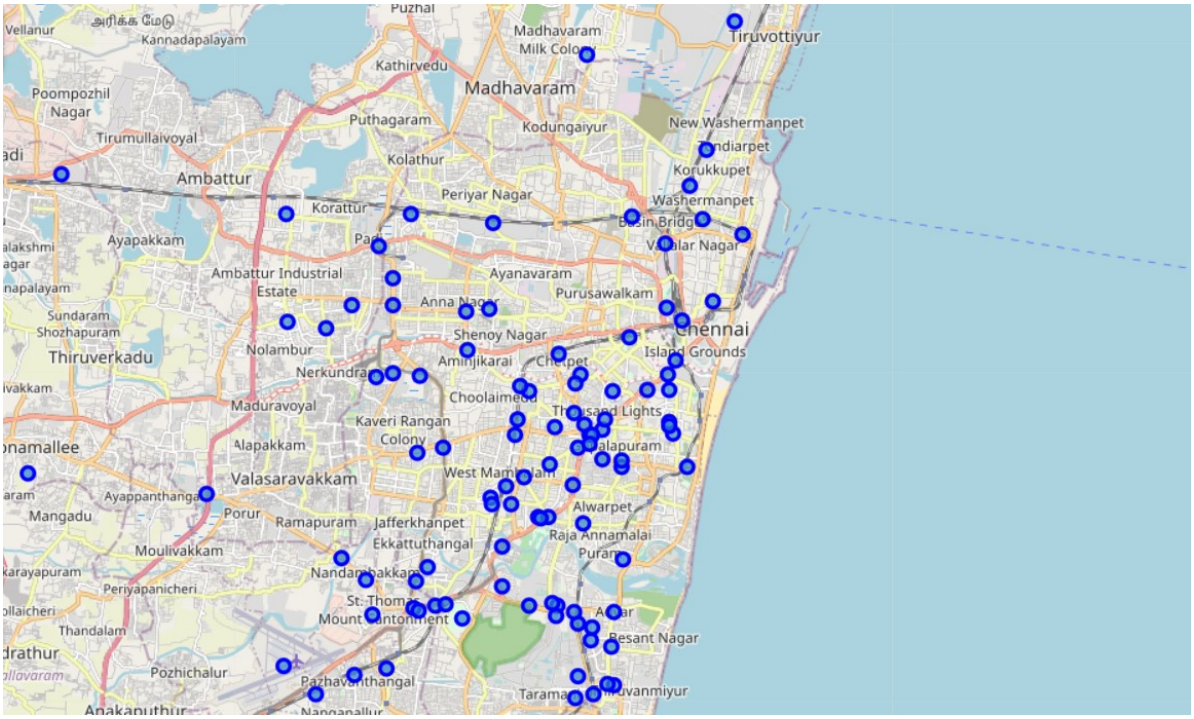
The Battle of Neighborhoods - Chennai



Applied Data Science Capstone by IBM on Coursera

(ROHITH S P)

1. INTRODUCTION: BUSINESS PROBLEM

This project deals with the major venue categories in the neighborhoods of **Chennai, The Detroit of India**. This project would specifically help Business personal plan to start new Restaurants, Hotels, etc. in Chennai, Tamil Nadu, India.

The **Foursquare API** is used to access the venues in the neighborhoods. Since, it returns less venues in the neighborhoods, we would be analyzing areas for which countable number of venues are obtained. Then they are clustered based on their venues using Data Science Techniques. Here the **k-means clustering algorithm** is used to achieve the task. The optimal number of clusters can be obtained using **silhouette score** metrics.

Folium visualization library can be used to visualize the clusters superimposed on the map of Chennai city. These clusters can be analyzed to help small scale business owners select a suitable location for their need such as Hotels, Shopping Malls, Restaurants or even specifically Indian restaurants or Coffee shops.

The major **Target Audience** would be small-scale business owners and stake holders planning to start their business at a location in Chennai. This project would help them find the optimal location based on the category of their business such as,

- What is the best location to start a new hotel in Chennai with restaurants around?
- Which area is best suitable for opening a Shopping Mall in Chennai?

2. DATA REQUIREMENTS

Chennai has multiple neighborhoods. The chennaiiq.com website has a dataset which has the list of locations in Chennai along with their Latitude and Longitude in Degrees Minute Seconds format. There is a total of 105 neighborhoods as shown in Fig 2.1.

1. https://chennaiiq.com/chennai/latitude_longitude_areas.asp

```
chennai_data.head()
```

```
(105, 3)
```

	Neighborhood	Latitude	Longitude
0	Adyar Bus Debot	12°59'50" N	80°15'25" E
1	Adyar Signal	13°00'23" N	80°15'27" E
2	Alandur	13°00'28" N	80°12'35" E
3	Ambattur	13°06'36" N	80°10'12" E
4	Anna Arch	13°04'28" N	80°13'06" E

Fig 2.1 Chennai Neighborhoods Dataset

But the Latitude and Longitude data obtained are in Degrees Minute Seconds format which needs to be converted to Decimal Degrees Format as shown in Fig. 2.2.

```
chennai_data.head()
```

```
(105, 3)
```

	Neighborhood	Latitude	Longitude
0	Adyar Bus Debot	12.997222	80.256944
1	Adyar Signal	13.006389	80.257500
2	Alandur	13.007778	80.209722
3	Ambattur	13.110000	80.170000
4	Anna Arch	13.074444	80.218333

Fig 2.2 Chennai Neighborhoods Dataset with Location Data in Decimal Degrees Format

Next the details of venues in each neighborhood namely **Venue, Venue Latitude, Venue Longitude, Venue Category** data needs to be obtained. Here, Foursquare API is used to obtain this data.

2. https://foursquare.com/

A total of 1130 venues data have been obtained from Foursquare. The resultant venues dataset, (shown in Fig 2.3) is used for the analysis process.

```
print(chennai_venues.shape)
chennai_venues.head()
```

```
(1130, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adyar Bus Debot	12.997222	80.256944	Zaitoon Restaurant	12.996861	80.256178	Middle Eastern Restaurant
1	Adyar Bus Debot	12.997222	80.256944	Kuttanadu Restaurant	12.997010	80.257799	Asian Restaurant
2	Adyar Bus Debot	12.997222	80.256944	Zha Cafe	12.999730	80.254806	Café
3	Adyar Bus Debot	12.997222	80.256944	Adyar Ananda Bhavan, Besant Nagar	12.996678	80.258275	Fast Food Restaurant
4	Adyar Bus Debot	12.997222	80.256944	Kovai Pazhamudir Nilayam	12.996522	80.259776	Fruit & Vegetable Store

A total of 1130 venues were obtained. Now lets check the number of venues returned per neighbourhood.

Fig 2.3 Chennai Venues Dataset

3. METHODOLOGY

Now, we have the neighborhoods data of Chennai (**105 neighborhoods**). We also have the most popular venues in each neighborhood obtained using Foursquare API. A total of **1130 venues** have been obtained in the whole city and **145 unique categories**. But as seen we have multiple neighborhoods with less than 10 venues returned. In order to create a good analysis let's consider only the ***neighborhoods with more than 10 venues***.

We can perform **one hot encoding** on the obtained data set and use it find the 10 most common venue category in each neighborhood. Then clustering can be performed on the dataset. Here **K - Nearest Neighbor** clustering technique have been used. To find the optimal number of clusters **silhouette score** metric technique is used.

The clusters obtained can be analyzed to find the major type of venue categories in each cluster. This data can be used to suggest business people, suitable locations based on the category.

4. ANALYSIS

Looking into the dataset we found that there were many neighborhoods with less than 10 venues which can be remove before performing the analysis to obtain better results. The following plot shows only the neighborhoods from which 10 or more than 10 venues were obtained. The resultant dataset consists of 37 neighborhoods as shown in Fig 4.1.

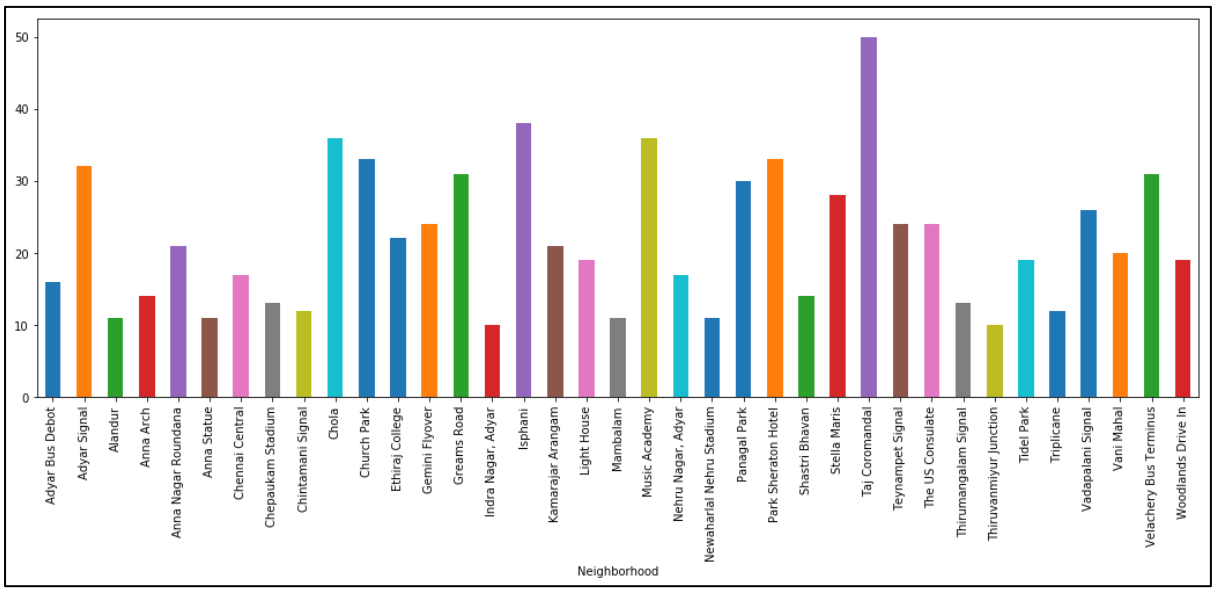


Fig 4.1 Filtered Neighborhood Dataset

Next, we will perform **one hot encoding** on the filtered data to obtain the venue categories in each neighborhood. Then group the data by neighborhood and take the mean value of the frequency of occurrence of each category. A sample output is shown in Fig 4.2.

```
chennai_grouped.head()
```

```
(37, 114)
```

	Neighborhood	Accessories Store	African Restaurant	Airport	American Restaurant	Amphitheater	Arcade	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery
0	Adyar Bus Debot	0.0	0.0	0.000000	0.0	0.0	0.00000	0.0	0.125000	0.0	0.0625	0.00000
1	Adyar Signal	0.0	0.0	0.000000	0.0	0.0	0.03125	0.0	0.031250	0.0	0.0000	0.03125
2	Alandur	0.0	0.0	0.090909	0.0	0.0	0.00000	0.0	0.000000	0.0	0.0000	0.00000
3	Anna Arch	0.0	0.0	0.000000	0.0	0.0	0.00000	0.0	0.000000	0.0	0.0000	0.00000
4	Anna Nagar Roundana	0.0	0.0	0.000000	0.0	0.0	0.00000	0.0	0.047619	0.0	0.0000	0.00000

Fig 4.2 Mean of frequency of occurrence of each category

The above dataset is used to obtain the top 10 most common venues in each neighborhood i.e. the 10 venues with the highest mean of frequency of occurrence. A sample for the first 5 neighborhoods is shown in Fig 4.3.

	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	Adyar Bus Debot	Indian Restaurant	Fast Food Restaurant	Asian Restaurant	Pizza Place	Sandwich Place	Breakfast Spot	Fruit & Vegetable Store	Restaurant	BBQ Joint	Middle Eastern Restaurant
1	Adyar Signal	Indian Restaurant	Electronics Store	North Indian Restaurant	Coffee Shop	Rock Club	Dessert Shop	Bookstore	Lounge	Café	Shoe Store
2	Alandur	Indian Restaurant	South Indian Restaurant	Hotel	Bus Station	Bus Line	Bar	Metro Station	Airport	Gym	Grocery Store
3	Anna Arch	Fast Food Restaurant	Clothing Store	Electronics Store	Mediterranean Restaurant	Café	Multiplex	Pub	Bookstore	Scenic Lookout	Shopping Mall
4	Anna Nagar Roundana	Indian Restaurant	Chinese Restaurant	South Indian Restaurant	Clothing Store	Paper / Office Supplies Store	Café	Electronics Store	Fast Food Restaurant	Middle Eastern Restaurant	Bookstore

Fig 4.3 Ten Most Common Venues in each Neighborhood

This dataset can be used for the clustering algorithm. Here, the K-Nearest Neighbor (KNN) clustering algorithm is used. It is an unsupervised machine learning technique that clusters the given data into K number of clusters. For optimal result we need to select the best value for K. Here, the silhouette score is used to find the best value for K. A range of values from 2 to 10 was considered, KNN clustering was performed on the dataset and the silhouette score was calculated and plotted on a line plot as shown in Fig 4.4. From the plot we can see that a K value of 8 provides the best score. This K value is used for the K-Means Clustering Technique.

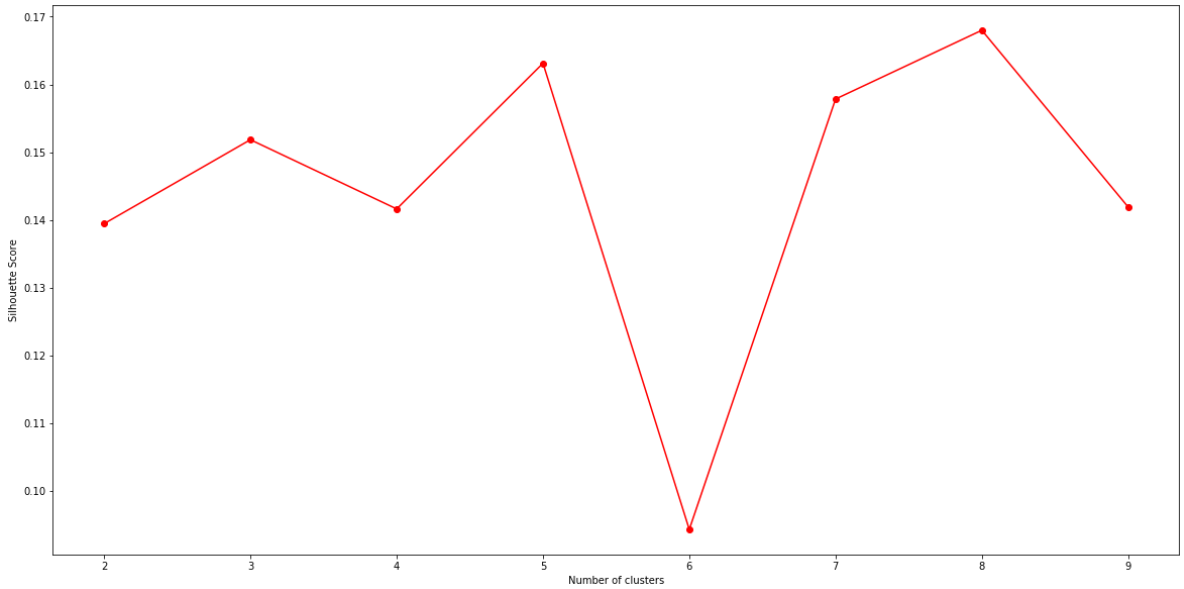


Fig 4.4 Silhouette Score for different Number of Clusters

The K-Means labels obtained were included in the top neighborhoods dataset for examining the characteristics of each cluster.

5. RESULTS

Let's examine the 8 clusters and find the discriminating venue categories that distinguish each cluster. For this purpose, lets also look into the five most common venue category in each cluster.

5.1. Cluster 1

The top venue categories in Cluster 1 are Indian Restaurant, Multiplex, Gym, Chinese Restaurant and Pizza Place.

	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
30	Thiruvanmiyur Junction	Indian Restaurant	Multiplex	Gym	Chinese Restaurant	Pizza Place	Clothing Store	Hotel	Hotel Bar	Hookah Bar	Donut Shop

5.2. Cluster 2

The top venue categories in Cluster 2 are Indian Restaurant, Hotel, Café, Chinese Restaurant and Juice 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
2	Alandur	Indian Restaurant	South Indian Restaurant	Hotel	Bus Station	Bus Line	Bar	Metro Station	Airport	Gym	Grocery Store
9	Chola	Indian Restaurant	Hotel	Ice Cream Shop	Concert Hall	Restaurant	Mexican Restaurant	Chinese Restaurant	Comfort Food Restaurant	Dessert Shop	Electronics Store
10	Church Park	Indian Restaurant	Multiplex	Café	Juice Bar	Middle Eastern Restaurant	Movie Theater	Chinese Restaurant	Bakery	Bengali Restaurant	Italian Restaurant
11	Ethiraj College	Hotel	Pizza Place	Juice Bar	Café	Kebab Restaurant	Indian Restaurant	Asian Restaurant	Athletics & Sports	Mexican Restaurant	Korean Restaurant
13	Greams Road	Multiplex	Indian Restaurant	Café	Middle Eastern Restaurant	Bakery	Movie Theater	Juice Bar	Chinese Restaurant	Pub	Buffet
19	Music Academy	Indian Restaurant	Hotel	Restaurant	Café	Concert Hall	Electronics Store	Chinese Restaurant	Comfort Food Restaurant	Dessert Shop	Lounge
25	Stella Maris	Indian Restaurant	Hotel	Bar	Ice Cream Shop	Italian Restaurant	Vietnamese Restaurant	Juice Bar	Garden	Kerala Restaurant	Mexican Restaurant
27	Teynampet Signal	Indian Restaurant	Hotel	Lounge	Italian Restaurant	Pub	Pizza Place	Chinese Restaurant	Diner	Mediterranean Restaurant	Café
33	Vadapalani Signal	Multiplex	Clothing Store	Indian Restaurant	Fast Food Restaurant	South Indian Restaurant	Asian Restaurant	Hotel	Shopping Mall	Movie Theater	Café

5.3. Cluster 3

The top venue categories in Cluster 3 are Indian Restaurant, Café, Chinese Restaurant, Fast Food Restaurant and Asian 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
0	Adyar Bus Debot	Indian Restaurant	Fast Food Restaurant	Asian Restaurant	Pizza Place	Sandwich Place	Breakfast Spot	Fruit & Vegetable Store	Restaurant	BBQ Joint	Middle Eastern Restaurant
1	Adyar Signal	Indian Restaurant	Electronics Store	North Indian Restaurant	Coffee Shop	Rock Club	Dessert Shop	Bookstore	Lounge	Café	Shoe Store
4	Anna Nagar Roundana	Indian Restaurant	Chinese Restaurant	South Indian Restaurant	Clothing Store	Paper / Office Supplies Store	Café	Electronics Store	Fast Food Restaurant	Middle Eastern Restaurant	Bookstore
8	Chintamani Signal	Indian Restaurant	Restaurant	Bakery	Café	Dessert Shop	Electronics Store	Middle Eastern Restaurant	Coffee Shop	Hookah Bar	Comfort Food Restaurant
24	Shastri Bhavan	Indian Restaurant	Chinese Restaurant	Japanese Restaurant	Theater	Convenience Store	Asian Restaurant	Multicuisine Indian Restaurant	Coffee Shop	Food	Concert Hall
26	Taj Coromandal	Indian Restaurant	Café	Chinese Restaurant	Sandwich Place	Italian Restaurant	Ice Cream Shop	Clothing Store	Asian Restaurant	Dessert Shop	Fast Food Restaurant
29	Thirumangalam Signal	Indian Restaurant	Bus Station	Smoke Shop	Vegetarian / Vegan Restaurant	Café	Tennis Court	Mobile Phone Shop	Jewelry Store	Pizza Place	Market
31	Tidel Park	Food Court	Café	Sandwich Place	Fast Food Restaurant	Office	Vegetarian / Vegan Restaurant	Chinese Restaurant	Asian Restaurant	Indian Restaurant	Platform
35	Velachery Bus Terminus	Indian Restaurant	Fast Food Restaurant	Chinese Restaurant	Restaurant	Accessories Store	Bar	Juice Bar	Kerala Restaurant	Dessert Shop	Multiplex

5.4. Cluster 4

The top venue categories in Cluster 4 are General Entertainment, Electronics Store, Multiplex, Indian Restaurant and Comfort Food 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
5	Anna Statue	Indian Restaurant	Multiplex	Dessert Shop	General Entertainment	Electronics Store	Flea Market	Women's Store	Food	Comfort Food Restaurant	Concert Hall
7	Chepaukam Stadium	Indian Restaurant	Bookstore	Breakfast Spot	Bar	Electronics Store	Mediterranean Restaurant	Multiplex	General Entertainment	Café	Hotel
32	Triplicane	Indian Restaurant	Dessert Shop	Multiplex	Hotel	General Entertainment	Electronics Store	Women's Store	Flea Market	Comfort Food Restaurant	Concert Hall

5.5. Cluster 5

The top venue categories in Cluster 5 are Multiplex, Shopping Mall, Fast Food Restaurant, Scenic Lookout and Business Service.

	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
3	Anna Arch	Fast Food Restaurant	Clothing Store	Electronics Store	Mediterranean Restaurant	Café	Multiplex	Pub	Bookstore	Scenic Lookout	Shopping Mall
17	Light House	Fast Food Restaurant	Coffee Shop	Snack Place	Multiplex	Business Service	Department Store	Sandwich Place	Beach	Bar	Shopping Mall

5.6. Cluster 6

The top venue categories in Cluster 6 are Jewellery Store, Miscellaneous Shop, Indian Restaurant, Concert Hall and Clothing 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
18	Mambalam	Clothing Store	Asian Restaurant	Boutique	Miscellaneous Shop	South Indian Restaurant	Indian Restaurant	Jewelry Store	Ice Cream Shop	Comfort Food Restaurant	Concert Hall
22	Panagal Park	Clothing Store	Indian Restaurant	Jewelry Store	Women's Store	Shopping Mall	Fast Food Restaurant	Miscellaneous Shop	Dessert Shop	Concert Hall	Coffee Shop

5.7. Cluster 7

The top venue categories in Cluster 7 are Chinese Restaurant, Café, Women's Store, Sandwich Place, Kids Store and Department 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
14	Indra Nagar, Adyar	Café	Women's Store	Sandwich Place	Juice Bar	Kids Store	Department Store	Chinese Restaurant	Breakfast Spot	Pizza Place	Athletics & Sports
20	Nehru Nagar, Adyar	Café	Pizza Place	Department Store	Indian Restaurant	Chinese Restaurant	Sandwich Place	Juice Bar	Kids Store	Women's Store	Ice Cream Shop

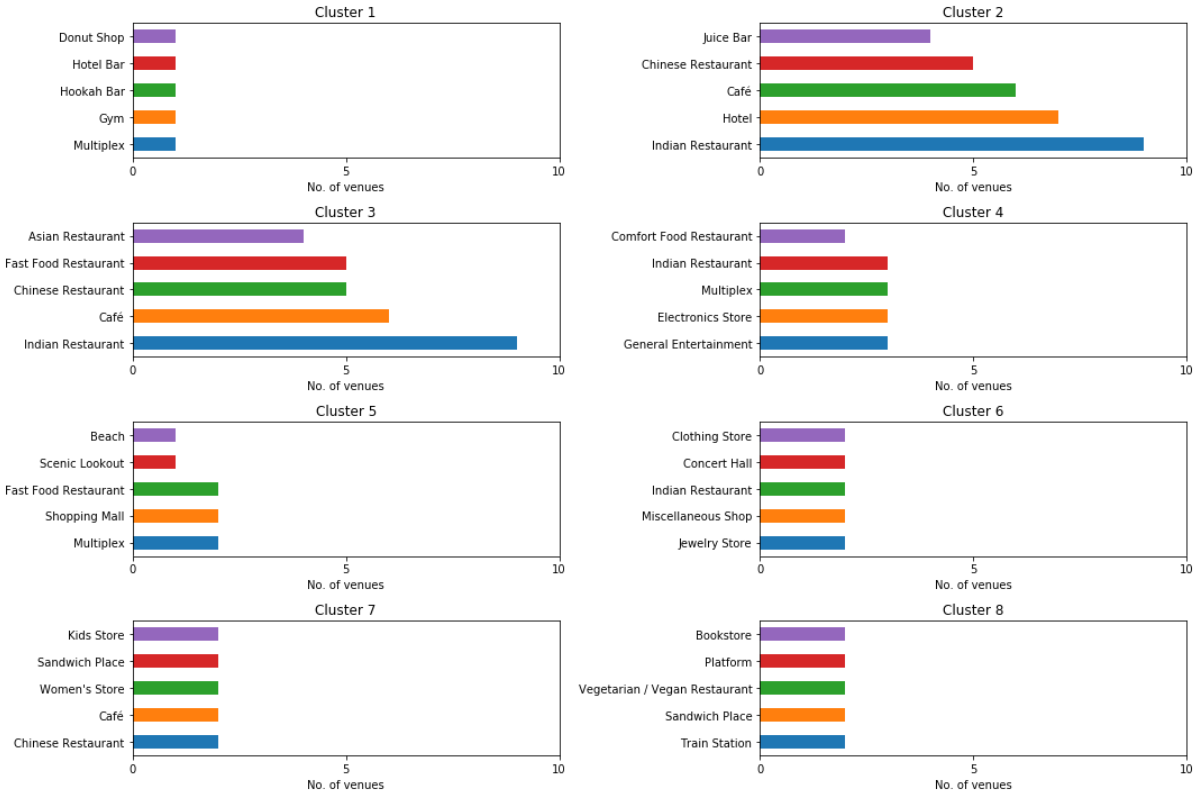
5.8. Cluster 8

The top venue categories in Cluster 8 are Sandwich Place, Vegan Restaurant, Platform, Bookstore and Indian 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
6	Chennai Central	Indian Restaurant	Train Station	Platform	Bookstore	Metro Station	Bus Station	Sandwich Place	Fast Food Restaurant	Nightclub	Vegetarian / Vegan Restaurant
21	Newaharlal Nehru Stadium	Indian Restaurant	Bookstore	Soccer Stadium	Vegetarian / Vegan Restaurant	Train Station	Café	Platform	Juice Bar	Sandwich Place	Electronics Store

6. DISCUSSION

Now that we have the clusters and the top venue categories let's visualize the top 5 venue category in each Cluster for comparison.



This plot can be used to suggest valuable information to Business persons. Let's discuss a few examples considering they would like to start the following category of business.

1. Hotel

The neighborhoods in cluster 2 has the greatest number of hotels, hence opening one here is not the best choice. So, is it best to open one at the neighborhoods in cluster 7 or 8? Not likely, since the place has a smaller number of food restaurants. Thus, an optimal place would be one which has less hotels, but also have restaurants and other places to explore. Considering all these facts, the best choice would be Cluster 3 and Cluster 4. such as the Adyar Bus Depot, Triplicane neighborhoods.

2. Shopping Mall

The neighborhoods 5 has notable number of shopping malls. By using the same procedure as above, the suitable cluster would be the Cluster 2 and Cluster 3, since it has not much shopping malls and also it has many Hotels and Restaurants which gives an advantage.

Similarly, based on the requirement suggestions can be provided about the neighborhood that would be best suitable for the business. Fig 6.1 shows a map of Chennai with the neighborhood clusters superimposed on top of it. This map can be used to suggest a vast location to start a new business based on the category.

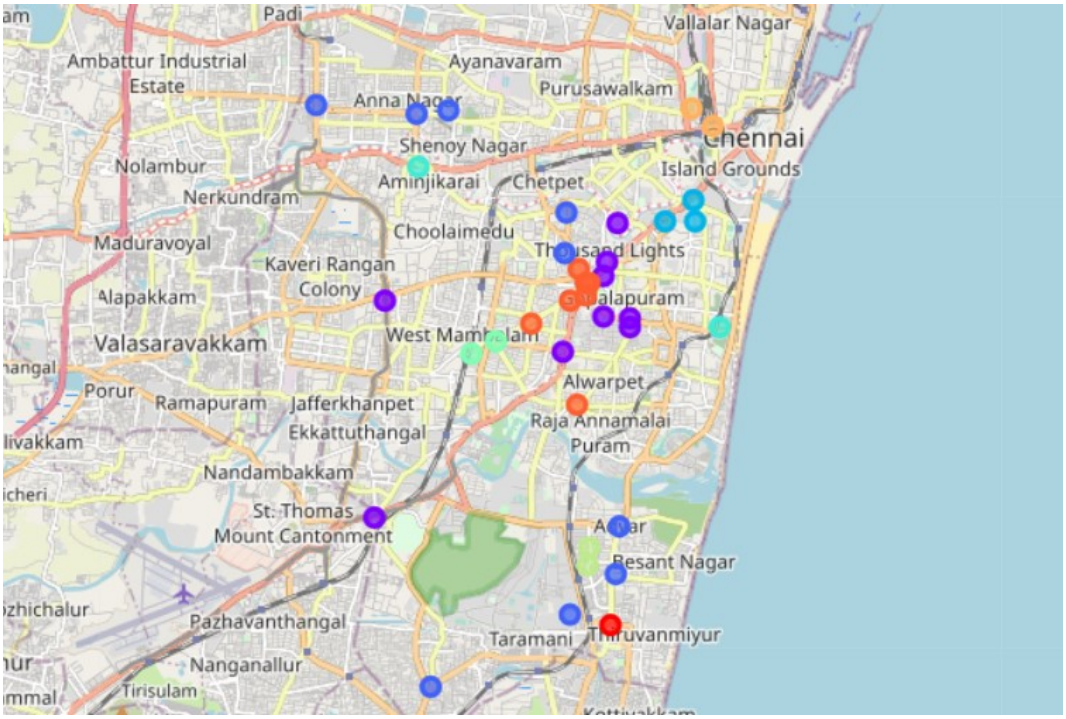


Fig 6.1 Map of Chennai with the Clusters Superimposed on Top

For example, the highlighted location shown in Fig 6.2 consists of Cluster 3 and Cluster 5, whose neighborhoods have many Restaurants and Shopping Malls but less Hotels. Thus, this would be a suitable location for building a hotel.

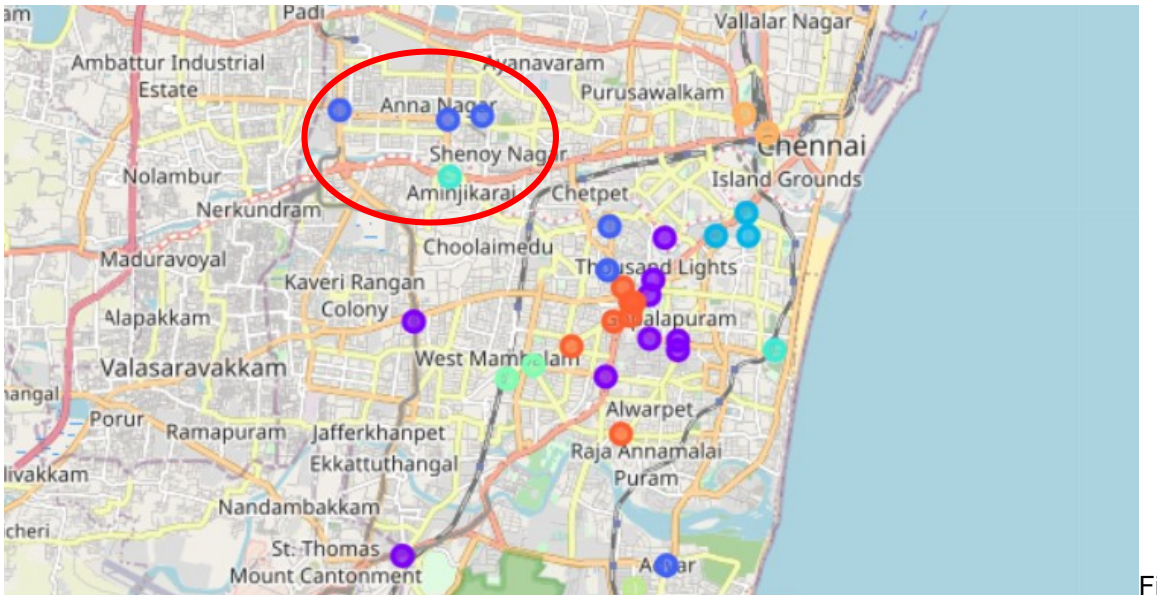


Fig 6.2 Location suitable to start a new Hotel

7. CONCLUSION

Purpose of this project was to analyze the neighborhoods of Chennai and create a clustering model to suggest personals places to start a new business based on the category. The neighborhoods data was obtained from an online source and the Foursquare API was used to find the major venues in each neighborhood. But we found that many neighborhoods had less than 10 venues returned. In order to build a good Data Science model, we filtered out these locations. The remaining locations were used to create a clustering model. The best number of clusters i.e. 8 was obtained using the silhouette score. Each cluster was examined to find the most venue categories present, that defines the characteristics for that particular cluster. A few examples for the applications that the clusters can be used for have also been discussed. A map showing the clusters have been provided.

Both these can be used by stakeholders to decide the location for the particular type of business. A major drawback of this project was that the Foursquare API returned only few venues in each neighborhood. As a future improvement, better data sources can be used to obtain more venues in each neighborhood. This way the neighborhoods that were filtered out can be included in the clustering analysis to create a better decision model.