**Project Notes-1 on**

***House Price Prediction***

***Submitted to***



**Great Learning Olympus**

***Prepared by***

**Mrinmoy Majumdar**

**June-D Batch**

**Post Graduate Program in Data Science & Business Analytics**

**From**

****

UT Austin

**June, 2022**

TABLE OF CONTENT

**1a. Defining problem statement** ……………………………………………….4

**1b. Need of the study/project………………………………….**……………….4

**1c. Understanding business/social opportunity………………………………...**4

###### **2a. Understanding how data was collected in terms of time, frequency and methodology**…………………………………………………………………....5

**2b. Visual inspection of data (rows, columns, descriptive details) …………...**.6

2c. Understanding of attributes (variable info, renaming if required) ………....8

3a. Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones) …………………………9

3b. Bivariate analysis (relationship between different variables, correlations) .21

3c. Removal of unwanted variables (if applicable) …………………………...27

3d. Missing Value treatment (if applicable) ………...…………………………27

3e. Outlier treatment (if required) …………………………………………….29

3f. Variable transformation (if applicable) ……………………………………31

3g. Addition of new variables (if required) ……………………………………35

4a. Is the data unbalanced? If so, what can be done? …………………………37

4b. Any business insights using clustering (if applicable) ……………………37

4c. Any other business insights ……………………………………………….38

**LIST OF FIGURES**

|  |  |
| --- | --- |
| **Images** | **Page No** |
| **Shape of dataset** | **6** |
| **Columns of dataset** | **6** |
| **Dataset information** | **8** |
| **Count plot** | **10** |
| **Boxplot** | **14** |
| **Histogram** | **19** |
| **Regression plot** | **22** |
| **Point plot** | **24** |
| **Correlation plot** | **24** |
| **Scatter plot** | **25** |
| **Pair plot** | **27** |
| **Null value presence** | **28** |
| **WSS plot** | **38** |

**LIST OF TABLES**

|  |  |
| --- | --- |
| Table 1 | 6 |
| Table 2 | 6 |
| Table 3 | 7 |
| Table 4 | 27 |
| Table 5 | 33 |
| Table 6 | 35 |
| Table 7 | 36 |
| Table 8 | 36 |

**1)Problem Understanding**

**a) Defining Problem Statement**

Every person resides in a house but due to some circumstances they have to sell their house. The owner of the house would quote the right price that is based on a number of factors. Real estate housing company has been tasked with the responsibility to analyse the various attributes and determine the price accordingly.

From a buyer’s perspective they would be interested in buying the house by taking a look at the property and its neighbourhood and whether it suits their daily requirements. It is not expected that they would agree with the price quoted right away. They will perform a thorough research on the recently sold properties in the area before bargaining on the price.

Data that is available includes information related to number of bedrooms and bathrooms, measurements of the house, sold, built and renovation years, proximity to coast, number of times property has been viewed and property location.

Explanation of data fields available in Data Dictionary, 'innercity.xlsx'

**b) Need of the study/project**

**The project is required from a learner’s standpoint to apply the concepts covered in the course in a real-life environment to solve problems in different domains. The objective of the study would be to predict the correct prices of residential properties by finding out the important features. As a data analyst it is our responsibility to predict an unbiased and fair price so as to ensure that both buyers and sellers are comfortable. House prices are on the rise every year and there is a need for a universal law to standardise the prices in the future.**

**c) Understanding business/social opportunity**

Real estate is a sophisticated business model whose main motive is to help customers provide suitable properties and associated appropriate price tag. They also act as a communicative bridge between sellers and buyers and maintains customer satisfaction if they are supplied with the right property and price. A positive customer feedback and word of mouth would yield profits for the business in the future. Poor experience and compromise on honesty will be an opportunity for customers to look for alternatives that matches their expectations. The industry is a major contributor to the economy and employment in the country. Hence, resolving house price problem is a great service to the nation.

**2)Data Report**

**a) Understanding how data was collected in terms of time, frequency and methodology**

Data is provided in the form of an excel sheet named as inner city by the course platform. The feature variables available are provided as below –

cid: a notation for a house

dayhours: Date house was sold

price: Price is prediction target

room\_bed: Number of Bedrooms/House

room\_bath: Number of bathrooms/bedrooms

living\_measure: square footage of the home

lot\_measure: square footage of the lot

ceil: Total floors (levels) in house

coast: House which has a view to a waterfront

sight: Has been viewed

condition: How good the condition is (Overall)

quality: grade given to the housing unit, based on grading system

ceil\_measure: square footage of house apart from basement

basement\_measure: square footage of the basement

yr\_built: Built Year

yr\_renovated: Year when house was renovated

zipcode: zip

lat: Latitude coordinate

long: Longitude coordinate

living\_measure15: Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area

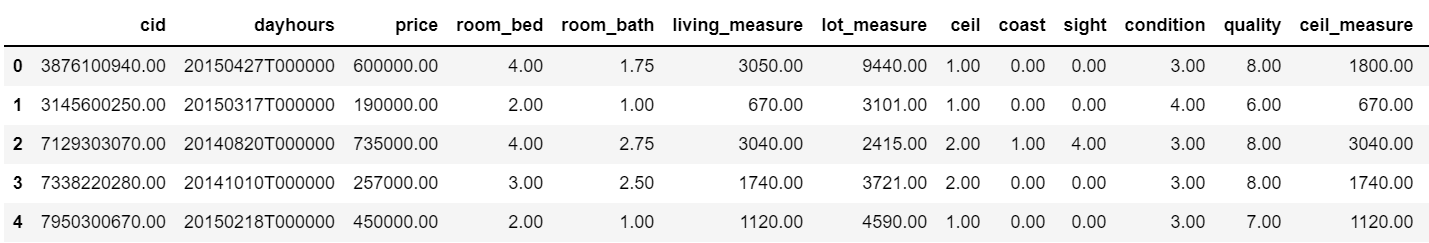
lot\_measure15: lotSize area in 2015(implies-- some renovations)

furnished: Based on the quality of room

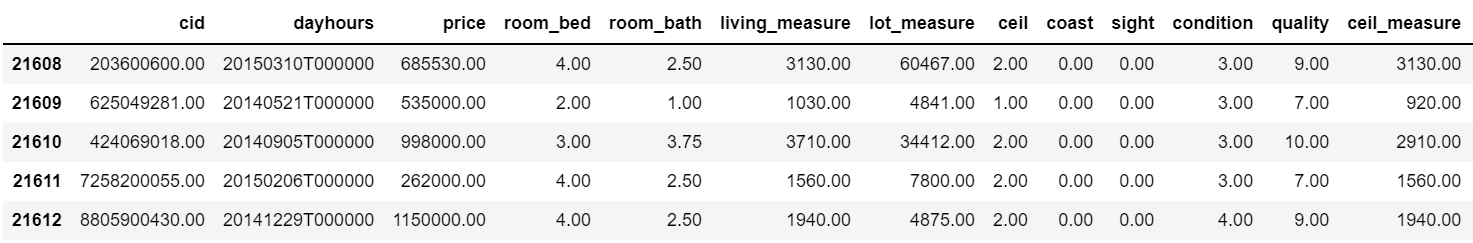
total\_area: Measure of both living and lot

The time period ranges from 2014 to 2015 over which the houses have been put up for sale and the corresponding prices quoted for the same. The variables are filtered to find their range, missing values and presence of special characters. Columns not required for analysis are determined from the set of values in the dataset. Numerical and categorical data types and probable outliers are identified based on the values available and the model type is also determined by having an in-depth look at the problem and data in hand. The count of each of the columns can also be figured out and total records are 21613.

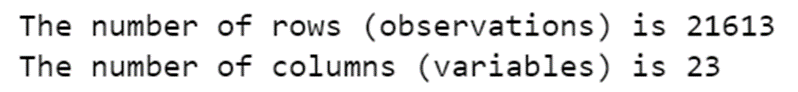
**b) Visual inspection of data (rows, columns, descriptive details)**



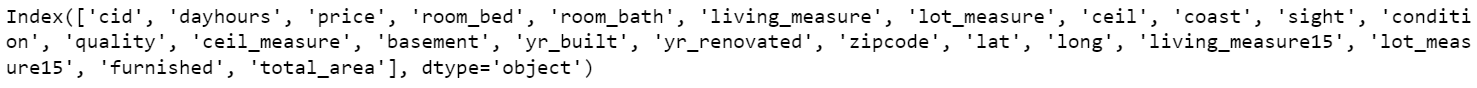
**Table 1: Head of dataset**



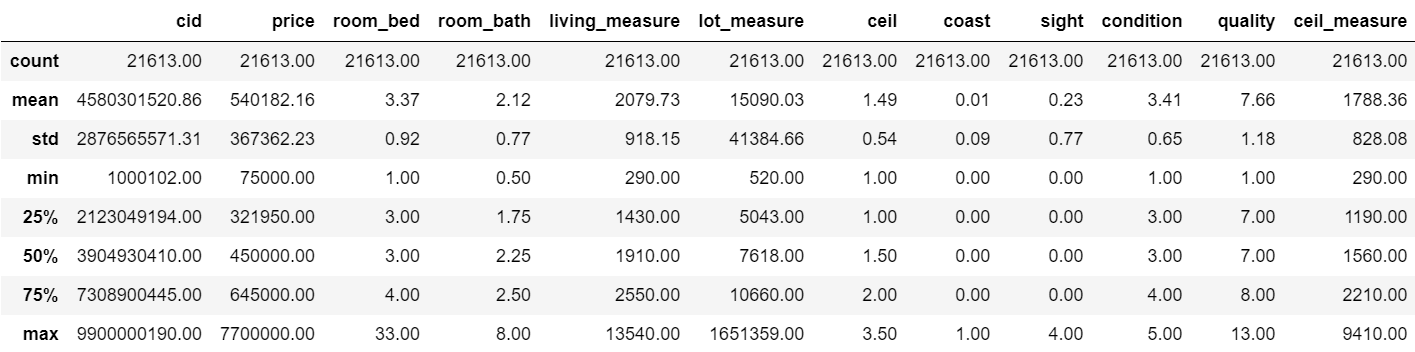
**Table 2: Tail of dataset**

****

**Figure 1: Shape of dataset**



**Figure 2: Columns of dataset**

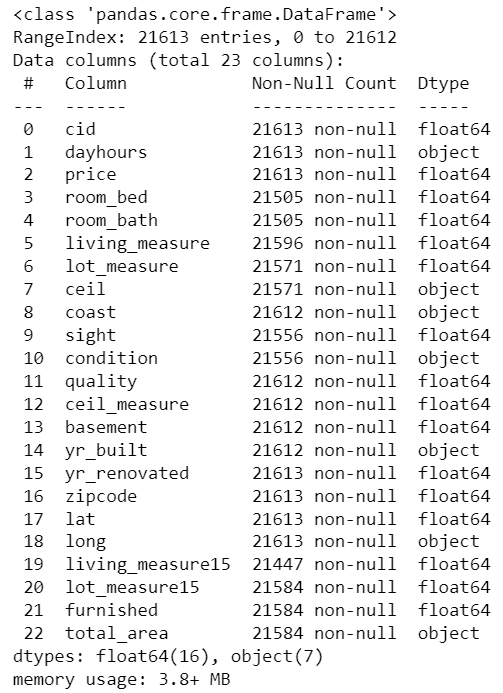


**Table 3: Description of dataset**

5-point summary of the numerical features are shown after removing null values and the inferences are as below –

* cid will not be useful for analysis so its description is meaningless from problem standpoint
* price distribution seems to be right skewed as mean > median
* room bed distribution is also right-skewed since mean > median
* room bath follows left skewed distribution due to mean < median
* Since mean > median, living measure is right skewed
* lot measure is highly skewed to the right
* ceil is normally distributed with mean and median having identical values
* coast is a categorical variable with binary values
* sight is positively skewed
* condition has values ranging from 1 to 5 with 1 indicating poor and 5 signifying the best condition of the house
* ceil measure, basement, zipcode, long, living measure15, lot measure15 and total area follow positively skewed distribution
* yr. built and lat follow negatively skewed distribution
* furnished and yr. renovated are categorical variables with 2015 being the year when house was renovated while majority of the houses have no renovation done till date
* Presence of outliers can also be suspected based on the wide range of values

**c) Understanding of attributes (variable info, renaming if required)**



**Figure 3: Dataset information**

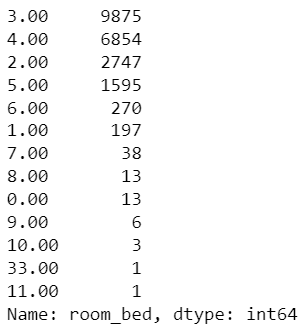
The dataset contains 23 features having a mixture of numerical and categorical datatypes. Total area is a measurement type of data which cannot be of object type and needs correction. Long also should be of integer/float datatype due to the location coordinate associated with it. Count of non-null values for all variables are not universal with total of 21613 records present which shows missing values and treatment for the same is a necessity. Duplicate records are absent from the given data.

We have chosen not to rename any columns as all the features are in line with the problem statement document provided.

**3) Exploratory data analysis**

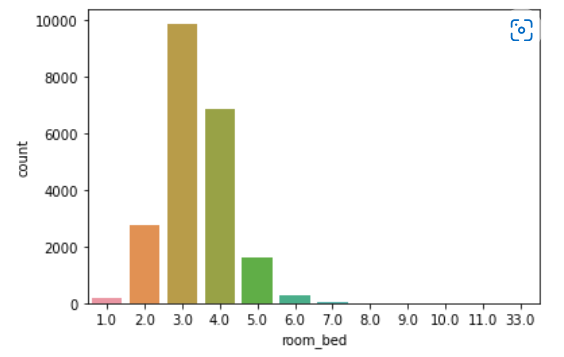
**a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)**

The analysis is performed with the help of different plots post missing value treatment and pre outlier removal. The motive of this exercise is to come up with inferences and plan the next steps of action.



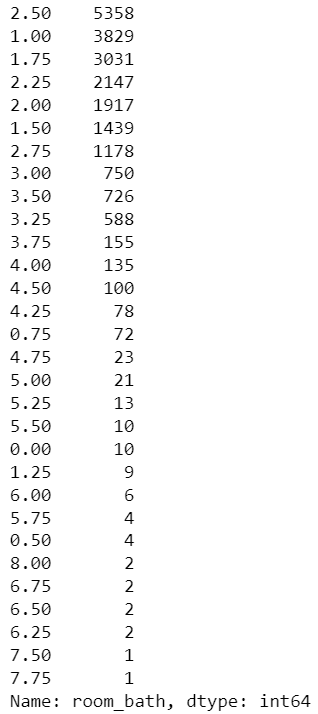
**Figure 4: Room bed count of values**

It is not possible to have a house with no bedroom and is treated similar to missing values. House with 33 beds suggests presence of outliers.



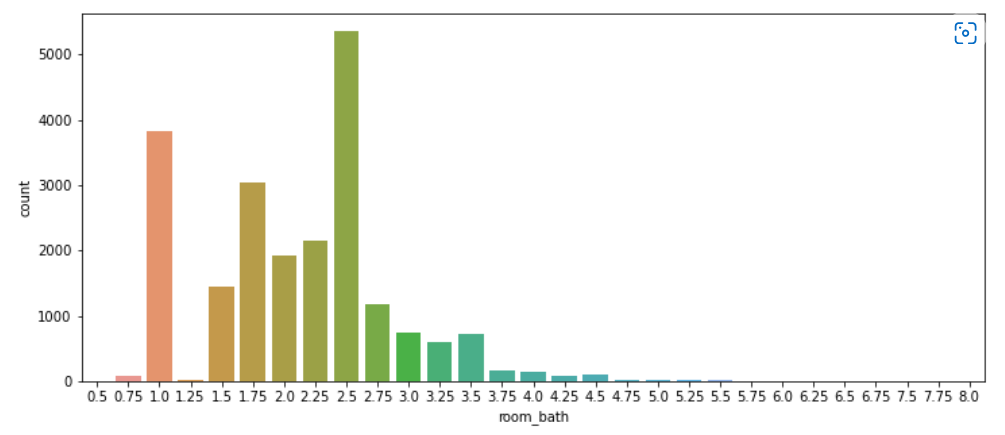
**Figure 4: Room bed count plot**

Majority of the houses consist of 3 bedrooms.



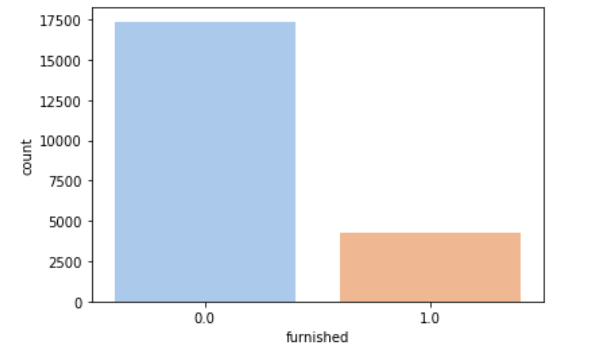
**Figure 5: Room bath count of values**

Room bath also has 10 values with 0 values which is practically not feasible in real world and treated in the same way as for room bed.



**Figure 6: Room bath count plot**

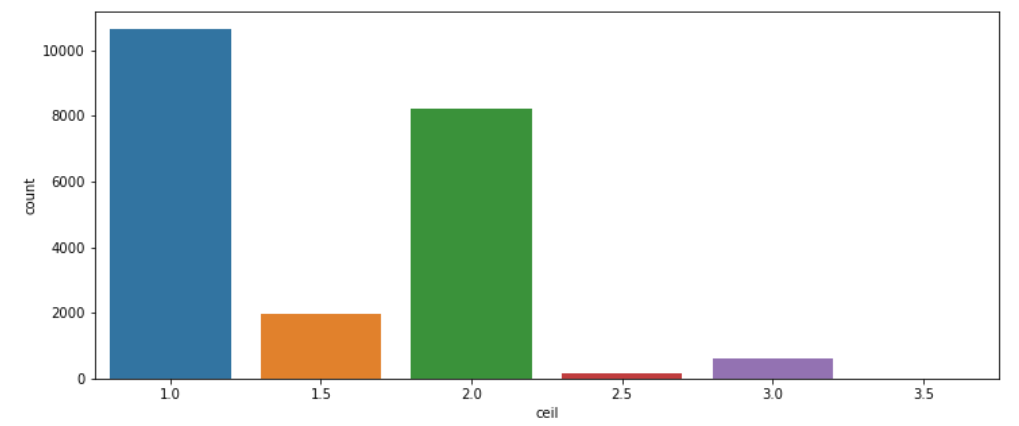
Customers prefer maximum 2 bathrooms or more based on their family size and requirements.



**Figure 7: Furnished count plot**

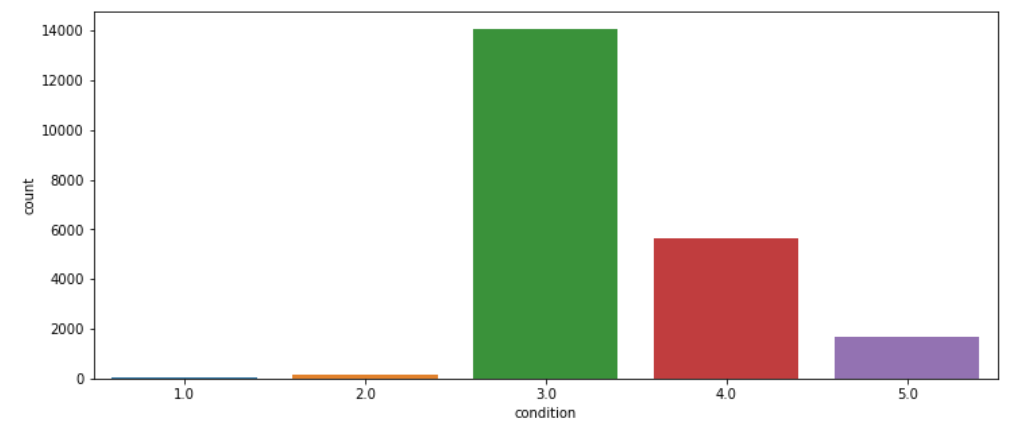
Most of the houses put up for sale are not furnished which may impact the prices going ahead considering the factor that buyers prioritize furnished residents for living.

The count of cid shows that though majority of the properties have been bought/sold once but there are 176 properties which have been put up for sale repeatedly in the past.



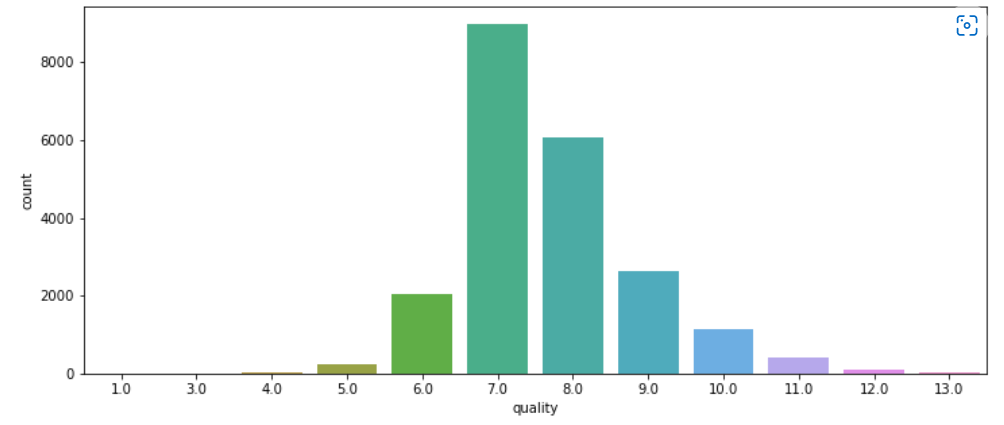
**Figure 8: Ceil count plot**

Majority of the houses have maximum 1 floor which may represent a nuclear family falling under the category of medium income group.



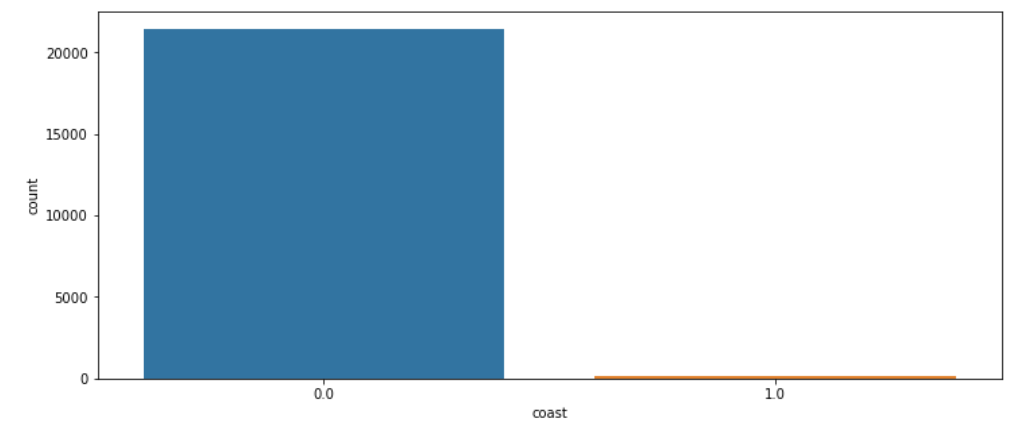
**Figure 9: Condition count plot**

Most of the buyers or real estate agents after looking at the houses have decided to give an average rating of 3 that could certainly influence potential buyers if shared on an online portal.



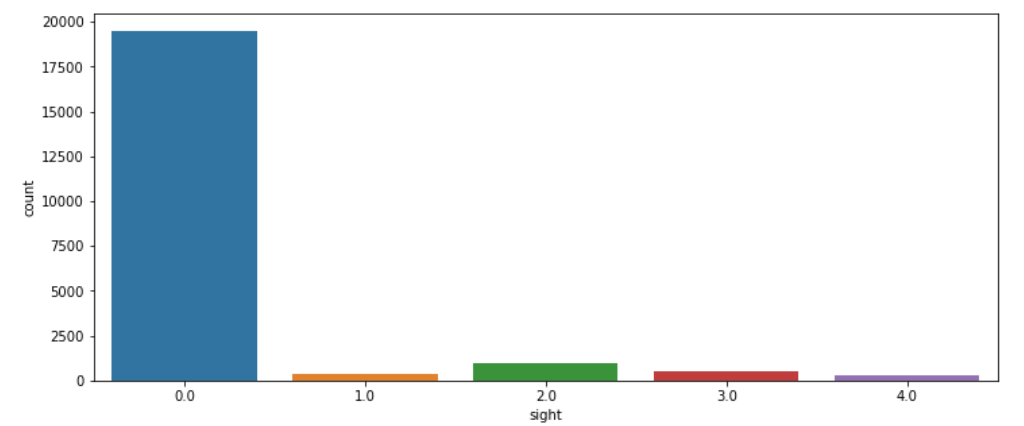
**Figure 10: Quality count plot**

13 is assumed to be the highest grade and 1 the lowest among the lot. Average grade of 7 have been given to housing unit that means more improvement needs to be done by taking feedback on the desired changes and implementing them to expect better prices. Only 13 properties have been able to attain the best grades.



**Figure 11: Coast count plot**

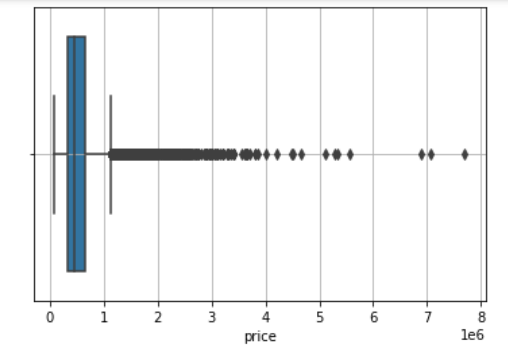
Houses mostly are not near to any waterbody which might be due to soaring prices of those in proximity to coast.



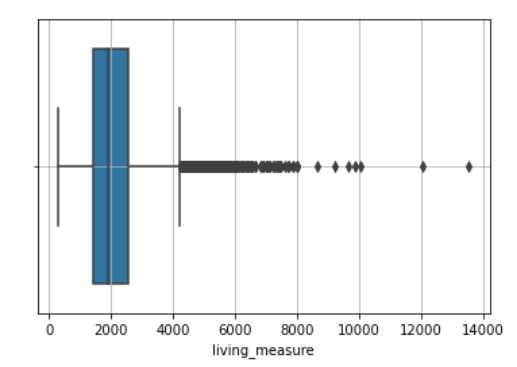
**Figure 12: Sight count plot**

Maximum customers are yet to view the property which might be causing delay in the process of finalizing the sale price.

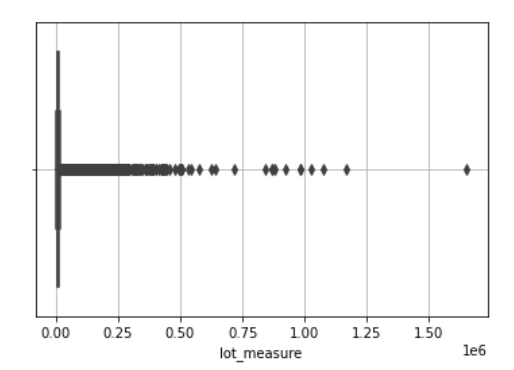
Most of the residents have no basement with 8487 properties having basement the reason for which may be huge cost of construction or lack of requirement or interest in the same.



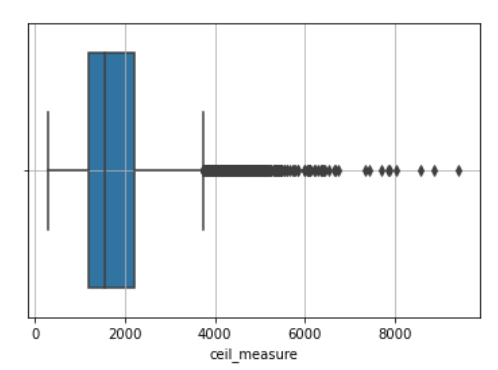
**Figure 13: Boxplot on price**



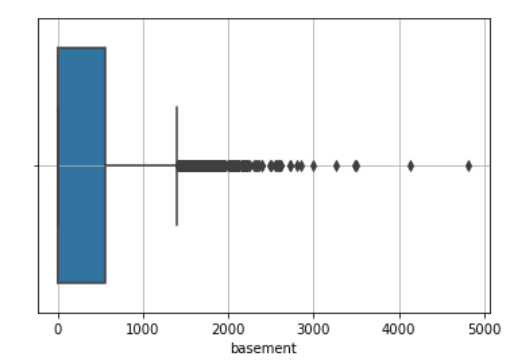
**Figure 14: Boxplot on living measure**



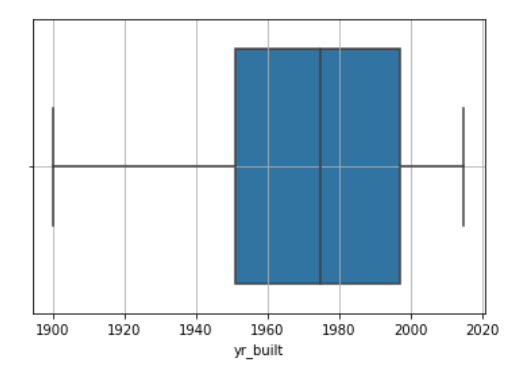
**Figure 15: Boxplot on lot measure**



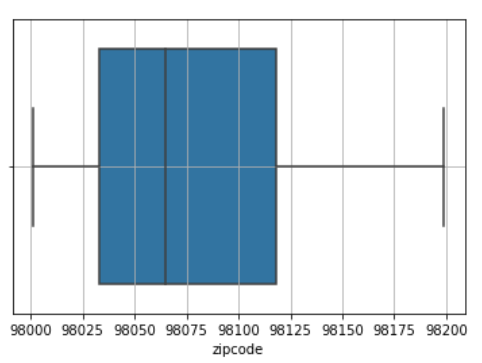
**Figure 16: Boxplot on ceil measure**



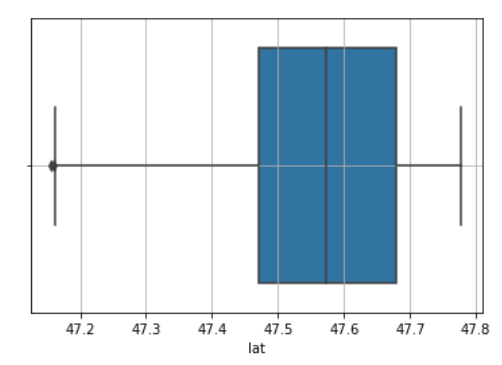
**Figure 17: Boxplot on basement**



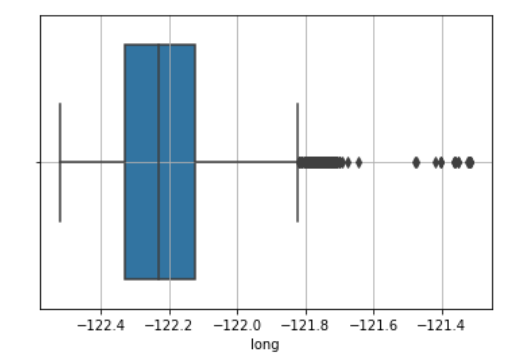
**Figure 18: Boxplot on yr. built**



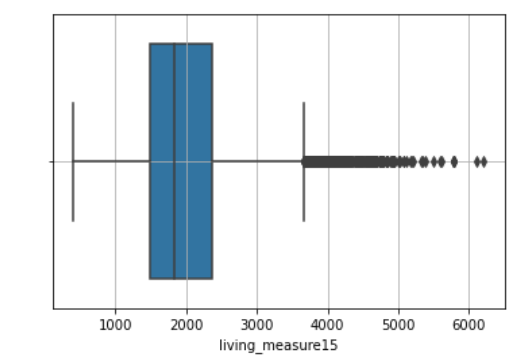
**Figure 19: Boxplot on zip code**



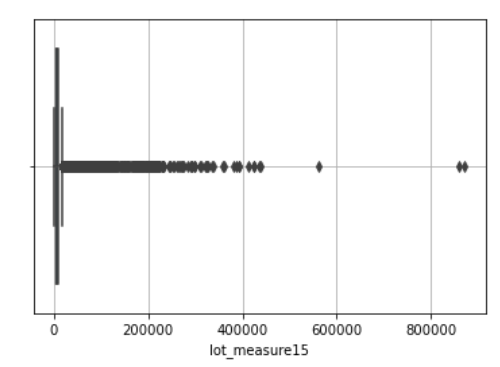
**Figure 20: Boxplot on lat**



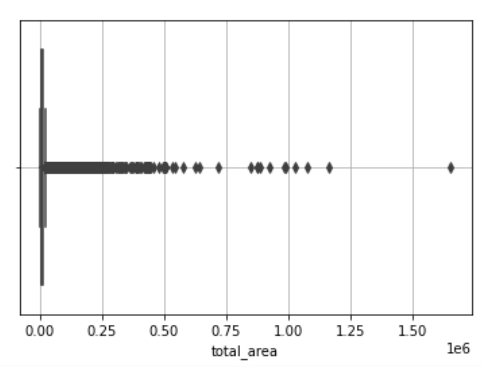
**Figure 21: Boxplot on long**



**Figure 22: Boxplot on living measure15**



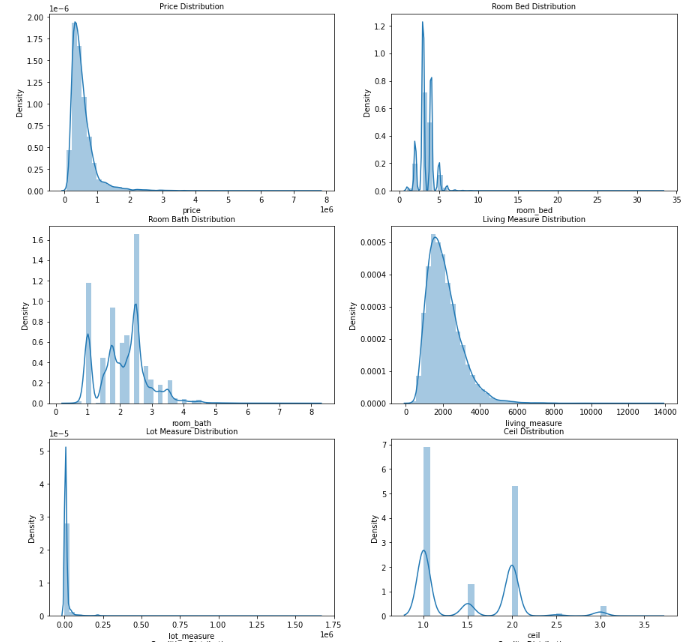
**Figure 23: Boxplot on lot measure15**

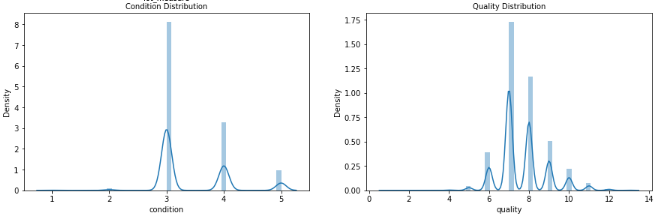


**Figure 24: Boxplot on total area**

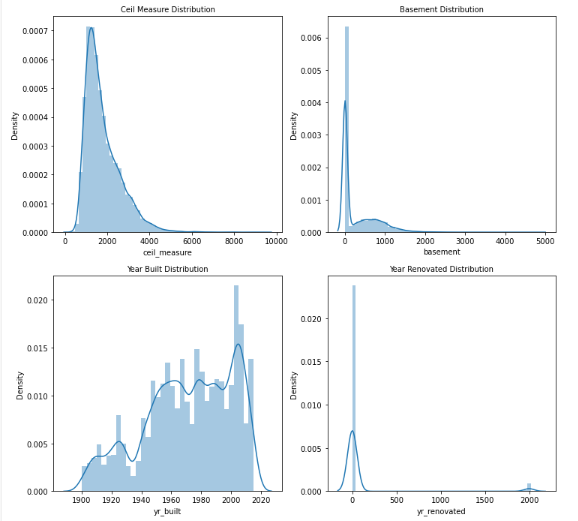
Inferences –

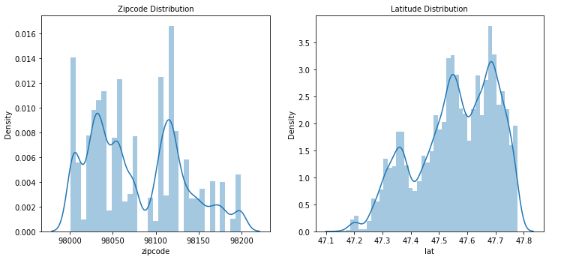
* Price, living measure, lot measure, ceil measure, basement, long, living measure15, lot measure15 and total area have significant outliers present
* yr. built, zip code and lat do not have outliers
* Outlier treatment will be done based on relation with target variable



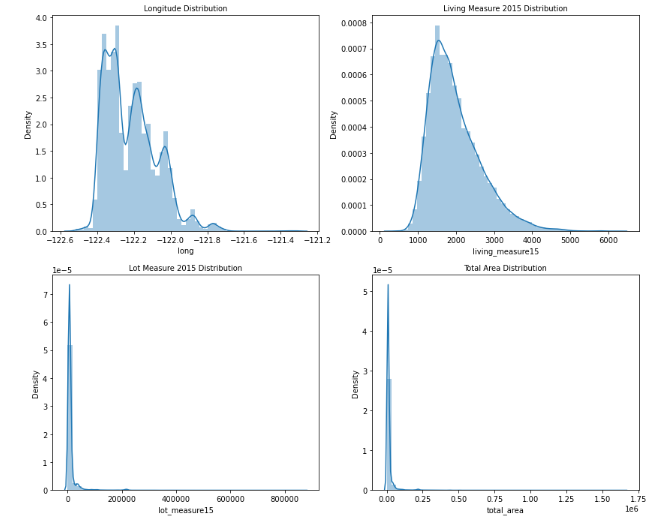


**Figure 25: Histogram distribution part I**





**Figure 26: Histogram distribution part II**



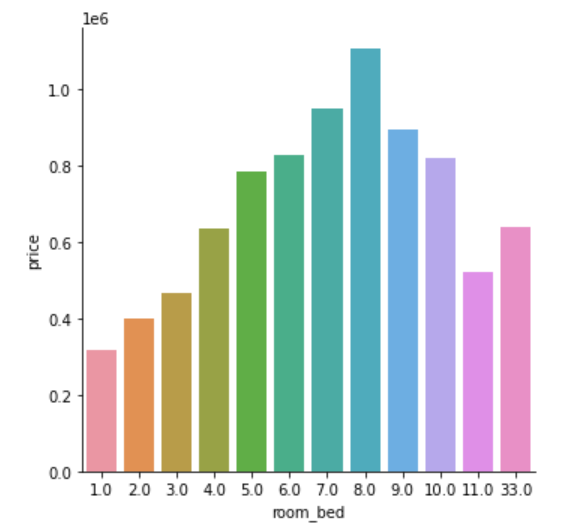
**Figure 27: Histogram distribution part III**

Observations –

* Skewness shown in the histogram of different features is same as concluded from description summary of data
* yr. renovated has a peak at 0 denoting houses not renovated
* Many columns categorical in nature show multiple peaks indicating multi-modal distribution

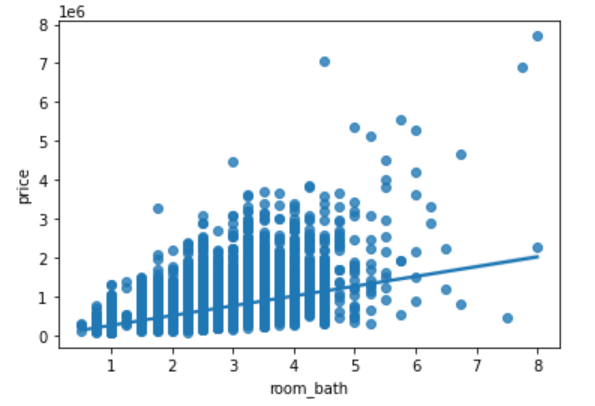
**b) Bivariate analysis (relationship between different variables, correlations)**

Bivariate analysis deals with trying to establish a relationship between two different columns with the help of various plots provided by seaborn and matplotlib libraries of python. Price is the target variable and its relation with remaining variables will be viewed with the intention of planning the next course of action.



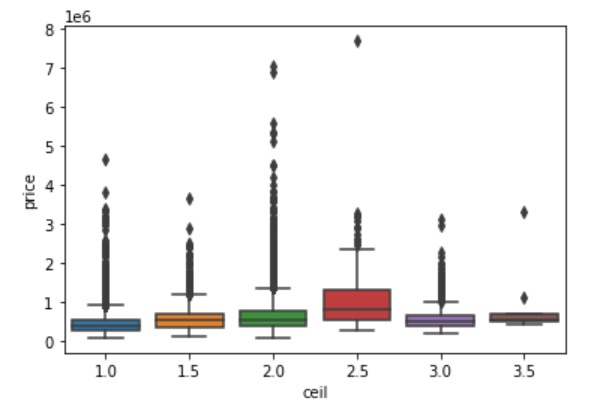
**Figure 28: Price vs room bed**

Houses with 8 bedrooms have the highest price while 1-bedroom residents are yielding low prices.



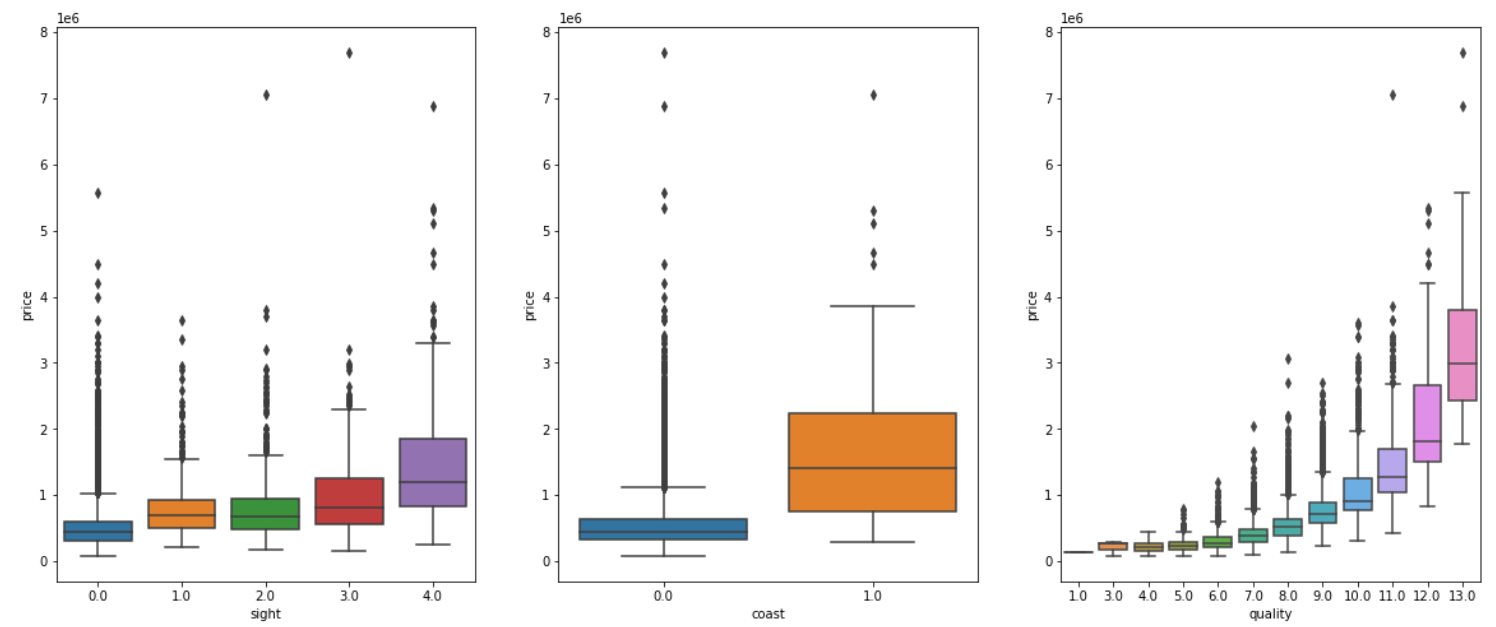
**Figure 29: Price vs room bath**

The observation points seem to be universally spread for 50% of room bath data while it fizzles out for the rest. Prices are increasing with rise in bathrooms and the regression line is a proof of the statement.



**Figure 29: Price vs ceil**

Average prices are quite similar for floors in the range of 1-2. The price slightly rises for floors greater than 2. The outliers are acceptable in the data due to its categorical nature.

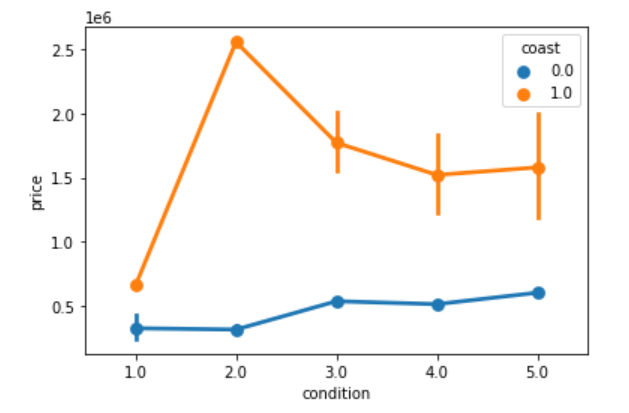


**Figure 30: Price vs three categories**

Median prices are highest for houses visited maximum number of times by buyers.

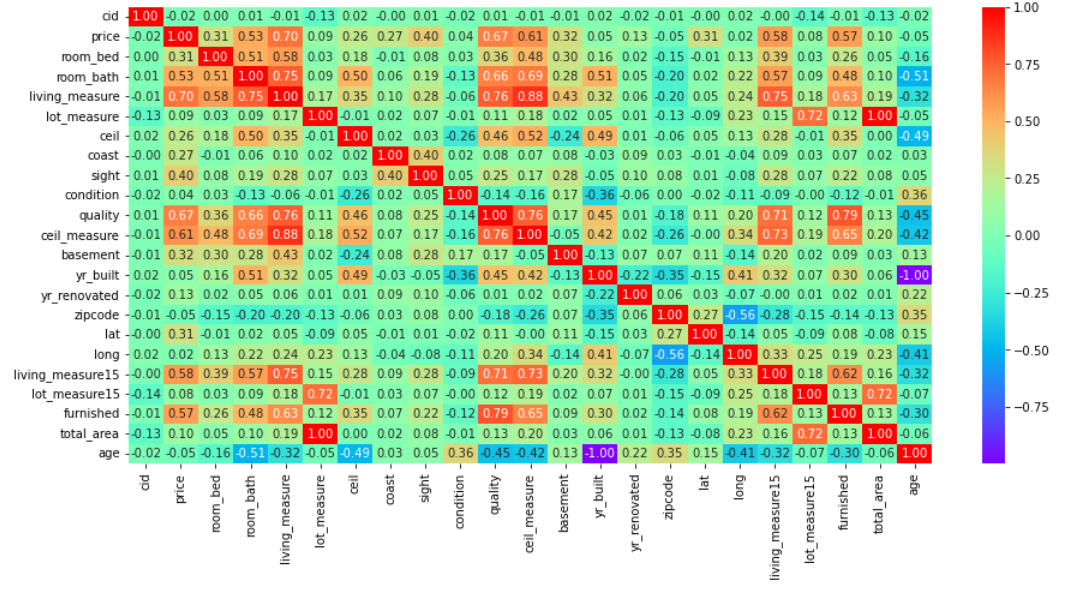
Sea facing houses are more likely to get sold and yield handsome returns for the sellers.

Customers are willing to select highly rated houses and thereby giving maximum prices for the same.



**Figure 31: Point plot with multiple variables**

House condition with no waterside view is not influencing prices whereas house with condition rating of 2 and coast view are experiencing a price shoot up is a matter of research.



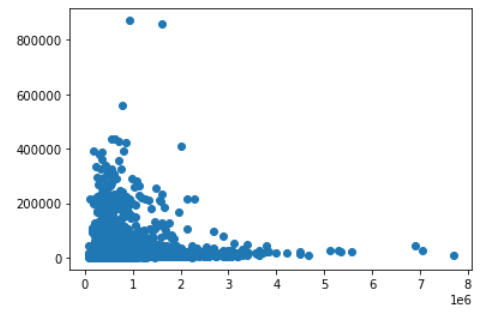
**Figure 32: Correlation plot**

Cid has poor relation with price as well as other variables and it is better to get rid of it before model building.

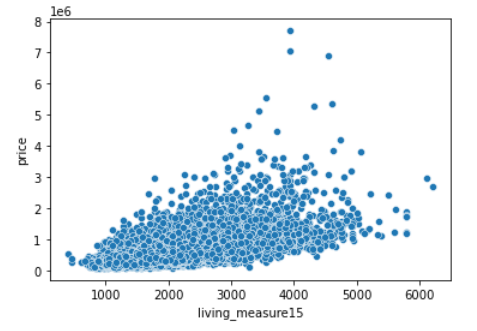
Price has strong correlation with room bath, living measure, quality, ceil measure, living measure15 and furnished as it ranges from 0.5 to 0.75.

Some variables are showing negative correlations which may be due to the fact that it was performed prior to outlier removal.

Multicollinearity behaviour is also observed for few variables like total area and lot measure, ceil measure and living measure, quality and furnished and quality and ceil measure which display significant correlation and either of these related variables need to be dropped post checking their VIF values.

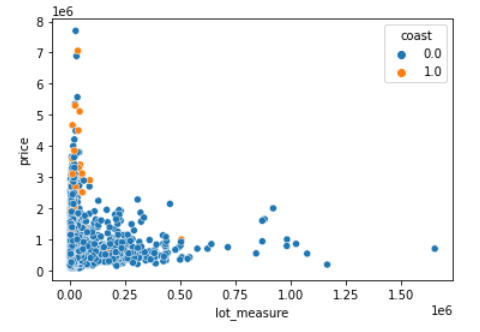


**Figure 33: Price vs lot measure15**



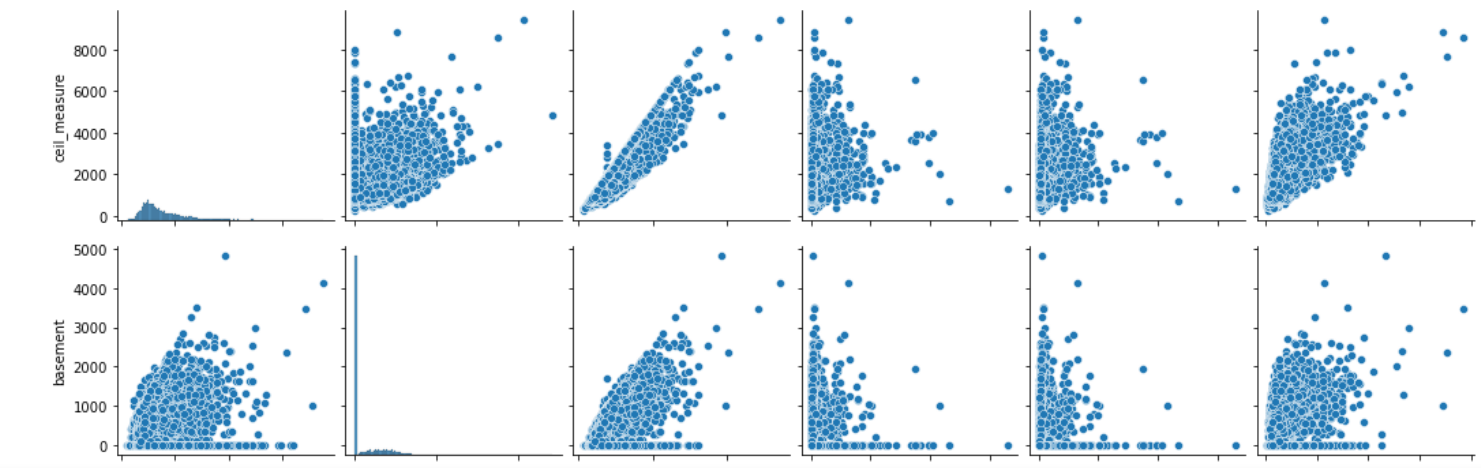
**Figure 34: Price vs living measure15**

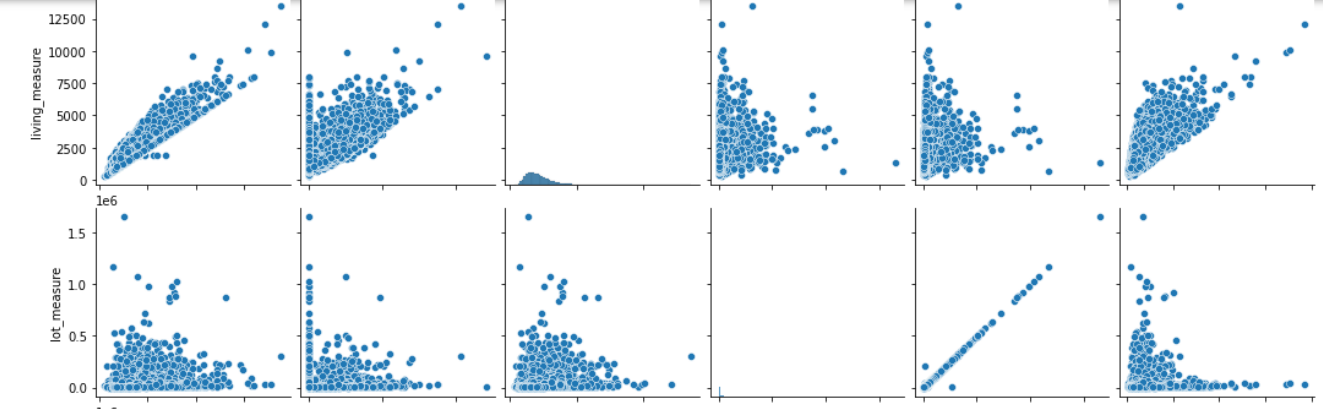
Scatterplot shows cloud like relationship between price and living measure15 and lot measure15 and no conclusion can be made out of these plots.

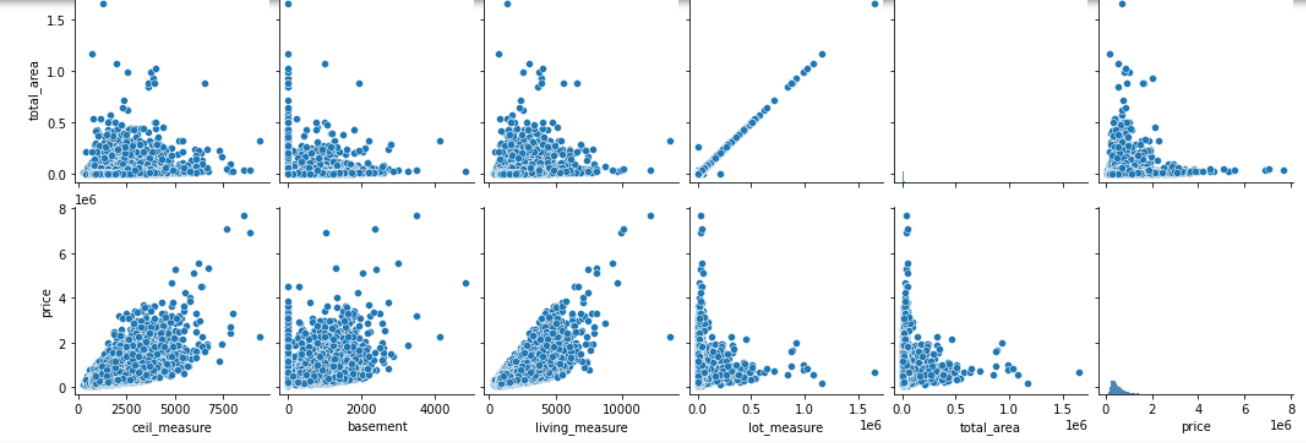


**Figure 35: Multivariate scatter plot**

Initial lot measure is densely populated with very minute coast view observations.





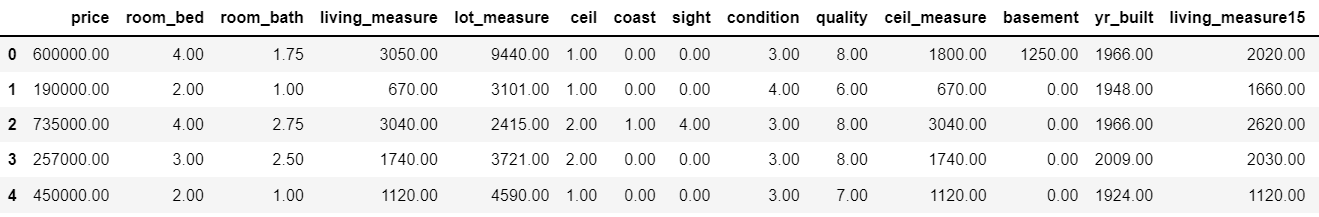


**Figure 36: Pair plot**

The numerical variables in pair plot are having linear relationship with target variable and hence could be important factors in model design going forward.

**c) Removal of unwanted variables (if applicable)**

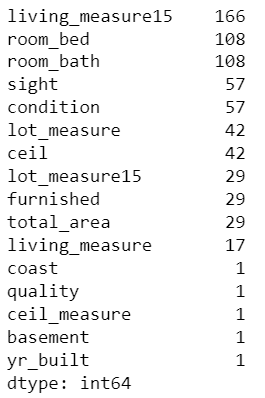
Cid, day hours, zip code, lat, long and yr. renovated were dropped from the analysis as they have no relation with price.



**Table 4: Head of dataset after dropping columns**

**d) Missing Value treatment (if applicable)**

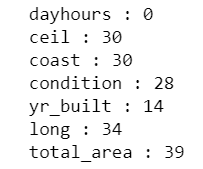
There is total 689 missing values in the dataset. Column wise null value presence are shown below –



**Figure 37: Null value count per column**

Percentage wise null value distribution makes it clear that we need to impute with median and mode values respectively for numerical and categorical fields.

Few columns are inappropriately labelled as object due to presence of special character “$” in the features and are displayed –

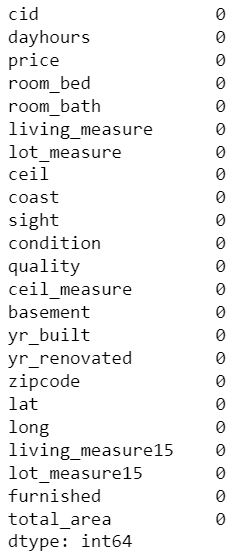


**Figure 38: Special character count**

Ceil, long and total area special characters are replaced with median while the remaining are replaced with mode values.

Living measure15, room bed, room bath, living measure, lot measure, lot measure15, total area, basement, ceil measure, quality, condition, coast, sight and furnished null values are imputed with median values.

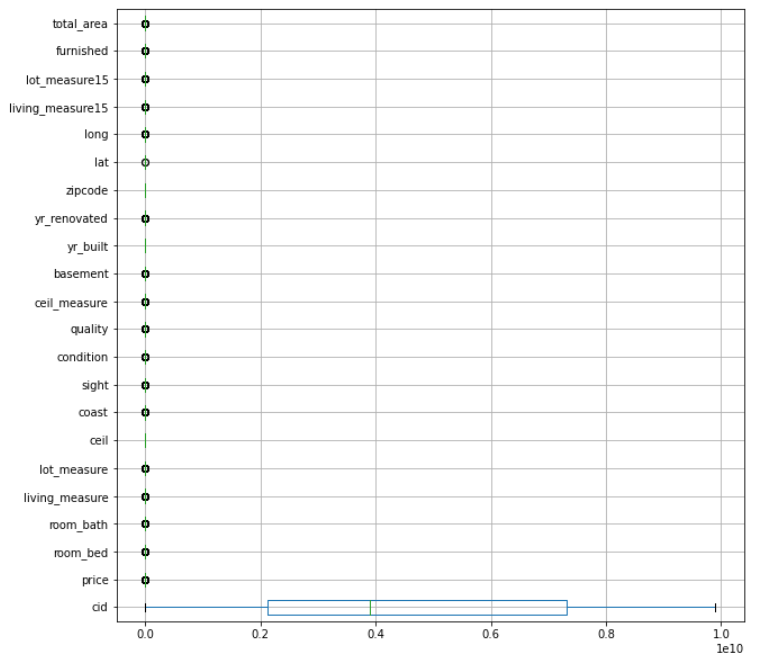
Ceil is imputed with median and yr. built with mode. Here we conclude our missing value treatment.



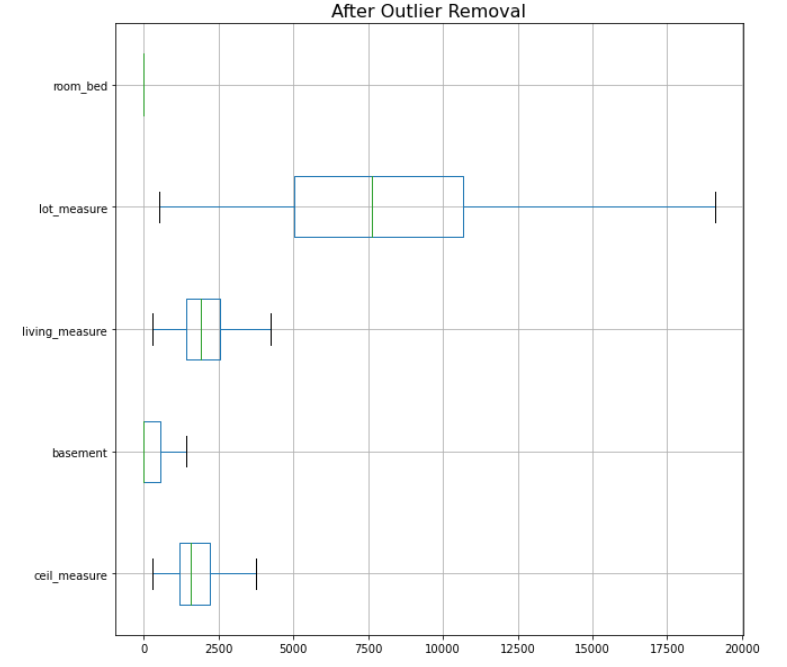
**Figure 39: No missing values**

**e) Outlier treatment (if required)**

Outlier treatment is necessary to ensure effectiveness of model. Boxplot is used to identify outliers. We have substituted the outlier present on the higher side with the 95th Percentile and the outliers which are present on the lower side with the 5th percentile.



**Figure 40: Outlier presence with boxplot**



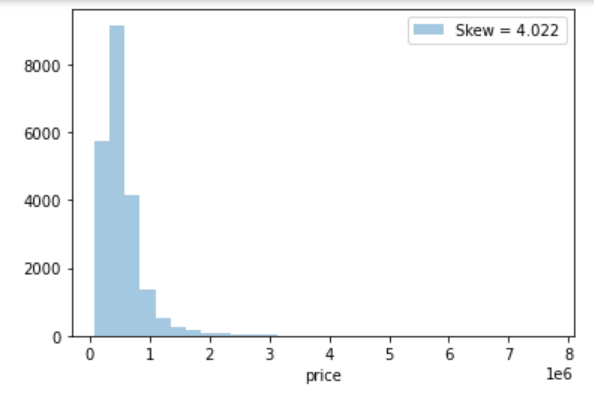
**Figure 41: Outlier treatment**

As seen previously room bed 33 was an outlier and hence it was treated.

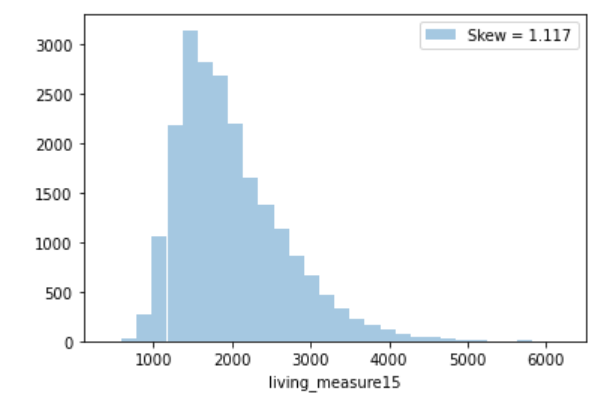
Lot measure, living measure, basement and ceil measure have good relation with target variable and hence they were treated.

Rest of the variable though contain outliers were not treated as they may lead to loss of valuable data.

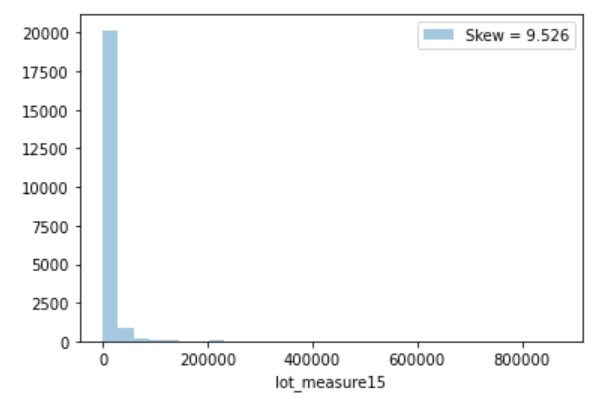
**f) Variable transformation (if applicable)**



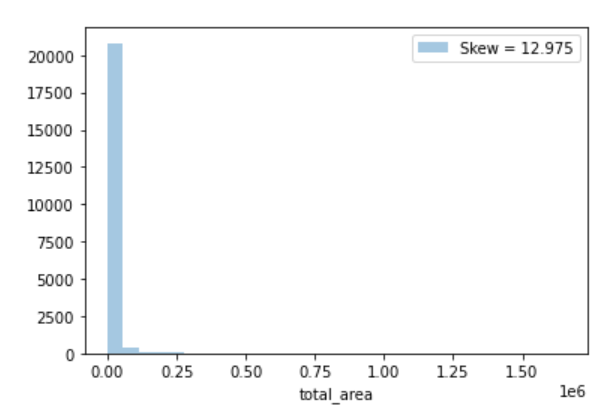
**Figure 42: Price skewness**



**Figure 43: Living measure15 skewness**

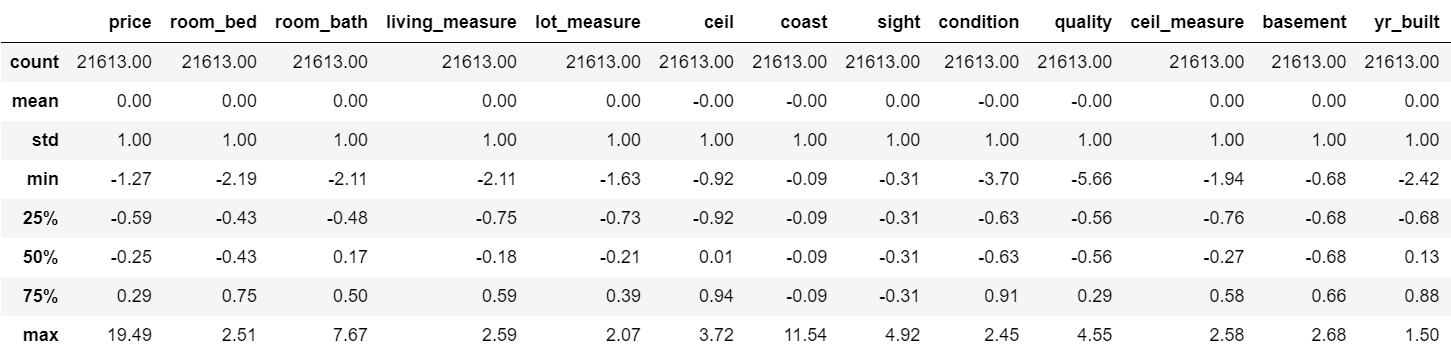


**Figure 44: Lot measure15 skewness**



**Figure 45: Total area skewness**

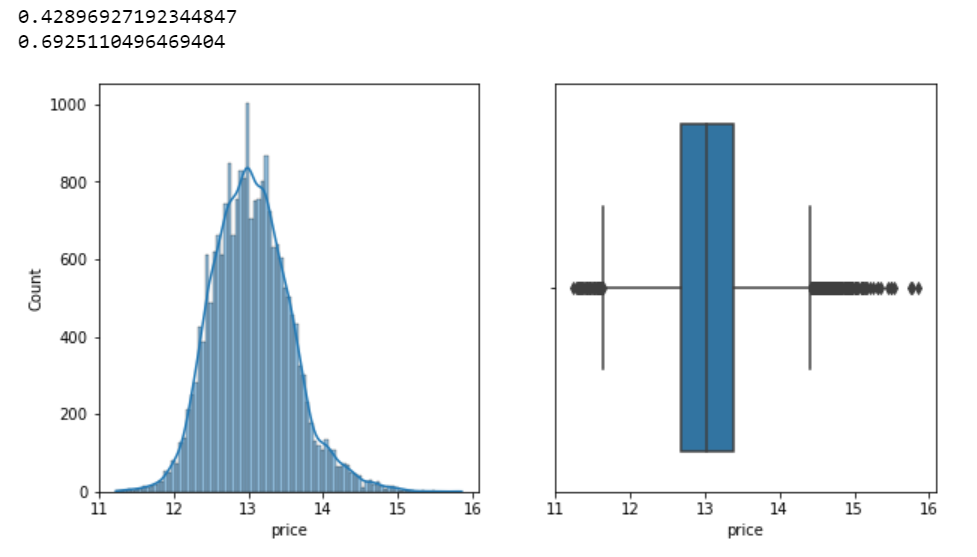
Numerical columns like price, living measure15, lot measure15, basement and total area have skewness more than 1. These variables need to be transformed to improve skewness and kurtosis. Skewness and kurtosis should be in the range of -1 to +1. Before applying variable transformation, scaling is performed since variables are of different scales. Standard scaler technique scales the data in such a way that the mean value of the features tends to 0 and the standard deviation tends to 1.



**Table 5: Description of scaled data**



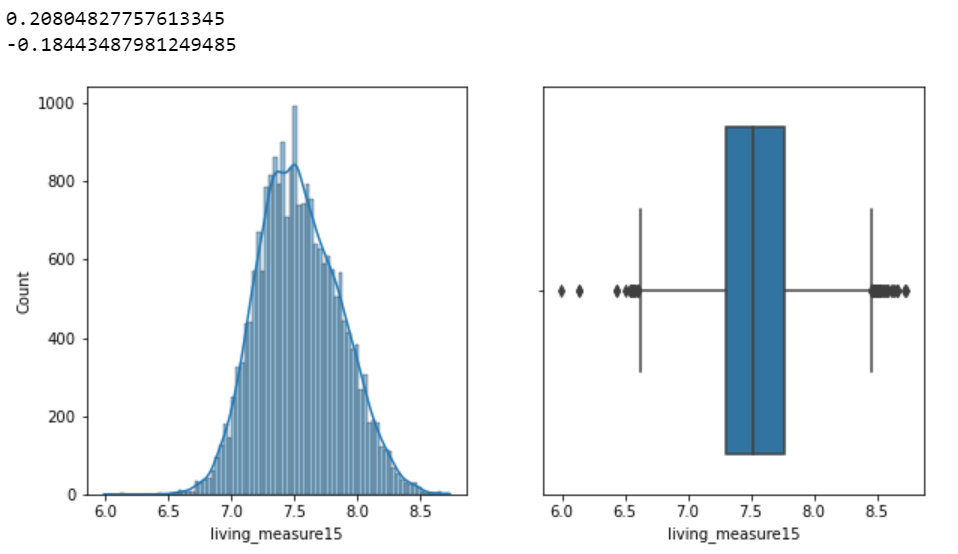
**Figure 46: Price skewness and kurtosis**



**Figure 47: Price log transformation**



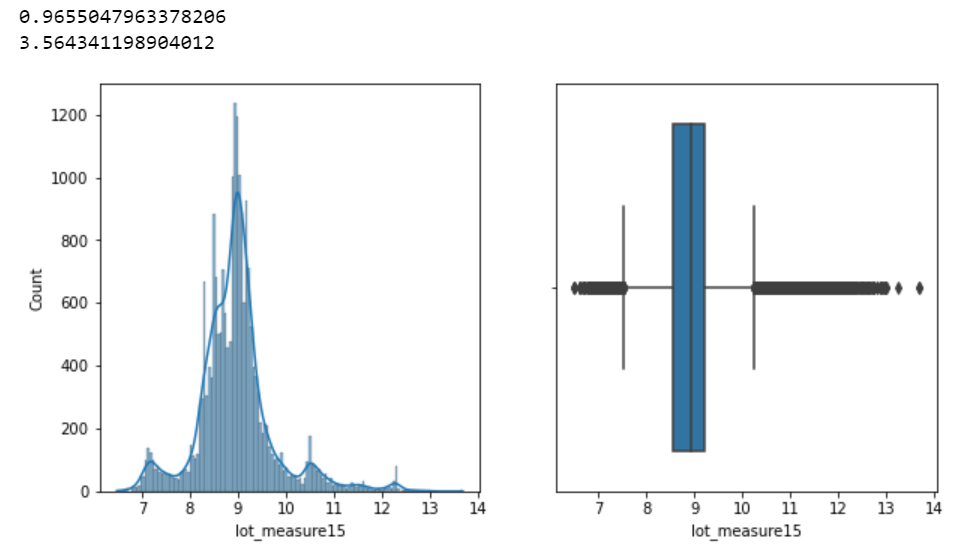
**Figure 48: Living measure15 transformation measures**



**Figure 49: Living measure15 log transformation**



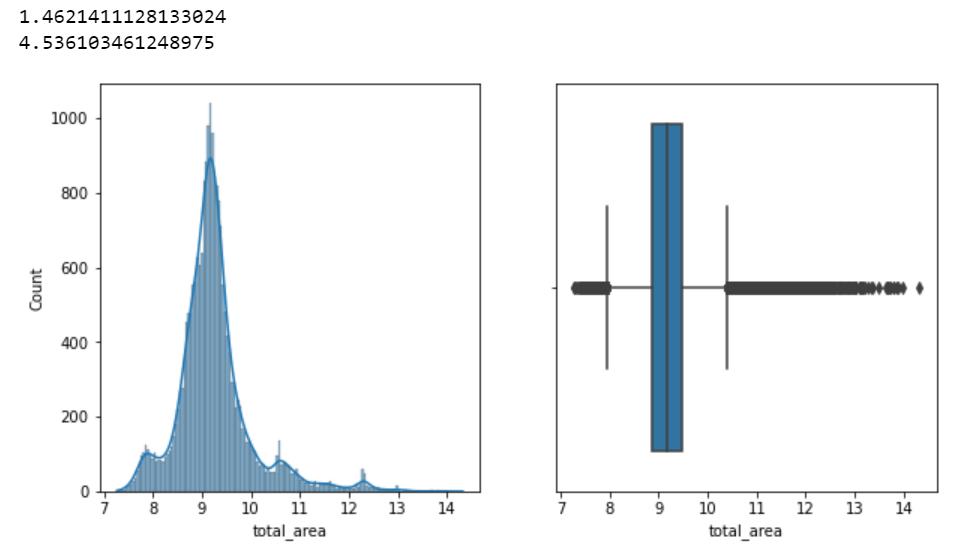
**Figure 50: Lot measure15 transformation measures**



**Figure 51: Lot measure15 log transformation**



**Figure 52: Total area transformation measures**



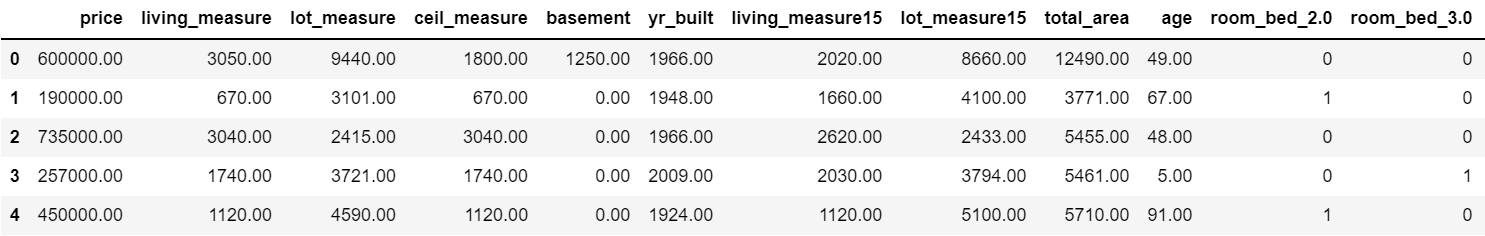
**Figure 53: Total area log transformation**



**Figure 54: Basement transformation measures**

Log transformation was performed to convert the features to follow normal distribution.

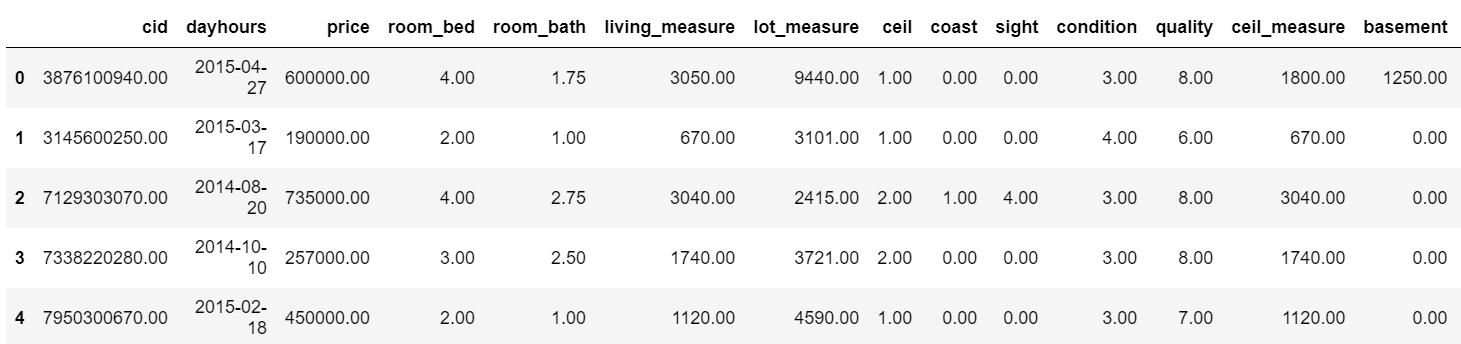
Dummy encoding is used to transform categorical columns like room bed, room bath, ceil, coast, sight, condition, quality and furnished.



**Table 6: Encoding categorical features**

**g) Addition of new variables (if required)**

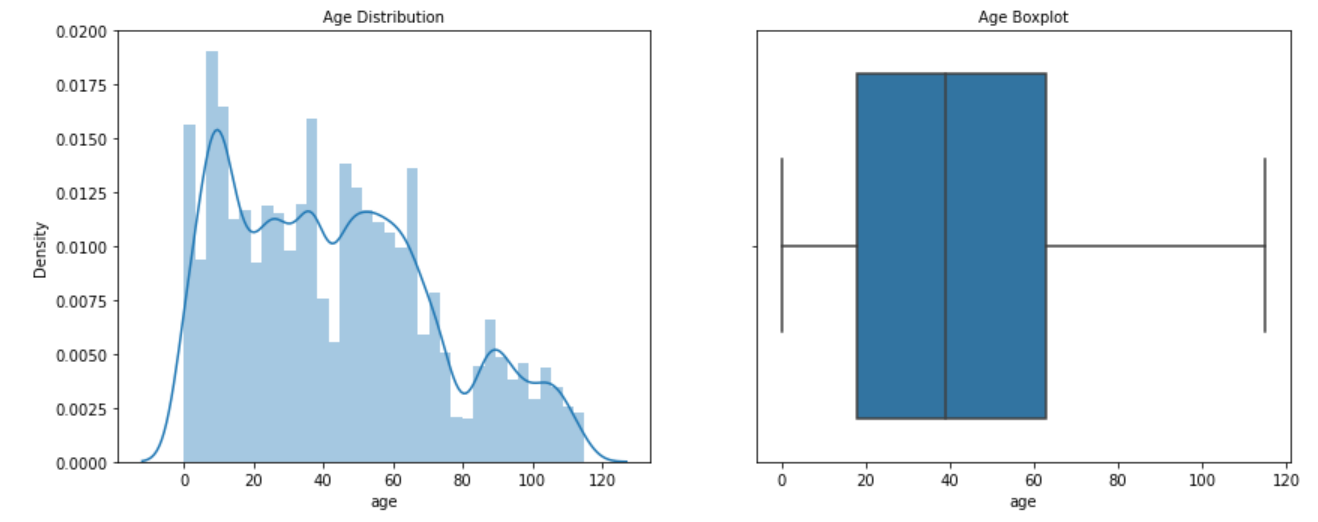
Age variable is added to the dataset which is the difference between extracted year from day hours and year built.



**Table 7: Timestamp removed from day hours**



**Table 8: Age head**



**Figure 55: Age distribution**

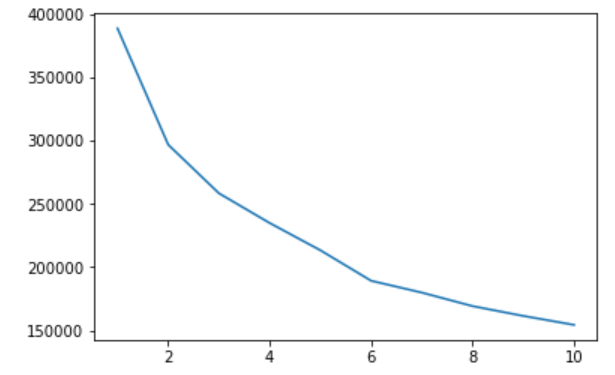
**4) Business insights from EDA**

**a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business**

The dataset is unbalanced as there are presence of missing values and outliers in the dataset which has been treated successfully. While plotting categorical variables against target column, it shows higher price for classes 1 than class 0. SMOTE can be applied to resolve data imbalance problem that will be taken care while doing train-test split. This results in either oversampling or under sampling the data.

**b) Any business insights using clustering (if applicable)**

K-means clustering is used to get any further business insights. Cluster output for all observations and within cluster sum of squares is calculated. Clusters are formed with different values of K and optimum clusters are selected and evaluated using elbow method and silhouette score.



**Figure 56: WSS plot**

2 clusters are chosen as the optimum one since the decline is the sharpest between K=1 and 2 values. Silhouette score also confirms the same. Cluster profiling would categorise the dataset into high and low sale prices.

**c) Any other business insights**

* The data sample consist of a mixture of categorical and numerical data. Regression algorithms will be the best suited model for predicting the prices.
* Incorrect data entry errors should be avoided as it impacts the data types which if ignored can lead to severe penalty.
* Strategies should be devised to investigate the reason for low sale prices of houses in a certain locality.