THE BATTLE OF NEIGHBORHOODS

Shrey Srivastava

June 2,2020

1. Introduction:

1.1 Background:

Bangalore, officially Bengaluru, is the capital of the Indian state of Karnataka. It has a population of over ten million, making it a megacity and the third-most populous city and fifth-most populous urban agglomeration in India. Bengaluru is sometimes referred to as the "Silicon Valley of India" (or "IT capital of India") because of its role as the nation's leading information technology (IT) exporter. Not only has it attracted almost all the major players in the IT industry, it is also a breeding ground for various indigenous companies and startups.

It is the second fastest-growing major metropolis in India, having one of the most highly educated workforces in the world. Its workforce comes from all the parts of the country and abroad owing to the increasing job opportunities.

1.2 Problem Description:

Since Bangalore is an IT hub, and is still growing, there are multiple job opportunities. In search of better opportunities and to have a better lifestyle employees sometimes change jobs. Now, the metropolis stretches across more than 700 sq km of land, so people have to relocate within the city.

When people relocate they look for a similar social environment, the same facilities and similar amenities. These facilities include Grocery Stores, Gyms, Restaurants of various cuisines, Metro/Bus stations for connectivity, etc.

Therefore, the aim of this project is to facilitate the decision making of people while choosing a new neighborhood to relocate to. The end product will be a map of Bangalore which will show neighborhoods of the same kind in the same color. Therefore, those who are looking to relocate within Bangalore can look at the map and choose a neighborhood to settle in, which is similar to their older neighborhood so as to maintain a similar lifestyle and to settle in easily.

1.3 Interest:

As mentioned earlier, this project would interest people in the workforce of the city who are changing jobs for the aforementioned reasons. Even people from outside the city who are looking to move to Bangalore, or tourists who want to explore the city should also find this project useful.

2. Data:

2.1 Data Acquisition and Description:

We will be dealing with location data for Bangalore. The following are the datasets I have used.

Data 1: Contains regions in Bangalore and the neighborhoods within them

- First we need the various neighborhoods in the different regions in Bangalore.
- I have manually copied the regions and their neighborhoods into a list and turned it into a dataframe, one for each region.
- Finally I have concatenated the individual region dataframes to arrive at the final dataframe for locations, which looks like this:

	Region	Neighborhood
0	Central	Cantonment
1	Central	Domlur
2	Central	Indiranagar
3	Central	Jeevanbheemanagar
4	Central	Malleswaram
...	...	...
5	Western	Nandini Layout
6	Western	Nayandahalli
7	Western	Rajajinagar
8	Western	Rajarajeshwari Nagar
9	Western	Vijayanagar

65 rows × 2 columns

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

Data 2: Contains the co-ordinates of the neighborhoods above

- To get the latitude and longitude values for each neighborhood, I have made use of the python package, Geocoders in Geopy.
- After using Geocoders and getting the latitudes and longitudes, our dataframe looks like this:

	Region	Neighborhood	Latitude	Longitude
0	Central	Cantonment	12.979120	77.591300
1	Central	Domlur	12.962467	77.638196
2	Central	Indiranagar	12.973291	77.640467
3	Central	Jeevanbheemanagar	12.964200	77.658100
4	Central	Malleswaram	13.002735	77.570325
...	...	...	...	...
5	Western	Nandini Layout	13.010406	77.537803
6	Western	Nayandahalli	12.941325	77.521212
7	Western	Rajajinagar	12.988234	77.554883
8	Western	Rajarajeshwari Nagar	12.927441	77.515522
9	Western	Vijayanagar	12.971889	77.545789

65 rows × 4 columns

Source/Tool: Python package Nominatim in geopy.geocoders.

Data 3: Contains the different venues in each of the neighborhoods

- To retrieve the venues in all the neighborhoods, the Foursquare API has been used.
- The final dataframe containing the regions, neighborhoods, their coordinates, the venues and the venue coordinates & category is as shown:

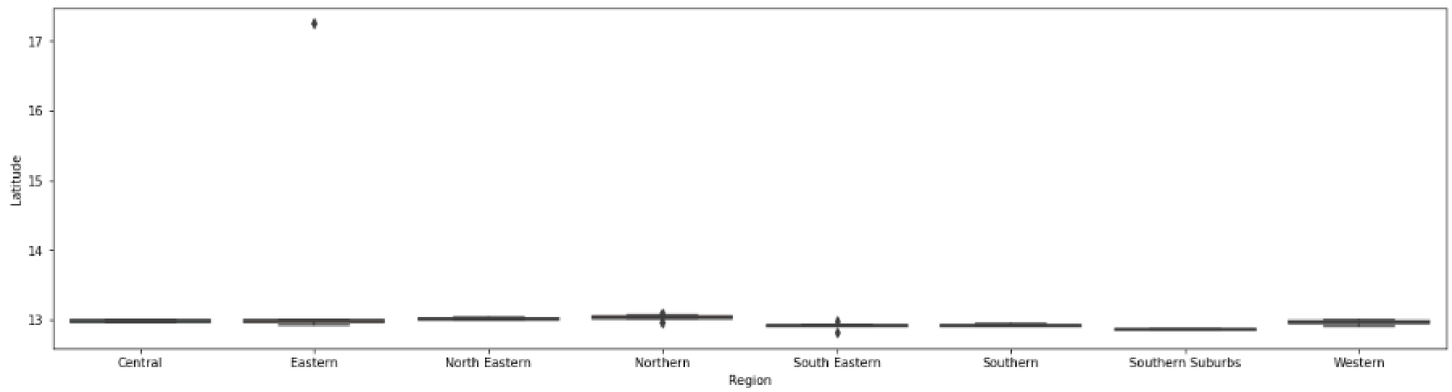
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Cantonment	12.979120	77.591300	Vidhana Soudha	12.979027	77.591881	Capitol Building
1	Cantonment	12.979120	77.591300	Cubbon Park	12.977042	77.595277	Park
2	Cantonment	12.979120	77.591300	Lobby @ ITC Gardenia	12.976568	77.589033	Garden
3	Cantonment	12.979120	77.591300	Wine Board	12.975636	77.590058	Vineyard
4	Domlur	12.962467	77.638196	Lavonne	12.963909	77.638579	Café
...	...	...	...	...	...	...	...
658	Vijayanagar	12.971889	77.545789	Natti style, vijayanagar	12.972908	77.544356	Indian Restaurant
659	Vijayanagar	12.971889	77.545789	Adichunchanagiri Kalyana Mantap	12.973807	77.545510	Arcade
660	Vijayanagar	12.971889	77.545789	Ruchi Bhel Centre	12.969259	77.544368	Indian Restaurant
661	Vijayanagar	12.971889	77.545789	reliance fresh	12.974825	77.548303	Department Store
662	Vijayanagar	12.971889	77.545789	just bake	12.973699	77.541899	Dessert Shop

Source/Tool: Foursquare API

2.2 Data Cleaning

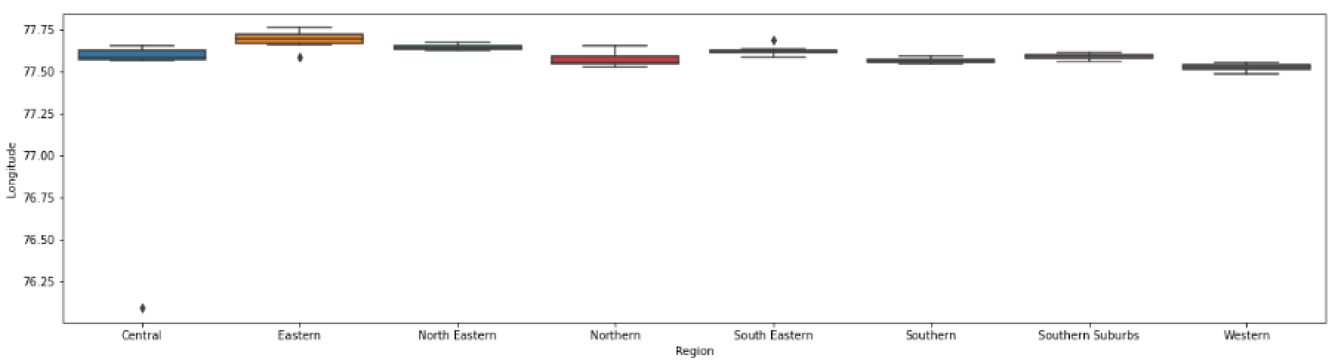
Since we have used the Geopy package wherein the addresses of various neighborhoods are used for finding the location coordinates, there is scope for ambiguity due to similarity in the names of neighborhoods in other locations.

Due to this some latitudes and longitudes were wrong. This was observed in the boxplots shown below:



Here we can see the inconsistencies in the latitude of a neighborhood in the Eastern region. Since it is just one neighborhood, I filtered it out and googled the correct coordinates.

After cleaning the data for latitudes, I observed the longitudes just to double check the data. Again there was an outlier as shown:



I found this neighborhood in the Central region, and googled the correct coordinates.

There was also a neighborhood for which Geopy did not have the coordinates available. I singled out this neighborhood through exception handling while retrieving the coordinates for all the neighborhoods. The code is shown below:

```
Latitude = []
Longitude = []

for neighborhood in neighborhoods['Neighborhood']:

    try:
        address = '{}, Bengaluru'.format(neighborhood)
        geolocator = Nominatim(user_agent="my_first_project")
        location = geolocator.geocode(address)
        Latitude.append(location.latitude)
        Longitude.append(location.longitude)


    except:
        Latitude.append('NA')
        Longitude.append('NA')




neighborhoods['Latitude'] = Latitude
neighborhoods['Longitude'] = Longitude

neighborhoods
```

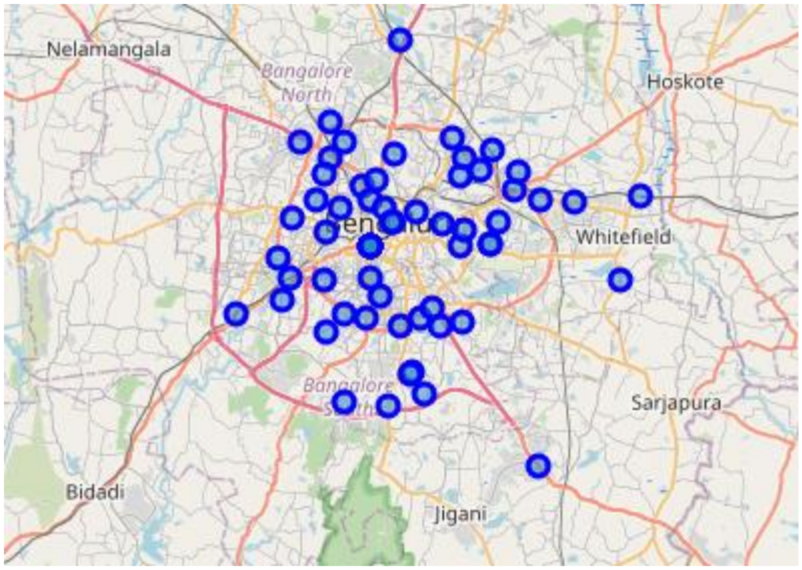

Again I googled the correct coordinates of this neighborhood.

Finally, the data procured was clean, and further manipulations were made on it.

3. Methodology

3.1 Pre Processing:

After cleaning the data and obtaining the desired datasets, I plotted the neighborhoods on the map of Bangalore, to see the neighborhoods geographically in the regions mentioned in the dataframe.



Now to achieve the purpose of this project and to solve the problem at hand, similar neighborhoods needed to be grouped together. This could be best done using clustering. So I used K-Means clustering.

To measure the similarity between neighborhoods, in terms of offering facilities and amenities to the dwellers, the comparison was made on the basis of the categories of venues in each neighborhood.

Therefore, I used one hot encoding for the various venue categories. This helped me in finding the frequency of each kind of venue in every neighborhood. The table below shows the categories after one hot encoding.

	Neighborhood	ATM	American Restaurant	Andhra Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	...	Tea Room	Temple	Tex-Mex Restaurant	Then Pa
0	Cantonment	0	0	0	0	0	0	0	0	0	...	0	0	0	
1	Cantonment	0	0	0	0	0	0	0	0	0	...	0	0	0	
2	Cantonment	0	0	0	0	0	0	0	0	0	...	0	0	0	
3	Cantonment	0	0	0	0	0	0	0	0	0	...	0	0	0	
4	Cantonment	0	0	0	0	0	0	0	0	0	...	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
668	Rajarajeshwari Nagar	0	0	0	0	0	0	0	0	0	...	0	0	0	
669	Rajarajeshwari Nagar	0	0	0	0	0	0	0	0	0	...	0	0	0	
670	Vijayanagar	0	0	0	0	0	0	0	0	0	...	0	0	0	
671	Vijayanagar	0	0	0	0	0	0	0	0	0	...	0	0	0	
672	Vijayanagar	0	0	0	0	0	0	0	0	0		0	0	0	

Now I grouped the dataframe above by ‘Neighborhoods’, I took the mean of the one hot encoded values while grouping them. In the result obtained, each value showed the exact frequency of occurrence of a type (category) of venue in the neighborhood, as shown:

	Neighborhood	ATM	American Restaurant	Andhra Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	...	Tea Room	Temple	Tex-Mex Restaurant	Th
0	Anjanapura	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	
1	Arekere	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	
2	BTM Layout	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	
3	Banashankari	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	
4	Banaswadi	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
60	Vidyaranyapura	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	
61	Vijayanagar	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	
62	Whitefield	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	
63	Yelahanka	0.333333	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	
64	Yeshwanthpur	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.058824	0.0	...	0.0	0.0	0.0	

This was the final dataframe obtained.

3.2 K-Means Clustering:

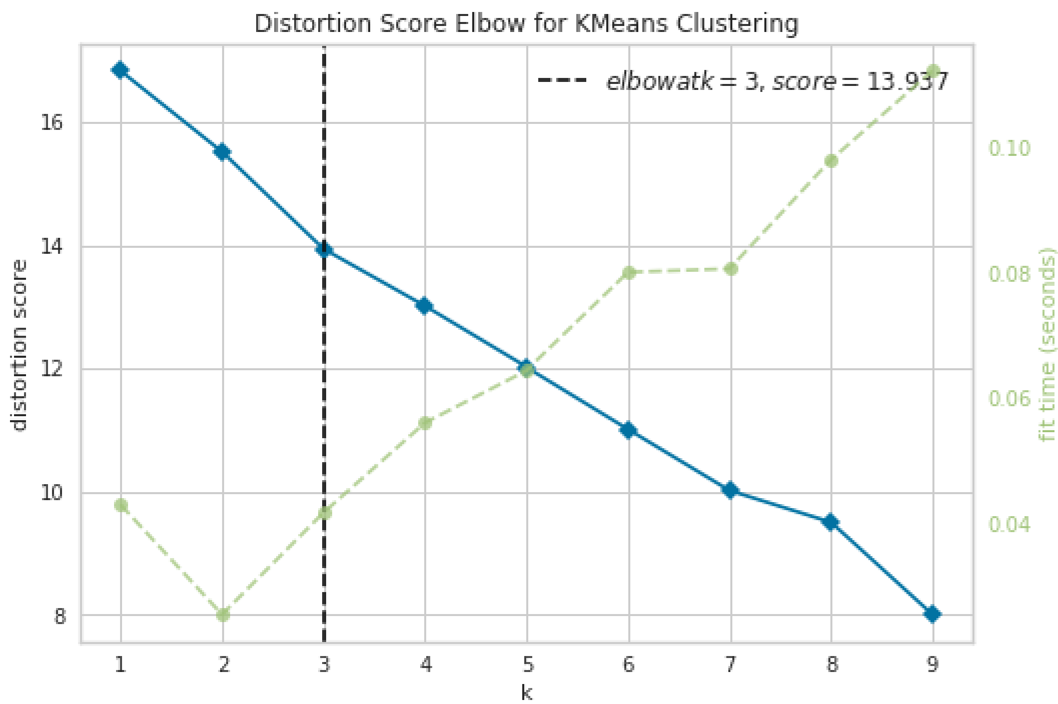
I performed K means clustering on the dataframe above. The distance between data points was measured in terms of Euclidean distance.

To find the best value of K, I made use of KElbowVisualizer in the yellowbrick package.

```
from yellowbrick.cluster import KElbowVisualizer
```

```
model = KMeans()
visualizer = KElbowVisualizer(model, k=range(1,10))
visualizer.fit(cluster_df)
visualizer.show()
```

On plotting the inertia for each value of K from 1 to 9, we got:



The elbow point was at K=3, therefore the model was optimum for 3 clusters.

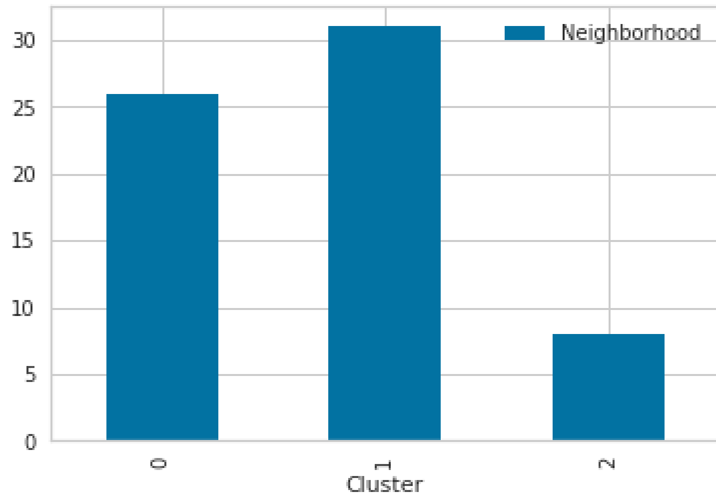
After performing K Means clustering for K=3, we obtained the clusters for each neighborhood.

4. Result:

All the neighborhoods of Bangalore were divided in 3 clusters;

- Cluster 0
- Cluster 1
- Cluster 2

as shown:



I formed dataframes to view the top 10 most common venues in all the neighborhoods in each cluster. The top 10 venues would give an idea about the kind of cluster. The dataframe for each cluster is shown below.

Cluster 1

	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	Indiranagar	Pub	Lounge	Ice Cream Shop	Cocktail Bar	Restaurant	Café	Cupcake Shop	Italian Restaurant	Indian Restaurant	Bakery
3	Jeevanbheemanagar	Bakery	Yoga Studio	Donut Shop	Food & Drink Shop	Flea Market	Fast Food Restaurant	Farmers Market	Electronics Store	Diner	French Restaurant
4	Malleswaram	Indian Restaurant	Vegetarian / Vegan Restaurant	Ice Cream Shop	Coffee Shop	Fast Food Restaurant	Donut Shop	Other Nightlife	Snack Place	Café	Social Club
5	Pete	Market	Indian Restaurant	Historic Site	Bookstore	Yoga Studio	Donut Shop	Flea Market	Fast Food Restaurant	Farmers Market	Electronics Store
8	Shivajinagar	Indian Restaurant	Men's Store	Donut Shop	Market	Clothing Store	Fast Food Restaurant	Asian Restaurant	Tea Room	South Indian Restaurant	Yoga Studio
9	Ulsoor	Hotel	Indian Restaurant	Café	Light Rail Station	Donut Shop	Flea Market	Fast Food Restaurant	Farmers Market	Electronics Store	Diner
11	Bellandur	Park	Indian Restaurant	Sandwich Place	Restaurant	Capitol Building	Cupcake Shop	Deli / Bodega	Department Store	Dessert Shop	Flea Market

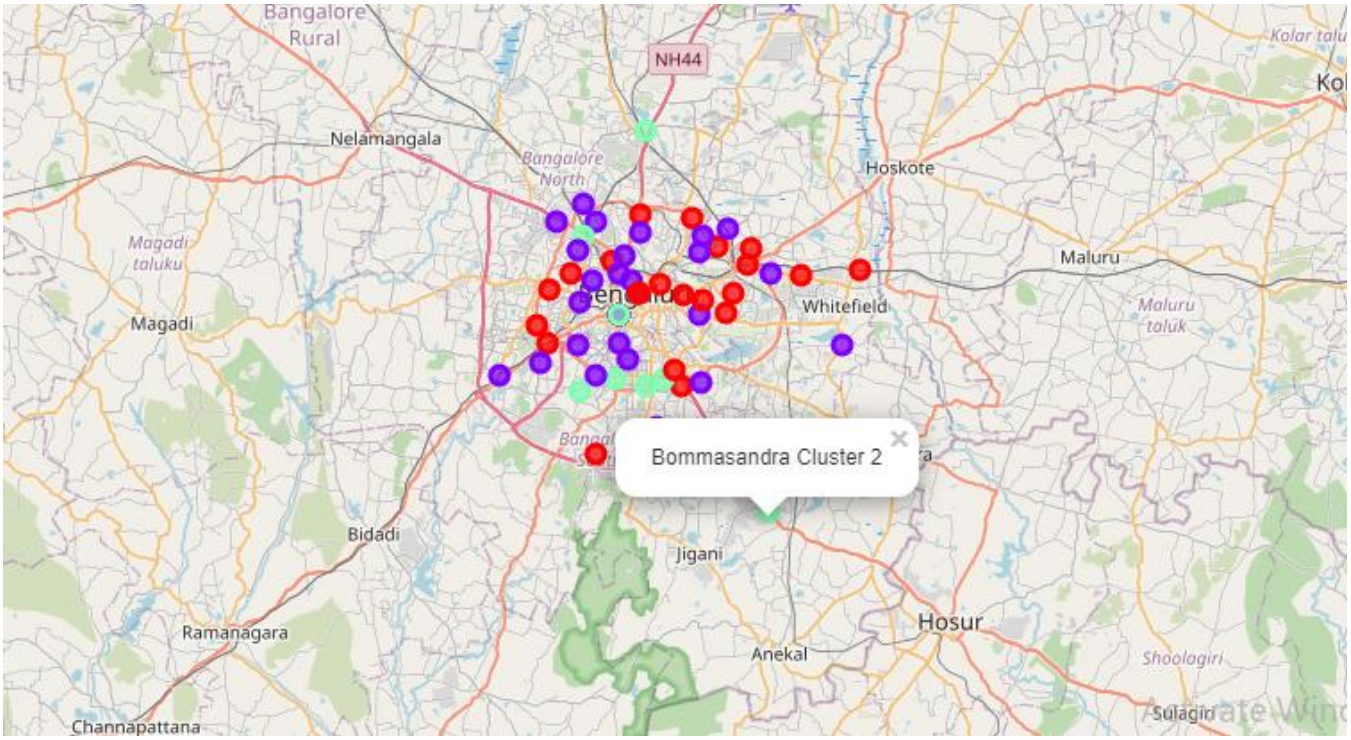
Cluster 2

	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	Cantonment	Park	Indian Restaurant	Sandwich Place	Restaurant	Capitol Building	Cupcake Shop	Deli / Bodega	Department Store	Dessert Shop	Flea Market
1	Domlur	Indian Restaurant	Café	Sports Bar	Liquor Store	Italian Restaurant	Restaurant	Lounge	Rajasthani Restaurant	Pub	Food & Drink Shop
6	Sadashivanagar	Coffee Shop	Indian Restaurant	Department Store	Café	Seafood Restaurant	Women's Store	Gym	Gourmet Shop	Dessert Shop	Performing Arts Venue
7	Seshadripuram	Clothing Store	Indian Restaurant	Coffee Shop	Department Store	Fast Food Restaurant	Juice Bar	Chaat Place	Shopping Mall	Restaurant	Miscellaneous Shop
10	Vasanth Nagar	Hotel	Indian Restaurant	Nightclub	Coffee Shop	Golf Course	Vietnamese Restaurant	Hotel Bar	Gym	Art Gallery	Art Museum
15	Mahadevapura	Coffee Shop	Movie Theater	Bar	Noodle House	Clothing Store	Multiplex	Fast Food Restaurant	Shopping Mall	French Restaurant	Convenience Store
17	Varthur	Candy Store	Department Store	Yoga Studio	Food & Drink Shop	Flea Market	Fast Food Restaurant	Farmers Market	Electronics Store	Donut Shop	Diner

Cluster 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
32	Yelahanka	ATM	Train Station	Motorcycle Shop	Diner	Flea Market	Fast Food Restaurant	Farmers Market	Electronics Store	Donut Shop	Dim Sum Restaurant
33	Yeshwanthpur	Bus Station	Train Station	Indian Restaurant	Fast Food Restaurant	Multiplex	Asian Restaurant	Light Rail Station	Miscellaneous Shop	Movie Theater	Shopping Mall
35	Bommasandra	Sporting Goods Shop	Yoga Studio	Diner	Flea Market	Fast Food Restaurant	Farmers Market	Electronics Store	Donut Shop	Dim Sum Restaurant	Food Court
36	BTM Layout	Indian Restaurant	Ice Cream Shop	Snack Place	Vegetarian / Vegan Restaurant	Bakery	Chinese Restaurant	Fast Food Restaurant	Sandwich Place	Coffee Shop	Italian Restaurant
41	Banashankari	Juice Bar	Indian Restaurant	Liquor Store	Metro Station	Donut Shop	Food & Drink Shop	Flea Market	Fast Food Restaurant	Farmers Market	Electronics Store
44	J. P. Nagar	Indian	Café	Pizza Place	Pub	Garden	Mediterranean	Shopping	Snack Place	Bookstore	South Indian

To make it more intuitive for the audience, I plotted these clusters on the map, by clicking on the circle marker, one can see the neighborhood and the cluster it is in:



5. Discussion:

The city's neighborhoods were divided into 3 clusters. Cluster 0 has 26 neighborhoods, Cluster 1 had 31 and Cluster 2 had 8. It can be seen in the map that these clusters are distributed all over the city. So anyone looking to relocate to any part of the city is likely to find a neighborhood similar to his/her old neighborhood.

The result will be helpful for people relocating within the city or even visiting or settling from outside, as one glance of the map will ease their decision making. Though there is still scope for more comprehensive results as the location data provided on Foursquare is limited for Indian cities.

6. Conclusion:

This was my Capstone project for the Data Science Professional Certificate offered by IBM on Coursera. I would like to thank all my instructors for sparking the interest for Data Science in me, helping me understand all the concepts thoroughly and building my practical skills. I hope this project is of help to anyone who is relocating to or within Bangalore and it finds them in good health, given the testing times due to the pandemic.