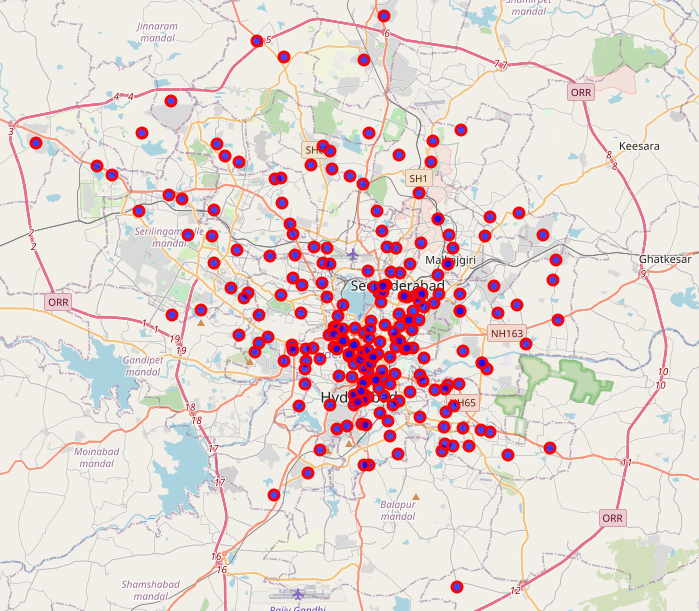
**Applied Data Science Capstone Project by IBM on Coursera**

**Clustering of Neighbourhoods – Hyderabad, India**



**Sujan Kumar Kummara**

Contents

[List of Figures 2](#_Toc56890637)

[Introduction 3](#_Toc56890638)

[Business Problem 3](#_Toc56890639)

[Data Sources 3](#_Toc56890640)

[Methodology 4](#_Toc56890641)

[Analysis 5](#_Toc56890642)

[Results 7](#_Toc56890643)

[Discussion 9](#_Toc56890644)

[Conclusion 11](#_Toc56890645)

[References 11](#_Toc56890646)

# List of Figures

[Figure 1 - Neighbourhoods of Hyderabad Dataset 3](#_Toc56890400)

[Figure 2 - Neighbourhoods with latitude and longitude coordinates 4](#_Toc56890401)

[Figure 3 - Venues in Hyderabad Dataset 4](#_Toc56890402)

[Figure 4 - Filtered Neighbourhoods 5](#_Toc56890403)

[Figure 5 - Average of frequency of each category 5](#_Toc56890404)

[Figure 6 - 10 most common venues in each neighbourhood 6](#_Toc56890405)

[Figure 7 - Silhouette Scores vs. Number of Clusters (k) 6](#_Toc56890406)

[Figure 8 - Cluster 1 7](#_Toc56890407)

[Figure 9 - Cluster 2 7](#_Toc56890408)

[Figure 10 - Cluster 3 8](#_Toc56890409)

[Figure 11 - Cluster 4 8](#_Toc56890410)

[Figure 12 - Visualization of top 10 venue categories of each cluster 9](#_Toc56890411)

[Figure 13 - Map of Clustered Neighbourhoods in Hyderabad 10](#_Toc56890412)

# 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**’

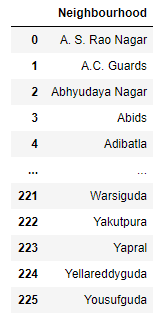


Figure - Neighbourhoods of Hyderabad Dataset

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

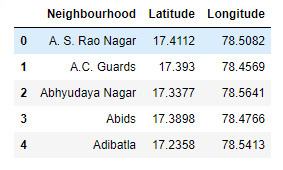


Figure - Neighbourhoods with latitude and longitude coordinates

1. **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.

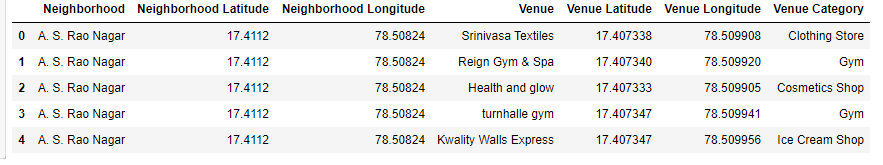


Figure - Venues in Hyderabad Dataset

# 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.

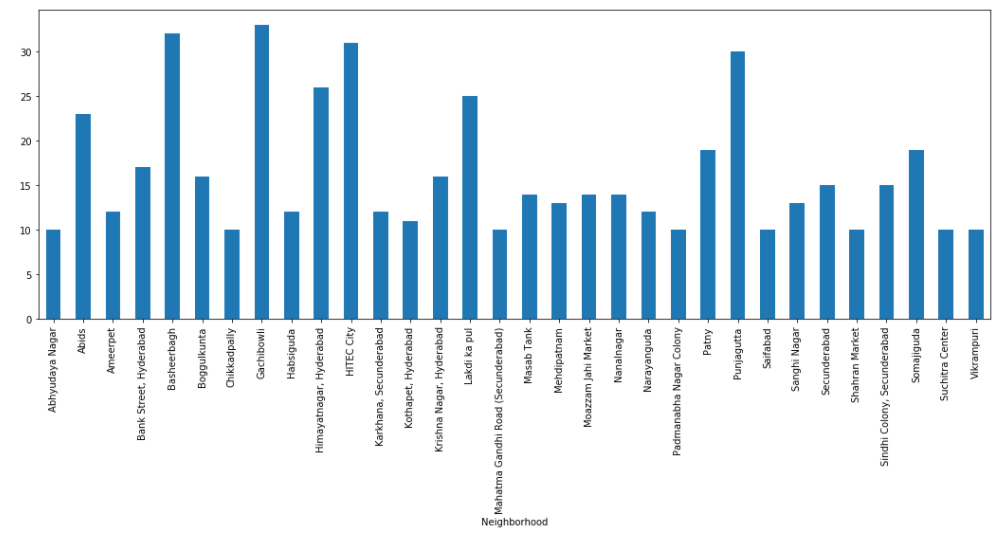


Figure - Filtered Neighbourhoods

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.

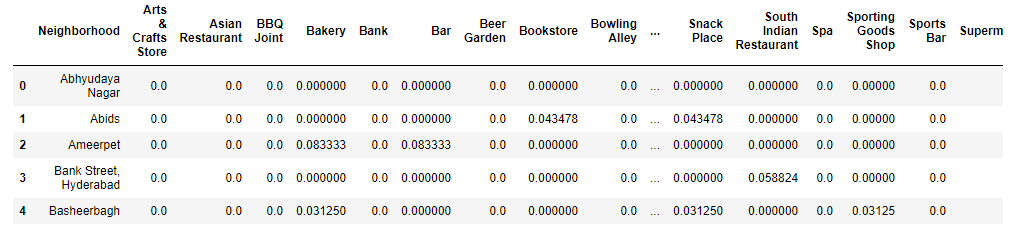


Figure - Average of frequency of each category

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



Figure - 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.

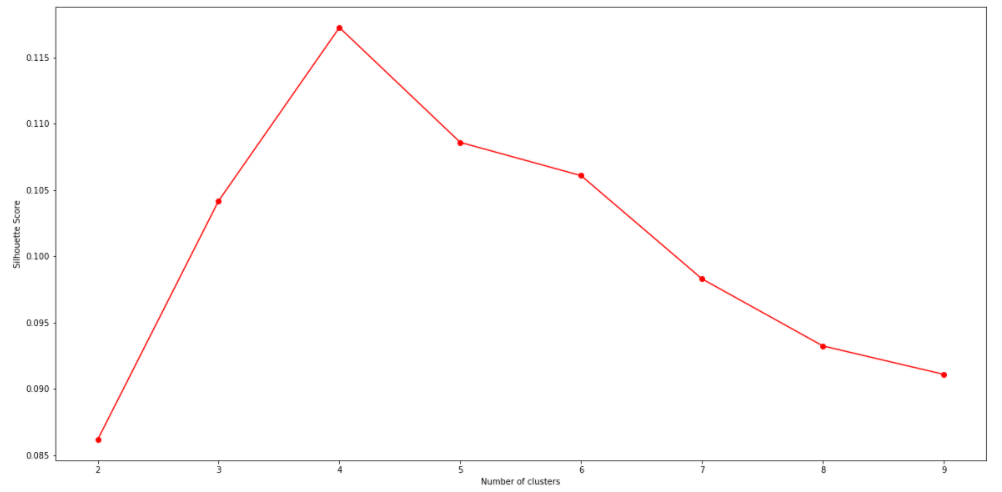


Figure - Silhouette Scores vs. Number of Clusters (k)

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.

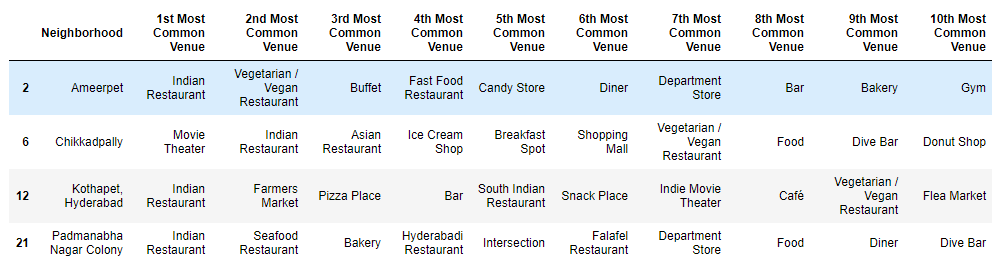


Figure - Cluster 1

**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.

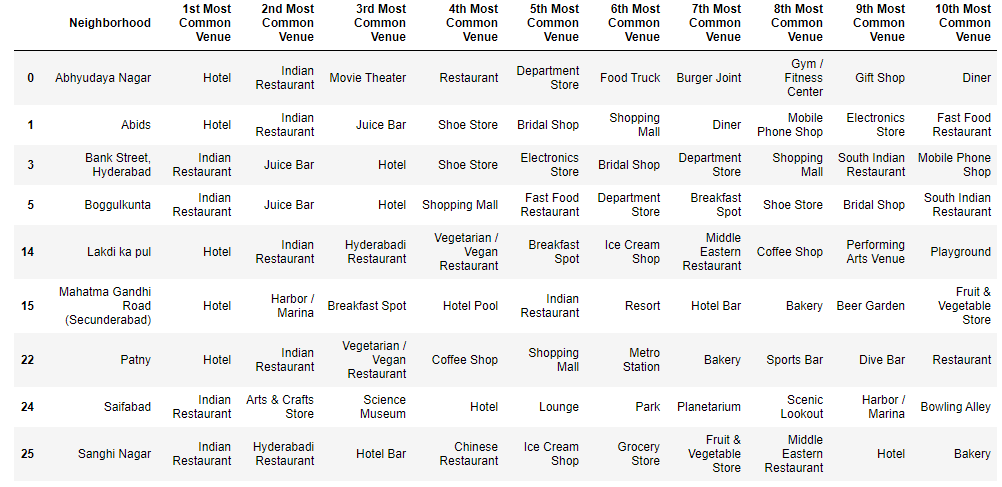


Figure - Cluster 2

**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.

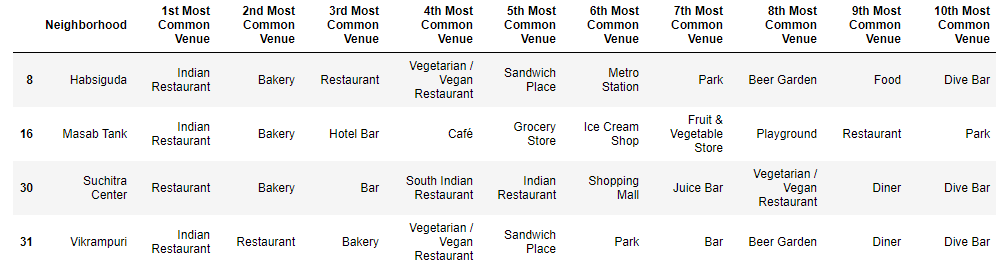


Figure - Cluster 3

**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.

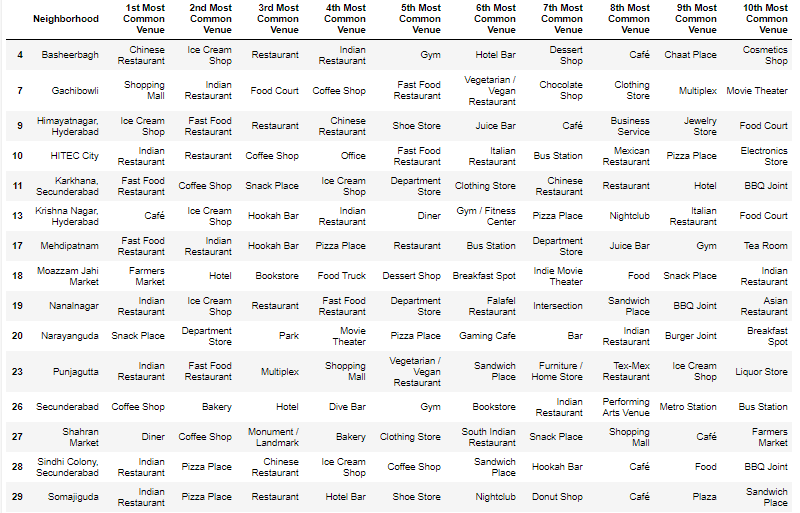


Figure - Cluster 4

# Discussion

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

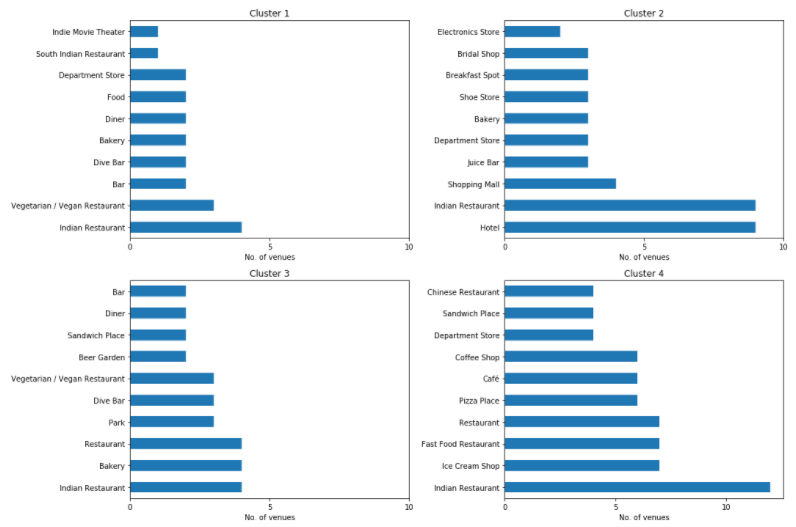


Figure - Visualization of top 10 venue categories of each cluster

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.

1. **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.

1. **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

1. **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.

1. **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.

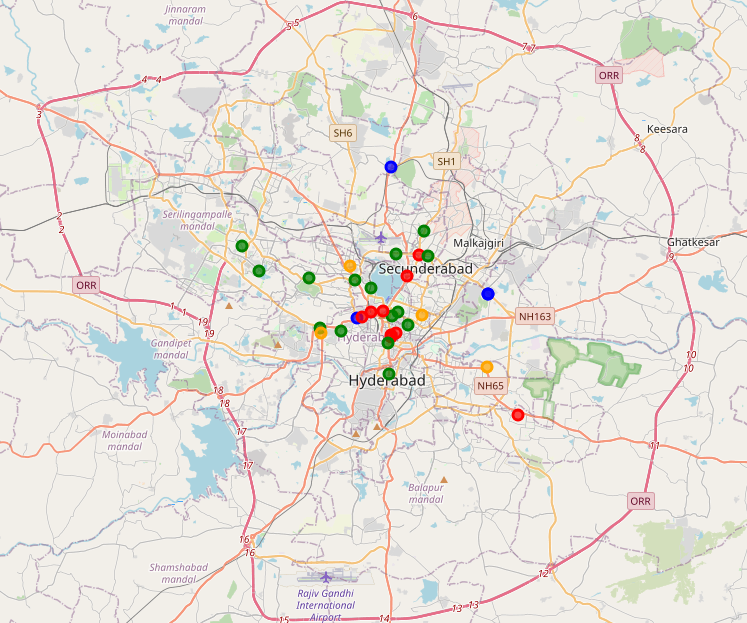


Figure - Map of Clustered Neighbourhoods in Hyderabad

# 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/>