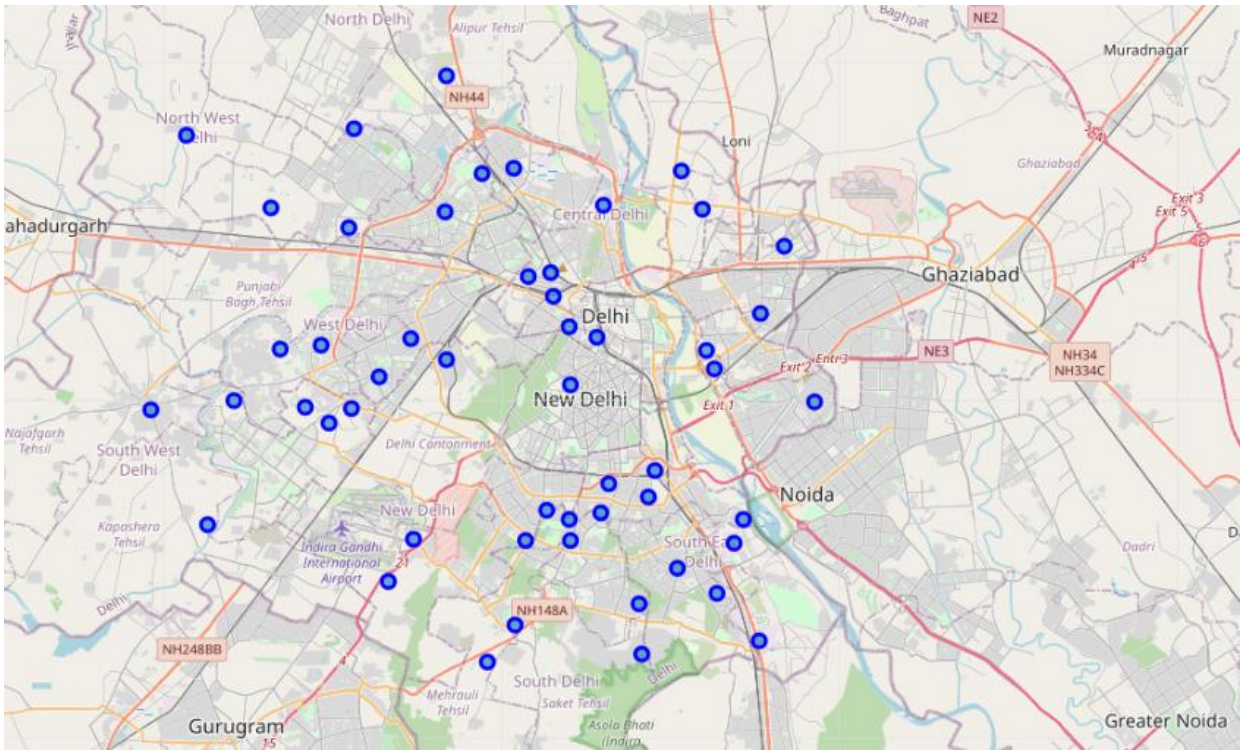
The Battle of Neighborhoods - Delhi



Applied Data Science Capstone by IBM on Coursera

Table of Contents

1. Introduction: Business Problem

This project deals with the major venue categories in the neighborhoods of Delhi, capital of India. This project would specifically help business personal plan to start new restaurants, hotels, etc. in Delhi, 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 Delhi 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 café shops.

The major target audience would be small-scale business owners and stake holders planning to start their business at a location in within Delhi. 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 Delhi with restaurants around?
- Which area is best suitable for opening a café shops in Delhi?

2. Data Requirement

Delhi has multiple neighborhoods. The https://www.latlong.net/category/districts-102-16.html website has a dataset which has the list of locations within India along with their Latitude and Longitude. There is a total of 53 neighborhoods as shown in Fig 2.1.

```
print (df_NDLS.shape)
df_NDLS.head()
```

```
(53, 3)
```

]:

	Neighborhood	Latitude	Longitude
0	Green Park	28.558899	77.202805
1	Rajouri Garden, New Delhi	28.641529	77.120918
2	Bindapur, New Delhi	28.610722	77.065971
3	Karkardooma, Anand Vihar	28.652946	77.302284
4	Dilshad Garden, New Delhi	28.683903	77.315094

Figure 2-1Delhi Neighborhood

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.

https://foursquare.com/

A total of 297 venues data have been obtained from Foursquare. The resultant venues dataset, (shown in Fig 2.2) is used for the analysis process.

```
print(delhi_venues.shape)
delhi_venues.head()
```

```
(296, 7)
```

]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Green Park	28.558899	77.202805	Adyar Ananda Bhavan	28.557293	77.202502	Indian Restaurant
1	Green Park	28.558899	77.202805	Gung The Palace	28.556827	77.205298	Korean Restaurant
2	Green Park	28.558899	77.202805	Green Park Market	28.557181	77.202821	Market
3	Green Park	28.558899	77.202805	Tamura	28.558154	77.206356	Japanese Restaurant
4	Green Park	28.558899	77.202805	Evergreen Sweets	28.556497	77.202411	Indian Restaurant

A total of 296 venues were obtained. Now lets check the number of venues returned per neighborhood.

Figure 2-2 Delhi values dataset

3. Methodology

Now, we have the neighborhoods data of Delhi (47 neighborhoods). We also have the most popular venues in each neighborhood obtained using Foursquare API. A total of 296 venues have been obtained in the whole city and 90 unique categories. But as seen we have multiple neighborhoods with less than 4 venues returned. In order to create a good analysis let's consider only the neighborhoods with more than 4 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, suitable locations based on the category.

4. Analysis

Looking into the dataset we found that there were many neighborhoods with less than 4 venues which can be remove before performing the analysis to obtain better results. The following plot shows only the neighborhoods from which 4 or more than 4 venues were obtained. The resultant dataset consists of 37 neighborhoods as shown in Fig 4.1.

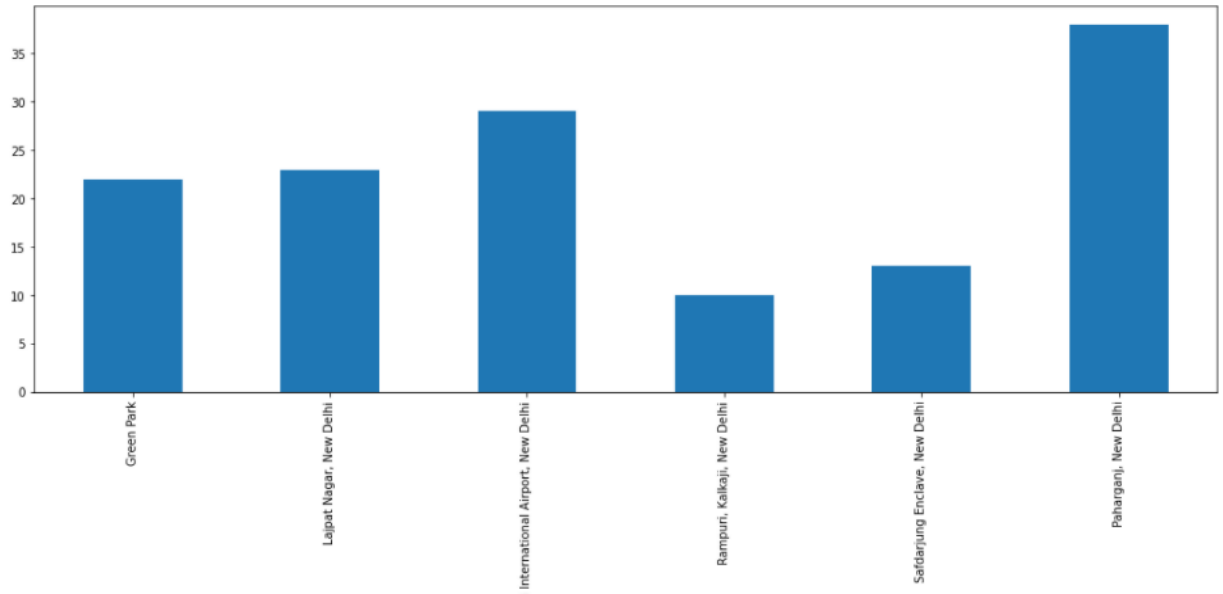


Figure 4-1 Filtered Neighborhood Dataset

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

```
delhi_grouped.head()
```

```
(6, 48)
```

	Neighborhood	Afghan Restaurant	Arcade	Asian Restaurant	Bakery	Bar	Bed & Breakfast	Buffet	Café	Chinese Restaurant	Clothing Store	Coffee Shop	Convenience Store	Department Store	Dessert Shop	Donut Shop	Fast Food Restaurant	Food Truck	French Restaurant
0	Aerocity, Indira Gandhi International Airport,...	0.000000	0.000000	0.000000	0.034483	0.000000	0.034483	0.034483	0.000000	0.000000	0.000000	0.068966	0.000000	0.034483	0.0	0.000000	0.000000	0.0	0.0
1	Green Park	0.000000	0.000000	0.045455	0.000000	0.045455	0.000000	0.000000	0.045455	0.045455	0.000000	0.181818	0.000000	0.000000	0.0	0.090909	0.000000	0.0	0.0
2	Lajpat Nagar, New Delhi	0.043478	0.043478	0.000000	0.000000	0.000000	0.000000	0.000000	0.086957	0.000000	0.043478	0.086957	0.130435	0.000000	0.0	0.086957	0.086957	0.0	0.0
3	Paharganj, New Delhi	0.000000	0.000000	0.000000	0.026316	0.000000	0.000000	0.000000	0.078947	0.026316	0.000000	0.026316	0.000000	0.000000	0.0	0.000000	0.052632	0.0	0.0
4	Rampuri, Kalkaji, New Delhi	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.100000	0.000000	0.000000	0.000000	0.1	0.000000	0.100000	0.1	0.0

Figure 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	Aerocity, Indira Gandhi International Airport,...	Hotel	Indian Restaurant	Coffee Shop	Hotel Bar	Train Station	Bed & Breakfast	Lounge	Italian Restaurant	Buffet	Punjabi Restaurant
1	Green Park	Indian Restaurant	Coffee Shop	Donut Shop	Pizza Place	Chinese Restaurant	Korean Restaurant	Japanese Restaurant	Park	Café	Bar
2	Lajpat Nagar, New Delhi	Convenience Store	Pizza Place	Fast Food Restaurant	Donut Shop	Market	Coffee Shop	Café	Afghan Restaurant	Indian Restaurant	Indie Movie Theater
3	Paharganj, New Delhi	Hotel	Platform	Café	Fast Food Restaurant	Restaurant	Indian Restaurant	Korean Restaurant	Train Station	Motel	Bakery
4	Rampuri, Kalkaji, New Delhi	Gym	Hotel	Clothing Store	Food Truck	Sandwich Place	Fast Food Restaurant	Dessert Shop	Pizza Place	Convenience Store	Gym / Fitness Center

Figure 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 2 provides the best score. This K value is used for the K-Means Clustering Technique

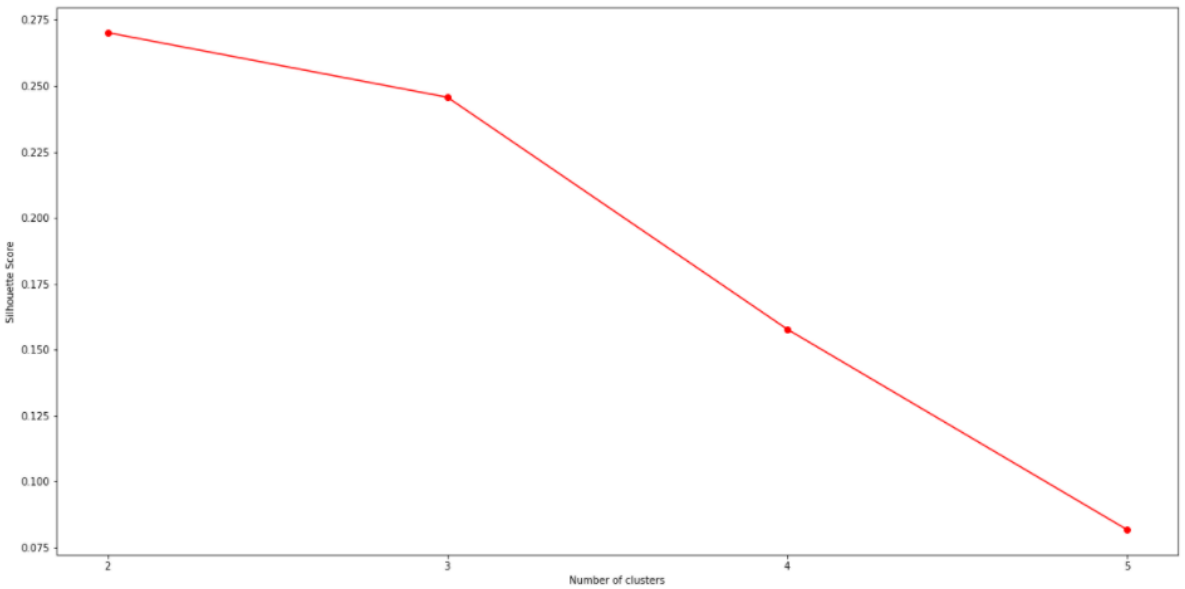


Figure 4-4 Silhouette Score for different Number of Cluster

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.

1. Cluster 1

The top venue categories in Cluster 1 are Hotel, Indian Restaurant, Coffee Shop, Hotel Bar, Train Station.

	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	Aerocity, Indira Gandhi International Airport,...	Hotel	Indian Restaurant	Coffee Shop	Hotel Bar	Train Station	Bed & Breakfast	Lounge	Italian Restaurant	Buffet	Punjabi Restaurant
3	Rampuri, Kalkaji, New Delhi	Gym	Hotel	Clothing Store	Food Truck	Sandwich Place	Fast Food Restaurant	Dessert Shop	Pizza Place	Convenience Store	Gym / Fitness Center
4	Safdarjung Enclave, New Delhi	Hotel	Spa	Park	Gym / Fitness Center	French Restaurant	Fast Food Restaurant	Coffee Shop	Nightclub	Stadium	Bed & Breakfast
5	Paharganj, New Delhi	Hotel	Platform	Café	Fast Food Restaurant	Restaurant	Indian Restaurant	Korean Restaurant	Train Station	Motel	Bakery

Figure 5-1 Cluster1

2. Cluster 2

The top venue categories in Cluster 2 are Indian Restaurant, Coffee Shop, Donut Shop, 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
0	Green Park	Indian Restaurant	Coffee Shop	Donut Shop	Pizza Place	Chinese Restaurant	Korean Restaurant	Japanese Restaurant	Park	Café	Bar
1	Lajpat Nagar, New Delhi	Convenience Store	Pizza Place	Fast Food Restaurant	Donut Shop	Market	Coffee Shop	Café	Afghan Restaurant	Indian Restaurant	Indie Movie Theater

Figure 5-2 Cluster2

3. Cluster 3

The top venue categories in Cluster 3 are Hotel, Indian Restaurant, Coffee Shop, Hotel 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	Aerocity, Indira Gandhi International Airport,...	Hotel	Indian Restaurant	Coffee Shop	Hotel Bar	Train Station	Bed & Breakfast	Lounge	Italian Restaurant	Buffet	Punjabi Restaurant
3	Rampuri, Kalkaji, New Delhi	Gym	Hotel	Clothing Store	Food Truck	Sandwich Place	Fast Food Restaurant	Dessert Shop	Pizza Place	Convenience Store	Gym / Fitness Center
4	Safdarjung Enclave, New Delhi	Hotel	Spa	Park	Gym / Fitness Center	French Restaurant	Fast Food Restaurant	Coffee Shop	Nightclub	Stadium	Bed & Breakfast
5	Paharganj, New Delhi	Hotel	Platform	Café	Fast Food Restaurant	Restaurant	Indian Restaurant	Korean Restaurant	Train Station	Motel	Bakery

Figure 5-3 Cluster 3

4. Cluster 4

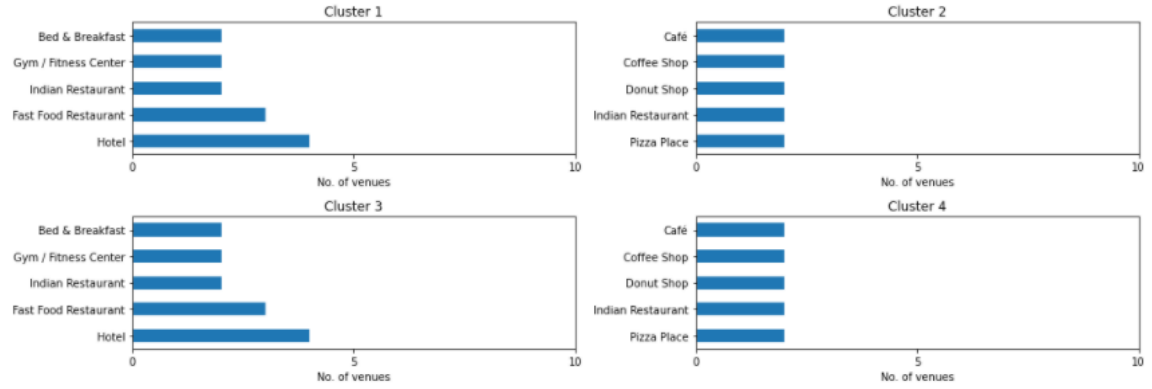
The top venue categories in Cluster 4 are Indian Restaurant, Coffee Shop, Donut Shop, Pizza Place

	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
0	Green Park	Indian Restaurant	Coffee Shop	Donut Shop	Pizza Place	Chinese Restaurant	Korean Restaurant	Japanese Restaurant	Park	Café	Bar
1	Lajpat Nagar, New Delhi	Convenience Store	Pizza Place	Fast Food Restaurant	Donut Shop	Market	Coffee Shop	Café	Afghan Restaurant	Indian Restaurant	Indie Movie Theater

Figure 5-4 Cluster 4

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. 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 highest number of hotels, hence opening one here is not the best choice. So, is it best to open one at the neighborhoods in cluster 1 or 3? 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 4.

2. Shopping Mall

The neighborhoods 4 has notable number of cafés. By using the same procedure as above, the suitable cluster would be the Cluster 1 and Cluster 3 since it has not much cafe 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.

Map of Delhi with the clusters superimposed on top.

This map can be used to find a suitable location to start a new business based on the category.

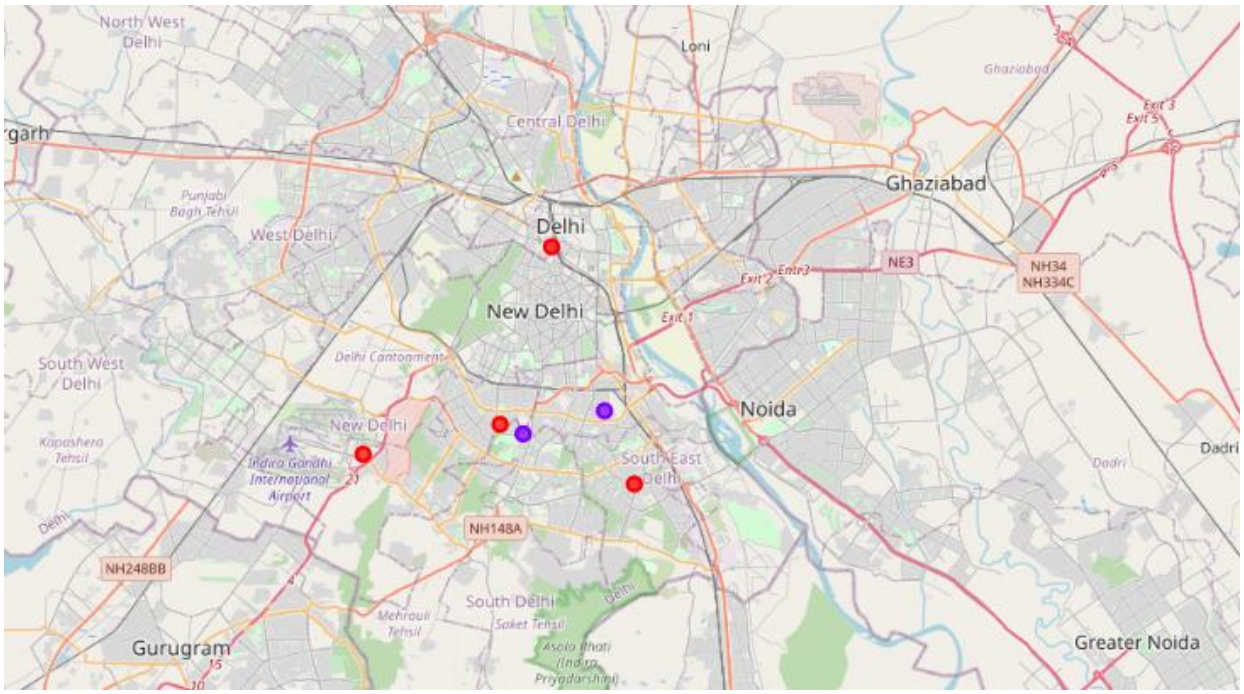


Figure 6-1Map of Delhi with the Clusters Superimposed on Top

7. Conclusion

Purpose of this project was to analyze the neighborhoods of Delhi 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. 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. 2 was obtained using the silhouette score. Each cluster was examined to find the most venue categories present, that defines the characteristics for that cluster. A few examples for the applications that the clusters can be used for having, also been discussed. A map showing the clusters have been provided.

Both these can be used by stakeholders to decide the location for the 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.