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.043478	0.0	...	0.043478	0.000000	0.0	0.00000	0.0	
2	Ameerpet	0.0	0.0	0.0	0.083333	0.0	0.083333	0.0	0.000000	0.0	...	0.000000	0.000000	0.0	0.00000	0.0	
3	Bank Street, Hyderabad	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.000000	0.0	...	0.000000	0.058824	0.0	0.00000	0.0	
4	Basheerbagh	0.0	0.0	0.0	0.031250	0.0	0.000000	0.0	0.000000	0.0	...	0.031250	0.000000	0.0	0.03125	0.0	

Figure 5 - Average of frequency of each category

Ten most common venues in each neighbourhood is obtained from the above data set. A sample of first five neighbourhoods is shown below.



	Neighborhood	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	Abhyudaya Nagar	Hotel	Indian Restaurant	Movie Theater	Restaurant	Department Store	Food Truck	Burger Joint	Gym / Fitness Center	Gift Shop	Diner
1	Abids	Hotel	Indian Restaurant	Juice Bar	Shoe Store	Bridal Shop	Shopping Mall	Diner	Mobile Phone Shop	Electronics Store	Fast Food Restaurant
2	Ameerpet	Indian Restaurant	Vegetarian / Vegan Restaurant	Buffet	Fast Food Restaurant	Candy Store	Diner	Department Store	Bar	Bakery	Gym
3	Bank Street, Hyderabad	Indian Restaurant	Juice Bar	Hotel	Shoe Store	Electronics Store	Bridal Shop	Department Store	Shopping Mall	South Indian Restaurant	Mobile Phone Shop
4	Basheerbagh	Chinese Restaurant	Ice Cream Shop	Restaurant	Indian Restaurant	Gym	Hotel Bar	Dessert Shop	Café	Chaat Place	Cosmetics Shop

Figure 6 - 10 most common venues in each neighbourhood

K-Means Clustering technique is applied on the above dataset to segregate the neighbourhoods into **k** number of clusters. In order to obtain a good result, the best value of **k** must be selected. Silhouette score is used as performance metric to select the best value of **k**. **k** will take values from 2 to 10. For each value of **k**, **K-Means** clustering is applied on the data set and silhouette scores are calculated. The silhouette scores are plotted against the **k**-values in the below figure.

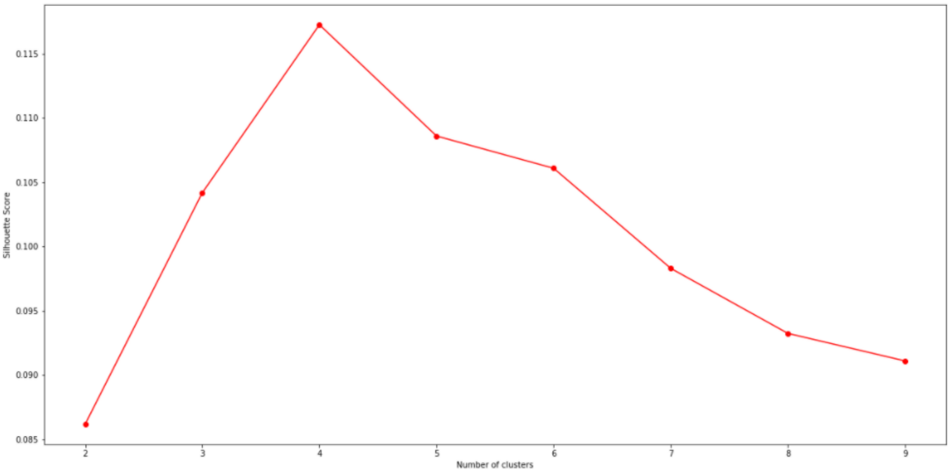


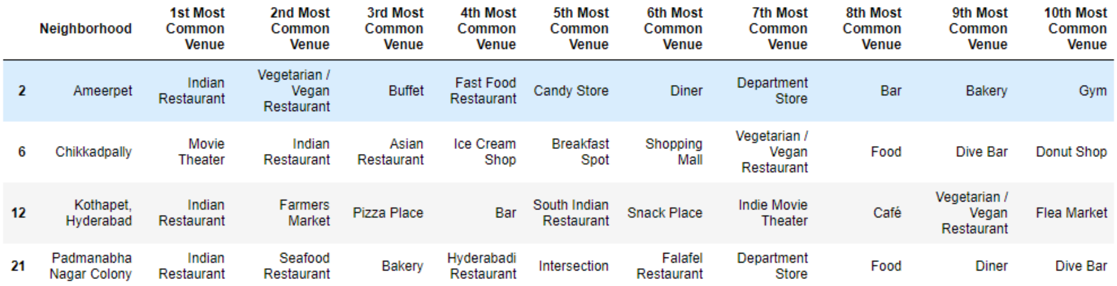
Figure 7 - Silhouette Scores vs. Number of Clusters (k)

From the above figure, the silhouette score is high for **k=4**. So, **K-Means Clustering** will be applied on the data set to segregate the neighbourhood into **four** clusters. The K-Means Labels obtained were included in the dataset for studying the characteristics of each cluster.

Results

Cluster 1

The top venue categories in **Cluster 1** are Indian Restaurant, Vegetarian/Vegan Restaurant, Bar, Dive Bar, Bakery, Diner, Food, Departmental Store, South Indian Restaurant and Movie Theatre.

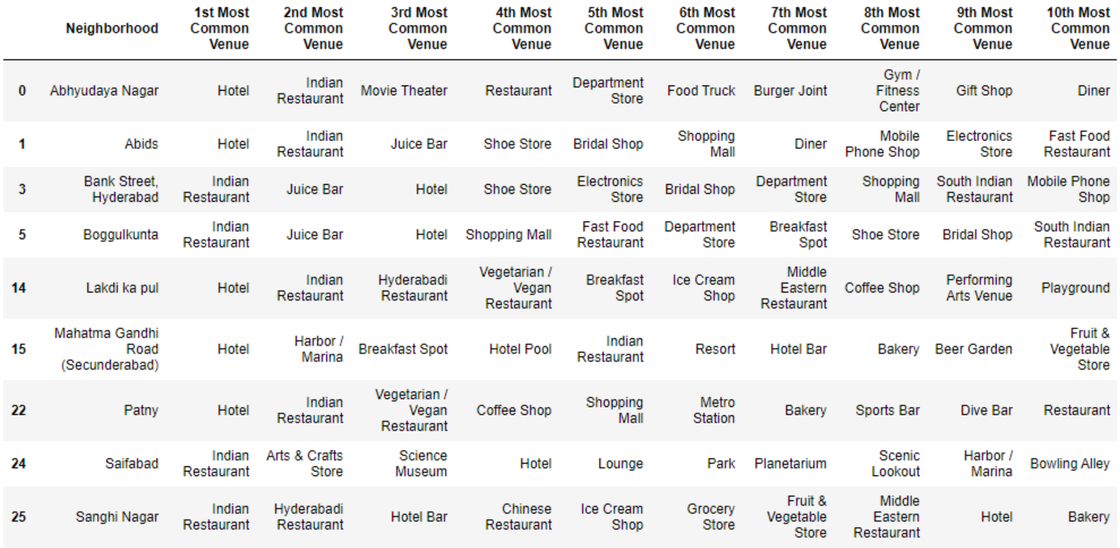


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Ameerpet	Indian Restaurant	Vegetarian / Vegan Restaurant	Buffet	Fast Food Restaurant	Candy Store	Diner	Department Store	Bar	Bakery	Gym
6	Chikkadpally	Movie Theater	Indian Restaurant	Asian Restaurant	Ice Cream Shop	Breakfast Spot	Shopping Mall	Vegetarian / Vegan Restaurant	Food	Dive Bar	Donut Shop
12	Kothapet, Hyderabad	Indian Restaurant	Farmers Market	Pizza Place	Bar	South Indian Restaurant	Snack Place	Indie Movie Theater	Café	Vegetarian / Vegan Restaurant	Flea Market
21	Padmanabha Nagar Colony	Indian Restaurant	Seafood Restaurant	Bakery	Hyderabadi Restaurant	Intersection	Falafel Restaurant	Department Store	Food	Diner	Dive Bar

Figure 8 - Cluster 1

Cluster 2

The top venue categories in **Cluster 2** are Hotel, Indian Restaurant, Shopping Mall, Juice Bar, Departmental Store, Bakery, Shoe Store, Breakfast Spot, Bridal Shop and Electronics Store.

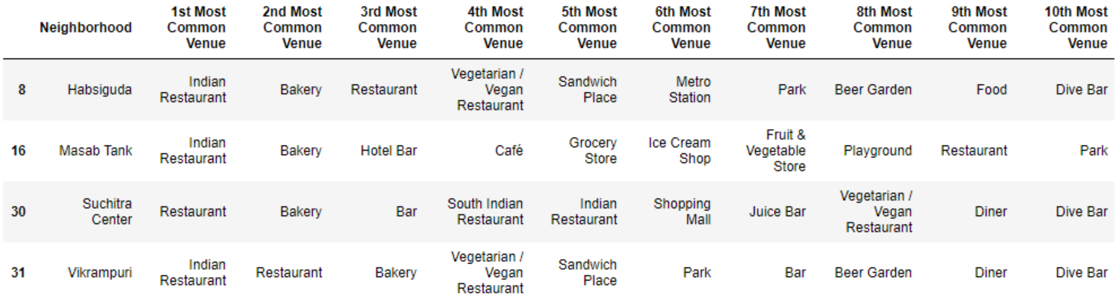


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abhyudaya Nagar	Hotel	Indian Restaurant	Movie Theater	Restaurant	Department Store	Food Truck	Burger Joint	Gym / Fitness Center	Gift Shop	Diner
1	Abids	Hotel	Indian Restaurant	Juice Bar	Shoe Store	Bridal Shop	Shopping Mall	Diner	Mobile Phone Shop	Electronics Store	Fast Food Restaurant
3	Bank Street, Hyderabad	Indian Restaurant	Juice Bar	Hotel	Shoe Store	Electronics Store	Bridal Shop	Department Store	Shopping Mall	South Indian Restaurant	Mobile Phone Shop
5	Boggulkunta	Indian Restaurant	Juice Bar	Hotel	Shopping Mall	Fast Food Restaurant	Department Store	Breakfast Spot	Shoe Store	Bridal Shop	South Indian Restaurant
14	Lakdi ka pul	Hotel	Indian Restaurant	Hyderabadi Restaurant	Vegetarian / Vegan Restaurant	Breakfast Spot	Ice Cream Shop	Middle Eastern Restaurant	Coffee Shop	Performing Arts Venue	Playground
15	Mahatma Gandhi Road (Secunderabad)	Hotel	Harbor / Marina	Breakfast Spot	Hotel Pool	Indian Restaurant	Resort	Hotel Bar	Bakery	Beer Garden	Fruit & Vegetable Store
22	Patny	Hotel	Indian Restaurant	Vegetarian / Vegan Restaurant	Coffee Shop	Shopping Mall	Metro Station	Bakery	Sports Bar	Dive Bar	Restaurant
24	Saifabad	Indian Restaurant	Arts & Crafts Store	Science Museum	Hotel	Lounge	Park	Planetarium	Scenic Lookout	Harbor / Marina	Bowling Alley
25	Sanghi Nagar	Indian Restaurant	Hyderabadi Restaurant	Hotel Bar	Chinese Restaurant	Ice Cream Shop	Grocery Store	Fruit & Vegetable Store	Middle Eastern Restaurant	Hotel	Bakery

Figure 9 - Cluster 2

Cluster 3

The top venue categories in **Cluster 3** are Indian Restaurant, Bakery, Restaurant, Park, Dive Bar, Vegetarian/Vegan Restaurant, Beer Garden, Sandwich Place, Diner and Bar.

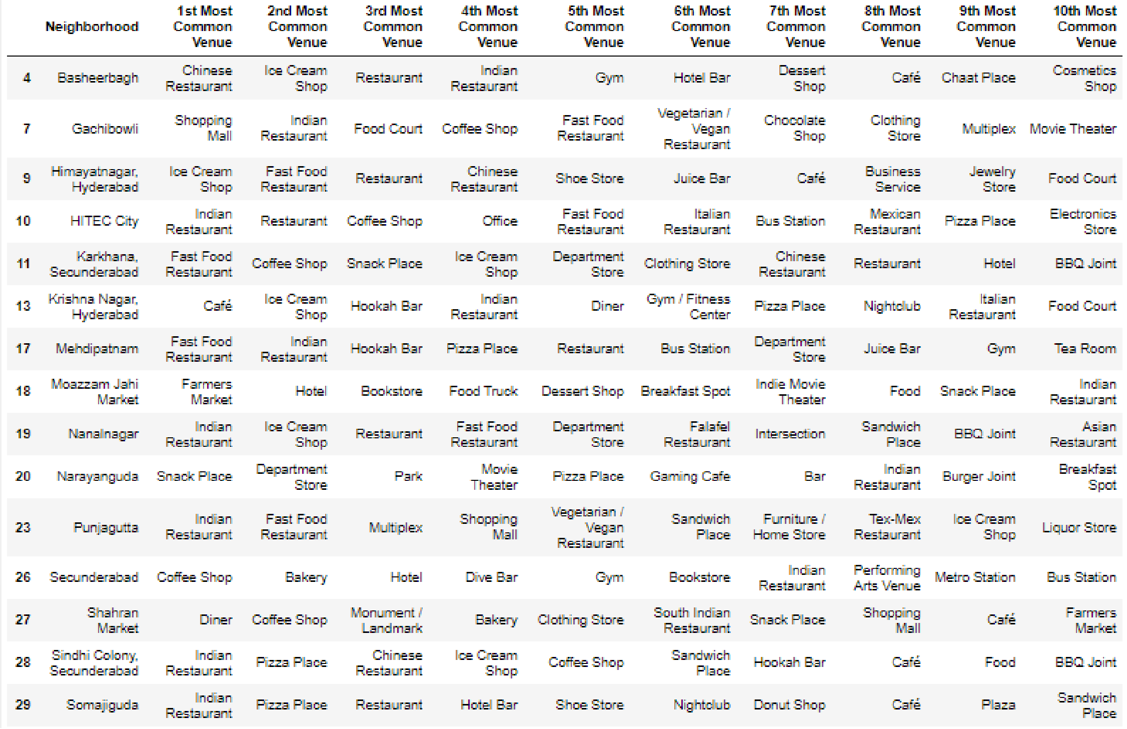


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
8	Habsiguda	Indian Restaurant	Bakery	Restaurant	Vegetarian / Vegan Restaurant	Sandwich Place	Metro Station	Park	Beer Garden	Food	Dive Bar
16	Masab Tank	Indian Restaurant	Bakery	Hotel Bar	Café	Grocery Store	Ice Cream Shop	Fruit & Vegetable Store	Playground	Restaurant	Park
30	Suchitra Center	Restaurant	Bakery	Bar	South Indian Restaurant	Indian Restaurant	Shopping Mall	Juice Bar	Vegetarian / Vegan Restaurant	Diner	Dive Bar
31	Vikrampuri	Indian Restaurant	Restaurant	Bakery	Vegetarian / Vegan Restaurant	Sandwich Place	Park	Bar	Beer Garden	Diner	Dive Bar

Figure 10 - Cluster 3

Cluster 4

The top venue categories in **Cluster 4** are Indian Restaurant, Ice cream shop, Fast Food Restaurant, Restaurant, Pizza Place, Café, Coffee Shop, Department Store, Sandwich Place and Chinese Restaurant.



	Neighborhood	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
4	Basheerbagh	Chinese Restaurant	Ice Cream Shop	Restaurant	Indian Restaurant	Gym	Hotel Bar	Dessert Shop	Café	Chaat Place	Cosmetics Shop
7	Gachibowli	Shopping Mall	Indian Restaurant	Food Court	Coffee Shop	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Chocolate Shop	Clothing Store	Multiplex	Movie Theater
9	Himayatnagar, Hyderabad	Ice Cream Shop	Fast Food Restaurant	Restaurant	Chinese Restaurant	Shoe Store	Juice Bar	Café	Business Service	Jewelry Store	Food Court
10	HITEC City	Indian Restaurant	Restaurant	Coffee Shop	Office	Fast Food Restaurant	Italian Restaurant	Bus Station	Mexican Restaurant	Pizza Place	Electronics Store
11	Karkhana, Secunderabad	Fast Food Restaurant	Coffee Shop	Snack Place	Ice Cream Shop	Department Store	Clothing Store	Chinese Restaurant	Restaurant	Hotel	BBQ Joint
13	Krishna Nagar, Hyderabad	Café	Ice Cream Shop	Hookah Bar	Indian Restaurant	Diner	Gym / Fitness Center	Pizza Place	Nightclub	Italian Restaurant	Food Court
17	Mehdipatnam	Fast Food Restaurant	Indian Restaurant	Hookah Bar	Pizza Place	Restaurant	Bus Station	Department Store	Juice Bar	Gym	Tea Room
18	Moazzam Jahi Market	Farmers Market	Hotel	Bookstore	Food Truck	Dessert Shop	Breakfast Spot	Indie Movie Theater	Food	Snack Place	Indian Restaurant
19	Nanalnagar	Indian Restaurant	Ice Cream Shop	Restaurant	Fast Food Restaurant	Department Store	Falafel Restaurant	Intersection	Sandwich Place	BBQ Joint	Asian Restaurant
20	Narayanguda	Snack Place	Department Store	Park	Movie Theater	Pizza Place	Gaming Cafe	Bar	Indian Restaurant	Burger Joint	Breakfast Spot
23	Punjagutta	Indian Restaurant	Fast Food Restaurant	Multiplex	Shopping Mall	Vegetarian / Vegan Restaurant	Sandwich Place	Furniture / Home Store	Tex-Mex Restaurant	Ice Cream Shop	Liquor Store
26	Secunderabad	Coffee Shop	Bakery	Hotel	Dive Bar	Gym	Bookstore	Indian Restaurant	Performing Arts Venue	Metro Station	Bus Station
27	Shahran Market	Diner	Coffee Shop	Monument / Landmark	Bakery	Clothing Store	South Indian Restaurant	Snack Place	Shopping Mall	Café	Farmers Market
28	Sindhi Colony, Secunderabad	Indian Restaurant	Pizza Place	Chinese Restaurant	Ice Cream Shop	Coffee Shop	Sandwich Place	Hookah Bar	Café	Food	BBQ Joint
29	Somajiguda	Indian Restaurant	Pizza Place	Restaurant	Hotel Bar	Shoe Store	Nightclub	Donut Shop	Café	Plaza	Sandwich Place

Figure 11 - Cluster 4

Discussion

The visualization of the top categories in each cluster are shown below for comparison against one another.

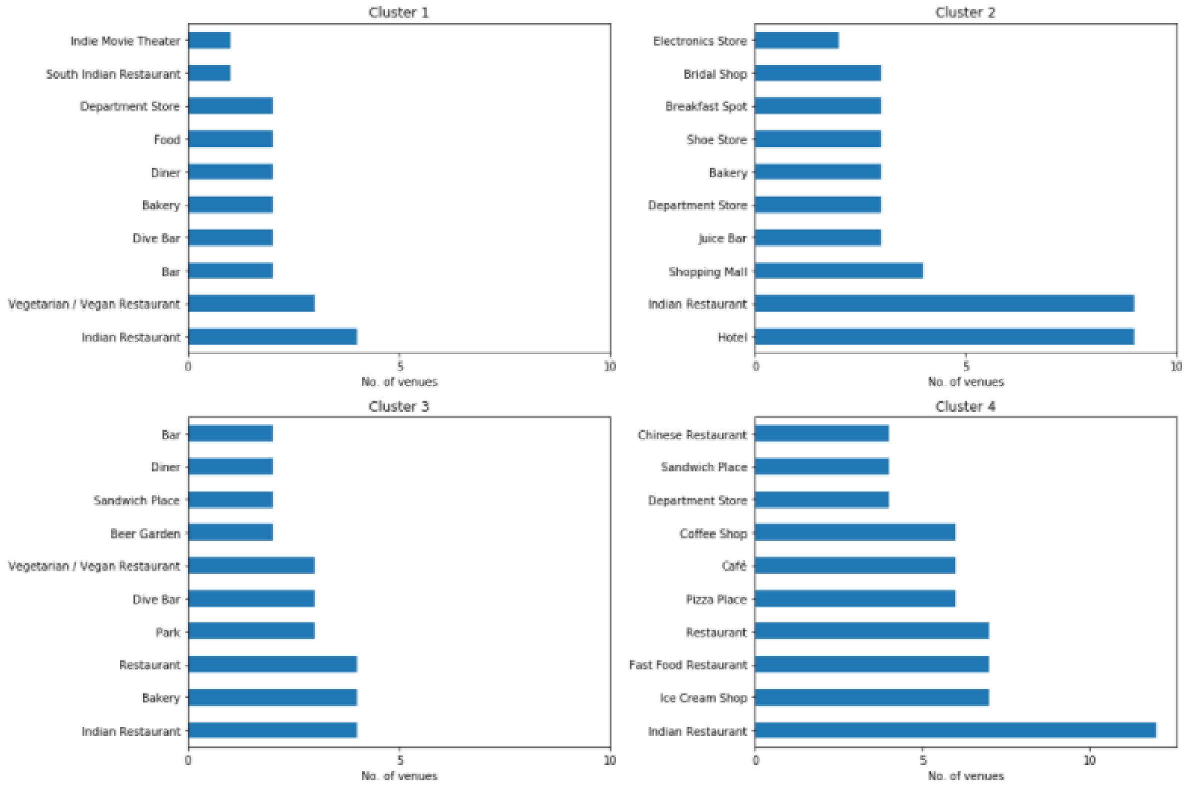


Figure 12 - Visualization of top 10 venue categories of each cluster

The above plots show some interesting insights which will be helpful for investors/developers and small business owners. It will help them to select an optimal location to set up their business. Following are few examples.

1. **Hotel**

   From the above plot, cluster-2 had more hotels. So, the neighbourhoods in cluster-2 are not the best locations to set up a hotel business. Cluster-1 and Cluster-3 have small number of restaurants. Whereas in Cluster-4, there are enough number of restaurants and coffee shops. The optimal places to set up a hotel is where there is less competition as well as restaurants nearby. So, neighbourhoods in Cluster-4 like HITEC City, Gachibowli, Panjagutta are the best options to open a hotel.

2. **Departmental Store or Convenient Store:**

   Based on the above plots all clusters except Cluster-3 are having departmental stores. So, if a small business owner wants to set up a departmental store, the neighbourhoods in Cluster-3, Habsiguda, Masab Tank, Suchitra and Vikrampuri will be good choice.

3. **Shopping Mall:**

   Cluster-2 have 4 shopping malls, whereas the remaining clusters do not have any shopping mall as the most common venue category. So, constructing a shopping mall would not be a best choice and neighbourhoods in Cluster-2 can be filtered out when selecting location to set up a shopping mall

4. **Restaurant:**

   Cluster-1 and Cluster-3 are having very few restaurants. So, if a business owner wants to open a restaurant, Cluster-1 and Cluster-3 would be good options.

5. **Coffee Shop:**

   Cluster-4 is having many coffee shops. So, if a business owner wants to open a coffee shop, Cluster-4 will not be a good option. Cluster-2 is having more shopping places and few restaurants. So, setting up a coffee shop in Cluster-2 would be best selection to open a coffee shop.

Below map along with the data obtained in the Results & Discussion section will be helpful for investors/developers and small business owners in selecting suitable location based on their business category.

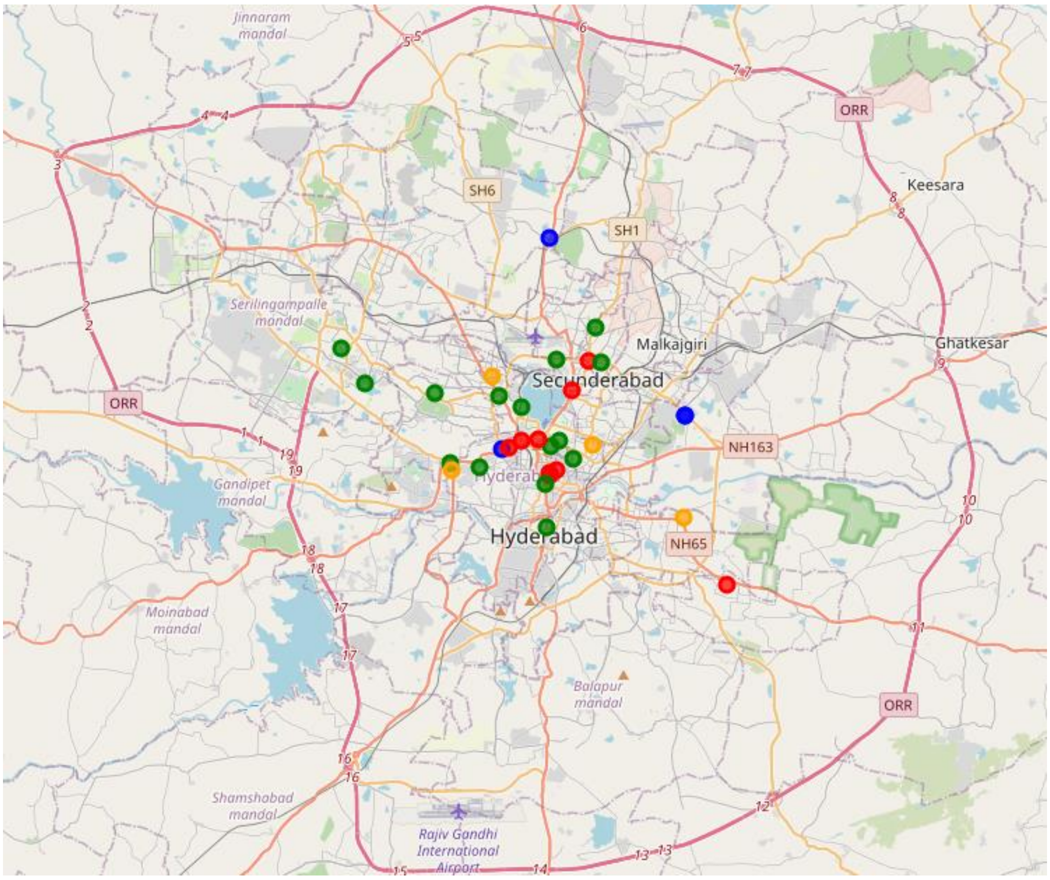


Figure 13 - Map of Clustered Neighbourhoods in Hyderabad

Conclusion

The outcomes of this project can be used as tool to select the optimal location for various business. The outcomes of this project combined with other data like population expenditure, demographics, income levels, and other location data such as upcoming big real estate projects or office complexes, proximity to nearby bus or metro stations, shopping malls etc., will help business owners to select a suitable location to open their business.

One of the main drawbacks is this project is few numbers of venues returned by the FourSquare API. The API had returned only 1170 venues which is small for a big city like Hyderabad. For some neighbourhoods, it returned venues as few as one. As the number of venues is less, the results obtained in this project might have been skewed. The same project, if done using Google API which had more places listed, may give good results. As a part of future extension of this project, the same methodology could be applied on better data sources to obtain the optimal locations.

References

1. Wikipedia page of Neighbourhoods in Hyderabad https://en.wikipedia.org/w/index.php?title=Category:Neighbourhoods_in_Hyderabad,_India&pageuntil=Sikh+Village#mw-pages
2. Documentation for FourSquare developers. https://developer.foursquare.com/docs/