APPLIED DATA SCIENCE CAPSTONE

THE BATTLE OF NEIGHBOURHOODS

Understanding the Neighborhoods of Delhi and recommending Neighbourhoods for Indian Restaurants

- Submitted By: Nishit Chaudhry

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

1. Introduction

1.1 Background

Delhi, officially the National Capital Territory of Delhi (NCT), is a city and a union territory of India containing New Delhi, the capital of India. The NCT covers an area of 1,484 square kilometres (573 sq mi). According to the 2011 census, Delhi's city proper population was over 11 million, the second-highest in India after Mumbai, while the whole NCT's population was about 16.8 million.

Delhi's urban area is now considered to extend beyond the NCT boundaries, and include the neighbouring satellite cities of Ghaziabad, Faridabad, Gurgaon and Noida in an area now called National Capital Region (NCR) and had an estimated 2016 population of over 26 million people, making it the world's second-largest urban area according to the United Nations.

Delhi is the second-wealthiest city in India after Mumbai and is home to 18 billionaires and 23,000 millionaires. Delhi ranks fifth among the Indian states and union territories in human development index. Delhi has the second-highest GDP per capita in India. It is one of the world's most polluted cities by particulate matter concentration.

1.2 Business Problem

This project will take a deeper look into the essence of this city which is known for its history, heritage, authentic food and diverse culture. We will explore the districts of Delhi by segmenting and clustering the districts based on its popular venues.

First half of this project will provide in detail analysis and visualisations at-a-glance to understand the different neighborhoods of each district and we will cluster these neighborhoods on the basis of top 5 venue categories found in each neighbourhood.

The next half will be related to restaurants where we will compare the neighbourhoods and segment them into clusters according to the types and frequencies of different food joints found in the neighbourhoods. This part will be of interest to stakeholders, businessmen, restaurant owner's who are either looking to expand their Indian restaurant chain to other neighbourhoods or planning to start a new Indian restaurant altogether. We will recommend neighbourhoods in Delhi that might be optimal for opening new Indian restaurants.

2. Data

2.1 Packages and Dependencies

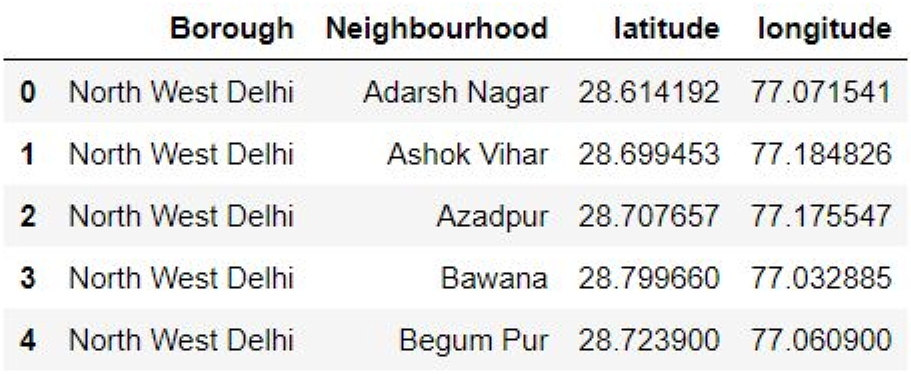
- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sklearn
- Folium
- Geopy
- Requests
- Json

2.2 Datasets

The dataset used here will comprise the Boroughs, Neighborhoods, Latitude and Longitude of Delhi. The dataset will be downloaded from an external source for the same.

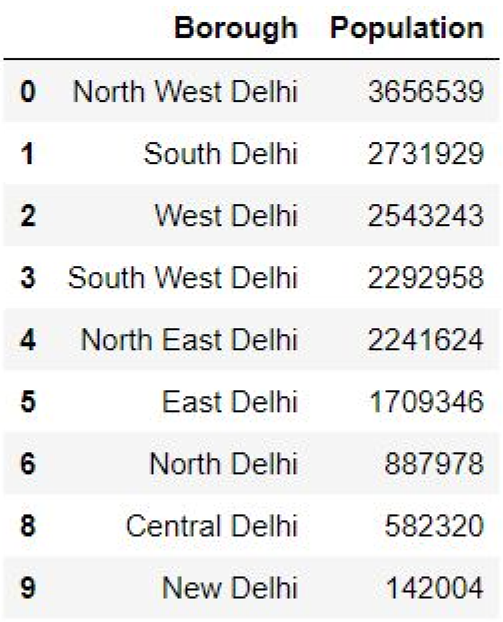
The longitude and latitude data will be used by the Foursquare API to help in exploring the neighborhood venues as well as to map the clusters on the map using the Folium package. The city venues will be compared on the basis of the Boroughs and Neighborhoods.

1. Neighbourhoods of Delhi



	Borough	Neighbourhood	latitude	longitude
0	North West Delhi	Adarsh Nagar	28.614192	77.071541
1	North West Delhi	Ashok Vihar	28.699453	77.184826
2	North West Delhi	Azadpur	28.707657	77.175547
3	North West Delhi	Bawana	28.799660	77.032885
4	North West Delhi	Begum Pur	28.723900	77.060900

2. Population of Delhi by Boroughs



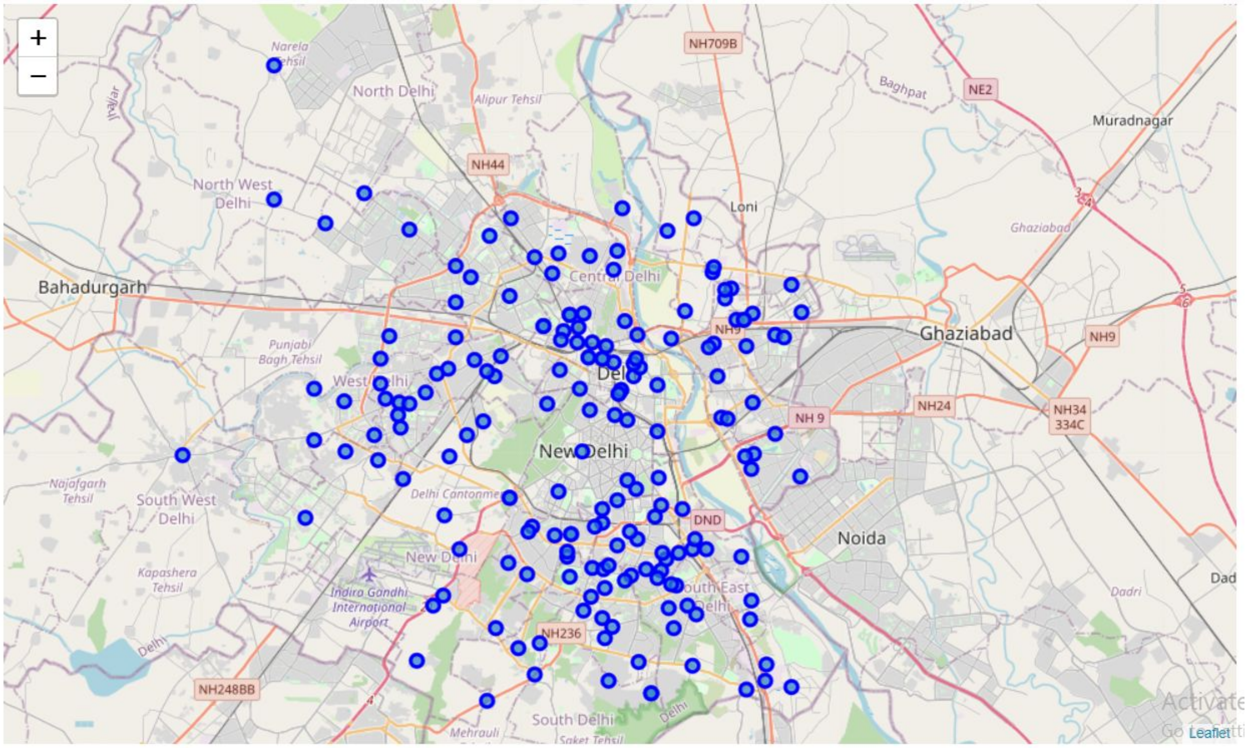
	Borough	Population
0	North West Delhi	3656539
1	South Delhi	2731929
2	West Delhi	2543243
3	South West Delhi	2292958
4	North East Delhi	2241624
5	East Delhi	1709346
6	North Delhi	887978
8	Central Delhi	582320
9	New Delhi	142004

2.3 Foursquare API

Foursquare API has a database of more than 105 million places updated in real-time. It is a very engaging platform with details from the users who tips to the details of the venues and places of cities, states. This project will use Foursquare API as its prime source of data gathering.

Foursquare uses the Latitude and Longitude of the data to provide a json file with the details of the venues like name, category, latitude and longitude. Also, Foursquare API uses the Client ID, Client secret(which is basically a password) and the Version(a date based on which you will get the json file)

Using the Folium package, we plotted the below map of Delhi containing the location of Neighborhoods derived from the dataframe above.



Neighbourhoods of Delhi

3. Methodology

In this project, we will explore the neighbourhoods of Delhi by taking a deeper look into its venue categories and places. First, we will get nearby venues of each Neighbourhood(185 to be exact) in the Boroughs. This step of the project will be completed with the help of Foursquare API. After which we will analyse the venues which results from the Foursquare data. After properly checking for the missing values in the external data, we will get dummies(one hot encoding) for each venue in the data preprocessing step and get the mean by grouping this data by Neighbourhood. Now, we will build a data frame containing the top 5 venues in each neighbourhood in terms of the frequency of their occurence.

This data is now ideal to be used to apply Unsupervised learning, in this case, KMeans clustering to get the segmentation of the data on the basis of the venues. We do this by first dropping the Neighbourhood column so only numerical data with all the venues are left to be inserted into the model. We will cluster this data into 5 clusters, hence k = 5 here. The data we have in the dataframe is unlabelled which indicates unsupervised nature of the data. Once we receive the cluster labels for each row, we will add back the Neighbourhood, Borough, Latitude and Longitude columns to the data which will help us map this dataframe to the Delhi map using the Folium package.

Once we have plotted the data, we will take a deeper look into the clusters to make sense of how the data is segmented into this cluster and what distinguishes them from each other. For this step we will plot each cluster into bar plots to highlight the venue category occuring in each cluster and as well as the most occuring venue categories to label these clusters. Okay, so to take a look at the cluster distribution at the Boroughs level, we will plot a box plot between clusters and boroughs.

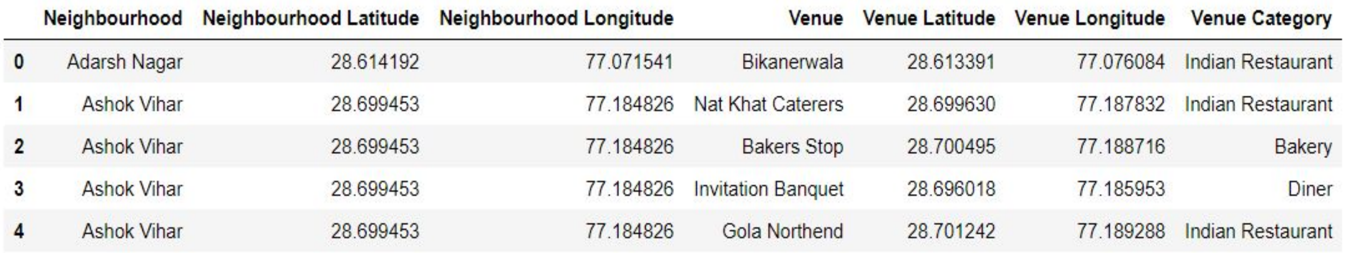
After this to answer the second part of the objective i.e recommending neighbourhoods for Indian restaurants, we will move our focus to food restaurants in the neighbourhoods. For this, we will further filter our data and take only categories related to food joints/restaurants. Then we will plot bar plots and pie plots to visualize this data. Now, we will drop all the venue categories except Indian Restaurants to take a look at the distribution throughout the Boroughs.

Okay to include some external features as well in our analysis, we will use the population census of Delhi to take a look at the population distribution. This step is not necessary for the project, but it can highlight the x-factor for the population, i.e to see if some boroughs are more densely populated than others or not. This might be useful in overall understanding of the audience that might interest the stakeholders and businessmen in better understanding Delhi as a city and to make further optimum recommendations as well.

At last, we will come to the recommendation part. Here we will again cluster the data which we filtered earlier related to just food joints. We will again use KMeans clustering with 3 clusters this time to cluster the data. After this we will again try to make a sense of the clusters by plotting the barplots and we will try to label the clusters.

4. Analysis

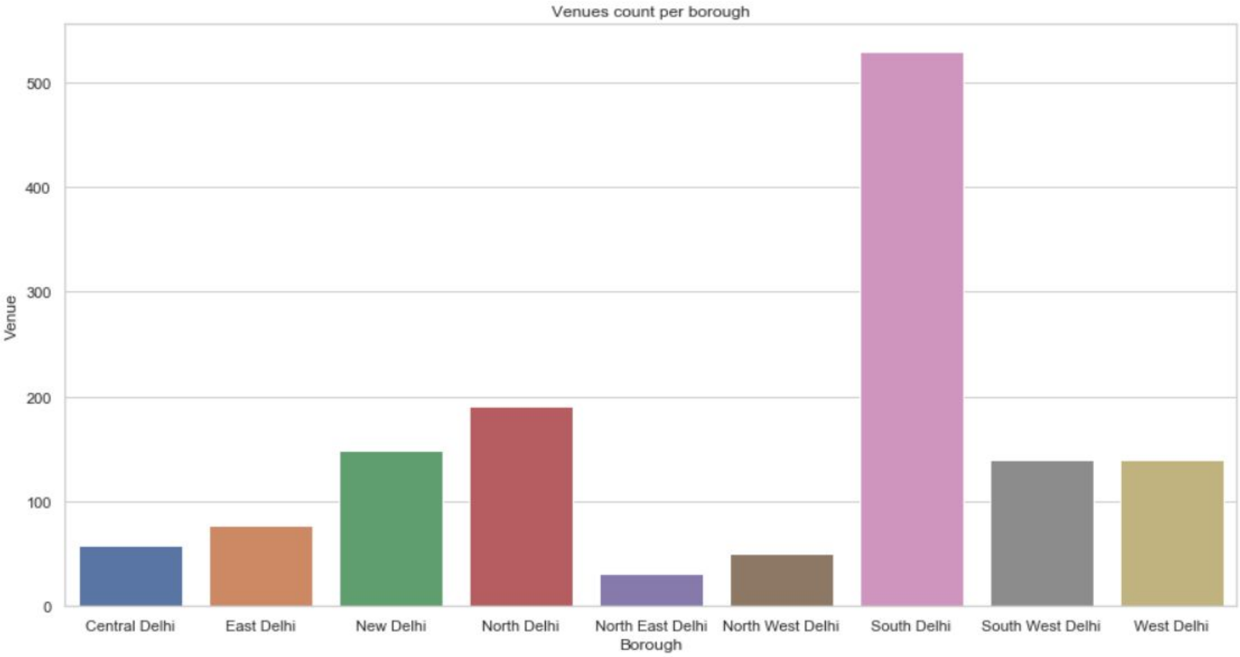
Let's start our Exploratory Data Analysis of the Neighbourhoods and Boroughs of Delhi to understand the city better and ultimately recommend neighbourhoods optimum to start a new Indian Restaurants.



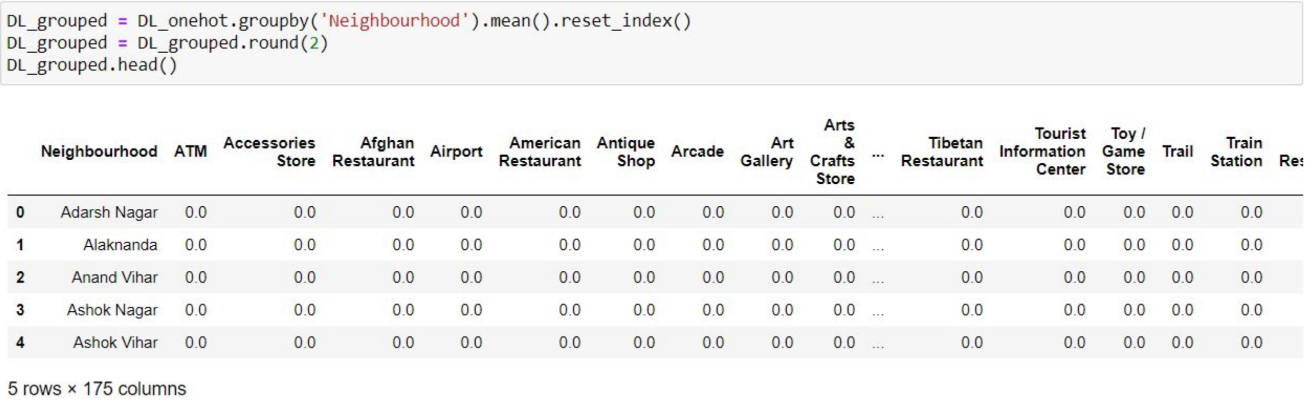
	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adarsh Nagar	28.614192	77.071541	Bikanerwala	28.613391	77.076084	Indian Restaurant
1	Ashok Vihar	28.699453	77.184826	Nat Khat Caterers	28.699630	77.187832	Indian Restaurant
2	Ashok Vihar	28.699453	77.184826	Bakers Stop	28.700495	77.188716	Bakery
3	Ashok Vihar	28.699453	77.184826	Invitation Banquet	28.696018	77.185953	Diner
4	Ashok Vihar	28.699453	77.184826	Gola Northend	28.701242	77.189288	Indian Restaurant

Delhi venue list extracted using Foursquare API

Part-1: Understanding the Neighbourhoods of Delhi



Bar graph showing number of venues in each Borough

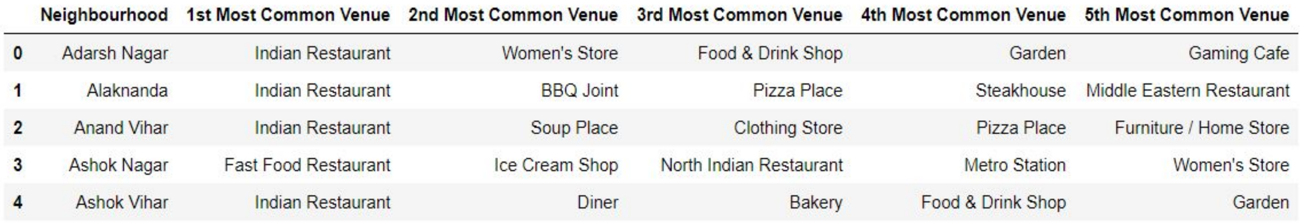


```
DL_grouped = DL_onehot.groupby('Neighbourhood').mean().reset_index()
DL_grouped = DL_grouped.round(2)
DL_grouped.head()
```

	Neighbourhood	ATM	Accessories Store	Afghan Restaurant	Airport	American Restaurant	Antique Shop	Arcade	Art Gallery	Arts & Crafts Store	...	Tibetan Restaurant	Tourist Information Center	Toy / Game Store	Trail	Train Station	Res
0	Adarsh Nagar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
1	Alaknanda	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
2	Anand Vihar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
3	Ashok Nagar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
4	Ashok Vihar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	

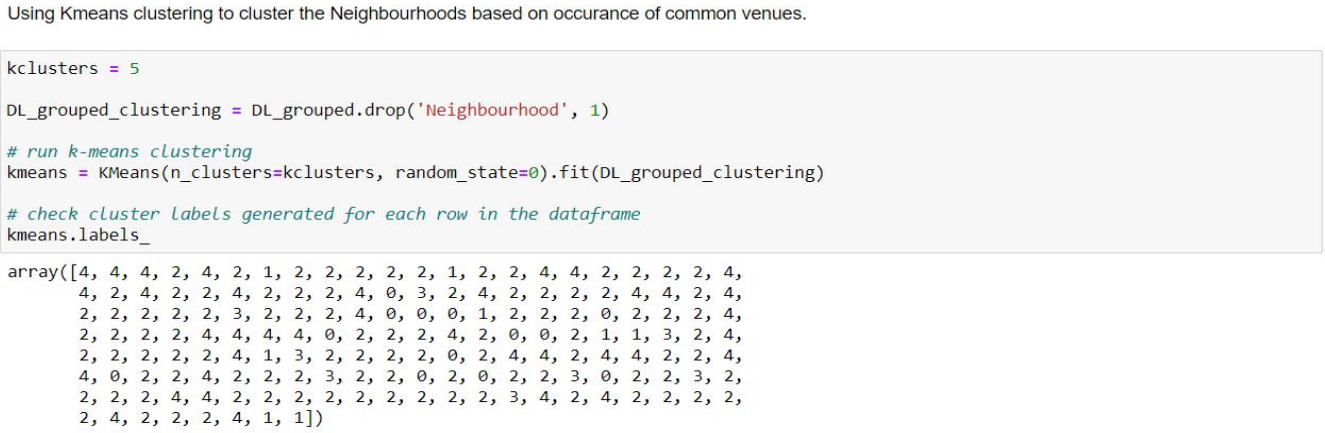
5 rows × 175 columns

One-hot encoding venues for Data modelling



	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Adarsh Nagar	Indian Restaurant	Women's Store	Food & Drink Shop	Garden	Gaming Cafe
1	Alaknanda	Indian Restaurant	BBQ Joint	Pizza Place	Steakhouse	Middle Eastern Restaurant
2	Anand Vihar	Indian Restaurant	Soup Place	Clothing Store	Pizza Place	Furniture / Home Store
3	Ashok Nagar	Fast Food Restaurant	Ice Cream Shop	North Indian Restaurant	Metro Station	Women's Store
4	Ashok Vihar	Indian Restaurant	Diner	Bakery	Food & Drink Shop	Garden

Neighbourhoods with top 5 most common venues

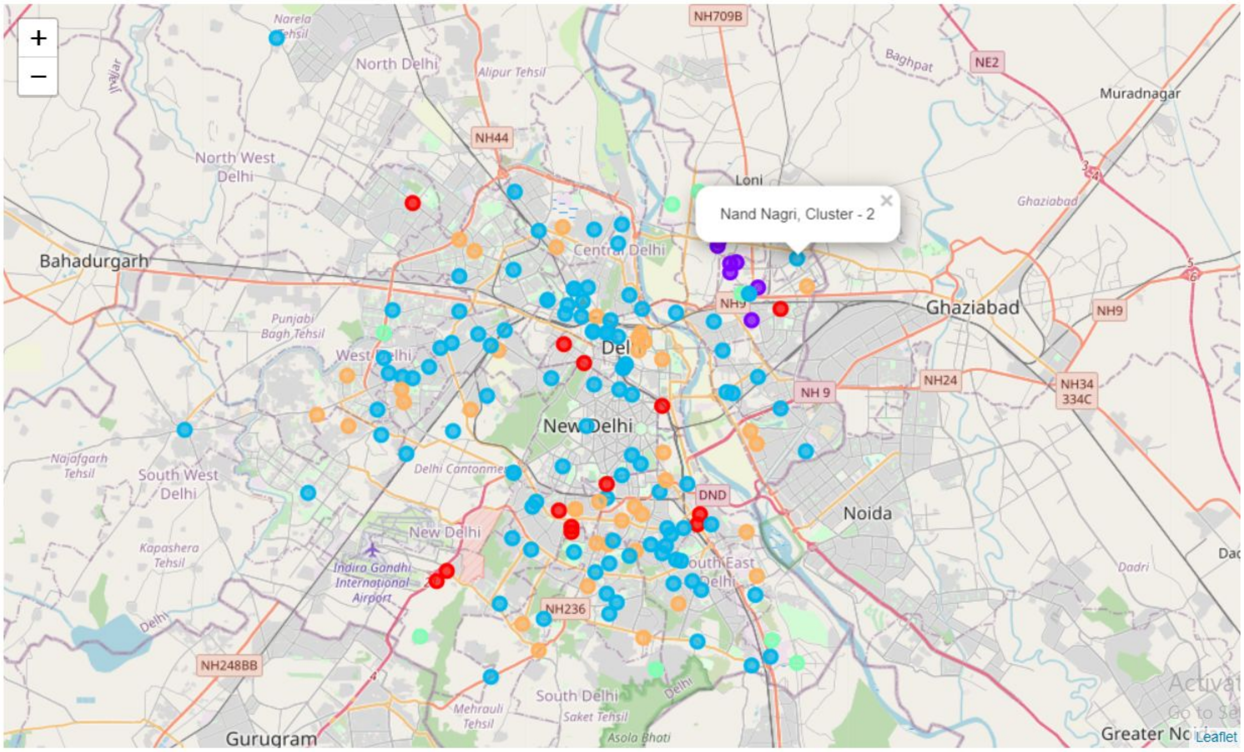


Using Kmeans clustering to cluster the Neighbourhoods based on occurance of common venues.

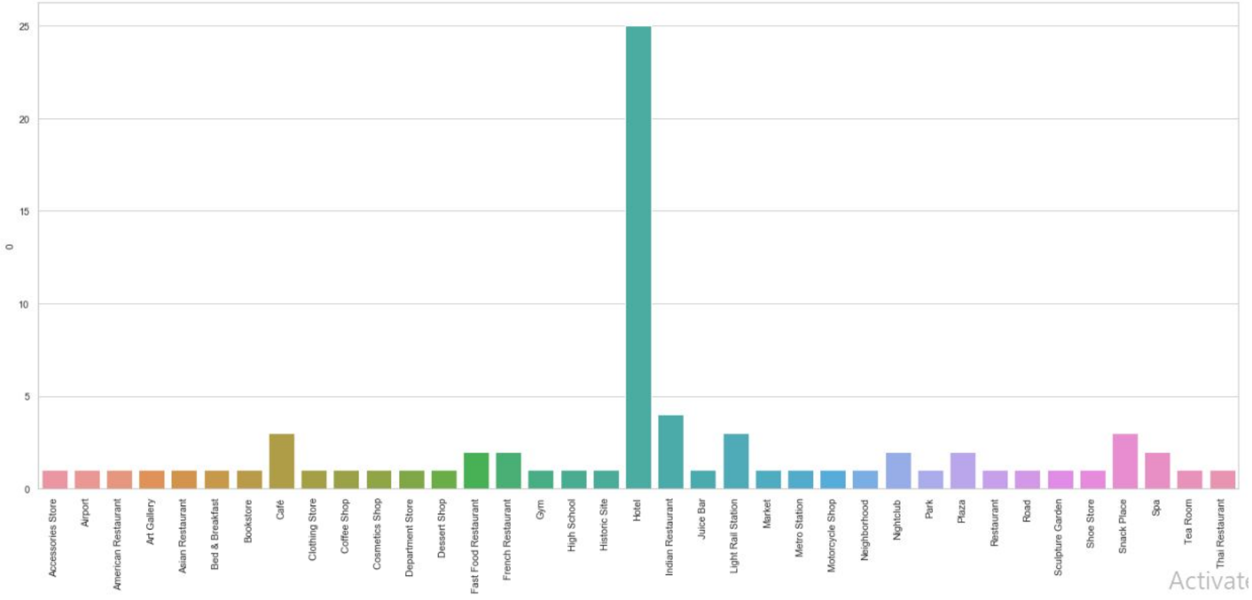
```
kclusters = 5

DL_grouped_clustering = DL_grouped.drop('Neighbourhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(DL_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_
```

```
array([4, 4, 4, 2, 4, 2, 1, 2, 2, 2, 2, 2, 1, 2, 2, 4, 4, 2, 2, 2, 2, 4,
       4, 2, 4, 2, 2, 4, 2, 2, 2, 4, 0, 3, 2, 4, 2, 2, 2, 2, 4, 4, 2, 4,
       2, 2, 2, 2, 2, 3, 2, 2, 2, 4, 0, 0, 0, 1, 2, 2, 2, 0, 2, 2, 2, 4,
       2, 2, 2, 2, 4, 4, 4, 4, 0, 2, 2, 2, 4, 2, 0, 0, 2, 1, 1, 3, 2, 4,
       2, 2, 2, 2, 2, 4, 1, 3, 2, 2, 2, 2, 0, 2, 4, 4, 2, 4, 4, 2, 2, 4,
       4, 0, 2, 2, 4, 2, 2, 2, 3, 2, 2, 0, 2, 0, 2, 2, 3, 0, 2, 2, 3, 2,
       2, 2, 2, 4, 4, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 4, 2, 4, 2, 2, 2, 2,
       2, 4, 2, 2, 2, 4, 1, 1])
```

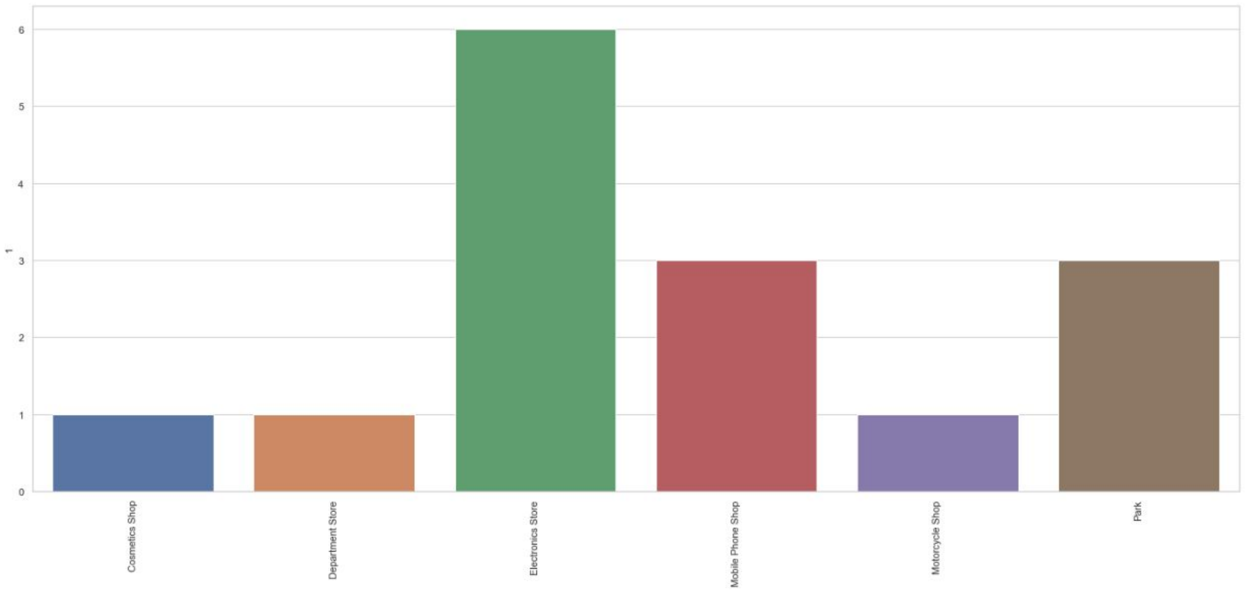
Clustering of data using K-Means clustering



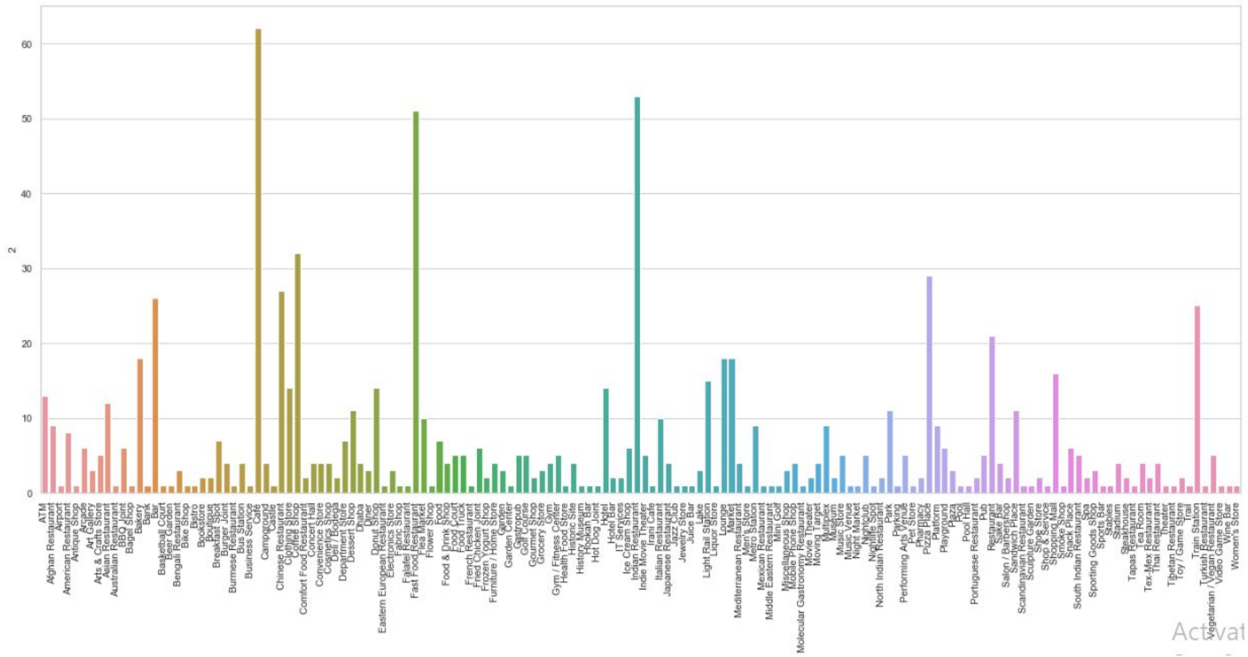
Clustered Neighbourhoods mapped on Delhi map



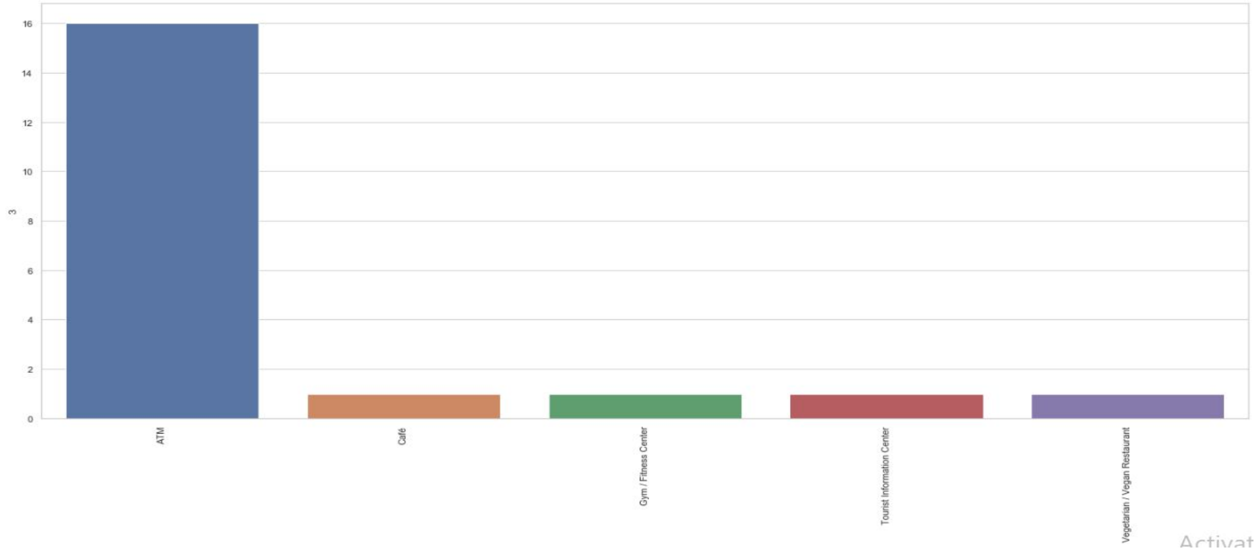
Cluster - 1



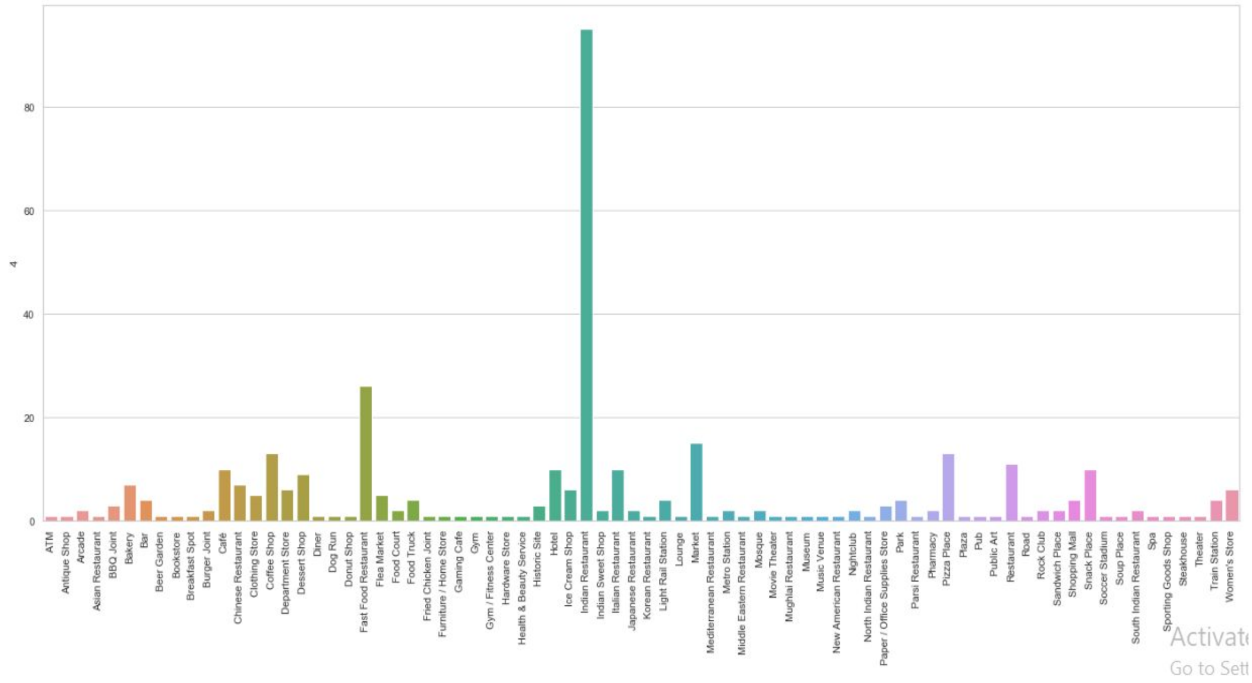
Cluster - 2



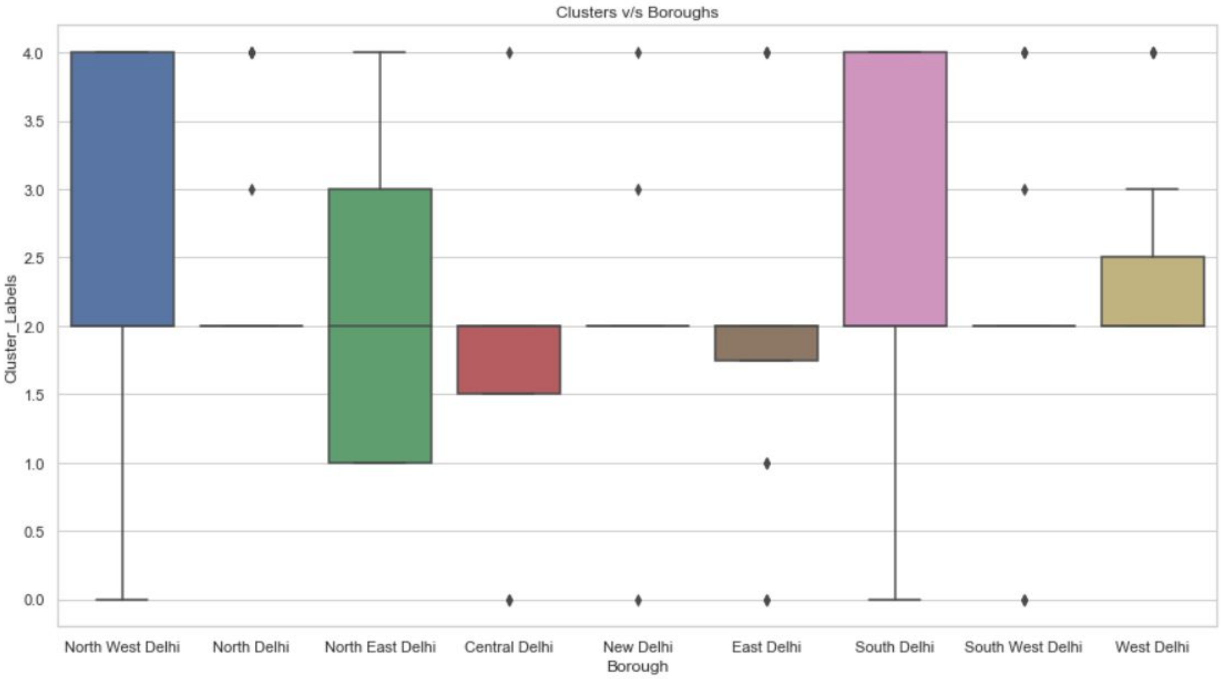
Cluster - 3



Cluster - 4

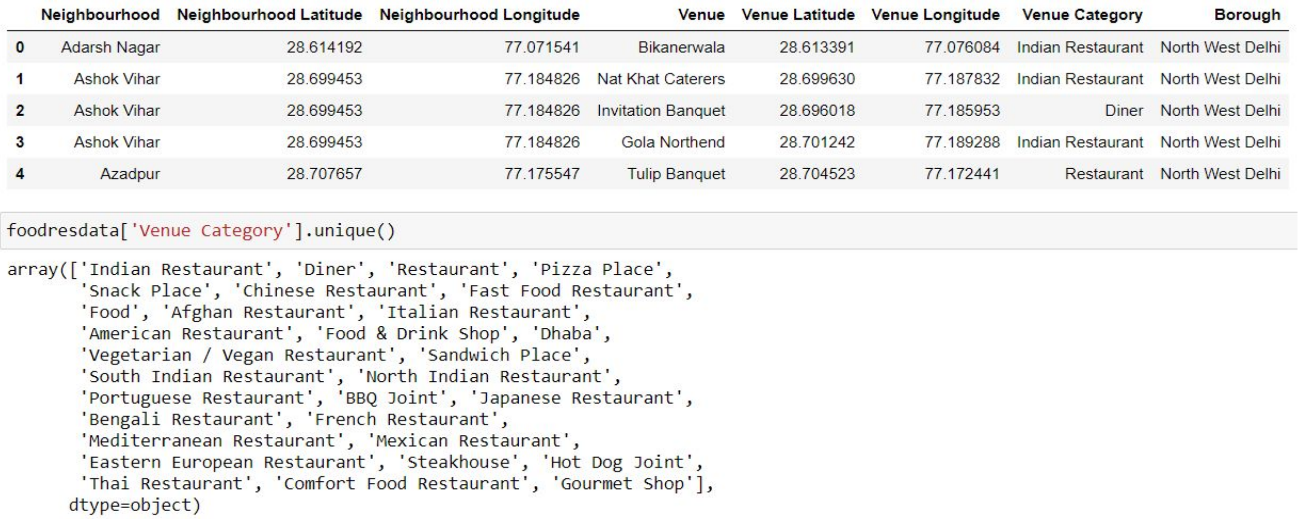


Cluster - 5



Clusters distribution among the 9 Boroughs

Part - 2 : Recommending Neighbourhoods for new Indian Restaurants

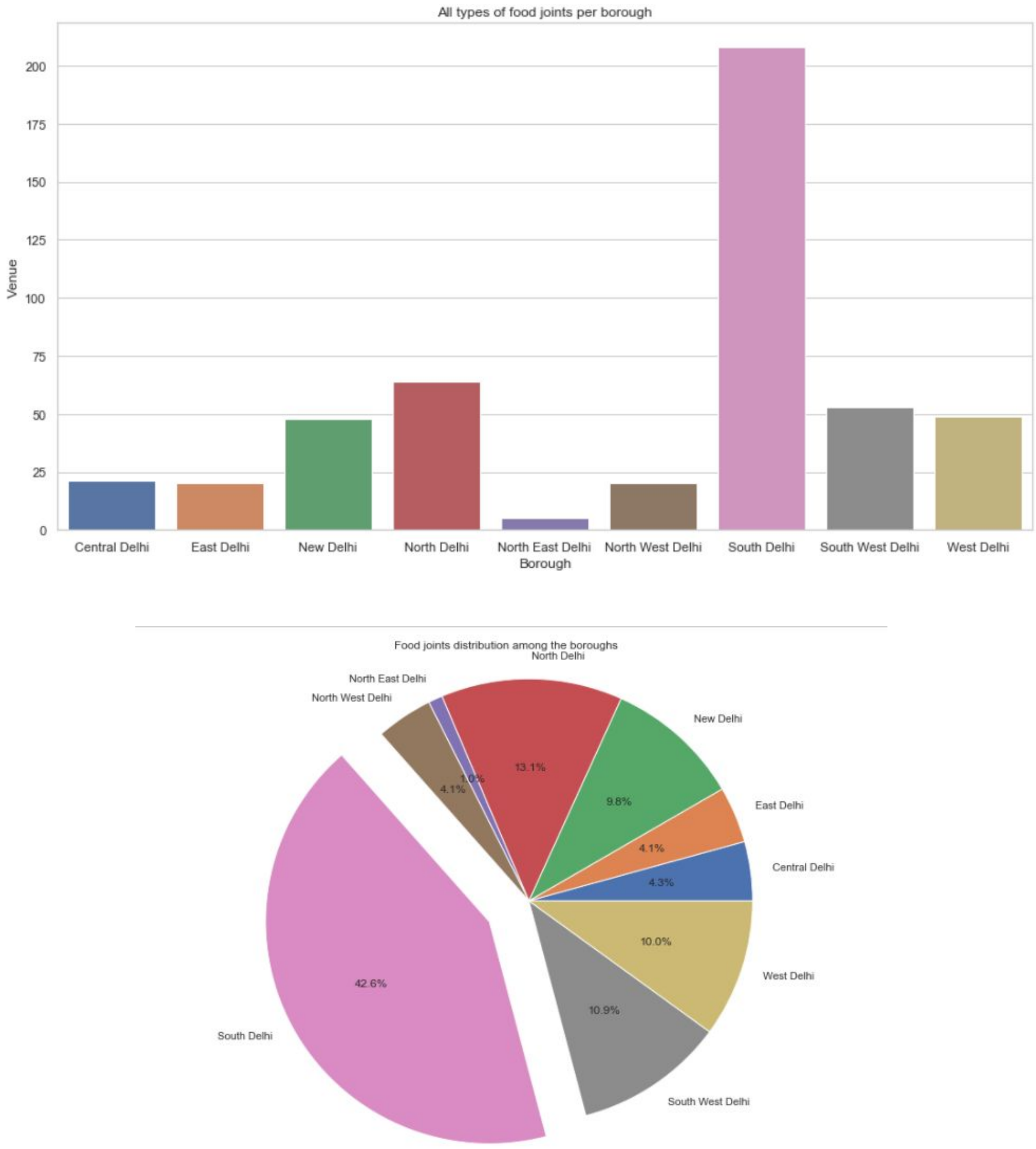


	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Borough
0	Adarsh Nagar	28.614192	77.071541	Bikanerwala	28.613391	77.076084	Indian Restaurant	North West Delhi
1	Ashok Vihar	28.699453	77.184826	Nat Khat Caterers	28.699630	77.187832	Indian Restaurant	North West Delhi
2	Ashok Vihar	28.699453	77.184826	Invitation Banquet	28.696018	77.185953	Diner	North West Delhi
3	Ashok Vihar	28.699453	77.184826	Gola Northend	28.701242	77.189288	Indian Restaurant	North West Delhi
4	Azadpur	28.707657	77.175547	Tulip Banquet	28.704523	77.172441	Restaurant	North West Delhi

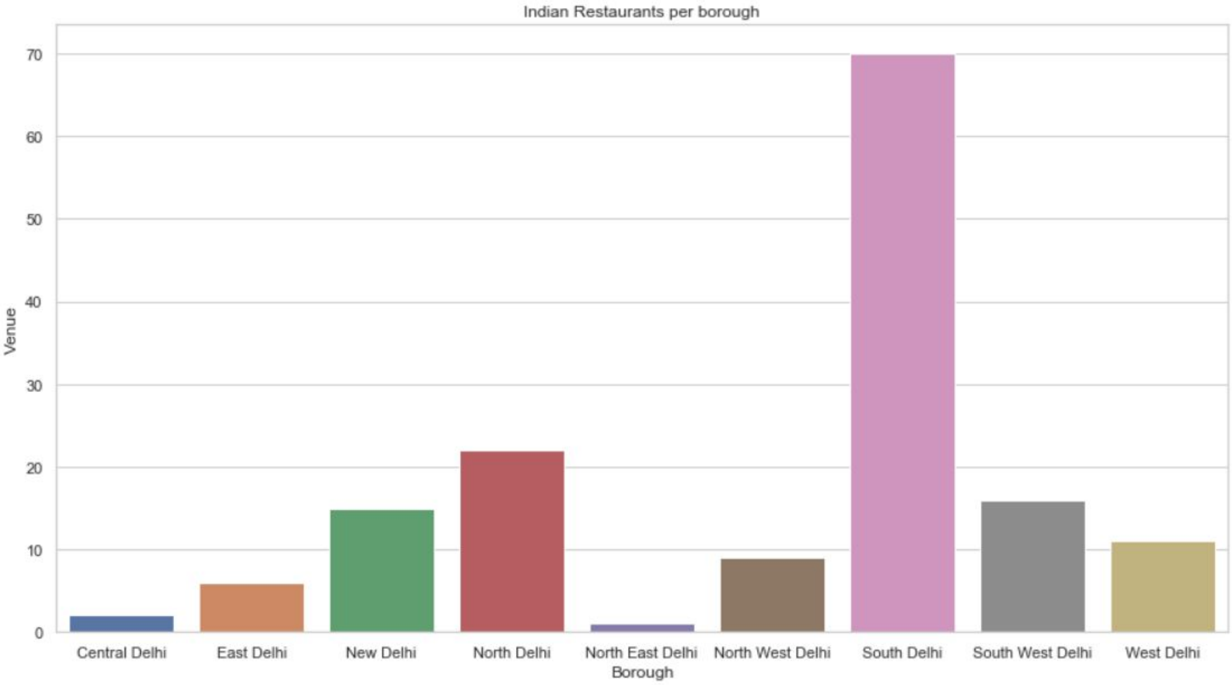
```
foodresdata['Venue Category'].unique()
```

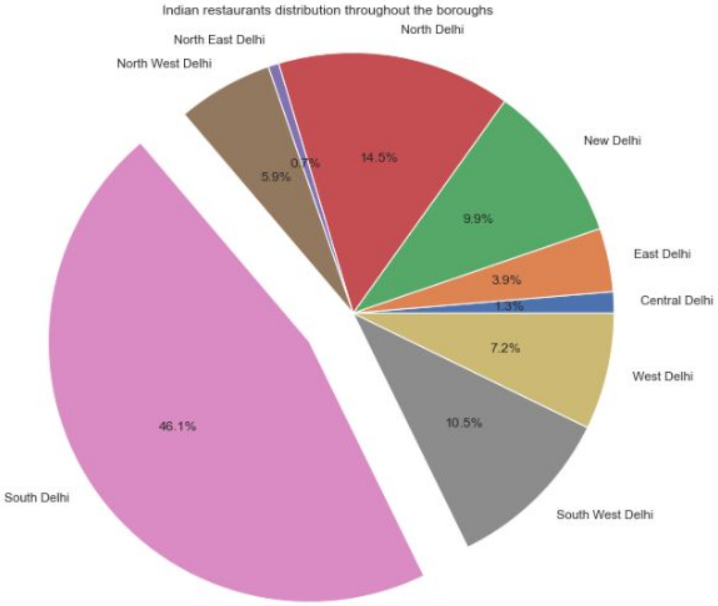
```
array(['Indian Restaurant', 'Diner', 'Restaurant', 'Pizza Place',
       'Snack Place', 'Chinese Restaurant', 'Fast Food Restaurant',
       'Food', 'Afghan Restaurant', 'Italian Restaurant',
       'American Restaurant', 'Food & Drink Shop', 'Dhaba',
       'Vegetarian / Vegan Restaurant', 'Sandwich Place',
       'South Indian Restaurant', 'North Indian Restaurant',
       'Portuguese Restaurant', 'BBQ Joint', 'Japanese Restaurant',
       'Bengali Restaurant', 'French Restaurant',
       'Mediterranean Restaurant', 'Mexican Restaurant',
       'Eastern European Restaurant', 'Steakhouse', 'Hot Dog Joint',
       'Thai Restaurant', 'Comfort Food Restaurant', 'Gourmet Shop'],
      dtype=object)
```

Food restaurants data frame

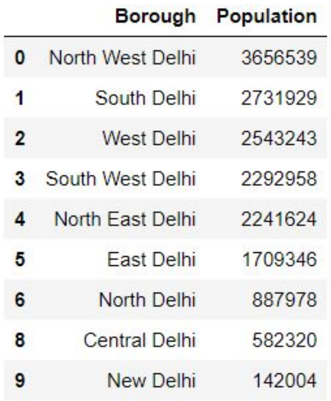


Food Joints distribution throughout the Boroughs of Delhi

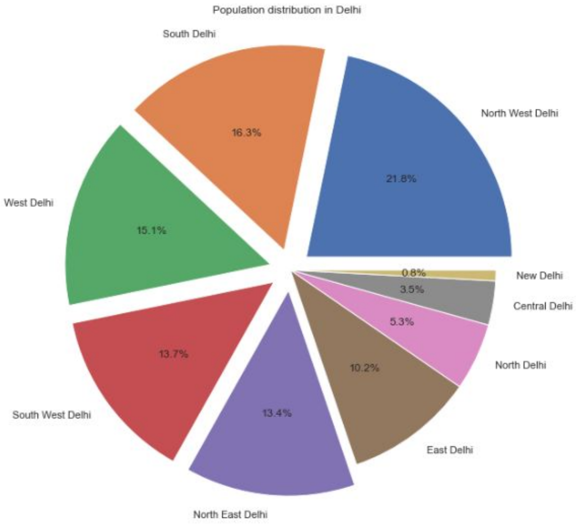




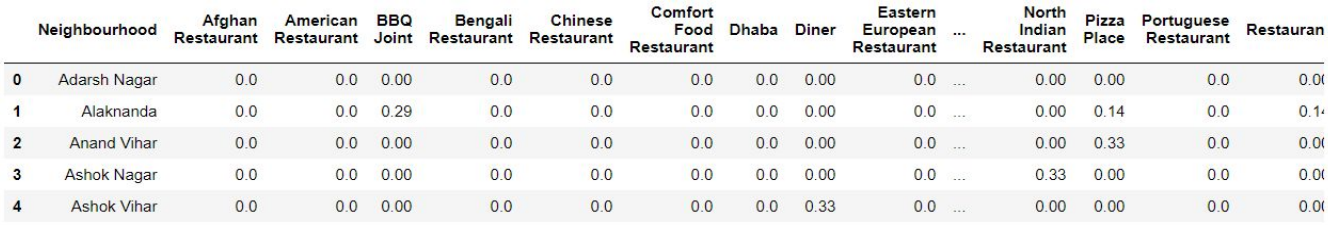
Indian Restaurants distribution throughout the Boroughs of Delhi



	Borough	Population
0	North West Delhi	3656539
1	South Delhi	2731929
2	West Delhi	2543243
3	South West Delhi	2292958
4	North East Delhi	2241624
5	East Delhi	1709346
6	North Delhi	887978
8	Central Delhi	582320
9	New Delhi	142004

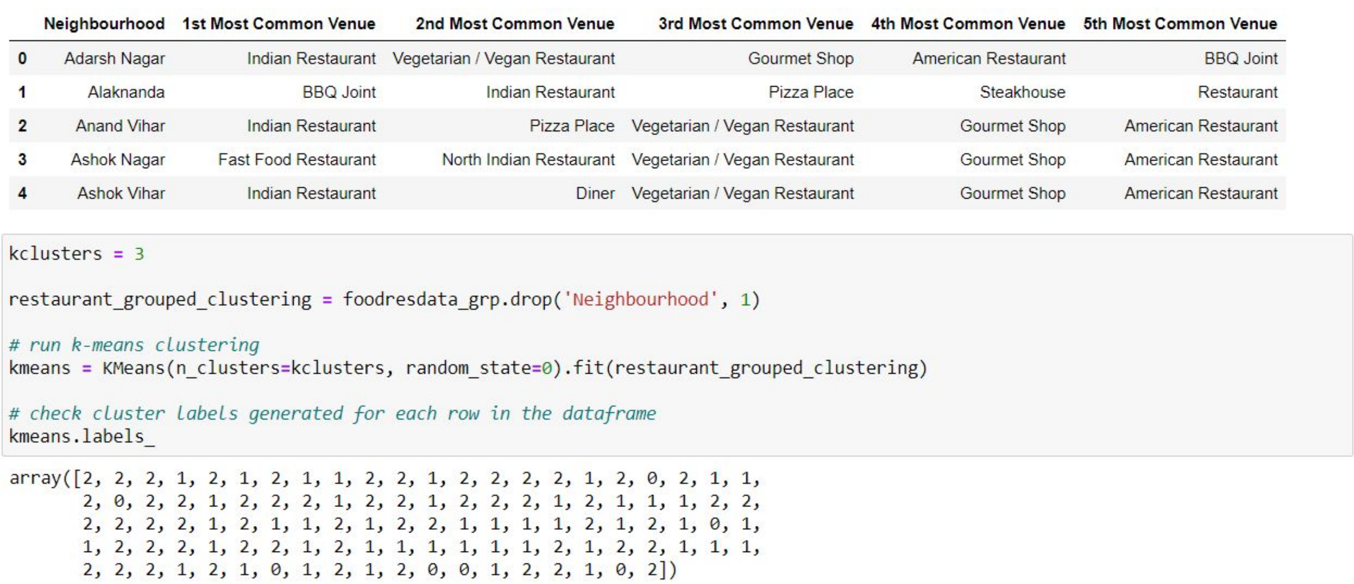


Population Distribution in Delhi



	Neighbourhood	Afghan Restaurant	American Restaurant	BBQ Joint	Bengali Restaurant	Chinese Restaurant	Comfort Food Restaurant	Dhaba	Diner	Eastern European Restaurant	...	North Indian Restaurant	Pizza Place	Portuguese Restaurant	Restauran
0	Adarsh Nagar	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.00	0.0	...	0.00	0.00	0.0	0.0
1	Alaknanda	0.0	0.0	0.29	0.0	0.0	0.0	0.0	0.00	0.0	...	0.00	0.14	0.0	0.1
2	Anand Vihar	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.00	0.0	...	0.00	0.33	0.0	0.0
3	Ashok Nagar	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.00	0.0	...	0.33	0.00	0.0	0.0
4	Ashok Vihar	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.33	0.0	...	0.00	0.00	0.0	0.0

One-hot encoding restaurant data for data modelling

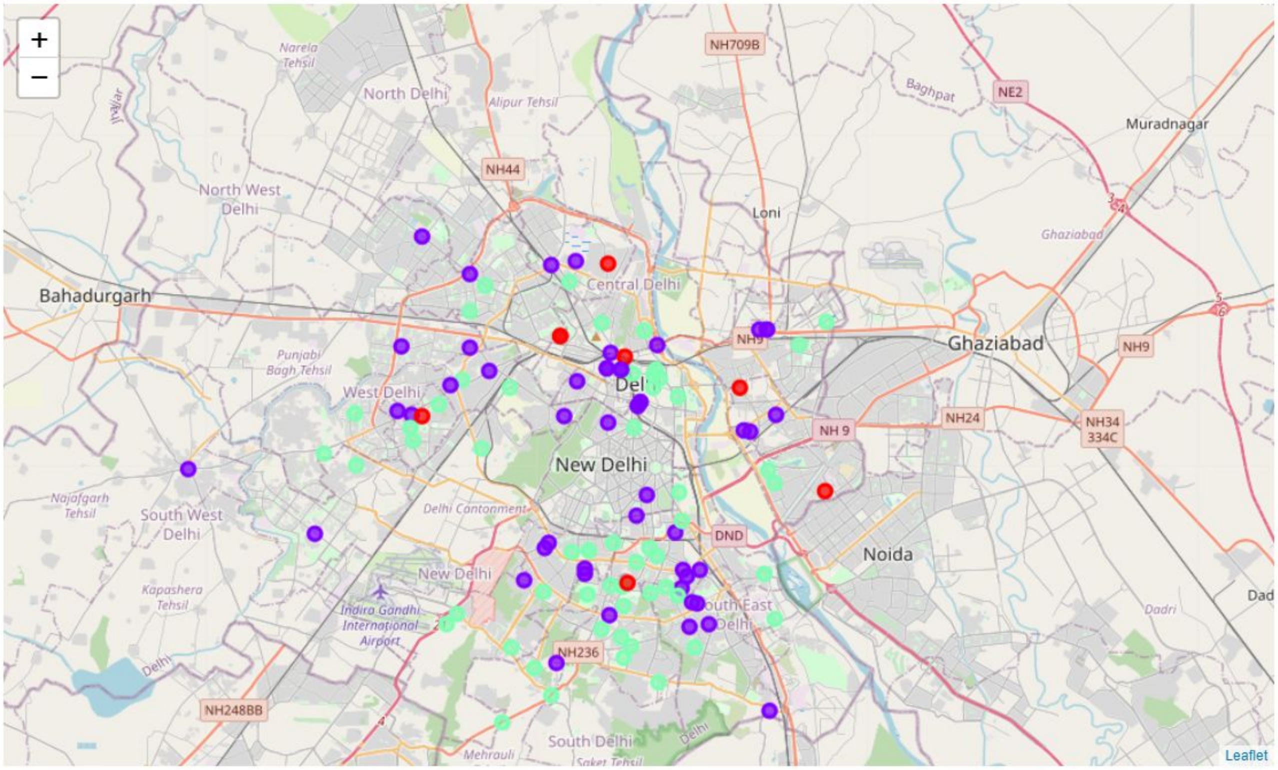


	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Adarsh Nagar	Indian Restaurant	Vegetarian / Vegan Restaurant	Gourmet Shop	American Restaurant	BBQ Joint
1	Alaknanda	BBQ Joint	Indian Restaurant	Pizza Place	Steakhouse	Restaurant
2	Anand Vihar	Indian Restaurant	Pizza Place	Vegetarian / Vegan Restaurant	Gourmet Shop	American Restaurant
3	Ashok Nagar	Fast Food Restaurant	North Indian Restaurant	Vegetarian / Vegan Restaurant	Gourmet Shop	American Restaurant
4	Ashok Vihar	Indian Restaurant	Diner	Vegetarian / Vegan Restaurant	Gourmet Shop	American Restaurant

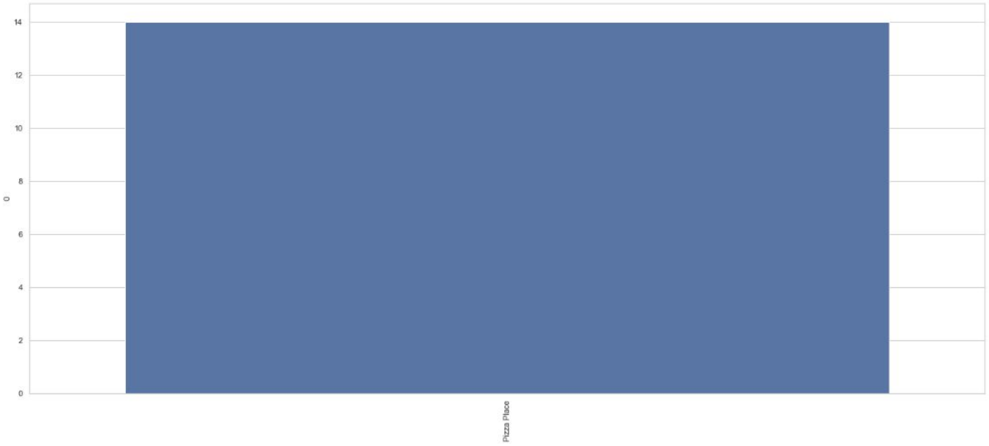
```
kclusters = 3

restaurant_grouped_clustering = foodresdata_grp.drop('Neighbourhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(restaurant_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_
```

```
array([2, 2, 2, 1, 2, 1, 2, 1, 1, 2, 2, 1, 2, 2, 2, 2, 1, 2, 0, 2, 1, 1,
       2, 0, 2, 2, 1, 2, 2, 2, 1, 2, 2, 1, 2, 2, 2, 1, 2, 1, 1, 1, 2, 2,
       2, 2, 2, 2, 1, 2, 1, 1, 2, 1, 2, 2, 1, 1, 1, 1, 2, 1, 2, 1, 0, 1,
       1, 2, 2, 2, 1, 2, 2, 1, 2, 1, 1, 1, 1, 1, 1, 2, 1, 2, 2, 1, 1, 1,
       2, 2, 2, 1, 2, 1, 0, 1, 2, 1, 2, 0, 0, 1, 2, 2, 1, 0, 2])
```

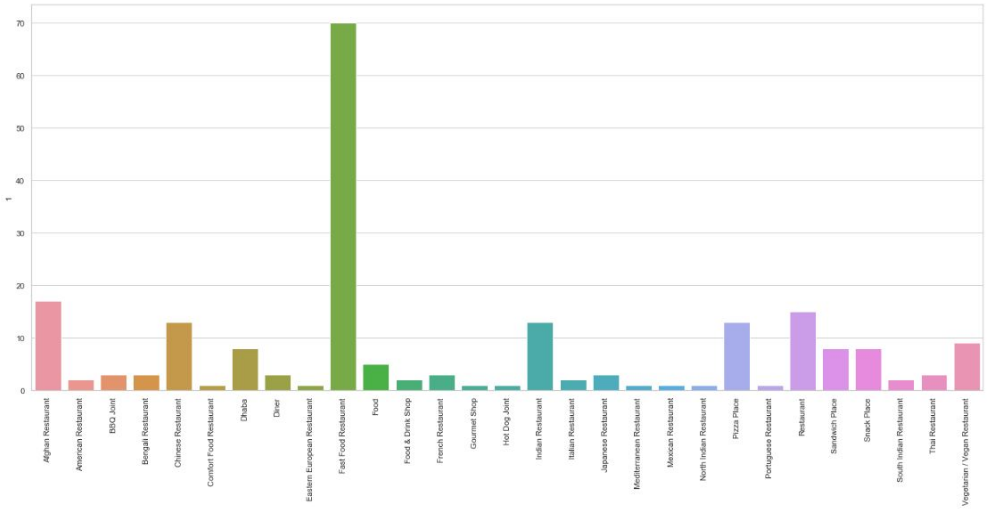
Data frame with Top 5 most common restaurant categories & Clustering data into 3 clusters



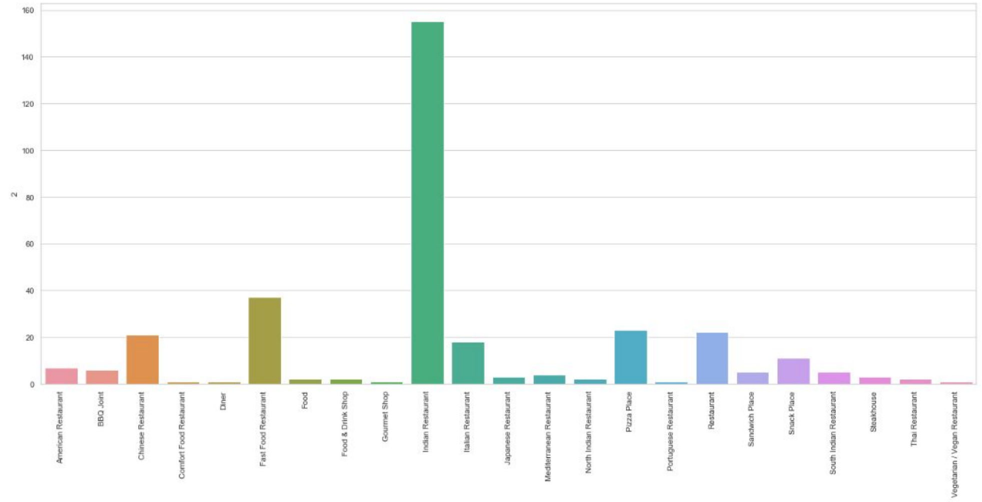
Clustered restaurant data mapped on Delhi map



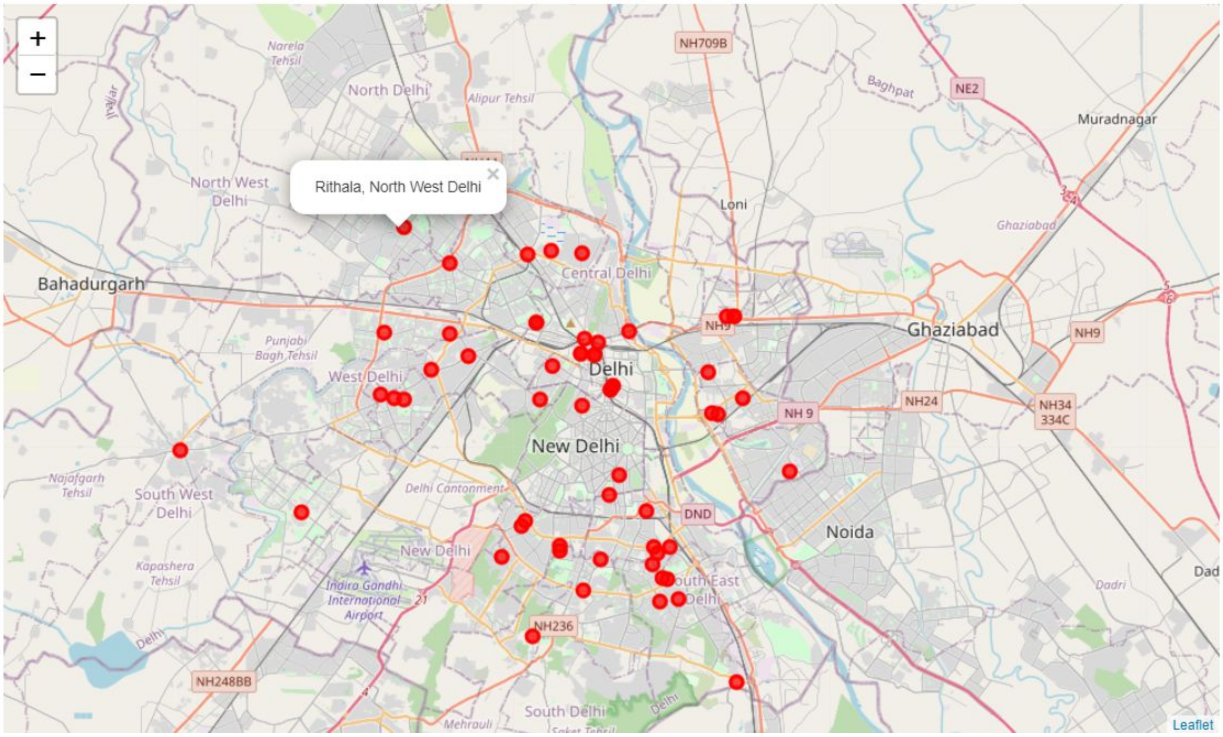
Cluster - 1



Cluster - 2



Cluster - 3



Recommended Neighbourhoods for Indian restaurants mapped on Delhi map

5. Results and Discussions

From the above analysis, we have found some exciting results. We took 185 Neighbourhoods in Delhi into account to firstly, explore the Neighbourhoods of Delhi by looking into its venues and places across different Boroughs. Secondly, we narrowed our analysis to look into the food venues in Delhi and how the distribution of total venues affect the food venues around different Boroughs as well.

We observed that although 5 Boroughs out of 9 accounts for more than 70% of the population but the same cannot be said about how the venues are distributed in Delhi. South Delhi was the one Borough which held most of the venues and the same was inferred from Food joints as well as Indian Restaurants as well.

We also observe that cluster-3(blue), which includes most of the neighbourhoods, reflects that a considerable amount of neighbourhoods are similar in terms of the Venue categories and venues each of them offers.

Below are the most occuring venues in each venue cluster and how we might label each cluster as:

- Cluster-1 : Hotels
- Cluster-2 : Electronic and Mobile stores
- Cluster-3 : Multiple Venue categories
- Cluster-4 : ATMs
- Cluster-5 : Indian Restaurants

Below are the most occuring restaurants in restaurant cluster and how we might want to label them:

- Cluster-1 : Pizza Places
- Cluster-2 : Fast Food Restaurants
- Cluster-3 : Indian Restaurants

6. Conclusions

From this project, after analysing 9 Boroughs with 185 Neighbourhoods and 1500+ venues we can conclude the following points which certainly depends on the limitation of Foursquare exploration of venues as well as the limitation of some venues in Delhi which are either small-scale businesses or are unregistered due to various reasons.

The pointers are as below:

- South Delhi Borough is densely populated with venues and has a lot of competition for Indian Restaurants in comparison to other Boroughs
- Most Boroughs offers a variety of venue categories and have shown signs of similarities when compared in terms of clusters formed
- Cluster 3 is the most occuring cluster labels which contain multiple venues and indicates that most Neighbourhood are equipped with densely populated multiple venues, which in turns

reflects Delhi's prosperity(2nd wealthiest city in India) as well as its high GDP per capita(2nd highest in India)

- The above map shows Neighbourhoods recommended by us for new Indian Restaurant in which we would further recommend Boroughs other than South Delhi, keeping in mind the population distribution in Delhi as well. Although, the large number of Indian Restaurants in South Delhi can also firmly mean that the restaurants here are a success, business wise, but we should keep in mind that it also offers a great amount of competition for new restaurants
- Also, we were successfully able to cluster food joints into broad categories i.e Pizza Places, Fast Food restaurants and Indian Restaurants which indicated the distribution of food joints all over Delhi.