****

**Opening a Restaurant in Delhi**

**AMIT KUMAR BEURA**

**February 08,2020**

**1. INTRODUCTION**

Delhi is the capital city of India and is known as the “Heart of the Nation”. It has a population of 11 million and metropolitan population of 16.3 million. This makes city the second most populous city in India. It covers an area of 573 square miles (1,480 km). Speaking about the variety of cuisines, one can easily get all variety of food items which will definitely follow their taste; all you have to do is to find out the right place. From a historic restaurant to a modernly outlined place serving the best food items are easy to find in India’s Capital Territory.

**1.1 Business Problem and Discussion of Background**

With such huge population and people from all of the country of different culture, one find its difficult to decide as where to open a restaurant.

In the Delhi Perspective, we will try to find out some of the issues listed below:

1. Which areas are more populated.

2. Comparing the different areas of Delhi.

3. The most Suitable location of having Restaurant.

4. The area which has the maximum number of restaurants.

* 1. **Beneficiaries**

1. All the Businessmen who are ready to invest their money in opening restaurant
2. All the general public and the tourists who wants to change their taste and try out different cuisines.
3. The jobseekers who can easily find out the work in the busiest areas of Delhi
4. The Marketing people who can have their big impact on public using the information after comparing cities.

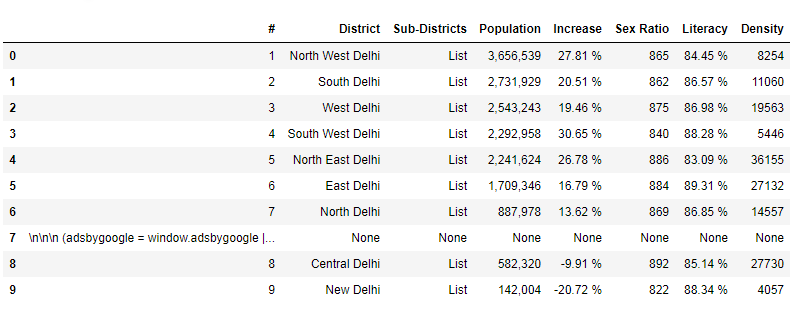
**2. DATA**

We will be using data as per:

1. Web scraping data from wikipedia of districts of delhi.
2. Web scraping data from other websites of delhi having population in it.
3. Using Geocoder package to get the latitudes and longitudes.
4. Using FourSquare location to fetch the details of venues nearby.

**2.1 Data preparation through web scraping**

At first we need to collect the information regarding all areas of Delhi. For that purpose,we will obtain the data by web scraping. The website that we have consult is <https://www.census2011.co.in/census/state/districtlist/delhi.html> . It contains the data of census 2011 of Delhi. After web scraping , we obtain the resulting dataframe.

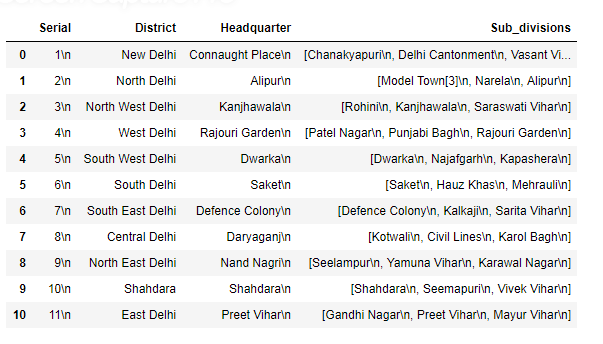
****

**Data cleaning:**

Our dataframe needs to get rid of various rows and columns in order to get the cleaned dataframe. The row having index 7 needs to be dropped as it contains the none type of data. The another column that needs to be dropped is of Sub-districts. It contains all the elements of list but the list does not solve any of our purpose.

The sub districts will be obtained using another website from Wikipedia. So, we will remove it. The “#” column containing the serial number is not required, we are having index number for it.

Next thing that we have to check is the data types of the dataframe . It is to be noted that we have to take the numerical datatype for the features population, increase,sex-ratio, literacy and density. The object datatype will be of District .

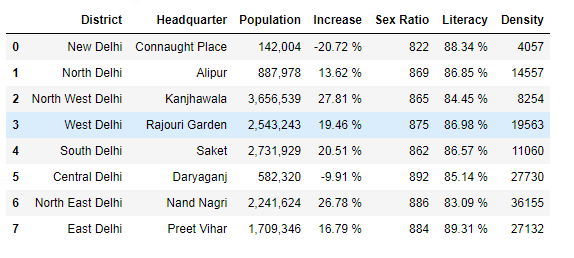
We need to know the sub districts of the delhi, for which we have consult the following website <https://en.wikipedia.org/wiki/List_of_districts_of_Delhi> . this website is scrapped and we obtain the another dataframe:

**Data cleaning:** The second dataframe is obtained contains the headquarter. Since Subdivisions and Serial column is not required, we are going to drop it.

The Headquarter column contains “\n|” which were extracted during scraping. They need to be removed.

**2.2Data Merging:**

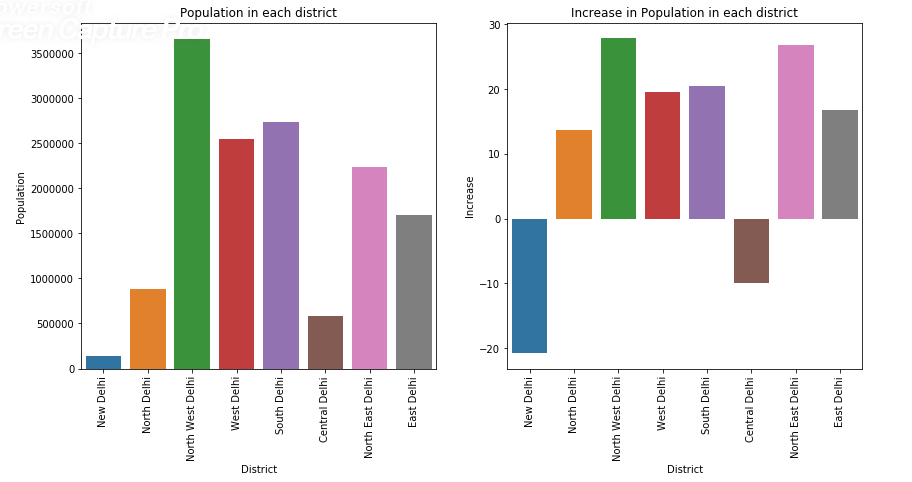
Our both the dataframes are ready and cleaned and we just need to merge them together. It should be kept in mind that both the tables have same index numbers. The tables that have the data in common is taken. We will reset the index and using merge option, we are going to merge the tables. After merging our dataframe df\_delhi is ready.



Finally, our dataframe with all the features required has been scraping.

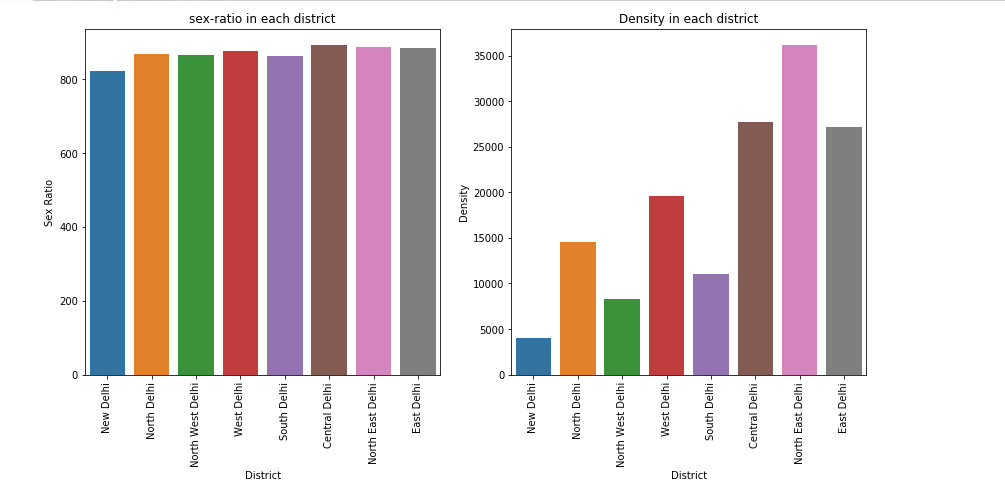
**3. DATA VISUALIZATION**:

The features are analyzed using bar plots. The bar plots are analyzed taken each of the districts and population, increase , sex ratio and density.



Looking at the bar plots, The population is maximum in the North-West Delhi having Headquarters at “**Kanjhawala”.**

The data is of 2011 census, the population has increased in the North-East Delhi having headquarters at “**Kanjhawala**”.



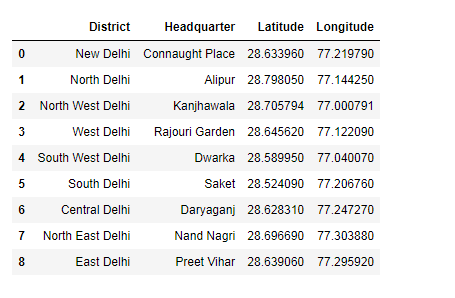
The sex-ratio is almost same in all the districts.

The density (number of people per land area) is maximum in Northeast delhi having its headquarters at **‘Nand-Nagri**”

We can explore the Delhi considering Kanjhawala as it is the highest populated district in all over Delhi. But first we need to get the latitudes and longitudes.

**Getting Latitudes and Longitudes:**

The geocoder package is used to obtain the latitudes and longitudes of all the districts. The obtained dataframes is as follows:



**FourSquare API :**

The four square api is used to get the 100 venues around these districts within a radius of 1000m.

**Single Neighborhood:**

We’ve chosen the **“Kanjhawala**” headquarter to get its venues as it has the highest population. But we found that there are no venues found in the area of 1km.

**Multiple Neighborhood:**

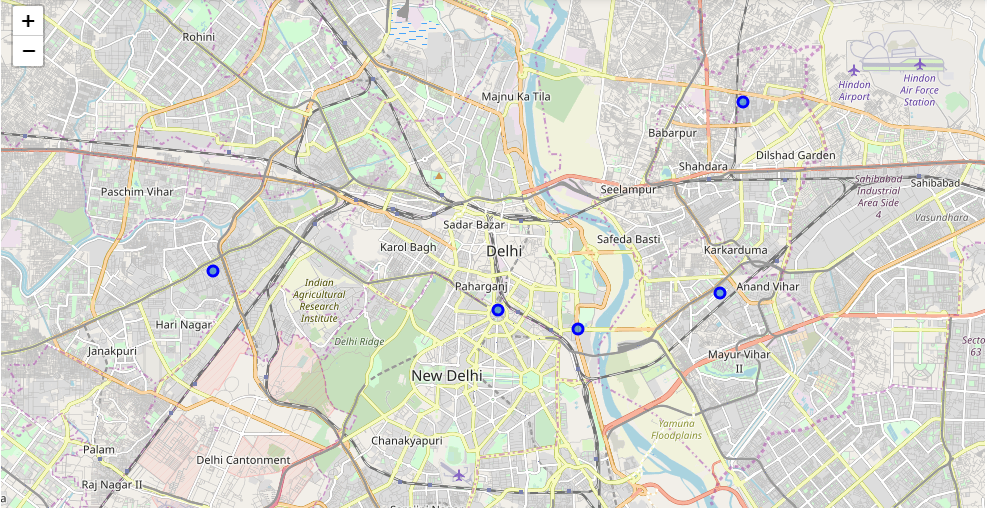
**Now the Battle of Neighborhoods begins:**

So, we will take all districts’ Headquarters to find which neighborhood has maximum number of venues within 1000metres.

**4. MAP VISUALIZATION:**

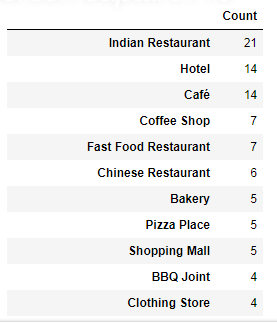
The folium Map library will be used to visualize the map of Delhi. We will define the latitudes and longitudes of Delhi and using information from our dataframe having latitudes and longitudes of all the districts, we are able to plot it.

**4.1 Districts of Delhi**



Now when we have obtained the venues using the four location data. We are interested in finding out the unique categories.

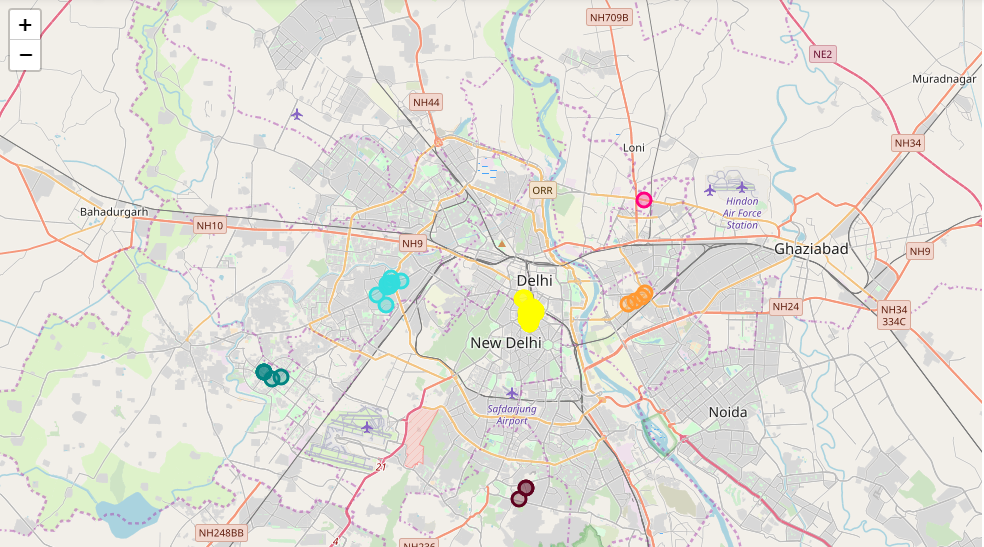
There are 77 unique categories found in the nearby Headquarters of Delhi.



There are maximum number of Indian Restaurants found in the nearby areas.

Lets figure out only the restaurants searched in the venue category. Now that we have all the restaurants found in all the district’s Headquarters , lets superimpose them on the map.

* 1. **Restaurants In all districts:**



If we visualize the map carefully, we will find that maximum number of Restaurants are shown by the yellow marker and which is “Connaught Place” of Delhi.

**5. CLUSTERING**

We will cluster the data obtained of all nearby venues with respect to each of its headquarters.

### Optimal Number of Clusters for K-means

To get the optimal number of clusters to be used for the K-mean, there are a number ways possible for the evaluation. Therefore, in this task, the following are used:

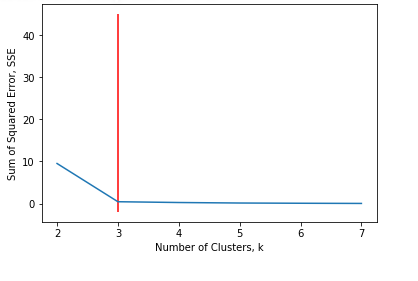
1.Elbow (Criterion) Method

2.Silhouette Coefficient

1.**Elbow Method**

The elbow method is used to solve the problem of selecting k. Interestingly, the elbow method is not perfect either but it gives significant insight that is perhaps not top optimal but sub-optimal to choosing the optimal number of clusters by fitting the model with a range of values for k.

The approach for this is to run the k-means clustering for a range of value k and for each value of k, the Sum of the Squared Errors (SSE) is calculated., calculate sum of squared errors (SSE). When this is done, a plot of k and the corresponding SSEs are then made. At the elbow (just like arm), that is where the optimal value of k is. And that will be the number of clusters to be used. The whole idea is to have minimum SSE.



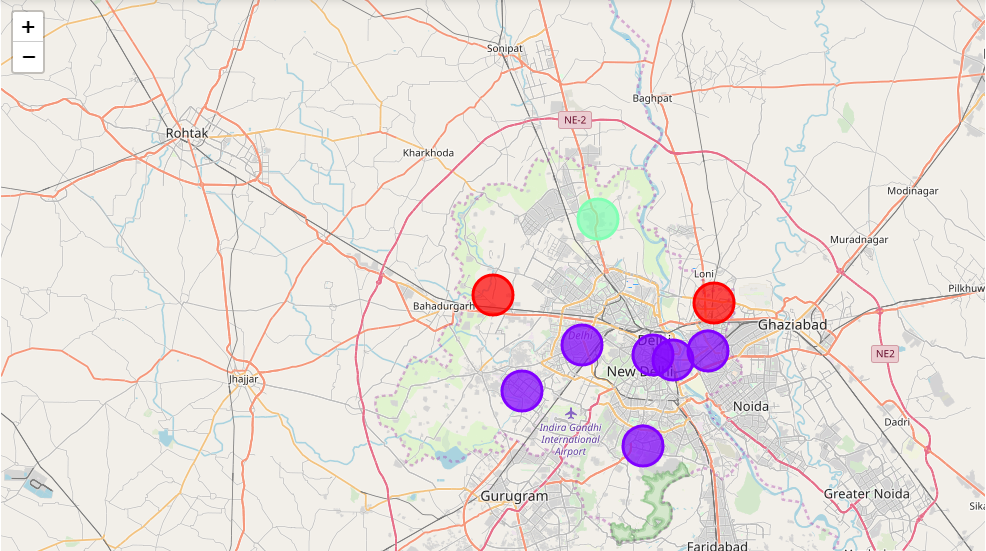
**2.Silhouette Coefficient**

To find the optimal value of the number of clusters, k, the number of clusters is iterated corresponding Silhouette Coefficient’s calculated for each of the k-values used. The highest Silhouette Coefficient gives the best match to its own cluster. Please see below:

From the result, the high the n\_clusters the better the silhouette coefficient.

For this project, a cluster value of 3 will be used.

After clustering, the data is shown on a map:

****

**6. RESULT:**

The following are the highlights obtained from above 3 clusters and the data analysis:

1. It is surprising to know that despite being the most populated area of delhi, we are not able to find any common venue near Kanjhawala.

2. Out of the all districts of Delhi, Indian Restaurant is found in the first venue category.

3. There are number of eating points in all districts than any other venues.

4. Although the clusters have variation, but the predominance of Indian restaurant was found 50%.

5. The number of Indian Restaurants were found at “Connaught Place”

**7. DISCUSSION AND CONCLUSION:**

It will be beneficial to open a restaurant at **Alipur, North delhi** as this will give least competition and does not have any restaurants nearby.

We were not able to find any venues nearby Kanjhawala , the reason is that Kanjhawala is a village and has very less amenities within 1kms.

Some drawback of these analyses is that clustering is done on only the most common venues which are obtained using four square location. Also, we have taken into consideration one city under each district .Furthermore, the result can vary if we use other clustering techniques like DBSCAN.

In conclusion, this project would have had better results if there were more data in terms of crime data within the area, price of the land, traffic access and allowance of more venues exploration with the Foursquare (limited venues for free calls).Also, getting the ratings and feedbacks of the current restaurants within the clusters would have helped in providing more insight into the best location.