

**PROJECT REPORT**  
**ON**  
**HOUSE PRICE PREDICTION IN METROPOLITAN AREAS OF INDIA**

Session 2019-2021



Project submitted for the award of the degree of

Master's in Science in Statistics

To

Gauhati University,

Guwahati-14, Assam

**Submitted By-**

Aditya Gautam

M.Sc. 4th semester

Roll no. PS-191-832-0029

Regd. No. 035844

Paper code – STA4056

Department of Statistics

Gauhati University

**Project Guide-**

Dr. Pallabi Medhi

Assistant Professor

Department of Statistics

Gauhati University

### **DECLARATION**

I hereby declare that the project report on *House price prediction in metropolitan areas of India*, is the record of bona fide project work done by me under the guidance and supervision of Dr. Pallabi Medhi, and that it has not previously formed the basis for the award of any degree, diploma, fellowship or any other similar title or recognition.

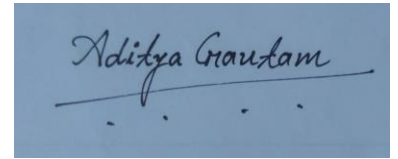
Date: 01/07/2021

Place: Guwahati

Aditya Gautam

Department of Statistics

Gauhati University

A rectangular box containing a handwritten signature in black ink. The signature is written in a cursive style and reads "Aditya Gautam". Below the signature, there are four small dots arranged horizontally.

## **ACKNOWLEDGEMENTS**

Individual effort can never contribute in totality, to the successful completion of any venture. It is with a sense of gratitude that I acknowledge the effort of a whole host of well-wishers who have in some way or the other contributed in their own special ways to the success of this effort.

At first, I would like to thank God for being able to complete this project with success. Then I wish to express my sincere gratitude to my project guide Dr. Pallabi Medhi, who has given me lots of support and encouragement and also for her valuable guidance and suggestions throughout the project.

I would also like to thank our Head of Department Sir, Prof. Kishore Kumar Das for giving us the opportunity to do this project.

Last but not the least I would like to thank my Parents who rendered me not only financial support but also a moral support without which I could not have completed my project.

I am also thankful to all the persons who helped me directly or indirectly.

Aditya Gautam

## **CONTENTS**

<b>CHAPTERS</b>	<b>PAGE NO.</b>
1.Chapter I	
1.1 Introduction	1-5
1.2 Objectives of the study	6
1.3. Review of Literature	7-10
1.4. Source of the data	11
1.5. Methodology	12-16
2. Chapter II	
2.1. Graphical Representation	17-23
2.2 Analysis of the dataset	24-33
3. Chapter III	
3.1 Result of the analysis	34-51
3.2 Conclusion	52
4. References	53

## *1. Chapter I*

### *1.1 Introduction*

The key of making money is through Investing, People often invest in Stocks, Liquid gold, crypto currency etc. One of the important investment is Real estate market, Real Estate Property is not only the basic need of a man but today it also represents the riches and prestige of a person. Investment in real estate generally seems to be profitable because their property values do not decline rapidly. Changes in the real estate price can affect various household investors, bankers, policy makers and many. Investment in real estate sector seems to be an attractive choice for the investments. Thus, predicting the real estate value is an important economic index. India ranks second in the world in number of households according to 2011 census with a number of 24.67 crore. According to the 2017 version of Emerging Trends in Real Estate Asia Pacific, Mumbai and Bangalore are the top-ranked cities for investment and development. These cities have supplanted Tokyo and Sydney. The house prices of 22 cities out of 26 dropped in the quarter from April to June when compared to the quarter January to March according to National Housing Bank's Residex(residential index). With the introduction of Real Estate Regulation Development Act (RERA) and Benami property Act throughout the country India, more number of investors are attracted to invest into real estate in India. The strengthening and modernizing of the Indian economy has made India as attractive Investment destination. However, past recessions show that real estate prices cannot necessarily grow. Prices of the real estate property are related to the economic conditions of the state In India, the real estate sector is the second-highest employment generator, after the agriculture sector. Real estate sector in India is expected to reach US\$ 1 trillion by 2030. By 2025, it will contribute 13% to country's GDP. Emergence of nuclear families, rapid urbanisation and rising household income are likely to remain the key drivers for growth in all spheres of real estate, including residential, commercial, and retail. Rapid urbanisation in the country is pushing the growth of real estate. >70-75% of India's GDP will be contributed by urban areas by 2020. According to India Ratings and Research (Ind-Ra), the Indian real estate sector may stage a sharp K-shaped recovery in FY22. However, the overall sales in FY22 could still be ~14% below the FY20 levels.

India's Global Real Estate Transparency Index ranking improved by a notch to 34 in 2019 on the back of regulatory reforms, better market data and green initiatives according to property consultant JLL.

Indian real estate attracted US\$ 5 billion institutional investments in 2020, equivalent to 93% of transactions recorded in the previous year. Investments from private equity (PE) players and VC funds reached US\$ 4.06 billion in 2020.

The office market in top eight cities recorded transactions of 22.2 msf from July 2020 to December 2020, whereas new completions were recorded at 17.2 msf in the same period. In terms of share of sectoral occupiers, Information Technology (IT/ITeS) sector dominated with a 41% share in second half of 2020, followed by BSFI and Manufacturing sectors with 16% each,

while Other Services and Co-working sectors recorded 17% and 10%, respectively. The office space leasing activity is expected to pick up in 2021 and is likely to be at par with the 10-year average, i.e., 30-31 million sq. ft.

Home sales volume across eight major cities in India jumped by 2x to 61,593 units from October 2020 to December 2020, compared with 33,403 units in the previous quarter, signifying healthy recovery post the strict lockdown imposed in the second quarter due to the spread of COVID-19 in the country.

According to the Economic Times Housing Finance Summit, about 3 houses are built per 1,000 people per year compared with the required construction rate of five houses per 1,000 population. The current shortage of housing in urban areas is estimated to be ~10 million units. An additional 25 million units of affordable housing are required by 2030 to meet the growth in the country's urban population.

The Government of India has been supportive towards the real estate sector. In August 2015, the Union Cabinet approved 100 Smart City Projects in India. The Government has also raised FDI (Foreign Direct Investment) limits for townships and settlements development projects to 100%. Real estate projects within Special Economic Zones (SEZ) are also permitted for 100% FDI. Construction is the third-largest sector in terms of FDI inflow. FDI in the sector (including construction development and construction activities) stood at US\$ 42.97 billion between April 2000 and September 2020. Exports from SEZs reached Rs. 7.96 lakh crore (US\$ 113.0 billion) in FY20 and grew ~13.6% from Rs. 7.1 lakh crore (US\$ 100.3 billion) in FY19. Indian real estate is expected to attract a substantial amount of FDI in the next two years with US\$ 8 billion capital infusion by FY22.

Government of India's Housing for All initiative is expected to bring US\$ 1.3 trillion investments in the housing sector by 2025. As of December 2019, under Pradhan Mantri Awas Yojana (Urban) [PMAY (U)], 1.12 crore houses were sanctioned in urban areas, with a potential to create 1.20 crore jobs. The scheme is expected to push affordable housing and construction in the country and give a boost to the real estate sector. On July 09, 2020, Union Cabinet approved the development of Affordable Rental Housing Complexes (AHRCs) for urban migrants and poor as a sub-scheme under PMAY-U.

Government has also released draft guidelines for investment by Real Estate Investment Trusts (REITs) in non-residential segment.

The Ministry of Housing and Urban Affairs has recommended all the states to consider reducing stamp duty of property transactions in a bid to push real estate activity, generate more revenue and aid economic growth.

To know more about Real estate Investment, major cities of the country are being highlighted which are as follows-

## **Mumbai**

Mumbai experiences similar urbanisation challenges as other fast growing cities in developing countries: wide disparities in housing between the affluent, middle-income and low-income segments of the population.

Highly desirable neighborhoods such as Colaba, Malabar Hill, Marine Drive, Bandra and Juhu house professionals, industrialists, Bollywood movie stars and expatriates. Up-scale flats have 3 or more bedrooms, ocean views, tasteful interior decoration, parking for luxury cars and sleeping quarters for maids and cooks. In 2007, Mumbai condominiums were the priciest in the developing world at around US\$9,000 to US\$10,200 per square metre.<sup>[6]</sup> Mumbai has more than 1,500 high rise buildings, many of which are just planned, but some already constructed or under construction.

Over 7 million people, over 42% of the population of Mumbai, live in informal housing or slums, yet they cover only 6–8% of the city's land area. Dharavi, Asia's second-largest slum is located in central Mumbai and houses over 1 million people. Slums are a growing tourist attraction in Mumbai.

Most of the remaining live in chawls and on footpaths. Chawls are a quintessentially Mumbai phenomenon of multi-storied terrible quality tenements, typically a bit higher quality than slums. 80 per cent of chawls have only one room. Pavement dwellers refers to Mumbai dwellings built on the footpaths/pavements of city streets.

With rising incomes, many residents of slums and chawls now have modern amenities such as mobile phones, access to electricity, often illegally, and television.

## **Delhi**

Delhi has witnessed rapid suburban growth over the past decade. South Delhi, Gurgaon and Noida have added thousands of apartment buildings, Affordable Homes, shopping centres and highways. New Delhi's famous Lutyens bungalows house the prime minister, members of his cabinet, top political and government leaders, military officials, senior judges and top bureaucrats. New Delhi is also home to thousands of diplomatic staff of foreign countries and the United Nations. With India's growth, Delhi has developed into a business center, especially for outsourcing, IT consultancy, high-tech, research, education and health care services. Employees of these institutions are the source of growing demand for high-end housing provided by major builders such as DLF.

## **Bangalore**

In the 1990s the information technology boom hit Bangalore. Y2K projects in America's IT industry resulted in shortages for skilled computer scientists and systems programmers.

Bangalore has transformed into the Silicon Valley of India as over 500,000 well-paying jobs for young college graduates were created. The demographics of the city changed, new high-rise were built, campus-style office parks sprouted, vast shopping centers started to thrive, streets became crowded with new cars and gated expatriate housing estates emerged.

Roughly 10% of Bangalore's population lives in slums.

## **Kolkata (Calcutta)**

The most sought-after neighbourhoods of Calcutta are generally centered around Lower Circular Road, Sarat Bose Road, Salt Lake, Ballygunge, Anwar Shah Road, Chowringhee and Golf Green. A recent building boom has converted sprawling British-era bungalows into high-rise condominiums and apartment-buildings with modern amenities. Kolkata currently has the second most number of highrises and tall buildings in the country, second only to Mumbai. The highest of them is at 65 floors (The 42). New suburbs are constantly being developed in Rajarhat and along the Eastern Metropolitan Bypass. These suburbs will consist of major condominiums, complete with penthouses, many designed primarily for NRIs, expats and affluent residents. The tallest buildings in the city, The South City Towers and Urbana towers, are also condominiums.

North Calcutta contains mansions built in the early 20th century during Calcutta's heyday as capital of British India. These buildings feature a courtyard surrounded by balconies, large rooms with tall ceilings, marble floors, tall pillars and crumbling artwork. Most of them are poorly maintained. The Marble Palace and other buildings received "heritage status" which provides them municipal funds and incentives to repair and restore. These mansions serve as reminder of the era of Bengali Renaissance.

It's estimated that 5% of the population live in slums. Although the number of slums are less than Mumbai, they are scattered all over Kolkata in between affluent areas giving the city a rustic and poverty driven look.

## **Hyderabad**

In Hyderabad, housing in modern ages in the 21st century is more modernized and developed than it has been in the past. The housing sector in Hyderabad has relatively sophisticated infrastructure. and is suitable for gated communities and villas, as well as higher-standard flats and condominiums. Hyderabad is home to several skyscrapers, including The Botanika, Lodha Belezza, etc. Roughly 15% of population is living in slum presently. Many residential infrastructure companies are well-established in Hyderabad.



## **Chennai**

In Chennai, Housing in the 21<sup>st</sup> century is also modernized and developed that the past. The real estate sector in Chennai has modernized and luxury infrastructure. The demographics of the city changed, new high-rise building were built, campus-style office parks sprouted, vast shopping centers started thrive, streets became crowded with new cars and gated expatriate housing estates emerged. Chennai is home to several skyscrapers, including Highliving District tower-H-SPR city, Anchorage, House of Hiranandani, Bayview House of Hiranandani, etc. Many residential infrastructure companies are well-established in Chennai. Roughly 26% of population is living in slum presently.

It is observed that in India there is privation of reliable economic method for price prediction of residential properties which results in inevitable way to trust the often-manipulated prices by middle person. The increasing purchasing power of subjects have led to increasing demand of property ultimately causing rise in prices of properties. However, the rate of fluctuation in prices should have a method for its traceability. Round the world, the ways such hedonic pricing method, multiple linear regression analysis, travel cost method, fuzzy logic system, AHP technique, ARIMA, ANN (Artificial Neural Network) techniques etc.

## ***1.2 Objectives of the study***

The objectives of these studies are-

- (1) To build a predictive model and to fit a linear regression model using ordinary least square method;
- (2) To build a predictive model using random forest methodology;
- (3) To compare the predictive mean of each city, and give an Insight about the costliest real estate in India;
- (4) To compare the linear regression model and random forest model, i.e. to find out the more accurate predictive model.

### **1.3. Review of Literature**

In Past few years, many researchers, institutions, and research centre have carried out many research works upon House price prediction and thereby enriched many literatures on it.

As Bhalla, Arora and Gill (2009) scrutinized the performance of housing sector as well as the problems and challenges faced by this sector. The study showed that due to continuous changes in the global financial environment banks and financial institutions have brought alteration in their strategies related to this sector so that slowly and gradually growth is shown by this sector. It was revealed that due to globalization process, India is witnessing competition among banks that has reduced the cost of finance for housing users.

Bhalla (2008) in a paper discussed the current scenario, development, performance, problems, challenges and prospects of housing finance as an industry segment. According to this study housing finance grow at the rate of 36 per cent. With the changes in strategies of banks and financial institution policies there is shift from buyer market to seller market.

Gundimeda (2005) studied the applicability of HPM to value water resources such as Bays, lakes and reservoirs, building of a new harbour, river views, restoration of urban stream, noise, landfills, dumping sites etc. on nearby property values. In India, hedonic price method has been employed in evaluating the relation between land prices and surface and ground water access (both in quality and quantity), (Gundimeda and Kathuria, 2005) and benefits if air quality improvement in India (Murty and Gulati, 2006). Kanojia Anita (2016) stated that the presence of Environmental Services such as parks are connected with additional environmental qualities. For example, Playground & open parks are often associated with higher air quality and lower noise levels and the presence of water can positively influence the climate of the surrounding areas. Therefore, the estimation of the capitalization of such “Services” in house prices might be biased by price effects of additional environmental qualities. It is worth noting that a park’s shape and area also have a significant effect on neighbourhood residential property values. For this reason, in the future, perspective of landscape ecology, including the landscape quality, diversity, and fragmentation should also be considered. This study can provide effective information for real estate developers, government (in terms of decision-making on environmental tax), urban and landscape planners or architects, and green space conservationists and managers. Nevertheless, in future planners have to consider such Environmental Services, besides their ecological benefits, as a source of utility for the inhabitants of cities.

Nihar Bhagat, Ankit Mohokar, Shreyash Mane (2016) studied linear regression algorithms for prediction of the houses. The goal of the paper is to predict the efficient price of real estate for customers with respect to their budgets and priorities. Analysis of past market trends and price ranges will predict future house pricing.

Sampath kumar and Santhi studied the land price trend of Sowcarpet which is the central part. They developed statistical model using economic factors and predicted that the annual rise in land price would be of 17%. Urmila reported that the past trends were analysed to ascertain the rate of growth or decline and the trends are used in forecasting. Economic parameters might be introduced to formulate more realistic relationship. Some of the other techniques they Mansural Bhuiyan and Mohammad Al Hasan 2016 use is regression, deep learning to learn the nature of models from the previous results (the

property/land which were sold off previously which are used as training data). There are different models used such as linear model data using only one feature, multivariate model, using several features as its input and polynomial model using the input as cubed or squared and hence calculated the root mean squared error (RMS value) for the model.

Ayush Varma, Abhijit Sharma, Sagar Doshi and Rohini Nair (2018) suggested that the use of neural networks along with linear and boosted algorithms improved prediction accuracy. The dataset used here contained various essential parameters. the dataset was cleaned up. Three algorithms were used namely Linear Regression, Forest Regression and Boosted Regression. The dataset was tested on all three and the results of all the above algorithms were fed as an input to the neural network. Neural networks were used mainly to compare all the predictions and display the most accurate result. A neural network along with Boosted Regression was used to increase the accuracy of the result.

Monk, Tang and Whitehead (2010) examined the social and economic impact of housing in Scottish country. Investment in housing finance impacts the economy directly and indirectly. Housing finance investment impacts the employment, GDP, productivity and many other important factors. The study revealed that the housing is an important indicator for increasing the wealth of nations. It was concluded that Scottish housing policy objective is to improve the quality standard of housing as well as to increase the investment in house old sector.

Hujia Yu, Jiafu Wu (2014) used classification and regression algorithms. According to analysis, living area square feet, roof content, and neighborhood have the greatest statistical importance in estimating the selling price for a home. And prediction analysis can be improved by the PCA technique.

Li Li and kai-Hsuan Chu (2017) studied various algorithms such as Backpropagation neural network (BPN) and Radial basis functional (RBF) neural networks. The use of RBF and BPN models is introduced to identify the difference between the house price index such as Cathy and sinny price index and complicated correlation function to detect the macroeconomic analysis.

The travel cost model is often used to measure the benefits provided by access to public recreation sites, e.g., national parks and national forests, which have relatively minor, if any, entrance fees (Oh, et al., 2005). Hotelling (1947) is credited with the initial development of the travel cost model. Using the travel cost model, observed travelers' net economic benefit, or consumer's surplus, from visiting a recreation site is calculated as the value of access to the recreation site less the travel cost and necessary entrance fees (Heberling and Templeton, 2009). The model assumes that people travel to a recreation site if the marginal value of accessing the site is at least as large as the marginal cost of traveling to the site. The estimated consumer surplus is often used as a monetary measure of consumer welfare. The aggregate net economic benefit of access to a recreation site is estimated by aggregating average individual consumer surplus per visit over all visits.

Eric Slone et.al. (2014) developed the relationships between various home characteristics and the asking price of a residential property were analysed using both a simple linear regression and a multiple linear regression using the method of ordinary least squares. Home square footage was utilized as the explanatory variable in the simple linear regression, and the multiple linear regression consisted of the addition of land parcel size, number of bedrooms, year of construction, and other explanatory variables. The results of the multiple linear regression proved the bias due to the omission of crucial factors in the

simple linear regression. Home square footage was found to be the most important factor in the determination of residential property price, while garage capacity proved to be the weakest factor Ezgi Candas et.al (2015) had found that if significance level is accepted as 0.05 all the 5 variables in the last regression model (Floor, Heating system, Earthquake Zone, Rental Value and Land Value) have a significant impact on the dependent variable Value. Land value and rental value have the highest impact on the housing price. Existing floor, heating system and earthquake zone are the following them. Although it is found that the other variable is not significant in this study, it can change according to the sample size. If sample size increases, regression model once again is recommended for further studies. The application of multiple regression analysis in a house data set explains or model's variation in house price which demonstrated good examples of strategic application of mathematical tool to aid analysis hence decision making in property investment.

Khamis et.al. (2014) compared the performance between Multiple Linear Regression (MLR) model and Neural Network model on estimate house prices in New York. A sample of 1047 houses is randomly selected and retrieved from the Math10 website. The factors in prediction house prices including living area, number of bedrooms, number of bathrooms, lot size and age of house. The methods used in this study are MLR and Artificial Neural Network. It was found that, the value of  $R^2$  in Neural Network model is higher than MLR model by 26.475%. The value of Mean Squared Error (MSE) in Neural Network model also lower compared to MLR model.

#### **1.4. Source of the data**

To study the Real estate Prices in the metropolitan areas of India, secondary dataset has been used.

Which has Data of six metro cities namely Mumbai, Hyderabad, Chennai, Bangalore, Kolkata, Delhi.

#### **Content**

This dataset comprises data that was scraped. It includes:

- collection of prices of new and resale houses located in the metropolitan areas of India
- the amenities provided for each house

#### **Inspiration**

With 40 explanatory variables describing various aspects of new and resale houses in the metropolitan areas of India, one can predict the final price of houses in these regions.

The dataset has been collected directly from the kaggle website:

<https://www.kaggle.com/ruchi798/housing-prices-in-metropolitan-areas-of-india>.

## 1.5. Methodology

To analyze the dataset, following methods have been used:

**1.5.1 Variance inflating factor:** A variance inflation factor(VIF) detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model; it's presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model. VIFs are usually calculated by software, as part of regression analysis. You'll see a VIF column as part of the output. VIFs are calculated by taking a predictor, and regressing it against every other predictor in the model. This gives you the R-squared values, which can then be plugged into the VIF formula. "i" is the predictor you're looking at (e.g.  $x_1$  or  $x_2$ ):

$$VIF = \frac{1}{1 - R_i^2}$$

### Interpreting the Variance Inflation Factor

Variance inflation factors range from 1 upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

A **rule of thumb** for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

Exactly how large a VIF has to be before it causes issues is a subject of debate. What is known is that the more your VIF increases, the less reliable your regression results are going to be. In general, a VIF above 10 indicates high correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above.

Sometimes a high VIF is no cause for concern at all. For example, you can get a high VIF by including products or powers from other variables in your regression, like  $x$  and  $x^2$ . If you have high VIFs for dummy variables representing nominal variables with three or more categories, those are usually not a problem.

### 1.5.2 Linear regression model:

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
- If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression ( $L^2$ -norm penalty) and lasso ( $L^1$ -norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

The linear regression model is-

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$$

Where,  $y_i$  = Dependent variable or response variable of the model

$x_i$  = Independent variable or explanatory variable of the model

$\beta_0$  = Intercept term of the model.

$\beta_i$  = The partial regression coefficients of the model.

$\varepsilon_i$  = Error term or Disturbance term of the model.



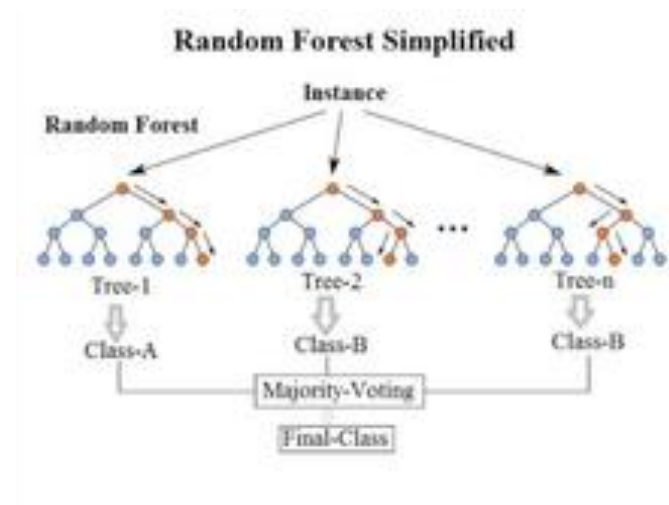
### 1.5.3 Random forest methodology:

Random forests are an ensemble learning technique that combines multiple **decision trees** into a forest or final model of decision trees that ultimately produces more accurate and stable predictions.

Random forests operate on the principle that a large number of trees operating as a committee (forming a strong learner) will outperform a single constituent tree (a weak learner). This is akin to the requirement in statistics to have a sample size large enough to be statistically relevant. Some individual trees may be wrong but as long as the individual trees are not making completely random predictions, their aggregate will form an approximation of the underlying data.

**Bagging** is the algorithmic technique used in the random forest scenario. This, we may recall, differs from the Gradient Boosting technique. Bagging trains individual decision trees on random samples of subsets of the dataset to reduce correlation. A benefit of bagging over boosting is that bagging can be performed in parallel while boosting is a sequential operation.

Individual decision trees are prone to overfitting and have a tendency to learn the noise in the dataset. Random Forests take an average of multiple trees -- so as long as the individual decision trees are not correlated, this strategy reduces overfitting and sensitivity to noise in the dataset.



### 1.5.4 ANOVA

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" within and between groups) used to analyze the differences among means. ANOVA was developed by the statistician Ronald Fisher. ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means

ANOVA is a form of statistical hypothesis testing heavily used in the analysis of experimental data. A test result (calculated from the null hypothesis and the sample) is called statistically significant if it is deemed unlikely to have occurred by chance, assuming the truth of the null hypothesis. A statistically significant result, when a probability (p-value) is less than a pre-specified threshold (significance level), justifies the rejection of the null hypothesis, but only if the a priori probability of the null hypothesis is not high.

"Classical" ANOVA for balanced data does three things at once:

1. As exploratory data analysis, an ANOVA employs an additive data decomposition, and its sums of squares indicate the variance of each component of the decomposition (or, equivalently, each set of terms of a linear model).
2. Comparisons of mean squares, along with an F-test allow testing of a nested sequence of models.
3. Closely related to the ANOVA is a linear model fit with coefficient estimates and standard errors.

#### Assumptions

The analysis of variance has been studied from several approaches, the most common of which uses a linear model that relates the response to the treatments and blocks. Note that the model is linear in parameters but may be nonlinear across factor levels. Interpretation is easy when data is balanced across factors but much deeper understanding is needed for unbalanced data.

The analysis of variance can be presented in terms of a linear model, which makes the following assumptions about the probability distribution of the responses:

- Independence of observations – this is an assumption of the model that simplifies the statistical analysis.
- Normality – the distributions of the residuals are normal.
- Equality (or "homogeneity") of variances, called homoscedasticity — the variance of data in groups should be the same.

The separate assumptions of the model imply that the errors are independently, identically, and normally distributed for fixed effects models, that is, that the errors ( $\varepsilon$ ) are independent and Normally distributed i.e.  $\varepsilon \sim N(0, \sigma^2)$

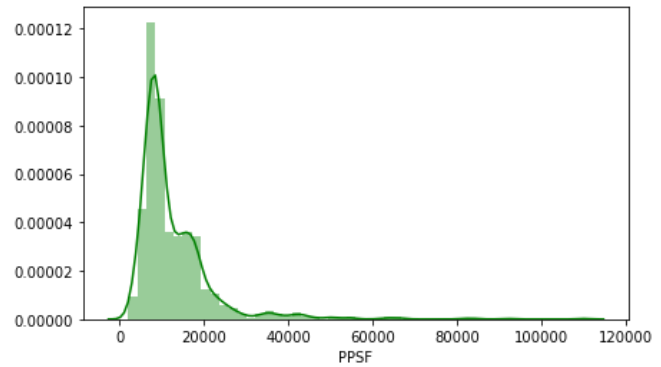
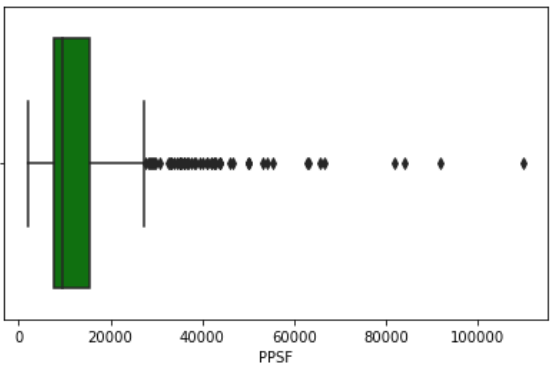
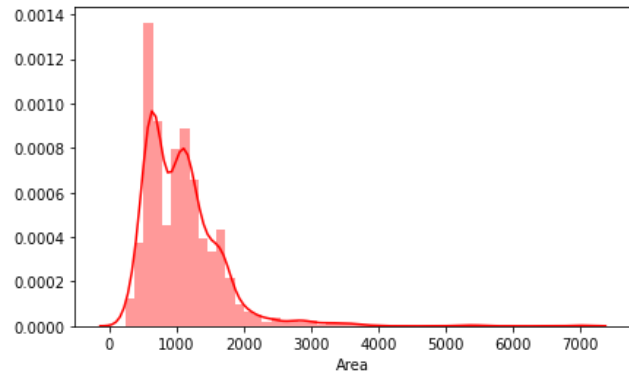
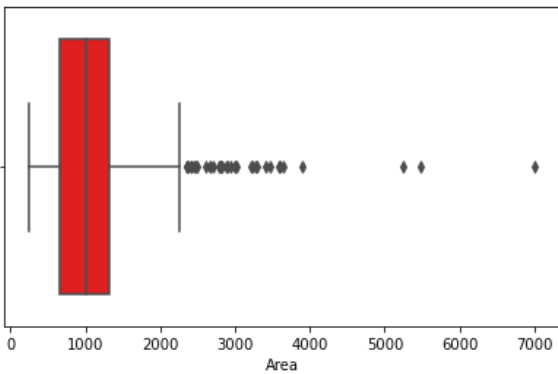
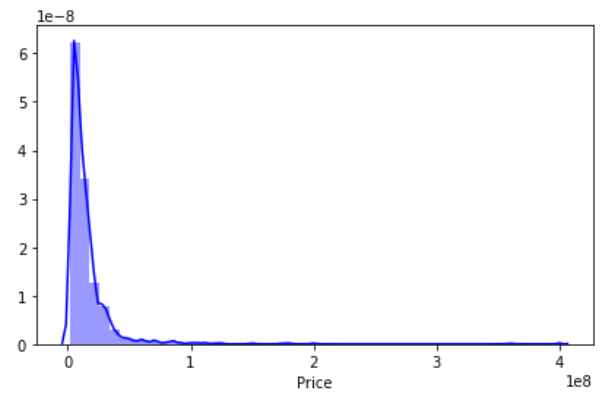
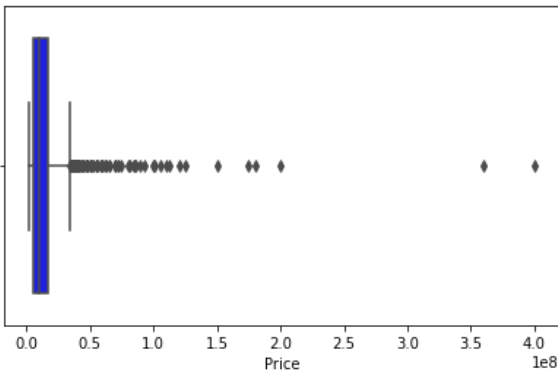
**1.5.5 Prediction mean**- It is the mean or average of all the prediction values of the model.

**1.5.6 Mean absolute error**- It is the difference between original value and prediction value of the model.

## 2. Chapter II

### 2.1. Graphical Representation

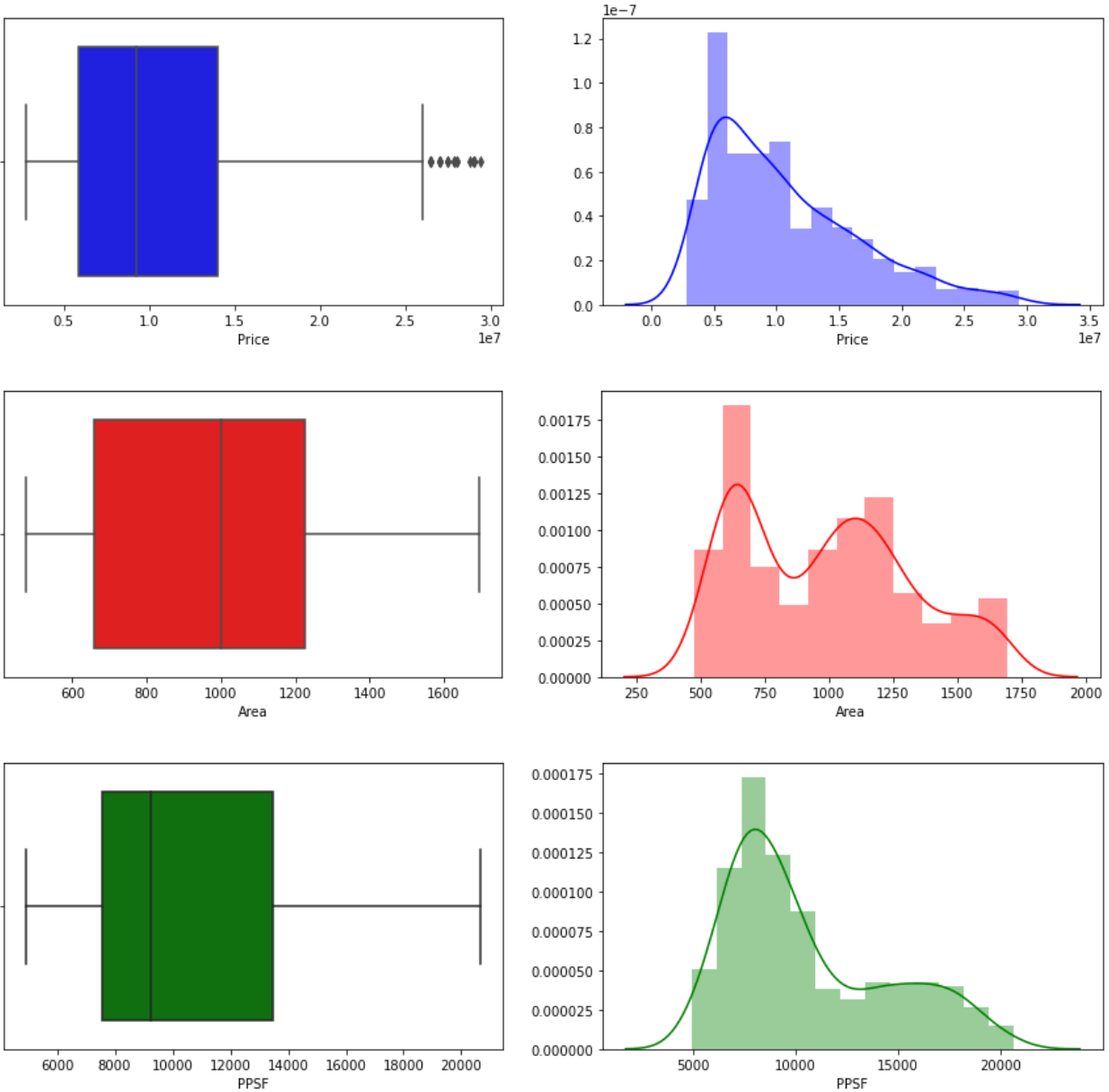
#### 2.1.1 Outlier present in the model



The above graphs shows that, the data is **heavily right-skewed** for Price and Area both. One would expect that PPSF would be normally distributed as it is a ratio, however it is not. This indicates either or both of the following:

1. There is much more skewness in Price compared to Area
2. The houses with high Prices are not necessarily the ones with the large Area, hence PPSF is not able to normalize such cases

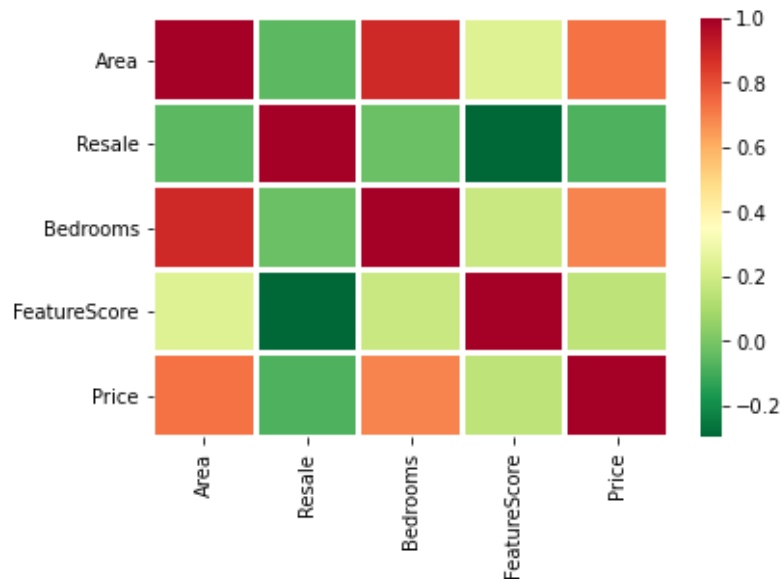
### **2.1.2 After outlier is removed**



As the data is heavily right skewed, we remove a higher percentile (10%) from the top versus bottom(5%), the outliers has been removed which can be seen in the above plots.

### 2.1.3 Heatmap for understanding correlation

<AxesSubplot:>



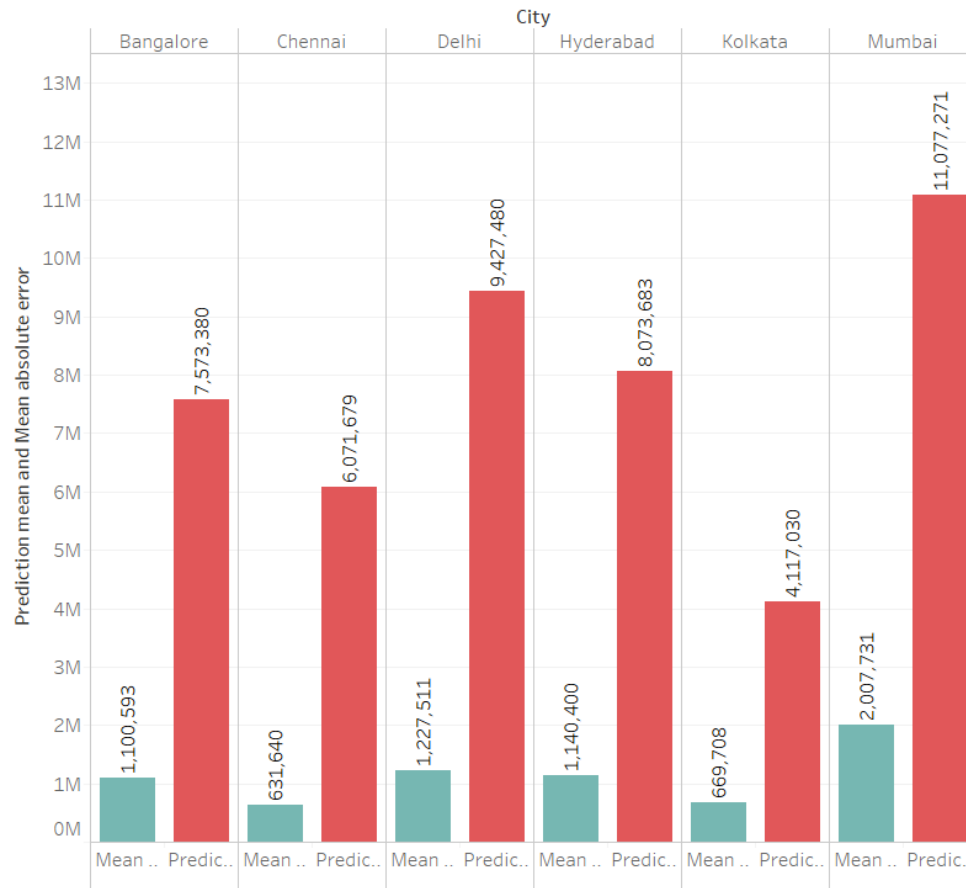
Interpretation:

Each of the square shows the correlation between the variables on each axis. Correlation ranges from -1 to +1, values closer to zero means there is no linear trend between the two variables. The value closes to 1 means it is positively correlated, i.e. as one increases so does the other and stronger will be there relation.

A correlation closer to -1 is similar, but instead of both increasing one variable will decrease as the other increases. The diagonal elements are all 1 because those squares are correlating with each other (i.e. perfect correlation). For the rest, larger the number, darker the colour and higher will be the correlation between two variables. The plot is also symmetrical about the diagonal since the same two variables are being paired together in those squares.

### 2.1.4 Prediction mean and mean absolute error of linear regression model:

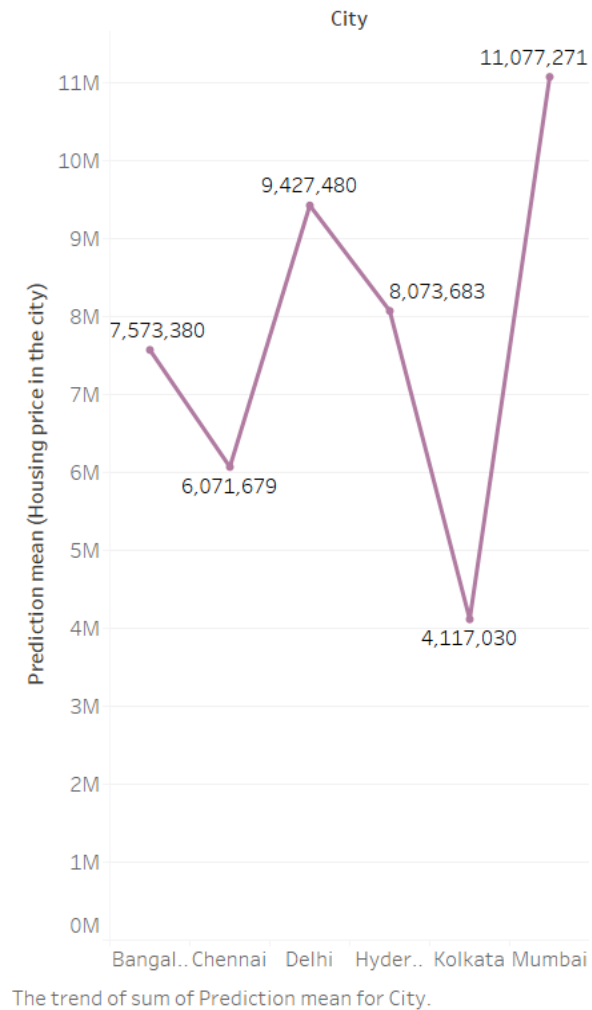
#### Linear Regression Model



Mean absolute error and Prediction mean for each City. Color shows details about Mean absolute error and Prediction mean.

**Measure Names**  
■ Mean absolute error  
■ Prediction mean

## Linear Regression Model

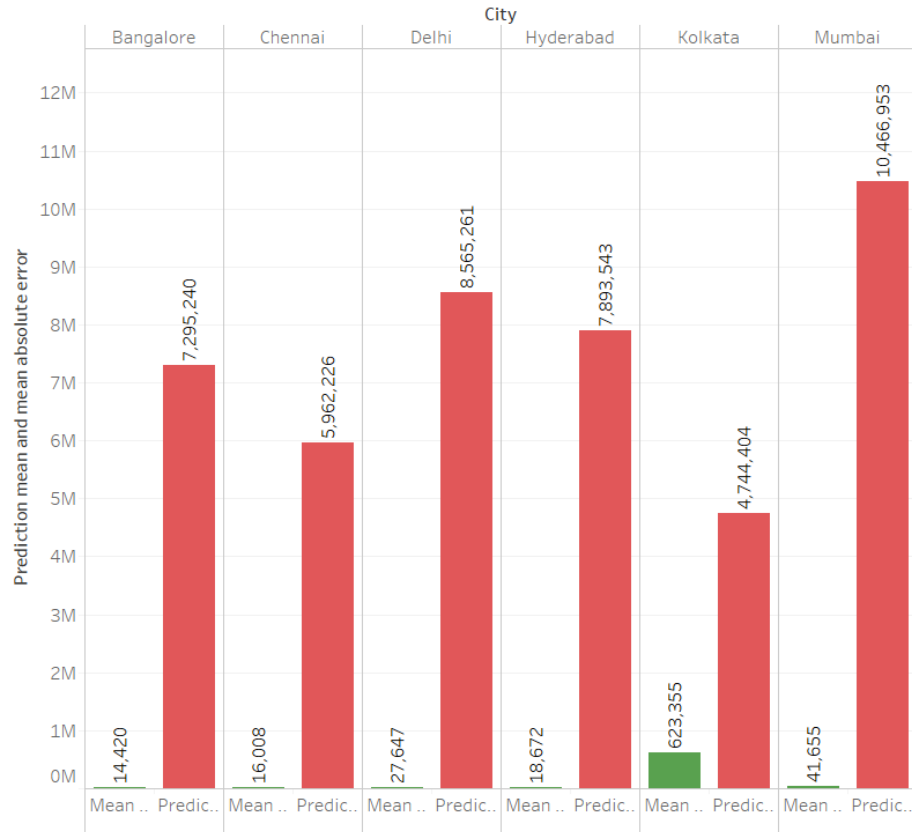


From the above graph it is seen that the highest prediction mean is Rs 11,077,271 million and the lowest prediction mean is Rs 4,117,030, so it is clear that Mumbai is costliest city and Kolkata is the cheapest city to buy a house and from the mean absolute error it is seen that Mumbai has the most error in the model and Chennai has the least error in the model.



### 2.1.5 Prediction mean and mean absolute error of Random forest model:

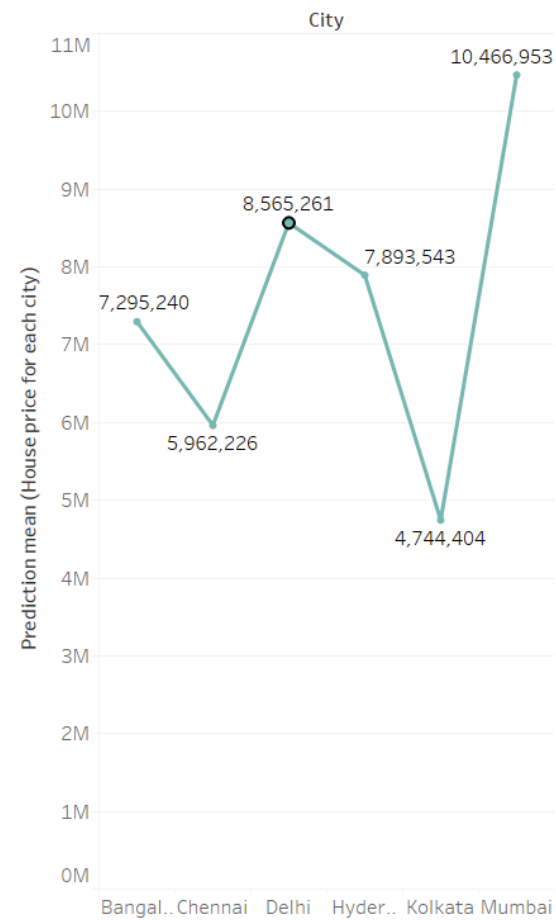
#### Random Forest Model



Mean absolute error and Prediction mean for each City. Color shows details about Mean absolute error and Prediction mean.

**Measure Names**  
■ Mean absolute error  
■ Prediction mean

### Random Forest Model



The trend of sum of Prediction mean for City.

From the above graph it is seen that the highest prediction mean is Rs 10,466,953 million and the lowest prediction mean is Rs 4,744,404, so it is clear that Mumbai is costliest city and Kolkata is the cheapest city to buy a house and from the mean absolute error it is seen that kolkata has the most error in the model and Chennai has the least error in the model.

## **2.2 Analysis of the dataset**

To analyze the dataset, the following methods has been carried out:

- (1) Exploratory Data analysis(EDA)
- (2) Feature Engineering
- (3) Feature selection
- (4) Model creation and deployment, Hyper parameter tuning, Prediction
- (5) Using the SPSS software, Draw the ANOVA table for further interpretation about the significance of the model.
- (6) Using the Tableau software, Carry out the necessary graph and plots for interpretation about the prediction mean and mean absolute error.

Tools that are used to analyze the dataset are:

- (1) Jupyter Notebook
- (2) Python
- (3) Pandas
- (4) NumPy
- (5) Statsmodels
- (6) Matplotlib
- (7) Seaborn
- (8) Scikit learn
- (9) Random forest methodology
- (10) Spss
- (11) Tableau

### 2.2.1 Exploratory data analysis (EDA)

Step-1 : Importing all the libraries such as pandas, Numpy, statsmodels, matplotlib, seaborn, sklearn etc.

Step-2: Import the dataset using pandas library.

Price	Area	Location	No. of Bedrooms	Resale	MaintenanceStaff	Gymnasium
0	4850000	720	Kharghar	1	1	1
1	4500000	600	Kharghar	1	1	1
2	6700000	650	Kharghar	1	1	1
3	4500000	650	Kharghar	1	1	1
4	5000000	665	Kharghar	1	1	1
...	...	...	...	...	...	...
7714	14500000	1180	Mira Road East	2	0	9
7715	14500000	530	Naigaon East	1	1	9
7716	4100000	700	Shirgaon	1	0	9
7717	2750000	995	Mira Road East	2	0	9
7718	2750000	1020	Mira Road East	2	0	9
7719 rows × 40 columns						

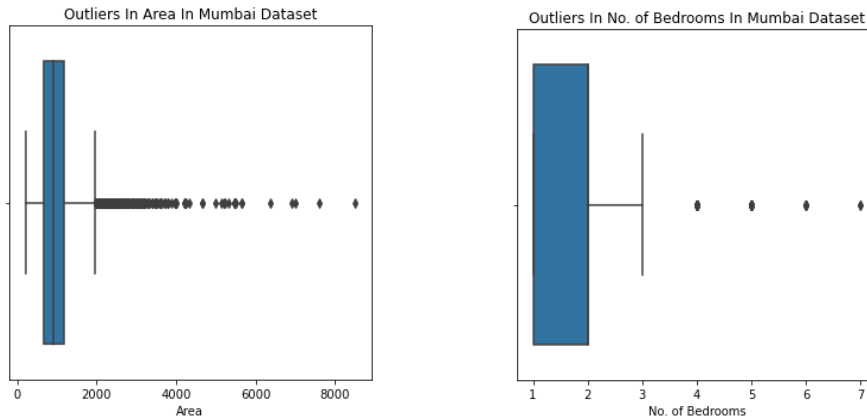
This Table is the short version of the original Table.

Step-3: Calculating all the descriptive statistics about the dataset.

Price	Area	No. of Bedrooms	Resale	MaintenanceStaff	Gymnasium
count	7.72E+03	7719	7719	7719	7719
mean	1.51E+07	998.40925	1.913331	0.647105	7.498899
std	2.05E+07	550.967809	0.855376	0.477901	3.197923
min	2.00E+06	200	1	0	0
25%	5.30E+06	650	1	0	9
50%	9.50E+06	900	2	1	9
75%	1.70E+07	1177	2	1	9
max	4.20E+08	8511	7	1	9

This Table is short version of the original Table.

Step-4: Using the visualization tool i.e. matplotlib and seaborn to visualize the outliers in the data-set.



From the above graph we can see that the dataset is heavily skewed for both Area and No. of bedrooms.

### **2.2.2 Feature Engineering**

Step-1: Handling the missing value and Duplicate value.

Step-2: A key metric in the housing industry, at least in India, is "Price Per Square Feet (Price/Area)", expressed as Rupees Per Square Feet. Almost everyone uses this metric to compare the prices as it eliminates the impact of house size and provides an easy comparison. Let us add a column that captures Price Per Square Feet. We will call this "PPSF" in short

Step-3: Handling the outlier, using the distribution plot to check outliers in the dataset and to remove the outliers from the dataset, remove a higher percentile (10%) from the top versus bottom(5%) by creating a filter which will remove the outliers from the dataset.

Step-4: Checking the features that is available and also clean up the names so that they do not cause any problems later, by renaming some of the names. Now calculating the feature score for each house based on the amenities.

Price	Area	Location	Bedrooms	Resale	PPSF	Feature score
0	4850000	720	Kharghar	1	6736.11111	10
1	4500000	600	Kharghar	1	7500	25
2	6700000	650	Kharghar	1	10307.6923	30
3 rows × 42 columns						

This table is the short version of the original Table.

Now, Reducing the number of features on which to run the regression model. These are:

1. Area (numeric)
2. Location (string).
3. Bedrooms (numeric)
4. Feature Score (numeric)
5. Resale (boolean)

Step-5: Modeling the Location as a numeric feature, Location is one of the most important indicator of house price. As a bit of feature engineering, calculate the 'Location Premium' for every Location. This is nothing but the PPSF for every location divided by the minimum PPSF. This sets the cheapest location as the base location (with a score of 1) and every other location has a premium as a multiple of that base location

Location	PPSF	Location premium
Vivek Vidyalaya Marg	20000	4.060606
no 9	19259.2593	3.910213
Samata Nagar Thakur Village	19259.2593	3.910213

<b>Mahatma Gandhi Road</b>	19200	3.898182
<b>raheja vihar</b>	18351.7094	3.725953
...	...	...
<b>Ambernath East</b>	5306.84932	1.077451
<b>Vasai east</b>	5197.13262	1.055175
<b>Taloja</b>	5160.79494	1.047798
<b>Badlapur West</b>	5149.65035	1.045535
<b>Koprol</b>	4925.37313	1
171 rows × 2 columns		

Now merge the location pivot with the Mumbai dataset on 'Location' column and As a final step calculate the **log** of LocationPremium.

<b>Price</b>	<b>Area</b>	<b>Location</b>	<b>Bedrooms</b>	<b>Resale</b>	<b>PPSF</b>	<b>FeatureScore</b>	<b>LocationPremium</b>	<b>LogPremium</b>
<b>0</b>	4850000	720	Kharghar	1	6736.111	10	1.73745	0.552419
<b>1</b>	4500000	600	Kharghar	1	7500	25	1.73745	0.552419
<b>2</b>	6700000	650	Kharghar	1	10307.69	30	1.73745	0.552419
3 rows × 44 columns								

Step-6: Now Convert categorical variable into dummy/indicator variables.

### **2.2.3 Feature selection**

Handling the Multicollinearity

Step-1: Calculate variance inflating factor (VIF)

Area	10.180599
Bedrooms	9.310975
Resale	3.374238
Gymnasium	7.029648
SwimmingPool	6.592871
...	
Location_no 9	1.407113
Location_raheja vihar	1.920898
Location_taloja panchanand	1.751822
Location_thakur village kandivali east	1.833091
Location_vasant vihar thane west	1.399950

Length: 208, dtype: float64

**A rule of thumb** for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

Step-2: Now, Remove Features that are not important or redundant

### **2.2.4 Model creation and deployment, Hyper parameter tuning and Prediction**

Step-1: Split the Dataset into Test and Train dataset.

Step-2: Define the parameters for hyper parameter tuning, which is an important part before building a model, which sets constraints on the parameter to obtain accuracy of the model.

Step-3: Now fit the linear regression model using ordinary least square method(OLS), and then formulate the summary from the model, which will give all the necessary values such R-square, adjusted R-square, F-statistic, p-value etc.



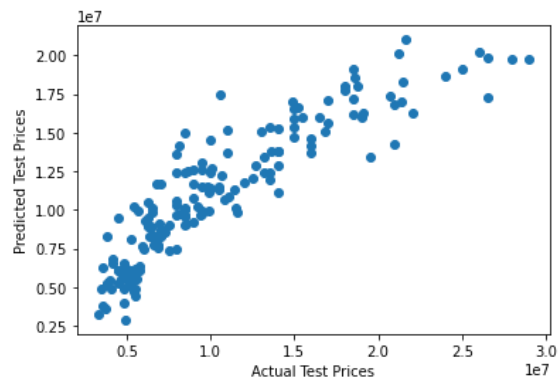
# OLS Regression Results

Dep. Variable:		Price		R-squared (uncentered):		0.953	
Model:		OLS		Adj. R-squared (uncentered):		0.953	
Method:		Least Squares		F-statistic:		2720.	
Date:		Thu, 03 Jun 2021		Prob (F-statistic):		0.00	
Time:		21:22:07		Log-Likelihood:		-10906.	
No. Observations:		673		AIC:		2.182e+04	
Df Residuals:		668		BIC:		2.184e+04	
Df Model:		5					
Covariance Type:		nonrobust					
		coef	std err	t	P> t	[0.025	0.975]
Area	7043.8083	666.201	10.573	0.000		5735.707	8351.909
FeatureScore	-5.342e+04	7900.886	-6.761	0.000		-6.89e+04	-3.79e+04
Resale	-3.027e+06	2.19e+05	-13.822	0.000		-3.46e+06	-2.6e+06
Bedrooms	8.726e+05	3.23e+05	2.701	0.007		2.38e+05	1.51e+06
LogPremium	8.599e+06	3.02e+05	28.502	0.000		8.01e+06	9.19e+06
Omnibus:	87.334	Durbin-Watson:	2.042				
Prob(Omnibus):	0.000	Jarque-Bera (JB):	148.759				
Skew:	0.821	Prob(JB):	4.98e-33				
Kurtosis:	4.616	Cond. No.	3.44e+03				

The above table is a summary table, which obtained after fitting the linear regression model using ordinary least square method.

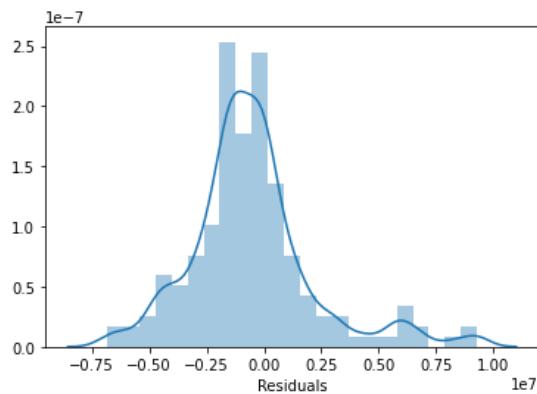
Step-4: Now checking the model performance on the test data.

```
Text(0, 0.5, 'Predicted Test Prices')
```



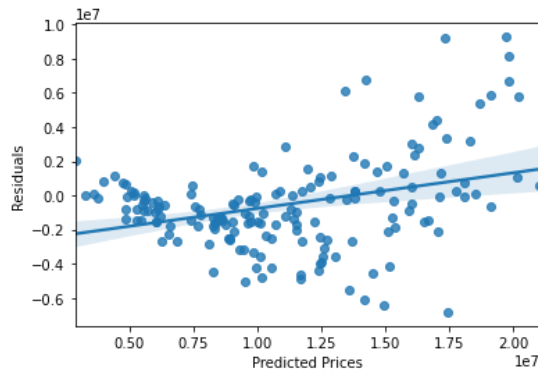
Although there seems to be a linear correlation between test set Prices and predicted Prices, there is also a fair bit of scatter. Now, Checking the residuals plot.

```
Text(0.5, 0, 'Residuals')
```



The residuals seem to be mostly normally distributed, except for a slight skew to the right. Residuals bordering on close to Rs 10 million are problematic. Now checking the plot of residuals against predicted prices.

```
Text(0, 0.5, 'Residuals')
```



Although there seems to be a linear correlation between predicted prices and the residuals, there is also a fair bit of scatter on both side of the line.

Step-5: Now, calculate the mean absolute error and the prediction mean for the linear regression model.

Step-6: Now, building a predictive model using Random forest methodology.

Step-7: Now, calculate the prediction mean, mean absolute error and evaluate the model performance.

```
array([ 8492593.42903682,  8039507.20554062,  9104596.06844743,
        11157357.66083356,  9426339.52016068,  7140142.88633867,
        13505996.98754307, 11008648.35160509,  3683770.74047355,
        5495447.97209172, 11611938.70428324, 19033307.00568063,
        16522965.5424884 ,  9028912.39139795,  5989966.11618862,
        4489513.27266645, 11113024.24433633, 12484968.99326487,
        4715465.77146974,  7508371.72578362,  5893331.46702895,
        5991009.8953643 ,  8492716.00116445,  9986735.75607399,
        13972839.02712946,  4917858.99322695,  6502032.79487798,
        4810540.35937733, 11978006.75357558, 13011739.3649576 ,

        6197160.45407657,  4297210.01461291,  9986116.70845494,
        20382377.15208115, 13531372.22524481,  4201755.50497324,
        6864173.52742656,  4299806.4121696 , 11003688.06810892,
        22533791.2646897 ,  6676803.76385585, 11539605.74699865,
        15013205.60716695,  3988920.50791917,  9835417.26804649,
        16562561.7769492 , 20386441.13832619,  8013728.27442541,
        7553636.30330291, 23175491.64640615,  9992177.67908153,
        7500507.64207565,  6000066.43064537, 10493736.43786405,
        3206070.50230652,  8489988.16970772,  4516535.9747848 ,
        8037594.35261805,  7006320.1325168 ,  8506553.25602668,
        4989588.7258773 ,  8023304.51016884,  9011123.46799496,
        11028845.3225226 ])
```

This is the short version of the original data.

Code:

```
from sklearn.metrics import r2_score  
  
r2_score(Y_train, random_forest_model.predict(X_train))
```

Output:

```
0.9997312058994824
```

The above value shows the model performance i.e. R-square value of the model.

### 3. Chapter III

#### 3.1 Result of the analysis

After completing the analysis of the Datasets for six cities, the following results have been obtained from the analysis which are:

##### 3.1.1 Variance inflating factor (VIF) for each cities:

###### VIF-table for Mumbai dataset:

Area	10.180599
Bedrooms	9.310975
Resale	3.374238
Gymnasium	7.029648
SwimmingPool	6.592871
...	
Location_no 9	1.407113
Location_raheja vihar	1.920898
Location_taloja panchanand	1.751822
Location_thakur village kandivali east	1.833091
Location_vasant viharthane west	1.399950
Length: 208, dtype: float64	

###### VIF-table for Bangalore dataset:

Area	4.218759
Bedrooms	3.180027
Resale	3.209462
Gymnasium	19.860166
SwimmingPool	26.932949
...	
Location_Varthur	inf
Location_Vidyaranyapura	inf
Location_Whitefield Hope Farm Junction	inf
Location_Yelahanka	inf
Location_Yerthiganahalli	inf
Length: 124, dtype: float64	

#### VIF-table for Delhi dataset:

Area	8.768304
Bedrooms	2.766580
Resale	4.477686
Gymnasium	17.260563
SwimmingPool	6.877760
...	
Location_Vasant Kunj	5.575837
Location_West Sagarpur	1.125693
Location_greater kailash Enclave 1	1.084679
Location_mayur vihar phase 1	1.135669
Location_nawada	3.946635

Length: 122, dtype: float64

#### VIF-table for Chennai dataset:

Area	5.766694
Bedrooms	3.579422
Resale	1.751947
Gymnasium	11.878468
SwimmingPool	6.694276
...	
Location_Vellakkal	2.429954
Location_Vengaivasal	1.331238
Location_Villivakkam	1.572014
Location_West Tambaram	2.174783
Location_tambaram west	8.072015

Length: 147, dtype: float64

#### VIF-table for Kolkata dataset:

Area	inf
Bedrooms	inf
Resale	inf
MaintenanceStaff	inf
Gymnasium	inf
...	
Location_Sonarpur	inf
Location_Tollygunge	inf
Location_Ultadanga	inf
Location_Uttarpara Kotrung	inf
Location_south dum dum	inf

Length: 64, dtype: float64

### VIF-table for Hyderabad dataset:

```
Area                3.409339
Bedrooms            3.028996
Resale              2.186807
Gymnasium           8.995529
SwimmingPool        7.915541

...
Location_muthangi   2.044062
Location_new nallakunta 4.075025
Location_nizampet road 3.069351
Location_raidurgam  2.315374
Location_west venkatapuram 2.018404
Length: 232, dtype: float64
```

A **rule of thumb** for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

From the above tables, it is clear that each of the six dataset contains some sort of multi-co-linearity, which varies with different dataset. To reduce multi-co-linearity certain remedies can be use such as dropping the variables, but dropping the significant variable may reduce the value of R-square which may not help as a remedy, then also reduce the multi-co-linearity by dropping some of the insignificant variable which are not significant for the model to be fitted.

### **3.1.2 Summary table of the models, ANOVA table of the models and model performances for each cities:**

#### **(i) Summary table for Mumbai dataset:**

OLS Regression Results			
<b>Dep. Variable:</b>	Price	<b>R-squared (uncentered):</b>	0.953
<b>Model:</b>	OLS	<b>Adj. R-squared (uncentered):</b>	0.953
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	2720.
<b>Date:</b>	Thu, 03 Jun 2021	<b>Prob (F-statistic):</b>	0.00
<b>Time:</b>	21:22:07	<b>Log-Likelihood:</b>	-10906.
<b>No. Observations:</b>	673	<b>AIC:</b>	2.182e+04
<b>Df Residuals:</b>	668	<b>BIC:</b>	2.184e+04

<b>Df Model:</b>	5					
<b>Covariance Type:</b>	nonrobust					
	<b>coef</b>	<b>std err</b>	<b>t</b>	<b>P&gt; t </b>	<b>[0.025</b>	<b>0.975]</b>
<b>Area</b>	7043.8083	666.201	10.573	0.000	5735.707	8351.909
<b>FeatureScore</b>	-5.342e+04	7900.886	-6.761	0.000	-6.89e+04	-3.79e+04
<b>Resale</b>	-3.027e+06	2.19e+05	-13.822	0.000	-3.46e+06	-2.6e+06
<b>Bedrooms</b>	8.726e+05	3.23e+05	2.701	0.007	2.38e+05	1.51e+06
<b>LogPremium</b>	8.599e+06	3.02e+05	28.502	0.000	8.01e+06	9.19e+06
<b>Omnibus:</b>	87.334	<b>Durbin-Watson:</b>	2.042			
<b>Prob(Omnibus):</b>	0.000	<b>Jarque-Bera (JB):</b>	148.759			
<b>Skew:</b>	0.821	<b>Prob(JB):</b>	4.98e-33			
<b>Kurtosis:</b>	4.616	<b>Cond. No.</b>	3.44e+03			

The required linear regression model of the Mumbai city dataset is:

$$Y[\text{Price}] = 7043.8083[\text{Area}_i] + (-5.342e+04)[\text{FeatureScore}_i] + (-3.027e+06)[\text{Resale}_i] + (8.726e+05)[\text{Bedrooms}_i] + (8.599e+06)[\text{LogPremium}_i]$$

Now, to test if the linear regression model is significant i.e. the model fits the data. One approach is compute the coefficient of determination R-square and other approach is to use the technique of ANOVA, so draw the ANOVA table using the SPSS software and interpret the results



**Null hypothesis for Mumbai Dataset:**

$H_0$ : Area = FeatureScore = Resale = Bedrooms = LogPremium = 0

i.e. None of the explanatory variable has significant contribution to the Dependent variable.

Against alternative hypothesis that:

$H_1$ : Area  $\neq$  FeatureScore  $\neq$  Resale  $\neq$  Bedrooms  $\neq$  LogPremium  $\neq$  0

i.e. There is at-least one explanatory variable, which has significant contribution to the dependent variable.

**ANOVA table for Mumbai city**

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2516596689005	5	5033193378011	1129.149	.000 <sup>b</sup>
		5340.000		068.000		
	Residual	3726477832295	836	4457509368774		
		340.000		.330		
	Total	2889244472235	841			
		0680.000				

a. Dependent Variable: Price

b. Predictors: (Constant), LogPremium, FeatureScore, Bedrooms, Resale, Area

Interpretation: The p-value of the model is less than 0.05, therefore reject the null hypothesis and conclude that at-least one of the explanatory variable, which has significant contribution to the dependent variable. Therefore the model significantly fits the data

From the summary table, it is clear that the coefficient of determination of the model is 0.953 i.e. around 95% which imply that 95% of the variable in Y is explained by the regression model.

Now, the Random forest methodology has been use to build a predictive model. After building the predictive model, check the performance of the model by evaluating the R-square-score which is 0.9997294868500382 for the Mumbai dataset, which is better than linear regression model.

**(ii) Summary table for Bangalore dataset:**

OLS Regression Results

<b>Dep. Variable:</b>	Price	<b>R-squared (uncentered):</b>	0.968
<b>Model:</b>	OLS	<b>Adj. R-squared (uncentered):</b>	0.968
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	5988.
<b>Date:</b>	Fri, 04 Jun 2021	<b>Prob (F-statistic):</b>	0.00
<b>Time:</b>	11:51:49	<b>Log-Likelihood:</b>	-15398.
<b>No. Observations:</b>	989	<b>AIC:</b>	3.081e+04
<b>Df Residuals:</b>	984	<b>BIC:</b>	3.083e+04
<b>Df Model:</b>	5		
<b>Covariance Type:</b>	nonrobust		
	<b>coef</b>	<b>std err</b>	<b>t</b> <b>P&gt; t </b> <b>[0.025</b> <b>0.975]</b>
<b>Area</b>	4572.9604	234.934	19.465 0.000 4111.932 5033.989
<b>FeatureScore</b>	346.1783	4235.076	0.082 0.935 -7964.640 8656.996
<b>Resale</b>	-7.743e+05	1.68e+05	-4.602 0.000 -1.1e+06 -4.44e+05
<b>Bedrooms</b>	-5.755e+05	1.12e+05	-5.157 0.000 -7.94e+05 -3.57e+05
<b>LogPremium</b>	5.77e+06	2.82e+05	20.433 0.000 5.22e+06 6.32e+06
<b>Omnibus:</b>	123.691	<b>Durbin-Watson:</b>	1.981
<b>Prob(Omnibus):</b>	0.000	<b>Jarque-Bera (JB):</b>	184.989
<b>Skew:</b>	0.877	<b>Prob(JB):</b>	6.76e-41
<b>Kurtosis:</b>	4.190	<b>Cond. No.</b>	8.89e+03

The required linear regression model of the Bangalore city dataset is:

$$Y[\text{Price}] = 4572.9604 [\text{Area}_i] + (346.1783) [\text{FeatureScore}_i] + (-7.743e+05) [\text{Resale}_i] + (-5.755e+05) [\text{Bedrooms}_i] + (5.77e+06) [\text{LogPremium}_i]$$

Now, to test if the linear regression model is significant i.e. the model fits the data. One approach is compute the coefficient of determination R-square and other approach is to use the technique of ANOVA, so draw the ANOVA table using the SPSS software and interpret the results

**Null hypothesis for Bangalore Dataset:**

$H_0$ : Area = FeatureScore = Resale = Bedrooms = LogPremium = 0

i.e. None of the explanatory variable has significant contribution to the Dependent variable.

Against the alternative hypothesis that:

$H_1$ : Area  $\neq$  FeatureScore  $\neq$  Resale  $\neq$  Bedrooms  $\neq$  LogPremium  $\neq$  0

i.e. There is at-least one explanatory variable, which has significant contribution to the dependent variable.

**ANOVA table for Bangalore city**

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6956212971484	5	1391242594296	1114.133	.000 <sup>b</sup>
		335.000		867.000		
	Residual	1537176617036	1231	1248721865992		
		598.000		.362		
	Total	8493389588520	1236			
		933.000				

a. Dependent Variable: Price

b. Predictors: (Constant), LogPremium, Resale, Bedrooms, FeatureScore, Area

Interpretation: The p-value of the model is less than 0.05, therefore reject the null hypothesis and conclude that at-least one of the explanatory variable, which has significant contribution to the dependent variable. Therefore the model significantly fits the data

From the summary table, it is clear that the coefficient of determination of the model is 0.968 i.e. around 96% which imply that 96% of the variable in Y is explained by the regression model.

Now, the Random forest methodology has been use to build a predictive model. After building the predictive model, check the performance of the model by evaluating the R-square-score which is 0.9996961469785727 for the Bangalore dataset, which is better than linear regression model.

**(iii) Summary table for Delhi dataset:**

OLS Regression Results							
Dep. Variable:		Price		R-squared (uncentered):		0.978	
Model:		OLS		Adj. R-squared (uncentered):		0.977	
Method:		Least Squares		F-statistic:		7987.	
Date:		Fri, 04 Jun 2021		Prob (F-statistic):		0.00	
Time:		14:52:43		Log-Likelihood:		-14404.	
No. Observations:		920		AIC:		2.882e+04	
Df Residuals:		915		BIC:		2.884e+04	
Df Model:		5					
Covariance Type:		nonrobust					
		coef	std err	t	P> t	[0.025	0.975]
Area		8695.0419	218.830	39.734	0.000	8265.575	9124.509
FeatureScore		1.873e+04	6867.942	2.727	0.007	5250.348	3.22e+04
Resale		1.518e+05	1.27e+05	1.198	0.231	-9.69e+04	4e+05
Bedrooms		-1.31e+06	6.37e+04	-20.578	0.000	-1.44e+06	-1.19e+06
LogPremium		4.745e+06	2.72e+05	17.419	0.000	4.21e+06	5.28e+06
Omnibus:	52.986	Durbin-Watson:		1.925			

**Prob(Omnibus):** 0.000    **Jarque-Bera (JB):** 169.199

**Skew:** 0.185    **Prob(JB):** 1.82e-37

**Kurtosis:** 5.068    **Cond. No.** 6.74e+03

The required linear regression model of the Delhi city dataset is:

$$Y[\text{Price}] = 8695.0419[\text{Area}_i] + (1.873e+04) [\text{FeatureScore}_i] + (1.518e+05) [\text{Resale}_i] + (-1.31e+06) [\text{Bedrooms}_i] + (4.745e+06) [\text{LogPremium}_i]$$

Now, to test if the linear regression model is significant i.e. the model fits the data. One approach is compute the coefficient of determination R-square and other approach is to use the technique of ANOVA, so we draw the ANOVA table using the SPSS software and interpret the results

**Null hypothesis for Delhi dataset:**

$H_0$ : Area = FeatureScore = Resale = Bedrooms = LogPremium = 0

i.e. None of the explanatory variable has significant contribution to the Dependent variable.

Against the alternative hypothesis that:

$H_1$ : Area  $\neq$  FeatureScore  $\neq$  Resale  $\neq$  Bedrooms  $\neq$  LogPremium  $\neq$  0

i.e. There is at-least one explanatory variable, which has significant contribution to the dependent variable.

### ANOVA table for Delhi city

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2999632395750 6912.000	5	5999264791501 382.000	3189.928	.000 <sup>b</sup>
	Residual	2151509219217 002.000	1144	1880689876937 .939		
	Total	3214783317672 3912.000	1149			

a. Dependent Variable: Price

b. Predictors: (Constant), LogPremium, Bedrooms, Resale, FeatureScore, Area

Interpretation: The p-value of the model is less than 0.05, therefore reject the null hypothesis and conclude that at-least one of the explanatory variable, which has significant contribution to the dependent variable. Therefore the model significantly fits the data.

From the summary table, it is clear that the coefficient of determination of the model is 0.968 i.e. around 96% which imply that 96% of the variable in Y is explained by the regression model.

Now the Random forest methodology has been used, to build a predictive model. After building the predictive model, check the performance of the model by evaluating the R-square-score which is 0.999937818688379 for the Delhi dataset, which is better than linear regression model.

### (iv) Summary table for Chennai dataset:

OLS Regression Results			
<b>Dep. Variable:</b>	Price	<b>R-squared (uncentered):</b>	0.984
<b>Model:</b>	OLS	<b>Adj. R-squared (uncentered):</b>	0.984
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	1.304e+04
<b>Date:</b>	Fri, 04 Jun 2021	<b>Prob (F-statistic):</b>	0.00
<b>Time:</b>	21:10:32	<b>Log-Likelihood:</b>	-15544.
<b>No. Observations:</b>	1035	<b>AIC:</b>	3.110e+04

Df Residuals:		1030		BIC:		3.112e+04	
Df Model:		5					
Covariance Type:		nonrobust					
		coef	std err	t	P> t	[0.025	0.975]
Area	4691.6198	164.685	28.488	0.000	4368.463	5014.776	
FeatureScore	1.77e+04	2231.535	7.933	0.000	1.33e+04	2.21e+04	
Resale	-1.377e+05	8.66e+04	-1.590	0.112	-3.08e+05	3.22e+04	
Bedrooms	-3.35e+05	6.88e+04	-4.869	0.000	-4.7e+05	-2e+05	
LogPremium	5.222e+06	1.83e+05	28.499	0.000	4.86e+06	5.58e+06	
Omnibus:		191.518	Durbin-Watson:		1.874		
Prob(Omnibus):		0.000	Jarque-Bera (JB):		443.865		
Skew:		1.009	Prob(JB):		4.13e-97		
Kurtosis:		5.494	Cond. No.		8.30e+03		

The required linear regression model of the Chennai city dataset is:

$$Y[\text{Price}] = 4691.6198 [Area_i] + (1.77e+04) [FeatureScore_i] + (-1.377e+05)[Resale_i] + (-3.35e+05) [Bedrooms_i] + (5.222e+06) [LogPremium_i]$$

Now to test if the linear regression model is significant i.e. the model fits the data. One approach is compute the coefficient of determination R-square and other approach is to use the technique of ANOVA, so we draw the ANOVA table using the SPSS software and interpret the results

**Null hypothesis of Chennai dataset:**

$H_0$ : Area = FeatureScore = Resale = Bedrooms = LogPremium = 0

i.e. None of the explanatory variable has significant contribution to the Dependent variable.

Against the alternative hypothesis that:

$H_1$ : Area  $\neq$  FeatureScore  $\neq$  Resale  $\neq$  Bedrooms  $\neq$  LogPremium  $\neq$  0

i.e. There is at-least one explanatory variable, which has significant contribution to the dependent variable.

**ANOVA table for Chennai city**

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5409195416422 680.000	5	1081839083284 536.000	2218.114	.000 <sup>b</sup>
	Residual	6281952162139 32.400	1288	487729205135. 041		
	Total	6037390632636 612.000	1293			

a. Dependent Variable: Price

b. Predictors: (Constant), LogPremium, Resale, Bedrooms, FeatureScore, Area

Interpretation: The p-value of the model is less than 0.05, therefore reject the null hypothesis and conclude that at-least one of the explanatory variable, which has significant contribution to the dependent variable. Therefore the model significantly fits the data.

From the summary table, it is clear that the coefficient of determination of the model is 0.984 i.e. around 98% which imply that 98% of the variable in Y is explained by the regression model.

Now the Random forest methodology has been used, to build a predictive model. After building the predictive model, check the performance of the model by evaluating the R-square-score which is 0.9992769442781928 for the Chennai dataset, which is better than linear regression model.



**(v) Summary table for Kolkata dataset:**

OLS Regression Results							
Dep. Variable:		Price		R-squared (uncentered):		0.984	
Model:		OLS		Adj. R-squared (uncentered):		0.981	
Method:		Least Squares		F-statistic:		301.6	
Date:		Thu, 03 Jun 2021		Prob (F-statistic):		7.75e-21	
Time:		21:08:39		Log-Likelihood:		-430.00	
No. Observations:		29		AIC:		870.0	
Df Residuals:		24		BIC:		876.8	
Df Model:		5					
Covariance Type:		nonrobust					
		coef	std err	t	P> t	[0.025	0.975]
Area		3730.7291	876.385	4.257	0.000	1921.959	5539.499
FeatureScore		1.149e+04	1.13e+04	1.017	0.319	-1.18e+04	3.48e+04
Resale		1.145e+06	3.64e+05	3.144	0.004	3.93e+05	1.9e+06
Bedrooms		-5.748e+05	3.23e+05	-1.777	0.088	-1.24e+06	9.27e+04
LogPremium		3.469e+06	5.42e+05	6.396	0.000	2.35e+06	4.59e+06
Omnibus:		0.882	Durbin-Watson:		1.674		
Prob(Omnibus):		0.643	Jarque-Bera (JB):		0.194		
Skew:		-0.146	Prob(JB):		0.908		
Kurtosis:		3.274	Cond. No.		4.69e+03		

The required linear regression model of the Kolkata city dataset is:

$$Y[\text{Price}] = 3730.7291 [\text{Area}_i] + (1.149\text{e}+04) [\text{FeatureScore}_i] + (1.145\text{e}+06) [\text{Resale}_i] + (-5.748\text{e}+05) [\text{Bedrooms}_i] + (3.469\text{e}+06) [\text{LogPremium}_i]$$

Now to test if the linear regression model is significant i.e. the model fits the data. One approach is compute the coefficient of determination R-square and other approach is to use the technique of ANOVA, so we draw the ANOVA table using the SPSS software and interpret the results

**Null hypothesis of Kolkata dataset:**

$$H_0: \text{Area} = \text{FeatureScore} = \text{Resale} = \text{Bedrooms} = \text{LogPremium} = 0$$

i.e. None of the explanatory variable has significant contribution to the Dependent variable.

Against the alternative hypothesis that:

$$H_1: \text{Area} \neq \text{FeatureScore} \neq \text{Resale} \neq \text{Bedrooms} \neq \text{LogPremium} \neq 0$$

i.e. There is at-least one explanatory variable, which has significant contribution to the dependent variable.

**ANOVA table for Kolkata city**

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1505014907619 01.660	5	3010029815238 0.332	68.321	.000 <sup>b</sup>
	Residual	1365764083961 5.338	31	440569059342. 430		
	Total	1641591316015 17.000	36			

a. Dependent Variable: Price

b. Predictors: (Constant), LogPremium, Area, FeatureScore, Resale, Bedrooms

Interpretation: The p-value of the model is less than 0.05, therefore reject the null hypothesis and conclude that at-least one of the explanatory variable, which has significant contribution to the dependent variable. Therefore the model significantly fits the data.

From the summary table, it is clear that the coefficient of determination of the model is 0.984 i.e. around 98% which imply that 98% of the variable in Y is explained by the regression model.

Now the Random forest methodology has been used, to build a predictive model. After building the predictive model, check the performance of the model by evaluating the R-square-score which is 0.9735968057224695 for the Kolkata dataset, which is better than linear regression model.

**(vi) Summary table for Hyderabad dataset:**

OLS Regression Results							
Dep. Variable:		Price		R-squared (uncentered):		0.971	
Model:		OLS		Adj. R-squared (uncentered):		0.971	
Method:		Least Squares		F-statistic:		7536.	
Date:		Fri, 04 Jun 2021		Prob (F-statistic):		0.00	
Time:		22:29:36		Log-Likelihood:		-17601.	
No. Observations:		1128		AIC:		3.521e+04	
Df Residuals:		1123		BIC:		3.524e+04	
Df Model:		5					
Covariance Type:		nonrobust					
		coef	std err	t	P> t	[0.025	0.975]
Area		5213.1300	188.054	27.722	0.000	4844.154	5582.106
FeatureScore		1.323e+04	3039.195	4.352	0.000	7262.735	1.92e+04
Resale		3.417e+05	1.11e+05	3.079	0.002	1.24e+05	5.59e+05
Bedrooms		-1.146e+06	1.05e+05	-10.942	0.000	-1.35e+06	-9.4e+05
LogPremium		5.728e+06	2.51e+05	22.864	0.000	5.24e+06	6.22e+06
Omnibus:		70.717	Durbin-Watson:		1.991		
Prob(Omnibus):		0.000	Jarque-Bera (JB):		109.480		

**Skew:** 0.495      **Prob(JB):** 1.69e-24

**Kurtosis:** 4.161      **Cond. No.** 8.99e+03

The required linear regression model of the Hyderabad city dataset is:

$$Y[\text{Price}] = 5213.1300 [\text{Area}_i] + (1.323e+04) [\text{FeatureScore}_i] + (3.417e+05) [\text{Resale}_i] + (-1.146e+06) [\text{Bedrooms}_i] + (5.728e+06) [\text{LogPremium}_i]$$

Now to test if the linear regression model is significant i.e. the model fits the data. One approach is compute the coefficient of determination R-square and other approach is to use the technique of ANOVA, so we draw the ANOVA table using the SPSS software and interpret the results

**Null hypothesis of Hyderabad dataset:**

$H_0$ : Area = FeatureScore = Resale = Bedrooms = LogPremium = 0

i.e. None of the explanatory variable has significant contribution to the Dependent variable.

Against the alternative hypothesis that:

$H_1$ : Area  $\neq$  FeatureScore  $\neq$  Resale  $\neq$  Bedrooms  $\neq$  LogPremium  $\neq$  0

i.e. There is at-least one explanatory variable, which has significant contribution to the dependent variable.

**ANOVA table for Hyderabad city**

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8489787061885	5	1697957412377	1003.857	.000 <sup>b</sup>
		649.000		129.800		
	Residual	1821673590462	1077	1691433231627		
		321.000		.039		
	Total	1031146065234	1082			
		7970.000				

a. Dependent Variable: Price

b. Predictors: (Constant), LogPremium, Resale, Bedrooms, FeatureScore, Area

Interpretation: The p-value of the model is less than 0.05, therefore reject the null hypothesis and conclude that at-least one of the explanatory variable, which has significant contribution to the dependent variable. Therefore the model significantly fits the data.

From the summary table, it is clear that the coefficient of determination of the model is 0.971 i.e. around 97% which imply that 97% of the variable in Y is explained by the regression model.

Now the Random forest methodology has been used, to build a predictive model. After building the predictive model, check the performance of the model by evaluating the R-square-score which is 0.9997239136064422 for the Hyderabad dataset, which is better than linear regression model.

**Table for comparison of R-square value of linear regression model and Random forest model**

City	R-square-value of Linear regression model	R-square-value of Random forest model
Mumbai	0.953	0.999729487
Bangalore	0.968	0.999696147
Delhi	0.978	0.999937819
Chennai	0.984	0.999276944
Kolkata	0.984	0.973596806
Hyderabad	0.971	0.999723914

Interpretation:

From the above table, it is clear that the R-square value of linear regression model of each city is less than the R-square value of Random forest model of each city except in case of Kolkata dataset. Hence the Random forest methodology provides more accurate model than the linear regression model, by comparing the values of R-square of both the models.

### **3.1.3 Comparison between the prediction mean and mean absolute error of each cities of Linear regression model and random forest model:**

**Table 1: prediction mean and mean absolute error of linear regression model:**

	Linear regression model	
City	Prediction mean	Mean absolute error
Mumbai	11077270.78	2007730.931
Bangalore	7573379.595	1100592.783
Delhi	9427479.741	1227510.507
Chennai	6071679.38	631639.5474
Kolkata	4117029.804	669708.2045
Hyderabad	8073683.209	1140399.936

**Table 2: prediction mean and mean absolute error of Random forest model:**

	Random forest model	
City	Prediction mean	Mean absolute error
Mumbai	10466953.09	41654.59069
Bangalore	7295239.812	14419.82601
Delhi	8565260.906	27646.91349
Chennai	5962225.508	16008.24813
Kolkata	4744404.163	623354.7196
Hyderabad	7893542.604	18671.62035

#### **Interpretation:**

From the above Tables and Graphs of Prediction mean and Mean absolute error of Linear regression model and Random forest model, it is clear that Mumbai is the costliest city among the six cities to buy a house with prediction mean of 11077270.78 from linear regression model and 10466953.09 from Random forest model which is approximately around Rs 11.0 million or Rs 1.1 crore.

Now from the mean absolute error of the two models, it is clear that Random forest model has least error than linear regression model, which provides more accuracy than the linear regression model.

### **3.2 Conclusion**

After analysis the six cities dataset, it is clear that:

- (1) For checking the multi-co-linearity in the dataset by using variance inflating factor (VIF), it is clear that multi-co-linearity exist in the dataset, which can be reduce by dropping some of the explanatory variable.
- (2) After fitting the linear regression model using OLS method and Building the predictive model using Random forest methodology, it is clear that comparing both the R-square values conclusion can be made that Random forest methodology builds more accurate predictive model.
- (3) After calculating the predictive mean and mean square error of both the models for six cities, it is clear that the Mumbai is the costliest city of all six cities to buy a house and Random forest methodology provides least error and accurate predictive model than the linear regression model.

#### **4. References**

1. Johnston, J., *Econometric Methods*, 3d ed., McGraw-Hill, New York, 1984.
2. Gujarati, N, D., and Dawn C. Porter, *Basic Econometrics*, 5d ed., McGraw-Hill, New York, 2008.
3. Wooldridge, Jeffrey M., *Introductory Econometrics*, 3d ed., South-Western College Publishing, 2006.
4. Dougherty, Christopher, *Introduction to Econometrics*, 3d ed., Oxford University Press, Oxford, 2007.
5. VanderPlas, J., *Python Data Science Handbook*, 1<sup>st</sup> ed., O'Reilly, CA, 2016.
6. Albon, C., *Machine Learning with Python Cookbook*, 1<sup>st</sup> ed., O'Reilly, CA, 2018.
7. Géron, A., *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2d ed., O'Reilly, CA, 2019.
8. Gundimeda, H., *Hedonic price method – A Concept Note*, Research paper, Madras School of Economics, Chennai, pp 1-12, 2005.
9. Kanojia, A., *Valuation of Residential Properties by Hedonic Pricing Method- A State of Art*, *International Journal of Recent Advances in Engineering & Technology (IJRAET)*, 2016.
10. Bhalla, A.K., (2008), “Housing finance in India: Development, growth and policy implications”. *PCMA Journal of Business*, 1(1), 51-63.
11. Bhalla, A.K., Arora P., & Gill P.S., (2009), “Competitive dynamics of Indian housing finance industry” *Paradigm*, 13 (1), 28-38.
12. Hujia, Yu., Jiafu Wu., “Real Estate Price Prediction with Regression and Classification”, CS 229 Autumn Project Final Report,(2016).
13. Housing in India, Wikipedia,  
[https://en.wikipedia.org/wiki/Housing\\_in\\_India](https://en.wikipedia.org/wiki/Housing_in_India)
14. Shinde, N., Gawande, K., “Valuation of House Prices Using Predictive Techniques”, *International Journal of Advances in Electronics and Computer Science*, ISSN: 2393-2835,(2018).