The Battle of Neighborhoods – Bangalore City.

Applied Data Science Capstone by IBM on Coursera.

Author:

Saideep R Naik.

1. INTRODUCTION: BUSINESS PROBLEM

This project deals with the major venue categories in the neighborhoods of **Bangalore, India**. This project would specifically help Business personal plan to start new Restaurants, Hotels, etc. in Bangalore, Karnataka, 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 Bangalore 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 Bangalore. 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 Bangalore with restaurants around?
- Which area is best suitable for opening a Shopping Mall in Bangalore?

2. Data Requirements:

Bangalore has multiple neighborhoods. The Kaggle website has a dataset which has the list of locations in Bangalore along with their Latitude and Longitude in degree format. There is a total of 352 neighborhoods as shown in Fig below,

1. https://www.kaggle.com/rmenon1998/bangalore-neighborhoods

```
bang_data.head()
```

	Location	Latitude	Longitude
0	Agram	45.813177	15.977048
1	Amruthahalli	13.066513	77.596624
2	Attur	11.663711	78.533551
3	Banaswadi	13.014162	77.651854
4	Bellandur	58.235358	26.683116

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. http://foursquare.com/

A total of 776 venues data have been obtained from Foursquare. The resultant venues dataset is used for the analysis process.

```
print(bang_venues.shape)
bang_venues.head()
```

```
(776, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agram	45.813177	15.977048	Amélie	45.813842	15.979011	Dessert Shop
1	Agram	45.813177	15.977048	Trg bana Josipa Jelačića	45.813032	15.976868	Plaza
2	Agram	45.813177	15.977048	Vinodol	45.811666	15.975643	Mediterranean Restaurant
3	Agram	45.813177	15.977048	Tržnica Dolac	45.814070	15.977261	Farmers Market
4	Agram	45.813177	15.977048	Cremme Zagreb	45.814987	15.976296	Dessert Shop

3. Methodology:

Now, we have the neighborhoods data of Bangalore (**352 neighborhoods**). We also have the most popular venues in each neighborhood obtained using Foursquare API. A total of **776 venues** have been obtained in the whole city and **179 unique categories**. But as seen we have multiple neighborhoods with less than 15 venues returned. In order to create a good analysis let's consider only the ***neighborhoods with more than 15 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 15 venues which can be remove before performing the analysis to obtain better results. The following plot shows only the neighborhoods from which 15 or more than 15 venues were obtained. The resultant dataset consists of 37 neighborhoods as shown in figure below,

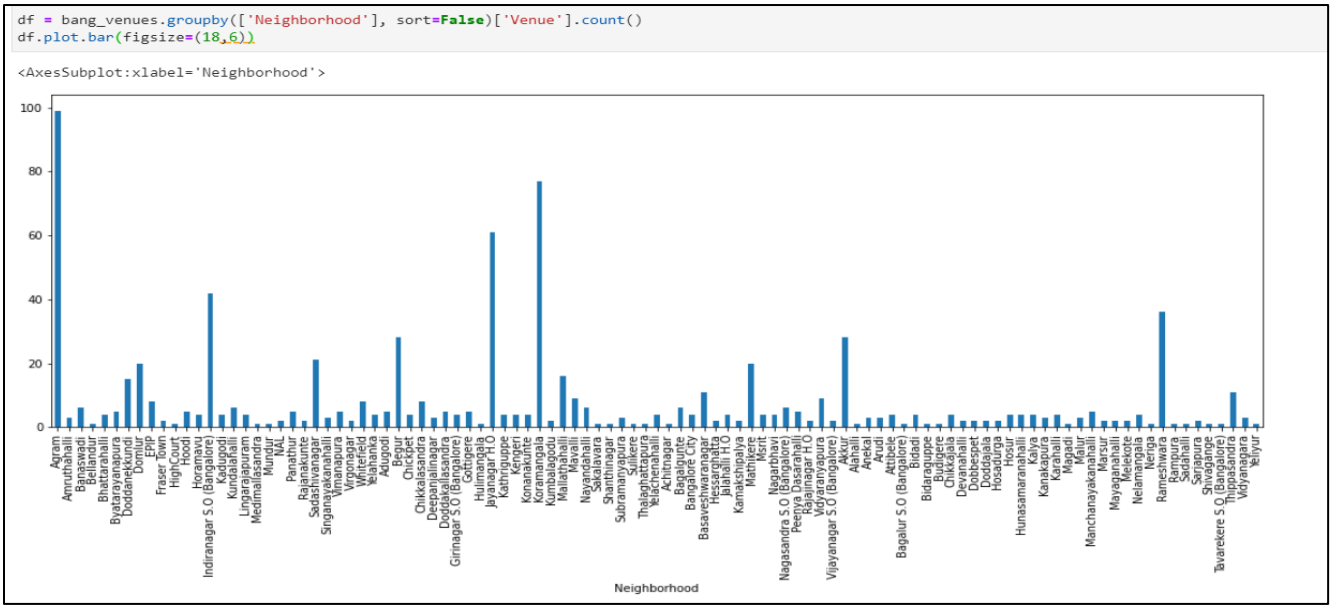


Figure representing all the Neighborhoods in Bangalore City

```
df = df[df >= 15]
df.plot.bar(figsize=(18,6))
```

```
<AxesSubplot:xlabel='Neighborhood'>
```

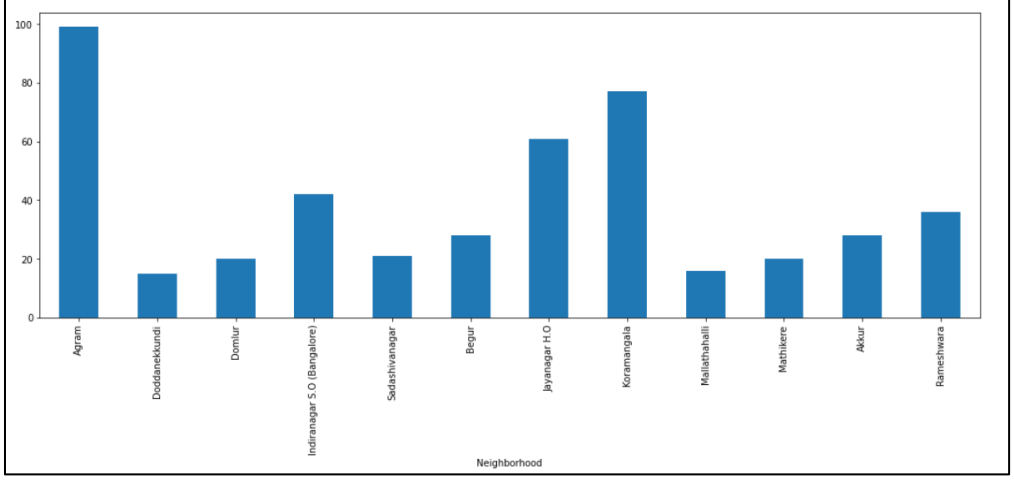


Figure representing Neighborhoods with more than 15 venues.

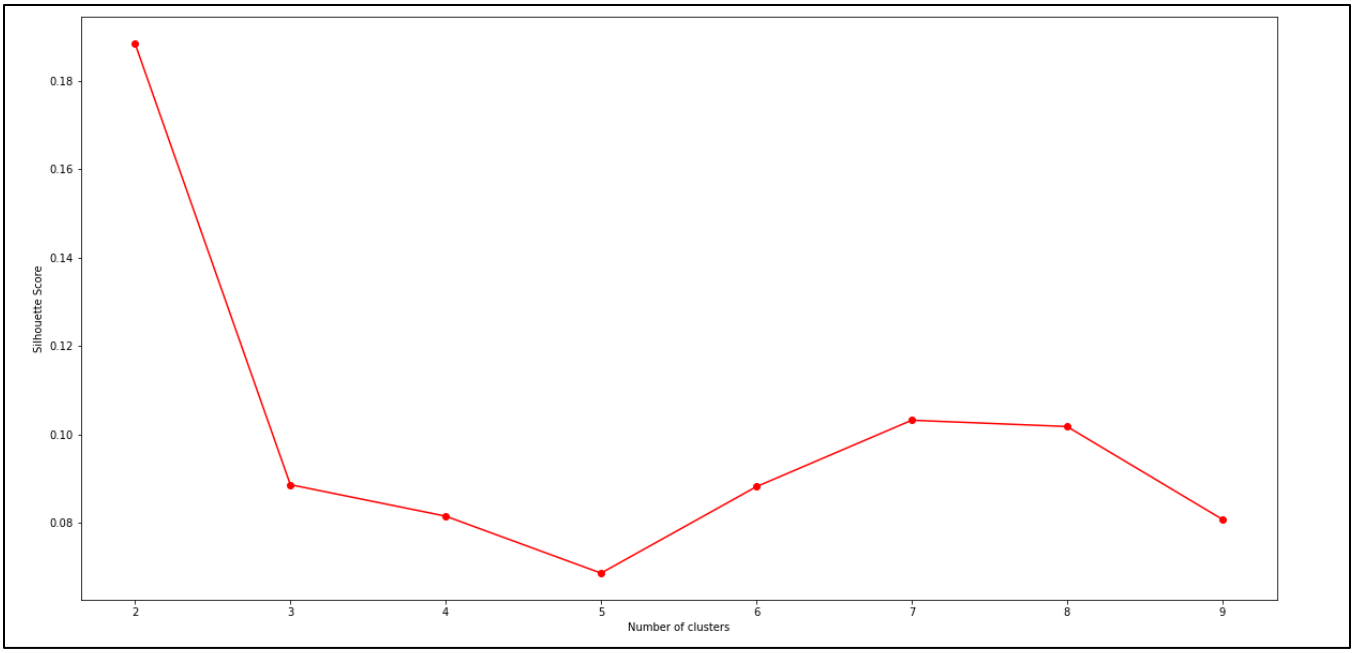
Next, we will perform **one hot encoding** on the filtered data to obtain the venue categories in each neighborhood. Then group the data by neighborhood and take the mean value of the frequency of occurrence of each category. And then we use this data to obtain the top 10 most common venues in each neighborhood i.e. the 10 venues with the highest mean of frequency of occurrence.

```
neighborhoods_venues_sorted.head()
```

	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	Agram	Bar	Plaza	Mediterranean Restaurant	Bistro	Historic Site	Café	Dessert Shop	Theater	Eastern European Restaurant	Pedestrian Plaza
1	Akkur	Supermarket	Pharmacy	Bar	Pet Store	Optical Shop	Discount Store	Clothing Store	Chinese Restaurant	Sandwich Place	Furniture / Home Store
2	Begur	Café	Supermarket	Mini Golf	Pizza Place	Burger Joint	Climbing Gym	Bar	Comedy Club	Korean Restaurant	Park
3	Doddanekkundi	Indian Restaurant	Hotel	Multiplex	Mediterranean Restaurant	Coffee Shop	Restaurant	Movie Theater	Motorcycle Shop	Fast Food Restaurant	Sports Bar
4	Domlur	Indian Restaurant	Café	Chinese Restaurant	Deli / Bodega	Sports Bar	Liquor Store	Pizza Place	Restaurant	Pub	BBQ Joint

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 the graph. From the plot we can see that a K value of 7 provides the best score. This K value is used for the K-Means Clustering Technique.



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 7 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.

1. Cluster 1:

```
bang_merged.loc[bang_merged['Cluster Labels'] == 0, bang_merged.columns[[0] + list(range(4, bang_merged.shape[1]))]]
```

	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	Domlur	Indian Restaurant	Café	Chinese Restaurant	Deli / Bodega	Sports Bar	Liquor Store	Pizza Place	Restaurant	Pub	BBQ Joint
6	Jayanagar H.O	Indian Restaurant	Café	Chinese Restaurant	Juice Bar	Fast Food Restaurant	Restaurant	Sandwich Place	Women's Store	Gym	Gastropub
7	Koramangala	Indian Restaurant	Café	Coffee Shop	Pub	Italian Restaurant	Juice Bar	Seafood Restaurant	Ice Cream Shop	Chinese Restaurant	Kerala Restaurant

2. Cluster 2:

```
bang_merged.loc[bang_merged['Cluster Labels'] == 1, bang_merged.columns[[0] + list(range(4, bang_merged.shape[1]))]]
```

	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	Begur	Café	Supermarket	Mini Golf	Pizza Place	Burger Joint	Climbing Gym	Bar	Comedy Club	Korean Restaurant	Park

3. Cluster 3:

```
bang_merged.loc[bang_merged['Cluster Labels'] == 2, bang_merged.columns[[0] + list(range(4, bang_merged.shape[1]))]]
```

	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	Agram	Bar	Plaza	Mediterranean Restaurant	Bistro	Historic Site	Café	Dessert Shop	Theater	Eastern European Restaurant	Pedestrian Plaza
3	Indiranagar S.O (Bangalore)	Lounge	Pub	Cocktail Bar	Cupcake Shop	Indian Restaurant	Italian Restaurant	Restaurant	Café	Kids Store	Music Venue
4	Sadashivanagar	Indian Restaurant	Spa	Coffee Shop	Department Store	Ice Cream Shop	Dessert Shop	Plaza	Electronics Store	Chinese Restaurant	Seafood Restaurant
11	Rameshwara	Clothing Store	Indian Restaurant	Ice Cream Shop	Fast Food Restaurant	Donut Shop	Electronics Store	Coffee Shop	American Restaurant	Miscellaneous Shop	Sandwich Place

4. Cluster 4:

```
bang_merged.loc[bang_merged['Cluster Labels'] == 3, bang_merged.columns[[0] + list(range(4, bang_merged.shape[1]))]]
```

	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
9	Mathikere	American Restaurant	Bakery	Café	Bus Station	Burger Joint	Gym	Market	Diner	Ice Cream Shop	Indian Restaurant

5. Cluster 5:

```
bang_merged.loc[bang_merged['Cluster Labels'] == 4, bang_merged.columns[[0] + list(range(4, bang_merged.shape[1]))]]
```

	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	Mallathahalli	Café	Indian Restaurant	Market	Breakfast Spot	Gym	South Indian Restaurant	Andhra Restaurant	Eastern European Restaurant	Electronics Store	Escape Room

6. Cluster 6:

```
bang_merged.loc[bang_merged['Cluster Labels'] == 5, bang_merged.columns[[0] + list(range(4, bang_merged.shape[1]))]]
```

	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
10	Akkur	Supermarket	Pharmacy	Bar	Pet Store	Optical Shop	Discount Store	Clothing Store	Chinese Restaurant	Sandwich Place	Furniture / Home Store

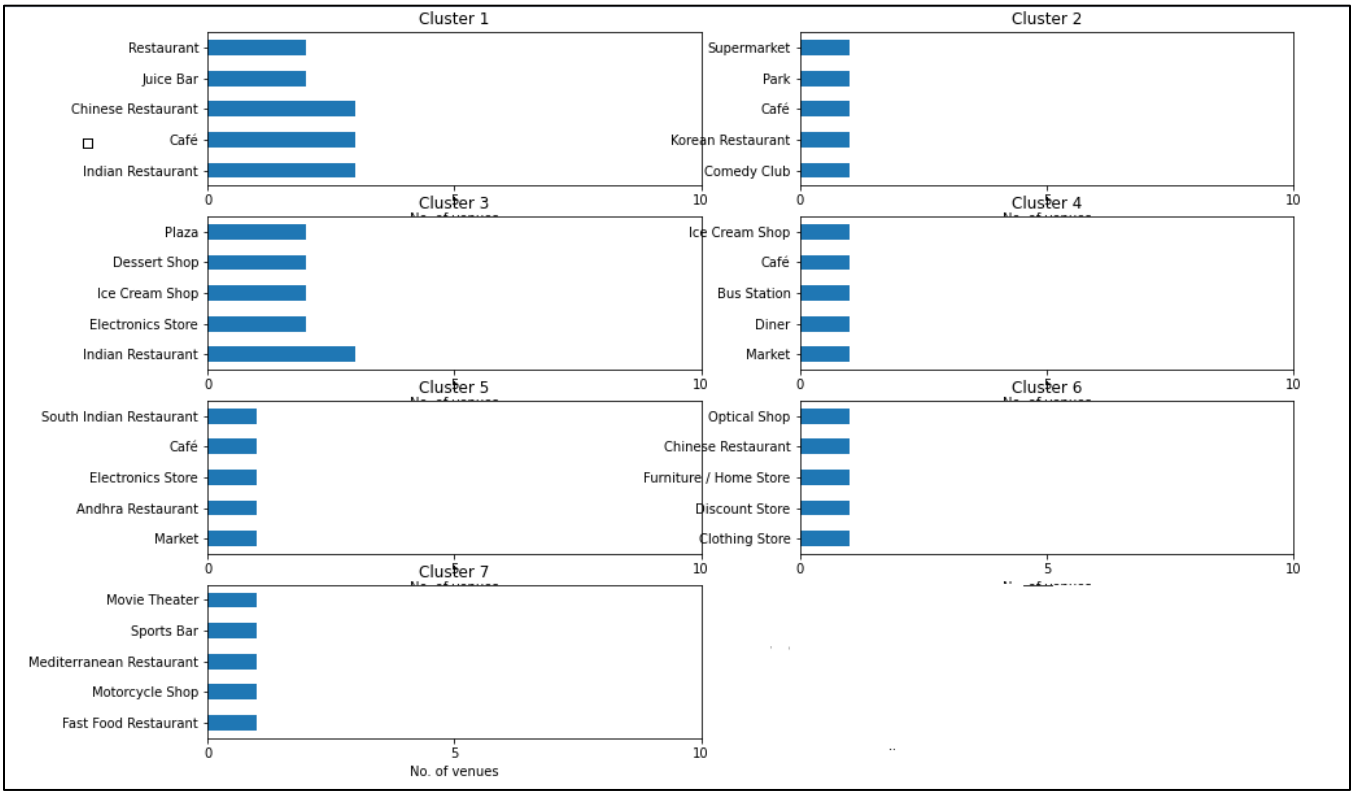
7. Cluster 7:

```
bang_merged.loc[bang_merged['Cluster Labels'] == 6, bang_merged.columns[[0] + list(range(4, bang_merged.shape[1]))]]
```

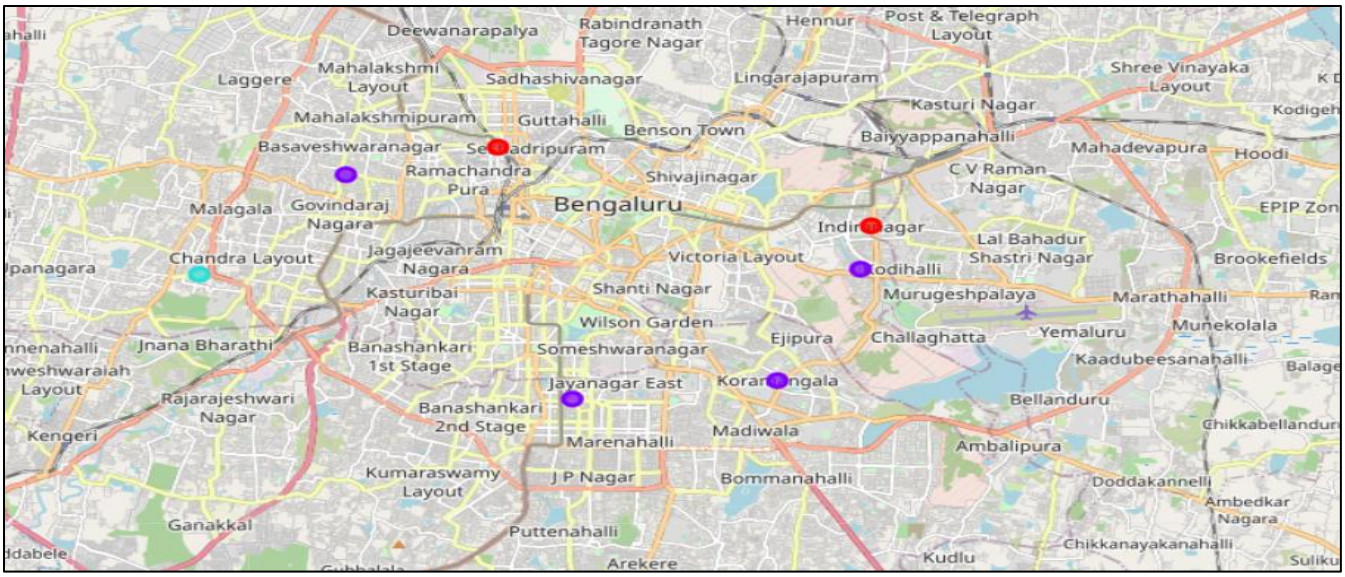
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Doddanekkundi	Indian Restaurant	Hotel	Multiplex	Mediterranean Restaurant	Coffee Shop	Restaurant	Movie Theater	Motorcycle Shop	Fast Food Restaurant	Sports Bar

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.



We can see that Cluster 1 is suitable for Chinese restaurant, Café and Indian restaurant. Also, Cluster 3 is best suitable for Indian restaurants. We can see that other clusters such as Cluster 2, Cluster 4, Cluster 5, Cluster 6 and Cluster 7 the values are same for all the category of business.



Map of Bangalore with the Clusters Superimposed on Top.

7. Conclusion:

Purpose of this project was to analyze the neighborhoods of Bangalore 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 15 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. 7 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.