IBM Applied Data Science Capstone Project

Recommending Housing price along with venue details for Home buyers in Kolkata, India

Naba Kumar Jana

24-August-2020

Business Problem

- A lot people found it difficulty to locate a suitable house for their leaving and most of them end up selecting their house where they are not appropriate with compare to all needs.
- This is all because they do not use appropriate systems and tools for their need.
- So, for them this is a data science analysis solution to help them make the right decisions.

Audience & Stakeholders

- People who want to purchased property or want to invest in Kolkata.
- Recommended the best localities for affordable price of house or best localities for luxury investment in Kolkata.

DATA

Data Required

- List of Localities of Kolkata
- Latitude and Longitude Coordinates of the Localities
- Venue Data related to Housing price of Locality

Data Source

- www.99acres.com (Web Scraping) for Locality and Housing price
- Geocoder package for Latitude and Longitude Coordinates
- Foursquare API for Venue details

Data Sample

Kolkata Property Price Data Set

	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

Kolkata Venue Data Set

LocalityName	Venue	Venue Category
Bally	Inox	Multiplex
Bally	Dakshineswar Railway Station	Train Station
Bally	Liluah Railway Station	Train Station
Belur	Dakshineswar Railway Station	Train Station
Bhadrakali	Bally Railway Station	Train Station
Bhadrakali	Dakshineswar Railway Station	Train Station
Chandannagar	chandannagar	Restaurant

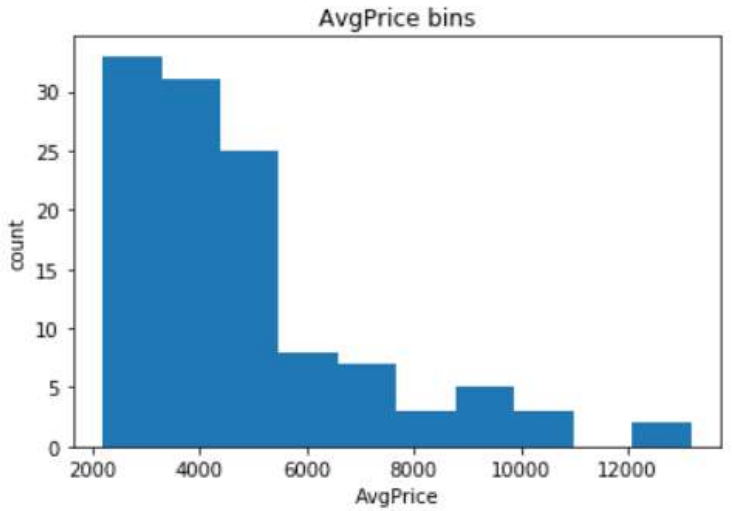
Methodology

Exploratory Data Analysis

- The property price is a continuous numerical variable ranging from 2189 to 13175, it has 81 (70%) unique values, that's why it has binned into 5 bins for better understanding.

- The venue category of each locality has been filtered and consider only those which are link to the housing price.

- New "Distance" feature has been introduced which is the kilometre distance from centre of Kolkata.

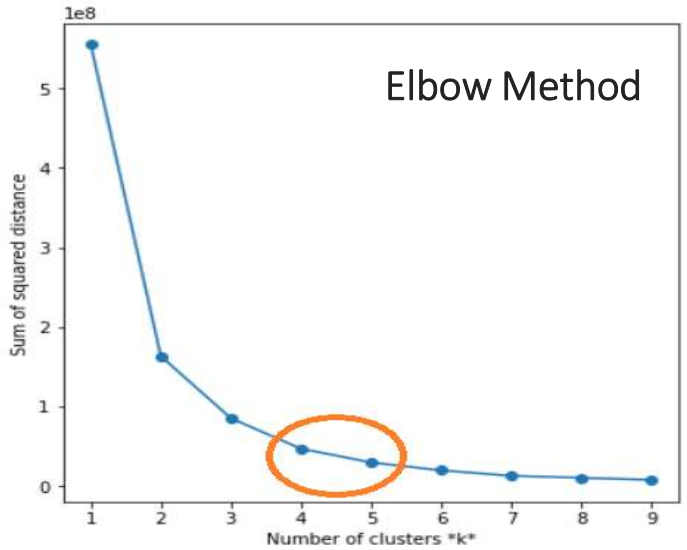


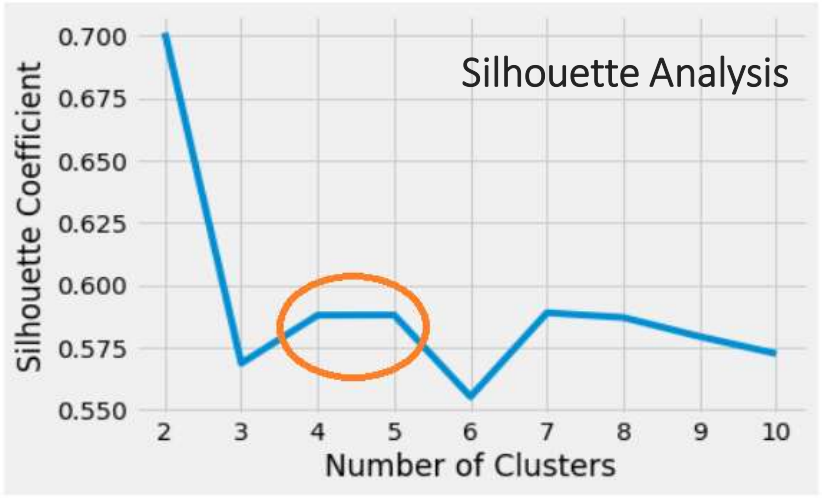
Modelling

We clustered our housing dataset using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size.

Model Evaluation

Since kmeans requires k as an input and doesn't learn it from data. So, its require to use some analysis metrics to measure it. Two metrics that may give us some intuition about k are Elbow Method and Silhouette Analysis





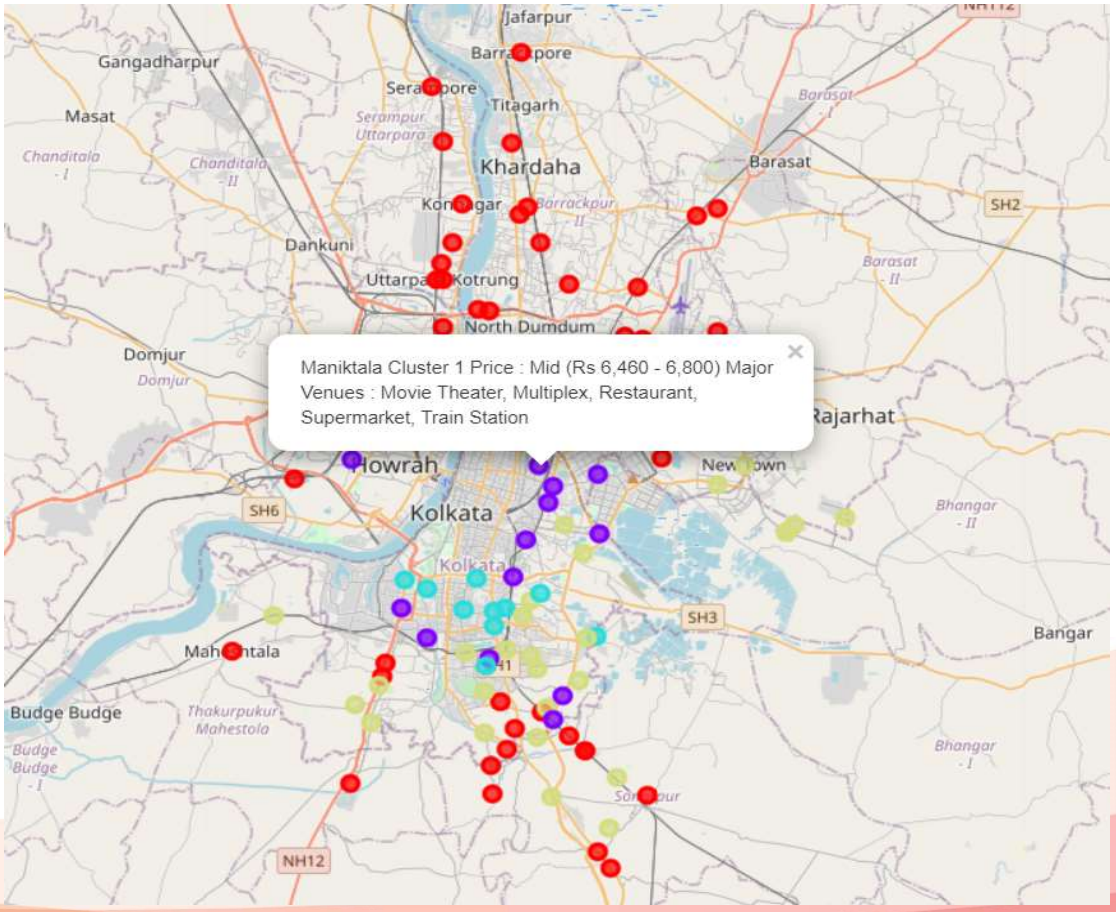
Finally K = 4, selected based on Elbow Method & Silhouette Analysis.

Result

Used python folium library to visualize Kolkata locality's along with housing price range, price category and related venues.

This choropleth map also display the below information's on **each marker**:

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



Result

Kolkata Locality Housing data has been Clustered by 4 and each cluster represent the category/range of price, let's look each cluster locality data sample

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 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 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 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

Discussion

- 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.

- There are some major venue category like school/Collage, Hospital which are significantly impact on housing price, but these venue categories are missing in Foursquare API data.

- The Elbow Method Analysis and Silhouette Analysis but both analyses return contradictory result for K=2, which require further details analysis to know the inside.

Conclusion

- This analysis considered the impact of relative location variables on housing prices. Particularly the distances from the center of Kolkata and more refined proximity measures to the types of location amenities (Venue Category) home buyers are likely to care about.

- This Data Analysis discover the housing price patterns compare to the venue details.