APPLIED DATA SCIENCE CAPSTONE-IBM

The Silicon Valley of India – Bangalore

JULY 28, 2021

AUTHORED BY
MEHATAB NABI S

IBM Applied Data Science Capstone

The Silicon Valley of India – Bangalore

1 Introduction

1.1 Background

Bangalore officially known as Bengaluru is the capital and the largest city of the Indian state of Karnataka. It has a population of more than 8 million and a metropolitan population of around 11 million, making it the third most populous city and fifth most populous urban agglomeration in India. Located in southern India on the Deccan Plateau, at a height of over 900 m (3,000 ft.) above sea level, Bangalore is known for its pleasant climate throughout the year. Its elevation is the highest among the major cities of India. Bangalore is widely regarded as the "Silicon Valley of India" (or "IT capital of India") because of its role as the nation's leading information technology (IT) exporter. Indian technological organizations are headquartered in the city. A demographically diverse city, Bangalore is the second fastest-growing major metropolis in India. Recent estimates of the metro economy of its urban area have ranked Bangalore either the fourth- or fifth-most productive metro area of India. As of 2017, Bangalore was home to 7,700 millionaires and 8 billionaires with a total wealth of $320 billion. It is home to many educational and research institutions. Numerous state-owned aerospace and defense organizations are located in the city. The city also houses the Kannada film industry. It was ranked the most livable Indian city with a population of over a million under the Ease of Living Index 2020.

1.2 Problem

The city suffers, however, from some of the perennial problems of many large expanding industrial cities like - air and water pollution, widespread areas of substandard housing, and overcrowding. With its diverse society, comes diverse infrastructure which decides the quality of living. Infrastructure in Bangalore is very spread out and unique - belonging to different categories like Drinking Water Plant, Waste Water/Sewage, Hospitals, Schools, Colleges, Railway Network, Electricity Power Plants, Telecommunication Support, Bank, Shopping malls, Supermarket, Gas Station, Hotels, Police Station, Café, medical shops, grocery shops, theatre, etc. One of the main problems, when one moves to a new city, is finding a good area to live in, settle down and grow prosperously.

1.3 Interest

The questions which I aim to answer through this analysis are the following:

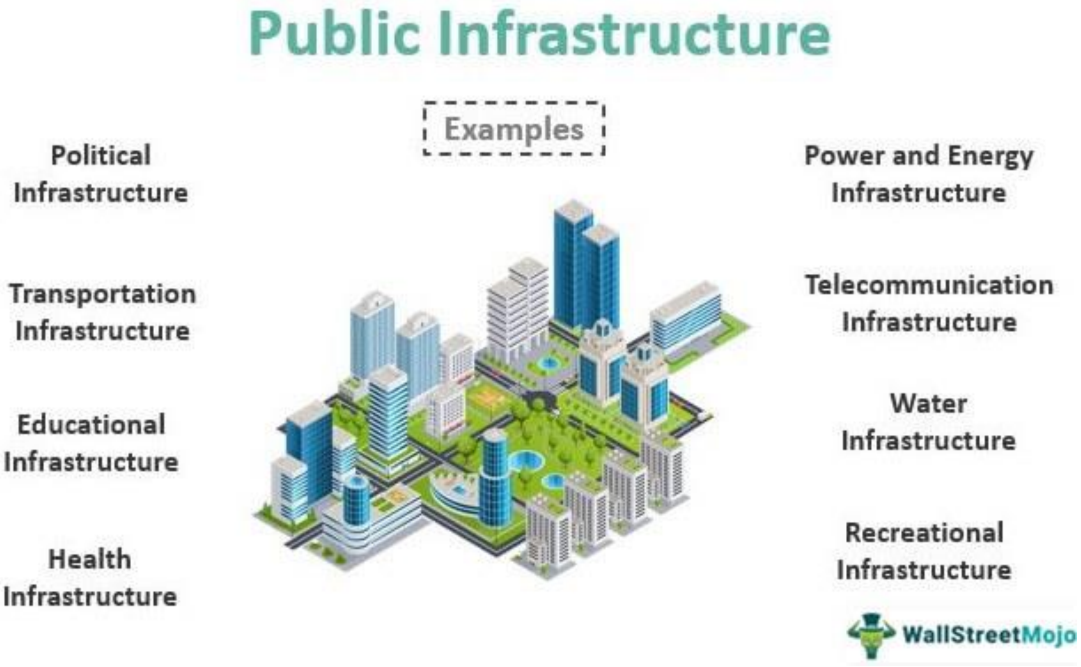
1. List and visualize all major parts of Bangalore with top existing infrastructure.

2. What are the best locations in Bangalore as per the existing infrastructure?

3. Which areas have the potential for the development of infrastructure of different kinds?

4. Which all areas lack the basic infrastructure facilities?

5. What is the best place to stay within the city for all vital infrastructure facilities?


Public Infrastructure
Examples
Political Infrastructure
Transportation Infrastructure
Educational Infrastructure
Health Infrastructure
Power and Energy Infrastructure
Telecommunication Infrastructure
Water Infrastructure
Recreational Infrastructure
WallStreetMojo

1.4 Target Audience

The purpose of this project is to help people in exploring better facilities around their neighborhood. It will help people making a smart and informed decision on selecting good neighborhoods in Bangalore, India. Lot of people migrate from various states of India and need lots of research for good housing prices, new business, and reputed professional and safe places for their children. This project is for those people who are looking for better neighborhoods and businesses. It will help people get the awareness of the area and neighborhood before moving to a new city, state, country, or place for their work or to start a new fresh life.

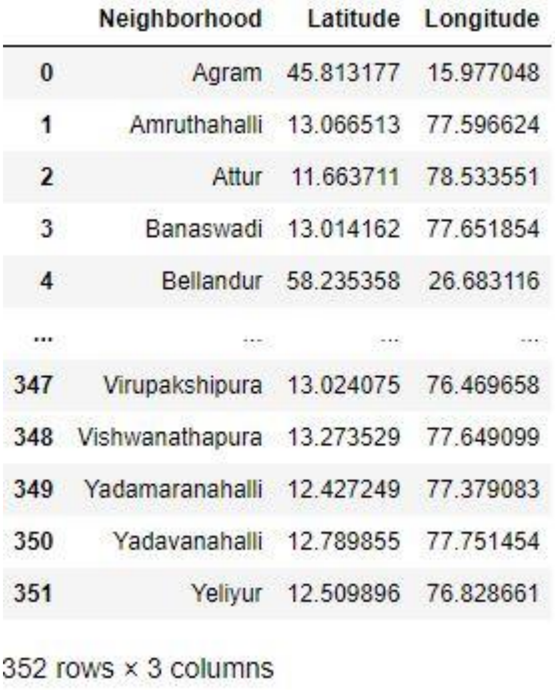
2 Data acquisition and cleaning

2.1 Data sources

Bangalore's demographics show that it is a large and ethnically diverse metropolis. With its diverse society, comes diverse infrastructure. There are many different kinds of infrastructure in the City, each belonging to different categories like Hospitals, Schools, Colleges, Hotels, etc.

For this project we need the following data:

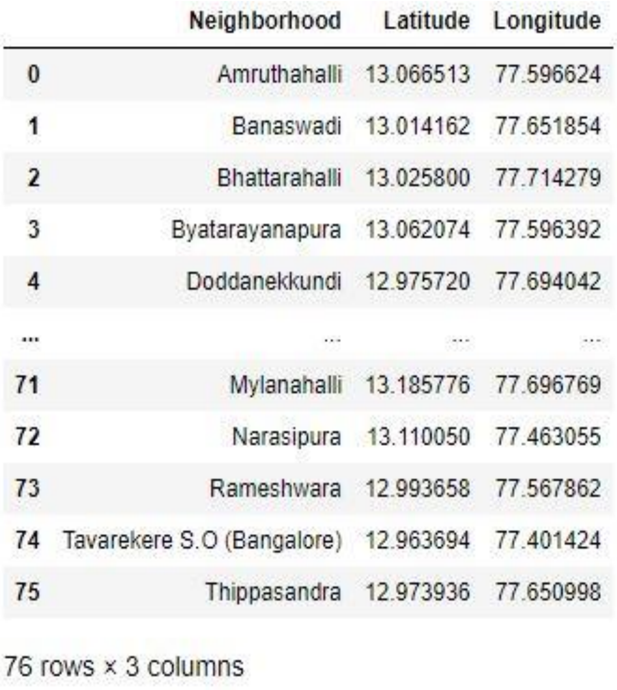
- Geospatial data (Collected from Kaggle datasets)
    - Data source: https://www.kaggle.com/rmenon1998/bangalore-neighborhoods
    - Description: Contain a list of Neighborhoods, latitude and longitude coordinates of the respective area
- Different kinds of infrastructures in each neighborhood of Bangalore.
    - Data source: Foursquare API
    - Description: By using this API we will get all the venues in each Neighborhood. We can filter this data to get different infrastructures and venues.



	Neighborhood	Latitude	Longitude
0	Agram	45.813177	15.977048
1	Amruthahalli	13.066513	77.596624
2	Attur	11.663711	78.533551
3	Banaswadi	13.014162	77.651854
4	Bellandur	58.235358	26.683116
...	...	...	...
347	Virupakshipura	13.024075	76.469658
348	Vishwanathapura	13.273529	77.649099
349	Yadamaranahalli	12.427249	77.379083
350	Yadavanahalli	12.789855	77.751454
351	Yeliyur	12.509896	76.828661

352 rows × 3 columns

Figure 1



	Neighborhood	Latitude	Longitude
0	Amruthahalli	13.066513	77.596624
1	Banaswadi	13.014162	77.651854
2	Bhattarahalli	13.025800	77.714279
3	Byatarayanapura	13.062074	77.596392
4	Doddanekkundi	12.975720	77.694042
...	...	...	...
71	Mylanahalli	13.185776	77.696769
72	Narasipura	13.110050	77.463055
73	Rameshwara	12.993658	77.567862
74	Tavarekere S.O (Bangalore)	12.963694	77.401424
75	Thippasandra	12.973936	77.650998

76 rows × 3 columns

Figure 2

2.2 DATA VISUALIZATION

Visualize the neighbourhoods in a map using Folium package as seen in Fig 1. This allows us to perform a sanity check to make sure that the geographical coordinate's data returned by Geocoder are correctly plotted in the city.

After reading in the Neighbourhoods data from the Kaggle dataset and visualizing the Neighbourhoods using Folium I immediately noticed that we have Neighbourhoods outside of the city and some even in other cities. So I decided to keep only the Neighbourhoods in and around Bangalore and removed all the other ones as visible in Fig 2 and Fig 4.

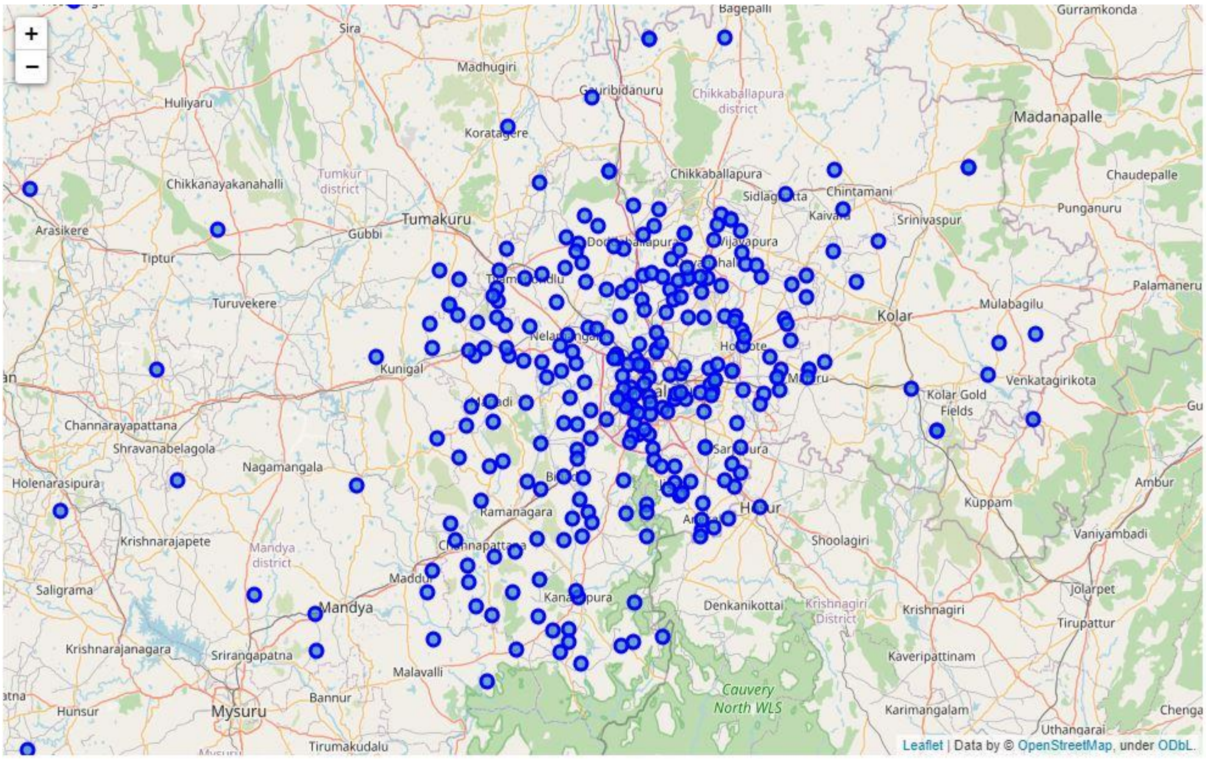


Figure 3

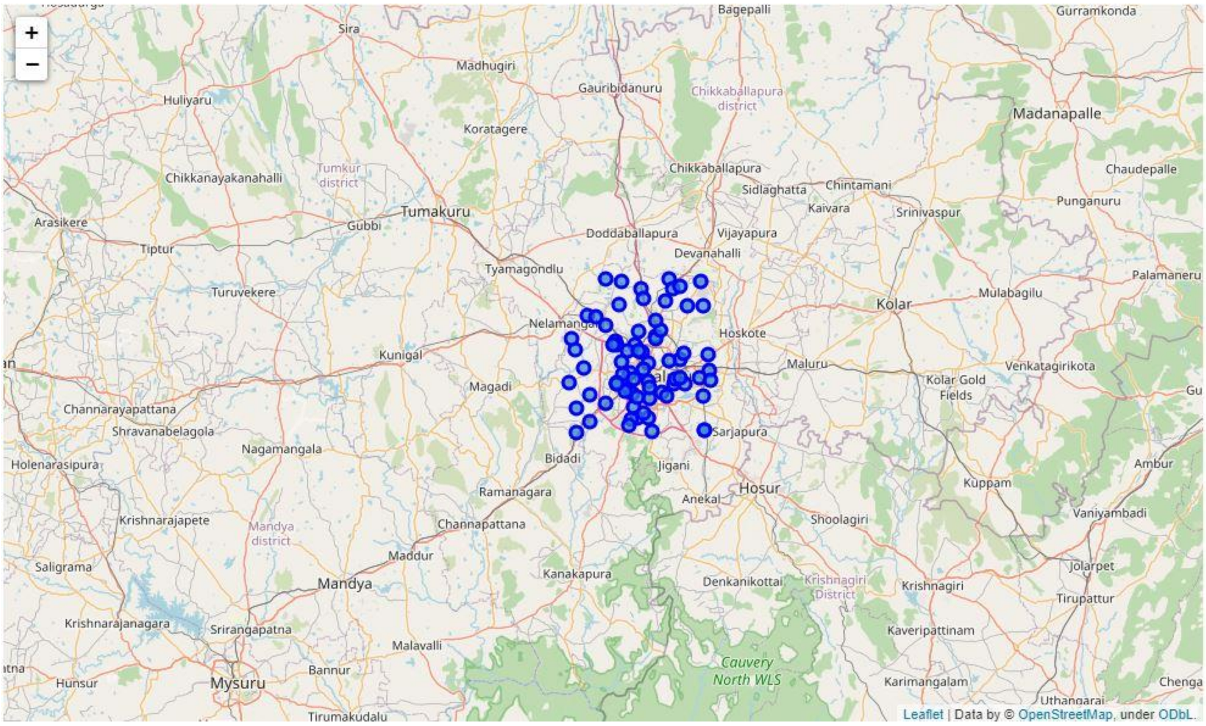
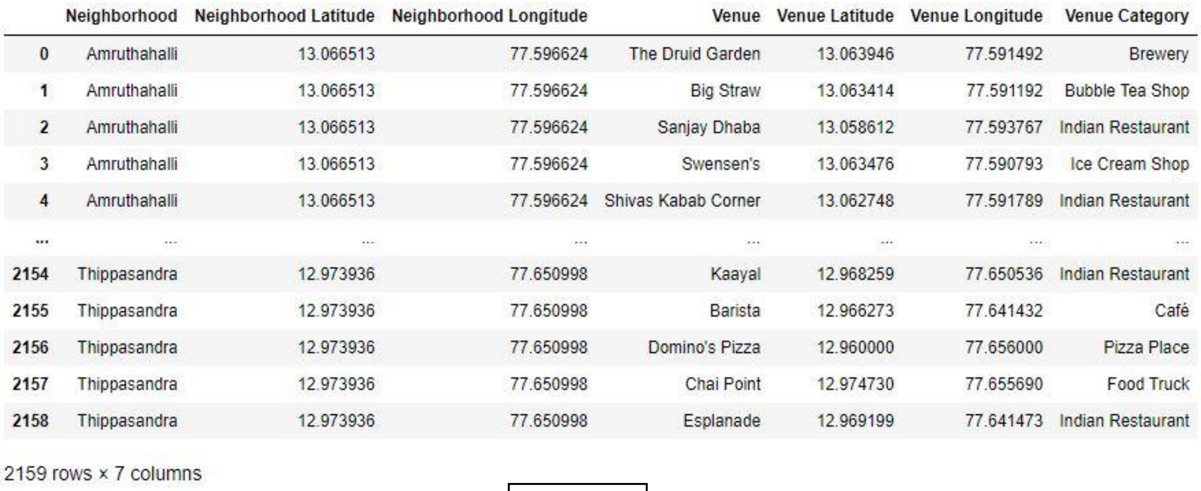


Figure 4

2.3 Explore the Neighborhoods and get the venues data using the Foursquare API

Next, we make use of Foursquare API to get the top 100 venues that are within a radius of 1.8 Km meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude, and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. The API returned back 2159 venues for the 76 Neighborhoods I had selected.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Amruthahalli	13.066513	77.596624	The Druid Garden	13.063946	77.591492	Brewery
1	Amruthahalli	13.066513	77.596624	Big Straw	13.063414	77.591192	Bubble Tea Shop
2	Amruthahalli	13.066513	77.596624	Sanjay Dhaba	13.058612	77.593767	Indian Restaurant
3	Amruthahalli	13.066513	77.596624	Swensen's	13.063476	77.590793	Ice Cream Shop
4	Amruthahalli	13.066513	77.596624	Shivas Kabab Corner	13.062748	77.591789	Indian Restaurant
...	...	...	...	...	...	...	...
2154	Thippasandra	12.973936	77.650998	Kaayal	12.968259	77.650536	Indian Restaurant
2155	Thippasandra	12.973936	77.650998	Barista	12.966273	77.641432	Café
2156	Thippasandra	12.973936	77.650998	Domino's Pizza	12.960000	77.656000	Pizza Place
2157	Thippasandra	12.973936	77.650998	Chai Point	12.974730	77.655690	Food Truck
2158	Thippasandra	12.973936	77.650998	Esplanade	12.969199	77.641473	Indian Restaurant

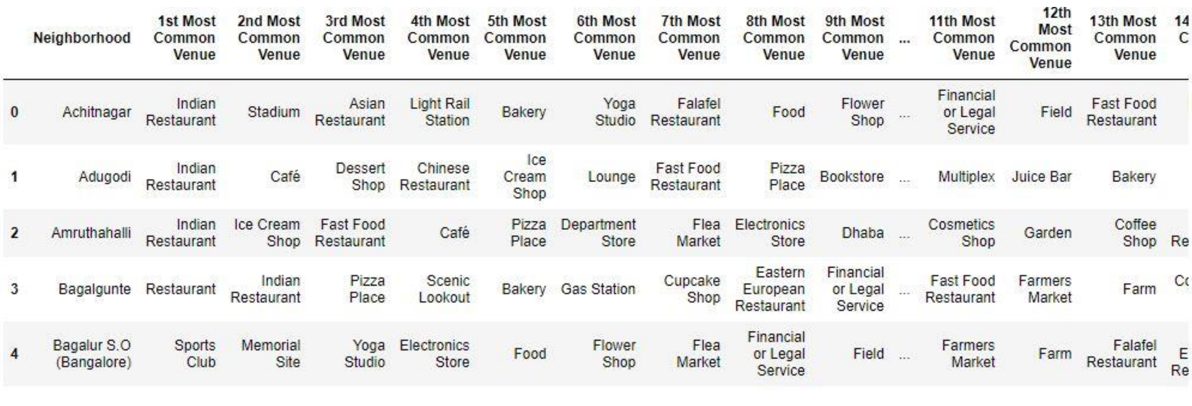
2159 rows × 7 columns

Figure 5

2.4 DATA WRANGLING

We are also preparing the data for use in selection. Based on the occurrence of infrastructures in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new infrastructures and which neighbourhoods are most suitable to visitors to stay.

Using the Venues data from the API I grouped all the Neighborhoods as in Fig 6 to display the top 20 venues of each Neighbourhood to better understand what's more popular there.



	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	...	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14 C
0	Achitnagar	Indian Restaurant	Stadium	Asian Restaurant	Light Rail Station	Bakery	Yoga Studio	Falafel Restaurant	Food	Flower Shop	...	Financial or Legal Service	Field	Fast Food Restaurant	
1	Adugodi	Indian Restaurant	Café	Dessert Shop	Chinese Restaurant	Ice Cream Shop	Lounge	Fast Food Restaurant	Pizza Place	Bookstore	...	Multiplex	Juice Bar	Bakery	
2	Amruthahalli	Indian Restaurant	Ice Cream Shop	Fast Food Restaurant	Café	Pizza Place	Department Store	Flea Market	Electronics Store	Dhaba	...	Cosmetics Shop	Garden	Coffee Shop	Re
3	Bagalgunte	Restaurant	Indian Restaurant	Pizza Place	Scenic Lookout	Bakery	Gas Station	Cupcake Shop	Eastern European Restaurant	Financial or Legal Service	...	Fast Food Restaurant	Farmers Market	Farm	Cc
4	Bagalur S.O (Bangalore)	Sports Club	Memorial Site	Yoga Studio	Electronics Store	Food	Flower Shop	Flea Market	Financial or Legal Service	Field	...	Farmers Market	Farm	Falafel Restaurant	E Re

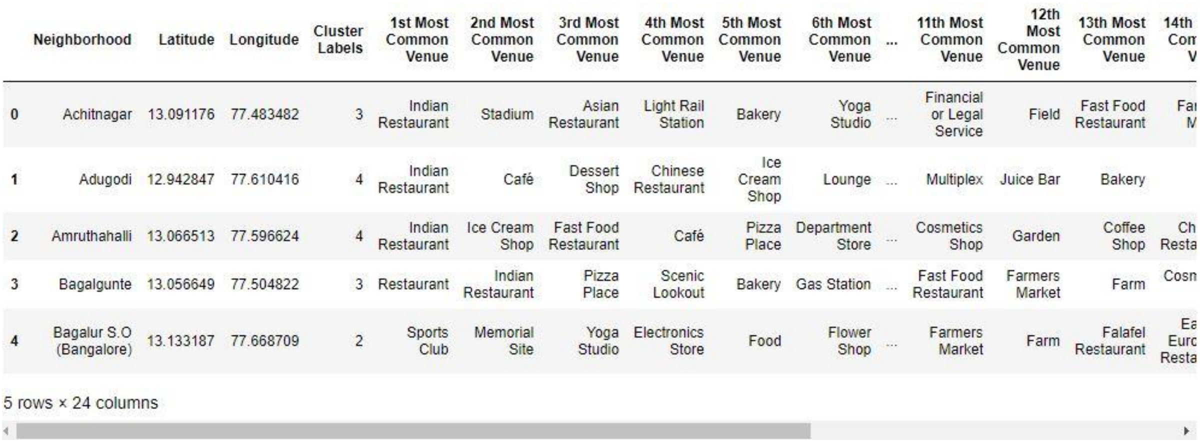
5 rows × 21 columns

Figure 6

3 Modelling & Discussion

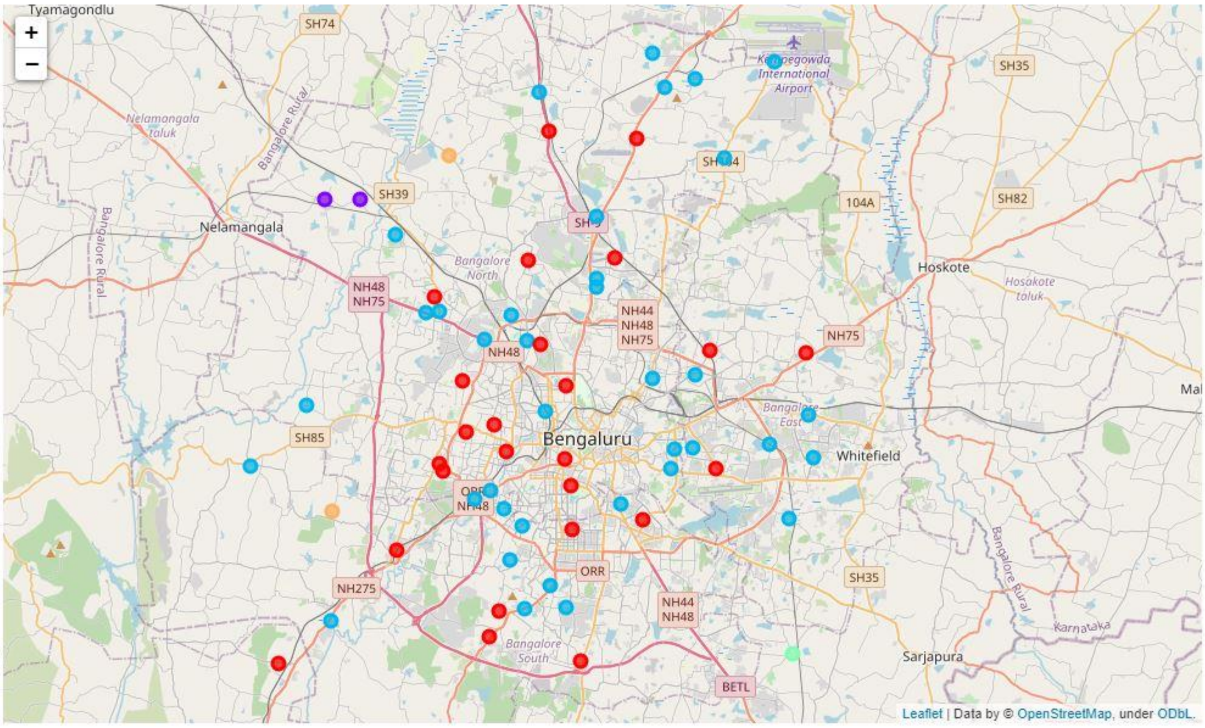
Finally, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies ‘K’ number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 5 clusters based on the similarity of their venue categories.

After applying the K-Means clustering algorithm and obtaining the labels for the Neighborhoods, I merged the labels to the venues data frame to get the final table below.

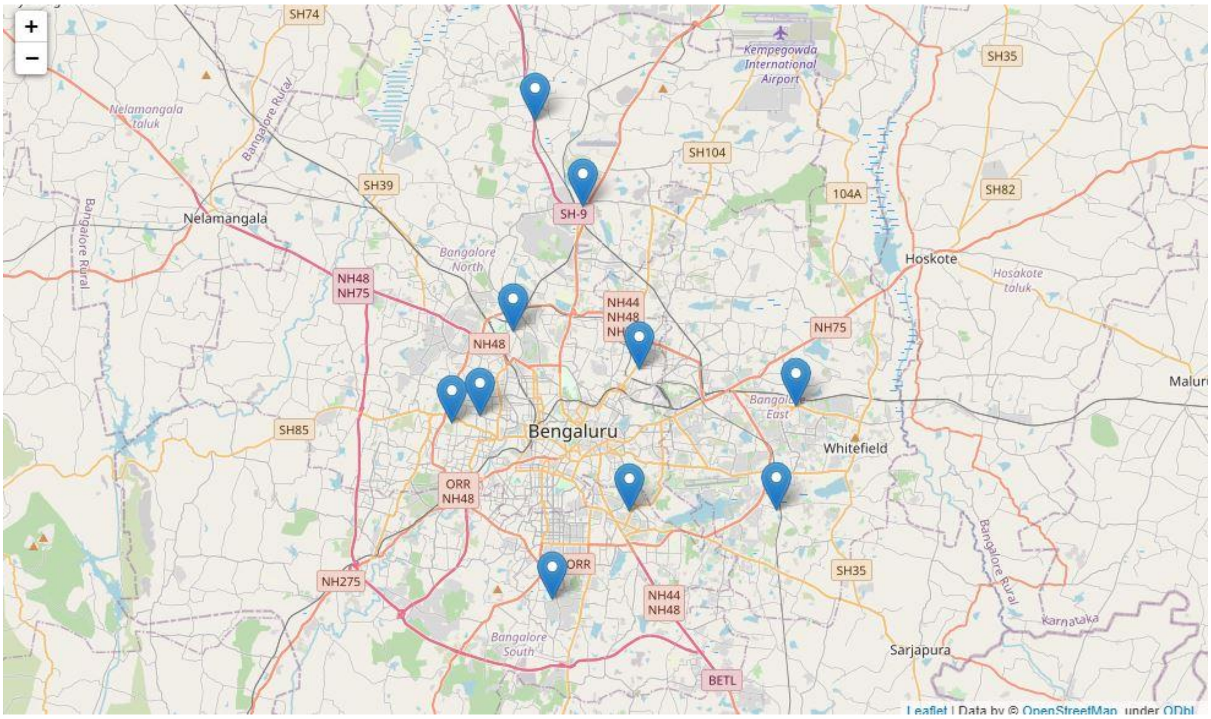


	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	...	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Com V
0	Achitnagar	13.091176	77.483482	3	Indian Restaurant	Stadium	Asian Restaurant	Light Rail Station	Bakery	Yoga Studio	...	Financial or Legal Service	Field	Fast Food Restaurant	Fa M
1	Adugodi	12.942847	77.610416	4	Indian Restaurant	Café	Dessert Shop	Chinese Restaurant	Ice Cream Shop	Lounge	...	Multiplex	Juice Bar	Bakery	
2	Amruthahalli	13.066513	77.596624	4	Indian Restaurant	Ice Cream Shop	Fast Food Restaurant	Café	Pizza Place	Department Store	...	Cosmetics Shop	Garden	Coffee Shop	Ch Resta
3	Bagalgunte	13.056649	77.504822	3	Restaurant	Indian Restaurant	Pizza Place	Scenic Lookout	Bakery	Gas Station	...	Fast Food Restaurant	Farmers Market	Farm	Cosn
4	Bagalur S.O (Bangalore)	13.133187	77.668709	2	Sports Club	Memorial Site	Yoga Studio	Electronics Store	Food	Flower Shop	...	Farmers Market	Farm	Falafel Restaurant	Ea Euro Resta

5 rows × 24 columns



From the above results we can observe that most of the Neighborhoods fall in the 3 and 4 clusters. As my objective was to find Neighborhoods which had more number of Parks and South Indian restaurants, I identified the Neighborhoods with Parks and South Indian restaurants included in the top 20 venues of each Neighbourhood. On observing the Neighborhoods with Parks and South Indian restaurants, I discovered that most of them belong to the 4th cluster and decided that the Neighborhoods in the 4th cluster would be the ideal locations to live in Bangalore and marked them on the map below.



Bangalore is a very big city and heavily populated, so to cluster neighbourhoods based on their venues is a challenging task especially when the all the neighbourhoods are very similar to each other. So in order to get a satisfactory result I set the location radius to 1.8 Km while making the API call for the venues. Most of the Neighborhoods had Indian restaurant as the top venue category. So they would have good Indian restaurants regardless but good South Indian restaurants are fewer compared to general Indian restaurants. Finally, I visualized my ideal locations in Bangalore for people to enjoy their stay in the city.

4 CONCLUSION

I hope this project will prove useful to anyone wanting to move to Bangalore or build a home in the city. Real estate firms can also benefit from such analysis and can take more informed decisions