Bengaluru Neighborhoods Clustering

Introduction

This project is aimed at clustering the neighbourhoods of Bengaluru, India to

- Find similar neighbourhoods.
- Help people find the most visited places in their neighbourhood
- Find which neighbourhoods have excess/ lack of restaurants etc
- Which neighbourhoods are good for residential purposes

Bengaluru is one of the largest metropolitan cities in India. This project aims to dive into the various neighbourhoods in the city. Finding the similarity between neighbourhoods can help people decide on questions like which neighbourhoods are the best for residential or business purposes. People who are unable to find a property in one of the neighbourhoods might be interested in this data which gives insight on the similarity between neighbourhoods.

Data

The data sources that I've used for this project are

- Wikipedia - https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore
- FourSquare API
- Geocoder API

Web scraping using beautiful soup

I obtained neighborhoods of Bangalore by scraping the tables from the wikipedia page on Bengaluru. I obtain the neighborhoods of Bangalore along with the region of Neighborhoods as shown below.

	Region	Neighbourhood
47	Southern	Padmanabhanagar, Bengaluru
34	South-Eastern	Bommanahalli, Bengaluru
31	Northern	Vidyaranyapura, Bengaluru
25	North-Eastern	Ramamurthy Nagar, Bengaluru
16	Eastern	Marathahalli, Bengaluru

Geocoder for coordinates data

Using geocoder package, I was able to obtain the coordinates for the various neighbourhoods of the bengaluru like the example shown below.

```
address = 'Koramangala, Bengaluru'

geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Koramangala are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of Koramangala are 12.9340114, 77.6222304.
```

Foursquare for venues

Using the Foursquare API, I was able to obtain the venues data for geographic coordinates. I fetched 25 venues per neighborhood from the foursquare data and used the same for my clustering analysis.

```
The geograpical coordinate of Shivajinagar, Bengaluru are 12.986391, 77.6075416.
The geograpical coordinate of Ulsoor, Bengaluru are 12.9778793, 77.6246697.
The geograpical coordinate of Vasanth Nagar, Bengaluru are 12.98872125, 77.5851687760182.
The geograpical coordinate of Bellandur, Bengaluru are 12.9791198, 77.5912997.
The geograpical coordinate of CV Raman Nagar, Bengaluru are 17.2510682, 80.1651978.
The geograpical coordinate of Hoodi, Bengaluru are 12.9919033, 77.7162015.
```

Methodology

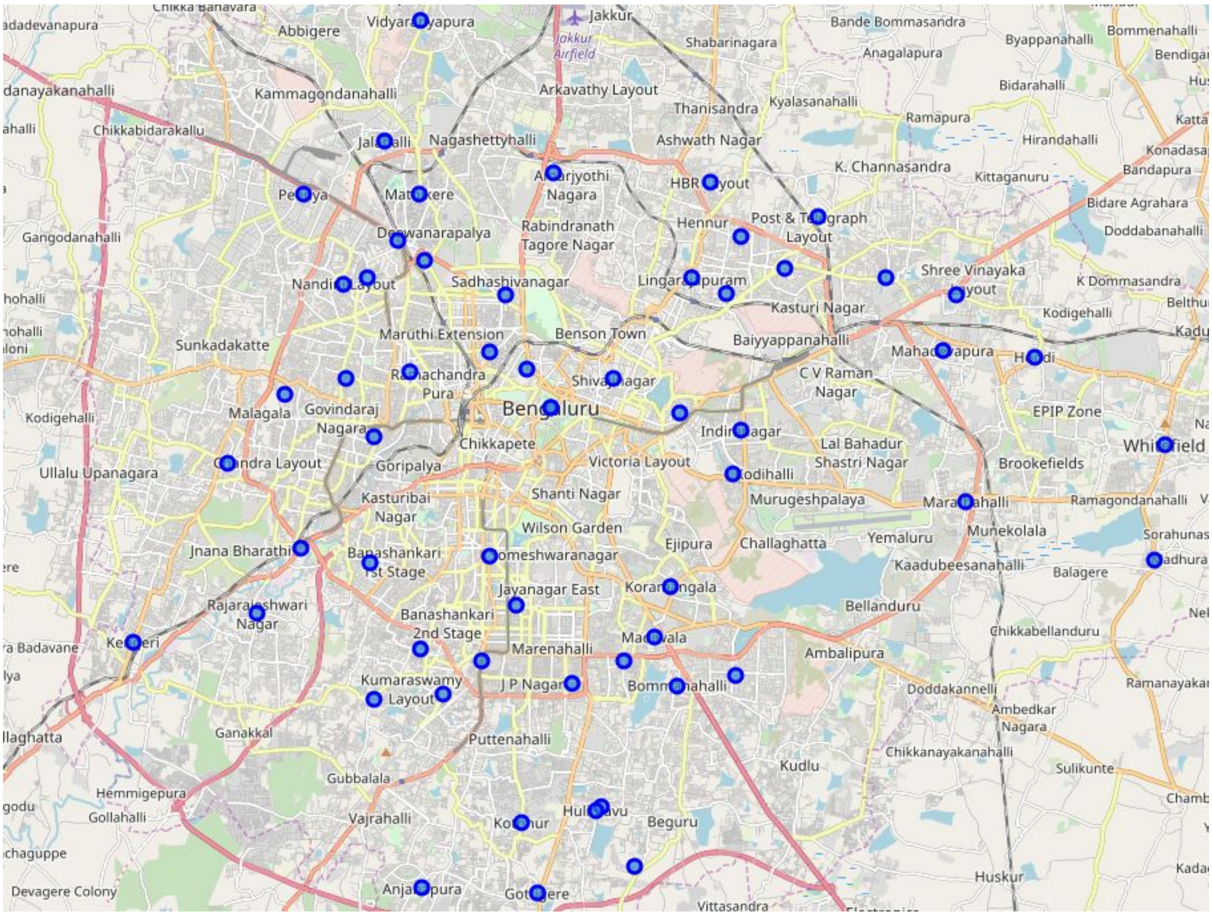
As a result of the data acquisition process explained above, I was able to to obtain a clean dataset with region, neighborhood, latitude and longitude values as shown below.

```
1 neighbourhood_data.head()
```

	index	Region	Neighbourhood	Latitude	Longitude
0	1	Central	Domlur, Bengaluru	12.962467	77.638196
1	2	Central	Indiranagar, Bengaluru	12.973291	77.640467
2	4	Central	Malleswaram, Bengaluru	13.016341	77.558664
3	6	Central	Sadashivanagar, Bengaluru	13.007708	77.579589
4	7	Central	Seshadripuram, Bengaluru	12.993188	77.575342

Plotting using folium

I plotted this data on a map using the folium library. The resulting map showed the various neighbourhoods of Bengaluru at their respective locations.



Getting the venues data from foursquare api

Using the location data, I obtained venue information from the foursquare api - upto 25 venues per neighborhood within a radius of 2000 metres. I ended up with a resulting dataset of 1297 rows and 6 columns.

One Hot Encoding the dataset

Having obtained the venues for neighborhoods, I one-hot encoded the dataset for venue categories to find the similarity and differences between neighborhoods and the resulting dataset looks like it does below.

	Neighborhood	Afghan Restaurant	American Restaurant	Andhra Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	...
0	Domlur, Bengaluru	0	0	0	0	0	0	0	0	0	...
1	Domlur, Bengaluru	0	0	0	0	0	0	0	0	0	...
2	Domlur, Bengaluru	0	0	0	0	0	0	0	0	0	...

To find which neighborhood has more of a particular type of venue categories, I grouped the one-hot encoded dataset by Neighborhood and found the average occurrence of each venue types. With this, I was able to obtain the most common types of venues for each neighborhood. The resulting dataset ended up as shown below

	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	Anjanapura, Bengaluru	Pool	Lounge	Train Station	Residential Building (Apartment / Condo)	Yoga Studio	Dive Bar	Falafel Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant
1	Arekere, Bengaluru	Ice Cream Shop	Pizza Place	Indian Restaurant	Café	Multiplex	Beer Garden	Department Store	Chinese Restaurant	Dumpling Restaurant	Rajasthani Restaurant
2	BTM Layout, Bengaluru	Ice Cream Shop	Indian Restaurant	Bakery	Indie Movie Theater	Burger Joint	Gym	Italian Restaurant	Garden	Furniture / Home Store	Mediterranean Restaurant
3	Banashankari, Bengaluru	Ice Cream Shop	Indian Restaurant	Fast Food Restaurant	South Indian Restaurant	Performing Arts Venue	Burger Joint	Seafood Restaurant	Breakfast Spot	Snack Place	Café
4	Banaswadi, Bengaluru	Indian Restaurant	Ice Cream Shop	Korean Restaurant	BBQ Joint	Bakery	Bistro	Pizza Place	Pub	Chinese Restaurant	Falafel Restaurant

K-means Clustering

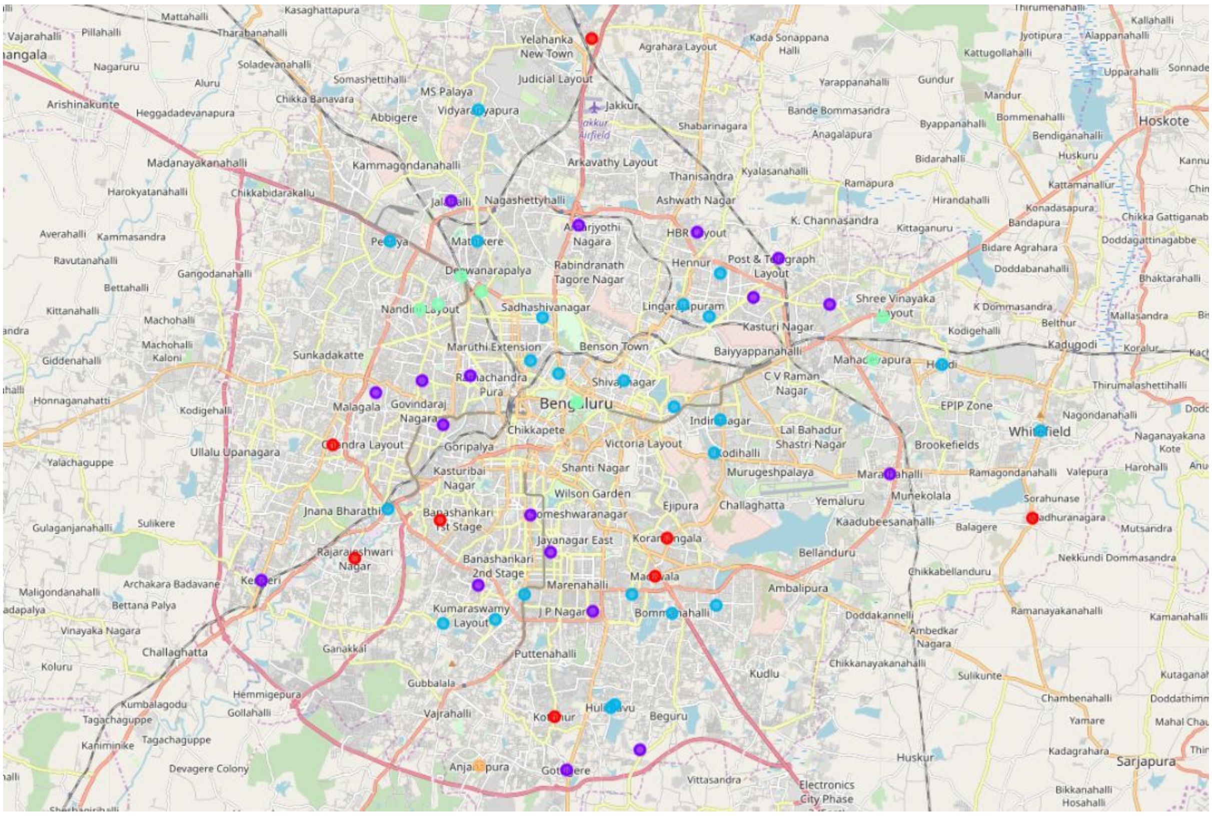
Setting the number of clusters to 5, I clustered the one-hot encoded data using k-means unsupervised clustering algorithm. Assigning the labels to the neighborhood values, I was able to obtain a clustered neighborhood dataset.

```
kclusters = 5
bengaluru_grouped_clustering = bengaluru_grouped.drop('Neighborhood', 1)
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(bengaluru_grouped_clustering)
kmeans.labels_[0:10]
```

```
array([4, 2, 2, 2, 1, 1, 1, 1, 3, 2], dtype=int32)
```

Results

Using the k-means clustering algorithm, the neighborhoods were successfully clustered into 5 clusters. The resulting clustered neighborhoods are shown over the map of Bengaluru in the figure below.



Discussion

The following interesting observations were obtained from this analysis

- There seem to be only 2 clusters near the center of the city.
- First Cluster comprising of neighborhoods - Varthur, Yelahanka, Koramangala, Madiwala, Girinagar, Kothnur, Nagarbhavi and Rajarajeshwari Nagar have a high frequency of cafes and breakfast spots. A new business in the area can take leverage of this information. For example, as this area is already saturated with cafes, a new cafe in the area will face a heavy competition from the established ones. However offering something which complements the cafe experience, may find a higher success rate in these areas.
- The second cluster comprising of Marathahalli, Banaswadi, HBR Layout, Horamavu, Ramamurthy Nagar, Hebbal, Jalahalli, Basavanagudi, J. P. Nagar, Jayanagar, Padmanabhanagar, Begur, Gottigere, Basaveshwaranagar, Kamakshipalya, Kengeri, Rajajinagar and Vijayanagar are away from the center and have an excess of Indian Restaurants and Ice cream shops. Again new players and established brands can leverage this information.

- The third cluster comprises of the following neighborhoods - Domlur, Indiranagar, Sadashivanagar, Seshadripuram, Shivajinagar, Ulsoor, Vasanth Nagar, Hoodi, Whitefield, Kalyan Nagar, Kammanahalli, Lingarajapuram, Mathikere, Peenya, Vidyaranyapura, Bommanahalli, Bommasandra, BTM Layout, HSR Layout, Banashankari, Kumaraswamy Layout, Uttarahalli, Arekere, Hulimavu, Nayandahalli. These neighborhoods have an abundance of ice cream shops, hotels and indian restaurants.
- The fourth cluster comprises of Malleswaram, Bellandur, Krishnarajapuram, Mahadevapura, Yeshwanthpur, Electronic City, Mahalakshmi Layout, Nandini Layout. These neighborhoods have an abundance of lounges, multiplexes and coffee shops.
- The fifth cluster comprises of a single neighborhood - Anjanapura.

Conclusion

In this study, I was able to cluster the different neighborhoods of Bengaluru. I leveraged the neighborhood data from wikipedia, obtained the location data through geocoder, and used the foursquare api to obtain the information about venues in each neighborhood. Then I clustered the various neighborhoods using the k-means unsupervised learning algorithm and obtained the clusters. Hence I was able to determine which neighborhoods were similar to each other and obtain the type of venues that are common in those neighborhoods.