Jitendra K Singh
Chennai
India

About 1000 words

**CAPSTONE PROJECT**

# HOUSING SALES RATE & NEIGHBORHOODS ANALYSIS OF CHENNAI, INDIA

By Jitendra K Singh

## A. Introduction

Chennai is one of the four major metro-cities of India. Area of the city is **426 square km**. Population of the city is close to **8.6 Million**. The city is divided in mainly Four Zones i.e.

a. Chennai North
b. Chennai South
c. Chennai West
d. Chennai East

These zones are further divided in several neighborhoods. All these neighborhoods have different housing rates and different amenities for public utilization.
For anyone new to the city it gets challenging to find a neighborhood of choice given the variation in housing rates and the available amenities.

## B. Problem Description

Chennai being a major city with huge area and several differences in neighborhoods, **buyers or investors** find it difficult to zero-in on the housing they must buy. As a solution we will create a map that would cluster the similar neighbourhoods based on the housing rate, available amenities and number of housing societies in each neighborhood.

## C. Data Description

To create a map with clustered neighbourhoods I will utilise the available data in the below websites.

a. The link here will be used to get the information of housing rates of various housing societies rate. Data for Housing Societies Rate. This will contain data of several housing societies and the range of prices for same.
b. The link here will be used to get the geo coordinates for various neighborhoods. Data for Neighborhood coordinates. This data will have the geo coordinates of neighborhoods.
c. I will use the **Foursquare API** to get the neighborhood venues for these neighborhoods. There are going to be some challenges about the how the extensive data is available.
   For example, the link here does not have geo coordinates for all the neighborhoods and **Foursquare.com** does not have very extensive list of venues of these neighborhoods.

For our analysis we are going to use the several variables for clustering neighborhoods.

## D. Approach

The below mentioned is the approach that I will utilise to get the desired result.

a. I will utilize **BeautifulSoup** library to extract relevant information for housing societies rate from here.
b. Then I will use **Pandas** library to process the data to create a data frame that has neighborhood name with average housing rates.
c. The Wikipedia Page will be utilized to get the geo coordinates for the neighborhoods. The neighborhoods for which geo coordinates will not be available, might have to be removed from data frame.
d. The **Foursquare API** will be utilized to get the top venues for these neighborhoods.
e. **K-Means** library and algorithm will be utilized to cluster these neighborhoods based on top venues and average rate of housing.
f. **Folium** library will be utilized to create a map of Chennai with marker of clustered neighborhoods.

This is the initial approach defined on how the data will be processed to arrive at conclusion. Depending on the data quality there can be minor changes on the methodology deployed.

# E. Methodology

In this section I will proceed step-by-step from web scraping to result.

## E.1 Import Required Libraries

In this step all the required libraries were imported.

## E.2 Web Scraping

In this step BeautifulSoup library was utilised to extract the data from the website and create a dataframe containing housing societies name and the price range as below.

| | Society name | Range |
|---|---|---|
| 0 | None | None |
| 1 | Akshaya Republic, Kovur | Rs. 4,972 - 5,398/sq. ft. |
| 2 | Altis Ashraya, Mangadu | Rs. 4,378 - 4,462/sq. ft. |
| 3 | Arihant Tiara, Nandambakkam | Rs. 6,375 - 6,375/sq. ft. |
| 4 | Bharathi Brikhouse, Vanagaram | Rs. 5,185 - 5,610/sq. ft. |

Fig.1 – Dataframe of housing societies name and price range

## E.3 Data Cleaning

The data frame had only two columns and first column was having Society name and neighborhood name combined in same column and price range column had price in string format. This required cleaning so that further calculation could be done on the input data. As I was interested to perform an analysis for each neighborhood, I needed a data frame that would contain average rates, number of housing societies, minimum and maximum rates for each neighborhood. After several steps of codes and I was able to get he data in cleaned format as below.

| | Neighbourhood | Number_of_societies | Average_min | Average_max | Mean_rate |
|---|---|---|---|---|---|
| 0 | Adambakkam | 1 | 7820.00 | 9392.00 | 8606.0 |
| 1 | Adyar | 1 | 13515.00 | 13898.00 | 13706.5 |
| 2 | Agraharam | 1 | 4420.00 | 5015.00 | 4717.5 |
| 3 | Alandur | 1 | 6672.00 | 6672.00 | 6672.0 |
| 4 | Ambattur | 4 | 4441.25 | 4908.75 | 4675.0 |

Fig.2 – Dataframe of neighborhoods data

## E.4 Geo coordinates for neighborhoods

The next step was to get the geo-coordinates for each of these neighborhoods. I utilized Nominatim library to get the geocoordinates for these neighborhoods.

| | Neighbourhood | Number_of_societies | Average_min | Average_max | Mean_rate | geo_loc | Latitude | Longitude |
|---|---|---|---|---|---|---|---|---|
| 0 | Adambakkam | 1 | 7820.00 | 9392.00 | 8606.0 | (Adambakkam, Ward 177, Zone 13 Adyar, Chennai,... | 12.982221 | 80.209121 |
| 1 | Adyar | 1 | 13515.00 | 13898.00 | 13706.5 | (Adyar, Ward 176, Zone 13 Adyar, Chennai, Chen... | 13.006450 | 80.257779 |
| 2 | Agraharam | 1 | 4420.00 | 5015.00 | 4717.5 | (Agraharam, Yelamanchili, Visakhapatnam, Andhr... | 17.568600 | 82.850391 |
| 3 | Alandur | 1 | 6672.00 | 6672.00 | 6672.0 | (Alandur, Tamil Nadu, India, (13.00282155, 80.... | 13.002822 | 80.171919 |
| 4 | Ambattur | 4 | 4441.25 | 4908.75 | 4675.0 | (Ambattur, Thiruvallur District, Tamil Nadu, I... | 13.112886 | 80.159862 |

Fig.3 – Dataframe of neighborhoods data with geo-coordinates

## E.5 Data selection

For few of these neighborhoods the geo-coordinates were not available from the Nominatim library. So those few neighborhoods had to be dropped form the data frame. The other neighborhood data that we would be needing was number of housing societies and average rates and geo-coordinates. So, only those columns were selected for the next steps.

| | Neighbourhood | Number_of_societies | Mean_rate | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | Adambakkam | 1 | 8606.0 | 12.982221 | 80.209121 |
| 1 | Adyar | 1 | 13706.5 | 13.006450 | 80.257779 |
| 2 | Agraharam | 1 | 4717.5 | 17.568600 | 82.850391 |
| 3 | Alandur | 1 | 6672.0 | 13.002822 | 80.171919 |
| 4 | Ambattur | 4 | 4675.0 | 13.112886 | 80.159862 |

Fig.4 – Selected data for further analysis

## E.6 Plotting on map

These neighborhoods were then plotted on map.

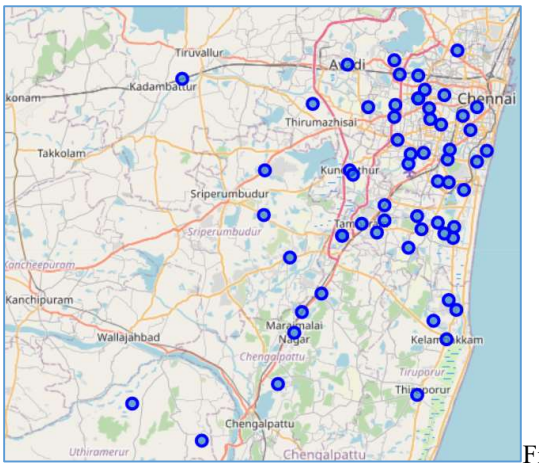


Fig.5 – Neighbourhoods plotted on map

## E.7 Neighborhood venues

Utilizing the Foursquare API, for each of these neighborhoods the nearby venues data was extracted. The relevant data for these venues were venue name, venue category and venue geo-coordinates. By using json normalize and few codes I was able to get the required venue data for each neighborhood.

Number of venues in Chennai is 362

| | Neighborhood | Neighborhood latitude | Neighborhood longitude | Venue name | Venue latitude | Venue longitude | Venue category |
|---|---|---|---|---|---|---|---|
| 0 | Adambakkam | 12.982221 | 80.209121 | arun icecream | 12.983447 | 80.207847 | Dessert Shop |
| 1 | Adambakkam | 12.982221 | 80.209121 | MedPlus | 12.980940 | 80.207356 | Pharmacy |
| 2 | Adambakkam | 12.982221 | 80.209121 | Sutherland | 12.981002 | 80.205200 | IT Services |
| 3 | Adambakkam | 12.982221 | 80.209121 | Bistro | 12.983193 | 80.205020 | Indian Restaurant |
| 4 | Adyar | 13.006450 | 80.257779 | Bombay Brassiere | 13.006961 | 80.256419 | North Indian Restaurant |

Fig.6 – Neighborhood venue data

## E.8 Top ten venues

For these neighborhoods I needed top ten venues for each of these neighborhoods to categorize and cluster the neighborhoods. So, after few steps of codes for aggregation of neighborhoods and one-hot coding, I was able to get the data for each neighborhood with top ten venues.

| | Neighbourhood | 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 | Adambakkam | IT Services | Pharmacy | Indian Restaurant | Dessert Shop | Food Service | Convenience Store | Cupcake Shop | Department Store | Electronics Store | Farmers Market |
| 1 | Adyar | Indian Restaurant | Electronics Store | North Indian Restaurant | Ice Cream Shop | Rock Club | Grocery Store | Vegetarian / Vegan Restaurant | Italian Restaurant | Juice Bar | Lounge |
| 2 | Alandur | Airport Service | Whisky Bar | Gastropub | Convenience Store | Cupcake Shop | Department Store | Dessert Shop | Electronics Store | Farmers Market | Fast Food Restaurant |
| 3 | Ambattur | Multiplex | Movie Theater | River | Whisky Bar | Coffee Shop | Convenience Store | Cupcake Shop | Department Store | Dessert Shop | Electronics Store |
| 4 | Anna Nagar | Indian Restaurant | Fast Food Restaurant | Electronics Store | Chinese Restaurant | Café | Park | Bistro | Vegetarian / Vegan Restaurant | Middle Eastern Restaurant | Department Store |

Fig.7 – Neighborhood venue data

## E.9 Clustering

i) Data

For the purpose of clustering I used the data of each neighborhood with frequency of each type of venue category and merged it with the data of number of housing societies in each neighborhood and the average rate.

| | Neighbourhood | Number_of_societies | Mean_rate | ATM | Afghan Restaurant | Airport Service | American Restaurant | Antique Shop | Art Gallery | Asian Restaurant | ... | Spa | Sports Club | Tattoo Parlor | Tea Room | Thai Restaurant |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Adambakkam | 1 | 8606.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 | Adyar | 1 | 13706.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.034483 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | Alandur | 1 | 6672.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.000000 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | Ambattur | 4 | 4675.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | Anna Nagar | 1 | 9010.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.043478 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

5 rows × 110 columns

Fig.7 – Neighborhood data for clustering

ii) Elbow method to find K for KMeans

I use the elbow method to find the optimum K that can be used for clustering in K-Means. I chose the k=4 for clustering as outcome.

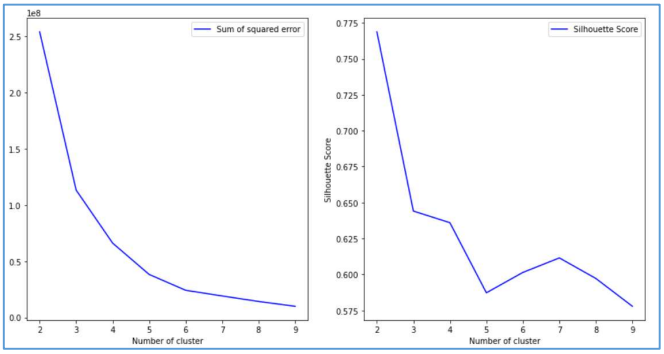


Fig.8 – Outcome of elbow method on plot

iii) Cluster labels

Using the K-means object, the labels for each of the neighborhoods was obtained. These labels were then taken in the data frame against each of the neighborhoods.

| | Neighbourhood | Number_of_societies | Mean_rate | Latitude | Longitude | Cluster label | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Adambakkam | 1 | 8606.0 | 12.982221 | 80.209121 | 2.0 | IT Services | Pharmacy | Indian Restaurant | Dessert Shop | Food Service | Convenience Store | Cupcake Shop |
| 1 | Adyar | 1 | 13706.5 | 13.006450 | 80.257779 | 1.0 | Indian Restaurant | Electronics Store | North Indian Restaurant | Ice Cream Shop | Rock Club | Grocery Store | Vegetarian / Vegan Restaurant |
| 2 | Alandur | 1 | 6672.0 | 13.002822 | 80.171919 | 0.0 | Airport Service | Whisky Bar | Gastropub | Convenience Store | Cupcake Shop | Department Store | Dessert Shop |
| 3 | Ambattur | 4 | 4675.0 | 13.112886 | 80.159862 | 0.0 | Multiplex | Movie Theater | River | Whisky Bar | Coffee Shop | Convenience Store | Cupcake Shop |
| 4 | Anna Nagar | 1 | 9010.0 | 13.087200 | 80.216442 | 2.0 | Indian Restaurant | Fast Food Restaurant | Electronics Store | Chinese Restaurant | Café | Park | Bistro |

Fig.9 – Neighborhoods labelled as per cluster.

iv) Visualize on map

The neighborhoods were again plotted on map and markers were colored as per the respective cluster.

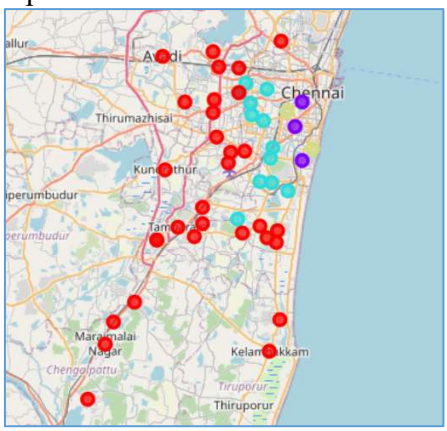


Fig.10 – Neighborhoods plotted and marked as per cluster.

### E.10 Average rates vs Distance from Chennai center

I further wanted to see how the average rate of housing in neighborhood varies w.r.t distance from city center. So, I used the geodesic library to get the distance of each of these neighborhoods from the center.

| | Neighbourhood | Number_of_societies | Mean_rate | Latitude | Longitude | Cluster label | Distance_from_center |
|---|---|---|---|---|---|---|---|
| 0 | Adambakkam | 1 | 8606.0 | 12.982221 | 80.209121 | 2.0 | 13.034431 |
| 1 | Adyar | 1 | 13706.5 | 13.006450 | 80.257779 | 1.0 | 8.650870 |
| 2 | Alandur | 1 | 6672.0 | 13.002822 | 80.171919 | 0.0 | 13.916017 |
| 3 | Ambattur | 4 | 4675.0 | 13.112886 | 80.159862 | 0.0 | 12.391949 |
| 4 | Anna Nagar | 1 | 9010.0 | 13.087200 | 80.216442 | 2.0 | 5.841288 |

Fig.11 – Neighborhoods plotted distance from center.

## F. Results

Now all the cluster were analysed and those were listed. Box plot was created to see the average rates of all the clusters. A scatter plot was also created to see the variation in average rates of each neighbourhoods w.r.t the distance from city center.

The results are as follows:

a. The number of neighborhoods in each cluster were listed and enumerated and we get the numbers as
   1. Cluster 1 – 33 neighborhoods
   2. Cluster 2 – 4 neighborhoods
   3. Cluster 3 – 12 neighborhoods
   4. Cluster 4 – 3 neighborhoods

b. When we look at the box plot of average rates for each cluster

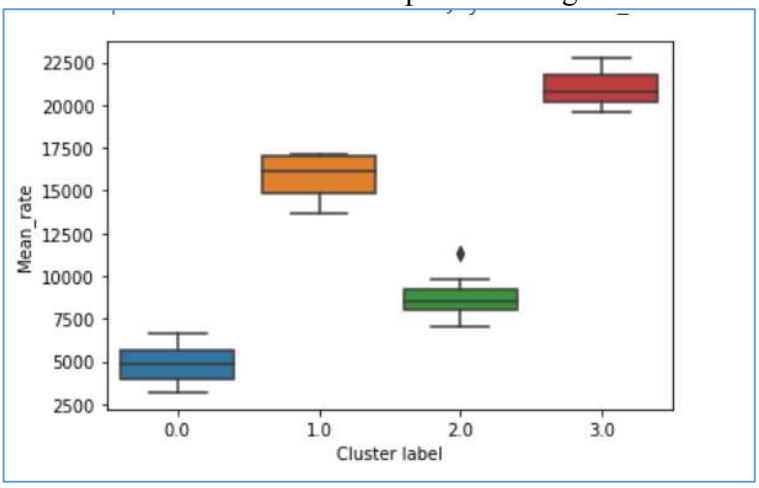


Fig.12 – Average rate of clusters

Hence, we can define.

1. Cluster 1 – Lowest average rates
2. Cluster 2 – second highest average rate
3. Cluster 3 – Second lowest average rate
4. Cluster 4 – Highest average rate

c. When we look at the scatter plot of average rate of neighborhoods vs distance from center

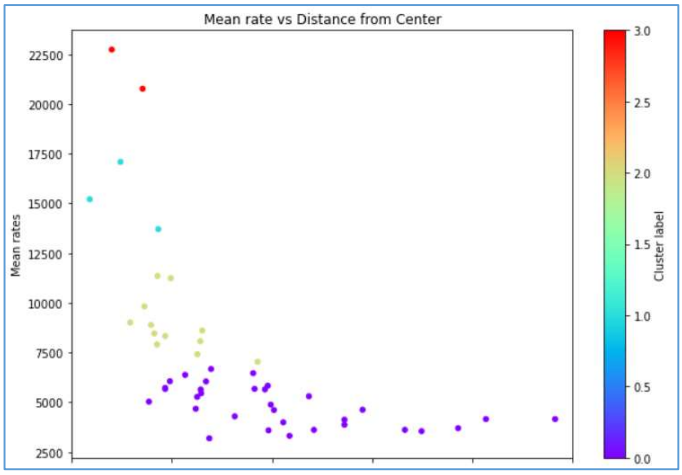


Fig.13 – Scatter plot

Hence, we can define.

1. Cluster 1 – Farthest from center
2. Cluster 2 – second closest to center
3. Cluster 3 – second far from center
4. Cluster 4 – nearest to center

Here we also observe that as the distance from center increases the average rates in the neighborhood is going down.

d. So, I looked at correlation between average rates and the distance from center and it was very negligible.

| | Distance_from_center | Mean_rate |
|---|---|---|
| **Distance_from_center** | 1.000000 | 0.357453 |
| **Mean_rate** | 0.357453 | 1.000000 |

Fig.14 – Average cost and distance correlation

Looking at this value we can say that there is very negligible correlation between the average rates and the distance from center. There can be other variables affecting the average price of housing.

## G. Discussion

When we look at the data and final results side by side, we have insights to define each cluster.

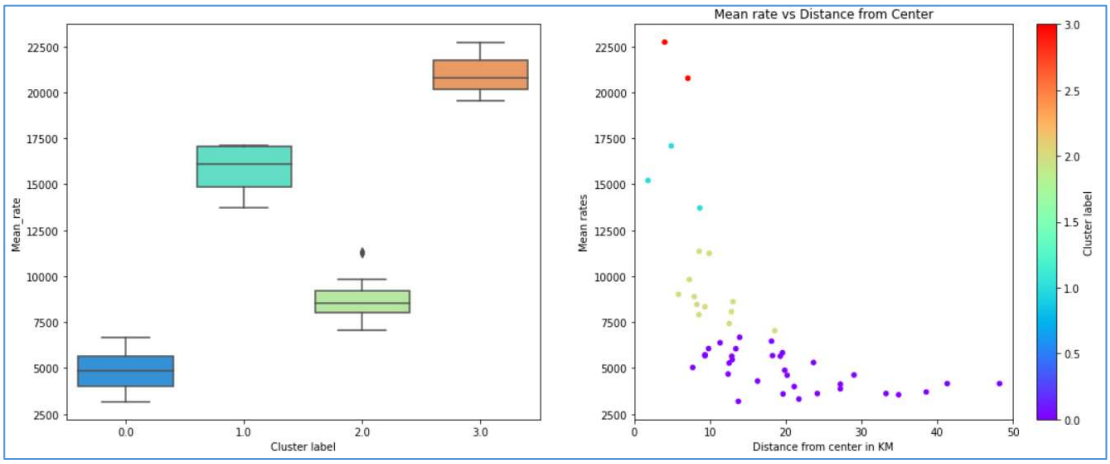


Fig.15 – Box plot and Scatter plot

We saw that the cluster 4 which is closest to the center is also having highest average rates and have very few neighborhoods.

We also saw that cluster 1 which is farthest from the center is also having the lowest average rates and has highest number of neighborhoods.

Based on this observation we can rank the clusters as below.

| Cluster Label | Mean_Rate Rank | Distance_from_Center rank | No. of Neighbourhoods |
|---|---|---|---|
| Cluster 4 | 1 | 4 | 3 |
| Cluster 2 | 2 | 3 | 4 |
| Cluster 3 | 3 | 2 | 12 |
| Cluster 1 | 4 | 1 | 33 |

Fig.16 – Clusters ranked, Rank 1 is highest and Rank 4 means lowest.

As we move away from center the average rates seem to be decreasing, but we could not define a direct correlation due to other variables that might be missing in the data.

## H. Conclusion

The intention of this project was to help investor and buyers to decide which neighborhood they may chose and what price they may have to pay. The whole process of investing gets cumbersome with so much variation in any major city. Through such a model it will be helpful for investors and buyers to gain a quick insight of the neighborhoods or locality or they may choose a locality depending on their budget.

Based on our analysis of neighborhoods of Chennai, we can define each of the clusters as below

| Cluster Label | Definition |
|---|---|
| **Cluster 1** | Lowest average rates, farthest from center with highest neighborhoods |
| **Cluster 2** | High average rates, closer to center with low neighborhoods |
| **Cluster 3** | Low average rates, far from center with higher neighborhoods |
| **Cluster 4** | Highest average rates, closest to center with least neighborhoods |

Table 1 – Cluster definition

## I. Future Directions

This analysis opens up many questions that can be further explored. Like,

a. Average rate of housing is dependent on what other variables?
b. Can we bring in other variable to estimate the price of any upcoming housing society?
c. It seems there is more supply of housing in the farther neighborhoods which are comparatively cheaper, are people preferring to flock to those neighborhoods as per their affordability?

An analysis of any major city with various available data can be conducted to help buyers and investors to make a decision.

End