**Housing Sales price & Neighborhood Analysis of Chennai**

Akila Balasundaram

November 11,2019

Synopsis:

1.Introduction:

2.Data

3.Methodology

4.Discussion

5.Conclusion

**1. Introduction**

**1.1 Background**

Chennai is one of the Metropolitan cities located in Southern part of India. Chennai is the capital of Tamil Nadu which serves as biggest cultural, economic and educational center of South India. According to 2011 census this city is sixth most populated city of India and together with adjoining regions is 36th-largest urban area by population in the world.

**1.2 Interest**

With above said growth rate and people moving towards Chennai urban area, we are trying to find best neighborhoods for migrated people to settle down based upon housing process and little details on neighborhood.

**2. Data**

**2.1 Description of data**

To consider the problem we can list the data as below,

* Obtained all the areas in Chennai and their pin codes from [here](https://www.mapsofindia.com/pincode/india/tamil-nadu/chennai).
* Most of the housing sale price details ,their areas and various stats from Kaggle data sets from [here](https://www.kaggle.com/nishant4k/chennai-house-pricing-).
* Using Four square API’s to determine the venues for each neighborhood.
* Using Geocoder function to get the Lat, Lon values for above dataset.

**2.2 Data cleaning**

Data downloaded or scraped from multiple steps has to be cleaned there were lot of missing values and data needed to be massaged into correct format.

For example the dataset which had the housing sale prices has been collected for entire Chennai metropolitan cities however the data that had pin codes of Chennai had only data’s of Chennai city. But our data analysis should be made for entire metropolitan area.

So the neighborhoods whose pin codes were missing was manually added to the data set.

Also many neighborhoods had entry with different spellings viz Anna Nagar had been entered as Ann nagar,Ana nagar for multiple records these kind of mistakes were identified for all the neighborhoods in the data and corrected.

**2.3 Feature selection**

After cleaning the data and merging the data from two different sources it consist of 6128 rows and 23 column. But looking at the columns we have to decide the columns which will be useful for analysis and columns which needs to be dropped .

Let us look into some of the features and how its decided that they will not contribute for any analysis.

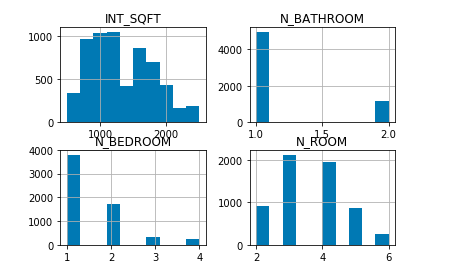
* State – This feature is not necessary since for all the records it has been given as Tamil Nadu and moreover
* District – This feature is not required because this data set belongs to various neighborhoods of Chennai

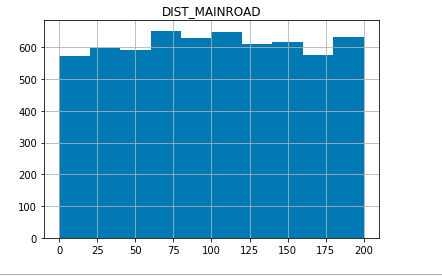
|  |  |  |
| --- | --- | --- |
| **Kept features** | **Dropped features** | **Reason for dropping features** |
| AREA | PRT\_ID | Clearly these features does not contribute to any trends into the sale price |
| IN\_SQFT | DATE\_SALE |
| N\_BEDROOM | QS\_ROOM | These fields are same as normal number of bedroom,bathroom,total number of rooms |
| N\_BATHROOM | QS\_BATHROOM |
| N\_ROOM | QS\_BEDROOM |
| PARK\_FACILTIES | REGFEE | These features are ambiguous whether they contribute to sales price |
| SALES\_PRICE | COMMIS |
| PINCODE | STREET |
| DIST\_MAIN |  |  |

**3. Exploratory Data Analysis (Methodology)**

Part 1: ***Trends of housing prices***

To test the trends of all the fields we just plotted them into histograms and saw the potential relation ship between a feature and sales price of house.





Where INT\_SQFT – inches in square feet.

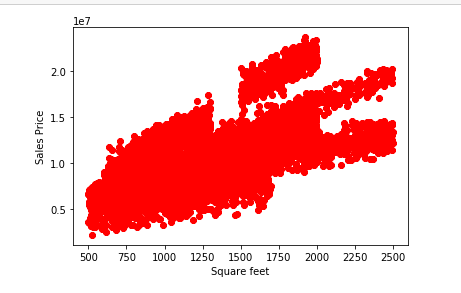
N\_BATHROOM – number of bathrooms

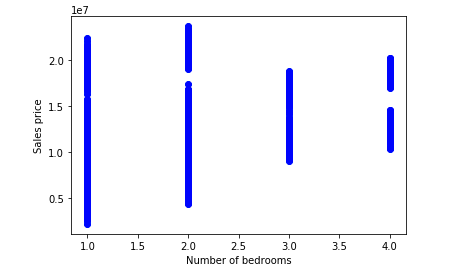
N\_BEDROOM – number of bedrooms

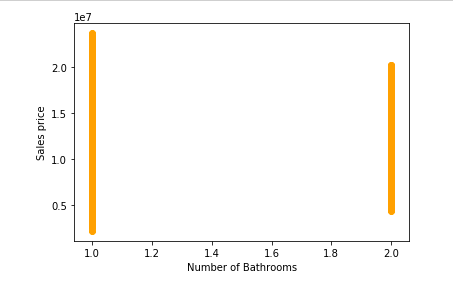
N\_ROOM – Total number of rooms.

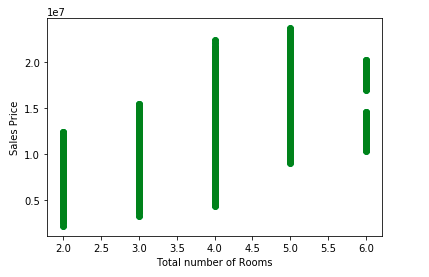
DIST\_MAINROOD – Distance from main road

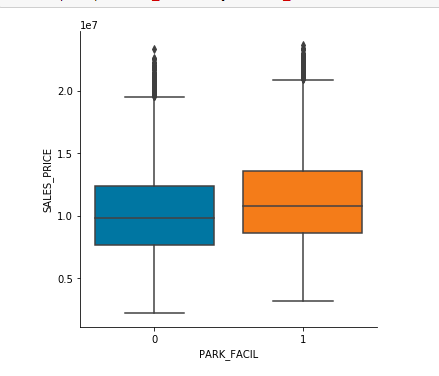
Lets check the linearity of the above features with sales price so that we can determine how the model might be built











Looking at the above charts there is no linearity between any of the features VS sales price

Let us first try **Multiple linear regression**.

**'INT\_SQFT','N\_BEDROOM','N\_BATHROOM','N\_ROOM','DIST\_MAINROAD' vs SALES\_PRICE**

The coefficients after applying the multiple linear regression are

Coefficients: [[ 3.39482061e+03 -1.42451424e+06 -2.27928903e+06 2.31744037e+06

8.35844995e+02]]

This has been evaluated with train and test data and evaluated the variance score

Of 0.64 which is good but there is still room for improvement.

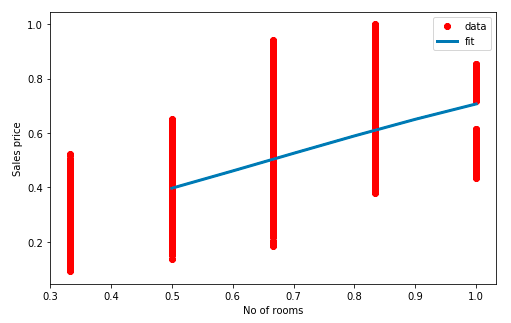
**80%** of the data is used for **training** and **20**% of data was used for **testing.**

So we proceeded multple with just using square feet inches and total number of rooms and calculated the variance which was about 0.45

Below given tabular column has various methodology used to build a model and their evaluation methods.all the models has used 80% of data as training data and 20 % as tesitng data

|  |  |  |
| --- | --- | --- |
| **Methodology** | **Features used** | **evaluation results** |
| Multiple linear regression | INT\_SQFT','N\_BEDROOM','N\_BATHROOM','N\_ROOM','DIST\_MAINROAD' vs SALES\_PRICE | variance=0.64 |
| 'INT\_SQFT','N\_ROOM' vs SALES\_PRICE | variance=0.45 |
| logistic Regression | N\_ROOM' vs' SALES\_PRICE' | R2 score=-0.31 |
| Polynomial (degree 7) | N\_ROOM' vs' SALES\_PRICE' | R2 score=-0.34 |

**Visual of logistic regression with the training data**



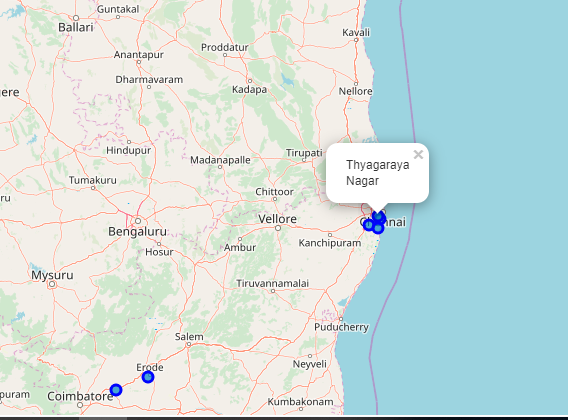
So it is decided that multiple linear regression of **'INT\_SQFT','N\_BEDROOM','N\_BATHROOM','N\_ROOM','DIST\_MAINROAD' vs SALES\_PRICE**

would be the best fit model to be used in this case of data and the same has been evaluated.

***Part 2: A peek into Neighborhoods of Chennai***

As a next part we are going to analyze the neighborhoods of Chennai and going to give a rough idea for the immigrants what are their choices.

I used python **folium** library to visualize geographic details of Chennai and its Neighborhoods and I created a map of Istanbul with boroughs superimposed on top. I used latitude and longitude values to get the visual as below

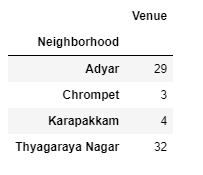


I utilized the Foursquare API to explore the boroughs and segment them. I designed the limit as **100 venue** and the radius **500 meter** for each boroughs from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API.

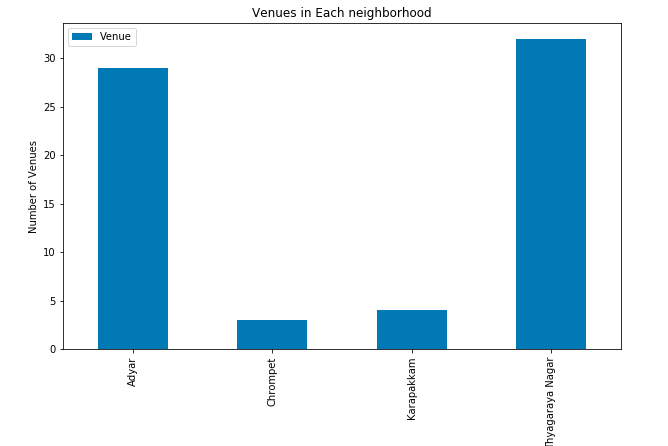
In summary of this data **29** venues were returned by Foursquare for adyar . Here is a merged table of boroughs and venues.



Looking at the table below we can notice that Adyar and Thyagaraya nagar has returned more than 25 venues but Chrompet and Karapakkam has returned venues that are less than 5.



The result doesn’t mean that inquiry run all the possible results in boroughs. Actually, it depends on given Latitude and Longitude information and here is we just run single Latitude and Longitude pair for each borough. We can increase the possibilities with Neighborhood information with more Latitude and Longitude information.

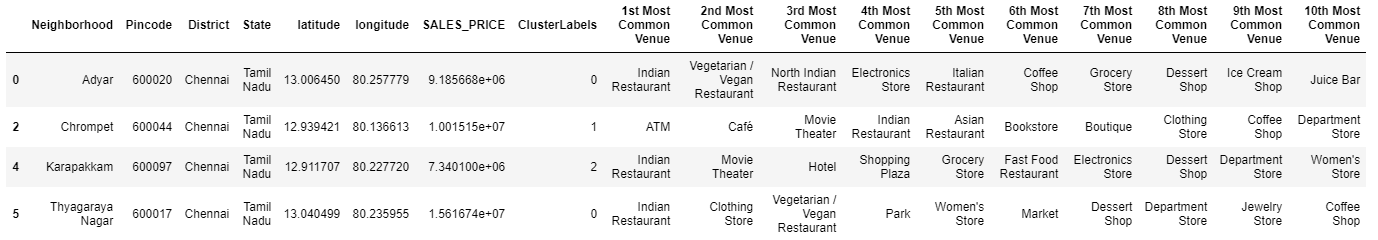


In summary of the above graph **71** unique categories were returned by Foursquare, then I created a table which shows list of top 10 venue category for each borough in below table.



We have some common venue categories in boroughs. In this reason I used unsupervised learning **K-means algorithm** to cluster the boroughs. K-Means algorithm is one of the most common cluster method of unsupervised learning.

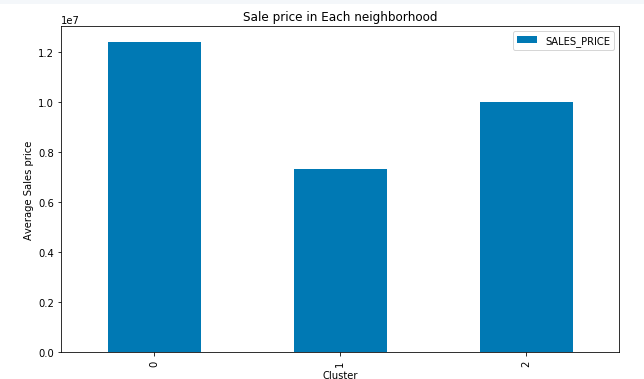
First, I will run K-Means to cluster the boroughs into **3** clusters because when I analyze the K-Means with elbow method it ensured me the 3 degree for optimum k of the K-Means.



After estimating at the number of 1st most common venue in each Cluster , we can label the clusters

* Cluster 0 : “Indian Cuisine Venues”
* Cluster 1 : “Entertainment & Social Venues”
* Cluster 2 : “Clothing,lifestyle & Intensive Cafe Venues”

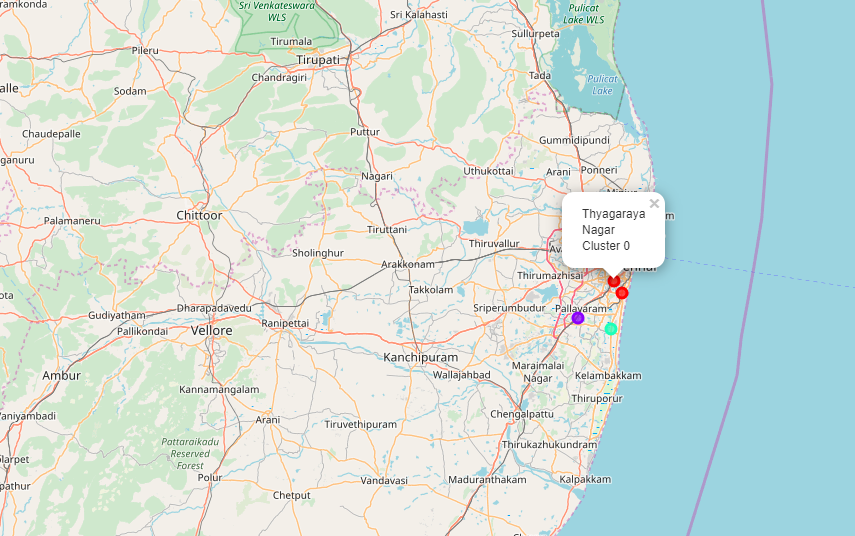
As it seems in above histogram, we can define the ranges as below:



* Cluster 0 : “High Level HSP”
* Cluster 1 : “Low Level HSP”
* Cluster 2 : “Mid Level HSP”

Results:

One of my aims was to give the details of various boroughs of Chennai so that people who are migrating will be able to know which places will be best for both their accommodation, lifestyle and business needs



**4. Discussion:**

As I mentioned before, Chennai is a big city with a high population density in a narrow area. The total number of measurements and population densities of the various districts in total can vary. As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high quality results for this metropol.

I used the Kmeans algorithm as part of this clustering study. When I tested the Elbow method, I set the optimum k value to 3. For more detailed and accurate guidance, the data set can be expanded and the details of the neighborhood or street can also be drilled.

I ended the study by visualizing the data and clustering information on the Chennai map. In future studies, web or telephone applications can be carried out to direct investors.

**5.Conclusion**

As a result, people are turning to big cities to start a business or work. For this reason, people can achieve better outcomes through their access to the platforms where such information is provided.

Not only for investors but also government officials can manage the city more regularly by using similar data analysis types or platforms.