# IBM DATA SCIENCE CAPSTONE PROJECT

## Property Price & Location Data Analysis of Kolkata, India

NABA KUMAR JANA

20-August-2020

# Table of Contents

1. Introduction ..... 2
1.1 Business Problem ..... 2
1.2 Audience & Stakeholders ..... 2
2. Data ..... 2
2.1 Data Sources ..... 2
**2.1.1 Web Scraping** ..... 2
**2.1.2 Geopy Client** ..... 3
**2.1.3 Foursquare API data** ..... 3
2.2 Data preparation ..... 4
**2.2.1 Average Property Price :** ..... 4
**2.2.2 Binning AvgPrice :** ..... 4
**2.2.3 Locality Latitude & Longitude:** ..... 4
**2.2.4 Locality distance from central Kolkata:** ..... 4
**2.2.5 Locality Venue details** ..... 5
**2.2.6 Marge Venue Category:** ..... 5
**2.3 Feature selection:** ..... 6
3 Methodology ..... 6
3.1 Model Selection ..... 6
3.2 Model Evaluation ..... 7
3.2.1 Elbow Method ..... 7
3.2.2 Silhouette Analysis ..... 8
4 RESULT ..... 10
5. DISCUSSION ..... 12
6. Conclusion ..... 12

# 1. Introduction

This project is a part of IBM data Science Capstone project where I am going to explore Kolkata base location data along with property price. Kolkata is situated on the brink of traditionalism and modernity. The real estate market of Kolkata is stable and growing on a positive note, but you can still get apartments at reasonable rates in Kolkata. Among a handful of cities where real estate prices are next to stable, the city brings profitable opportunities for property purchase.

## 1.1 Business Problem

Clustered the localities of Kolkata by considering the average property price along with the location area details. Like the other major city in India, the property price in Kolkata mostly depend on two things - the property feature details itself and the location area details.

The property details include – the common facilities like club house, banquet hall, library, swimming pool, playground, Jogging track, car parking etc and the interior property features like number and size of bedroom, number of toilets, size of dining area etc.

Location area details - public transport facility like metro station, school & college, hospitals, supermarket, bank, amenities and services (e.g. restaurants, shops, and cinemas) and the most important is the distance from the centre of Kolkata.

The scope of this project is to consider the location area details along with the average price of property of each location and finally clustered these by data science methodology.

## 1.2 Audience & Stakeholders

People who want to purchased property or want to invest in Kolkata. This data analysis project tells us - what are the best localities for affordable price of house or best localities for luxury investment in Kolkata. This analysis also helps the user to find out the other similar locality along with property. It also gives us the idea about the average property price of most localities of Kolkata along with other relevant information's.

# 2. Data

This section describes the data sourced for this project, as well as the data cleansing and preparation for subsequent exploration.

## 2.1 Data Sources

The require data has been collected by Web Scraping, Geopy Client as well as Foursquare API data. This section describes each of these data sources and provides examples of the data.

### 2.1.1 Web Scraping

Web scraping is used to collect Kolkata localities Name and the minimum and maximum property price of each location from websites - https://www.99acres.com This site provide the real time live property price as below (sample)

| | LocalityName | MinPrice | MaxPrice |
|---|---|---|---|
| 0 | Bally | 2,338 | 2,932 |
| 1 | Belur | 2,508 | 3,442 |
| 2 | Bhadrakali | 2,338 | 2,678 |

### 2.1.2 Geopy Client

A python library that used to get the latitudes and longitudes of all localities of Kolkata. Below are the sample data by combing with the above list.

| | LocalityName | MinPrice | MaxPrice | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | Bally | 2,338 | 2,932 | 22.647 | 88.3436 |
| 1 | Belur | 2,508 | 3,442 | 22.6357 | 88.3398 |
| 2 | Bhadrakali | 2,338 | 2,678 | 22.6744 | 88.3433 |

### 2.1.3 Foursquare API data

This project is used to access Foursquare venue data for each locality of Kolkata using latitude and longitude. The Foursquare venue data particularly seek to identify venues of categories like public transit (like train, metro), school & college, hospital, shopping mall, bank etc that are categorized base on availability. These data will then be used for subsequent comparison and categorization to provide insight to the business problem.

Following is a sample of the imported data showing particularly the venues (by name) and the respective venue categories for each neighbourhood.

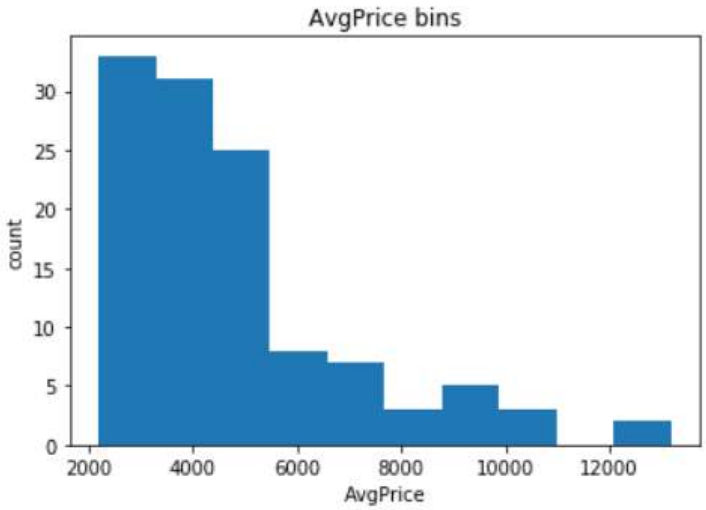
| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Bally | 22.646958 | 88.343612 | China Gaon | 22.646231 | 88.350838 | Chinese Restaurant |
| 1 | Bally | 22.646958 | 88.343612 | Bally Ghat Railway Station | 22.652423 | 88.347909 | Train Station |
| 2 | Bally | 22.646958 | 88.343612 | Bally Halt | 22.652869 | 88.347239 | Bus Station |
| 3 | Bally | 22.646958 | 88.343612 | Thatware | 22.651007 | 88.337460 | Business Service |
| 4 | Bally | 22.646958 | 88.343612 | Axis Bank ATM | 22.642562 | 88.350811 | ATM |
| 5 | Bally | 22.646958 | 88.343612 | Bally Railway Station | 22.655027 | 88.340431 | Train Station |
| 6 | Belur | 22.635732 | 88.339822 | Belur Railway Station | 22.635719 | 88.339820 | Train Station |
| 7 | Belur | 22.635732 | 88.339822 | Theatre Road Puchkas | 22.636763 | 88.346969 | Food Truck |
| 8 | Belur | 22.635732 | 88.339822 | Noxx | 22.635457 | 88.349100 | Diner |
| 9 | Bhadrakali | 22.674365 | 88.343289 | KFC | 22.669260 | 88.345520 | Restaurant |
| 10 | Bhadrakali | 22.674365 | 88.343289 | Axis Bank ATM | 22.680790 | 88.343353 | ATM |
| 11 | Bhadrakali | 22.674365 | 88.343289 | Mio Amore | 22.668330 | 88.345830 | Bakery |
| 12 | Bhadrakali | 22.674365 | 88.343289 | Uttarpara Station | 22.680636 | 88.341430 | Light Rail Station |
| 13 | Bhadrakali | 22.674365 | 88.343289 | Apollo Pharmacy | 22.666499 | 88.346984 | Pharmacy |

## 2.2 Data preparation

2.2.1 Average Property Price : Received min & max property price by Web scraping of https://www.99acres.com and now its require to insert another column name *"AvgPrice"* calculated by min & max price and then change the *"AvgPrice"* data type to 'int'.

2.2.2 Binning AvgPrice : Binning is a process of transforming continuous numerical variables into discrete categorical 'bins', for grouped analysis.

Let's plot the histogram of *"AvgPrice"* feature, to see what the distribution of price looks like.




In our property price dataset, *"AvgPrice"* is a continuous numerical variable ranging from 2189 to 13175, it has 81 (70%) unique values. To simplify analysis and as per above histogram, we can define five 'bins' with the ranges as below:

- Price less than Rs. 4386 -> Eco
- Rs. 4386 to Rs. 6583 : Low
- Rs. 6583 to Rs. 8780 : Mid
- Rs. 8780 to 10977 : High
- Price more than 10977 : MHigh

Insert new column *"AvgPrice-Binned"* into property price dataset.

2.2.3 Locality Latitude & Longitude: Using Geopy Client, we get the Latitude & Longitude value of each locality Name and added as new feature into property price dataset

2.2.4 Locality distance from central Kolkata: For each locality, calculate kilometre distance from central Kolkata (*Latitude = '22.5726', Longitude = '88.3639'*) by haversine distance calculation method.

The final property price data set looks like below.

| | LocalityName | MinPrice | MaxPrice | AvgPrice | AvgPrice-binned | Latitude | Longitude | Distance |
|---|---|---|---|---|---|---|---|---|
| 0 | Bally | 2,338 | 2,932 | 2635 | Eco | 22.646958 | 88.343612 | 8.53 |
| 1 | Belur | 2,508 | 3,442 | 2975 | Eco | 22.635732 | 88.339822 | 7.44 |
| 2 | Bhadrakali | 2,338 | 2,678 | 2508 | Eco | 22.674365 | 88.343289 | 11.51 |
| 3 | Chandannagar | 2,550 | 3,018 | 2784 | Eco | 22.861472 | 88.370607 | 32.13 |
| 4 | Hindmotor | 1,955 | 2,550 | 2252 | Eco | 22.683216 | 88.348237 | 12.40 |

### 2.2.5 Locality Venue details

This project will access Foursquare venue data for each locality of Kolkata. Base on real estate data, the property price linearly impacts to the specific category of venues up to 4 kilometres that's why we used RADIUS=4000 to retrieve Foursquare venue data. The major venue categories which are strongly impact into the property price are: Public transit (like train, metro), School & College, Hospital, Supermarket/Shopping mall, Bank, Restaurant, Multiplex.

### 2.2.6 Marge Venue Category:

Since the property price might have dependency on venue called 'Restaurant' but not any specific type of restaurant like *Chinese Restaurant, Indian Restaurant*. To simplify, I have merged all 'Restaurant' type value category's into single "Restaurant' category. We apply the same marge concept into venue category *'Grocery Store', 'Supermarket', 'Mall'* into single category 'Supermarket'.

Finally, the Kolkata Venue data set with total **1862 venues for 117 Locality's** looks like below.

| LocalityName | Venue | Venue Category |
|---|---|---|
| Bally | Inox | Multiplex |
| Bally | Dakshineswar Railway Station | Train Station |
| Bally | Liluah Railway Station | Train Station |
| Belur | Inox | Multiplex |
| Belur | Liluah Railway Station | Train Station |
| Belur | Dakshineswar Railway Station | Train Station |
| Bhadrakali | Bally Railway Station | Train Station |
| Bhadrakali | Bally Halt | Bus Station |
| Bhadrakali | Konnagar Station | Train Station |
| Bhadrakali | Dakshineswar Railway Station | Train Station |
| Chandannagar | chandannagar | Restaurant |

### 2.3 Feature selection:

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of model. Upon examining the meaning of each feature from two dataset *'kol_housing_price_final' & 'kolkata_venues_final',* below are the feature's consider for modelling.

1. Average Property Price (Binned) – the average price of housing sale in each locality.
2. Distance – the Kilometres distance of each locality from the centre of Kolkata.
3. Venue Category – The venue categories which are really impact into the property price for this data analysis: Restaurant, Multiplex, Supermarket, public transit (Metro, Rail, Bus, Ferry), Bank.

## 3 Methodology

The data set has built with the combination Kolkata locality housing price along with Foursquare venue data. Both source data has gone through the above 'Data Preparation' stage and the final data set which are ready for Machine learning Algorithm as below (sample)

| LocalityName | Airport Lounge | Airport Service | Airport Terminal | Bus Station | Indie Movie Theater | Metro Station | Movie Theater | Multiplex | Restaurant | Supermarket | Train Station | AvgPrice | Distance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Action Area 1A | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.500000 | 0.500000 | 0.000000 | 0.000000 | 4781 | 13.20 |
| Action Area 1B | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.500000 | 0.500000 | 0.000000 | 0.000000 | 4441 | 13.20 |
| Action Area 1C | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.500000 | 0.500000 | 0.000000 | 0.000000 | 4632 | 13.20 |
| Action Area 1D | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.500000 | 0.500000 | 0.000000 | 0.000000 | 4399 | 13.20 |
| Action Area I | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.300000 | 0.700000 | 0.000000 | 0.000000 | 4632 | 19.40 |
| Action Area II | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.222222 | 0.666667 | 0.111111 | 0.000000 | 5270 | 23.70 |
| Action Area III | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 | 0.000000 | 4526 | 27.60 |
| Agarpara | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 | 2189 | 12.46 |
| Airport | 0.066667 | 0.066667 | 0.066667 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.133333 | 0.466667 | 0.133333 | 0.066667 | 3187 | 14.00 |
| Alipore | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.052632 | 0.078947 | 0.789474 | 0.052632 | 0.026316 | 13175 | 5.29 |

## 3.1 Model Selection

Based on our data set exploratory data analysis, we are going to use K-means clustering which is one of the simplest and popular unsupervised machine learning algorithms. This method divides or partitions the data points, final data set into a pre-determined "k" number of clusters where each data point belongs to only one group. The algorithm is initialized with randomly chosen **k = 5 centres** or centroids. Below is the sample of Kolkata Housing data set with K-Mean cluster (K=5) Levels.

| LocalityName | MinPrice | MaxPrice | AvgPrice | AvgPrice-Binned | Latitude | Longitude | Distance | Vanues | Cluster Labels |
|---|---|---|---|---|---|---|---|---|---|
| Bally | 2,338 | 2,932 | 2635 | Eco | 22.646958 | 88.343612 | 8.53 | Bus Station, Multiplex, Train Station | 4.0 |
| Belur | 2,508 | 3,442 | 2975 | Eco | 22.635732 | 88.339822 | 7.44 | Bus Station, Multiplex, Train Station | 4.0 |
| Bhadrakali | 2,338 | 2,678 | 2508 | Eco | 22.674365 | 88.343289 | 11.51 | Bus Station, Train Station | 4.0 |
| Chandannagar | 2,550 | 3,018 | 2784 | Eco | 22.861472 | 88.370607 | 32.13 | Train Station | 4.0 |
| Hindmotor | 1,955 | 2,550 | 2252 | Eco | 22.683216 | 88.348237 | 12.40 | Bus Station, Train Station | 4.0 |

## 3.2 Model Evaluation

Contrary to supervised learning where we have the ground truth to evaluate the model's performance, clustering analysis doesn't have a solid evaluation metric that we can use to evaluate the outcome of different clustering algorithms. Moreover, since kmeans requires k as an input and doesn't learn it from data, there is no right answer in terms of the number of clusters that we should have in any problem.

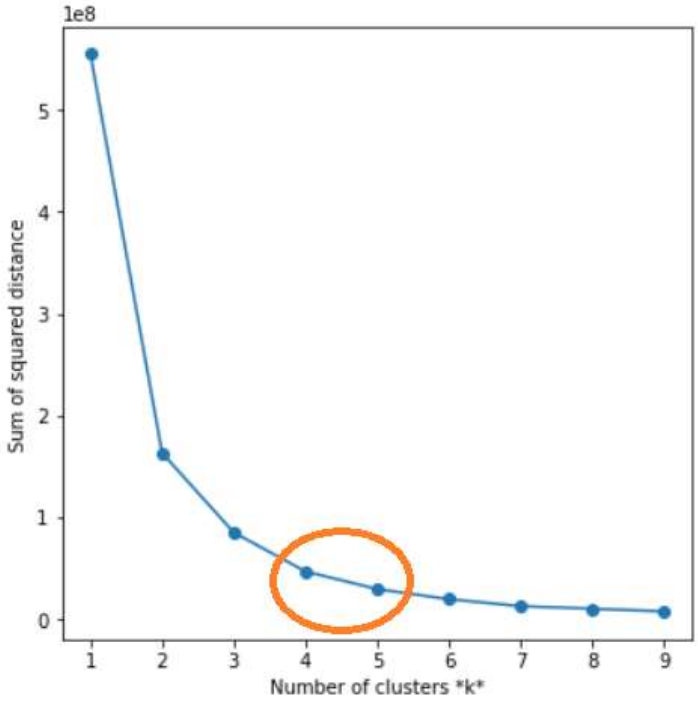
Two metrics that may give us some intuition about k:

- **Elbow Method**
- **Silhouette Analysis**

### 3.2.1 Elbow Method

Elbow method gives us an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids.

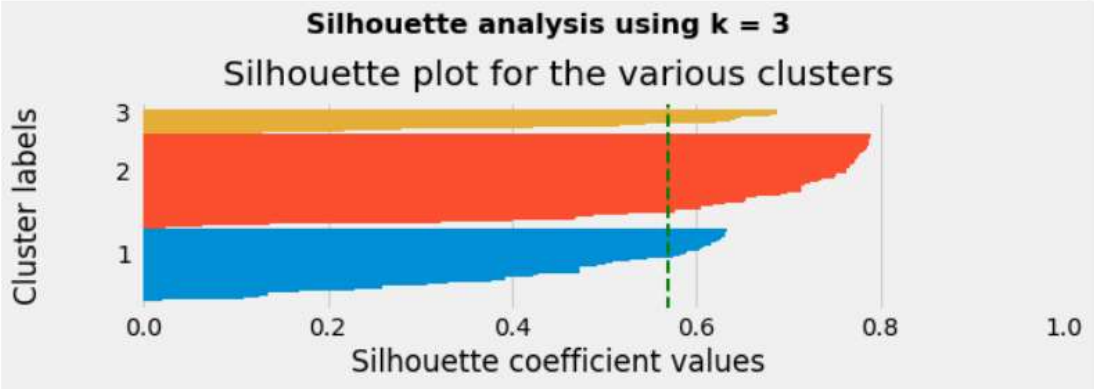
Start with a small cluster value, let's say 1. Train the model using 1 cluster, calculate the inertia for that model, and finally plot it in the graph. increase the number of clusters **up to 10**, train the model again, and plot the inertia value. This is the plot we get:

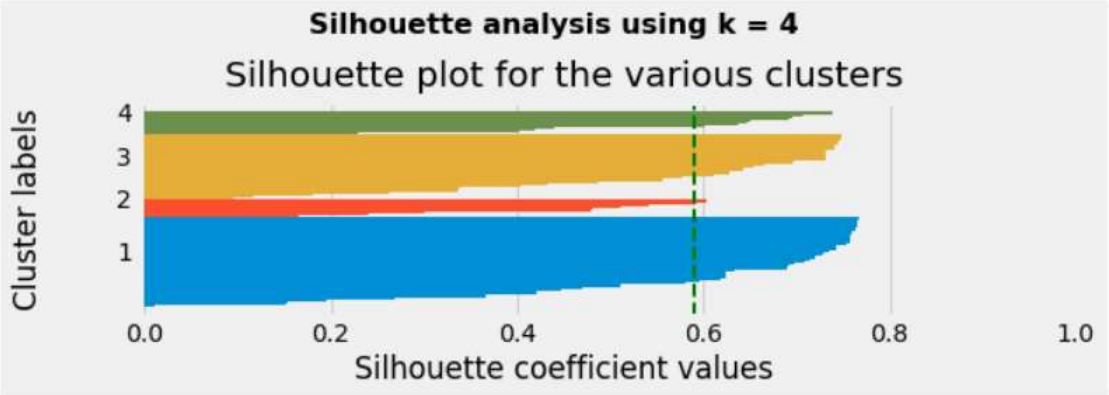


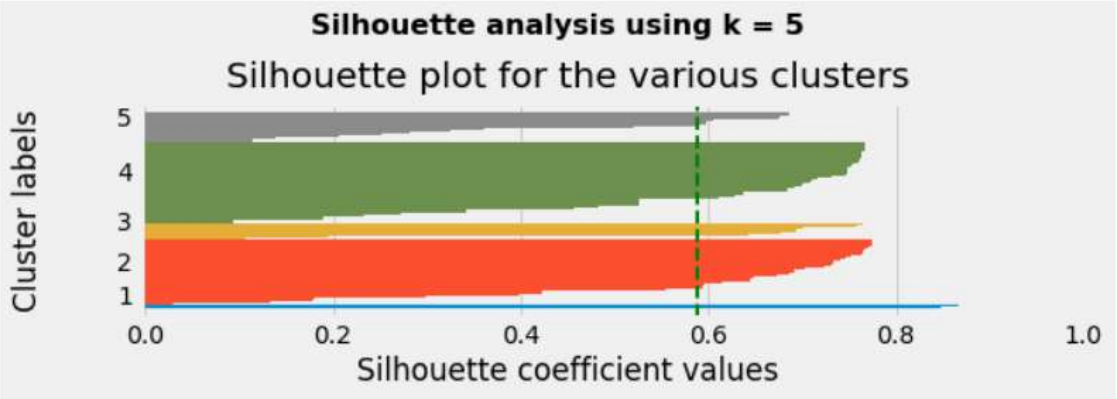

**Here, as per Elbow graph, we can choose number of clusters either 4 or 5**

### 3.2.2 Silhouette Analysis

Silhouette analysis is used to determine the degree of separation between clusters. Silhouette coefficient values range between -1 and 1. Larger numbers indicate that samples are closer to their clusters than they are to other clusters.




As the above plots show, number of clusters either 4 or 5 has the best average silhouette score also, the thickness of the silhouette plot gives an indication of how big each cluster is.

**Finally, based on Elbow Method & Silhouette Analysis we are going to used number of clusters K = 4 for our Kolkata base housing data clustering.**

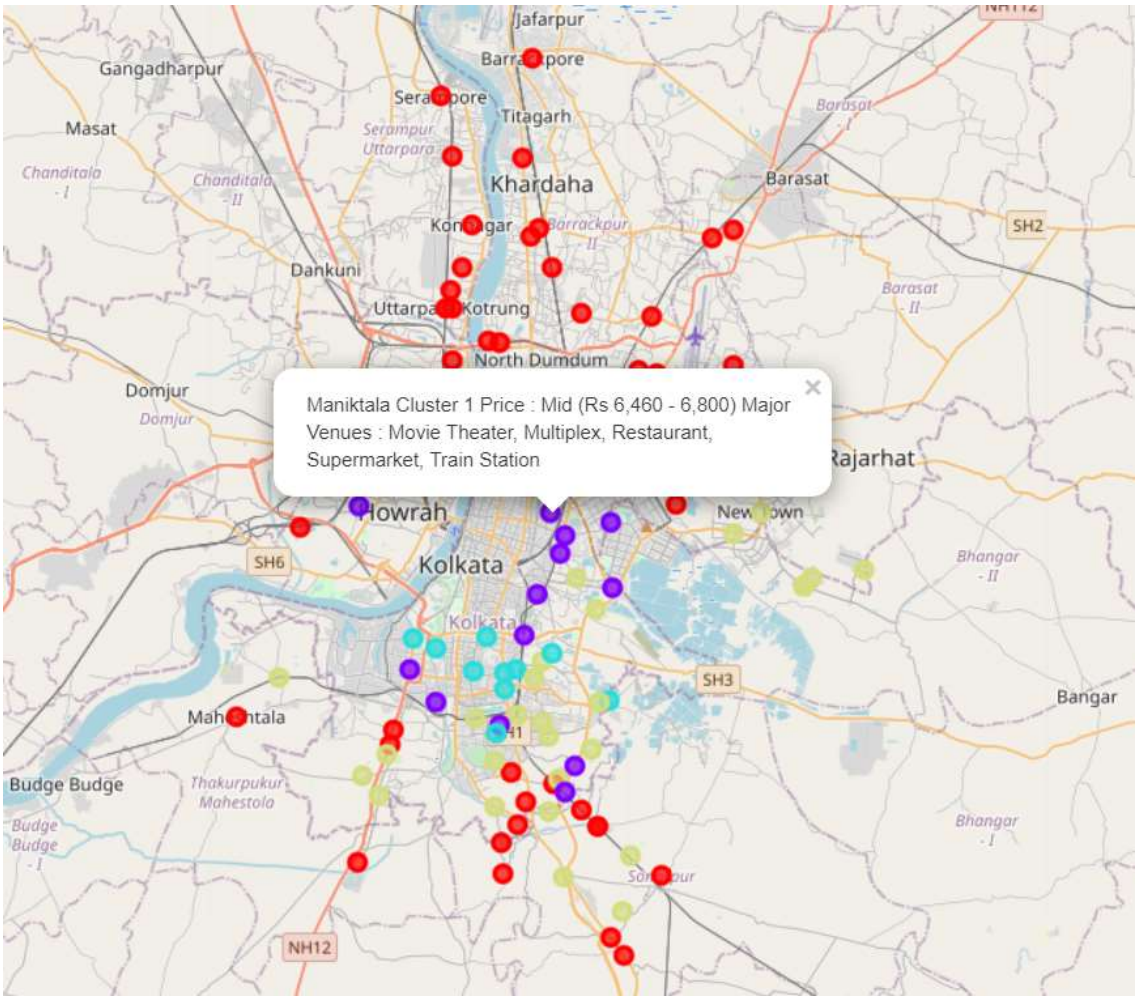
Below is the sample of Kolkata Housing data set with K-Mean cluster (K=4) Levels.

| LocalityName | MinPrice | MaxPrice | AvgPrice | AvgPrice-Binned | Latitude | Longitude | Distance | Vanues | Cluster Labels |
|---|---|---|---|---|---|---|---|---|---|
| Bally | 2,338 | 2,932 | 2635 | Eco | 22.646958 | 88.343612 | 8.53 | Multiplex | 0.0 |
| Belur | 2,508 | 3,442 | 2975 | Eco | 22.635732 | 88.339822 | 7.44 | Multiplex | 0.0 |
| Bhadrakali | 2,338 | 2,678 | 2508 | Eco | 22.674365 | 88.343289 | 11.51 | Bus Station, Train Station | 0.0 |
| Chandannagar | 2,550 | 3,018 | 2784 | Eco | 22.861472 | 88.370607 | 32.13 | Train Station | 0.0 |
| Hindmotor | 1,955 | 2,550 | 2252 | Eco | 22.683216 | 88.348237 | 12.40 | Bus Station, Train Station | 0.0 |

# 4 RESULT

This results section provides an overview of the outcomes of the methodology and their relevance to the original problem of identifying locality of Kolkata along with relevant details.

**Used python folium library to visualize Kolkata locality's along with housing price range and price category and related venues as bellows**




In the above choropleth map, which also display the below information's on **each marker** to the map:

- **Locality Name**
- **Cluster Level**
- **Price Range and Category (Eco, Low, Mid, High…)**
- **Housing Price related venues category**

The total Kolkata Locality Housing data has been **Clustered by 4** and each cluster represent the category of price, let's look each cluster locality data

**Cluster 1**: Housing Price Category – **ECONOMY**

| LocalityName | MinPrice | MaxPrice | AvgPrice | AvgPrice-Binned | Latitude | Longitude | Distance | Vanues | Cluster Labels |
|---|---|---|---|---|---|---|---|---|---|
| Bally | 2,338 | 2,932 | 2635 | Eco | 22.646958 | 88.343612 | 8.53 | Multiplex | 0.0 |
| Belur | 2,508 | 3,442 | 2975 | Eco | 22.635732 | 88.339822 | 7.44 | Multiplex | 0.0 |
| Bhadrakali | 2,338 | 2,678 | 2508 | Eco | 22.674365 | 88.343289 | 11.51 | Bus Station, Train Station | 0.0 |
| Chandannagar | 2,550 | 3,018 | 2784 | Eco | 22.861472 | 88.370607 | 32.13 | Train Station | 0.0 |
| Hindmotor | 1,955 | 2,550 | 2252 | Eco | 22.683216 | 88.348237 | 12.40 | Bus Station, Train Station | 0.0 |

**Cluster 2**: Housing Price Category – **MIDIUM**

| LocalityName | MinPrice | MaxPrice | AvgPrice | AvgPrice-Binned | Latitude | Longitude | Distance | Vanues | Cluster Labels |
|---|---|---|---|---|---|---|---|---|---|
| Tara Park | 6,375 | 6,842 | 6608 | Mid | 22.590686 | 88.304942 | 6.38 | Multiplex, Restaurant, Supermarket, Train Station | 1.0 |
| Kankurgachi | 6,885 | 8,670 | 7777 | Mid | 22.578972 | 88.391517 | 2.92 | Movie Theater, Multiplex, Restaurant, Supermar... | 1.0 |
| Maniktala | 6,460 | 6,800 | 6630 | Mid | 22.588090 | 88.385283 | 2.79 | Movie Theater, Multiplex, Restaurant, Supermar... | 1.0 |
| Phoolbagan | 6,970 | 8,330 | 7650 | Mid | 22.572159 | 88.389421 | 2.62 | Movie Theater, Multiplex, Restaurant, Supermarket | 1.0 |
| Salt Lake | 6,588 | 7,140 | 6864 | Mid | 22.584470 | 88.410394 | 4.95 | Movie Theater, Multiplex, Restaurant, Supermar... | 1.0 |

**Cluster 3**: Housing Price Category – **HEIGH & VERY HEIGH**

| LocalityName | MinPrice | MaxPrice | AvgPrice | AvgPrice-Binned | Latitude | Longitude | Distance | Vanues | Cluster Labels |
|---|---|---|---|---|---|---|---|---|---|
| Alipore | 11,262 | 15,088 | 13175 | MHigh | 22.539171 | 88.327278 | 5.29 | Movie Theater, Multiplex, Restaurant, Supermarket | 2.0 |
| Anandapur | 8,500 | 9,308 | 8904 | High | 22.514839 | 88.409788 | 14.60 | Bus Station, Multiplex, Restaurant, Supermarket | 2.0 |
| Ballygunge | 9,392 | 12,452 | 10922 | High | 22.525881 | 88.366047 | 5.20 | Multiplex, Restaurant, Supermarket | 2.0 |
| Ballygunge Circular Road | 11,092 | 14,280 | 12686 | MHigh | 22.539594 | 88.358439 | 3.71 | Multiplex, Restaurant, Supermarket | 2.0 |
| Ballygunge Place | 9,350 | 10,795 | 10072 | High | 22.526946 | 88.370324 | 5.12 | Multiplex, Restaurant, Supermarket | 2.0 |

**Cluster 4**: Housing Price Category – **LOW**

| LocalityName | MinPrice | MaxPrice | AvgPrice | AvgPrice-Binned | Latitude | Longitude | Distance | Vanues | Cluster Labels |
|---|---|---|---|---|---|---|---|---|---|
| Action Area 1A | 4,462 | 5,100 | 4781 | Low | 22.559524 | 88.491716 | 13.2 | Multiplex, Restaurant | 3.0 |
| Action Area 1B | 4,208 | 4,675 | 4441 | Low | 22.559524 | 88.491716 | 13.2 | Multiplex, Restaurant | 3.0 |
| Action Area 1C | 4,420 | 4,845 | 4632 | Low | 22.559524 | 88.491716 | 13.2 | Multiplex, Restaurant | 3.0 |
| Action Area 1D | 4,208 | 4,590 | 4399 | Low | 22.559524 | 88.491716 | 13.2 | Multiplex, Restaurant | 3.0 |
| Action Area I | 4,420 | 4,845 | 4632 | Low | 22.579769 | 88.462189 | 19.4 | Bus Station, Multiplex, Restaurant | 3.0 |

## 5. Discussion

In the beginning, the Web Scraping technique is used to collect the Locality and the price details of data from one of the best retail site (https://www.99acres.com) but I see there are lots of Locality are missing. So, it's better to analysis more details about the source to find out better data source point with complete set of Localities & Price details of Kolkata.

Kmeans machine learning supervised algorithm is used to cluster our data set. In the evaluation of model to choose the correct value of K, we used **Elbow Method Analysis and Silhouette Analysis** but both analyses return contradictory result for **K=2**, which require further details analysis to know the inside.

As a residence of Kolkata, the School/Collage and Hospital are the major category which are significantly impact on housing price, but these venue categories are missing in Foursquare API data. So again, it requires more analysis to find out the alternative source.

This decision framework, particularly the identification of factors resulting from the data analysis, provides guidance to the home purchaser or investor to make the decision. The data analysis and framework highlight how various factors can be link to the housing price that target and effectively service the demand in real estate.

## 6. Conclusion

In this analysis, we have considered the impact of relative location variables on house prices. In particular, we have considered distances from the centre of Kolkata and more refined proximity measures to the types of location amenities (Venue Category) home buyers are likely to care about.

In all, our analysis shows that relative location is important to understanding house price patterns in the locality of Kolkata. In fact, our analysis indicates that relative location is so important, that a failure to incorporate it may severely hinder the pursuit of providing high value locations.