

The Battle of Neighborhoods Delhi vs Pune.

Abhinav Sharma

24th April 2020

1. Introduction and Problem Statement

In this project, we will study, analyze, cluster and compare neighborhoods of two Important locations of India: Delhi and Pune. This will act as a guide for people who leave their native place and come to a new city to pursue their careers.

This is applicable for any city; here we consider that a newly Graduated Engineer has left his native place in Delhi and has come to Pune to pursue his career. We will primarily investigate how similar is the area (specific Postal Code, here we will use 110095 from Delhi) that he has left to the area (Postal Code in Pune 411057) where he is now living.

Doing this project will also enable us to understand similarities and dissimilarities between two Locations as we will compare neighborhoods of Delhi and Pune.

Delhi officially the National Capital Territory of Delhi (NCT), is a union territory of India containing New Delhi, the capital of India. It is bordered by the state of Haryana on three sides and by Uttar Pradesh to the east. The NCT covers an area of 1,484 square kilometers (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 neighboring 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. As of 2016, recent estimates of the metro economy of its urban area have ranked Delhi either the most or second-most productive metro area of India. 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.

Pune also called Poona, the official name until 1978 is the second largest city in the Indian state of Maharashtra, after Mumbai. It is the ninth most populous city in the country with an estimated population of 6.4 million. Along with its extended city limits Pimpri Chinchwad and the three cantonment towns of Pune, Khadki and Dehu Road, Pune forms the urban core of the eponymous Pune Metropolitan Region (PMR). According to the 2011 census, the urban area has a combined population of 5.05 million while the population of the metropolitan region is estimated at 7.27 million. Situated 560 meters (1,837 feet) above sea level on the Deccan plateau on the right bank of the Mutha river, Pune is also the administrative headquarters of its namesake district. In the 18th century, the city was the seat of the Peshwas, the prime ministers of the Maratha Empire and so was one of the most important political centers on the Indian subcontinent. Pune is ranked the number one city in India in the ease of living ranking index.

2. Data Acquisition and Preparation

In this section, the processes of acquiring, cleaning, and preparing each dataset used in this project for next stages will be specified. To be able to do this project, two types of data are needed:

Neighborhood Data: Datasets that lists the names of the neighborhoods of Delhi and Pune and their latitude and longitude coordinates. We performed Web Scraping for this. Below are the URLs used.

For Delhi: <https://www.mapsofindia.com/pincode/india/delhi/>

For Pune: <https://www.mapsofindia.com/pincode/india/maharashtra/pune/>

Below is an example how final data frame looks like.

	Location	Pincode	State	District	Latitude	Longitude
0	A.G.c.r.	110002	Delhi	Central Delhi	28.640964	77.245468
1	A.K.market	110055	Delhi	Central Delhi	28.652506	77.213391
2	Delhi High court	110003	Delhi	Central Delhi	28.598730	77.222995
3	Hauz Qazi	110006	Delhi	Central Delhi	28.655984	77.231623
4	Rajender Nagar	110060	Delhi	Central Delhi	28.638871	77.186188

	Location	Pincode	State	District	Latitude	Longitude
0	A.R. shala	411004	Maharashtra	Pune	18.5087	73.8309
1	Afmc	411040	Maharashtra	Pune	18.484	73.9017
2	Adhale Bk	410506	Maharashtra	Pune	18.6071	73.6912
3	Adivare	410509	Maharashtra	Pune	19.1053	73.6923
4	Agoti	413132	Maharashtra	Pune	18.2188	74.8849

Venues data: Data that describes the top 50 venues (restaurants, cafes, parks, museums, etc.) in each neighborhood of the two cities. The data should list the venues of each neighborhood with their categories. Below is an example how final data frame looks like.

For Delhi:

	Location	Location Latitude	Location Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	A.G.c.r.	28.640964	77.245468	Changezi Chicken	28.643751	77.240486	Indian Restaurant
1	A.G.c.r.	28.640964	77.245468	Bengali Market বাংলা মাঠ বাংলা বাজার	28.629498	77.232020	Indian Restaurant
2	A.G.c.r.	28.640964	77.245468	Karim's करीम كاريम (Karim's)	28.649498	77.233691	Indian Restaurant
3	A.G.c.r.	28.640964	77.245468	Hotel Broadway New Delhi	28.641058	77.237908	Hotel
4	A.G.c.r.	28.640964	77.245468	Feroz Shah Kotla Stadium फिरोज शाह कोटला स्...	28.637907	77.241869	Cricket Ground

For Pune:

	Location	Location Latitude	Location Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	A.R. shala	18.508727	73.830869	Nisarg (निसर्ग)	18.506737	73.832017	Seafood Restaurant
1	A.R. shala	18.508727	73.830869	Cafe Cofee Day	18.508026	73.831636	Café
2	A.R. shala	18.508727	73.830869	Aasing's Kitchen	18.514030	73.828750	Burger Joint
3	A.R. shala	18.508727	73.830869	Le Plaisir	18.514205	73.838551	Bistro
4	A.R. shala	18.508727	73.830869	Pastry Corner	18.508920	73.830565	Bakery

This data will be retrieved from Foursquare which is one of the world largest sources of location and venue data. Foursquare API will be utilized to get and download the data.

Neighborhood Data

a. Delhi

A dataset that specifies the neighborhood data for Delhi was fetched using the mapsofindian.com website where the data set was distinguished on the basis of Districts in Delhi. Data consisted of Location, Pin code, State and District. Below is what raw data looks like for one of the districts of Delhi.

	Pincode Details	Pincode Details.1	Pincode Details.2	Pincode Details.3
0	Location	Pincode	State	District
1	F F c okhla	110020	Delhi	New Delhi
2	New Delhi	110001	Delhi	New Delhi
3	New Delhi.	110001	Delhi	New Delhi

From the above it is clear the first Row is the Column Header and there could be duplicate values for Pin code.

So, first row was made as Column header for all the districts data frames and then they were appended to make one final data frame for Delhi. Once all the data frames were appended, duplicate values were removed.

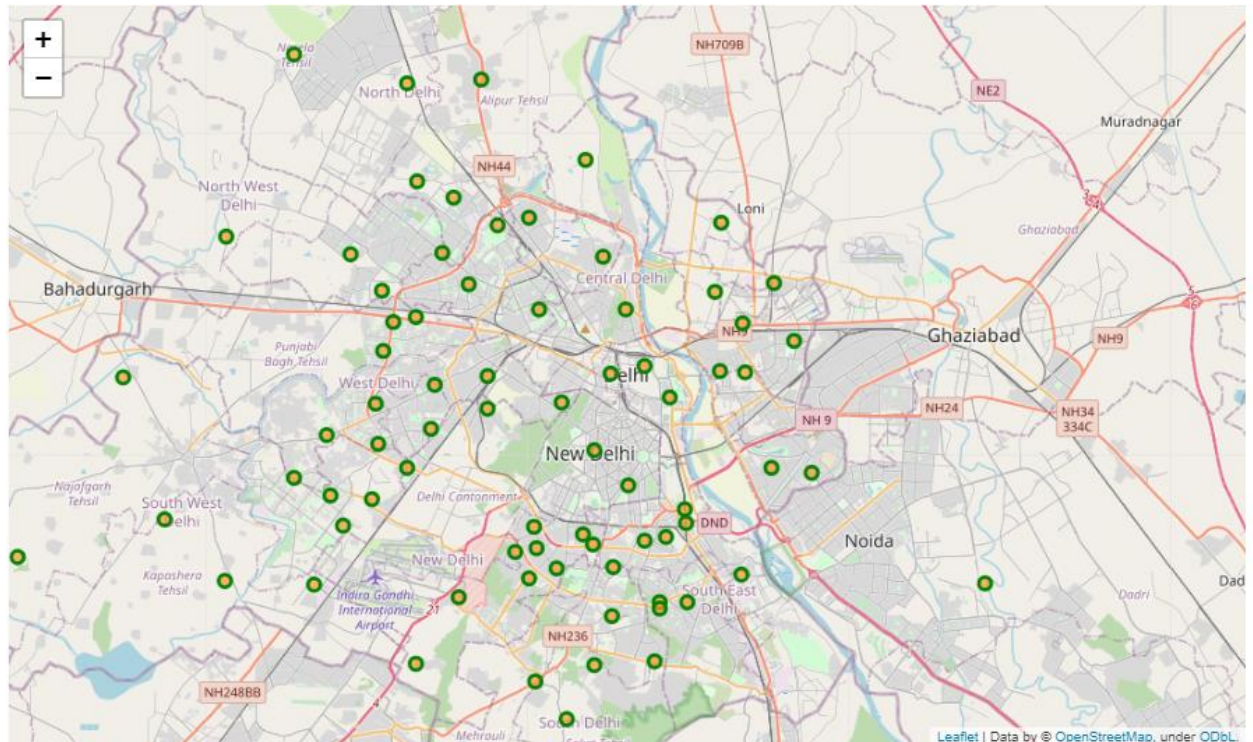
And below is how data frame looked.

	Location	Pincode	State	District
0	A.G.c.r.	110002	Delhi	Central Delhi
1	A.K.market	110055	Delhi	Central Delhi
2	Anand Parbat	110005	Delhi	Central Delhi
3	Baroda House	110001	Delhi	Central Delhi
4	Dada Ghosh bhawan	110008	Delhi	Central Delhi

To fetch the Latitude and Longitude, created a function using geopy's Nominatim. And then using "apply" method of Python we fetched Latitude and Longitudes of different Postal Codes and added columns to main Delhi Data Frame. Below is a snip of Data frame.

	Location	Pincode	State	District	Latitude	Longitude
0	A.G.c.r.	110002	Delhi	Central Delhi	28.640964	77.245468
1	A.K.market	110055	Delhi	Central Delhi	28.652506	77.213391
2	Delhi High court	110003	Delhi	Central Delhi	28.598730	77.222995
3	Hauz Qazi	110006	Delhi	Central Delhi	28.655984	77.231623
4	Rajender Nagar	110060	Delhi	Central Delhi	28.638871	77.186188

Having data of the coordinates of Delhi neighborhoods, it is possible to draw a map using Folium Python package. In this map each circle represents the location of one neighborhood.



b. Pune

A dataset that consists of Neighborhood data of Pune. We did Web Scrapping from mapsofindia.com. Since Pune is a district in State of Maharashtra, we were able to fetch all the postal code at once. Below is how raw data looks like.

	Pincode Details	Pincode Details.1	Pincode Details.2	Pincode Details.3
0	Location	Pincode	State	District
1	A.R. shala	411004	Maharashtra	Pune
2	Afmc	411040	Maharashtra	Pune
3	Adhale Bk	410506	Maharashtra	Pune
4	Adivare	410509	Maharashtra	Pune

Again, from above its clear that first row is column header and there could be duplicates as well. After making first row as header and removing duplicates, below is how data frame looked like.

	Location	Pincode	State	District
0	A.R. shala	411004	Maharashtra	Pune
1	Afmc	411040	Maharashtra	Pune
2	Adhale Bk	410506	Maharashtra	Pune
3	Adivare	410509	Maharashtra	Pune
4	Agoti	413132	Maharashtra	Pune

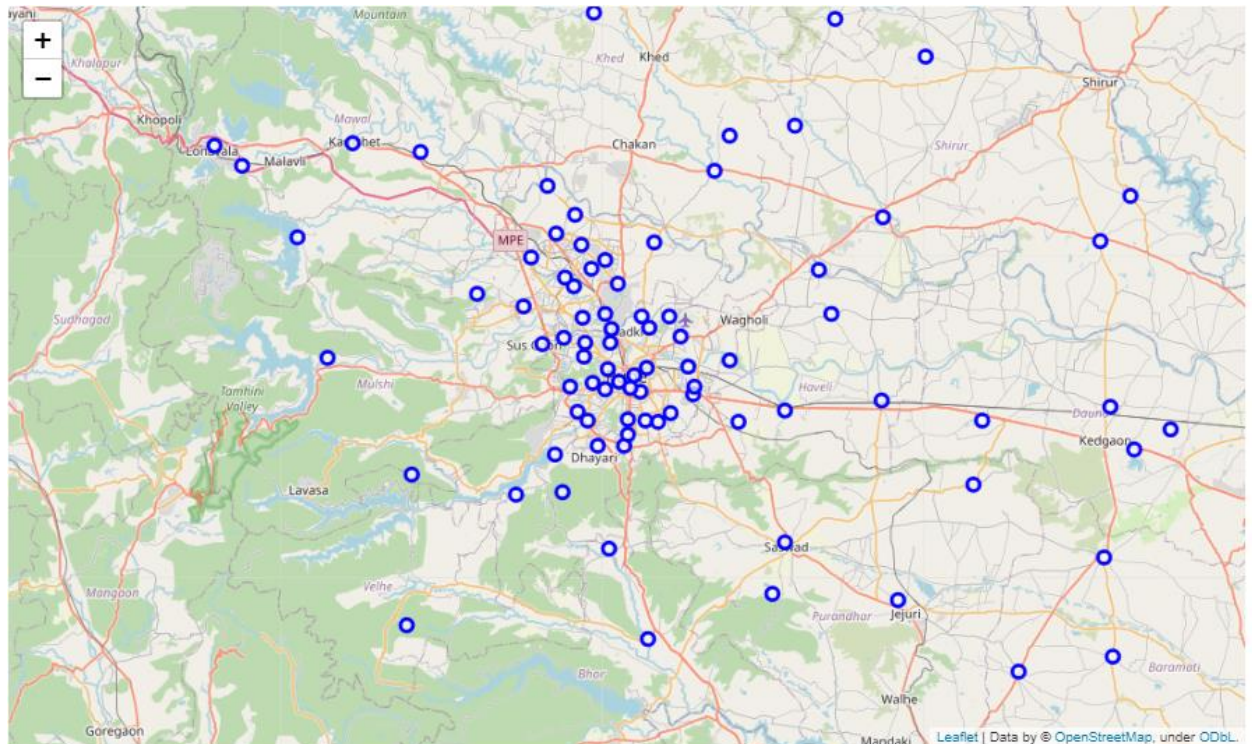
Again, to fetch the Latitude and Longitude, I used same function created above using geopy's Nominatim. And then using "apply" method of Python we fetched Latitude and Longitudes of different Postal Codes and added columns to main Pune Data Frame. Below is a snip of Data frame.

	Location	Pincode	State	District	Latitude	Longitude
0	A.R. shala	411004	Maharashtra	Pune	18.5087	73.8309
1	Afmc	411040	Maharashtra	Pune	18.484	73.9017
2	Adhale Bk	410506	Maharashtra	Pune	18.6071	73.6912
3	Adivare	410509	Maharashtra	Pune	19.1053	73.6923
4	Agoti	413132	Maharashtra	Pune	18.2188	74.8849

While fetching the Latitude and Longitude I received null values for some of the postal codes for Pune. Which were filtered out.

	Location	Pincode	State	District	Latitude	Longitude
28	Arvi	412401	Maharashtra	Pune	NA	NA
35	Avsari Kh	412405	Maharashtra	Pune	NA	NA
55	Bibi	410513	Maharashtra	Pune	NA	NA
61	C M e	411031	Maharashtra	Pune	NA	NA
70	Daundaj	412305	Maharashtra	Pune	NA	NA
71	Dawanewadi	412311	Maharashtra	Pune	NA	NA
82	Gangapur Kh	410516	Maharashtra	Pune	NA	NA
100	Kandali	412412	Maharashtra	Pune	NA	NA
102	Karegaon	412220	Maharashtra	Pune	NA	NA
115	Lohogaon	411047	Maharashtra	Pune	NA	NA
131	Srpf	411022	Maharashtra	Pune	NA	NA
132	Theur	412110	Maharashtra	Pune	NA	NA
133	Tondal	412312	Maharashtra	Pune	NA	NA

Having data of the coordinates of Pune neighborhoods, a map was using Folium Python package was created. In this map each circle represents the location of one neighborhood.



Venues

For each city, data that describes the venues of its neighborhoods and the categories of these venues is needed. Venues data will be retrieved from Foursquare which is a popular

source of location and venue data. Foursquare API service will be utilized to access and download venues data.

To retrieve data from Foursquare using their API, a URL should be prepared and used to request data related a specific location.

An example URL is the following:

https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}

where explore indicates the API endpoint used, client_id and client_secret are credentials used to access the API service and are obtained when registering a Foursquare developer account, v indicates the API version to use, ll indicates the latitude and longitude of the desired location, radius is the maximum distance in meters between the specified location and the retrieved venues, and limit is used to limit the number of returned results if necessary.

a. Delhi

Using the function get_near_by_venues with Delhi neighborhood data, retrieved data about 1931 venues in Delhi neighborhoods. For each venue, venue name, category, latitude, and longitude were retrieved. The head of the data frame returned by the function for Delhi is shown. We can see that each row in the data frame contains data about one venue: the venue name, coordinates (latitude and longitude), and category in addition to the neighborhood in which the venue is located and the coordinates of the neighborhood.

	Location	Location Latitude	Location Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	A.G.c.r.	28.640964	77.245468	Changezi Chicken	28.643751	77.240486	Indian Restaurant
1	A.G.c.r.	28.640964	77.245468	Bengali Market বাংলা মার্কেট বাংলা বাজার	28.629498	77.232020	Indian Restaurant
2	A.G.c.r.	28.640964	77.245468	Karim's करीम کرمی (Karim's)	28.649498	77.233691	Indian Restaurant
3	A.G.c.r.	28.640964	77.245468	Hotel Broadway New Delhi	28.641058	77.237908	Hotel
4	A.G.c.r.	28.640964	77.245468	Feroz Shah Kotla Stadium फिरोज शाह कोटला स्टेडियम	28.637907	77.241869	Cricket Ground

b. Pune

Like what has been done for Delhi, a data frame that describes the venues of Pune neighborhoods was created. The data frame contains data for more than 1547 venues in Pune.

	Location	Location Latitude	Location Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	A.R. shala	18.508727	73.830869	Nisarg (निसर्ग)	18.506737	73.832017	Seafood Restaurant
1	A.R. shala	18.508727	73.830869	Cafe Cofee Day	18.508026	73.831636	Café
2	A.R. shala	18.508727	73.830869	Aasing's Kitchen	18.514030	73.828750	Burger Joint
3	A.R. shala	18.508727	73.830869	Le Plaisir	18.514205	73.838551	Bistro
4	A.R. shala	18.508727	73.830869	Pastry Corner	18.508920	73.830565	Bakery

Now let's first compare specific Postal Codes from both Locations: 110095(Dilshad Garden, Delhi) vs 411057(Rajiv Gandhi Infotech Park, Pune).

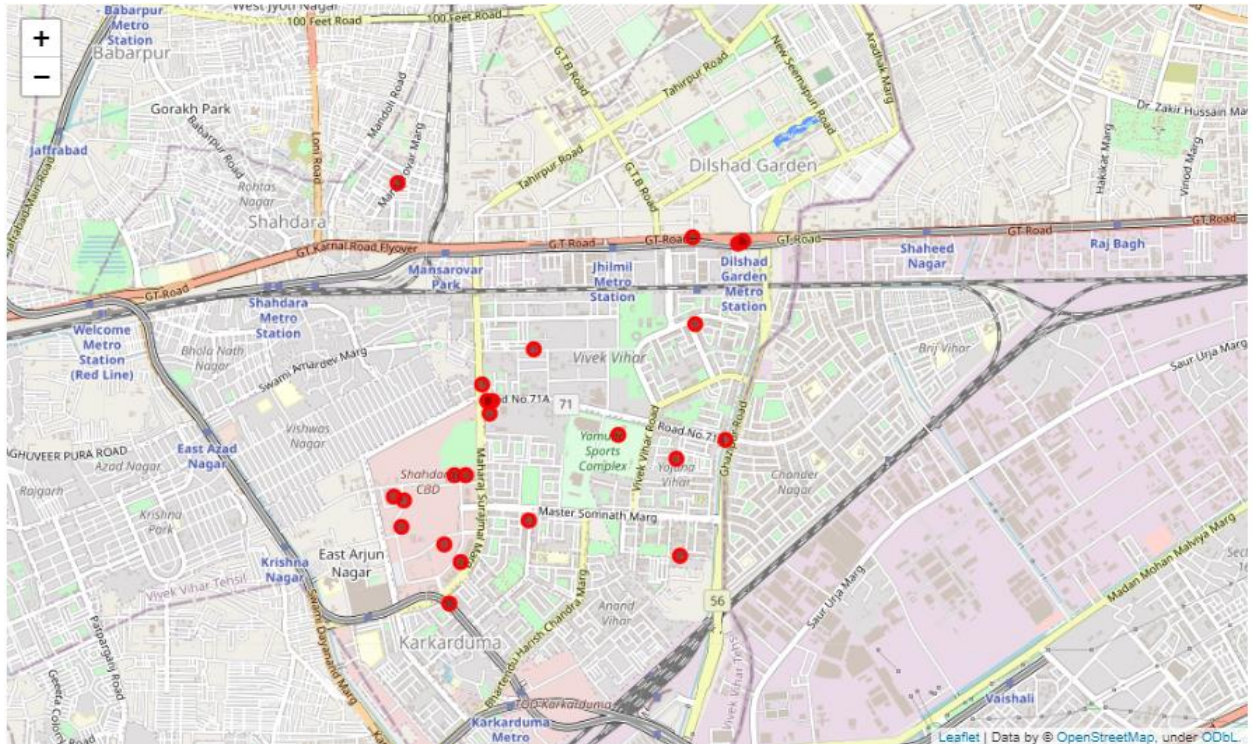
I filtered out Dilshad Garden from Delhi data frame and Infotech park from Pune Data frame. Following are the resulting Data Frames.

	Location	Location Latitude	Location Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Dilshad Garden	28.668089	77.313461	Y.S.C - Yamuna Sports Complex	28.664500	77.312847	Athletics & Sports
1	Dilshad Garden	28.668089	77.313461	Yamuna Sports Complex	28.664257	77.320179	Pool
2	Dilshad Garden	28.668089	77.313461	Cafe Wink	28.657311	77.317098	Italian Restaurant
3	Dilshad Garden	28.668089	77.313461	The Leela Ambience Delhi Convention Hotel	28.662089	77.302429	Hotel
4	Dilshad Garden	28.668089	77.313461	Barbeque Nation	28.665840	77.304094	BBQ Joint

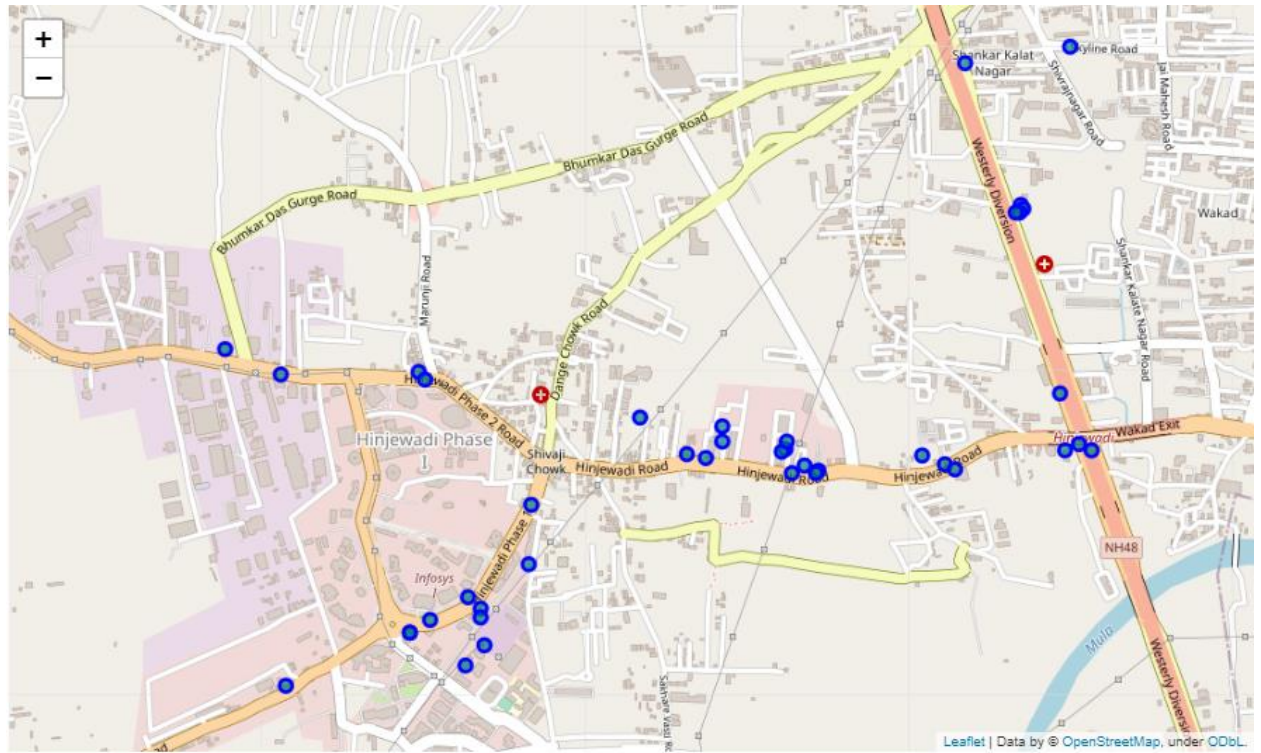
	Location	Location Latitude	Location Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Infotech park (hinjawadi)	18.59393	73.741764	Little Italy	18.591513	73.743668	Italian Restaurant
1	Infotech park (hinjawadi)	18.59393	73.741764	MoMo Cafe	18.591650	73.747011	Indian Restaurant
2	Infotech park (hinjawadi)	18.59393	73.741764	Courtyard by Marriott	18.591591	73.746877	Hotel
3	Infotech park (hinjawadi)	18.59393	73.741764	The Gateway HOTEL	18.591914	73.744877	Hotel
4	Infotech park (hinjawadi)	18.59393	73.741764	Natural Ice Cream	18.591192	73.752440	Ice Cream Shop

Since we have the coordinates. I used Folium to plot these venues.

Following map is of Dilshad Garden with **23 venues**.



And below map is of Rajiv Gandhi Infotech park (Hinjawadi) with **45 Venues**.

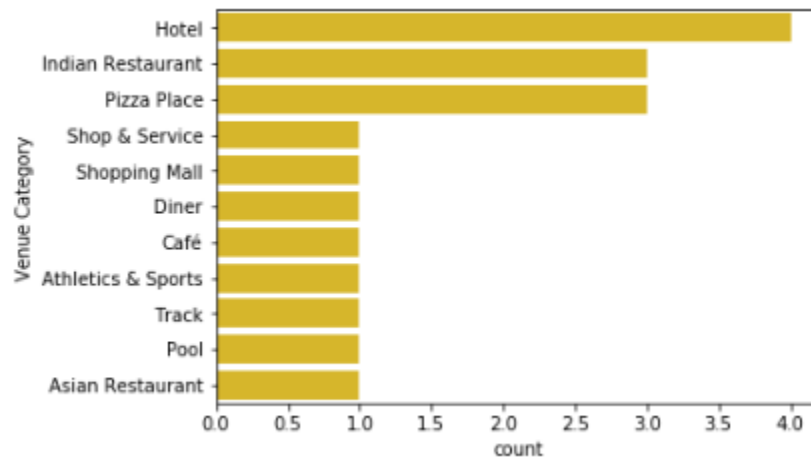


3. Exploratory Data Analysis

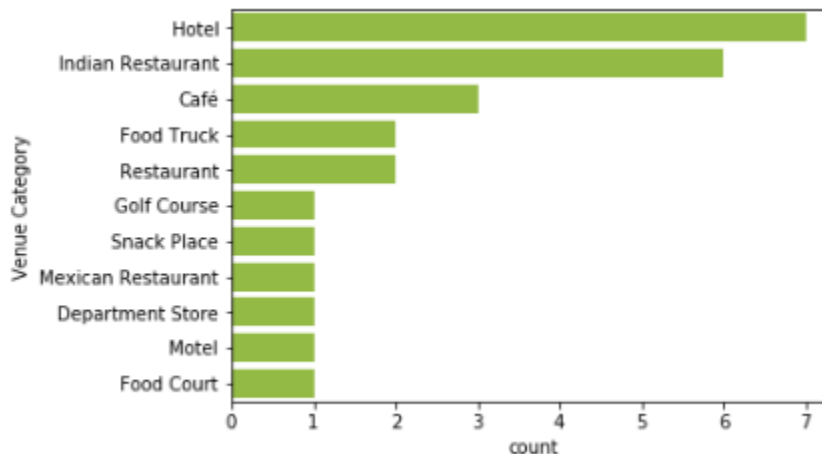
In this section, the datasets produced in the previous section will be explored via effective visualizations to understand the data better.

a. Dilshad Garden vs Rajiv Gandhi Infotech Park.

Below count plot is depicting top 10 venue categories in Dilshad Garden.



And, below count plot is depicting top 10 venue categories in Infotech Park.



We can clearly see "Hotel" is the most popular place in both Postal Codes, but Rajiv Gandhi Infotech Park (Hinjawadi) has a greater number of hotels as compared to Dilshad Garden. Also, Rajiv Gandhi Infotech Park (Hinjawadi) has a greater number of Food Venues as compared to Dilshad Garden.

One possible reason could be that Rajiv Gandhi Infotech Park (Hinjawadi) is an IT hub. All big IT companies are present here, which justifies a greater number of hotels, multicuisine restaurants, cafe etc. Whereas Dilshad Garden is more of a residential area.

	Location	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	11th Most Common Venue	12th Most Common Venue
0	Dilshad Garden	Hotel	Pizza Place	Indian Restaurant	Athletics & Sports	Shopping Mall	Shop & Service	BBQ Joint	Pool	Park	Diner	Asian Restaurant	Track
1	Infotech park (hinjawadi)	Hotel	Indian Restaurant	Café	Food Truck	Restaurant	Deli / Bodega	Golf Course	Food Court	Fast Food Restaurant	Department Store	Vegetarian / Vegan Restaurant	Chinese Restaurant
		13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	16th Most Common Venue	17th Most Common Venue	18th Most Common Venue	19th Most Common Venue	20th Most Common Venue				
		Café	Juice Bar	Italian Restaurant	Big Box Store	Deli / Bodega	Chaat Place	Chinese Restaurant	Vegetarian / Vegan Restaurant				
		Ice Cream Shop	Buffet	Breakfast Spot	Bar	Bakery	BBQ Joint	Chaat Place	IT Services				

Both Locations Dilshad Garden and Infotech Park have number of eating options. Both locations have Restaurants, BBQ Joints, Cafes etc. in common which makes it a lot easier for someone to adjust in terms of Food.

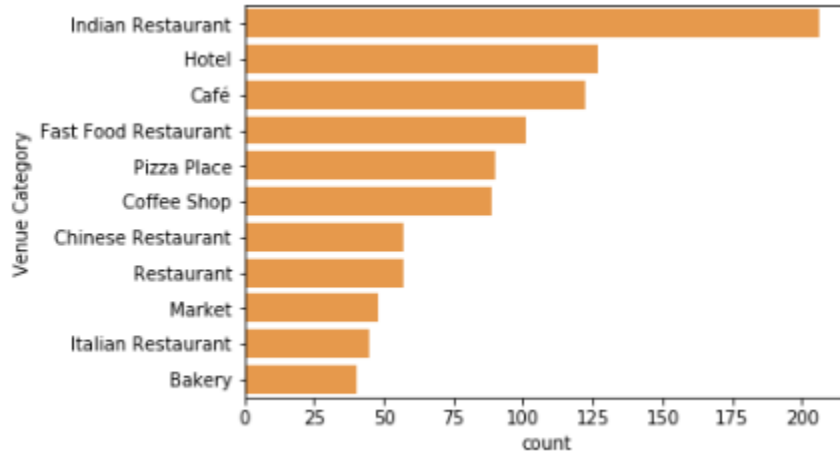
One of the Biggest Difference is presence of IT services in Pune. Infotech Park as its name says is an IT hub. Whereas there is no such venue in Dilshad Garden.

Therefore many people from different parts of country to Pune as it is one of the IT Hubs in India.

b. Delhi vs Pune

I used count plot of Seaborn Lib to find top 10 venue categories in Delhi and Pune.

Below is the count plot with top 10 venue categories of Delhi.



And, below is the count plot with top 10 venue categories in Pune.



We see, "Indian Restaurant" is clearly a favorite choice in both Delhi and Pune.

But in Delhi, second most opened place is "Hotel" whereas in Pune its "Cafe". One of the Possible reasons could be that Delhi is capital of India so it's more likely to have tourists. Hence a greater number of Hotels.

Also, both locations have tremendous number of food options. Which is one of the similarities between two places.

4. Clustering of Neighborhoods

In this section, clustering will be applied on Delhi and Pune neighborhoods to find similar neighborhoods in the two cities. Clustering is the process of finding similar items in a dataset based on the characteristics (features) of items in the dataset. K-means clustering algorithm of the Scikit-learn Python library will be used.

K-Means clustering algorithm identifies k number of centroids and allocates every data point to nearest clusters while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithm and apt for this Project.

a. Feature Selection

The goal of the clustering is to cluster neighborhoods based on the similarity of venue categories in the neighborhoods. This means that the two things of interest here are the neighborhood and the venue categories in the neighborhood. Thus, the following two features will be selected out of the data frames “Locations (Neighborhood)” and “Venue Category”.

But still after that, the data is not ready for the clustering algorithm because the algorithm works with numerical features.

Before doing any data preparation, I combined both data frames of Delhi and Pune.

And, I applied “one-hot encoding” to Venue Category feature of the combined data frame and resulting data frame is shown below.

	Location	ATM	Accessories Store	Airport Food Court	Airport Service	Airport Terminal	American Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Astrologer	Athletics & Sports	Australian Restaurant	Aut Garag
0	A.G.c.r.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	A.G.c.r.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	A.G.c.r.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	A.G.c.r.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	A.G.c.r.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Shape of this resulting data frame was (3478,213), which means total of 213 Venue categories are present in data frame. Basis of these Venue Categories clustering will be performed.

b. Grouping of Locations

Next, with one-hot encoded data frame I grouped locations by taking the mean of the frequency of each occurrence of each category. Shape of resulting data frame was (163,213). By doing so, I was also preparing data for Clustering.

c. Most common categories of each Location

Using grouped data frame another data frame is created to display the top 10 venues of each Location.

	Location	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	A F rajokari	Hotel	Indian Restaurant	Restaurant	Fast Food Restaurant	Shopping Mall	Café	Multiplex	Bar	American Restaurant	Gym
1	A-3 Janak puri	Indian Restaurant	Pizza Place	Fast Food Restaurant	Café	Sandwich Place	Coffee Shop	Chinese Restaurant	Metro Station	Restaurant	BBQ Joint
2	A.G.c.r.	Indian Restaurant	Hotel	Flea Market	Bakery	Café	Mosque	Frozen Yogurt Shop	Stadium	Restaurant	Light Rail Station
3	A.K.market	Hotel	Indian Restaurant	Snack Place	Dessert Shop	Market	Pizza Place	Fast Food Restaurant	Restaurant	Indian Chinese Restaurant	Café
4	A.R. shala	Café	Indian Restaurant	Snack Place	Vegetarian / Vegan Restaurant	South Indian Restaurant	Seafood Restaurant	Fast Food Restaurant	Gym / Fitness Center	Pizza Place	Bistro

d. Clustering

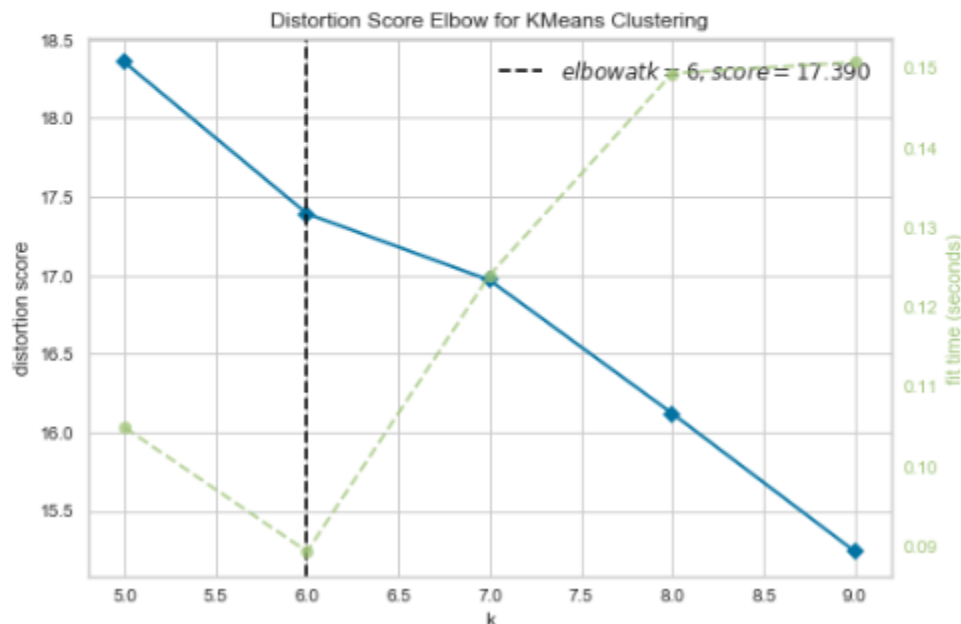
To use KMEANS, we need k value i.e. number of clusters. For this we will use yellowbrick module's method `kelbow_visualizer()`.

Yellowbrick Module's `KELbow_visualizer()` helps to determine the k value i.e. how many clusters should be created. Following is code snippet and result.

```

1 kmeans = KMeans()
2
3 #import yellowbrick's KELbowVisualizer
4 from yellowbrick.cluster import KELbowVisualizer
5 visualizer = KELbowVisualizer(kmeans, k=(5,10))
6
7 visualizer.fit(delve_grouped_clustering)
8 visualizer.show()

```



The vertical black dashed line tells us the apt number of clusters aka k value. Above results clearly indicate to use k = 6 i.e. to have 6 clusters.

Here the parameter passed in `visualizer.fit()` i.e. `delne_grouped_clustering` is the data frame that is prepared by grouping the locations, just the Location column is dropped of it. This was done as clustering algorithm doesn't accept non-numerical columns.

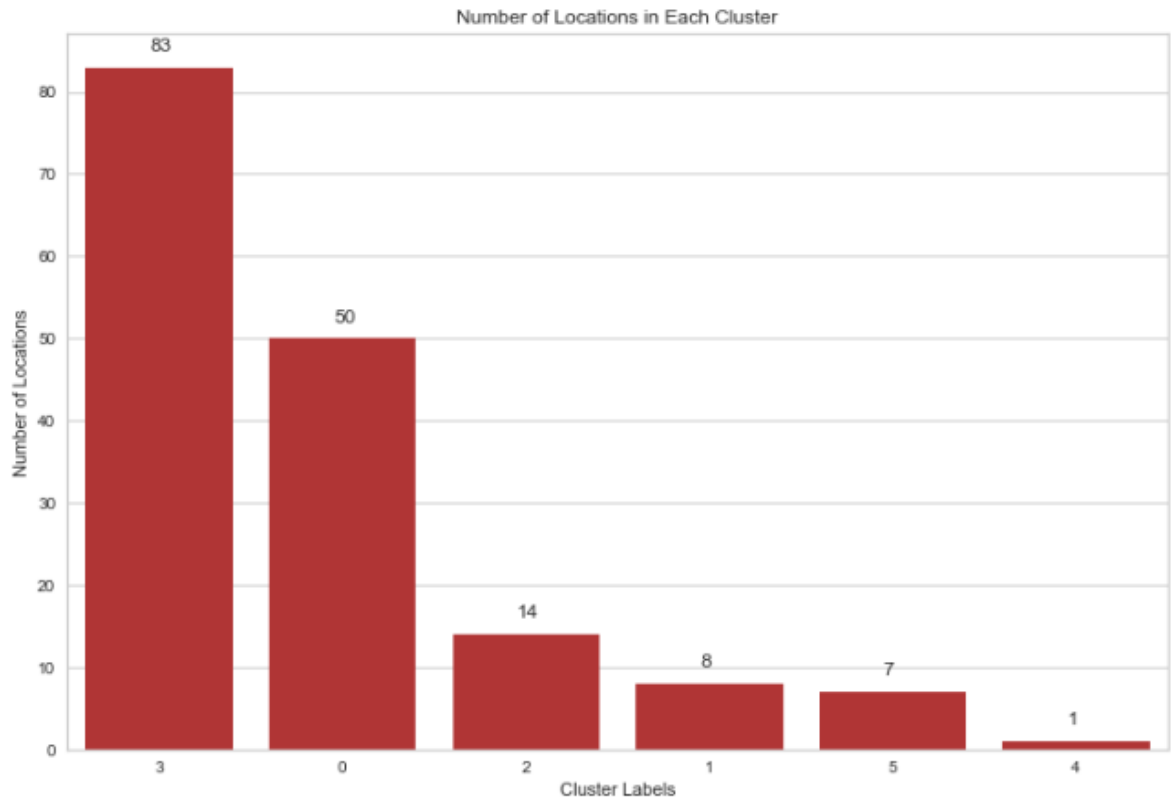
Clustering algorithm produced cluster-labels; these labels denote the cluster of each record (i.e. each neighborhood) in the data. Using these labels and the top 10 venues data frame, a data frame is constructed to show the neighborhoods of Delhi and Pune, the cluster to which each neighborhood belongs, and the most common venue categories in each neighborhood.

Cluster Labels		Location	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	3	A.F. rajokari	Hotel	Indian Restaurant	Restaurant	Fast Food Restaurant	Shopping Mall	Café	Multiplex	Bar	American Restaurant	Gym
1	0	A-3 Janak puri	Indian Restaurant	Pizza Place	Fast Food Restaurant	Café	Sandwich Place	Coffee Shop	Chinese Restaurant	Metro Station	Restaurant	BBQ Joint
2	0	A.G.c.r.	Indian Restaurant	Hotel	Flea Market	Bakery	Café	Mosque	Frozen Yogurt Shop	Stadium	Restaurant	Light Rail Station
3	0	A.K.market	Hotel	Indian Restaurant	Snack Place	Dessert Shop	Market	Pizza Place	Fast Food Restaurant	Restaurant	Indian Chinese Restaurant	Café
4	3	A.R. shala	Café	Indian Restaurant	Snack Place	Vegetarian / Vegan Restaurant	South Indian Restaurant	Seafood Restaurant	Fast Food Restaurant	Gym / Fitness Center	Pizza Place	Bistro

5. Clustering Results

The output of the clustering operation is 6 clusters with cluster labels 0, 1, 2, 3, 4 and 5. Each cluster is expected to contain a group of similar neighborhoods based on the categories of the venues in each neighborhood.

Following Seaborn count plot depicts number of locations in each cluster in descending order.

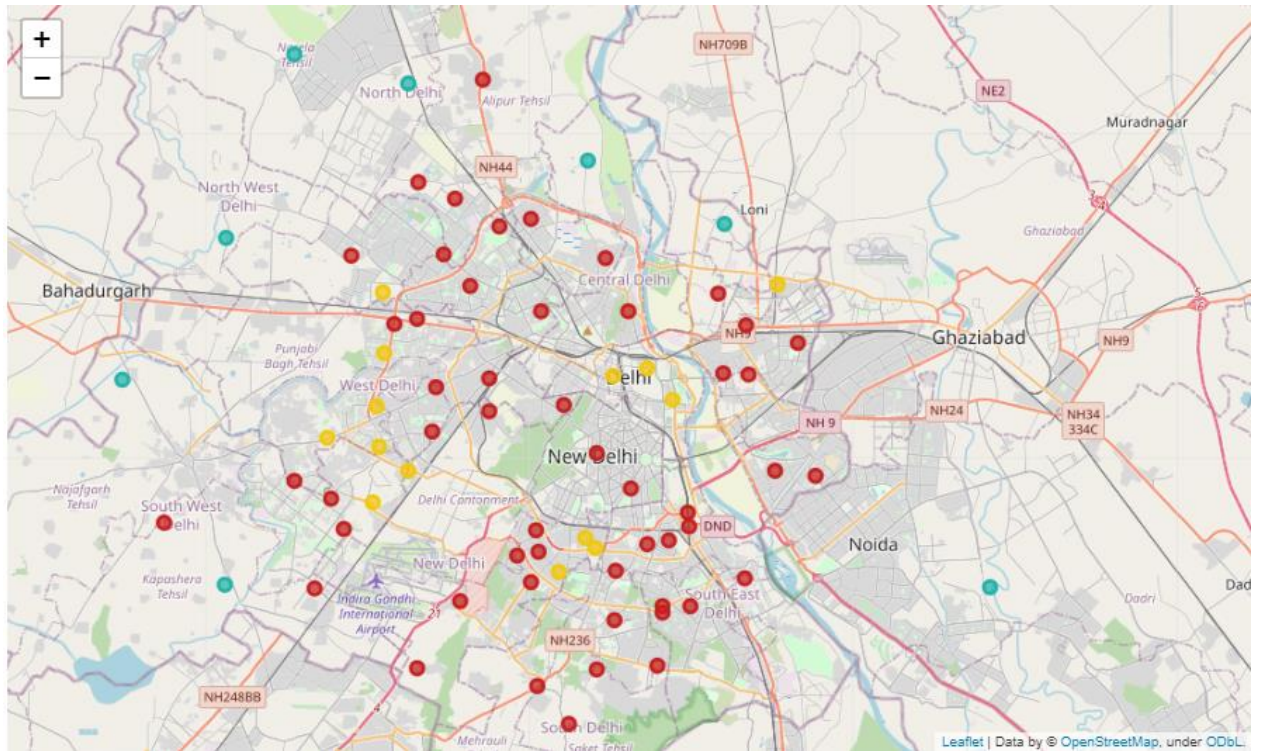


Above plot clearly says most locations are in Cluster 3 and least number of locations are in Cluster 4. One location in Cluster 4 is Bhigwan as shown below.

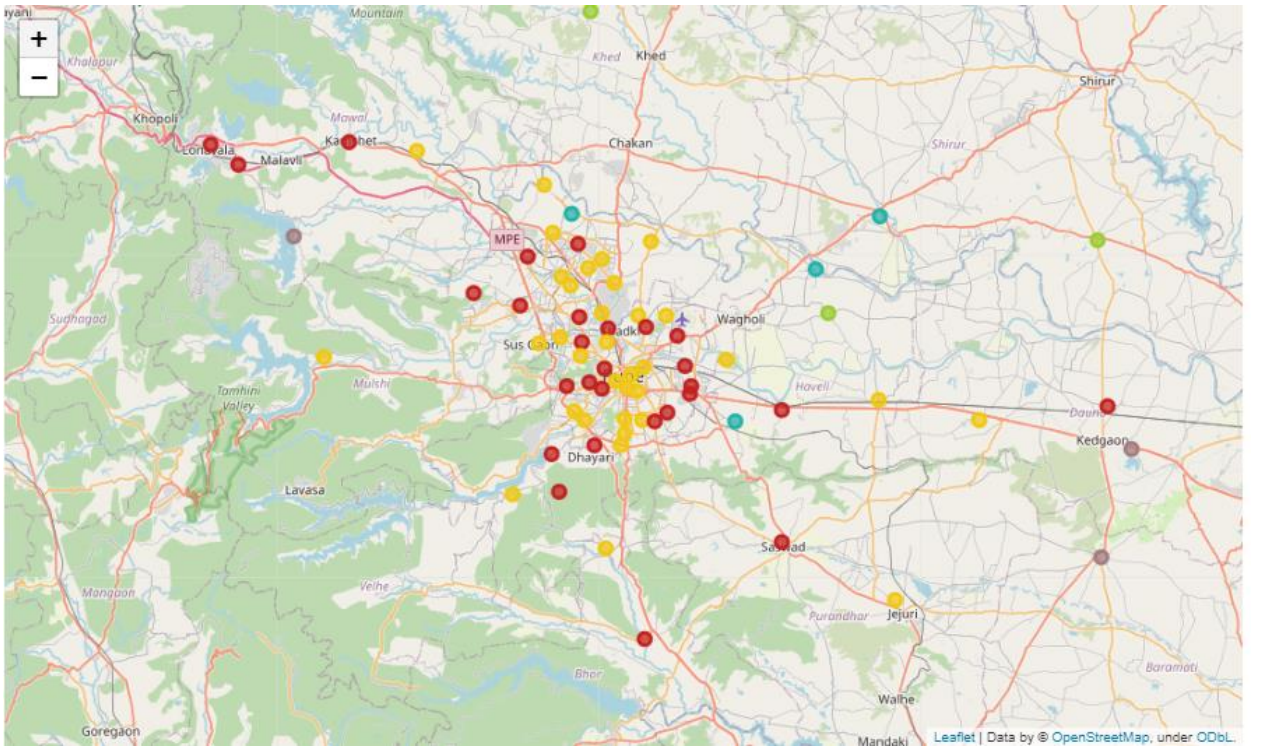
1 delne_venues_sorted[delne_venues_sorted['Cluster Labels'] == 4]												
Cluster Labels	Location	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	
63	4 Bhigwan	Diner	Zoo	Farm	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Food	


Since location coordinates are available, map for Delhi and Pune is plotted. To make sure that color of clusters remains constant in both map a function is created and used while plotting the map.

Delhi Cluster Map:





Pune Cluster Map:





Cluster 0 = Moon Yellow Color 

Cluster 1 = Yellow Green Color 

Cluster 2 = Light Sea Green color 

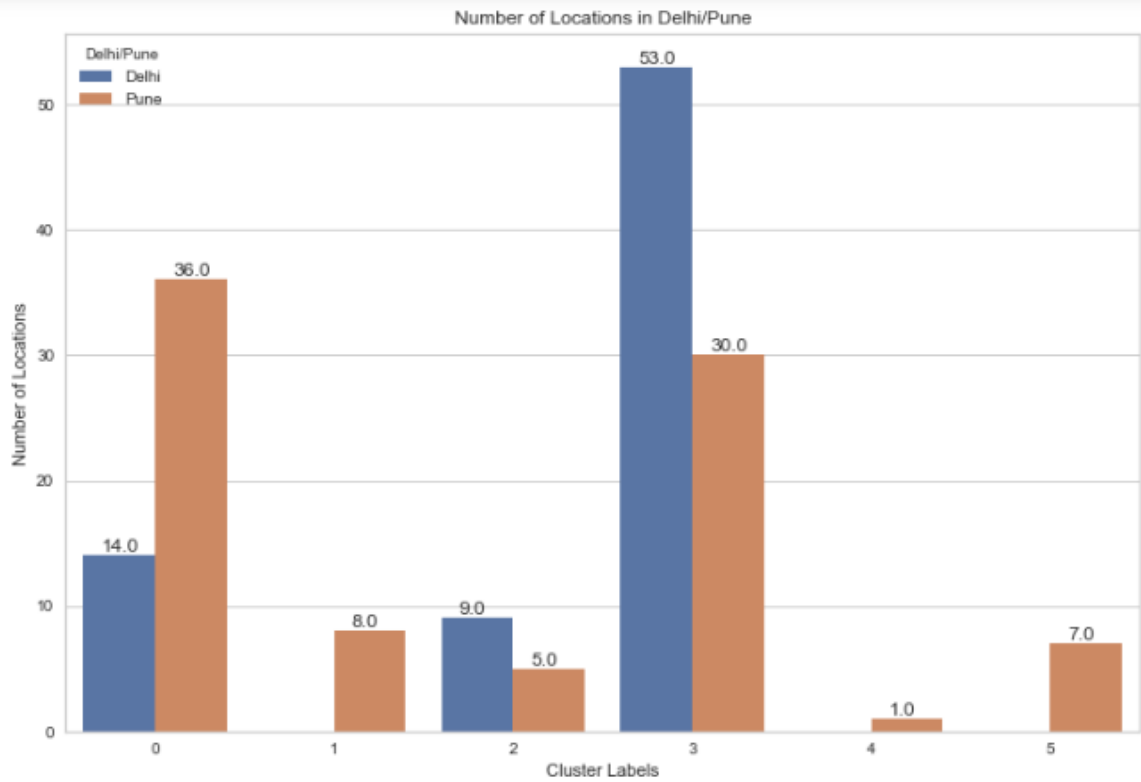
Cluster 3 = Fire Brick 

Cluster 4 = Neon Carrot 

Cluster 5 = Turkish Rose 

- By Observing above maps, clearly both locations have neighborhoods in common. Specially in cluster 3 and cluster 0.
- Coming to Dilshad Garden and Infotech park, both are clustered under cluster 3. This implies someone whose native place is Dilshad Garden, Delhi and comes to Pune, Maharashtra to pursue his/her career will not face many challenges in terms of adjusting to new place. Both places are almost similar.

To find how many locations of Delhi and Pune each are present in each cluster I used seaborn's count plot.



From above, composition of most dense clusters cluster 3 and cluster 0 can be seen along with other clusters.

- Cluster 0 is dominated by Locations Pune with 36 Locations and 14 Locations from Delhi.
- Cluster 3 is dominated by Location Delhi with 53 Locations and 30 locations from Pune.
- Out of 163 neighborhoods collectively, 143 neighborhoods are similar Between Delhi and Pune I.e. 90.19% similarity rate.
- 16 neighborhoods (cluster 1, 4 and 5) in Pune do not have any similar neighborhoods in Delhi. These neighborhoods are shown below.

Agoti
Ala
Alegaon
Amondi
Ane
Ashtapur
Audar
Avasari Bk
Babhurdi
Bawada
Bhigwan
Chilewadi
Hivare Bk
Kadethan
Kurwandi
Pawananagar

6. Limitations and Recommendations

In this project, we only considered clustering locations on the similarity of Venues. We didn't consider Population Density, Standard of living, Accessibility, Connectivity etc. Which can be crucial for clustering Locations.

It will be recommended for future research purpose to devise a methodology to estimate such data to be used in clustering algorithm.

This project made use of Sandbox Tier account of Foursquare API that came with limitations as to the API calls and results returned. Future research can make use of paid account to bypass these limitations.

7. Conclusion

In this project, the neighborhoods of Delhi and Pune were clustered into multiple groups based on the categories (types) of the venues in these neighborhoods. The results showed that there are venue categories that are more common in some cluster than the others; the most common venue categories differ from one cluster to the other. If a deeper analysis—taking more aspects into account—is performed, it might result in discovering different style in each cluster based on the most common categories in the cluster.