How Property Amenities and Locality Affect Housing Prices in Bangalore

BY: Thomas Jones

For: Springboard Data Science Career Track

10/7/2021

Contents

[Introduction 3](#_Toc84779720)

[Data 3](#_Toc84779721)

[Data Wrangling 4](#_Toc84779722)

[Exploratory Data Analysis 4](#_Toc84779723)

[Data Visualization 5](#_Toc84779724)

[Correlation Analysis 7](#_Toc84779725)

[The problem statement 8](#_Toc84779726)

[Preliminary Analysis 8](#_Toc84779727)

[Conclusion of Simple MLR Model 9](#_Toc84779728)

[Machine Learning Modeling Methods: Linear Regression of the price with Gradient Boosting 10](#_Toc84779729)

[Training Data 10](#_Toc84779730)

[Prediction of Property Prices 11](#_Toc84779731)

[Future Research 13](#_Toc84779732)

# Introduction

***Bangalore***, officially known as *Bengaluru*, is the capital and the largest city of the Indian state of Karnataka. It has a population of more than 8 million and a metropolitan population of around 11 million, making it the third most populous city and fifth most populous urban agglomeration in India. Located in southern India on the Deccan Plateau, at a height of over 900 m (3,000 ft) above sea level, Bangalore is known for its pleasant climate throughout the year. Its elevation is the highest among the major cities of India.

The goal of this project is to predict the price of a house in Bangalore. In recent years, the development of real estate industry has become an important driving engine of economic growth, but the real estate industry is also suffered criticism (Pyhrrey et al., 2004). The factor that the price of house is unaffordable draws the government and public attention (Case and Shiller, 2003). The factors influencing the prices of real estate are diversified and complex. The development of different cities in India project the same issues and challenges with every other country. I was interested to identify the predictors of the real estate market prices using different methods of predictive analysis and machine learning algorithms.

Many factors are effecting the real estate costs. Yihong (2016) introduced Real estate investment, Land price, loan interest rates and completed residential area as the variables to describe the supply model, Population, GDP and income as variable indicators to describe the model of demand. Considering the needs of research and data availability, from the above variables, I choose real estate investment, interest rate as my supply factors; population, GVA, and income as demand factors. My goal was to find a complete data set on features and amenities of different properties.

The data set that I used was found in Kaggle <https://www.kaggle.com/ruchi798/housing-prices-in-metropolitan-areas-of-india?select=Kolkata.csv> it was satisfying the requirements for this project.

I completed my analysis using MLR with gradient, random forest, and plain MLR that provided me with very similar results identifying the main predictors being the ***square footage and number of bedrooms*** and in addition variables identifying that if the property was in a cosmopolitan or urban area.

# Data

The data set that I used was found in Kaggle <https://www.kaggle.com/ruchi798/housing-prices-in-metropolitan-areas-of-india?select=Kolkata.csv> it was satisfying the requirements for this project.

The data set in Kaggle included a total of 6208 properties. Unfortunately many of them had missing data including a major one for my research. I had to reduce the data set to 1856 properties that included information for most of the 40 different variables. The variable had descriptions of amenities and features for each property. For each property there was information on the price, the neighborhood that it was located in, the square footage, the region/state the property was located in, number of times it was sold. In addition information regarding amenities included in the building was also in the data set such as elevator, maintenance, cafeteria, garden, gymnasium, etc. Extraneous data regarding the appliances and furniture that the apartment or house offered were also included in some cases, and conclusions on if the apartment was furnished or not, was possible to be reached.

The most important variables that presented were the following:

1. **Price**: Numeric variable expressed in local currency Rupees
2. **Area**: Numeric variable expressed in Square footage
3. **Number of Bedrooms**: Integer variable
4. **Region**: Categorical with 7 levels *(North, South, East West, Central, North East, South East)*
5. **Resale**: Boolean variable identifying if the property was sold or only had one owner
6. **Multipurpose Room**: Boolean variable identifying if the building has a multipurpose room
7. **Rainwater Harvest**: Boolean variable identifying if the building has a rainwater harvest
8. **Car Parking**: Boolean variable identifying if the building has car parking
9. **Intercom**: Boolean variable identifying if the building has an intercom
10. **Cafeteria**: Boolean variable identifying if the building has a cafeteria
11. **Landscaped Gardens**: Boolean variable identifying if the building has a landscaped garden
12. **Fitness Facilities**: Integer variable enumerating the number of fitness amenities the building or neighborhood offers
13. **Vaastu Compliant**: Boolean variable identifying if the building is vaastu compliant is a science (shastra) of arranging the five elements – earth, water, fire, air and sky in complete harmony. Very important in Indian tradition
14. **Multiple other variables that were not of great important for our project**

## Data Wrangling

There were a few variables such as **Wifi** and **AC** that were indicating that these were not available for any of the apartments in the data set and they were removed from the working data set.

I cleaned up my data by getting rid of some factors that could skew analysis based on missing data. I then assessed how I could bundle some categories together to make larger, more useful ones. The four I came up with were furniture, fitness, appliances, and entertainment. Knowing that in the US an apartment or house comes with appliances, it was interesting to identify observations in my dataset where the number of appliances was zero. This was one of the reasons why I combined these together. Regarding the furniture, I wanted to create a variable that was focusing on furnished vs unfurnished apartments and their differences. Unfortunately the number of furnished apartments in the dataset was limited.

I wanted to identify categories of apartments based on the luxury and I focused on some variables that could be categorized as luxury amenities. I split those into two categories **fitness** and **entertainment**. It was interesting to see that fitness was a major predictor on the price of the apartment.

# Exploratory Data Analysis

I ran the summary of descriptive analytics for each one of the variables. I reviewed the results and identified and initial subset of variables to include in my research. In addition I reviewed the correlations between variables to see if there is multicollinearity in my data before I run the regression.

The mean price of properties in my dataset is approximately 10m rupees. Within the dataset there are outliers of properties costing 2B rupees. Most of the properties in my dataset are new and had only 1 owner. The maximum number of bedrooms is 5 while the average number is 2.5. While the square footage of the properties have an average of 1500 with a maximum of 10,000. In general it seems that most of the properties are in buildings with gymnasiums, swimming pools, landscaped gardens and mostly located in urban, higher end neighborhoods and they are overall unfurnished. Note: Based on the dataset, properties in Bangalore are rented without appliances.

|  |
| --- |
| **Price Area Bedrooms Resale**  **count** 1,856 1856 1856 1856  **mean** 9,880,561 1509 2.5 0.1  **standard** 13,218,820 776.8 0.7 0.3  **deviation** |

# Data Visualization

My goal now is to better visualize the relations between the regions and the amenities that the properties offer as well as the relationship between the regions and prices of the properties. I would like to dissect my dataset in a way to identify the best predictors of prices and the most important amenities for location of the properties.

The box plot below shows how extremely high the prices in the Central region are compared to everywhere else even though median prices seem relatively stable.

|  |  |
| --- | --- |
|  |  |

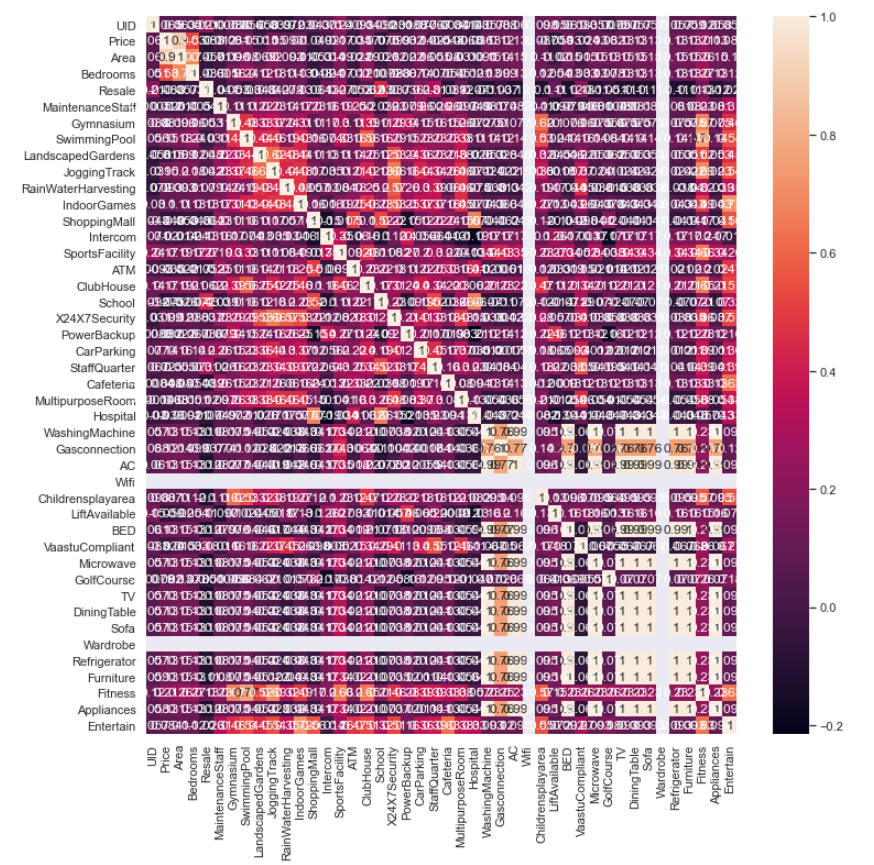
All the extreme outliers are in the Central region while the rest of the properties that are located in all other regions have prices that are closer to the mean. The West and Northeast regions have the most uniformed prices towards the lower end.

|  |  |
| --- | --- |
|  |  |

Analyzing all the different factors that affect the price of the properties led us to focus on the finance rules of supply and demand and analyzed the number of properties that are available in each region as well as the size of properties. We can clearly see that the most populated area is South with the 2nd largest being the Central while the Northeast seems to have the lowest density. And we can clearly see that the rule of supply and demand affects the price with Northeast having the lowest prices overall while Central and South have the highest prices overall.

# Correlation Analysis

We are reviewing the different variables in order to identify possible correlation between different variables. For example we'd expect that properties with a jogging track and gymnasium to also include a swimming pool. We would like to identify variables that show multicollinearity to ensure that they do not affect our linear model. Due to the fact of having more than 40 different variables, reviewing the matrix manually is difficult. This is why we also decided to create a heatmap that appears under the table.



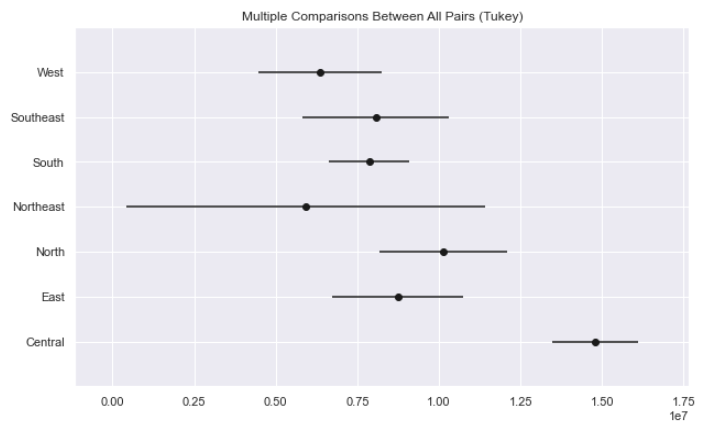
# The problem statement

Not being a native of Bangalore I used the data provided to better understand the real estate market in the area with a goal to help my business customer form an opinion on how to best invest in Bangalore real estate. The problem is multifaceted as every business problem the focus is to limit cost and increase revenue. In order to do this efficiently we need to understand all the underlying factors that shape the prices of the real estate today and in the past but also those that will affect the prices in the future. The main goal is to identify the factors that affect the real estate price in Bangalore but Bangalore is not just a single neighborhood but rather is a state with a population of 11 million. The prices of the real estate have different trends and are affected by different factors depending on the region of Bangalore they are. The goal is to find the median price of the real estate in the entire region.

# Preliminary Analysis

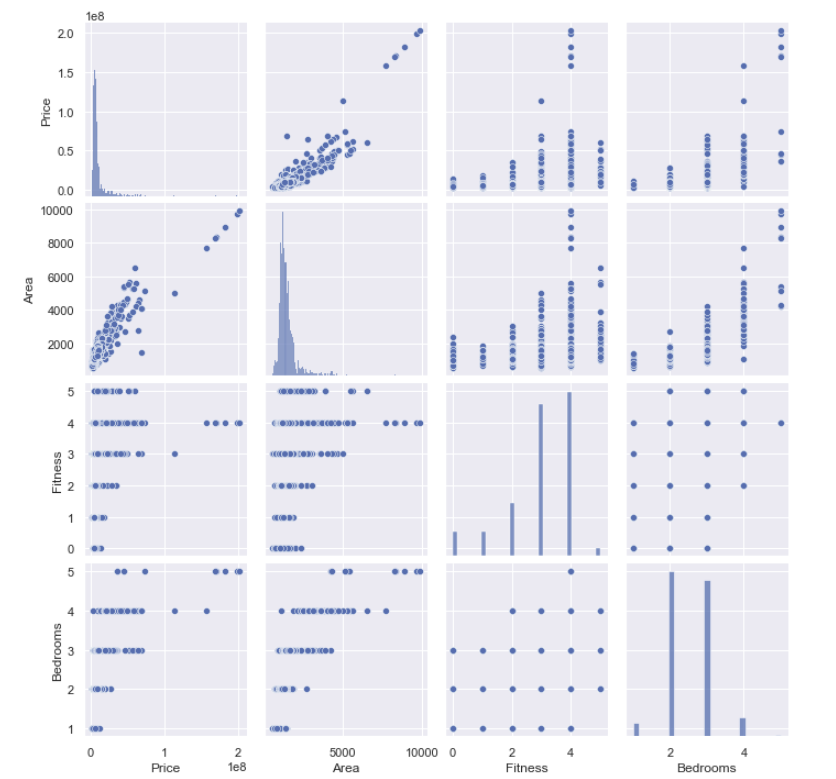
I broke the regions down based on the concentration of neighborhood groupings. (Central, East, North, Northeast, South, Southeast, and West). We can clearly see that the Central region is the most expensive with an average prices almost double that of the other regions. The **number of houses in that region** was 505 that made it the 2nd largest **after the Southern** region**, which** was actually a region with a low average price. The Central region has also the highest standard deviation that tells me that there are many different types of housing in that area and both ends of the prices. The maximum price in Central is 4 times the maximum price of the next region which is **the** South.

**In order to complete the statistical analysis to support my findings I ran an ANOVA and a Tukey test.** The graph shows us that the difference in average price is small but higher cost in the Central in comparison with every other region. The East, Southeast, South, and West regions have similar averages and ranges of prices. North region has slightly higher prices. Northeast has a big range. While Central is **not similar to any of the others** with the mean price being much higher than all the others.



Focusing only on the variables that are predictors for the prices I further reviewed their correlations I created a heat map and correlation matrix.

Based on heatmap it appears that the number of bedrooms and total area within the housing are the most important factors in the pricing. The number of bedrooms has a moderate impact. The number of fitness centers have almost no impact.



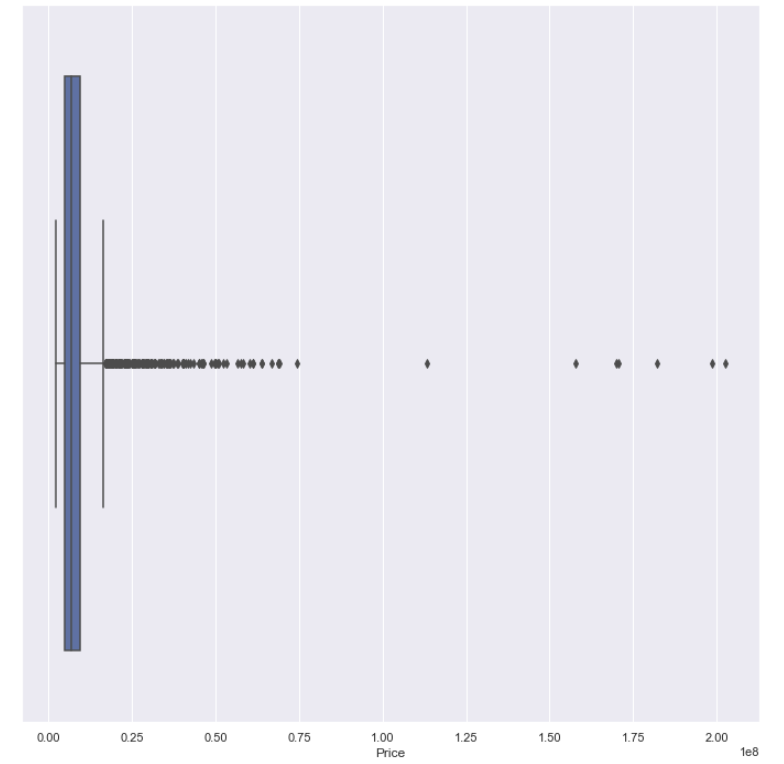
# Conclusion of Simple MLR Model

I started with lots of data and some of it was messy. But after cleaning it up a bit and performing the exploratory data analysis I was able to see strong trends. One is particular was how much more expensive it is to be in the central region compared to everywhere else. I can get a sense of income levels and what types of areas these are but not totally conclusive. There is more work I can do to find that out. But housing prices are strongly tied to the size of a home and the area it's located. There are several other factors such as number of bedrooms that have a moderate impact and others like fitness centers in the area that have almost no impact. Also some area have a much higher degree of variance in prices than others.

# Machine Learning Modeling Methods: Linear Regression of the price with Gradient Boosting

**We** run an OLS, a linear regression model with gradient boosting. The model will learn to predict prices based on the most important variables. Initially it runs with the prices of all the properties. The learned model gave me standard outliers in our predictions. In order to make the model more accurate and not dependent on outliers I reviewed the prices and identified 117 extreme outliers. This was less than 6% of the total observations. I decided to drop the outliers from the initial model and run the learning without them.

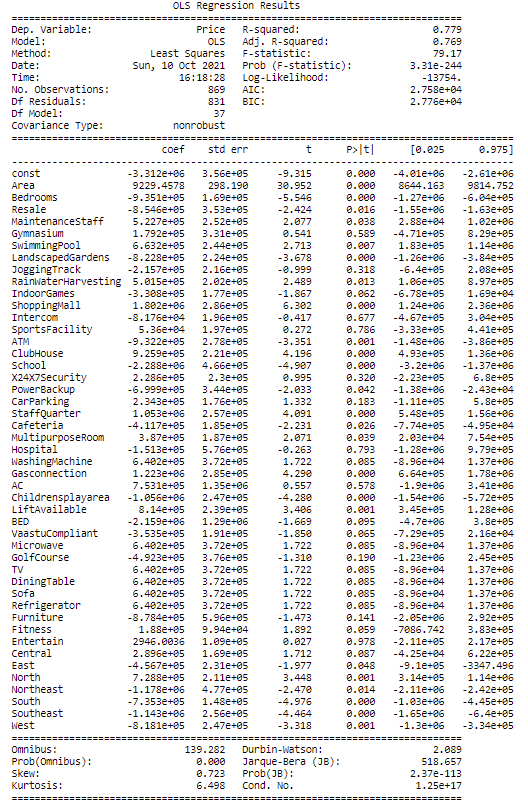
### Training Data



# Prediction of Property Prices

Using a learning rate of 0.5 we get a mean absolute error of 1.35 x10^6 that is a good error since our min apartment price is around $10 million.

|  |  |
| --- | --- |
|  |  |

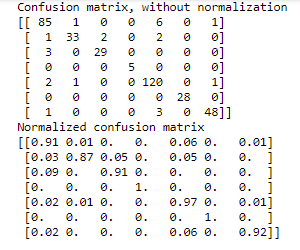


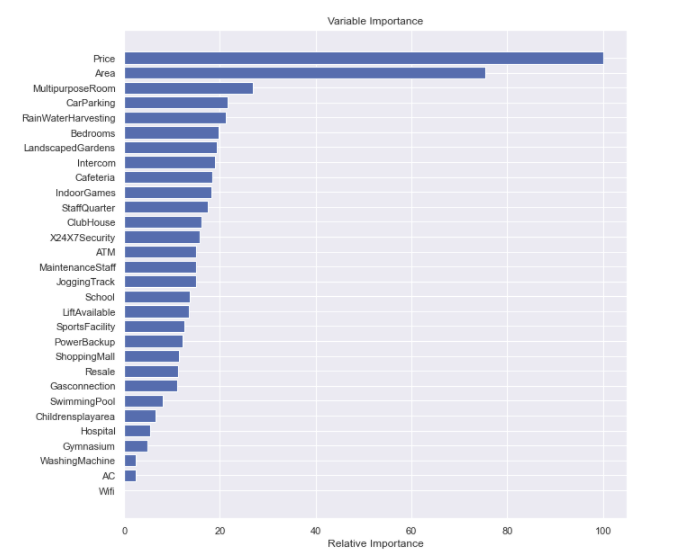
These are the results of the regression with gradient boosting. R-squared equals 79% can provide us a prediction of prices based on all the above factors.

# Future Research

One of the major questions that I have to answer for my customer is not only the average price and the predictors for the price of a property but also the coherency of the region based on different factors. My focus this time is to predict the region that the property comes from based on the 48 different variables that I have. In order to do that I chose to run a logistic regression, response variable "Region" identifying predictors. I will start my machine learning part of the project trying to verify the results of the previous work. My goal here is to build a process that learns the region where a property is located using all the other variables including the price of the properties. The region is a qualitative variable with 7 different values. We have already seen that the prices on some of the regions overlap and predicting prices based on the amenities of the properties was not easy. What I'd like to do here is to reveal any possible predictors (amenities) or price that would identify the region of the property. For this purpose I will use a logistic regression.

We prepare to run a logistic regression. Our response variable is the region. We are going to develop a model based on the data received. It will predict the region that the property lays.





My logistic regression predicts the region of a given apartment/house using the set of 40 variables. The results were quite interesting and had a very high accuracy of 93.5%. The main predictor of the region is the price of the apartment. Although somebody would expect the area to also be one of the most important predictors this doesn't seem to be the case. We can see that variables that define if a region is cosmopolitan area are the ones that are good predictors.