**HOUSE VALUE PREDICTION**

**Final Report**

##### Submitted in partial fulfilment of the requirements of

##### Post Graduate Program in Machine Learning

by

Amitej Singh Sandhu

Ashwin P

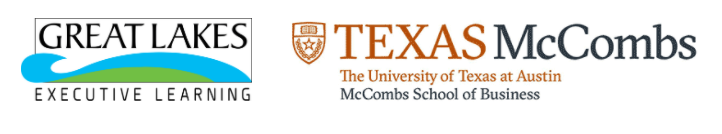
Indrajit Basu

Jai Kishan A

Priyanka kumari Sharma

Mentors:

Mr. Survesh Satish Chauhan

****

**ABSTRACT**

Investment in real estate generally seems to be profitable because their property values do not decline rapidly. Generally, property values rise with respect to time and its appraised value need to be calculated. Understanding an appropriate market value is critical during the buy or sale of a property or while applying for the loan and for the marketability of the property. These appraised values are determined by the professional appraisers. However, drawback of this practice is that these appraisers could be biased due to bestowed interests from buyers, sellers or mortgages. This document presents the implementation of price prediction project using machine learning methods and models to help predict the property values without any bias. This automated model is aimed at proving a good first estimate to the buyers and sellers alike and help less experienced customers to understand whether the property rates are overrated or underrated. The goal of this project is to critically analyze all possible variables that can affect the house price, evaluate their interrelation and effect on the house price; run multiple regression models; use methods to improve quality of the models to finally predict the house prices with high level of accuracy.

**Table of Contents**

Chapter1: Project Definition and Planning4

Introduction4

1.1 Problem Statement4

1.2 Objective4

1.3 Project Flowchart4

Chapter2: Data Understanding and Pre-Processing6

2.1 Data Collection6

2.2 Data Cleaning / Data Editing7

2.3 EDA8

2.3.1 Data Distribution8

2.3.2 Univariate Analyses – by Box Plots9

2.3.3 Univariate Analyses – Visualizing skewness 11

2.3.4 Outlier Treatment 12

2.3.5 Bivariate Analyses: Analyzing target attribute with other variables 17

2.3.6 Multivariate Analyses: Correlation among variables 22

Chapter3: Machine Learning Models25

3.1 Methodology25

3.2 Model Metrices and Cross Validation25

3.3 Dataset Split prior to Modeling26

3.4 Base Model Generation: Linear Regression 27

3.5 Other Models and Performance and Performance Vis-à-vis the Base Model 28

3.6 Hyperparameter Tuning & Final Model Selection31

3.7 Implication & Application of the Machine Learning Model32

3.8 Conclusion and Reflection 33

3.9 Closing Remarks and Scope for Improvement 33

APPENDIX: List of Figures34

**Chapter 1: Project Definition and Planning**

**Introduction**

People are looking to buy a new home tend to be more conservative with their budgets and market strategies. Ideally, to find house prices, one usually tries to find similar properties in their neighborhood and based on gathered data they will try to assess the house price. A house value is simply more than location and square footage. There are multiple features that will come into the consideration while estimating the house price. The customer's expectations differ from one individual to another. As there are multiple features that decide the house values, it is difficult for buyer/seller to set the right target price. The seller has to fix the price based on the market value by considering all aspects that give a house its value. It can’t be too low or too high. Similarly, buyers should be aware of their target price based on all the desired requirements. House price prediction also helps the seller determine the selling price of a house so that the house is sold as quickly as possible and for the buyers to make a prompt and well-informed decision to buy a buy that meet their requirements.

**1.1 Problem Statement**

* Housing sales price is determined by numerous factors such as area of the property, location of the house, number of bedrooms and so on. The goal is to predict efficient house pricing for real estate with respect to customers budgets and priorities.
* As sizeable data are typically available in terms of factors affecting house prices, machine learning methods could efficiently work on those multitude of factors and use the intricate inter-relationships of the independent variables to estimate the approximate house costs with high accuracy.

**1.2 Objective**

* To predict the selling price of the houses using various machine learning algorithms.
* To analyze and compare model’s performance in order to choose the best model.

**1.3 Project Flowchart**

Designing the project plan and process flow is the blueprint of the project and defines the end-to-end flow of steps required to carry out the project. Housing price prediction is a Regression problem hence we’ll deploy the individual elements of the process accordingly, keeping in line with broader machine learning guiding principles throughout the process.

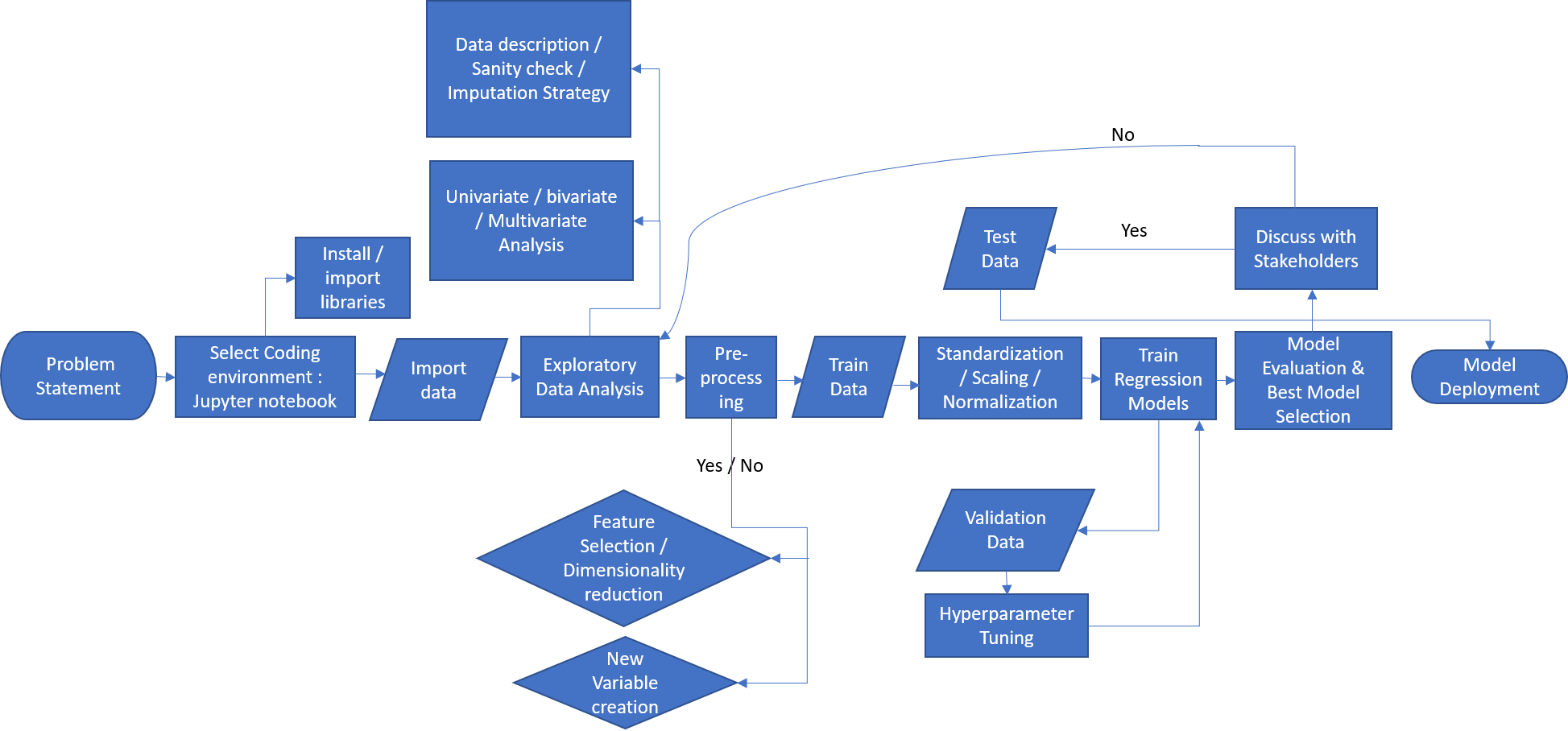


Figure 1.1: Process Flowchart

**Chapter 2: Data Understanding and Pre-Processing**

**2.1 Data Collection**

In this project, a housing dataset is use to describes the sales of residential units in Seattle region of the state of Washington, USA, based on the latitude-longitude data and zip codes. The dataset contains a large number of variables that are involved in determining a house price. A csv copy of the data was obtained from https://olympus.greatlearning.in/. The dataset includes 21613 records and it is qualified by 23 important different features of which Price is the target variable.

The description of the data set is given below:

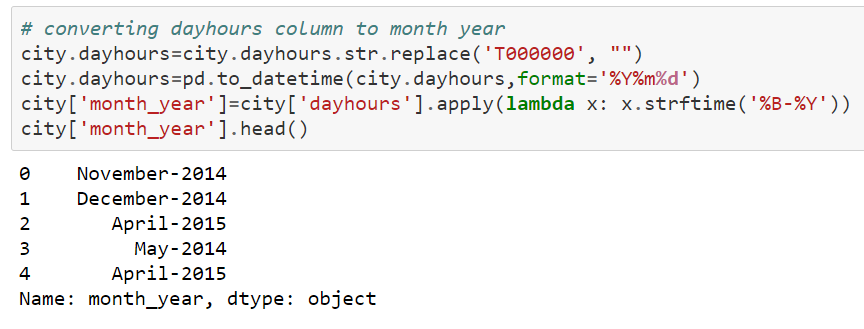
|  |
| --- |
| 1. cid: a notation for a house |
| 1. dayhours: Date house was sold |
| 1. room\_bed: Number of Bedrooms/House, |
| 1. room\_bath: Number of bathrooms/bedrooms, |
| 1. living\_measure: square footage of the home, |
| 1. lot\_measure: square footage of the lot, |
| 1. ceil: Total floors (levels) in house, |
| 1. coast: House which has a view to a waterfront, |
| 1. sight: Has been viewed, |
| 1. condition: How good the condition is (Overall), |
| 1. quality: grade given to the housing unit, based on grading system, |
| 1. ceil\_measure: square footage of house apart from basement, |
| 1. basement\_measure: square footage of the basement, |
| 1. yr\_built: Built Year, |
| 1. yr\_renovated: Year when house was renovated, |
| 1. zipcode: zipcode |
| 1. lat: Latitude coordinate, |
| 1. long: Longitude coordinate, |
| 1. living\_measure15: Living room area in 2015(implies--some renovations) This might or might not have affected the lot size area, |
| 1. lot\_measure15: lot Size area in 2015(implies--some renovations), |
| 1. furnished: Based on the quality of room, |
| 1. total\_area: Measure of both living and lot. |
| 1. price: Price is prediction target (Target variable) |

**2.2 Data Cleaning / Data Editing**

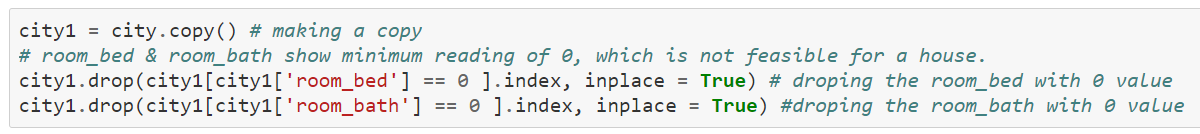
It is a process of transforming the raw, complex data into systematic understandable knowledge. It involves the process of finding out missing and redundant data in the dataset.

Entire dataset is checked for Null values and whichever observation consists of Null values will be deleted. Thus, this brings uniformity in the dataset.

* In our dataset, there was no missing values found that means each record was constituted its corresponding feature values.
* There is no duplicate data found in data set.
* dayhours is present as object datatype, needs to be converted into consumable numeric data format. Also, house sold dates have been converted into month / year format in order to check if some seasonal variability exists in terms of house sale price.



* Number of bedrooms and bathrooms read minimum values of 0, which is not possible for a residential property; hence those values have been removed during EDA.



* Yr-renovated has been converted as renovated versus non-renovated format during EDA to see whether renovated houses have any impact on the house price vis-a-vis non-renovated houses having similar other house parameters.

**2.3 EDA**

Before applying any model to the dataset, we need to find out characteristics of the dataset. Thus, we need to analyze the dataset and find relationship between different features and the target variable.

**2.3.1 Data Distribution**

Key observations from the 5-point summary are listed below which can provide a quality overview of overall data distribution, which are further analyzed using univariate and bivariate plots in the later sections.

1. price: Target column value ranges from 75k - 7700k. As Mean > Median, it's Right-Skewed.
2. room\_bed: Number of bedrooms range from 0 - 33. As Mean slightly > Median, it's slightly Right-Skewed.
3. room\_bath: Number of bathrooms range from 0 - 8. As Mean slightly < Median, it's slightly Left-Skewed.
4. living\_measure: As Mean > Median, it's Right-Skewed.
5. lot\_measure: As Mean almost double of Median, it's highly Right-Skewed.
6. ceil: Number of floors range from 1 - 3.5 As Mean ~ Median, it's almost Normal Distributed.
7. coast: From above analysis we got know, very few houses has waterfront view.
8. sight: Value ranges from 0 - 4. As Mean > Median, it's Right-Skewed
9. condition: Represents rating of house which ranges from 1 - 5. As Mean > Median, it's Right-Skewed
10. quality: Representing grade given to house which range from 1 - 13. As Mean > Median, it's Right-Skewed.
11. ceil\_measure: As Mean > Median, it's Right-Skewed.
12. basement: As Mean >> Median, it's Highly Right-Skewed.
13. yr\_built: House built year ranges from 1900 - 2015. As Mean < Median, it's Left-Skewed.
14. yr\_renovated: This column can be used as to knowing whether house is renovated or not.
15. zipcode: House Zip Code ranges from 98001 - 98199. As Mean > Median, it's Right-Skewed.
16. lat: Latitude, As Mean < Median, it's Left-Skewed.
17. long: Longitude, As Mean > Median, it's Right-Skewed.
18. living\_measure15: As Mean > Median, it's Right-Skewed.
19. lot\_measure15: As Mean highly > Median, it's Highly Right-Skewed.
20. furnished: Representing whether house is furnished or not.
21. total\_area: As Mean is almost double of Median, it's Highly Right-Skewed

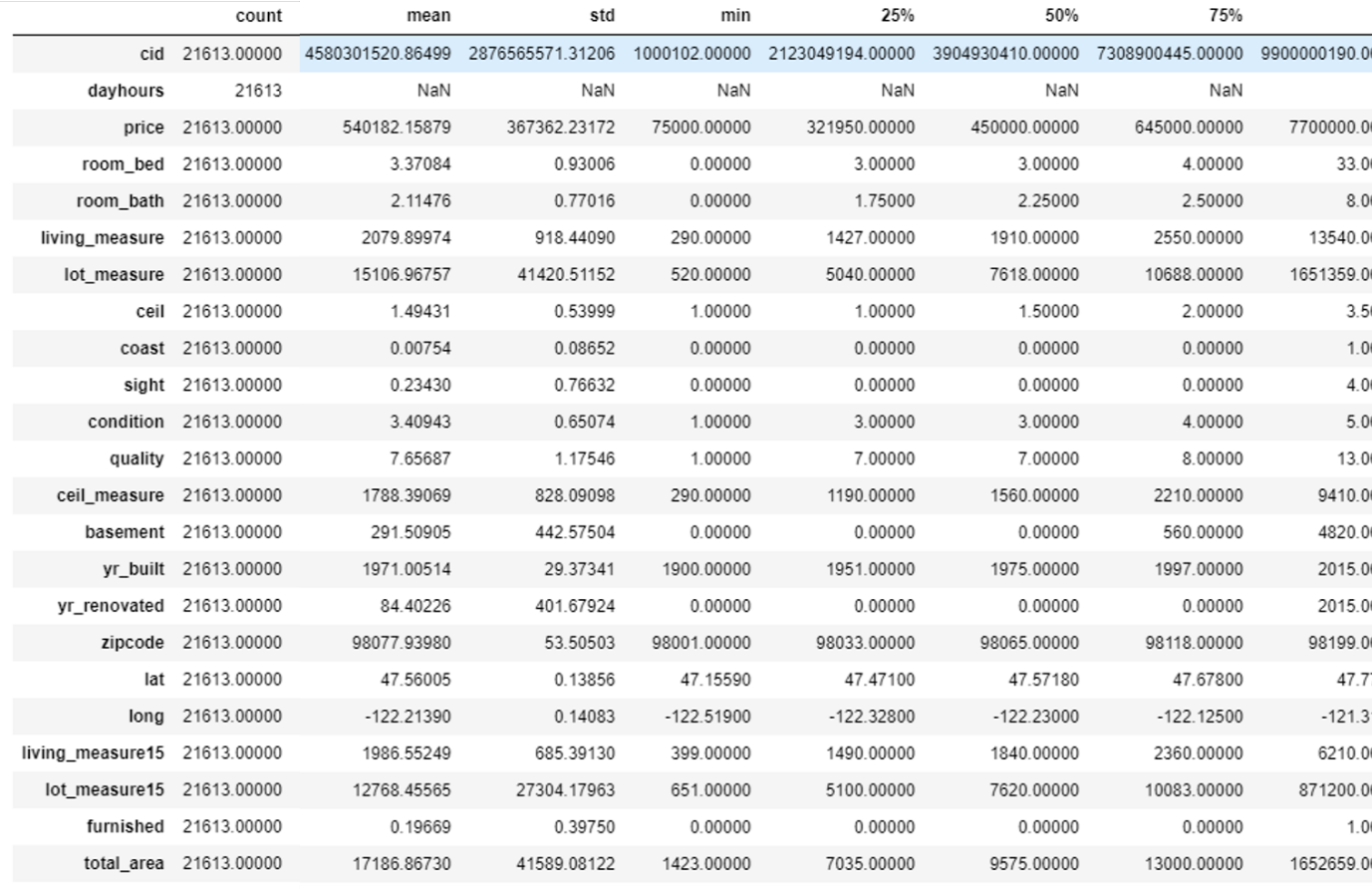


Figure 2.1: 5-point summary of the dataset

**2.3.2 Univariate Analyses - By BoxPlots**

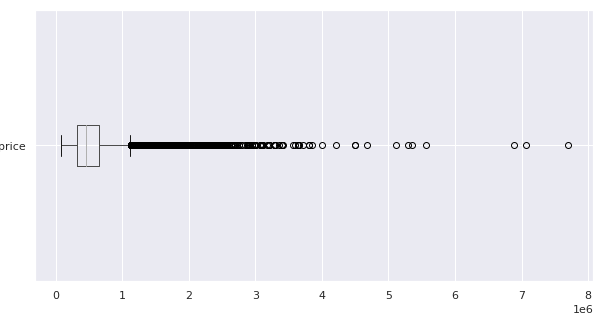
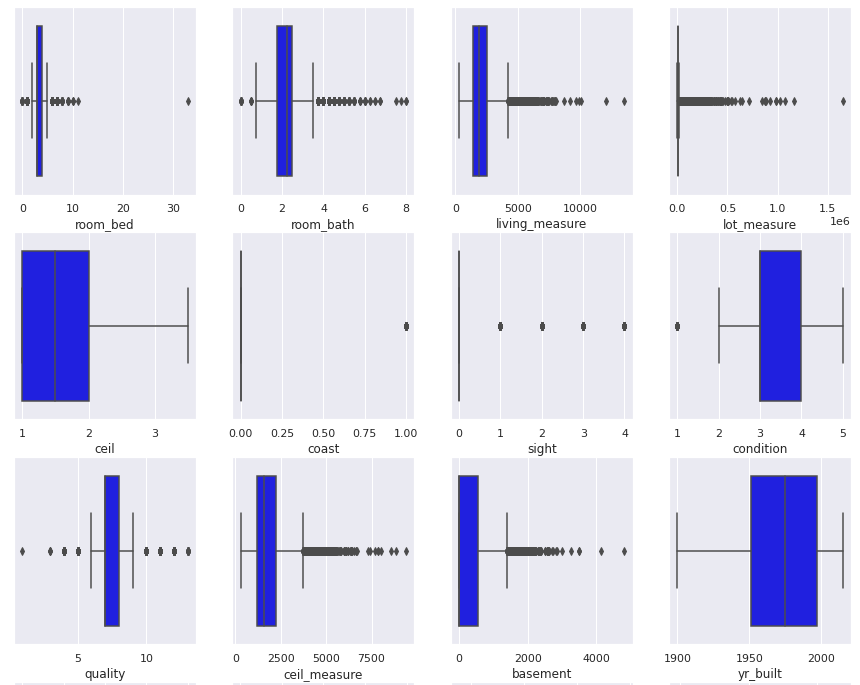


Figure 2.2: Box plot of all house price

The plot above and information from the raw data file indicates Price (target variable) have a of values are > $1 mm but only a dozen values are >$4mm out of a total of 21613 readings.



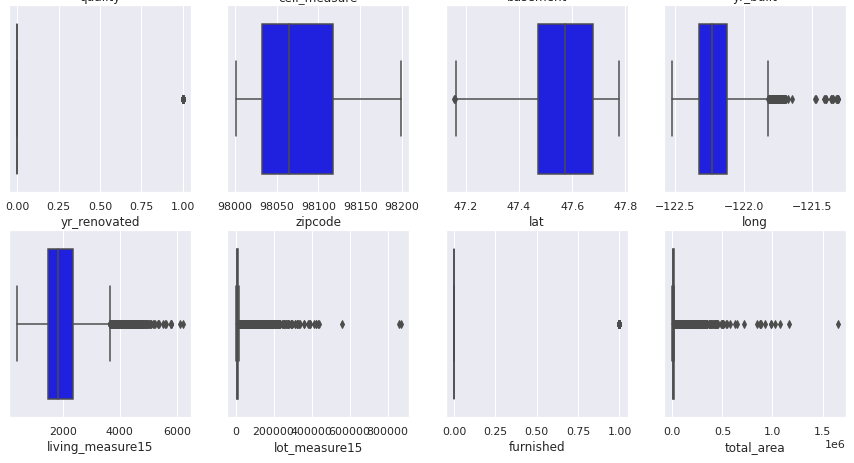
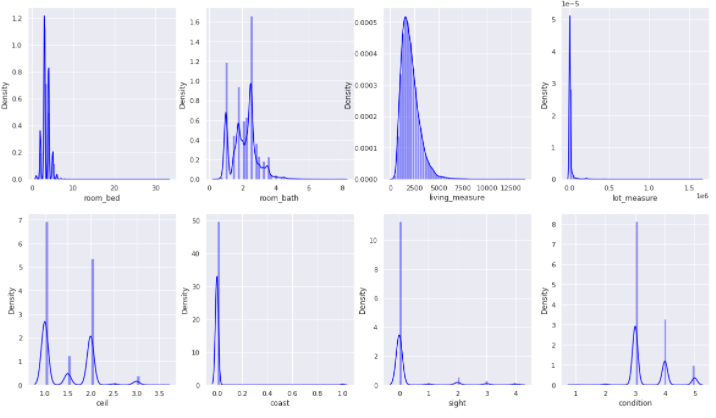


Figure 2.3: Box plot of all attributes prior to outlier removal

Above boxplots indicate that the following variables have values beyond whiskers (i.e., > 1.5 IQR & < 1.5 IQR) i.e., some outliers are present in the dataset.

* We have a more no of outliers in the data set Total\_area, longitude, latitudes, living\_measures15, lot\_measure15, basement, ceil\_measure, quality, lot\_measure, living measure, room bath and room bed.
* Furnished, yr\_renovated and coast columns are having 0 or 1 value
* No outliers in ceil, yr\_built and Zipcode columns

**2.3.3 Univariate Analyses: Visualizing Skewness**



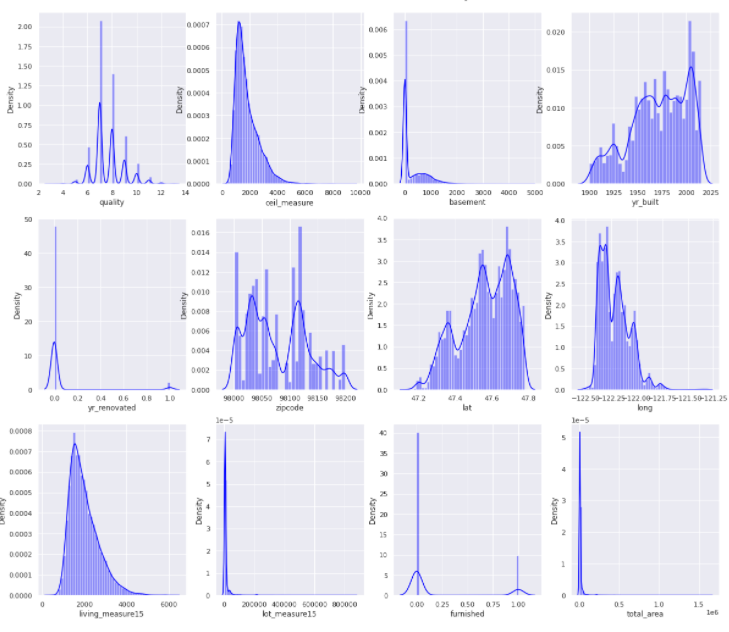


Figure 2.4: skewness of all attributes prior to outlier removal

**Key observations:**

* price, room\_bed, living\_measure, lot\_measure, sight, condition, quality, ceil\_measure, basement, zipcode, longitude, living\_measure15, lot\_measure15, total\_area: are Right-Skewed.
* room\_bath, yr\_built, lat are Left-Skewed.
* ceil: it's almost Normal Distributed.
* Max room\_bed value of 33.
* A lot of houses don't have basement (reads 0)

**2.3.4 Outlier treatment**

We’ve removed house prices > $4mm as they’re dozen out of 21000 plus values as below:

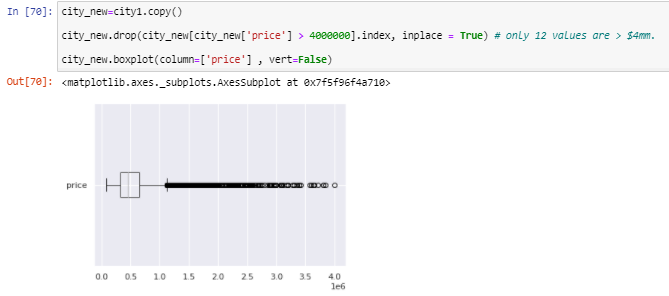


Figure 2.5: Outlier removal of price

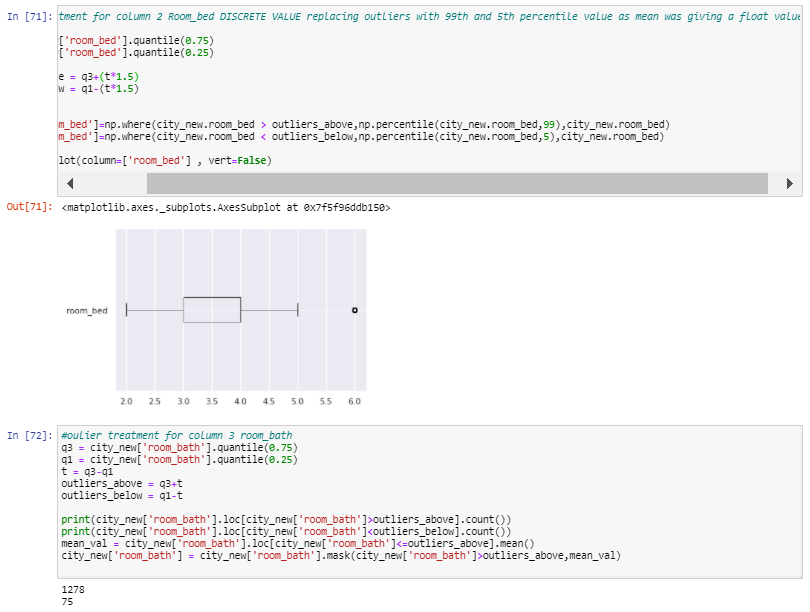
Different outlier treatment mechanisms have been adopted for the independent variables as listed below:

Three principal methods of outliers treatments have been adopted:

1. removing values < 5th & > 99th percentile

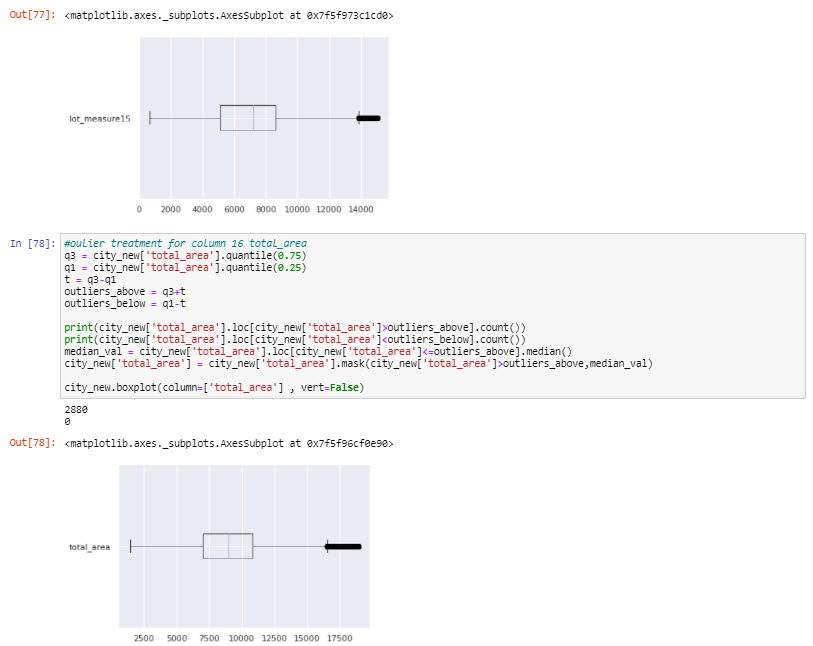
2. replacing outliers with median values for skewed variables

3. replacing outliers with mean values









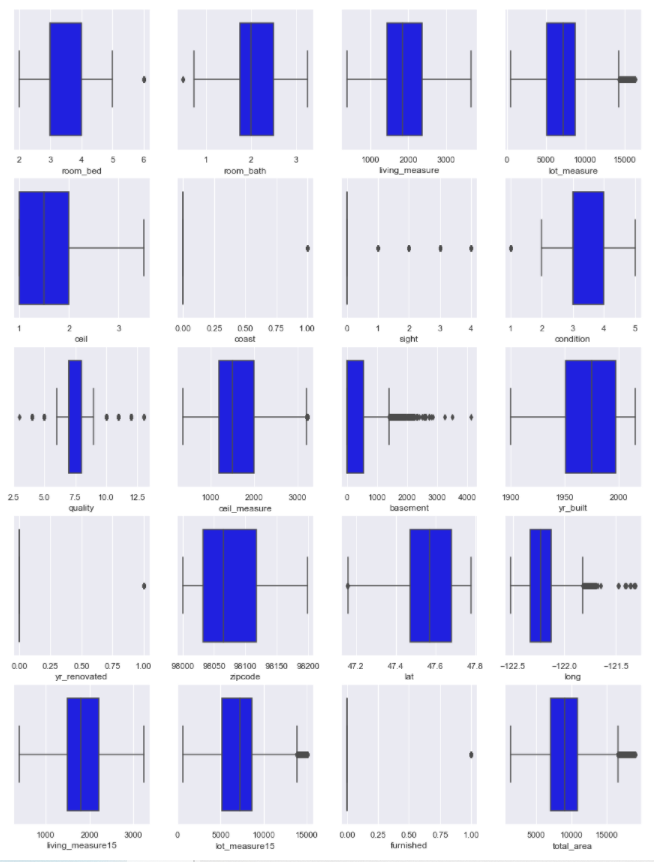
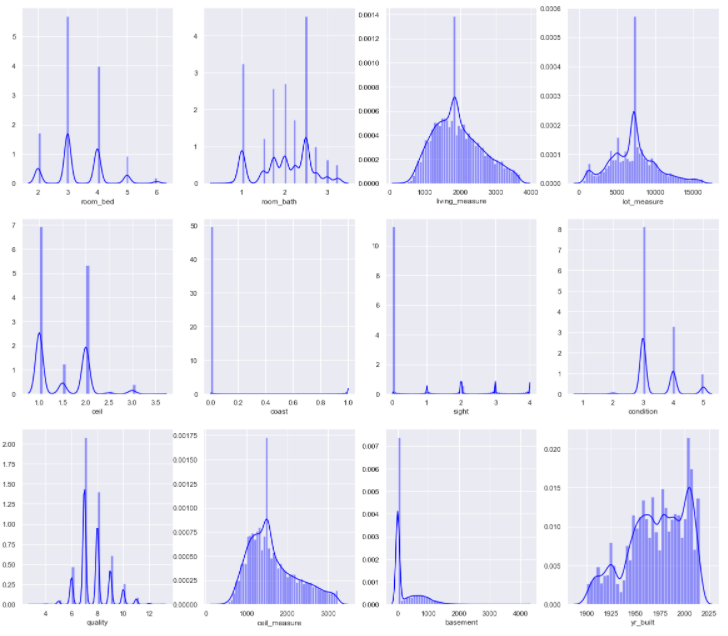


Figure 2.6: Box plot of all attributes after outlier removal



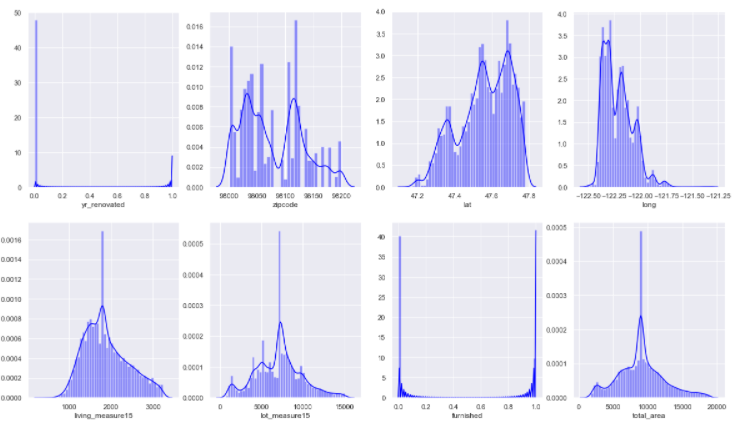


Figure 2.7: skewness of all attributes after outlier removal

**2.3.5 Bivariate Analysis: Analyzing Target Attribute with other variables**

Most significant of the independent variables versus house price along with some interrelationship among variables are discussed below. Rest of the details and comments may be found in the code file.

#### **Price vs. sale time:**

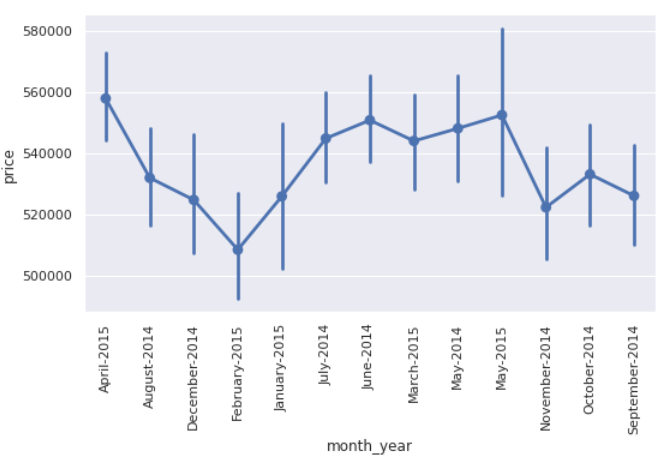


Figure 2.8: Sale time vs. price

* The mean price of the houses tends to be high during April, May

#### **Bedroom vs. Price:**

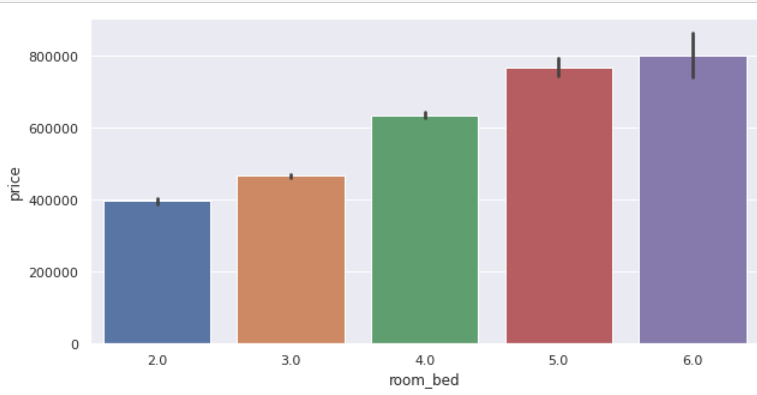


Figure 2.9: Bedroom vs. price

* Higher number of rooms draw higher price.

#### **Different Area Measures vs. Price:**

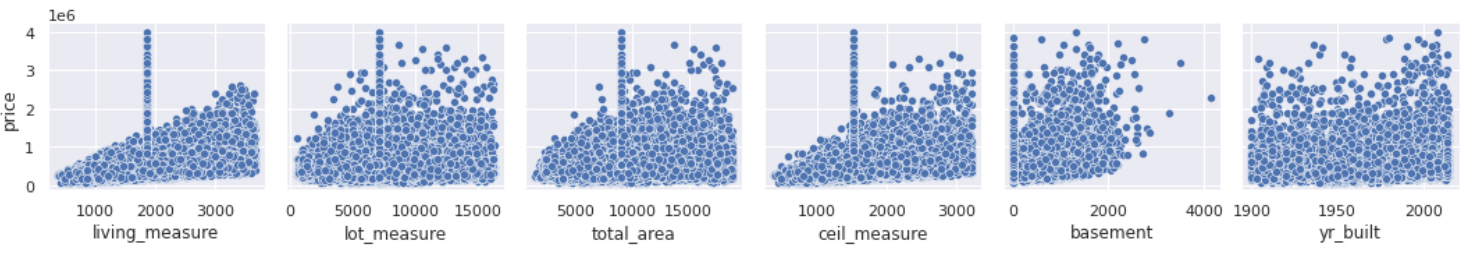


Figure 2.10: Different Area Measures vs. price

* living\_measure shows positive correlation with price
* We may either chose 'lot\_measure' or 'total\_area' owing to their high inter-correlation and similar response to 'price'
* 'ceil\_measure' is positively correlated with house price.
* basement' is positively correlated with house price but data cloud shows larger scatter.
* Price shows a gradual increase in recent years.

#### **Inter-relationship among area measures:**

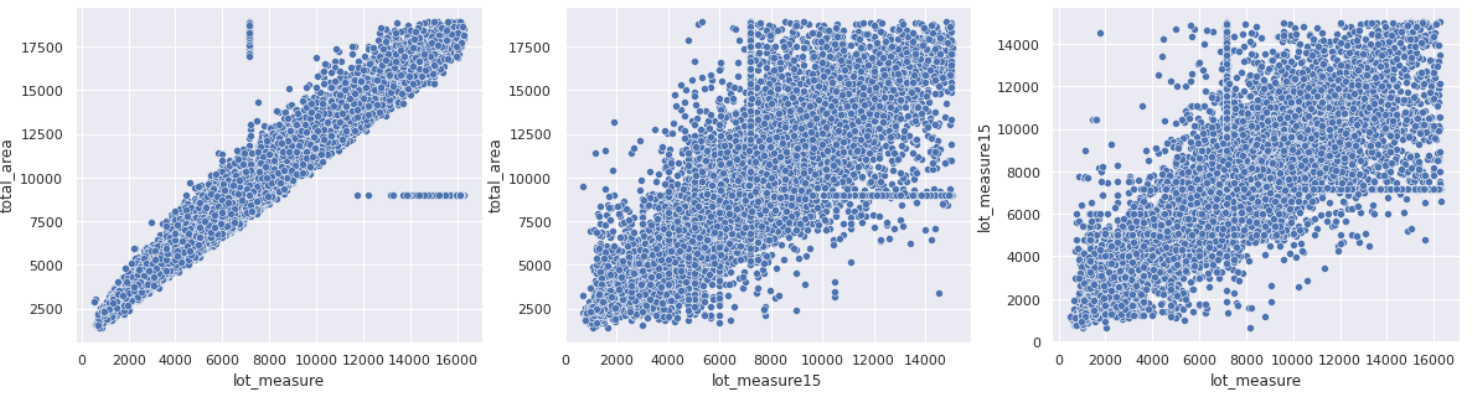


Figure 2.11: Interrelationship among different area measures

* We may drop lot\_measure15 if lot\_measure15 has strong correlation with lot\_measure
* lot\_measure has better correlation with total area with lesser spread.

#### **Renovation vs. Price & Area Measures**

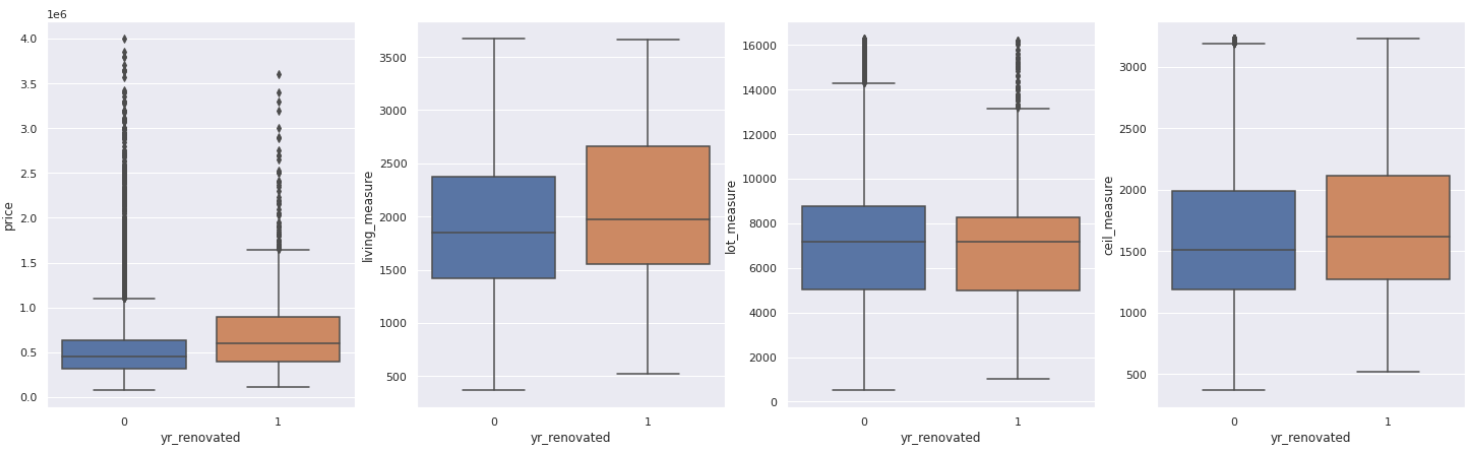


Figure 2.12: Renovation vs. price

* renovated houses have high price range in general.
* yr\_renovated doesn't have any observable correlation with other independent variables related to area measure.

#### **Furnished vs. Price:**

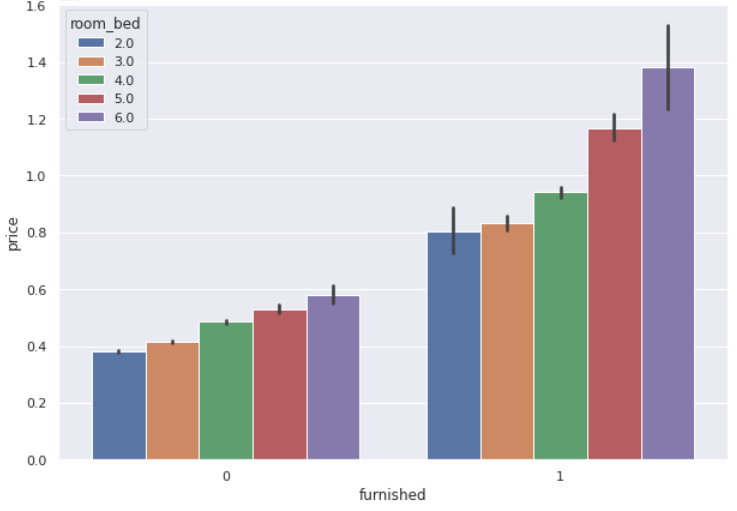
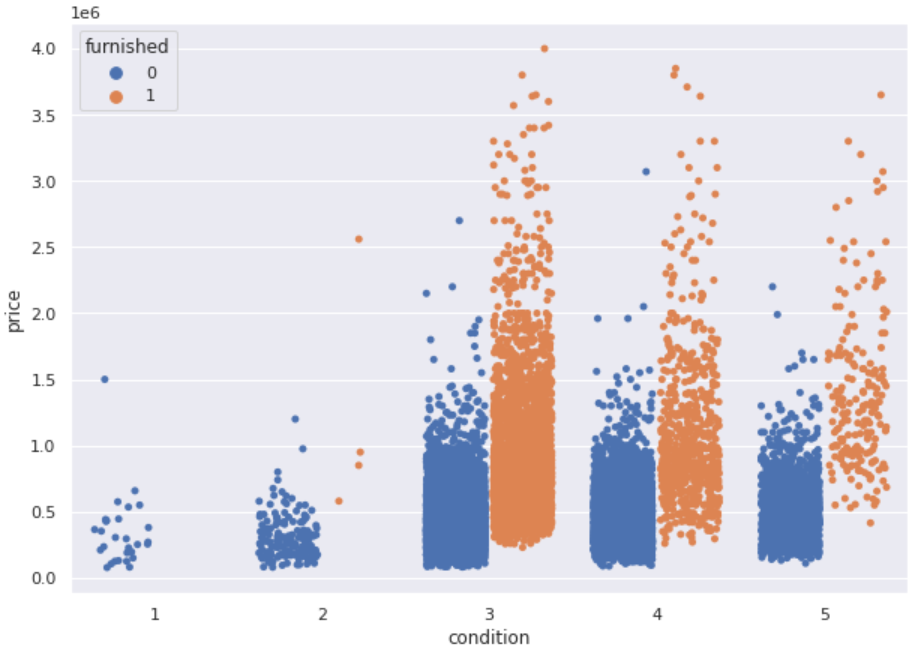


Figure 2.13: Furnished houses vs. price

* Furnished Plots are priced above $800,000, with minimum no. of bedroom/house.
* Also, as the no. of bedroom/house increases, the price of house increases as well, signifying positive correlation with price.

#### **Condition vs. Price:**

* Houses with condition 1-2 have negligible furnished houses thus they have low prices.
* Houses with condition 3-4-5 have higher number of furnished house and thus are in the higher price slab.
* Both furnished and unfurnished houses in any of the individual 3-4-5 conditions fall in the same price range for a specific "condition" category.
* Furnished houses starting prices are higher than unfurnished ones but higher end of house price is not impacted by furnishing status.
* Overall, furnished flats have better condition thus have higher price.



#### Figure 2.14: Condition vs. Price

#### **Quality vs. Price:**

* Houses with quality higher than 8 are furnished.
* Price of houses increases with Quality and the rate of increase of house price vs. quality is higher for higher price range.

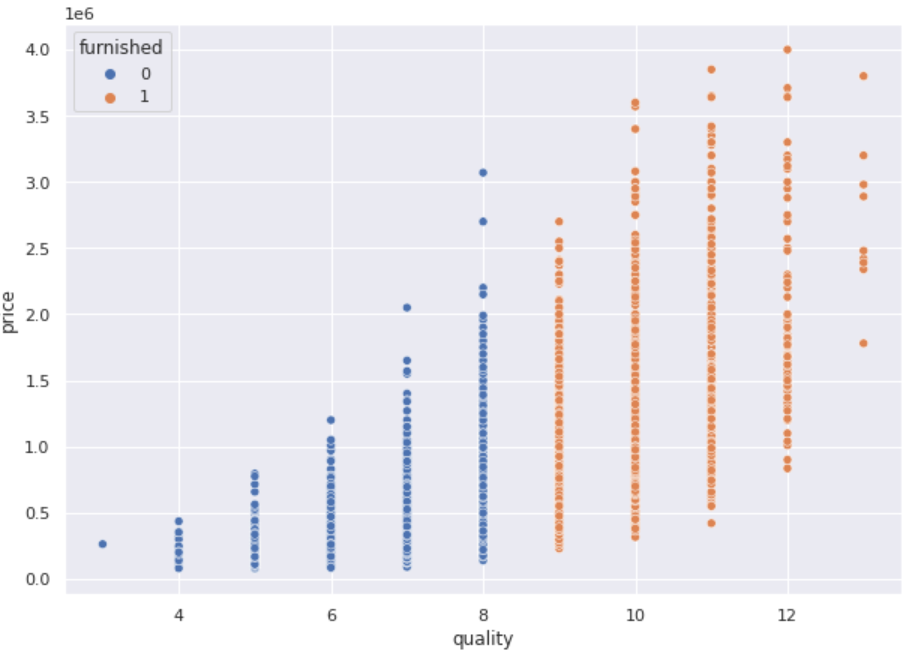
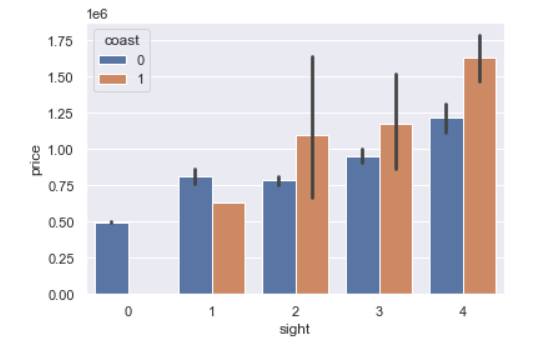


Figure 2.15: Quality & furnished vs. Price

#### **Sight vs. Price:**

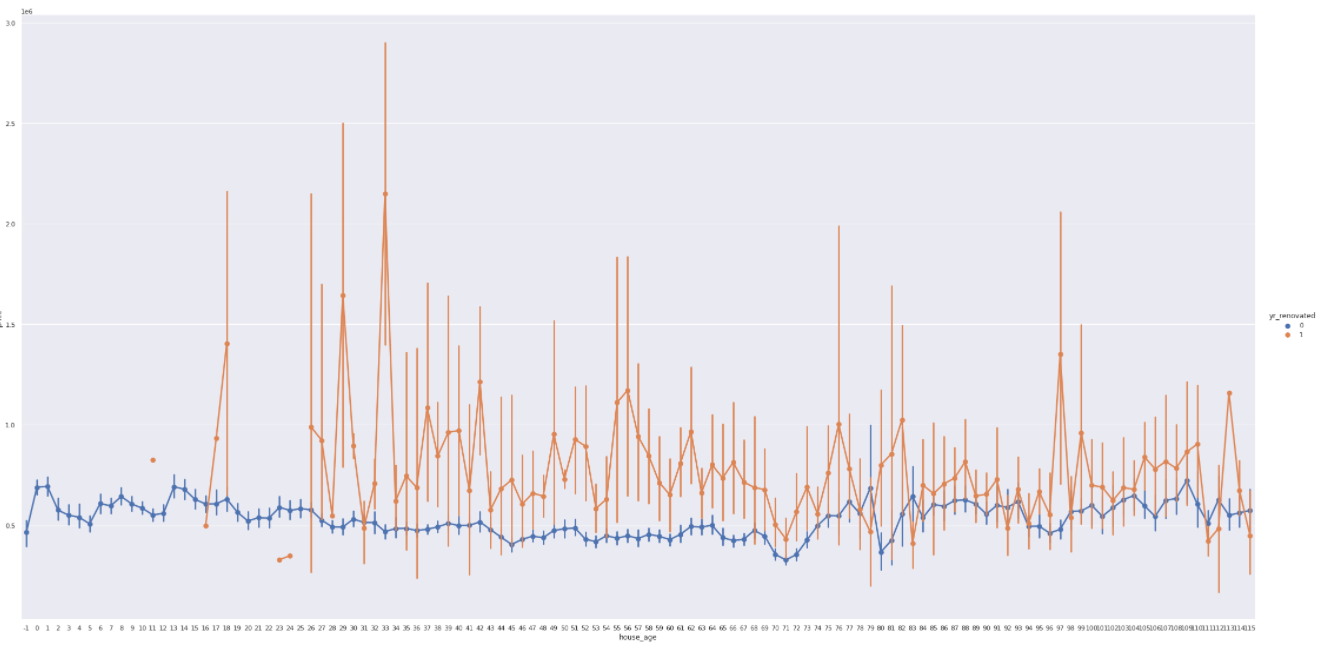
* Houses in the vicinity of a waterfront are slightly higher priced, especially for visits > = 2.

****

#### Figure 2.16: Sight vs. Price

#### **House Age vs. Price:**

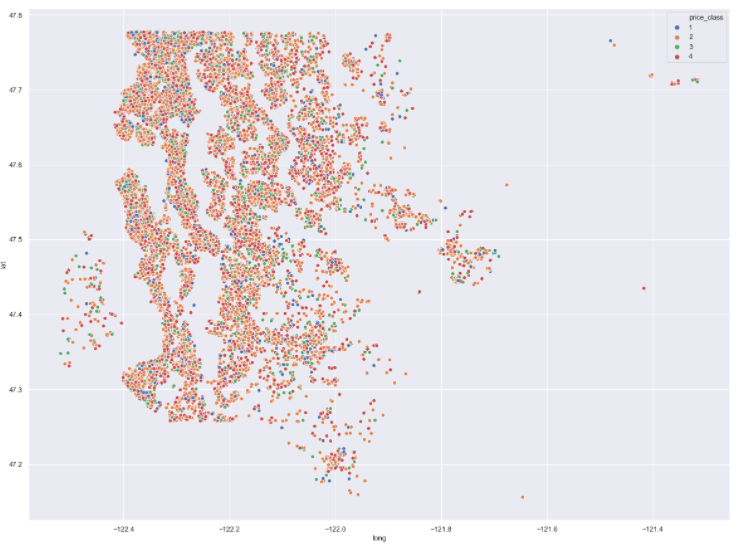
* House age between 26 - 77 years are having higher price for renovated houses.



#### Figure 2.17: House Age vs. Price

#### **Effect of latitude and longitude on house price:**

Target variable is divided into 4 classes with price value from <$240000, 240000-480000, 480000-750000 & >750000. Class 2 (240000-480000) & Class 4 (>750000) have a lot of overlap across and doesn’t show much clarity. However, this is a large area to be considered and hence a challenging task whether to include or drop these variables into model creation.

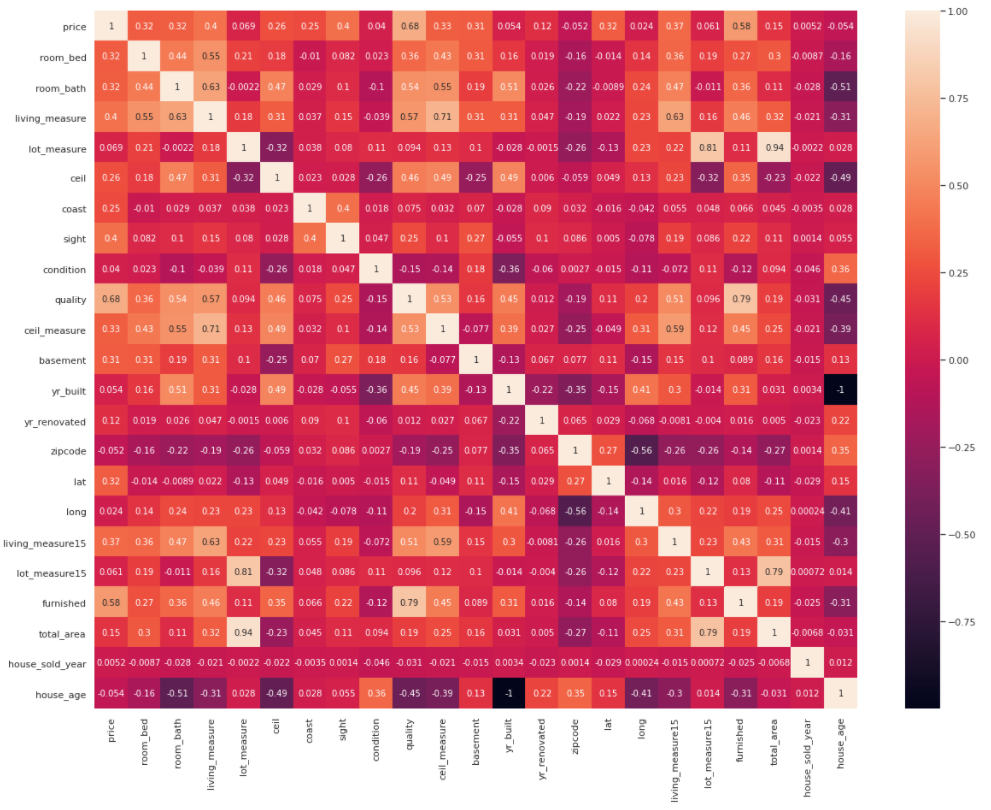


#### Figure 2.18: Longitude- Latitude vs. Price

#### **2.3.6 Multivariate Analyses: Correlation among variables**

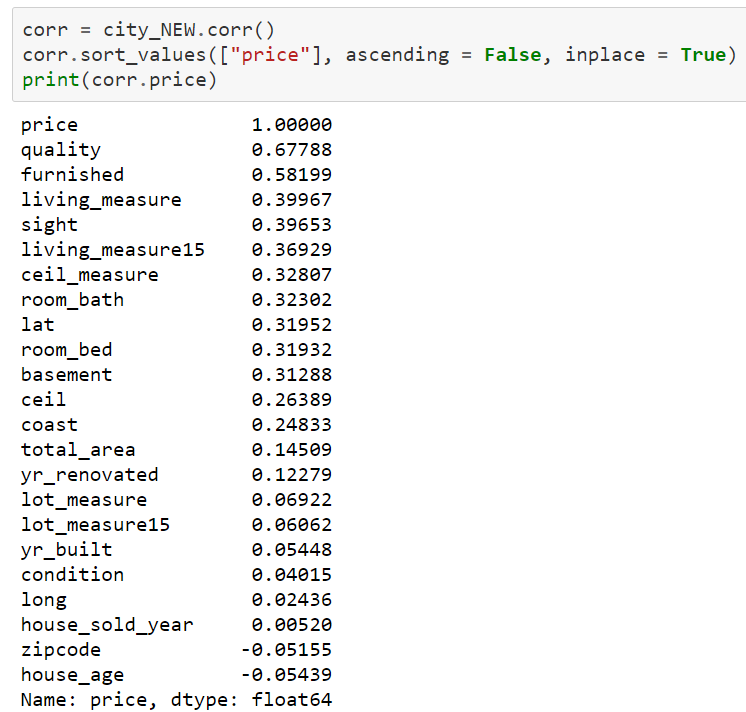
Pairplots and correlation matrix have been used to decipher interrelation among dependent and independent variables. For sake of visual clarity heatmap is shown below as pairplots of a large number of variables are not easy to visualize and multiple bivariate analyses above covers a great deal of the pairplot as such.

For correlation matrix heatmap, correlation number gives the degree of association between two variables. Correlation is represented as a value between -1 and +1 where +1 denotes the highest positive correlation, -1 denotes the highest negative correlation, and 0 denotes that there is no correlation.



#### Figure 2.19 Correlation matrix

Correlation values vs. price are:



#### Based on the data analysis and observations, following recommendations can be made for inputs into regression modelling exercise:

1. Either Latitude-longitude or zip code may be eventually dropped as they all contain location information. This will be decided based on modeling exercise outcomes.
2. We can drop continuous numeric variables that correlated <0.05 or < 0.1 with house price, e.g., lot\_measure15, based on the further analyses during modeling.
3. Living\_measure15 might be dropped as well as this doesn't add any additional value as compared to living\_measure.
4. CID can be dropped for sure as it doesn’t add any value to the modeling.

**Chapter 3: Machine Learning Models**

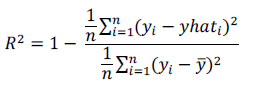
**3.1 Methodology**

House price evaluation dataset have 22 different features to choose from. Based on the earlier data evaluation, we’ve adopted a strategy to use as many variables as we deem optimum to start with.

House price is a continuous number; hence, the “Regression” type of supervised learning method is used as the price prediction strategy. We have used several models, starting with Linear regression and have calculated model accuracy, range of uncertainty through validation techniques and calculated errors for each method in order to estimate, compare and finally to determine the best likely model for house price prediction.

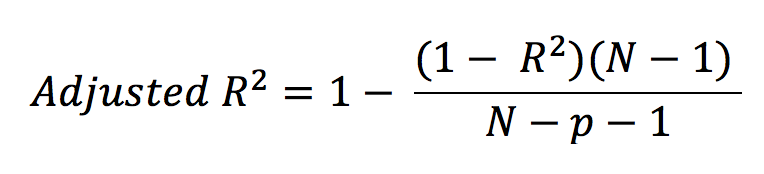
**3.2 Model Performance Metrics and Cross-Validation**

Performance metrics are the way to quantify and compare the efficiency of any machine earning model. The least-square regression uses R2 (R-squared) and Radj2 (Adjusted R-Square) metrics to measure the performance of the regression model. Radj2 (Adjusted R-Square) is used with multiple linear regression. Both of these metrics denotes the power of the ability to explain the selected independent variable(s) to the variation of the response variable. The equation of R2 (R-squared) is given by:



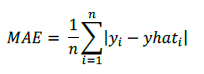
The numerator term gives the average of squares of residuals and the denominator shows the variance in y(response) value. A small value for R2 or higher mean residual error denotes a poor model.

The formula for calculating adjusted R squared is as follows, where:

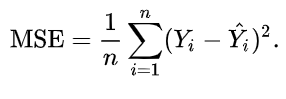


* R2: Sample R Squared
* p: Number of predictors
* N: Sample size

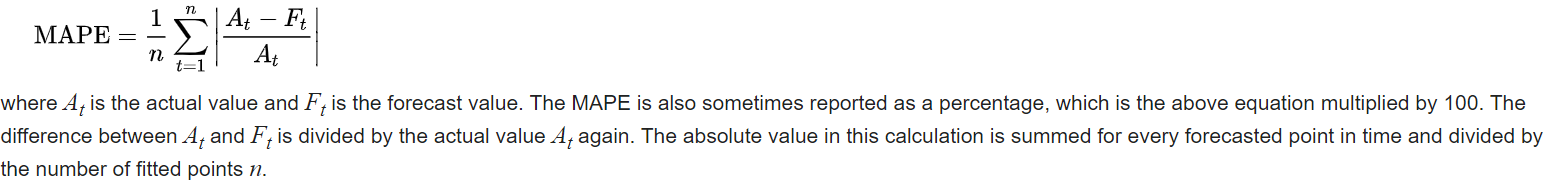
**Mean Absolute Error (MAE)** - Mean Absolute Error is the average of the difference between the actual and predicted value of the target variable.



**Mean Square Error (MSE) -** defined as: measures the [average](https://en.wikipedia.org/wiki/Expected_value) of the squares of the [errors](https://en.wikipedia.org/wiki/Error_(statistics)), that is, the average squared difference between the estimated values and the actual value.



**Mean Absolute Percentage Error (MAPE**), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in [statistics](https://en.wikipedia.org/wiki/Statistics), for example in [trend estimation](https://en.wikipedia.org/wiki/Trend_estimation), also used as a [loss function](https://en.wikipedia.org/wiki/Loss_function) for regression problems in [machine learning](https://en.wikipedia.org/wiki/Machine_learning). It usually expresses the accuracy as a ratio defined by the formula:



**Cross-Validation:**

Validation is required to estimate the likely range of performance for a machining learning (ML) model. Usually, the available data is not sufficient to represent the population but only a sample set. One such technique is cross-validation (k-fold cross-validation in this example) to have a better idea of model performance and associated uncertainty while put into production.

[K-Fold](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html#sklearn.model_selection.KFold) divides all the samples in k groups of samples, called folds (if k=n, this is equivalent to the *Leave One Out* strategy), of equal sizes (if possible). The prediction function is learned using k−1 folds, and the fold left out is used for the test. K-Fold gives a range of model accuracy and hence confidence in the mo0del accuracy.

**3.3 Dataset Split prior to Modeling**

The first step involved in the modelling is the splitting of the dataset into two parts in a 70:30 ratio. 70% of the data will be used for training the model, and the remaining 30% will be used to test the performance of the model. This splitting of data in Test/Train data is a general step.

On this split data, all the models used for prediction will fit.

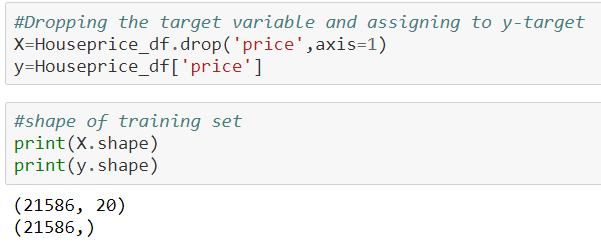
Also, a necessary step is to drop the target variable from the train data. In this case, the ‘Price’ column is dropped.

In terms of input variables, we’ve dropped the ‘month\_year’, ‘house\_sold year’, which is replaced by ‘house age’ and ‘zip code’ for the base model 1.

Houseprice\_df.drop('month\_year', axis=1 ,inplace=**True**)

Houseprice\_df.drop('house\_sold\_year', axis=1 ,inplace=**True**)

Houseprice\_df.drop('zipcode', axis=1 ,inplace=**True**)



Input data was then split into 70:30 train/test to build and then fit the model.

**3.4 Base Model Generation: Linear Regression**

**Linear Regression** will be our base model for predicting housing prices. Then we’ll be using it as a benchmark to achieve better accuracy using other models.

Linear regression is a way to identify a relationship between two or more variables and use these relationships to predict values for one variable for a given value(s) of other variable(s). Linear regression assumes the relationship between variables can be modelled through a linear equation or an equation of a line. The variable, which is used in prediction is termed as independent/explanatory/regressor where the predicted variable is termed as dependent/target/response variable. Linear regression assumes that independent variables are related linearly to the response variable.

𝑦 = 𝑐 + 𝑚𝑥

In machine learning and regression literature above equation is used in the form:

𝑦 = 𝑤! + 𝑤"𝑥

Where w0 is intercept on y-axis, w1 is slope of line, x is an explanatory variable and y is the response variable.

**Results**

**Following are Linear Regression Performance parameters:**

Performance on training data using Linear Regression: 0.6583557916669898

Performance on testing data using Linear Regression: 0.6619022567965154

**R2 LR: 0.6619022567965154**

MSE: 40653577961.26467

MAE: 129141.33585612064

MAPE: 26.420582854891734

**Pros and cons of Linear Regression:**

**Pros-** Linear regression models are very simple and easy to implement. These models are said to be most interpretable.

**Cons-** Linear regression models are largely affected by the presence of outlier in training data. These models assume a linear relationship between target and explanatory variables which are sometimes is not true.

**3.5 Other models & Performance vis-à-vis the Base Model:**

For the purpose of this project, other models used for house price predictions are **Lasso Regression, Ridge Regression, Ada boosting, Gradient Boosting, Random Forest, KNN Regressor and Support Vector Regressor.**

### **Lasso Regression:**

### Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean.:

**Results**

Performance on training data using Lasso Regression: 0.6575327321071504

Performance on testing data using Lasso Regression: 0.6606042151552541

**R2 las: 0.6606042151552541**

MSE: 40809657196.10363

MAE: 129373.98243261973

MAPE: 26.42135270290683

### **Ridge Regression:**

### Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity.

**Results**

Performance on training data using Ridge Regression: 0.6583557687460491

Performance on testing data using Ridge Regression: 0.6618995534093641

R2 rid: 0.6618995534093641

MSE: 40653903022.17536

MAE: 129140.28379564053

MAPE: 26.420141332082046

**Ada Boosting:**

“Boosting” in machine learning is a way of combining multiple simple models into a single composite model.

Ada-boosting model combines multiple classifiers to increase the accuracy of classifiers. AdaBoost is an iterative ensemble method.

**Results**

Performance on training data using Ada-boosting: 0.2555089836520954

Performance on testing data using Ada-boosting: 0.23922273417078643

**R2 AD: 0.23922273417078643**

MSE: 91477445529.5061

MAE: 270636.53732493066

MAPE: 68.65769166792306

### **Gradient Boosting:**

A Gradient Boosting combines the predictions from multiple decision trees to generate the final predictions.

**Results**

Performance on training data using Gradient-boosting: 0.8693122149954108

Performance on testing data using Gradient-boosting: 0.8517790439141942

**R2 GB: 0.8517790439141943**

MSE: 17822396969.09453

MAE: 79485.73126554099

MAPE: 15.233207786883064

### **Random Forest:**

### Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.

**Results**

Performance on training data using Random Forest: 0.9802712811283774

Performance on testing data using Random Forest: 0.8668975068697298

**R2 RF: 0.8668975068697298**

MSE: 16004521444.124107

MAE: 71650.96774879408

MAPE: 13.515210599042039

### **KNN Regressor:**

### KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

**Results**

Performance on training data using Random Forest: 0.9802712811283774

Performance on testing data using Random Forest: 0.8668975068697298

**R2 RF: 0.8668975068697298**

MSE: 16004521444.124107

MAE: 71650.96774879408

MAPE: 13.515210599042039

### **Support vector regressor:**

### Support vector regressor classifies the new point depending on whether it lies on the positive or negative side of the hyperplane depending on the classes to predict.

**Results**

Performance on training data using Support vector regressor: -0.06404950544757315

Performance on testing data using Support vector regressor: -0.060654196144387384

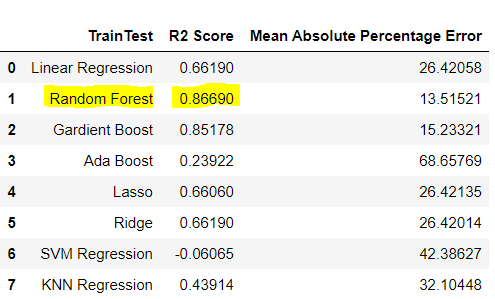
**R2 SVM: -0.060654196144387384**

MSE: 127535273215.19827

MAE: 218035.25094781036

MAPE: 42.38626825638349

**Models Scores & Errors have been tabulated below:**

****

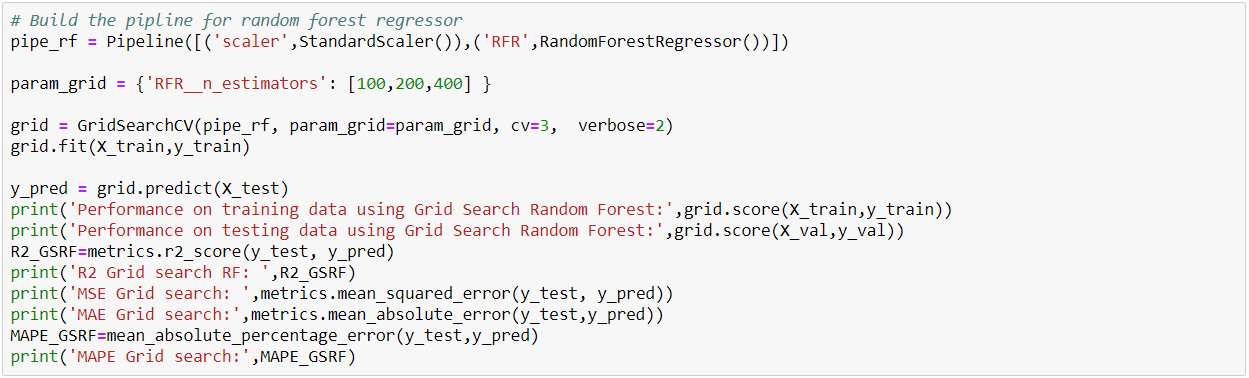
Based on the above comparison, we can see that the base model, ie. Linear regression has R2 score of 0.66, whereas other models, such as Gradient Boosting & Random Forest model, outperform.

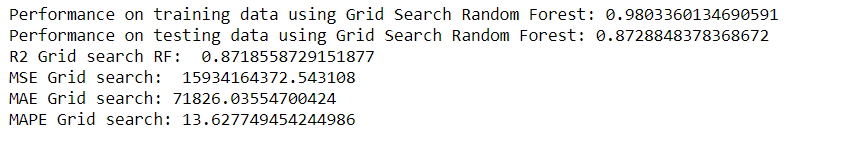
So, we will choose Random forest as the best Model with a maximum R2 score of 0.86 and a low Mean absolute percentage error of 13.5

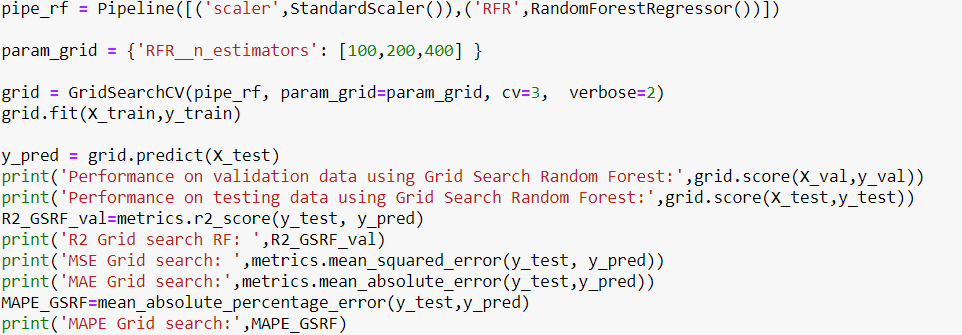
**3.6 Hyperparameter Tuning & Final Model Selection:**

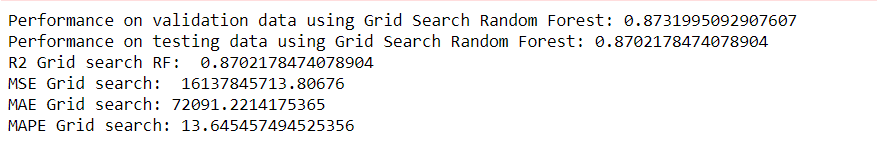
Machine learning models have hyperparameters that can be set to customize the models and improve their performance. The result of a hyperparameter tuning is a single set of well-performing hyperparameters that can be used to configure the model, thus increasing its overall performance. Hyperparameter tuning is required to get the most out of machine learning models.

***Grid Search for Regression- Hyperparameter*** was used, which affected the overall model accuracy.

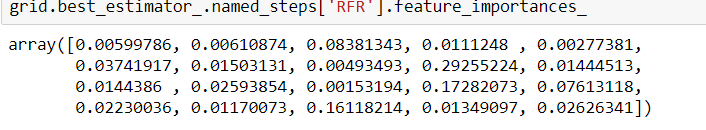




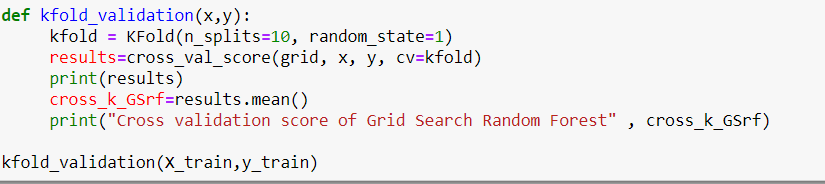


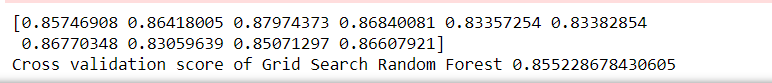


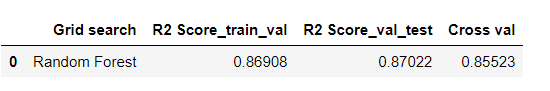
**Feature importance from pipleline**



***K-Fold cross-validation for Random forest***







***Observations: the relationship between y\_test and y\_pred***

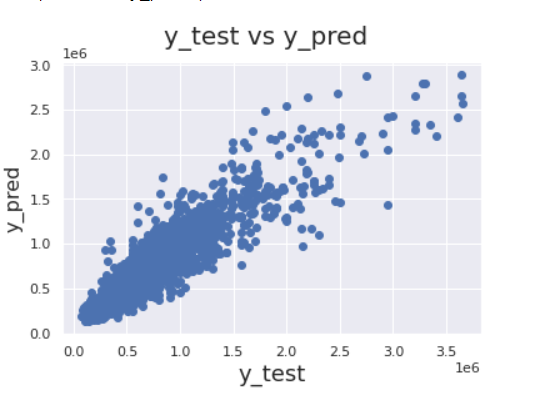
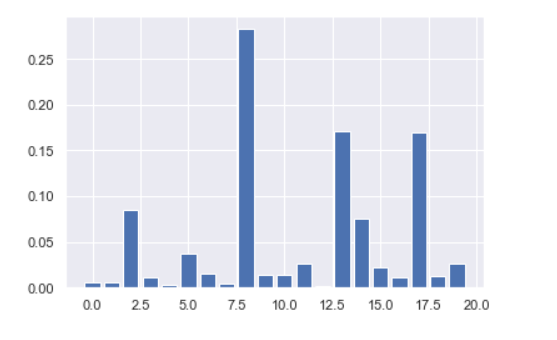


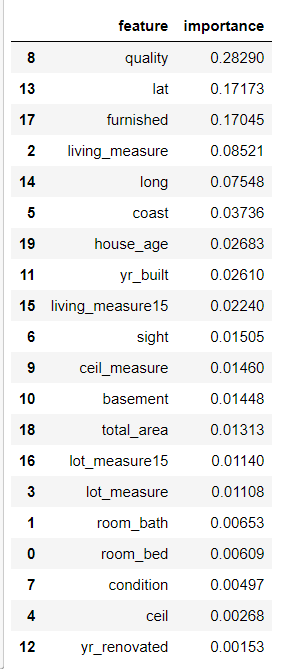
Figure 3.1: House value test versus prediction scatter plot.

* After analyzing variables in EDA, Price (Target Variable) have a linear relationship with living\_measure, living\_measure15 and ceil\_measure
* We are able to get almost horizontal line can be an indication of linear relationship ,thus taking Random forest regression algorithm linear regression seems to be the best option.
* Hyperparameter Tuning improves the accuracy of the model.

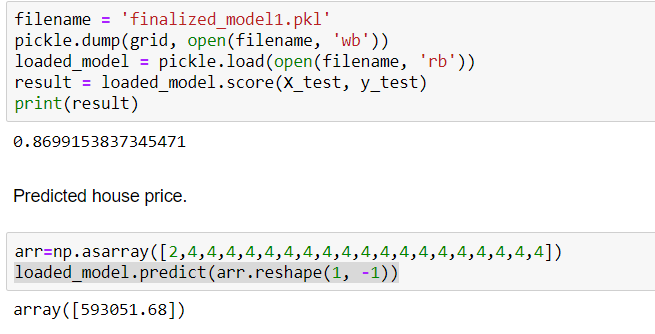
**3.7 Feature Importance**







**3.8 Saving the model and Prediction**



**3.9 Implication & Application of the Machine Learning Model**

The most significant aspect of any machine learning model is the unbiased and a variety of data-driven approach that is repeatable. It can be used as a benchmark for buyers and sellers alike. Real estate agents can also get their groundwork done using such house valuation models.

Parameter selection can be another important finding out of such models where the relation between the house valuation versus independent variables can be adjudged. Such models can be widely used as methods for both market survey and appraisal phases so that the buyers can judiciously spend their time on selective properties.

**4.0 Conclusion and Reflection:**

The main objective of this machine learning project is building a mechanism based on the given set of variables to generate an optimum prediction on the houses that gives buyers and sellers fair confidence independent of real estate agents.

We’ve used 21 variables in the exploratory data analysis after necessary data curation to analyze their impact on house price. We’ve also listed the hierarchy of parameters supposedly affecting the house price were then listed.

These feature set were then given as an input to the different regression modelling algorithms to predict house prices. We’ve calculated the performance of each model using different performance metrics and compared them based on these metrics Based on the above comparison, we can see that Gradient Boosting Regression model and Random Forest model outperforms all the other regression models. So, we will choose **Random forest** as the best Regression Model for this problem, **at the time of preparation of the preliminary report**.

**4.1 Closing Remarks and Scope for Improvement:**

* Every model like this price prediction can be improved by adding more attributes like surroundings, marketplaces and many other related variables and amenities related to the houses (e.g., swimming pool, energy efficient houses, Developers) by studying market trends and published data.
* The predicted data can be stored in the databases and an app can be created for the people so they would have a sound basis before they invest the money when taking a big decision involving hundreds of thousands of dollars.
* The model can be further enhanced if a real-time data the data be fed into it and necessary adjustments can be made that will keep the model relevant to the changing real estate market.
* More domain knowledge itself can be a definite advantage during data curation or while selecting or dropping parameters may ensure a better training versus test results to ensure higher level of confidence as the model goes into production.
* A larger dataset could be another factor that could improve the model quality as well. Last but not the least is the possibility to use other methods and deep-learning techniques like Artificial Neural Network that could prove useful in generating more accurate models. House price data range over a longer period would be very useful for data analysis, further insights and robust model building.

**APPENDIX: List of Figures**

Figure 1.1: Process Flowchart

Figure 2.1: 5-point summary of the dataset

Figure 2.2: Box plot of all house price

Figure 2.3: Box plot of all attributes prior to outlier removal

Figure 2.4: Skewness of all attributes prior to outlier removal

Figure 2.5: Outlier removal of price

Figure 2.6: Box plot of all attributes after outlier removal

Figure 2.7: Skewness of all attributes after outlier removal

Figure 2.8: Sale time vs. price

Figure 2.9: Bedroom vs. price

Figure 2.10: Living Measure vs. price

Figure 2.11: Interrelationship among different area measures

Figure 2.12: Renovation vs. price & area measures

Figure 2.13: Furnished houses vs. price

Figure 2.14: Condition vs. Price

Figure 2.15: Coast & Sight vs. Price

Figure 2.16: Quality & furnished vs. Price

Figure 2.17: House Age vs. Price

Figure 2.18: Longitude- Latitude vs. Price

Figure 2.19: Correlation matrix

Figure 3.1: House value test versus prediction scatter plot.