Bangalore Apartment Venue and Pricing Data Analysis

JULY 12

Authored by: **Ankit Parikh**

Table of Content

1. Introduction

1.1 Background

Bangalore, officially known as Bengaluru is the capital of the Indian state of Karnataka. Bengaluru 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 such as ISRO, Infosys, Wipro, and HAL are headquartered in the city. Bangalore is the second fastest-growing major metropolis in India. It is home to many educational and research institutions in India, such as Indian Institute of Science (IISc), Indian Institute of Management (Bangalore) (IIMB), International Institute of Information Technology, Bangalore (IIITB), National Institute of Fashion Technology, Bangalore, National Institute of Design, Bangalore (NID R&D Campus), National Law School of India University (NLSIU) and National Institute of Mental Health and Neurosciences (NIMHANS). Numerous state-owned aerospace and defense organizations, such as Bharat Electronics, Hindustan Aeronautics, and National Aerospace Laboratories are located in the city. [1]

Given the huge employment and educational opportunities, this demographically diverse city has a population of about 10 million and a metropolitan population of about 8.52 million with a population density of 12,000 people per square kilometer, making it the third most populous city and fifth most populous urban agglomeration in India. [1]

1.2 Problem

Given the huge population that resides in Bangalore, there is a huge requirement and a burgeoning housing market in the city. The home buyers are looking at investing in the area that can help maximize return on investment (where the current cost of ownership is low, but the potential for growth is high), at the same time they will also like to choose an area where their social needs are taken care as well. The shop owners (e.g. Restaurant, Grocery, Entertainment, etc.) are also looking at investing in areas of potential high population density. However, it is difficult to obtain this information that can guide the investors.

1.3 Interest

Potential investors in real estate, be it the domestic investors looking at buying a home or a retail investor looking at setting up a business, is the key audience that will be interested in this information set. This can also be helpful for government agencies to plan and ensure adequate infrastructure and services are made available.

2. Data Acquisition and Cleaning

2.1 Data Acquisition (Data Source)

To address the problem on hand, the data requirements are stated as below:

a. Bangalore Apartment Location Data:

This data is obtained from Kaggle which can be accessed here. This data set provides us with a list of all the apartments listed on Commonfloor website along with the geo-coordinates.

	names	lat	lon	geometry
0	Purva Skydale bangalore	12.894033	77.662362	POINT (77.66236169999999 12.894033)
1	Salarpuria Sattva Cadenza bangalore	12.889441	77.640221	POINT (77.64022109999999 12.889441)
2	Shriram Summitt bangalore	12.836068	77.667242	POINT (77.6672418 12.8360678)
3	Shriram Luxor bangalore	13.085249	77.654915	POINT (77.65491539999999 13.085249)
4	Ecolife Elements Of Nature bangalore	12.938728	77.731126	POINT (77.73112619999999 12.9387277)

b. Bangalore Apartment Data:

This data is obtained from Kaggle which can be accessed here. This data set provides us with a list of the apartments listed on Commonfloor website along with the price, area of the apartment, and unit type (1BHK, 2 BHK, etc). BHK – stands for Bedroom Hall and Kitchen. 1 BHK – is a unit with one bedroom, 2BHK is a unit with two bedrooms, and so on.

	names	Price	Area	Unit Type
0	Salarpuria Sattva Cadenza	39 L -41.65 L	755 sq.ft	1 BHK Apartment
1	Salarpuria Sattva Cadenza	55 L -75 L	1175-1275 sq.ft	2 BHK Apartment
2	Salarpuria Sattva Cadenza	70.04 L -73.30 L	1335-1340 sq.ft	2.5 BHK Apartment
3	Salarpuria Sattva Cadenza	65 L -95 L	1365-1595 sq.ft	3 BHK Apartment
4	Purva Skydale	76.25 L -1.75 Cr	1273-1371 sq.ft	2 BHK Apartment

c. Bangalore Neighborhood Data:

This data is obtained from Kaggle and can be accessed here. This data set provides us details of Bangalore Neighborhood (Area), along with the geo-coordinates of the area.

	Neighborhood	Latitude	Longitude
0	Agram	12.958000	77.630800
1	Amruthahalli	13.066513	77.596624
2	Attur	13.107000	77.566300
3	Banaswadi	13.014162	77.651854
4	Bellandur	12.930400	77.678400

d. Venues nearby the area:

This data is obtained by using Foursquare API (Foursquare), this API helps us get the details of the venues in the defined vicinity(radius) of the area of our interest.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bommanahalli	12.903	77.6242	Pizza Hut	12.899474	77.631437	Pizza Place
1	Bommanahalli	12.903	77.6242	Faaso's	12.899975	77.622621	Indian Restaurant
2	Bommanahalli	12.903	77.6242	Fooday kabab center	12.905933	77.629228	Indian Restaurant
3	Bommanahalli	12.903	77.6242	Hotel Ibis	12.901015	77.632125	Hotel Bar
4	Bommanahalli	12.903	77.6242	Ananda Honda	12.909018	77.627175	Auto Garage

2.2 Data Cleaning

The data that was required for this analysis was not available from a single location, and I had to work with data from different sources. The data that was available from these sources needed to be analyzed and corrections were required for the cases where the data was missing. To address this issue I decided to drop the data points with missing value as it was not feasible for us to get the missing data.

For the Bangalore apartment location data, the details of the area/address of the apartment were missing. To address this missing data I wrote a piece of code to get the address of the apartment based on the geo-coordinated via a reverse geo-coordinate approach.

For the Bangalore Neighborhood (Area) data, the Neighborhood (Area) column had few additional characters over and above the area name (e.g. S.O. – Sub Office and others). I wrote a code to remove the unwanted/additional characters. In this data set, I also observed that the location data for many of the areas were incorrect this had to be manually corrected.

For the Bangalore Apartment Data, this data set had multiple line items for the same apartment based on the price, apartment size, and unit type (1BHK, 2BHK, etc.). However, to ensure one single point on the map per apartment and not to miss out on the details of the apartment, I had to write a query to group data by the apartment name while ensuring that the details of the apartment are not lost.

For venues nearby that area, I had to write a query to extract useful venue information from the json response that was received by calling the Foursquare API. Also, keeping in mind the performance aspect of the response from Foursquare API I had to break down the data for query into smaller chunks to ensure that the API call does not time out or fail.

2.3 Data Integration

As mentioned earlier, the data that was required for our analysis was not always directly available for the data acquired. Even on the data that was available from the data sets, various integration approaches were required to be put in place for data to be used for final analysis and not all the integration was straightforward.

The following are some of the data integration approaches followed as part of this project to make data ready for use.

1. Address of the apartment from the geo-code: The data available from Kaggle for Bangalore apartment did not have the details of the Neighborhood (Area) to which the apartment belongs. This was an important feature for the analysis as venue details based on the area would provide a better basis for grouping based on venues rather than identifying venues near each apartment and then trying to group them. To overcome this challenge, I wrote a code for reverse geo-coding (identifying address based on the geo-coordinates)

2. Mapping the area and area geo-coordinates to the apartments: The Bangalore Neighborhood (Area) data did provide a neighborhood and geo-coordinates details, but the challenge was to map the area to the apartment. The first step above gives details on how a part of the challenge was addressed. The other part of the challenge was to map the Area with the address, and the address of each apartment was not in the same format (e.g. Name of the apartment, area, zip code). Different apartments followed a different format of the address that made it challenging to map the area with the apartment. To overcome this challenge, I wrote a piece of code to map two tables based on the address field in one table and area field in another table.

3. Integrating venue details with apartment data: Here I had to write a code to call Foursquare API to get venue details around the area, the response from API is a json file which form which relevant data needed to be picked up and corresponding values updated in the table

4. Integrating apartment details with the clustered apartment data: Once the clustering of the apartment is completed, after the K-means is applied, the apartment details needed to be populated to ensure that when the apartment clusters are mapped on to the map when we click on location apartment all the details of the apartments are visible like the cluster, apartment name, price, area of the house, and unit type. This required me to write a code to ensure that the data is integrated with the clustered data correctly.

3. Methodology

I used Jupiter Notebook to accomplish this project, for writing code as well as for the storage of my data. The code was then uploaded to the GitHub library. The main components of my master data consist of apartment name, area, and area's geo-coordinates. These main components are achieved by the following mechanism.

1. Apartment Name – obtained from "Bangalore Apartment Location" data set

	names	lat	lon	geometry
0	Purva Skydale bangalore	12.894033	77.662362	POINT (77.66236169999999 12.894033)
1	Salarpuria Sattva Cadenza bangalore	12.889441	77.640221	POINT (77.64022109999999 12.889441)
2	Shriram Summitt bangalore	12.836068	77.667242	POINT (77.6672418 12.8360678)
3	Shriram Luxor bangalore	13.085249	77.654915	POINT (77.65491539999999 13.085249)
4	Ecolife Elements Of Nature bangalore	12.938728	77.731126	POINT (77.73112619999999 12.9387277)

2. Area and its co-ordinates – these details are obtained by the following steps:

   a. Obtaining address for the apartment in the above data set, by using reverse geo technique – Here we passed the latitude and longitude details of the apartment to obtain the address for that apartment (Data formatting operations are also done)

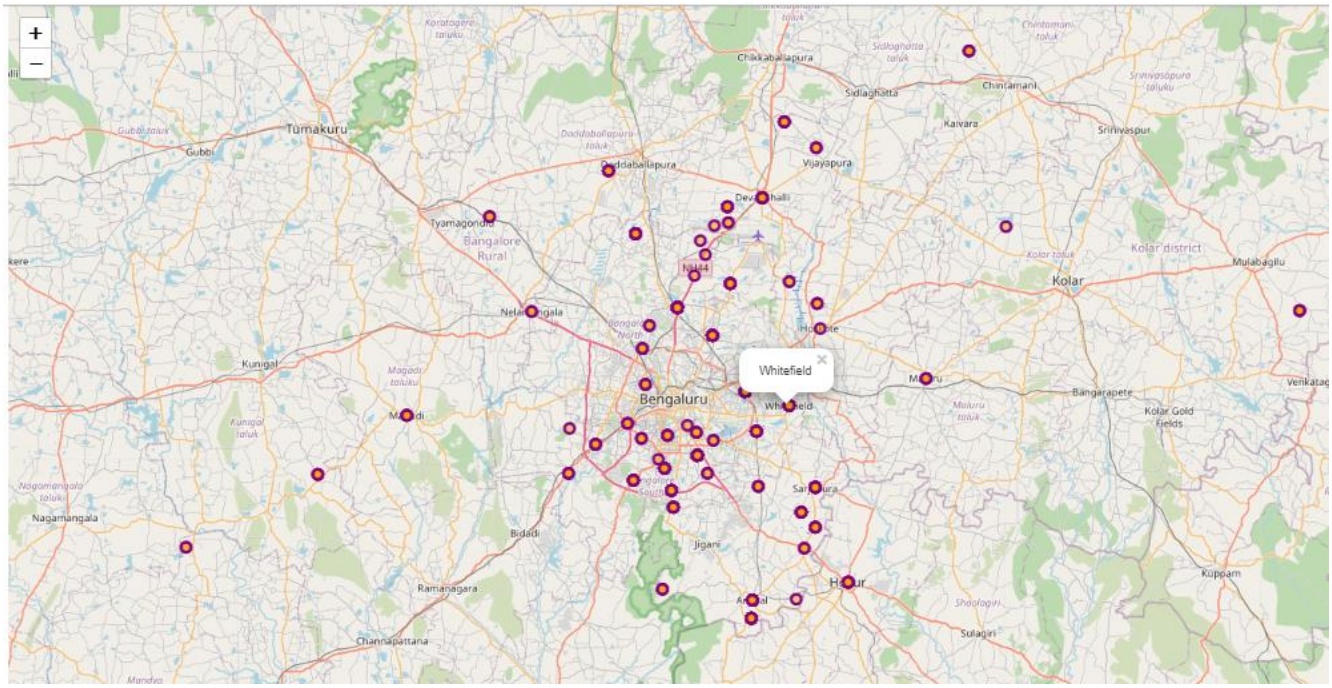
	names	lat	lon	geometry	Cordinated	Address
0	Purva Skydale bangalore	12.894033	77.662362	POINT (77.66236169999999 12.894033)	12.894033,77.66236169999998	Purva Skydale Apartments, Singasandra, Bommana...
1	Salarpuria Sattva Cadenza bangalore	12.889441	77.640221	POINT (77.64022109999999 12.889441)	12.889441,77.64022109999998	Kudlu Road, Mangammanapalya, Bommanahalli Zone...
2	Shriram Summitt bangalore	12.836068	77.667242	POINT (77.6672418 12.8360678)	12.836067800000002,77.6672418	Hebbagodi, Electronics City Phase 2 (West), Hu...
3	Shriram Luxor bangalore	13.085249	77.654915	POINT (77.65491539999999 13.085249)	13.085249000000001,77.65491540000001	Kannuru, Bangalore East, Bangalore Urban, Karn...
4	Ecolife Elements Of Nature bangalore	12.938728	77.731126	POINT (77.73112619999999 12.9387277)	12.9387277,77.73112619999998	Varthuru, Mahadevapura Zone, Bengaluru, Bangal...

   b. Next, we use "Banglore Neighborhood" data as shown below, and map the area latitude and longitude to the above data set – Here Address field in the table above and Neighborhood field in the table below are used as the joining keys – The result can be seen in the next screenshot

	Neighborhood	Latitude	Longitude
0	Agram	12.958000	77.630800
1	Amruthahalli	13.066513	77.596624
2	Attur	13.107000	77.566300
3	Banaswadi	13.014162	77.651854
4	Bellandur	12.930400	77.678400

	names	lat	lon	geometry	Cordinated	Address	Neighborhood	Nbh_lat	Nbh_lon
0	Purva Skydale bangalore	12.894033	77.662362	POINT (77.66236169999999 12.894033)	12.894033,77.66236169999998	Purva Skydale Apartments, Singasandra, Bommana...	Singasandra	12.685041	77.697563
1	Salarpuria Sattva Cadenza bangalore	12.889441	77.640221	POINT (77.64022109999999 12.889441)	12.889441,77.64022109999998	Kudlu Road, Mangammanapalya, Bommanahalli Zone...	Bommanahalli	12.903000	77.624200
2	Shriram Summitt bangalore	12.836068	77.667242	POINT (77.6672418 12.8360678)	12.8360678000000002,77.6672418	Hebbagodi, Electronics City Phase 2 (West), Hu...	Anekal	12.708637	77.699397
3	Shriram Luxor bangalore	13.085249	77.654915	POINT (77.65491539999999 13.085249)	13.085249000000001,77.65491540000001	Kannuru, Bangalore East, Bangalore Urban, Karn...	Kannur	11.946689	75.353877
4	Ecolife Elements Of Nature bangalore	12.938728	77.731126	POINT (77.73112619999999 12.9387277)	12.9387277,77.73112619999998	Varthuru, Mahadevapura Zone, Bengaluru, Bangal...	Mahadevapura	12.988000	77.689500

The above neighborhood (area) co-ordinate data is then used to plot the areas of Bangalore on a map. I have used the Folium library to visualize the areas of Bangalore, by superimposing the neighborhood (areas) on to the map as below.



Next, I utilized Foursquare API to explore the venues nearby the areas of my data set. I designed a query to limit the number to **30 venues** within a radius of **1000 meters** of the area (neighborhood). Below is an excerpt from the data set of venues (**Note:** to ensure optimal performance, we divided the data set into set of 100 line items for API calls)

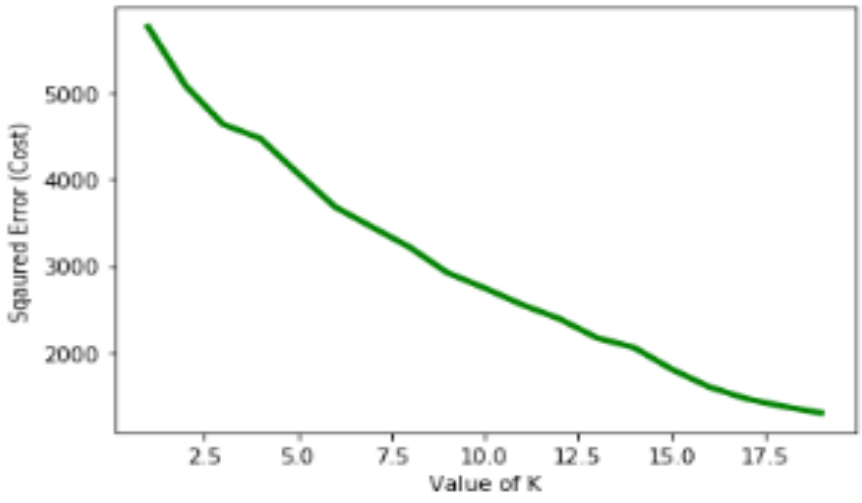
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bommanahalli	12.903	77.6242	Pizza Hut	12.899474	77.631437	Pizza Place
1	Bommanahalli	12.903	77.6242	Faaso's	12.899975	77.622621	Indian Restaurant
2	Bommanahalli	12.903	77.6242	Fooday kabab center	12.905933	77.629228	Indian Restaurant
3	Bommanahalli	12.903	77.6242	Hotel Ibis	12.901015	77.632125	Hotel Bar
4	Bommanahalli	12.903	77.6242	Ananda Honda	12.909018	77.627175	Auto Garage

From the list of all the venues that were received for each area from Foursquare, I then identified the top 10 most frequent venues in a given area (neighborhood).

	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	Adugodi	Indian Restaurant	Dessert Shop	Lounge	Multiplex	Coffee Shop	Café	Brewery	Donut Shop	Juice Bar	Clothing Store
1	Agara	Indian Restaurant	Pizza Place	Ice Cream Shop	Italian Restaurant	Chinese Restaurant	Café	Bike Shop	Indie Movie Theater	Food Truck	Japanese Restaurant
2	Anekal	ATM	Indian Restaurant	Camera Store	Movie Theater	Business Service	Farm	Cosmetics Shop	Creperie	Department Store	Dessert Shop
3	Attibele	ATM	Bakery	Indian Restaurant	South Indian Restaurant	Castle	Antique Shop	Art Gallery	Cosmetics Shop	Creperie	Department Store
4	Bagalur	Memorial Site	Food Truck	Farm	Coffee Shop	Convenience Store	Cosmetics Shop	Creperie	Department Store	Dessert Shop	Dhaba

As we want to cluster the area (neighborhood) of Bangalore, based on the venues in the given radius of 1 Km, I have used an approach of unsupervised clustering. Kmeans being one of the most popular and widely used clustering algorithm, I have used Kmeans as a clustering algorithm form clustering my data set.

To identify an optimal value of clusters, I ran the algorithm for different values of K to find the error associated with each value of K. below is the plot for different values of K and corresponding error.



For the graph above we can see that the optimal has not yet converged.

However, As as a data science product owner, we have a job of playing a balancing act between finding optimal clusters and restricting the clusters to make the end user's (consumer of our service/application) life easy and to peak their interest.

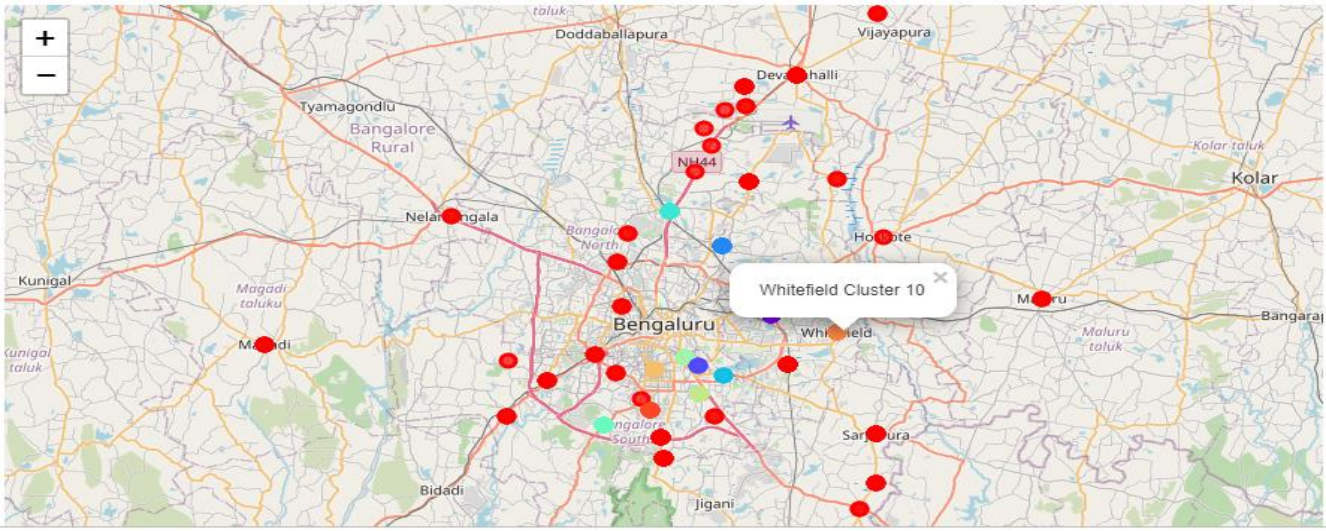
Playing the balancing act between usability and accuracy, we will use the value of 12 clusters based on the above graph - Here, the number of clusters is moderate (not too few or not too many) and that has an acceptable error.

Below is the merged table after clustering using Kmeans and number of clusters equal to 12

	Neighborhood	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	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adugodi	7	Indian Restaurant	Dessert Shop	Lounge	Multiplex	Coffee Shop	Café	Brewery	Donut Shop	Juice Bar	Clothing Store
1	Agara	4	Indian Restaurant	Pizza Place	Ice Cream Shop	Italian Restaurant	Chinese Restaurant	Café	Bike Shop	Indie Movie Theater	Food Truck	Japanese Restaurant
2	Anekal	0	ATM	Indian Restaurant	Camera Store	Movie Theater	Business Service	Farm	Cosmetics Shop	Creperie	Department Store	Dessert Shop
3	Attibele	0	ATM	Bakery	Indian Restaurant	South Indian Restaurant	Castle	Antique Shop	Art Gallery	Cosmetics Shop	Creperie	Department Store
4	Bagalur	0	Memorial Site	Food Truck	Farm	Coffee Shop	Convenience Store	Cosmetics Shop	Creperie	Department Store	Dessert Shop	Dhaba

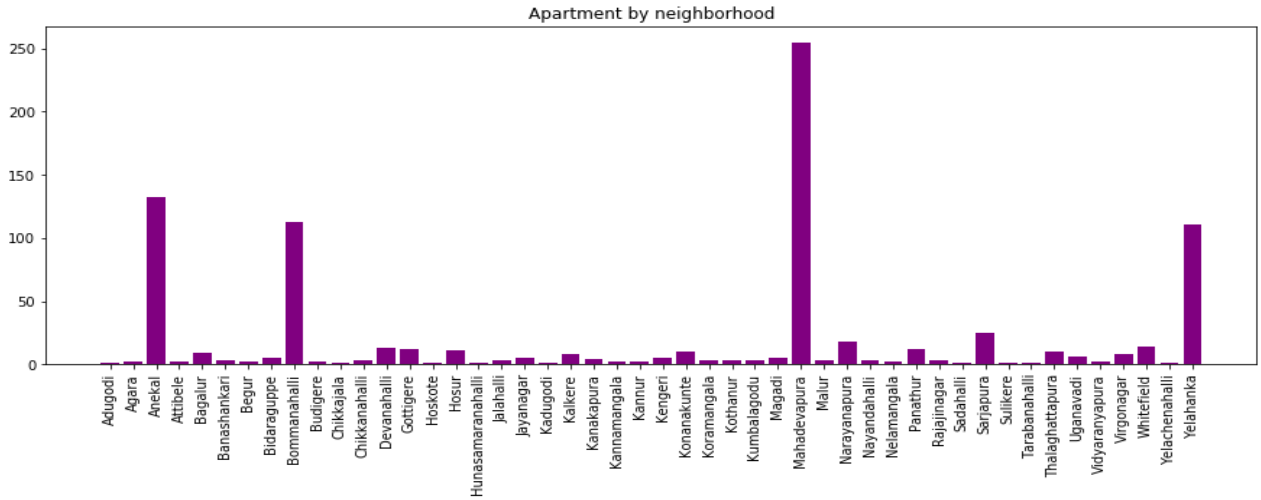
4. Results

Based on the data set generated from the clustering analysis, we can see that the areas (neighborhood) of Bangalore are grouped into 12 clusters as below. The image below is a color-coded superimposition of the clustered area on the map, along with cluster label.



The above area clustering information alone is not very useful to the inventory, although it may give them a high-level overview of the similar areas and provide a comparative indication of the area that may have growth potential based on the grouping with a known area that is currently considered favorable.

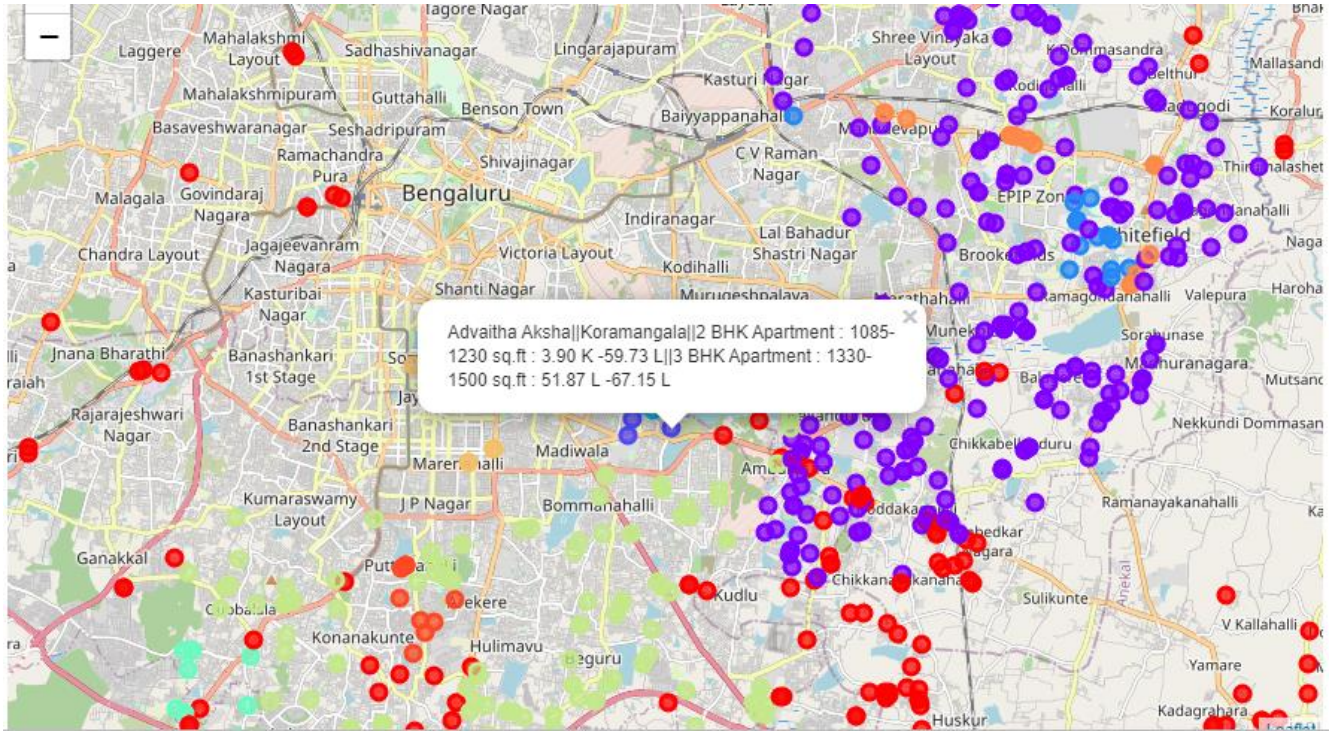
To provide more insights, we have grouped the data based on the area (neighborhood) to identify the number of apartments in the given area (neighborhood). For the graph below we can see that the **Mahadevpura** area of Bangalore has the highest number of apartments with **250+** apartments. This information is a useful density concentration indication.



To provide further insights for decision making, we are color-coded the apartment in the area based on the area clusters, while also providing apartment details like.

- Apartment Name
- Area (Neighborhood)
- Apartment type (1BHK, 2BHK, etc.)
- Area of apartment
- Price

These details are superimposed on the interactive map, that allows the user to check for the apartment details in the given neighborhood while doing comparative analysis in a similar neighborhood based on their requirements.



5. Discussion

As mentioned in the introduction section, Bangalore with its huge employment and educational opportunity is a melting pot of demographically diverse population. With this influx of migrant population, Bangalore also has a widespread and burgeoning real estate market. Given this fact, investors will appreciate any enables that can help them make better investment decisions.

The approach that I took for this project is generalized, no special preferences or weightage is given a particular feature (venues in our case) for clustering. Also, given the clustering problem on hand, various clustering techniques can be utilized to arrive at a solution and not all techniques will yield high-quality results.

I have used the Kmeans algorithm as a clustering approach for the project. I tried clustering based on a couple of different approaches.

1. Clustering base on venues near the apartment
2. Clustering based on venues near the area (Neighborhood) of the apartment

Finally, I decided to proceed with clustering base on venues near the area (neighborhood) of the apartment even though the clustering based on venues near the apartment provided better convergence and lower error. The reason for choosing this approach was to provide a generalized overview of the area (neighborhood) rather than a specific apartment.

There was also a challenge to play a balancing act between the number of clusters to be selected and the accuracy of the clusters. The elbow diagram of error vs cluster size of Kmeans that I plotted did not converge to optimal even at a larger clustering size, due to this I had to play a balancing of a trade-off between the number of clusters to be selected and the accuracy. However, keeping in mind the usability aspect for end-user I decided to cluster Bangalore areas (neighborhood) into 12 clusters.

I used the folium library to visualize the clustered areas on the map using color coding based on the cluster group they belong to. To provide greater details, I superimposed the apartment details like apartment name, area, price, build area, and unit type on the map and color-coding the apartment on the map in line with the area (neighborhood) cluster.

As a future improvement, we can provide an interactive interface for the end-users – This can further improve their experience and help better with a better investment decision. Another key future improvement can be to have clustering based on user preference (e.g. ability to select venues (feature) based on which clustering can be done or providing the capability to mark

weightage of a feature for clustering – e.g. for some people distance from the school or a hospital or a metro station has higher weightage as compared to other features)

6. Conclusion

As more people migrate to Bangalore, in search of employment and educational opportunity, an application or platform like this can prove useful for them by providing consolidated information on a click, that can aid them in informed decision making.

Such a platform can also be used by retail investors looking at business opportunities like opening a restaurant, gymnasium, and others, as this platform can provide information on high population density areas and the type of services that are in shorter supply in those areas

This platform can also be used by the government agencies and civic bodies to better manage the development, infrastructure, and service needs of an area.

7. Reference

[1] **Bangalore wiki**

2 **Foursquare API**