CAPSTONE PROJECT FINAL BUSINESS REPORT

Table of Contents

1) Introduction of the business problem

This section aims at introducing the project and providing the basic understanding of the project and the objectives of this analysis. The analysis deals with target to understand the real estate market of the geographical location given. Prediction of house prices is not only depend upon square foot of space that it occupies but, different other factors like, number of bedrooms, bathrooms, floors, basement area, condition of house, quality of house, year of build, age of the house, age of renovation of the house, etc., are few of the important points that play a major role in determining its cost.

a) Defining Problem Statement

The goal of this analysis is to understand the relationship between the features of the house and how those features can predict the house price. A house value is simply more than location and square footage. For example, you want to sell a house and you don't know the price which you may expect – it can't be too low or too high. To find house price you usually try to find similar properties in your neighbourhood and based on gathered data you will try to assess your house price.

b) Need of the study/project

This section aims at understanding the attributes in the data set which are not explained well in the problem.

Ceil - 1 indicates the level/floor of house which is lowest in the attributes and 3.5 indicates the maximum levels/floor of house.
Coast -0 indicates closer to waterfront and 1 indicates farther to waterfront
Condition - 1 indicates Poor Condition and 5 indicates Best Condition
Quality - 1 indicate Poor Quality and 13 indicates Best Quality
Furnished - 0 indicates not furnished and 1 indicates furnished

Different houses have different features, features of more than two houses can help evaluate relevant prices. Hence, analysing the bulk of data can help predict the house price. To get the profitable pricing for the houses and buildings, so that neither the seller nor the buyer are at a loss.

c) Understanding business/social opportunity

This section aims at understanding that how will such kind of a project or a study generate business profitability or social benefits. Real estate is a booming sector that contributes hugely to the country's economy. It is also one of the sectors that contribute substantially to generating the employment. When we talk about employment, it's not only for the brokers of the houses or the builders, rather it also accounts those laborers who help with construction of the building. Now, if a sector is contributing such heavily into the economy and employment, then it's fair to have an honest and viable pricing of the product that the sector generates, in our case, houses. Any unfair pricing will be injustice not only to buyer and the seller but also to the workers who are contributing building the real estate. Not only this, big companies who are into building, buying and selling of the properties which means that the major turnover of these companies are from the pricing of the houses. These houses maybe newly built or selling of an already existing house. Also this is the investment option chosen by majority of the public.

2) Data Report

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

This section provides us to understand how the data was collected. The data has houses built from 1900 to 2015. We have data of houses from 1900 – 2015.

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

1. cid: a notation for a house
2. dayhours: Date house was sold
3. price: Price is prediction target
4. room_bed: Number of Bedrooms/House
5. room_bath: Number of bathrooms/bedrooms
6. living_measure: square footage of the home
7. lot_measure: quare footage of the lot
8. ceil: Total floors (levels) in house
9. coast: House which has a view to a waterfront
10. sight: Has been viewed
11. condition: How good the condition is (Overall)
12. quality: grade given to the housing unit, based on grading system
13. ceil_measure: square footage of house apart from basement
14. basement_measure: square footage of the basement
15. yr_built: Built Year
16. yr_renovated: Year when house was renovated
17. zipcode: zip
18. lat: Latitude coordinate
19. long: Longitude coordinate
20. living_measure15: Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area
21. lot_measure15: lotSize area in 2015(implies-- some renovations) 22. furnished: Based on the quality of room
23. total_area: Measure of both living and lot

Figure 1 Dataset Head

	0	1	2	3	4	5	6	
cid	3876100940	3145600250	7129303070	7338220280	7950300670	8016250080	510002519	16240592
dayhours	20150427T000000	20150317T000000	20140820T000000	20141010T000000	20150218T000000	20140709T000000	20140715T000000	20140618T0000
price	600000	190000	735000	257000	450000	245000	466000	11600
room_bed	4.0	2.0	4.0	3.0	2.0	3.0	2.0	4
room_bath	1.75	1.0	2.75	2.5	1.0	2.5	1.5	3
living_measure	3050.0	670.0	3040.0	1740.0	1120.0	1610.0	1140.0	468C
lot_measure	9440.0	3101.0	2415.0	3721.0	4590.0	7223.0	1058.0	970C
ceil	1	1	2	2	1	2	3	
coast	0	0	1	0	0	0	0	
sight	0.0	0.0	4.0	0.0	0.0	0.0	0.0	C
condition	3	4	3	3	3	3	3	
quality	8.0	6.0	8.0	8.0	7.0	7.0	7.0	1C
ceil_measure	1800.0	670.0	3040.0	1740.0	1120.0	1610.0	1140.0	336C
basement	1250.0	0.0	0.0	0.0	0.0	0.0	0.0	132C
yr_built	1966	1948	1966	2009	1924	1994	2005	20
yr_renovated	0	0	0	0	0	0	0	
zipcode	98034	98118	98118	98002	98118	98030	98103	980
lat	47.7228	47.5546	47.5188	47.3363	47.5663	47.3661	47.6608	47.57
long	-122.183	-122.274	-122.256	-122.213	-122.285	$	-122.333	-122.1
living_measure15	2020.0	1660.0	2620.0	2030.0	1120.0	1610.0	1170.0	280C
lot_measure15	8660.0	4100.0	2433.0	3794.0	5100.0	7162.0	1116.0	12343
furnished	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1
total_area	12490	3771	5455	5461	5710	8833	2198	143

23 rows × 25 columns

The columns have different factors affecting the price of the house. Many have different meaning and impact to the price of the house

Figure 2 Datatype information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 23 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   cid               21613 non-null  int64
 1   dayhours          21613 non-null  object
 2   price             21613 non-null  int64
 3   room_bed          21505 non-null  float64
 4   room_bath         21505 non-null  float64
 5   living_measure    21596 non-null  float64
 6   lot_measure       21571 non-null  float64
 7   ceil              21571 non-null  float64
 8   coast             21612 non-null  float64
 9   sight             21556 non-null  object
 10  condition         21556 non-null  float64
 11  quality           21612 non-null  object
 12  ceil_measure      21612 non-null  float64
 13  basement          21612 non-null  float64
 14  yr_built          21612 non-null  float64
 15  yr_renovated      21613 non-null  int64
 16  zipcode           21613 non-null  int64
 17  lat               21613 non-null  float64
 18  long              21613 non-null  object
 19  living_measure15  21447 non-null  float64
 20  lot_measure15     21584 non-null  float64
 21  furnished         21584 non-null  object
 22  total_area        21584 non-null  float64
dtypes: float64(14), int64(4), object(5)
memory usage: 3.8+ MB
```

There are different data types present in the data, we have
Int 64 – 4
Object – 5
Float 64 – 14

We also see that there are null values present in the data, room bath, living measure, lot measure, ceil, coast, sight, condition, quality, ceil measure, basement, year built, living measure15, lot measure 15, furnished and total area have null values

Figure 3 Data Description

	count	mean	std	min	25%	50%	75%	max
cid	21613.0	4.580302e+09	2.876566e+09	1.000102e+06	2.123049e+09	3.904930e+09	7.308900e+09	9.900000e+09
price	21613.0	5.401822e+05	3.673622e+05	7.500000e+04	3.219500e+05	4.500000e+05	6.450000e+05	7.700000e+06
room_bed	21505.0	3.371355e+00	9.302886e-01	0.000000e+00	3.000000e+00	3.000000e+00	4.000000e+00	3.300000e+01
room_bath	21505.0	2.115171e+00	7.702481e-01	0.000000e+00	1.750000e+00	2.250000e+00	2.500000e+00	8.000000e+00
living_measure	21596.0	2.079861e+03	9.184961e+02	2.900000e+02	1.429250e+03	1.910000e+03	2.550000e+03	1.354000e+04
lot_measure	21571.0	1.510458e+04	4.142362e+04	5.200000e+02	5.040000e+03	7.618000e+03	1.068450e+04	1.651359e+06
ceil	21571.0	1.492050e+00	5.424017e-01	0.000000e+00	1.000000e+00	1.500000e+00	2.000000e+00	3.500000e+00
coast	21612.0	7.449565e-03	8.599076e-02	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
condition	21556.0	3.404899e+00	6.617790e-01	0.000000e+00	3.000000e+00	3.000000e+00	4.000000e+00	5.000000e+00
ceil_measure	21612.0	1.788367e+03	8.281025e+02	2.900000e+02	1.190000e+03	1.560000e+03	2.210000e+03	9.410000e+03
basement	21612.0	2.915225e+02	4.425808e+02	0.000000e+00	0.000000e+00	0.000000e+00	5.600000e+02	4.820000e+03
yr_built	21612.0	1.969733e+03	5.811458e+01	0.000000e+00	1.951000e+03	1.975000e+03	1.997000e+03	2.015000e+03
yr_renovated	21613.0	8.440226e+01	4.016792e+02	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.015000e+03
zipcode	21613.0	9.807794e+04	5.350503e+01	9.800100e+04	9.803300e+04	9.806500e+04	9.811800e+04	9.819900e+04
lat	21613.0	4.756005e+01	1.385637e-01	4.715590e+01	4.747100e+01	4.757180e+01	4.767800e+01	4.777760e+01
living_measure15	21447.0	1.987066e+03	6.855196e+02	3.990000e+02	1.490000e+03	1.840000e+03	2.360000e+03	6.210000e+03
lot_measure15	21584.0	1.276654e+04	2.728699e+04	6.510000e+02	5.100000e+03	7.620000e+03	1.008700e+04	8.712000e+05
total_area	21584.0	1.716098e+04	4.159747e+04	0.000000e+00	7.020000e+03	9.562500e+03	1.298200e+04	1.652659e+06

CID: House D/Propertv ID. Not used tor analvsis
price: Our target column value is in 75k - 7700k range. As Mean > Median, it's Right-Skewed
room_bed: Number of bedrooms range from 0 - 33. As Mean slightly > Median, it's slightly Right-Skewed.
room_bath: Number of bathrooms range from 0 - 8. As Mean slightly < Median, it's slight Left-Skewed
lIving_measure: square footage of house ranges from 290 - 13.540. As Mean > Median it's Right-Skewed
lot measure: Square footage of lot range from 520 - 16,51,359. As Mean almost double or Median, it's Highly Right-skewed.
ceil: Number or floors range trom 1 - 3.5 As Mean Median, It's almost Normal Distributed.
coast: As this value represent whether house has waterfront view or not It's categorical column. From above analysis we got know, very few houses has Waterfront view
sight: Value ranges from 0 - 4. As Mean > Median, it's Right-Skewed
condition: Represents rating of house which ranges from 1 - 5. As Mean > Median it's Right-Skewed

quality: Representing grade given to house which range from 1 - 13. As Mean > Median, it's Right-Skewed
ceil measure: square foot of house apart from basement range in 290 - 9.410. As Mean > Median. it's Right-Skewed
basement: Square footage house basement ranges in 0 - 4,820. As Mean highly > Median. it's Highly Right-Skewed.
yr_built: House built year ranges from 1900 - 2015. As Mean < Median, it's Left- Skewed
yr_renovated: House renovation year only 2015. So, this column can be user as Categorical Variable for knowing whether house is renovated
Zipcode: House zipcode ranges trom yoUVl- y819y. As Mean > Median, It's Right Skewed
lat: Latitude ranges from 47.1559 - 47.7776 As Mean < Median it's Left-Skewed
long: Longitude ranges from -122 5190 to > Median, it's Right Skewed.
Living_measure 15: Value ranges trom 399 to 6.210. As Mean > Median. it's Right- Skewed
lot measure15: Value ranges from 651 to 8.71.200. As Mean highly > Median,
it's Highly Right-Skewed
furnished: Representing whether house is furnished or not. It's a Categorical Variable
total_area: Total area of house ranges from 1,423 to 16,52,659. As Mean is almost double of Median. it's Highlv Right-Skewed.

3) Exploratory data analysis

Univariate data analysis

Univariate analysis is the easiest way to analyse data

Figure 4 CID analysis

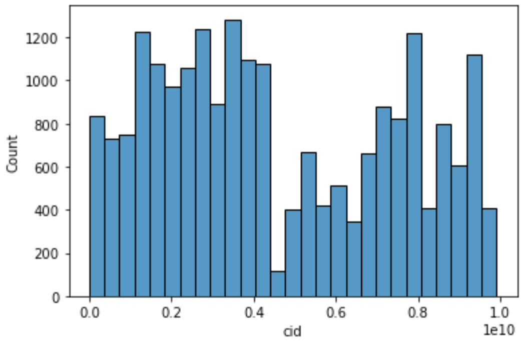


Figure 5 CID analysis boxplot

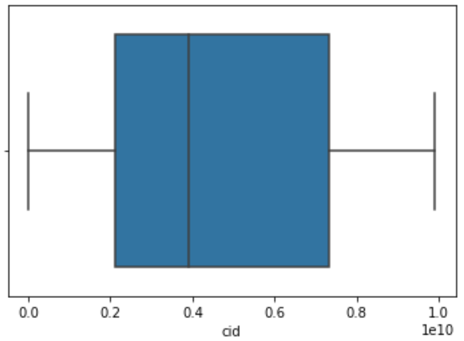


Figure 6 price analysis

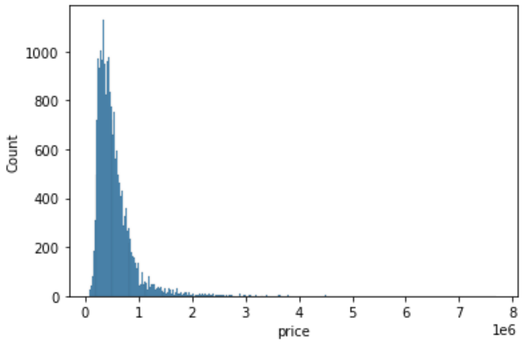


Figure 7 price analysis boxplot

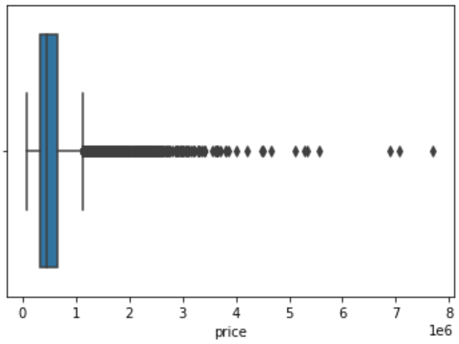


Figure 8 living measure analysis

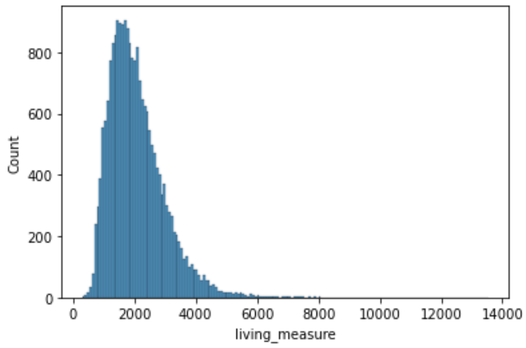


Figure 9 living measuFigure 8re boxplot analysis

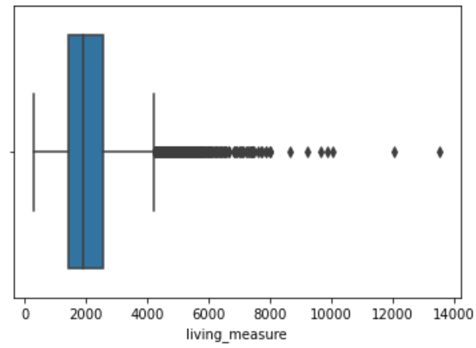


Figure 10 lot measure analysis

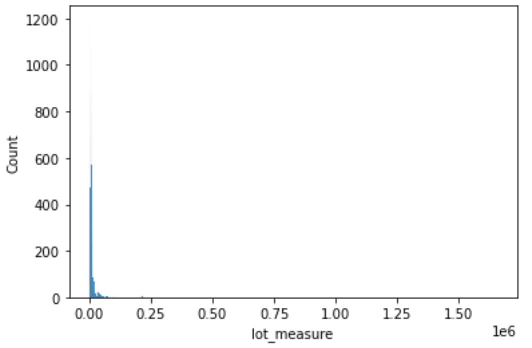


Figure 11 lot measure boxplot analysis

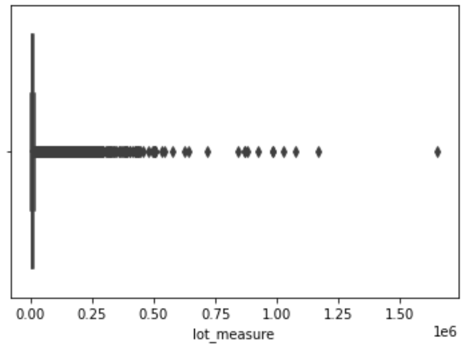


Figure 12 ceil analysis

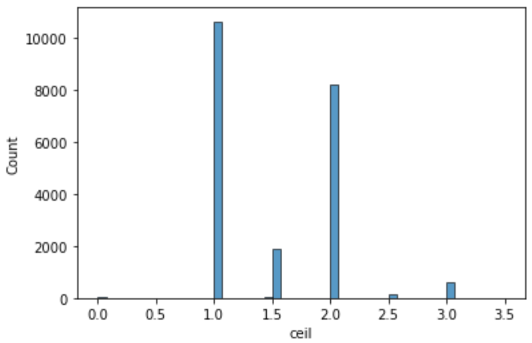


Figure 13 ceil boxplot analysis

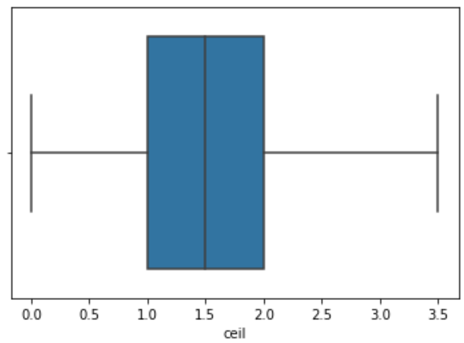


Figure 14 ceil measure analysis

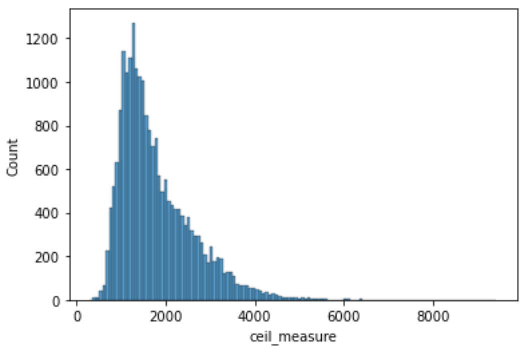


Figure 15 ceil measure boxplot analysis

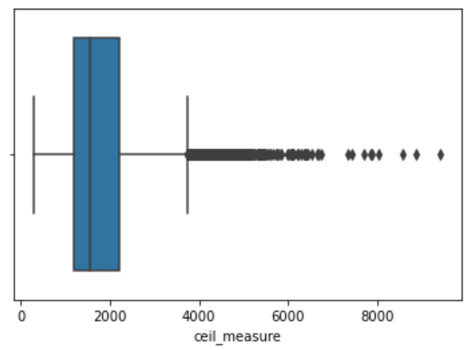


Figure 16 basement analysis

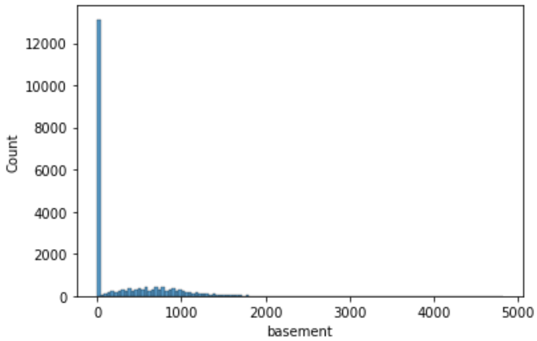


Figure 17 basement boxplot analysis

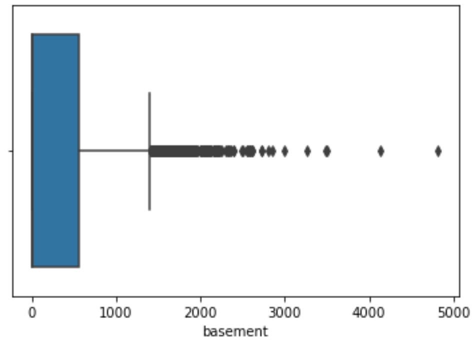


Figure 18 year built analysis

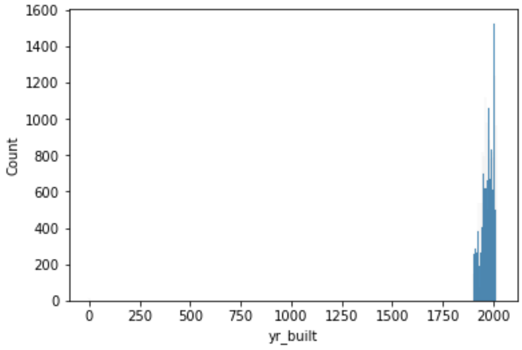


Figure 19 year built boxplot analysis

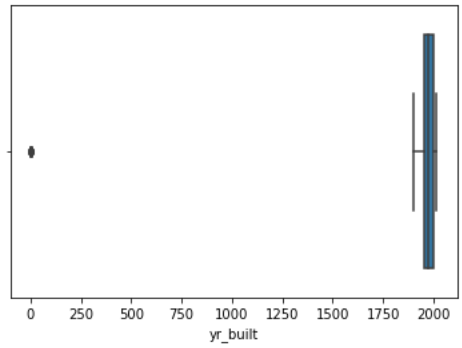


Figure 20 year renovated analysis

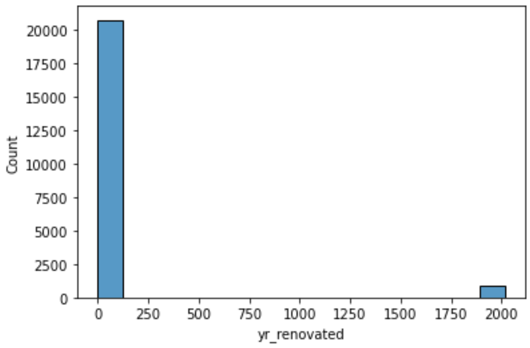


Figure 21 year renovated boxplot analysis

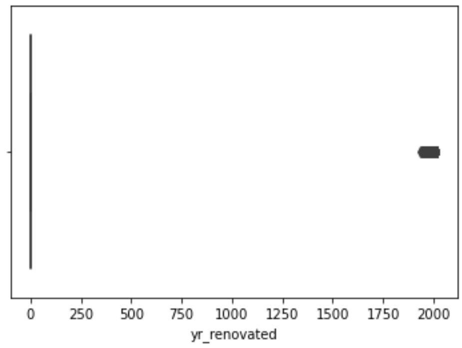


Figure 22 living measure 15 analysis

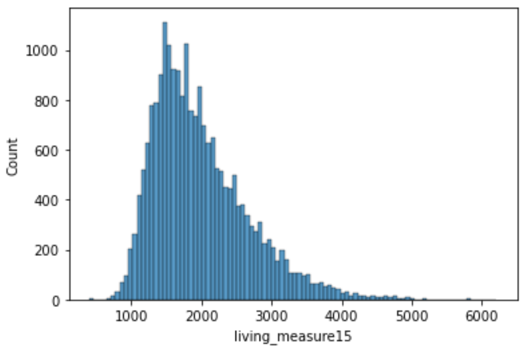


Figure 23 living measure15 boxplot analysis

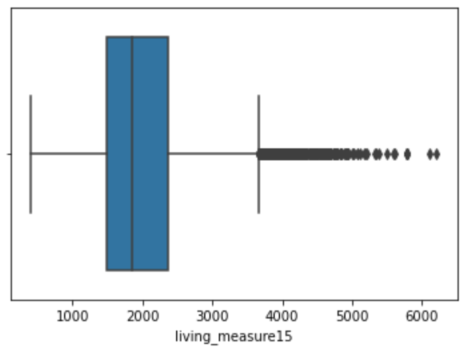


Figure 24 lot measure15 analysis

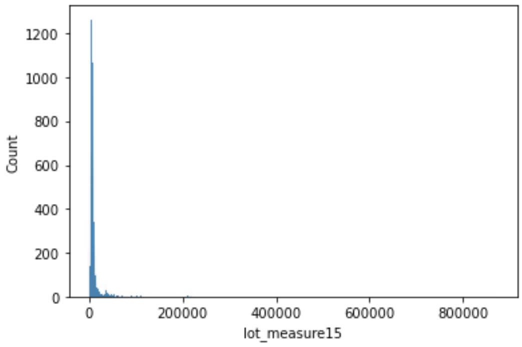


Figure 25 lot measure15 boxplot analysis

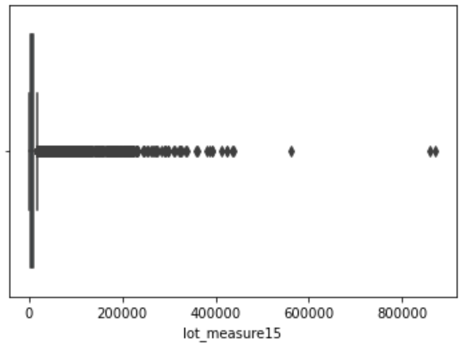


Figure 26 total area analysis

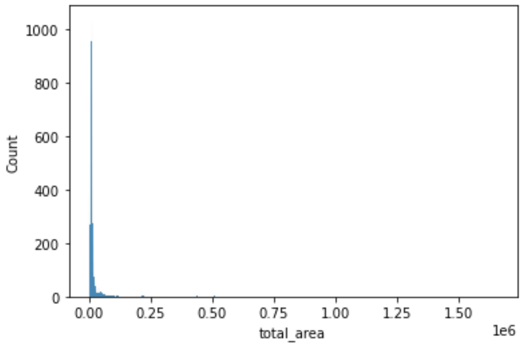
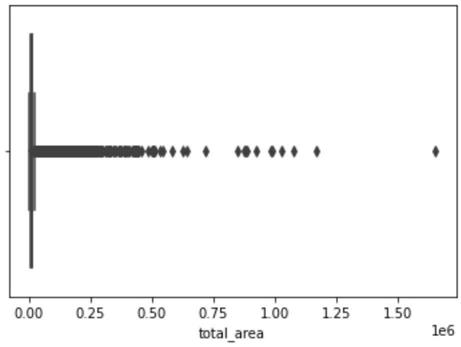


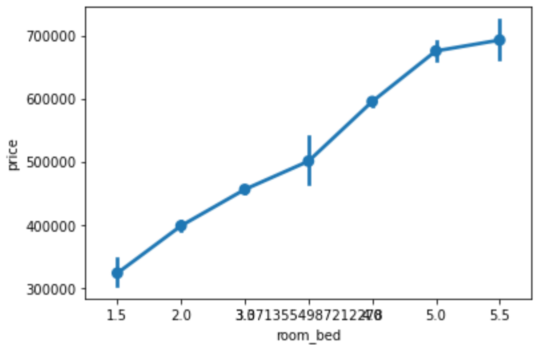
Figure 27 total area boxplot analysis



Very few houses are renovated, only 914 houses are renovated out of total 21613 records house with no sight or 0 record is more, after that we have house few more houses with 2 sights, house with 1 or 4 site is very minimal. Most of the houses in the dataset has bedroom within the range of 0 to 5 more no of houses are built from year 2000 onwards, from the year 1900 to 1950 we can see less no of house got constructed more no of unfurnished house are there in data set .
17500 house are unfurnished and near about only 4000 houses are furnished Most of the houses are non-coast in the dataset and very few houses negligible amount of houses are near the coast.

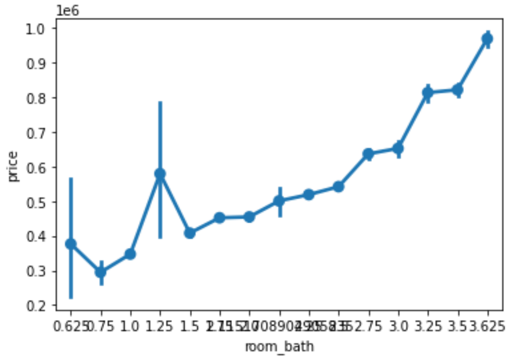
Bivariate Analysis

Figure 28 room_bed vs price



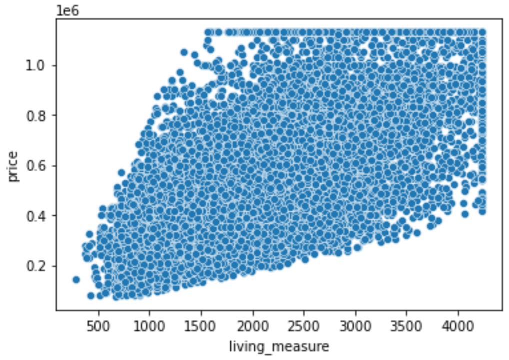
We can see an increasing trend in price, with increase in number of bedrooms. For high number of bedrooms the price of the house is also high

Figure 29 room bath vs price



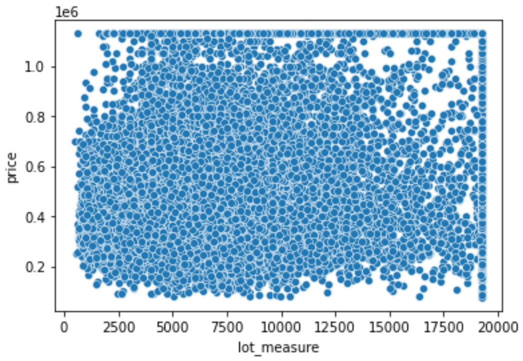
We can see an increasing trend in price, with increase in number of bathrooms. For high number of bathrooms the price of the house is also high

Figure 30 living measure vs price



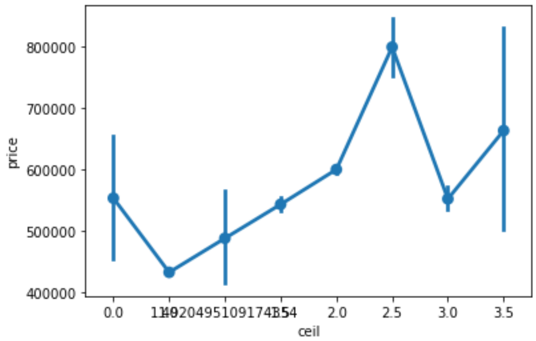
The price is high for houses with higher living measure. Big houses are at a costlier price

Figure 31 lot measure vs price



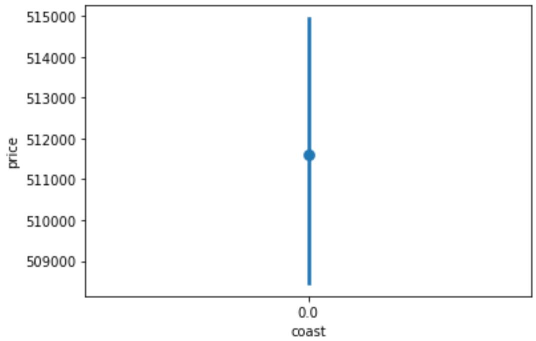
Price for Lot measure is almost same for all size.

Figure 32 ceil vs price



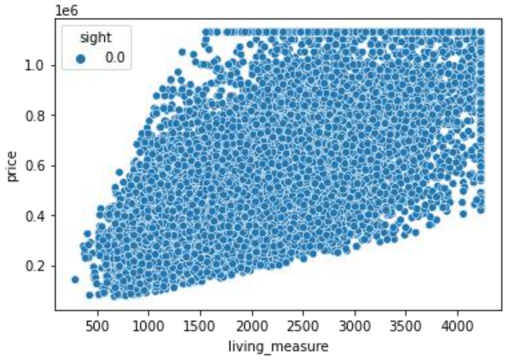
House with high number of floors have a higher price.

Figure 33 coast vs price



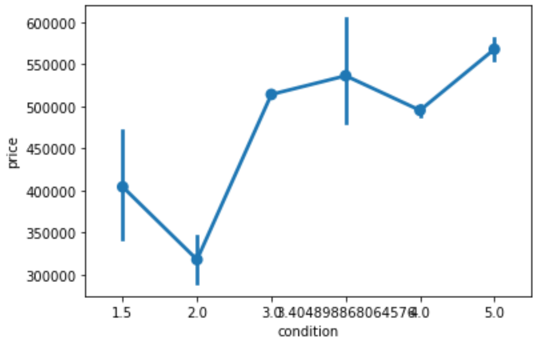
House with a coast view attracts higher price.

Figure 34 living measure vs price



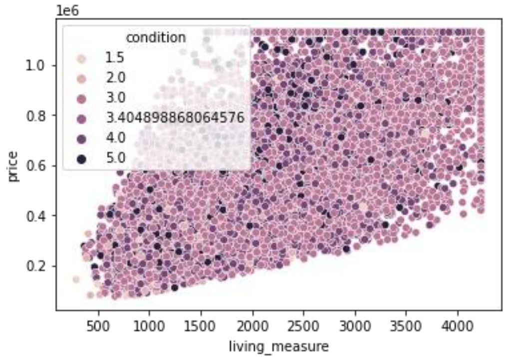
The sight view is high for houses with high living measure and the price is also high for such houses.

Figure 35 condition vs price



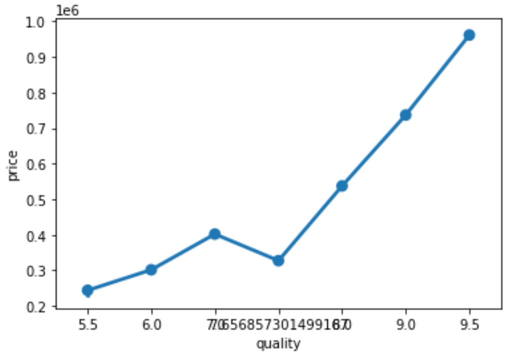
Houses with good condition attracts high price when compared to houses which are not in a good condition.

Figure 36 living measure vs price, hue= condition



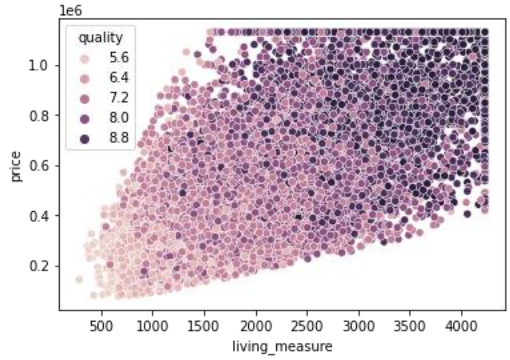
House with high living measure are mostly in a good condition and the price of those houses are also high.

Figure 37 quality vs price



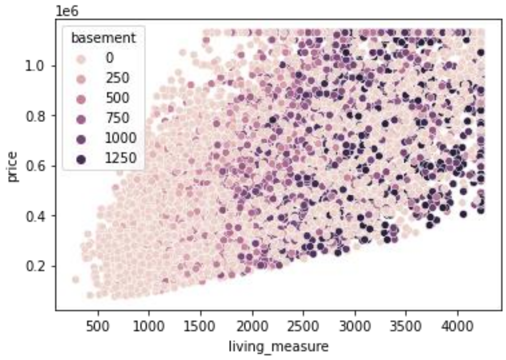
Quality of house is of a high concern. Houses with good quality are of higher price when compared to houses with low quality.

Figure 38 living measure vs price, hue= quality



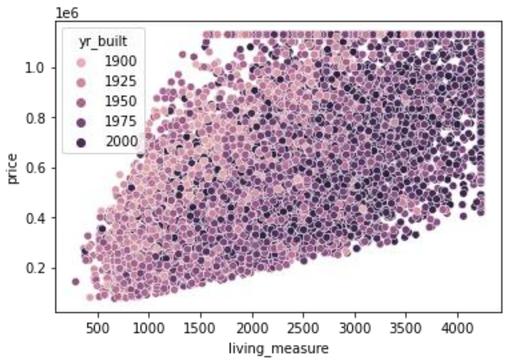
House with high living measure are mostly in a good quality and the price of those houses are also high.

Figure 39 living measure vs price, hue= basement



House with high living measure are mostly having a he basement and the price of those houses are also high.

Figure 40 living measure vs price, hue year built



House with high living measure are mostly built in recent years and the price of those houses are also high.

Missing Value treatment (if applicable)

Figure 41 missing values

cid	0
dayhours	0
price	0
room_bed	108
room_bath	108
living_measure	17
lot_measure	42
ceil	42
coast	1
sight	57
condition	57
quality	1
ceil_measure	1
basement	1
yr_built	1
yr_renovated	0
zipcode	0
lat	0
long	0
living_measure15	166
lot_measure15	29
furnished	29
total_area	29
dtype: int64	

There are missing values present in room bath, living measure, lot measure, ceil, coast, sight, condition, quality, ceil measure, basement, year built, living measure15, lot measure 15, furnished and total area.

Figure 42 treated missingvalues

cid	0
dayhours	0
price	0
room_bed	0
room_bath	0
living_measure	0
lot_measure	0
ceil	0
coast	0
sight	0
condition	0
quality	0
ceil_measure	0
basement	0
yr_built	0
yr_renovated	0
zipcode	0
lat	0
long	0
living_measure15	0
lot_measure15	0
furnished	0
total_area	0
dtype: int64	

All the missing Values are treated.

Outlier treatment (if required)

Figure 43 Outliers treated for cid

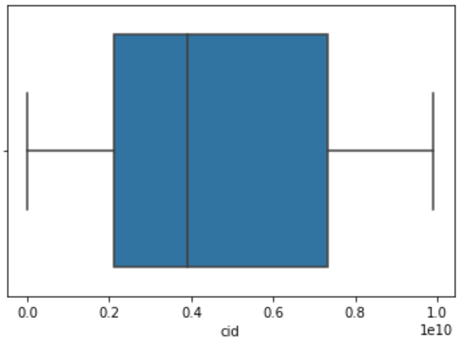


Figure 44 Outliers treated for price

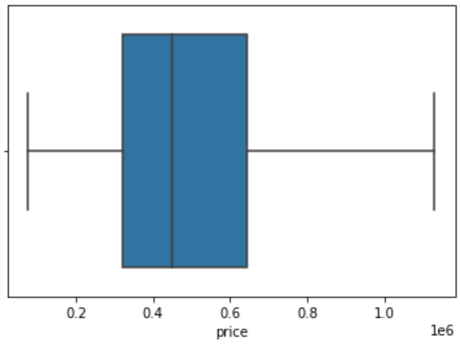


Figure 45 Outliers treated for living measure

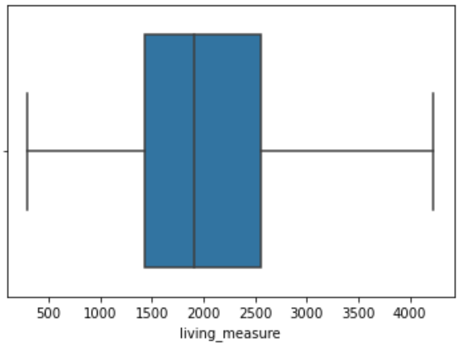


Figure 46 Outliers treated for lot measure

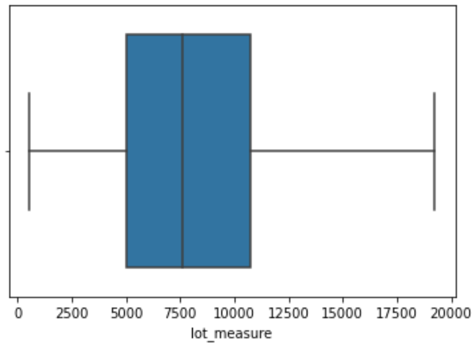


Figure 47 Outliers treated for ceil

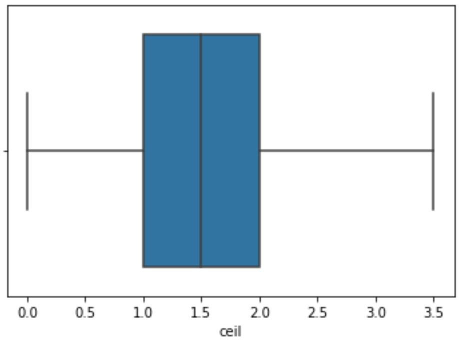


Figure 48 Outliers treated for ceil measure

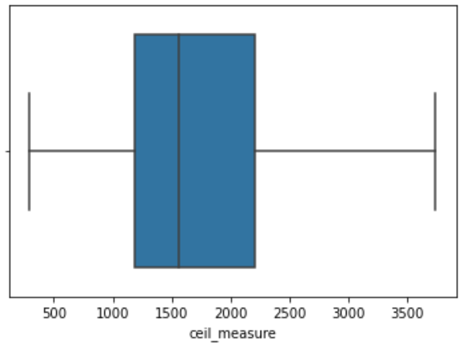


Figure 49 Outliers treated for basement

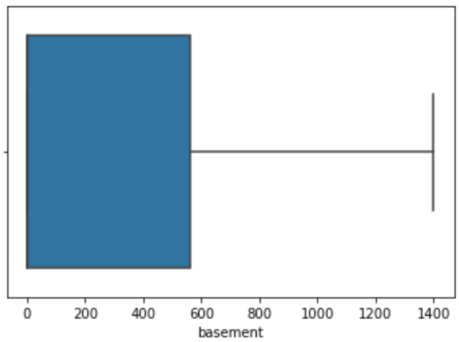


Figure 50 Outliers treated for year built

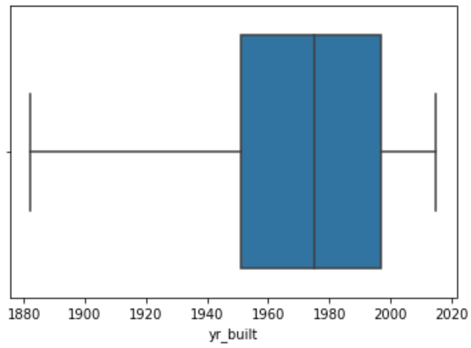


Figure 51 Outliers treated for lot measure 15

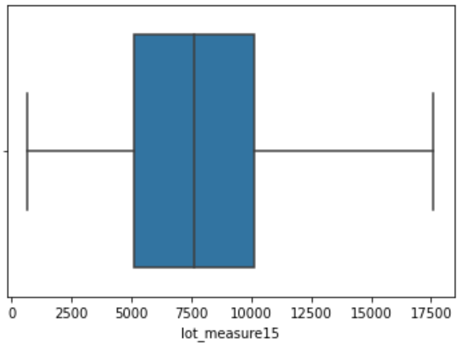
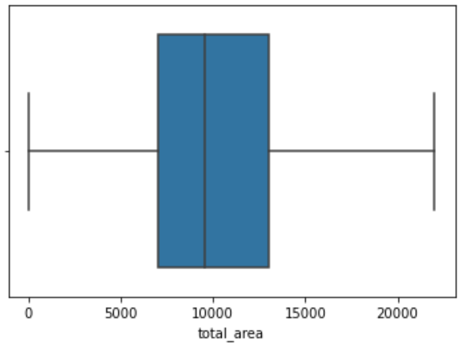


Figure 52 Outliers treated for total area



All the columns are treated with outliers. As we can see no black dots lying in the boxplot

Value counts of independent variables data does seems to be unbalanced due to the outliers. IQ method is used for treating the outliers. Outliers can degrade the efficiency of the data. - It results in overestimation or underestimation

Multivariate Analysis

Figure 53 Outliers treated for total area

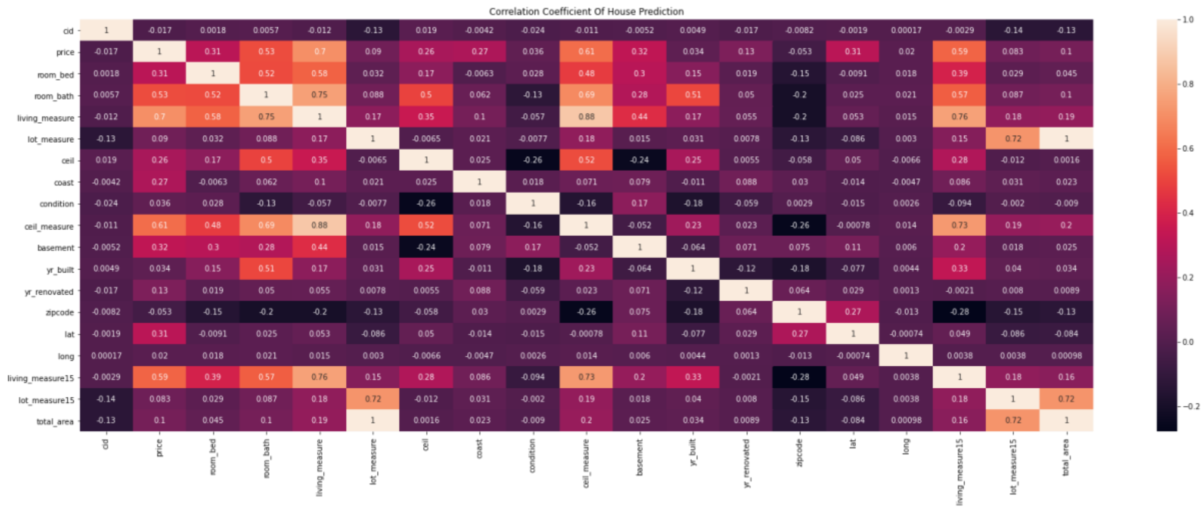


Figure 53 Columns with dummies

```
Index(['cid', 'dayhours', 'price', 'living_measure', 'lot_measure',
       'ceil_measure', 'basement', 'yr_built', 'yr_renovated', 'zipcode',
       'lat', 'long', 'living_measure15', 'lot_measure15', 'total_area',
       'room_bed_2.0', 'room_bed_3.0', 'room_bed_3.3713554987212278',
       'room_bed_4.0', 'room_bed_5.0', 'room_bed_5.5', 'room_bath_0.75',
       'room_bath_1.0', 'room_bath_1.25', 'room_bath_1.5', 'room_bath_1.75',
       'room_bath_2.0', 'room_bath_2.1151708904905835', 'room_bath_2.25',
       'room_bath_2.5', 'room_bath_2.75', 'room_bath_3.0', 'room_bath_3.25',
       'room_bath_3.5', 'room_bath_3.625', 'ceil_0', 'ceil_1', 'ceil_2',
       'ceil_3', 'ceil_4', 'ceil_5', 'ceil_6', 'coast_0', 'coast_1', 'coast_2',
       'condition_0', 'condition_1', 'condition_2', 'condition_3',
       'condition_4', 'condition_5', 'quality_6.0', 'quality_7.0',
       'quality_7.656857301499167', 'quality_8.0', 'quality_9.0',
       'quality_9.5'],
      dtype='object')
```

Convert categorical variable into dummy/indicator variables. As many columns will be created as distinct values. This is also known as one hot coding.

Figure 54 Dataset with Dummies

	cid	dayhours	price	living_measure	lot_measure	ceil_measure	basement	yr_built	yr_renovated	zipcode	...	condition_2	condition_3	condi
0	3.876101e+09	351	600000.0	3050.0	9440.0	1800.0	1250.0	66	0.0	98034.0	...	1	0	
1	3.145600e+09	310	190000.0	670.0	3101.0	670.0	0.0	48	0.0	98118.0	...	0	1	
2	7.129303e+09	110	735000.0	3040.0	2415.0	3040.0	0.0	66	0.0	98118.0	...	1	0	
3	7.338220e+09	161	257000.0	1740.0	3721.0	1740.0	0.0	109	0.0	98002.0	...	1	0	
4	7.950301e+09	283	450000.0	1120.0	4590.0	1120.0	0.0	24	0.0	98118.0	...	1	0	

5 rows × 57 columns

We can see that dummies are created for categorical variable. Number of columns are increased in number.

Now we need to divide the data into Test data and Train data. The data must be divided into 30 percent Test data and 70 percent into Train data.

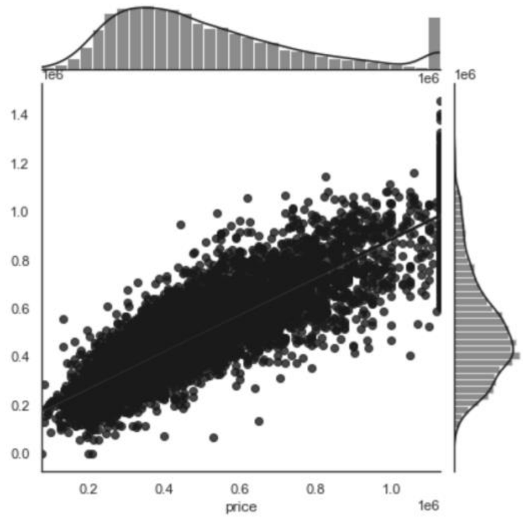
4) Model Building

Linear Regression Model

Figure 55 Model score of LR

	Method	Test Score	RMSE_te	MSE_te	MAE_te	train Score	RMSE_tr	MSE_tr	MAE_tr
0	Linear Reg Model1	0.752434	122843.896374	1.509062e+10	93444.075608	0.755254	124360.822285	1.546561e+10	93598.338471

Figure 56 Joint plot of LR



Inference:

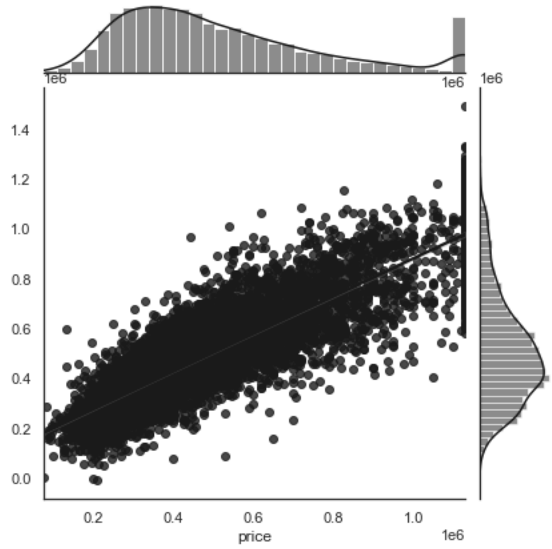
Linear Regression model score for Train data: 0.755254
Linear Regression RMSE score for Train data: 124360.822285
Linear Regression MSE score for Train data: 1.546561e+10
Linear Regression MAE score for Train data: 93598.338471

Linear Regression model score for Test data: 0.752434
Linear Regression RMSE score for Test data: 122843.896374
Linear Regression MSE score for Test data: 1.509062e+10
Linear Regression MAE score for Test data: 93444.075608

The model score for Test and Train data is almost same around 75 percent. The Root mean square error, mean square error and mean absolute error is also almost same for test and train data.

Lasso Regressor

Figure 57 Joint plot for Lasso



Linear Regression model score for Train data: 0.745304
Linear Regression RMSE score for Train data: 124600.136129
Linear Regression MSE score for Train data: 1.552519e+10
Linear Regression MAE score for Train data: 94331.658967

Linear Regression model score for Test data:0.749215
Linear Regression RMSE score for Test data: 125885.573343
Linear Regression MSE score for Test data: 1.584718e+10
Linear Regression MAE score for Test data: 94575.736010

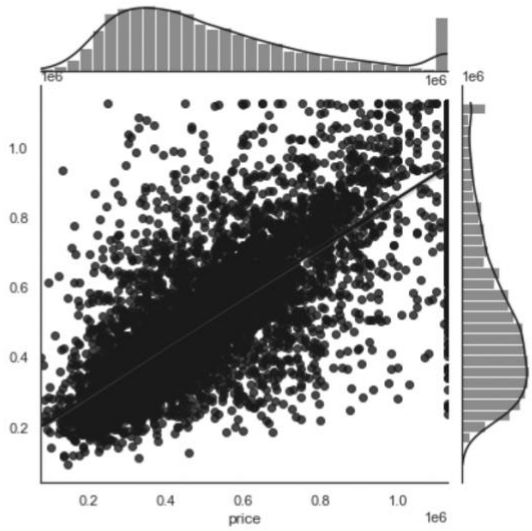
The model score for Test and Train data is almost same around 75 percent. The Root mean square error, mean square error and mean absolute error is also almost same for test and train data.

KNN Regressor Model

Figure 58 Model score of KNN

	Method	Test Score	RMSE_te	MSE_te	MAE_te	train Score	RMSE_tr	MSE_tr	MAE_tr
0	Linear Reg Model1	0.752434	122843.896374	1.509062e+10	93444.075608	0.755254	124360.822285	1.546561e+10	93598.338471
0	knn1	0.581503	159718.103624	2.550987e+10	104056.025321	0.997910	11493.213340	1.320940e+08	2859.934655

Figure 59 Joint plot of KNN



Inference:

KNN Regression model score for Train data: 0.997910
KNN Regression RMSE score for Train data: 11493.213340
KNN Regression MSE score for Train data: 1.320940e+08
KNN Regression MAE score for Train data: 2859.934655

KNN Regression model score for Test data: 0.581503
KNN Regression RMSE score for Test data: 159718.103624
KNN Regression MSE score for Test data: 2.550987e+10
KNN Regression MAE score for Test data: 104056.025321

The model score for Train data is high around 99 percent and model score for Test data is around 58 percent which is very low. Train data model seems to be over fitted.

Decision Tree Regressor

Figure 60 Model score of DT

	Method	Test Score	RMSE_te	MSE_te	MAE_te	train Score	RMSE_tr	MSE_tr	MAE_tr
0	Linear Reg Model1	0.752434	122843.896374	1.509062e+10	93444.075608	0.755254	124360.822285	1.546561e+10	93598.338471
0	knn1	0.581503	159718.103624	2.550987e+10	104056.025321	0.997910	11493.213340	1.320940e+08	2859.934655
0	DT1	0.766752	119238.595775	1.421784e+10	79823.611505	0.998861	8484.654337	7.198936e+07	723.499636

Figure 61 Joint plot of DT

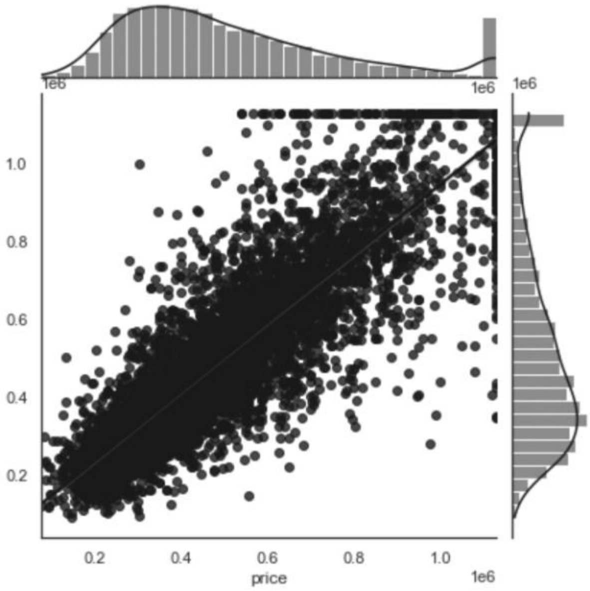
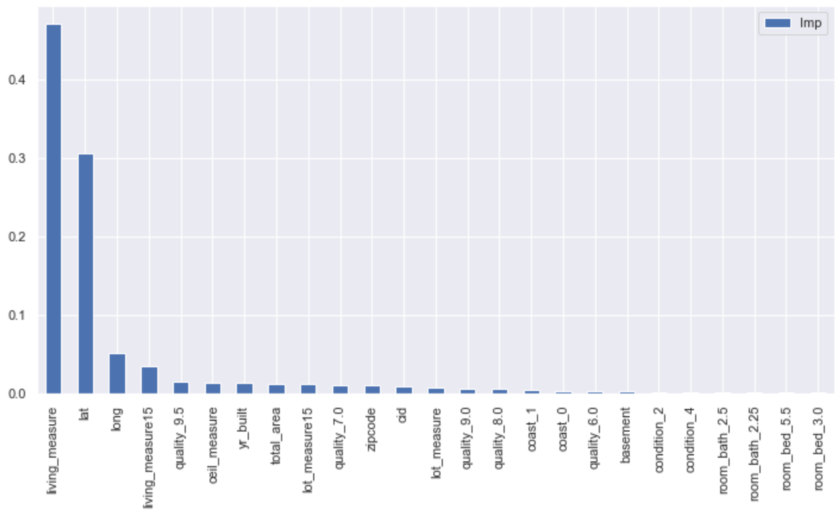


Figure 62 Feature importance of DT



Decision Tree Regression model score for Train data: 0.998861
Decision Tree Regression RMSE score for Train data: 8484.654337
Decision Tree Regression MSE score for Train data: 7.198936e+07

Decision Tree Regression MAE score for Train data: 723.499636

Decision Tree Regression model score for Test data: 0.762380
Decision Tree Regression RMSE score for Test data: 120350.930610
Decision Tree Regression MSE score for Test data: 1.448435e+10
Decision Tree Regression MAE score for Test data: 80127.949954

The model score for Train data is high around 99 percent and model score for Test data is around 76 percent. Train data model seems to be over fitted. While Test model seems to be fine, but accuracy around 80 percent would be better.

Clearly, our model is over fitted in the above model building techniques, test score is around 75 and 76 percent. While train model score is around 99 percent, which clearly says the model has overfitted. Hence, it might be a big red flag. KNN regressor model and decision tree models have not performed well in comparison with linear regression models.

5). Model Tuning and business implication

Model Tuning

Generally ensemble models are used to avoid problems of overfitting but in this model may be while sampling with replacements some observations got repeated in each subset. Hence, our model is over fitting.

Gradient Boosting Regressor

Figure 63 Model score of GB

	Method	Test Score	RMSE_te	MSE_te	MAE_te	train Score	RMSE_tr	MSE_tr	MAE_tr
0	Linear Reg Model1	0.752434	122843.896374	1.509062e+10	93444.075608	0.755254	124360.822285	1.546561e+10	93598.338471
0	knn1	0.581503	159718.103624	2.550987e+10	104056.025321	0.997910	11493.213340	1.320940e+08	2859.934655
0	DT1	0.766752	119238.595775	1.421784e+10	79823.611505	0.998861	8484.654337	7.198936e+07	723.499636
0	GB1	0.875409	87146.608006	7.594531e+09	62468.761622	0.893289	82116.324538	6.743091e+09	58814.031951

Figure 64 Joint plot of GB

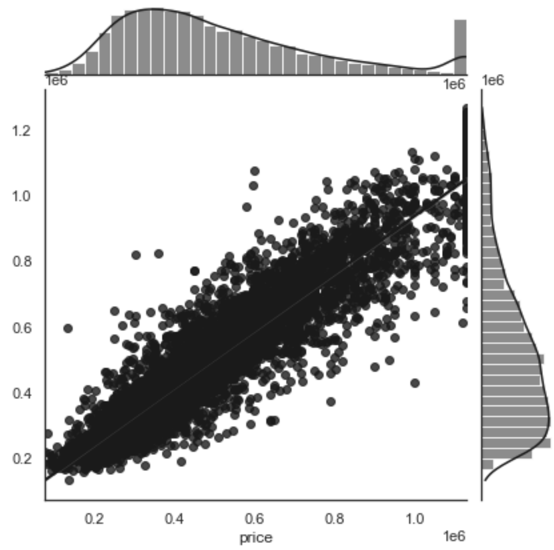
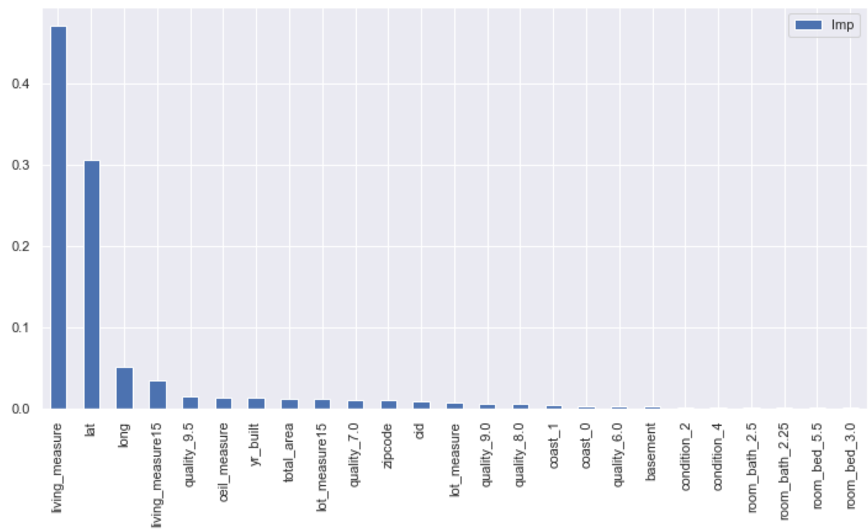


Figure 65 Feature Importance of GB



Gradient Boosting Regression model score for Train data: 0.893289
Gradient Boosting Regression RMSE score for Train data: 82116.324538
Gradient Boosting Regression MSE score for Train data: 6.743091e+09
Gradient Boosting Regression MAE score for Train data: 58814.031951

Gradient Boosting Regression model score for Test data: 0.875409
Gradient Boosting Regression RMSE score for Test data: 87146.608006
Gradient Boosting Regression MSE score for Test data: 7.594531e+09
Gradient Boosting Regression MAE score for Test data: 62468.761622

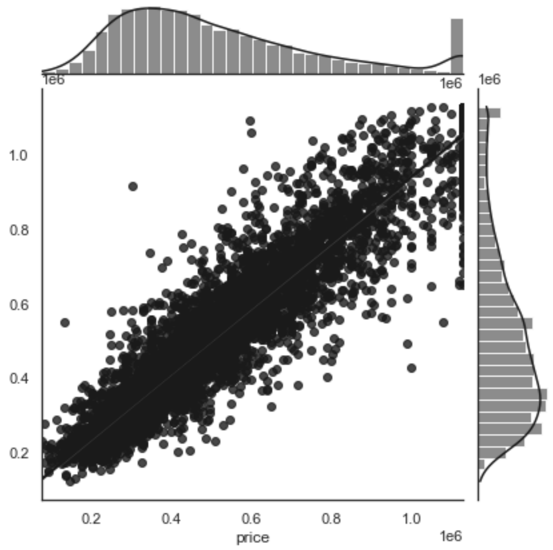
Gradient Boosting Regression techniques has the best test and train score model, the model is also not over fitted. Overall till now, Gradient Boosting Regression seems to have best train and test score.

Bagging Regressor

Figure 66 Model score of BG

	Method	Test Score	RMSE_te	MSE_te	MAE_te	train Score	RMSE_tr	MSE_tr	MAE_tr
0	Linear Reg Model1	0.752434	122843.896374	1.509062e+10	93444.075608	0.755254	124360.822285	1.546561e+10	93598.338471
0	knn1	0.581503	159718.103624	2.550987e+10	104056.025321	0.997910	11493.213340	1.320940e+08	2859.934655
0	DT1	0.766752	119238.595775	1.421784e+10	79823.611505	0.998861	8484.654337	7.198936e+07	723.499636
0	GB1	0.875409	87146.608006	7.594531e+09	62468.761622	0.893289	82116.324538	6.743091e+09	58814.031951
0	BGG1	0.884589	83874.952985	7.035008e+09	57263.357410	0.983000	32775.155253	1.074211e+09	21826.481459

Figure 67 Joint plot of BG



Bagging Regression model score for Train data: 0.983000
Bagging Regression RMSE score for Train data: 32775.155253
Bagging Regression MSE score for Train data: 1.074211e+09
Bagging Regression MAE score for Train data: 21826.481459

Bagging Regression model score for Test data: 0.884589
Bagging Regression RMSE score for Test data: 83874.952985
Bagging Regression MSE score for Test data: 7.035008e+09
Bagging Regression MAE score for Test data: 57263.357410

Bagging Regressor has very good Test model score but when it comes to Train score it seems to be overfitted with 98 percent score.

Random Forest Regressor

Figure 68 Model score of Random Forest Regressor

	Method	Test Score	RMSE_te	MSE_te	MAE_te	train Score	RMSE_tr	MSE_tr	MAE_tr
0	Linear Reg Model1	0.752434	122843.896374	1.509062e+10	93444.075608	0.755254	124360.822285	1.546561e+10	93598.338471
0	knn1	0.581503	159718.103624	2.550987e+10	104056.025321	0.997910	11493.213340	1.320940e+08	2859.934655
0	DT1	0.766752	119238.595775	1.421784e+10	79823.611505	0.998861	8484.654337	7.198936e+07	723.499636
0	GB1	0.875409	87146.608006	7.594531e+09	62468.761622	0.893289	82116.324538	6.743091e+09	58814.031951
0	BGG1	0.884589	83874.952985	7.035008e+09	57263.357410	0.983000	32775.155253	1.074211e+09	21826.481459
0	RF1	0.885672	83480.450491	6.968986e+09	56860.804667	0.983387	32400.712495	1.049806e+09	21523.813211

Figure 69 Joint Plot of Random Forest Regressor

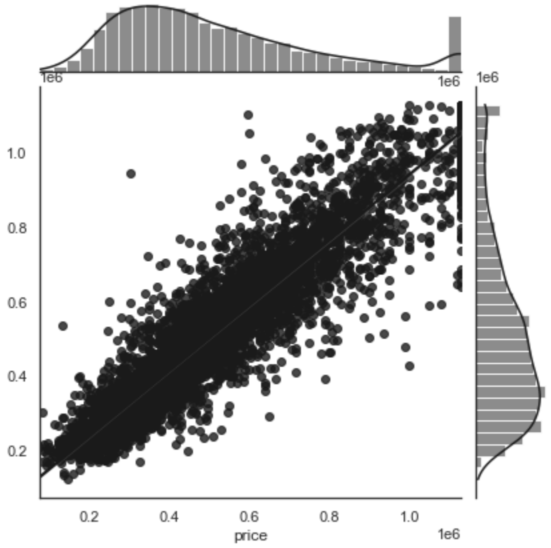
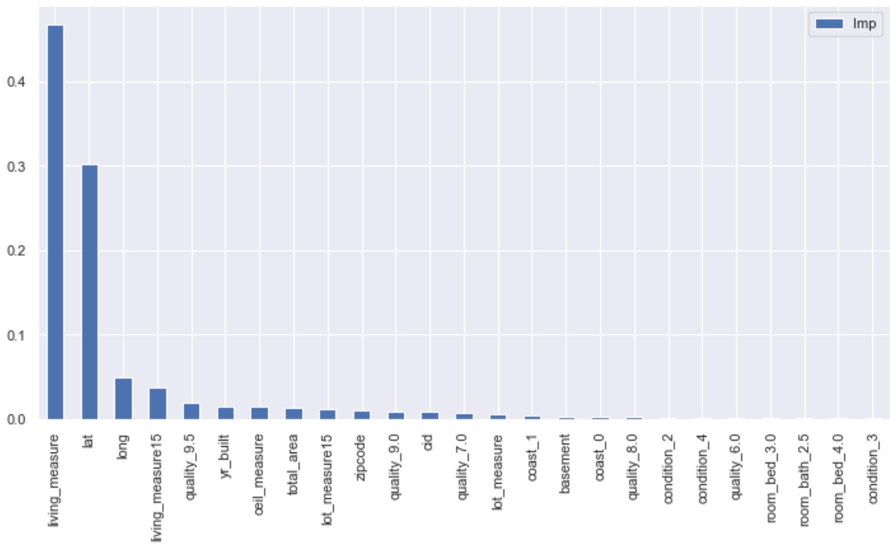


Figure 70 Feature importance of Random Forest Regressor



Random Forest Regression model score for Train data: 0.983676
Random Forest Regression RMSE score for Train data: 32116.834008
Random Forest Regression MSE score for Train data: 1.031491e+09
Random Forest Regression MAE score for Train data: 21406.778425

Random Forest Regression model score for Test data 0.885479
Random Forest Regression RMSE score for Test data: 83550.589739
Random Forest Regression MSE score for Train data 6.980701e+09
Random Forest Regression MAE score for Test data: 56895.874490

Random Forest Regression has very good Test model score but when it comes to Train score it seems to be overfitted with 98 percent score. Hence considering this model will not be affective.

Hypertuning

Figure 71

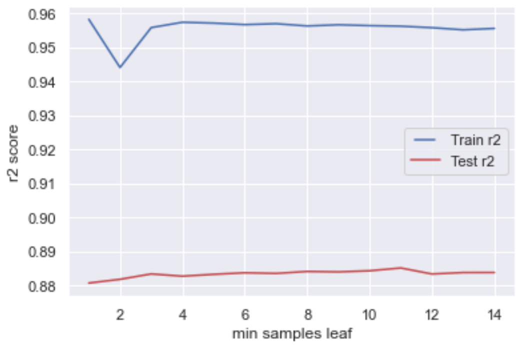


Figure 72

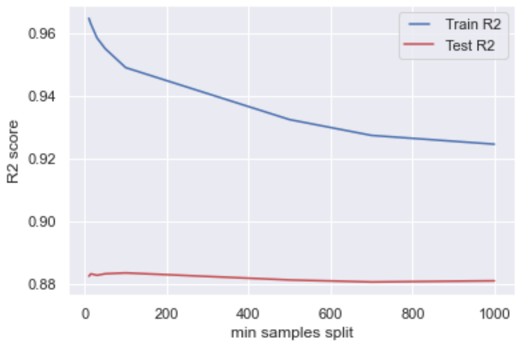


Figure 73

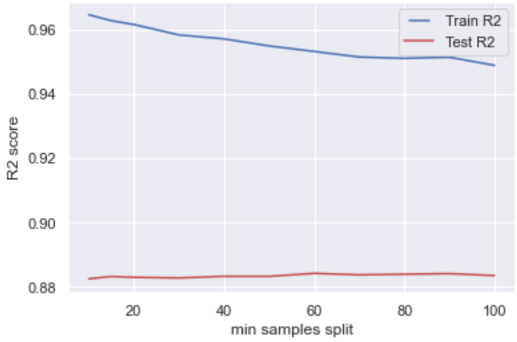


Figure 74

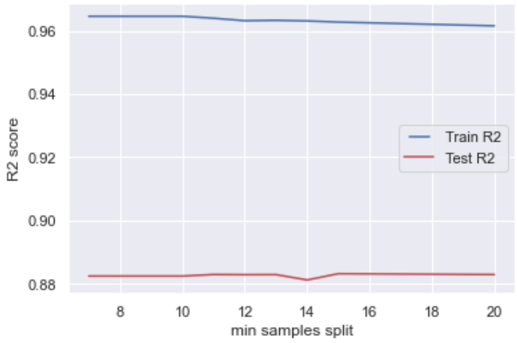


Figure 75

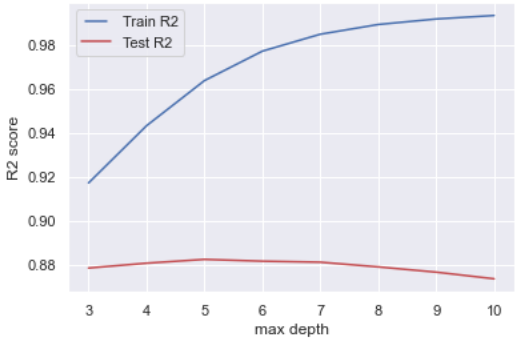
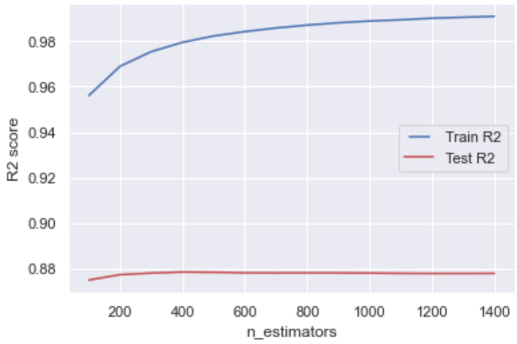


Figure 76



	Method	train Score	RMSE_tr	MSE_tr	test Score	RMSE_te	MSE_te
0	GBRF	0.958318	51321.601676	2.633907e+09	0.893551	80552.529106	6.488710e+09

6) Interpretation of the most optimum model and its implication on the business

Gradient Boosting Regressor can be considered as the best model with best Test model and Train model score with 87 percent and 89 percent respectively when compared to other models.

We have 15 important features important in buying a house and we found that most of the consumers are interested in the following below features.
It is evident from EDA that an ideal house would be the one with 2-3 bedrooms and 3 bathrooms, even though houses with 8 and >8 bedrooms and bathrooms have sold for a higher price a lot of people doesn't seem to be buying them, higher number of records are sold with three-bedroom houses hence an equal or even more revenue could be obtained by selling more houses with three bedrooms and bathrooms.
Although majority of houses are not furnished, it is seen in bivariate analysis that furnished houses produce more revenue compared to unfurnished ones. From the above analysis, we can conclude that, high quality house has the highest house price. These features combined, can help estimate the house price.

Living measure plays an important role in purchase of the house. High the living measure of the house, higher is the price and depending on the living measure purchase of the house is based on.

Latitude and longitude also plays a important role in the house price prediction. Many customers also check for the latitude and longitude .i.e. the area, locality where the house is situated and buy the house. If the house is situated in a higher locality, the price of the house is higher.

Many customers are also concerned about the quality of the house. Quality of the house also plays a vital role in the purchase of the house and the quality of the house also decides the price of the house. House with top quality will have a higher price and customers also prefer a house with good quality.

Year in which the house is built also plays an important role in house price. House built in recent past will have no renovation work or any repairs and the customers prefer to buy house were they need not get anything repaired. Hence this is also an important feature in prediction of price of the house.

Likewise Ceiling measure and total area of the house are also deciding factors in the purchase of the house.