The Battle of Neighbourhoods

Capstone Project

Introduction/Business Problem :

Because Bangalore is known as silicon valley of India, So many people from different parts of country moves to Bangalore but India is very big and Diverse every part of a country is different from each other be it language, culture, food or anything else. A large number of people who move to Bangalore are from New Delhi and nearby area (NCR, UP, Haryana, Punjab, Rajasthan). They know New Delhi very well but Bangalore is totally different. So we will try to cluster New Delhi and Bangalore Based on similarity. That is we will cluster New Delhi on the basis of Neighbourhood and cluster Bangalore on the basis of Neighbourhood and will tell which places of New Delhi is similar to which places of Bangalore. This will help not only the new people moving to Bangalore but also those who are already living in Bangalore but are looking for change. So anyone who wants to shift to Bangalore from North India might be interested in this project

Data :

We Scrapped Wikipedia to get Neighbourhood names of New Delhi and Bangalore. we get the list of district of New Delhi and their sub district and one more page for the Neighbourhood of New Delhi. speaking of Bangalore we got a Wikipedia page about Bangalore neighbourhood which is just the name of areas of Bangalore. Moreover we will use foresquare api to get the information about various amenities or facilities available in these places based on which we will cluster theses cities. We will clustered New Delhi then predict on Bangalore to get the similarity between these two cities.

Following is how our data looks like :

```
In [4]: delhi.head()
```

Out[4]:

	Key	District	Neighbourhood	latitude	longitude
0	0	New Delhi	Connaught Place	28.6315	77.2167
1	1	New Delhi	Chanakyapuri	28.5939	77.1887
2	2	New Delhi	Delhi Cantonment	28.5961	77.1587
3	3	New Delhi	Vasant Vihar	28.5603	77.1617
4	4	North Delhi	Narela	28.8540	77.0918

```
In [8]: banglore.head()
```

Out[8]:

	Key	District	Neighbourhood	latitude	longitude
0	2	Central	Domlur	12.962467	77.638196
1	3	Central	Indiranagar	12.973291	77.640467
2	4	Central	Jeevanbheemanagar	12.962900	77.659500
3	5	Central	Malleswaram	13.016341	77.558664
4	7	Central	Rajajinagar	12.990100	77.552500

About Columns :

In both of the Dataframes Key is a column to uniquely identify a particular tuple in a Dataframe, District column is only the area where the place is present for example Connaught Place is in New Delhi, and Indiranagar is in Central Bangalore. Neighbourhood is the place which we are concerned about and latitude and longitude are there coordinates respectively.

Sources of DATA :

following are the links which we have used in order to get the data :

https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi

https://en.wikipedia.org/wiki/List_of_districts_of_Delhi

https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore

Foresquare API

Project Methodology :

DATA Collection :

To reach to our goal we needed DATA and we utilized python's web scrapping power in order to get the data from sources like Wikipedia. And then we collected the data from Foresquare API in order to get some venues in our data. We follow following steps in order to collect data :

1. We scrapped the Wikipedia web page with the help of Pandas read_html method
2. We stored the result in a variable which we utilized in order to extract table out from it.
3. We then processed the respective tables to make them in proper format and saved them with the help of csv file.
4. We repeated the same process multiple times to append the csv file with the needed Data.
5. We did this to get different regions of New Delhi and Bangalore
6. Our Data is partially ready, But we are still left with coordinates of respective regions.
7. We utilized geopy library in order to get the respective coordinates of our Neighbourhoods.
8. We stored them in a csv file
9. Due to geopy's Terms and some other reasons we were unable to get all coordinates So we entered them manually
10. We now have another Dataframe which consist of coordinates of Neighbourhoods.

Now We are ready to get the nearby venues of these Neighbourhoods from Foresquare API we did it in following steps :

1. Import the required library to send request and then parse the JSON
2. Prepare the url with client id and secret key
3. Form the request and send it to Foresquare to get the JSON response back
4. Parse the json and store the information in a Dataframe

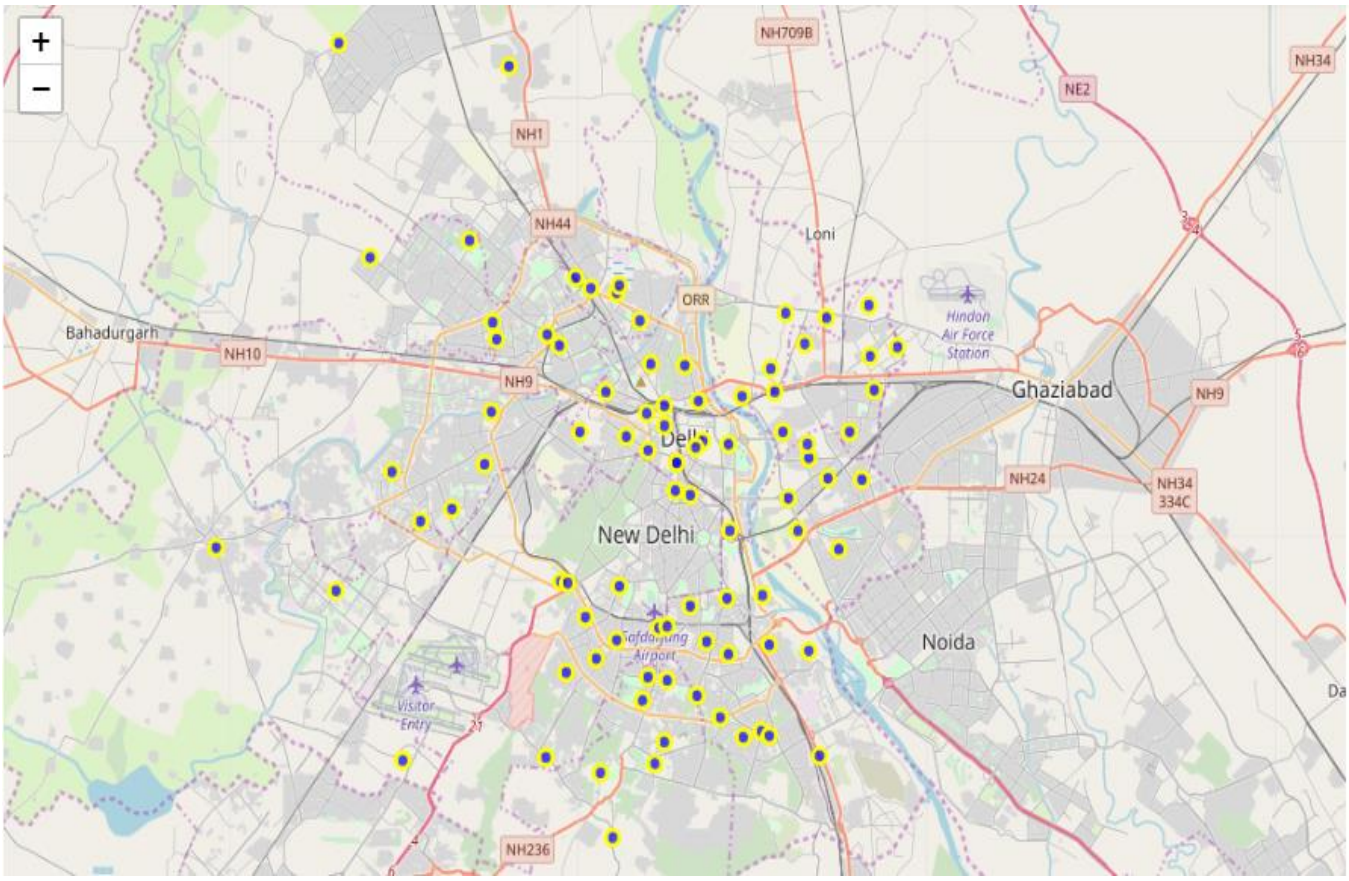
Data Preprocessing :

Before we start clustering we need to make data in proper format that is we need to convert each of the categorical variable into numerical variable and take the mean of frequency of occurrence of each category. So we achieve our this goal as follows :

1. group rows by neighborhood and by take the mean of the frequency of occurrence of each category
2. Sort them in descending order
3. Create a Dataframe with top 10 venues for each Neighborhood
4. Perform One Hot Encoding
5. We repeated the same thing with Bangalore data

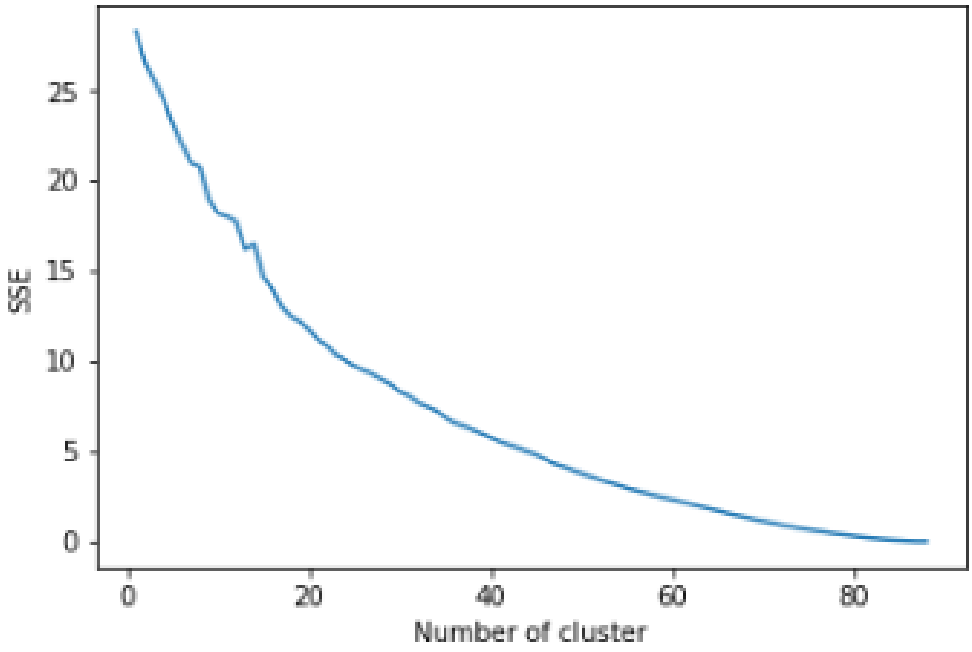
Implementation :

Now we are ready to implement K-Means Clustering on our Data so initially we cluster New Delhi into 10 clusters Following fig shows Delhi before Clustering



Refinement :

We need to find correct value of K for our model this is one of a disadvantage of using K Means but to find the optimum number of K we draw Elbow Plot which shows correct value of K

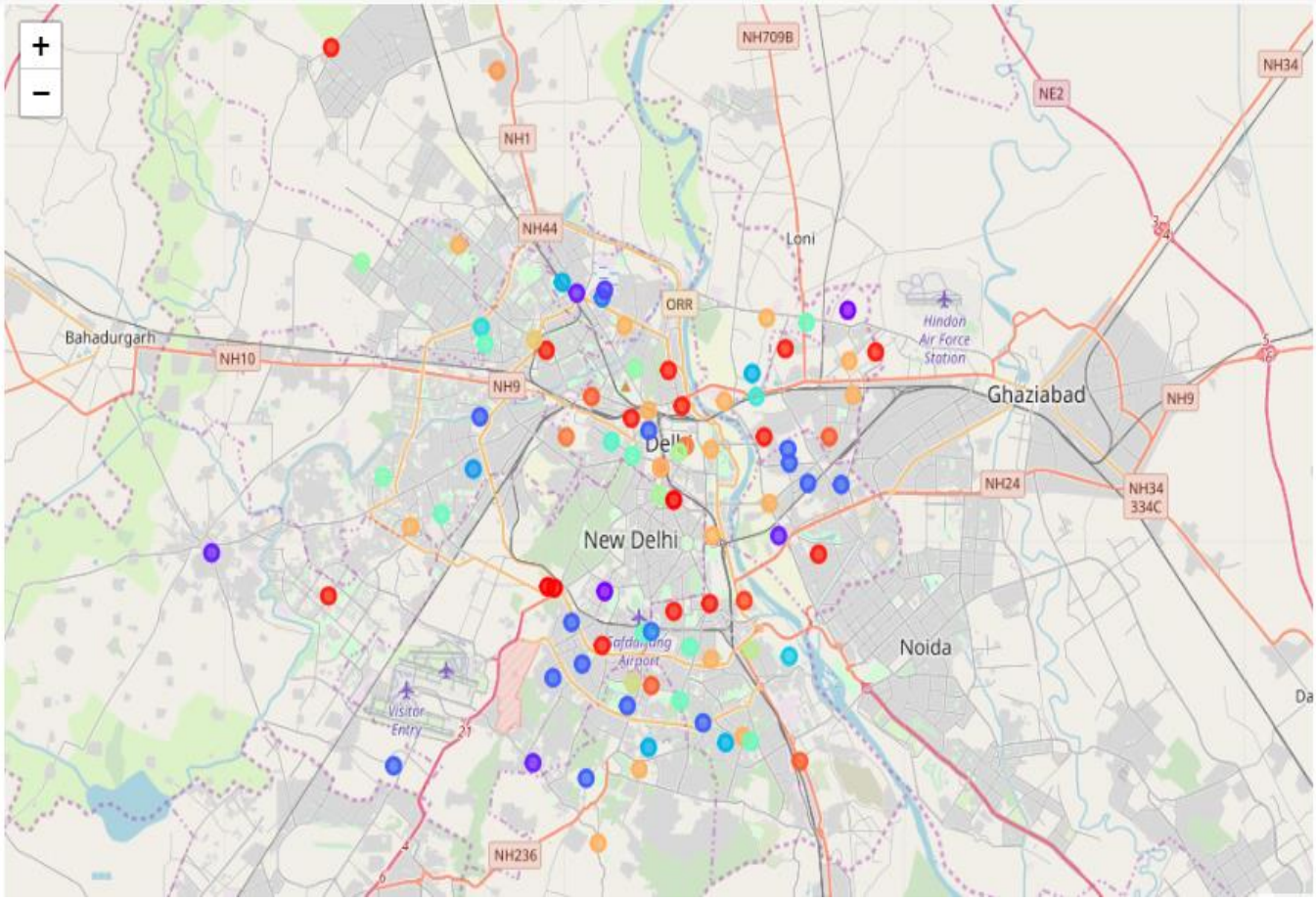


From Elbow Plot we found out that right number of clusters **K = 25**

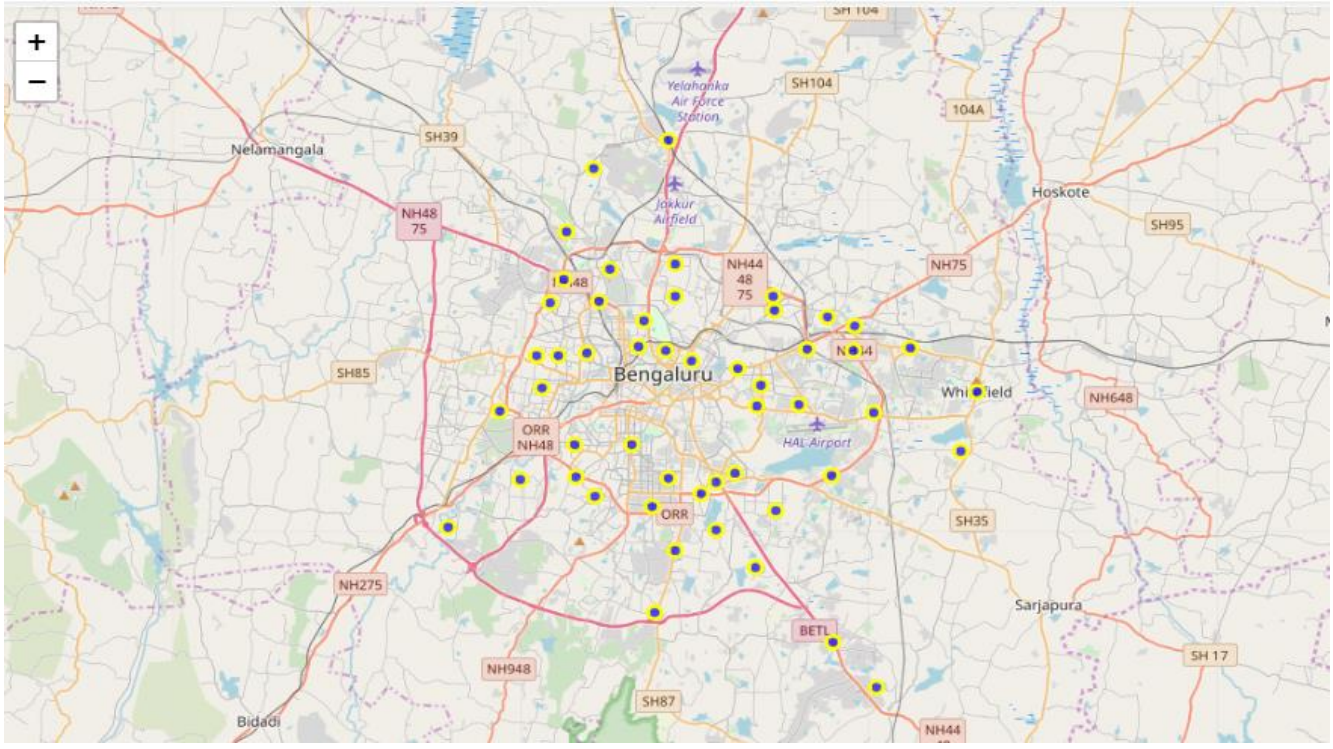
Therefore we will build our model again with *K = 25*

Implementation :

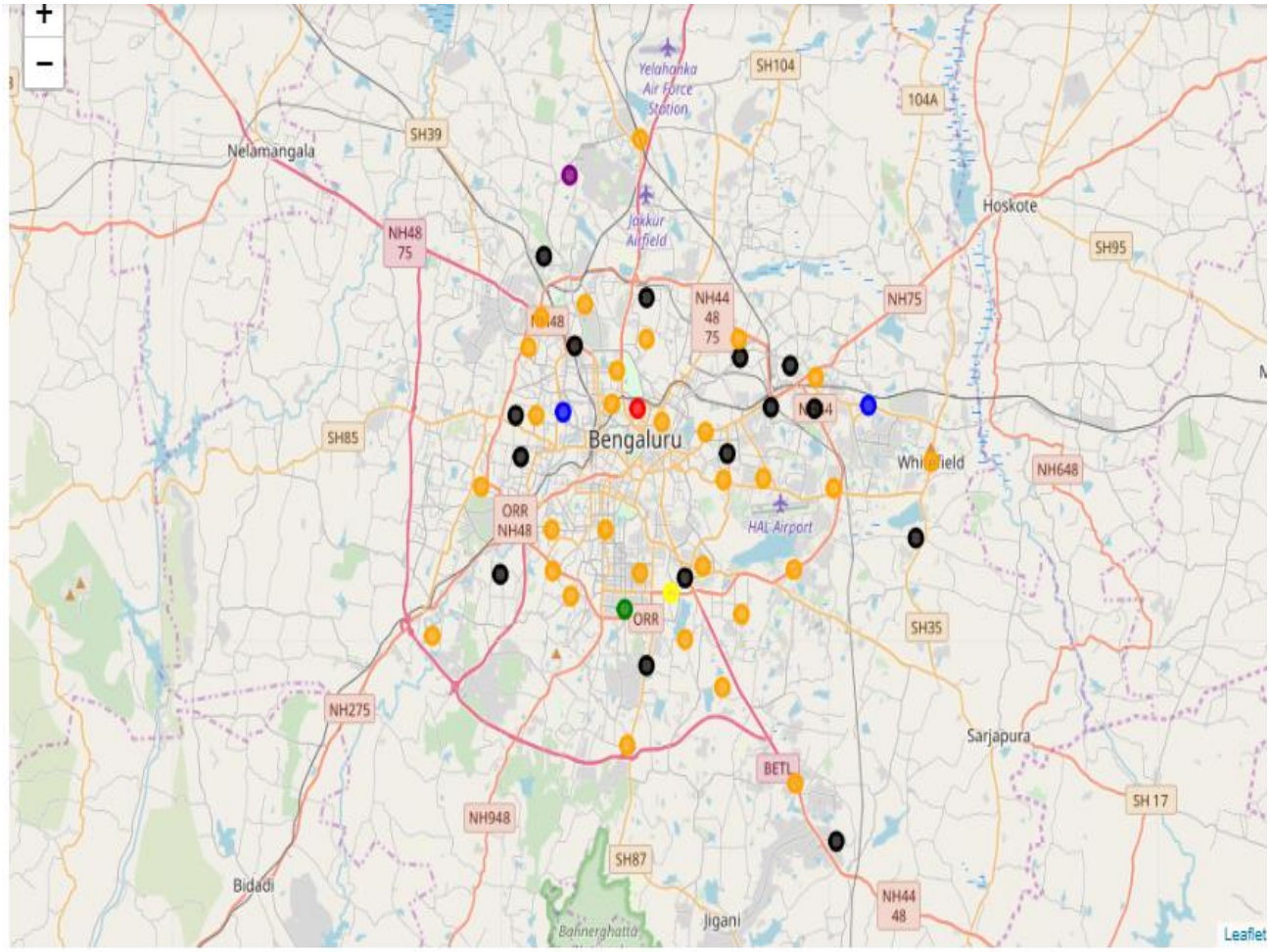
Implementing with K = 25 the following figure shows clusters


NH709B
NH34
NE2
NH1
NH44
Loni
ORR
Hindon Air Force Station
Bahadurgarh
NH10
NH9
Ghaziabad
NH9
NH24
NH34 334C
New Delhi
Noida
Visitor Entry
NH236

Prediction :


Yelahanka Air Force Station
SH104
104A
SH39
Nelamangala
Jakkur Airfield
Hoskote
NH48 75
SH95
NH44 48 75
NH75
SH85
Bengaluru
NH648
ORR NH48
HAL Airport
ORR
SH35
NH275
Sarjapura
BETL
SH 17
NH948
SH87
Bidadi
NH44

Results :



We used K Means clustering in order to find the simillarities between New Delhi and Bangalore. One of the reasons that we used K Means was that it allows us to predict on some unknown datasets like a supervised learning. As this is an Unsupervised Learning their are not many ways to evaluate the putcome of our model but we Evaluated it with the help of silhouette coefficient which computes the compactness of a cluster, where higher is better, with a perfect score of 1. Which is not necessarily true for our case.But we clustered with different algorithms like DBSCAN, Agglomarative and K Means where the result of Aglomarative and K Means where almost same. Apart from that after we succesfully clustered both the cities I manually checked for the clusters which were quite simillar. So overall our model is working and we can tell anyone which part of New Delhi is simillar to Bangalore.

For Example if someone lives in *Geeta Colony, New Delhi* then that person can easily live in *Hoodi, Bangalore* as they are part of one Cluster and thus have lots of Dessert Shop, Donot Shop and Dinner places in common

Conclusion :

We collected our data from multiple Wikipedia pages with the help of pythons web scrapping capabilities. Then we utilize geopy to some extent to get the coordinates of some of the locations of New Delhi and Bangalore and rest I filled manually. After that we merged both together to query the Foresquare API then we parsed the result and stored it into a Dataframe. Later we used that to sort the most common venues everywhere. One more step to go before we can start building our model that is Data Pre-processing where we converted the categorical variables into numerical with the help of One Hot Encoding and also calculated the mean frequency of occurrence of each category of the venues of the Neighbourhoods. Finally we built our model with the help of K Means Clustering algorithm where we clustered New Delhi and then used our model to Predict the values or cluster for Bangalore.