Segmenting Apartment complexes in Bangalore

Aravind Ganesan

June 14th 2020

1. Introduction

1.1 Background

Bangalore is a city in the southern part of India and is commonly known as Silicon Valley of India, because of being a technology and start-up capital of India. The city has also seen a real estate boom on the onset of millennium due to an increase in the affordability of the middle class. There is a value in understanding the real estate packages on offer based on the size and nearby venues. This will allow us to understand what are the various price segments in which apartments are offered and is it actually providing enough value for the price.

1.2 Problem Statement

This analysis will be focusing on a hypothetical prospective customer named Mr. Omar Little who is moving to Bangalore with family for a long time and is looking for an apartment.
He has the following requirements:

a. He is looking for a moderately sized 3 Bedroom apartment which is moderately (not premium) prized
b. He is looking for the apartments which have High schools and hospitals in the vicinity (within 2 Kms of the apartment)

1.3 Who is this for?

This analysis is for potential homebuyers in any big city across India who are home-hunting and are looking for apartments which meet their affordability criteria. This helps us in understanding how apartment offerings are packaged in the metro cities in India such as Bangalore.

2. Description of the Data

2.1 Source of Data

The data was sourced from Kaggle (here) where a user has collected the real estate data from a website. The data had was available as 2 CSV files:

a. ***apartment* dataset*:*** *datasets_151373_358770_apartment_data.csv* – List of apartments in Bangalore with their longitude and latitude
b. ***apartment details* dataset*:*** *datasets_151373_358770_blore_apartment_data.csv* – The size (number of rooms), area, price of the apartments ()

The foursquare API was then be used to get the neighbourhood venues (within 5000m radius) to these apartments.

The foursquare API was also used to fetch the hospitals and schools near this apartment by specific queries on category hospitals and schools.

2.2 Data Cleaning

2.2.1 Data Cleansing for apartments.csv:

a. The apartment dataset has just 4 columns with the *apartment name*, *latitude* and *longitude* and *geometry*. The geometry column was dropped as it was a repeat of the coordinates again.

b. The names column had city appended to each name of the column. The city name was parsed out of the *names* columns

2.2.2 Data Cleansing for apartment details.csv

a. The apartment details csv had four columns namely *names, Price, Area, Unit type.*

b. The *Price* column were coded as alphanumeric (e.g. 100000 as 1 L, 1000000 as 1C) and the lower and upper prices were coded in the same column. The column was split based on '- 'and the amounts were converted to a numeric value.

c. The *Unit Type* had BHK appended to it. To convert it to a numeric value, the BHK was split off.

d. Some of the apartments had types such as 1 RK Studio Apartment, 1 RK Apartment, 1RK, Studio. These were coded as 0.5 BHK apartments.

e. There were some properties which had an entry simply as a Plot. These entries were dropped.

f. The Price and Area fields were converted to Float data types. The low and high price were made same if one of them was not available.

g. The Mean price and Mean area per property was determined to standardize and cluster it later.

2.2.3 Merging

a. The data frames created above were merged into a single final data set which was used for further processing.

2.2.4 Summary of pre-processing

Input file name	Input Columns	Columns Dropped	Columns Created	Output Data frame Name
datasets_151373_358770_ apartment_data.csv	*names,lat,lon,geometry*	*geometry*	NA	bang_apt
datasets_151373_358770_blore_ apartment_data.csv	*names,price,area,Unit type*	*price,area,Unit type*	Size(BHK),Area-L,Area-H,Price-L-num,Price-H-num	bang_apt_details

2.3 Four Square data fetch

2.3.1 Venues search

a. The two dataframe (*bang_apt* and *bang_apt_details*) were merged and final Bangalore data set(*bang_fin*) was created.

b. Since the interest is in 3BHK apartments, a new dataframe *df_surya* was created for 3BHKs with average price and areas averaged.

c. The foursquare API was then called for each of these apartments to get the venue details(*df_venues*) which was then one-hot coded(*bg_onehot*) and clustered.

d. This allowed clustering of the apartments based on their price and nearby venues and allowed us to select the cluster based on moderate pricing.

2.3.2 **Hospital and Schools Data**

a. The cluster data which corresponded to the moderate pricing was moved to a new dataframe(*bg_cluster1*)

b. The **query search** feature of Four Square was used to get hospitals(*df_hosps*) and schools(*df_schools*) in the 2Km radius of the apartments.

c. This data was then grouped based on number of hospitals(*df_hosps_grouped*) and schools(*df_schools_grouped*) near the apartments and the mean distance of those hospitals and schools.

d. The data was then merged to create a dataset(*bg_final_sel*) of apartments which had both hospitals and schools near the apartments which yielded 147 such apartments.

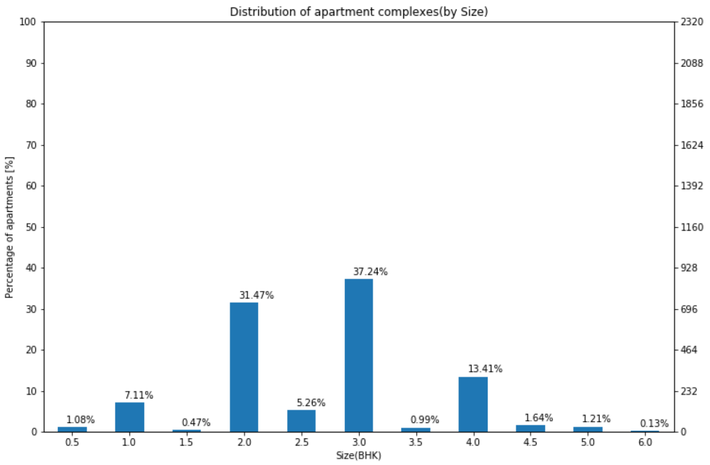
e. This data was then sub-clustered again to understand the best deal possible which satisfies the requirements and cluster labels were added.

f. The top 20 venues near each of these apartments were added again to provide a holistic view of the apartments with all the venues and the average distance of hospitals and schools near them.

3. Methodology

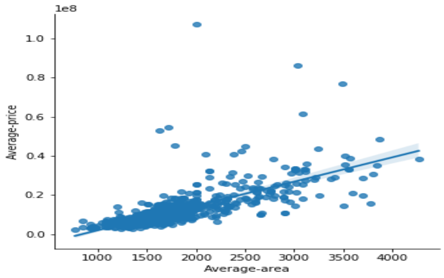
3.1 Descriptive Analysis

3.1.1 **Count by Unit Type**: The data was analysed to understand the data availability for each type of apartment. This is to understand if we have sufficient samples to make valid inferences.



There is a good percentage of data available for 3BHKs which is the focus of this analysis.

3.1.2 **Relationship between Price and Area:** The relationship between price and area of the apartment was plotted for 3 BHKs to understand how much the area determines what the price is.



We can infer that the relationship is mostly linear which says that the area is one of the features which explains the price of the apartment.

3.2 Methodology

3.2.1 K-Means Clustering for identifying the segment of apartments to satisfy the price requirement.

a. The venues in vicinity of 5 Kms for each of the 3 BHK apartments were fetched using the foursquare API venues feature to build a venues dataframe(*df_venues*).

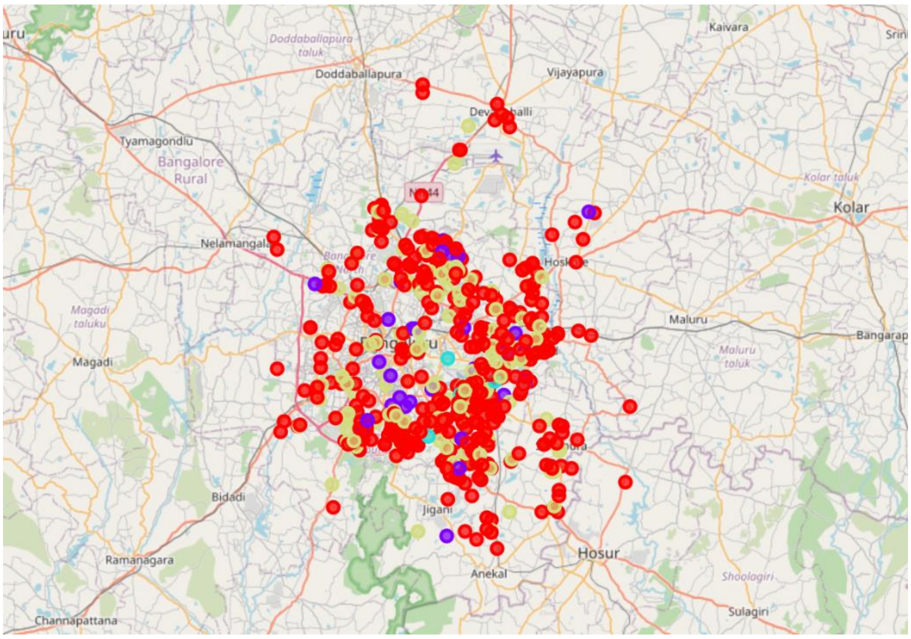
b. These venues were then one hot coded and a new dataframe was built by scaling the values on the dataframe using Standard Scalar(*bg_cluster_final*).

c. Various values of K were tried when eventually 4 meaningful clusters could be derived.

d. Using K Means clustering the dataset was clustered which resulted in the following 4 Clusters:

   a. **Cluster 0** - Medium size 3 BHKs with Moderate to premium pricing in proximity to Restaurants, Supermarkets, Yoga Studio and Gyms
   b. **Cluster 1** - Large size 3 BHKs with Premium to ultra-premium pricing
   c. **Cluster 2** - Large size 3 BHKs with Moderate to premium pricing in proximity to Restaurants, Coffee shops
   d. **Cluster 3** - Medium size 3 BHKs with premium to ultra-premium pricing in proximity to Shopping Malls, Hotels, pubs

e. The segments thus derived can be visualized as below:



f. Based on the pricing requirement, it is clear that the analysis needs to focus on **Cluster 0** which consists of moderately priced apartments as Mr. Omar wishes.

3.2.2 Building data for only apartments with hospitals and schools

a. The second requirement was to identify apartments which had a hospital or a high school in the 2kms radius to the apartment.

b. The query feature of Foursquare API was used to pull only Hospitals and Schools. The datasets were then grouped on number of Hospitals and schools and their average distance from the apartment.

c. This data was merged with the cluster1 dataset to arrive at sub-clusters.

3.2.3 Creating Sub Clusters by based on vicinity to hospitals, schools and other venues along with price.

a. The dataset this built was clustered again using k-means to identify sub-clusters and their value proposition.

b. The fields considered for clustering were only Price, Area, Number of hospitals, Average distance from hospitals, Number of schools and Average distance from Schools.

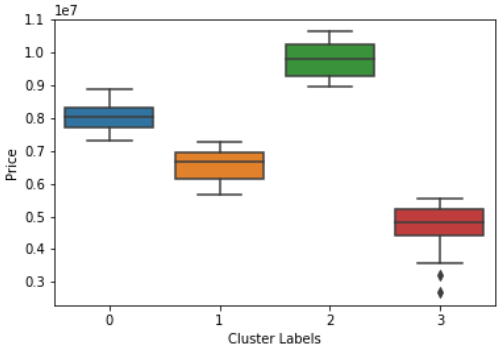
c. The subclusters were thus obtained were as follows:

   a. **Sub Cluster 0:** Sub Premium Priced 3BHKs (INR 7 Million to 8.5 Million)

   b. **Sub Cluster 1:** Moderately Priced 3BHKs (INR 5.6 Million to 7 Million)

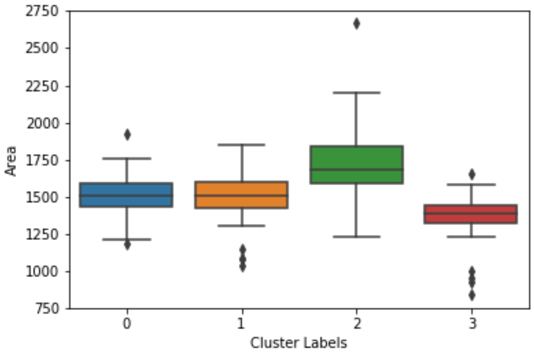
   c. **Sub Cluster 2:** Premium Priced 3BHKs (INR 8.6 Million to 10.3 Million)

   d. **Sub Cluster 3:** Value priced 3 BHKs (INR 2.5 Million to 5.5 Million)



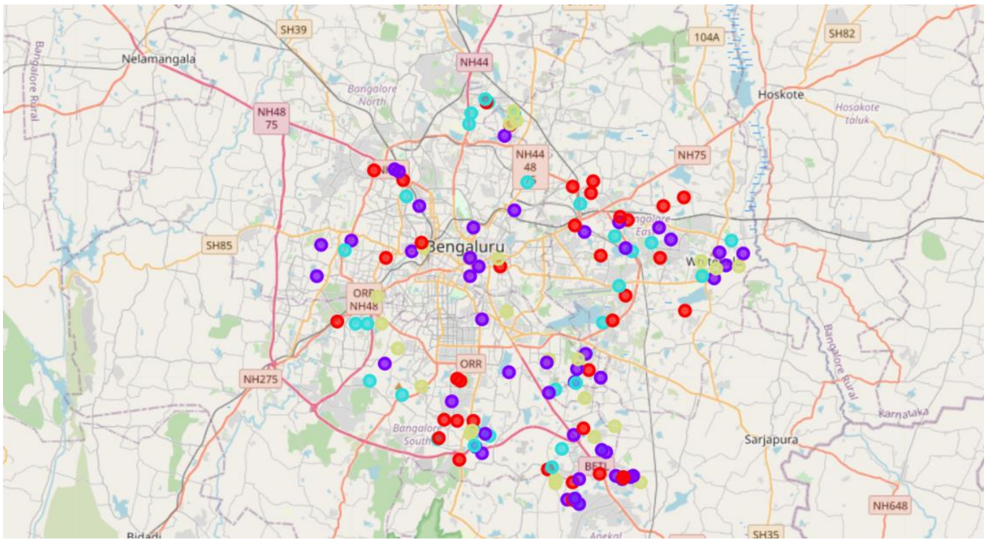
This shows that the clusters are clearly demarcated on Price

d. Lets analyze if the area of the apartments has any bearing on the pricing.



As it is clear that although pricing is different the area size of the apartment is not much different.

e. The clusters thus created can be visualized as below on the map.



4. Results

1. It was noted that based on the three criteria provided in the problem statement, a variety of offerings could be provided to the customer with each offering having its clear demarcation on price and having venues such as Restaurants, Supermarkets, Yoga Studio and Gyms.
    - Sub Premium Priced 3BHKs (INR 7 Million to 8.5 Million)
    - Moderately Priced 3BHKs (INR 5.6 Million to 7 Million)
    - Premium Priced 3BHKs (INR 8.6 Million to 10.3 Million)
    - Value priced 3 BHKs (INR 2.5 Million to 5.5 Million)
2. The offerings also consider the requirement for the proximity to High Schools and Hospitals and provide the other venues which are in the proximity to the apartments.
3. The customer can thus choose the candidate apartments from these packages and evaluate them for purchase based on his affordability.

5. Discussion

1. Based on his budget, Omar Little can make a purchase decision on any of these sub-segments as they are clearly segregated by Price.
2. The nearby venues do not seem to have a major impact on the price of the apartments as all of them seem to have a restaurant and cafes. This may lead us to believe that the venues might be coming up because these apartments are there and not vice versa.
3. The builders in each of the price subsegments are different. This leads us to believe that the brand plays a large role in the pricing that they command regardless of the area of the apartments.
4. There are other factors such as amenities, location in the city which could be contributing to the pricing of the apartment which can be the subject of a future analysis

6. Conclusion

This study creates a basic framework in understanding how the apartment offerings in the city of Bangalore are packaged by using the price, area, size of apartments and nearby venues as basis. I used a problem statement as basis to cluster this data and built clearly demarcated packages which will aid a future purchaser in making a purchase decision for these apartments.

7. Future Studies

There is a paucity of data in terms of amenities and ratings of venues in and around the apartments. If that information can be obtained, then this study can be used to develop a pricing model for apartments in Bangalore based on the parameters such as Area, Size, Average ratings of venues and types of amenities provided in the society. The factor of the brand is still an intangible which plays a major role in how these apartments are priced as this study as demonstrated. There is a domain knowledge also needed to further enhance this study in terms of what drives the prices of the apartments apart from the factors mentioned above some of which could be closeness to IT or industrial corridors. These features can be used further to enhance the study.