CAPSTONE PROJECT

By Elton Rebello (JAN22A)

Date: 08-01-2023

INDEX -

TABLES -

FIGURES -

1) Introduction of the business problem -

A house value is simply more than location and square footage. Like the features that make up a person, an educated party would want to know all aspects that give a house its value. 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 prices you usually try to find similar properties in your neighborhood and based on gathered data you will try to assess your house price.

We are given data of 21613 houses and their location, lot area, price, number of rooms, condition etc. There are 23 variables explaining about the features that give a house its value.

This study/ project aims to provide the price of a house considering all the features and conditions the house has to offer. This will help the purchasing party understand all the aspects that the house has and can buy the house for the right price. Similarly, the Seller can also quote an appropriate price for the house.

This project gives us the opportunity to study how various variables can affect the price of a property being sold or bought, also how the values appreciate and depreciate over time for houses.

Data Report -

Looks like the Data was collected from the records and documentation of the houses sold in the past, the data contains information about the house built between 1900 and 2015 which tells us the data is from that period.

During the selling and buying of houses the records and documentations are updated and maintained properly, this data is collected from such documents and most of it seems to be accurate as its legal documentation however there is always a chance of error or mistakes in the data.

Sample Of the Data set -

	0	1	2	3	4
cid	3876100940	3145600250	7129303070	7338220280	7950300670
dayhours	20150427T000000	20150317T000000	20140820T000000	20141010T000000	20150218T000000
price	600000	190000	735000	257000	450000
room_bed	4	2	4	3	2
room_bath	1.75	1	2.75	2.5	1
living_measure	3050	670	3040	1740	1120
lot_measure	9440	3101	2415	3721	4590
ceil	1	1	2	2	1
coast	0	0	1	0	0
sight	0	0	4	0	0
condition	3	4	3	3	3
quality	8	6	8	8	7
ceil_measure	1800	670	3040	1740	1120
basement	1250	0	0	0	0
yr_built	1966	1948	1966	2009	1924
yr_renovated	0	0	0	0	0
zipcode	98034	98118	98118	98002	98118
lat	47.7228	47.5546	47.5188	47.3363	47.5663
long	-122.183	-122.274	-122.256	-122.213	-122.285
living_measure15	2020	1660	2620	2030	1120
lot_measure15	8660	4100	2433	3794	5100
furnished	0	0	0	0	0
total_area	12490	3771	5455	5461	5710

Table 1 – Sample of the dataset

In the above table we can see a sample of the data, the data contains 21613 rows and 23 columns. (We have written the transpose of the data so that it fits)

Descriptive Details -

	count	mean	std	min	25%	50%	75%	max
cid	21613	4580301521	2876565571	1000102	2123049194	3904930410	7308900445	9900000190
price	21613	540182	367362	75000	321950	450000	645000	7700000
room_bed	21505	3.4	0.9303	0	3	3	4	33
room_bath	21505	2.1	0.77	0	1.75	2.25	2.5	8
living_measure	21596	2080	918	290	1429	1910	2550	13540
lot_measure	21571	15105	41423	520	5040	7618	10685	1651359
ceil	21571	1.5	0.5	0	1	1.5	2	3.5
coast	21612	0.00745	0.09	0	0	0	0	1
sight	21556	0.2344	0.8	0	0	0	0	4
condition	21556	3.4	0.7	0	3	3	4	5
quality	21612	7.7	1.2	1	7	7	8	13
ceil_measure	21612	1788	828	290	1190	1560	2210	9410
basement	21612	292	443	0	0	0	560	4820
yr_built	21612	1970	58	0	1951	1975	1997	2015
yr_renovated	21613	84	402	0	0	0	0	2015
zipcode	21613	98078	54	98001	98033	98065	98118	98199
lat	21613	48	0.14	47.2	47.5	47.6	47.7	47.8
long	21613	-122.02	5	-122.5	-122.3	-122.2	-122.1	0
living_measure15	21447	1987	686	399	1490	1840	2360	6210

lot_measure15	21584	12767	27287	651	5100	7620	10087	871200
furnished	21584	0.2	0.40	0	0	0	0	1
total_area	21584	17161	41597	0	7020	9563	12982	1652659

Table 2 – Descriptive Details

The Descriptive details show us the mean, median, min, max etc. of various variables. (We have written the transpose of the data so that it fits).

- Price has a mean of 540182 and a standard deviation of 367362, most of the price is around the mean. The min price of a house is 75000 and max is 7700000, only 25% of the houses are sold over 64500.
- More than 25% of the houses have 3 bedrooms. bedrooms have a 3.4 mean and around 0.93 standard deviation which indicates most of the houses have 2 to 4 bedrooms. Only one house has 33 bedrooms up to 75% of the houses have 4 or less bedrooms.
- Mean bathrooms are 2.1 with a standard deviation of 0.77, more than 75% of the houses have less than 2.5 bathrooms. Max bathrooms are 8.
- Houses have a mean of 2080 sq footage of living space with a standard deviation of 918. Min living area is about 290 and max is about 13540.
- around 75% of the houses have less than 2 floors and the mean floors been 1.5 with a standard deviation of 0.5. max floors are 3.5.
- The mean total area is 17161 with a standard deviation of 41597. max total area is 1652659 whereas min total area is 0.

Columns in the dataset -

```
 #   Column             Non-Null Count  Dtype
 0   cid                21613 non-null  int64
 1   date_house_sold    21613 non-null  datetime64[ns]
 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  float64
```

```
10  condition          21556 non-null  float64
11  quality            21612 non-null  float64
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  float64
19  living_measure15   21447 non-null  float64
20  lot_measure15      21584 non-null  float64
21  furnished          21584 non-null  float64
22  total_area         21584 non-null  float64
```

We can see that some rows contain null values in them, Dollar signs in the data have been replaced with 0's. Data contains Float, Int and datetime datatypes.

2) EDA and Business Implications -

Univariate Analysis -

Boxplot and Histplot for price -

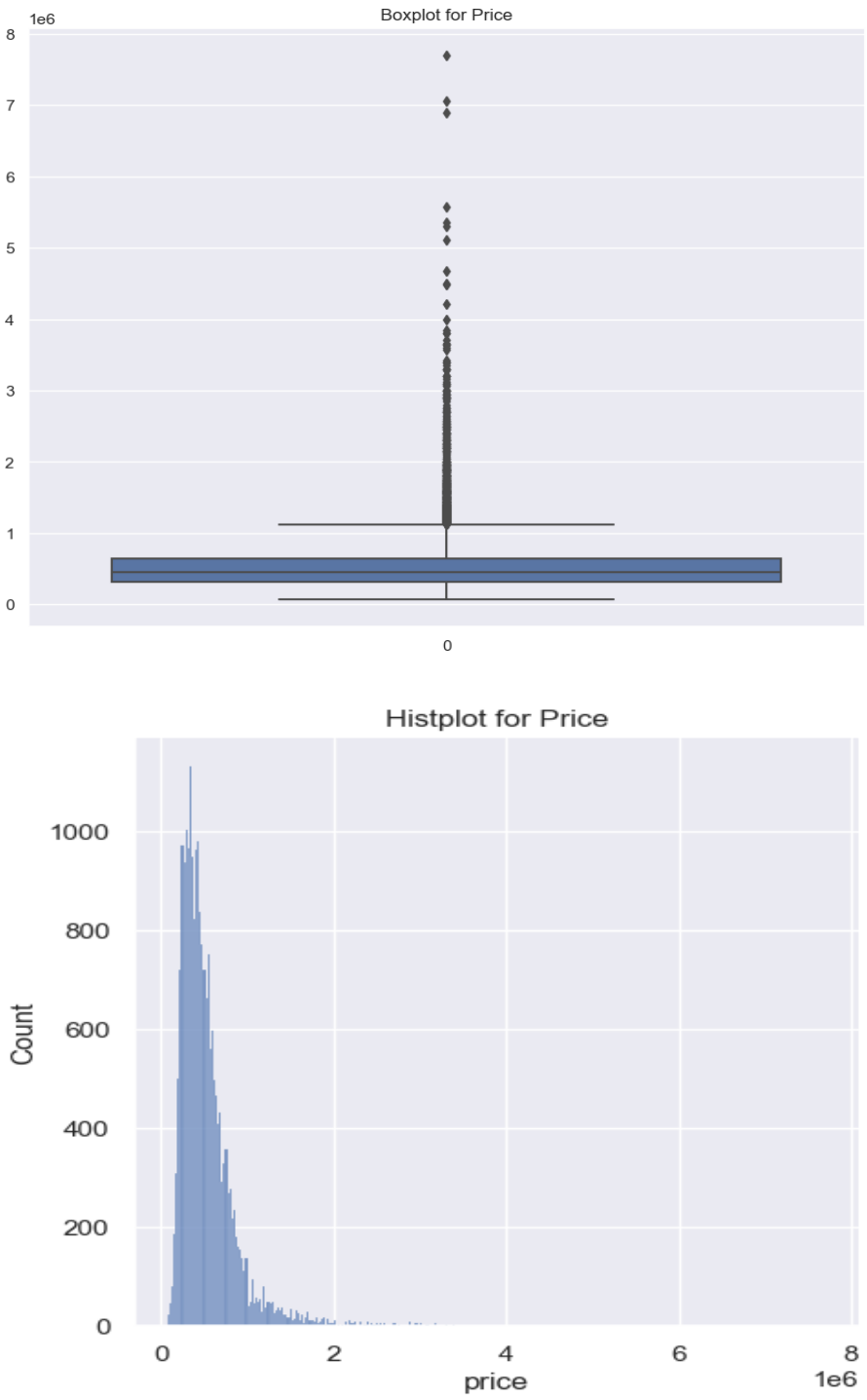


Fig 1 – Boxplot and Histplot for price

The price ranges in between 75000 and 7700000. We can see that the data is right skewed.

Boxplot and Histplot for room_bed -

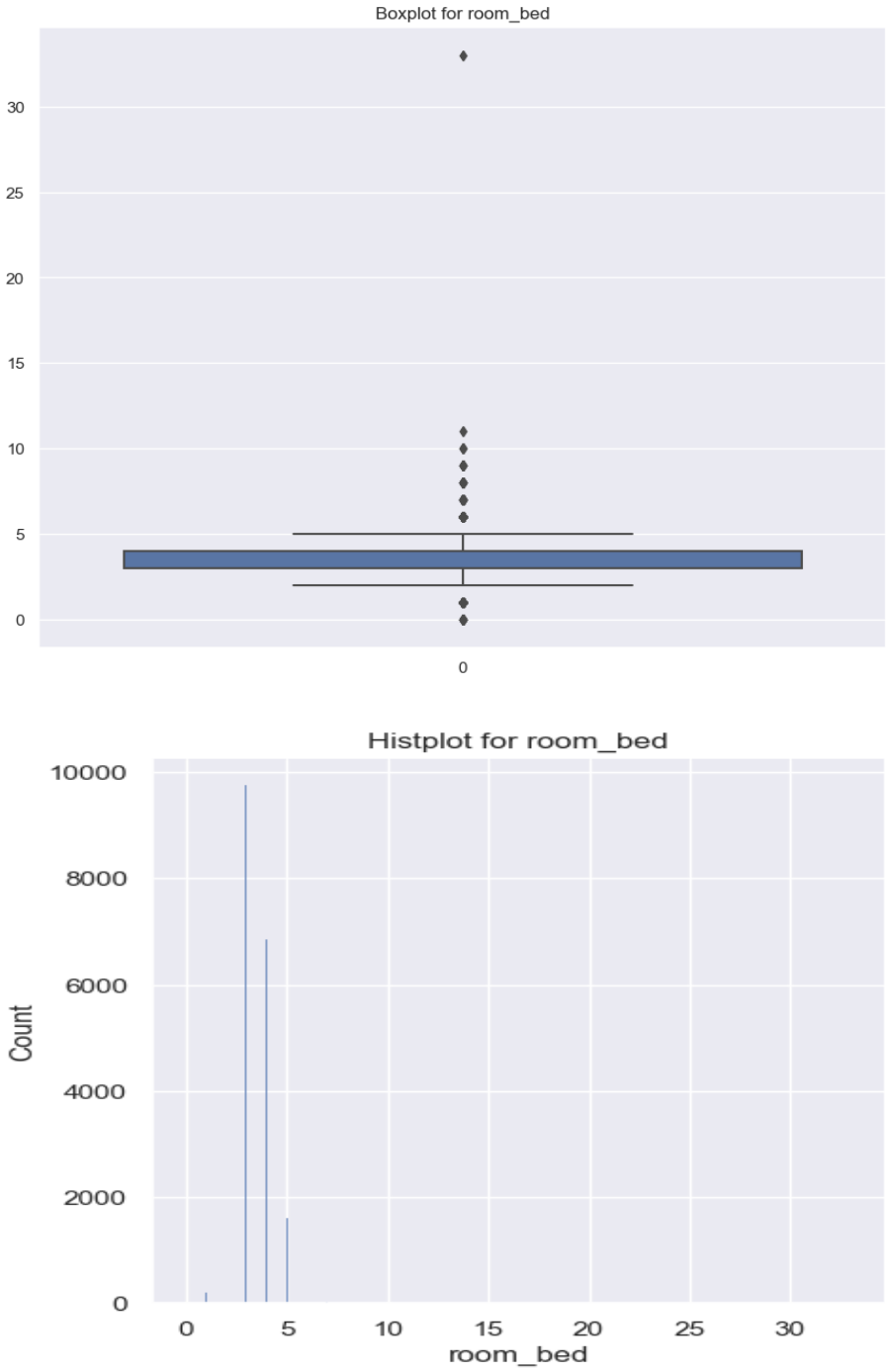
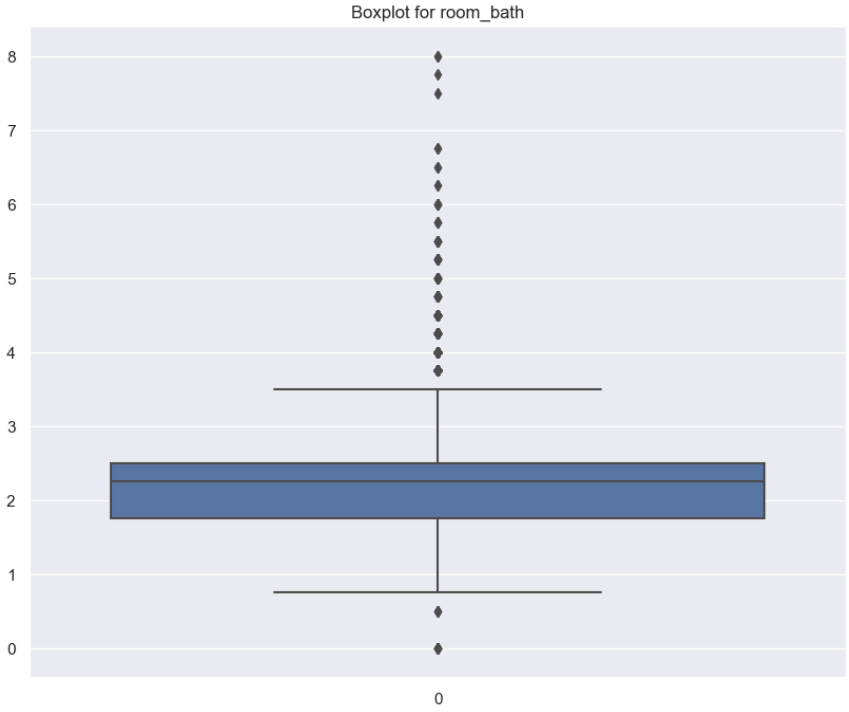


Fig 2 – Boxplot and Histplot for room_bed

We can see that the bedrooms range between 0 and 11, only one house has 33 bedrooms. 50% of the houses have 3 to 4 bedrooms.

The house with 33 bedrooms is valued at 640000 and has a total area of 7620, the price and area for the house do not match with the bedrooms hence we will drop this row.

Boxplot and Histplot for room_bath -



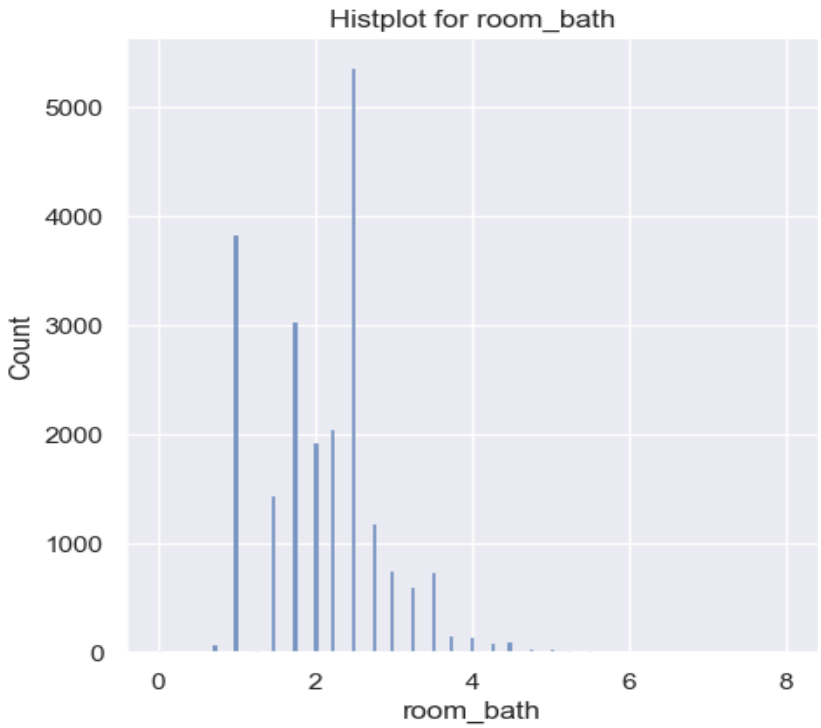
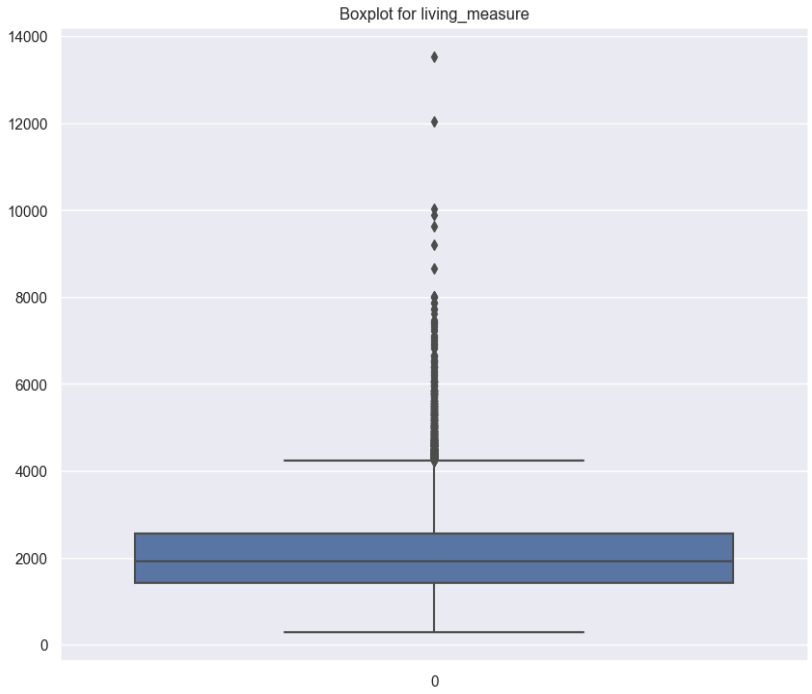


Fig 3 – Boxplot and Histplot for room_bath

Most of the houses have 1 to 2.75 bathrooms. 16 houses have more than 5.75 bathrooms.

Boxplot and Histplot for living_measure -



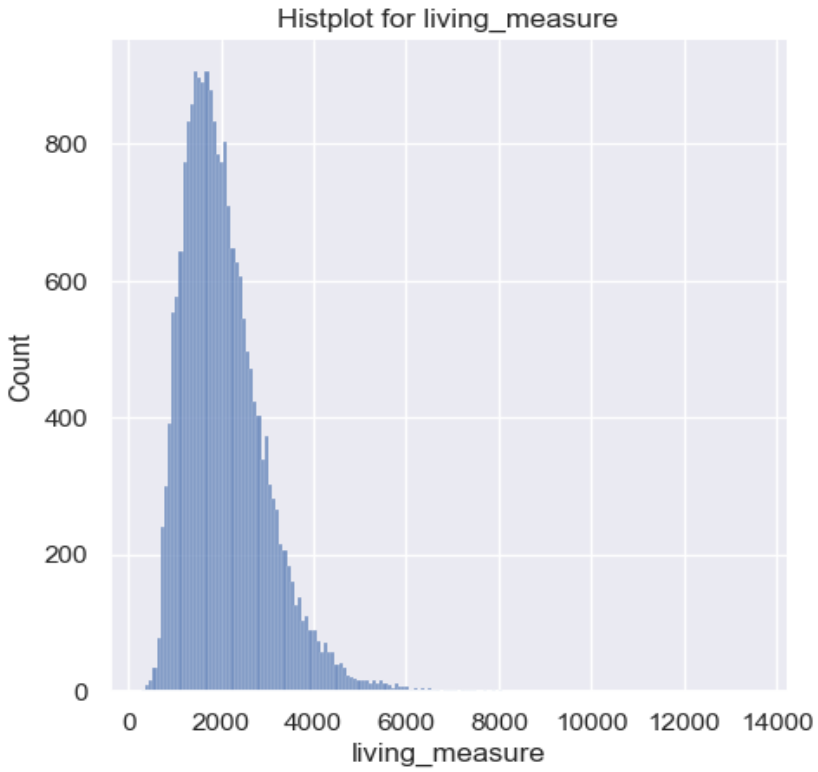


Fig 4 – Boxplot and Histplot for living_measure

We can see that living measure is right skewed. There are 3 properties with greater than 10000 square footage of living area. These are outliers that can be treated. Most of the houses have around 2000 square footage of living area.

Boxplot and Histplot for lot_measure -

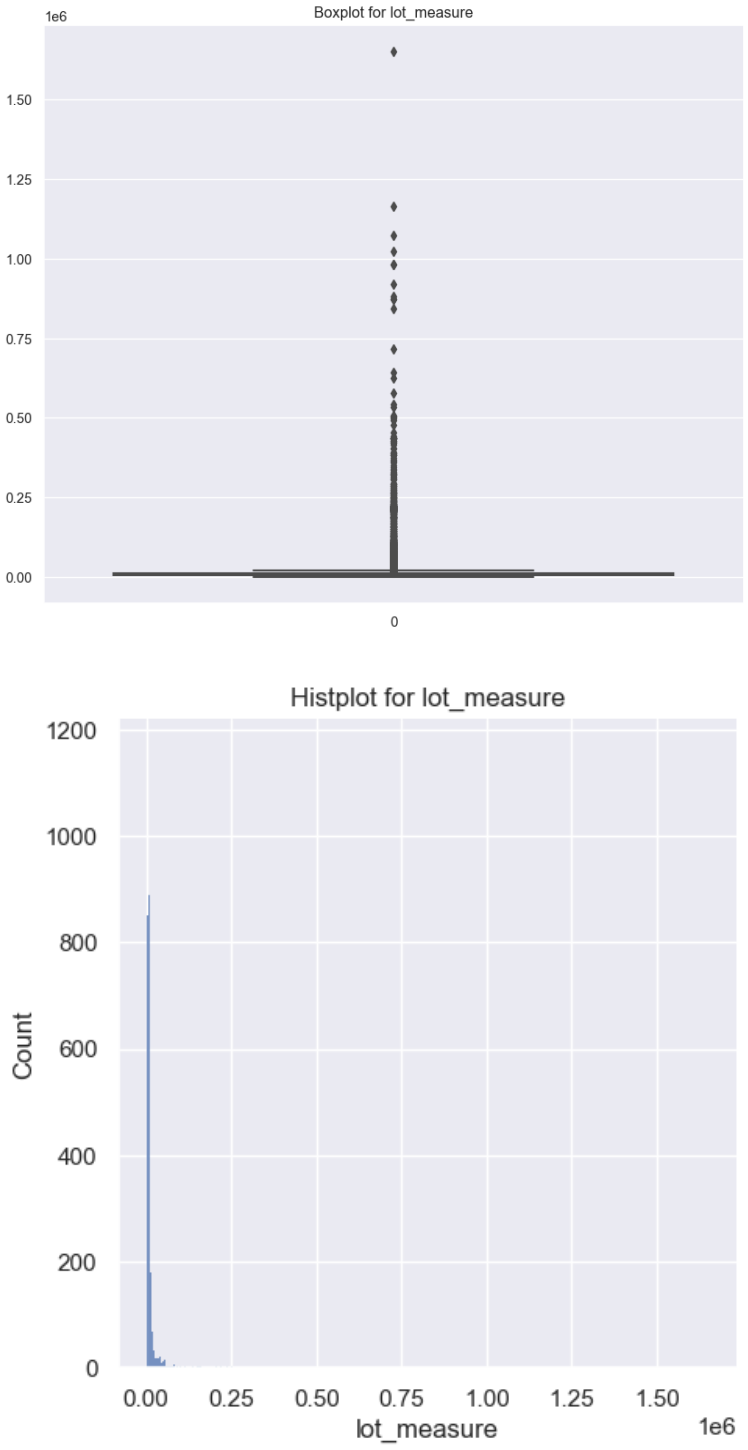
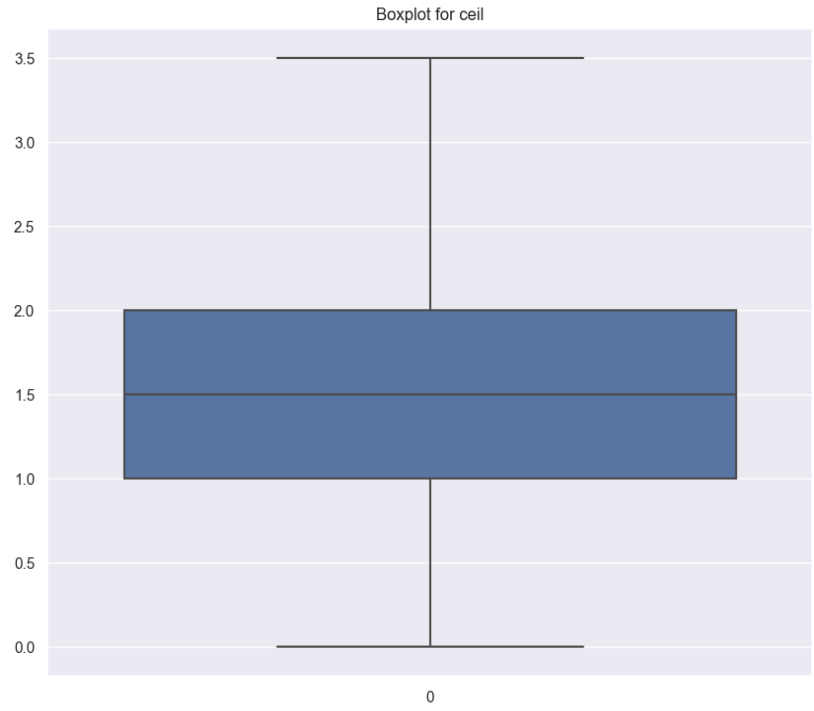


Fig 5 – Boxplot and Histplot for lot_measure

We can see that the data is right skewed. 4 houses have a lot of measure greater than 1000000 square feet, these outliers will be treated.

Boxplot and Histplot for ceil -



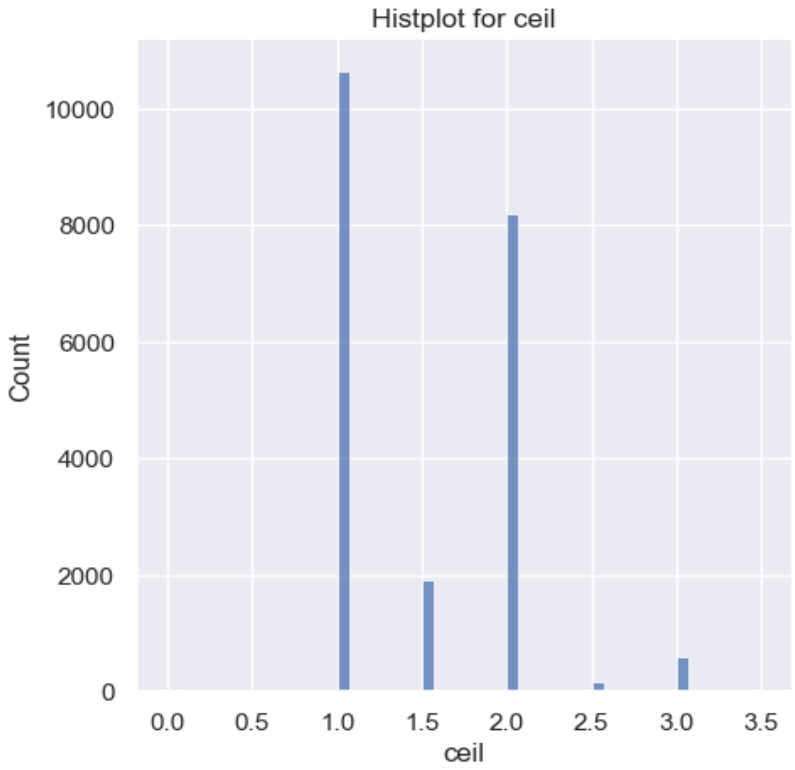


Fig 6 – Boxplot and Histplot for ceil

Most of the houses have between 1 and 2 floors. Only 8 houses have 3.5 floors.

Boxplot and Histplot for coast -

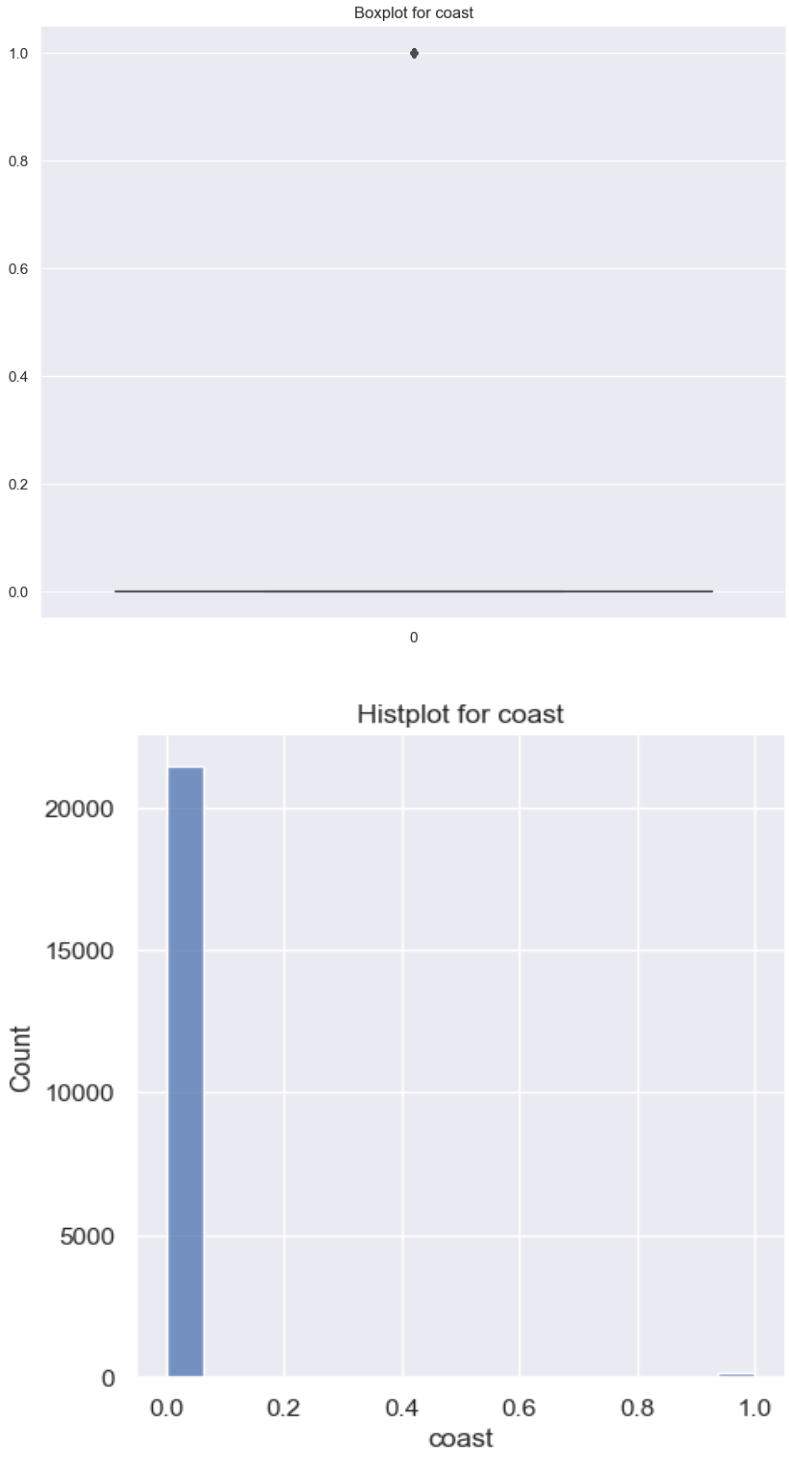
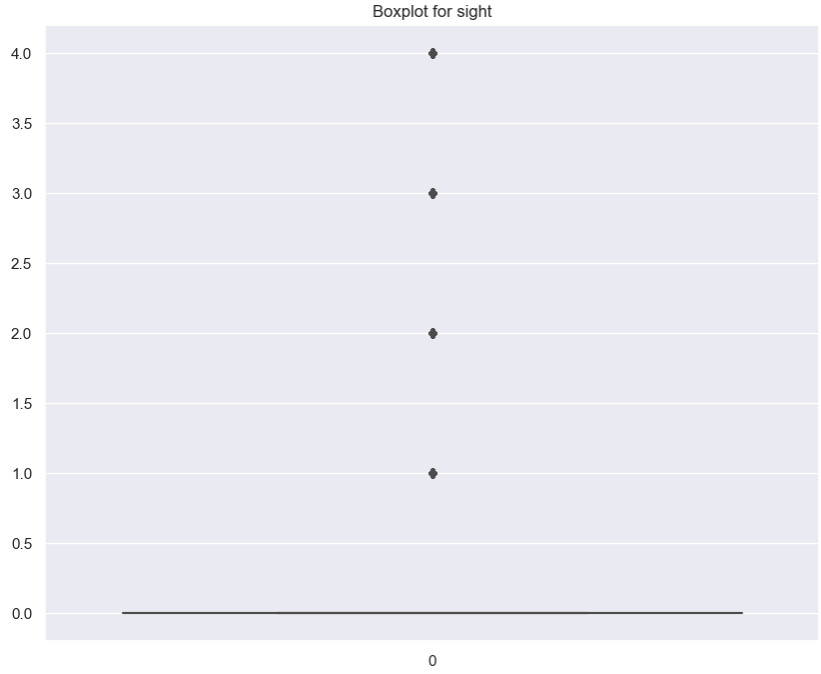


Fig 7 – Boxplot and Histplot for coast

Only 161 houses have a waterfront view.

Boxplot and Histplot for sight -



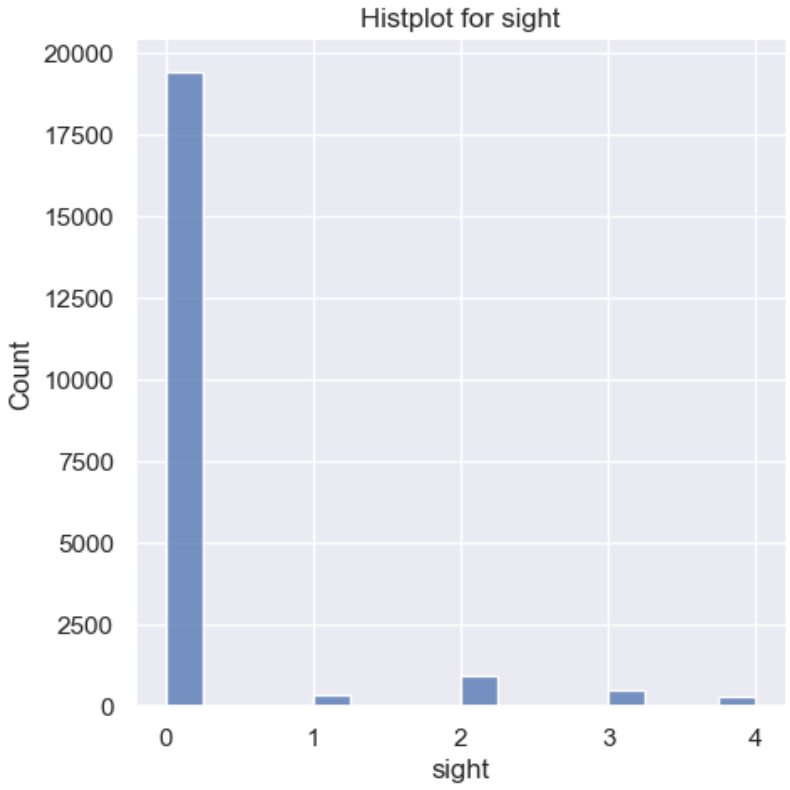
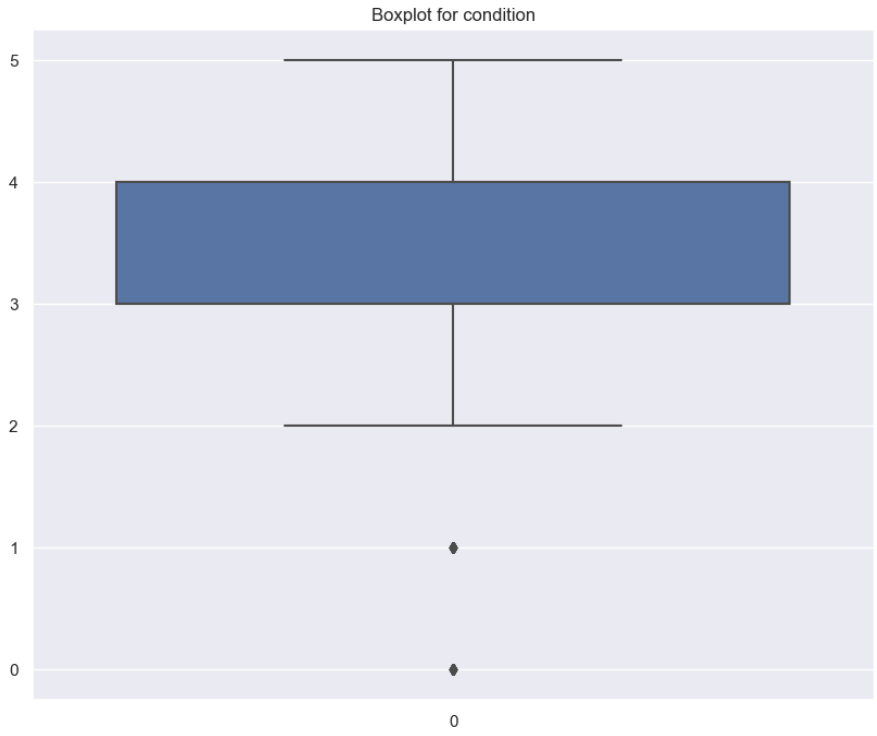


Fig 8 – Boxplot and Histplot for sight

We can see that most of the sights have not been viewed even once.

Boxplot and Histplot for condition -



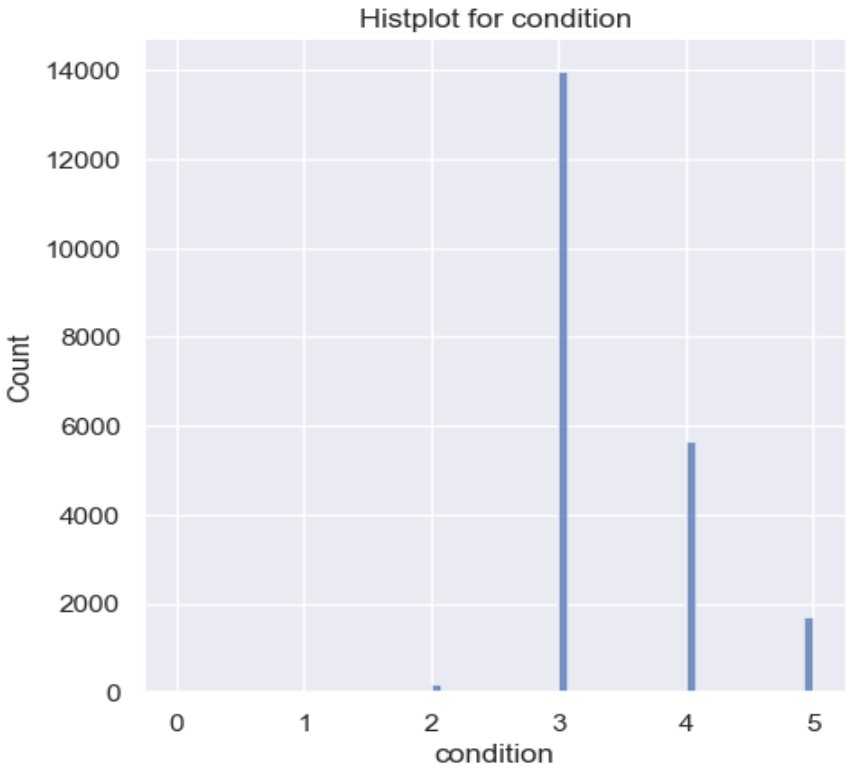
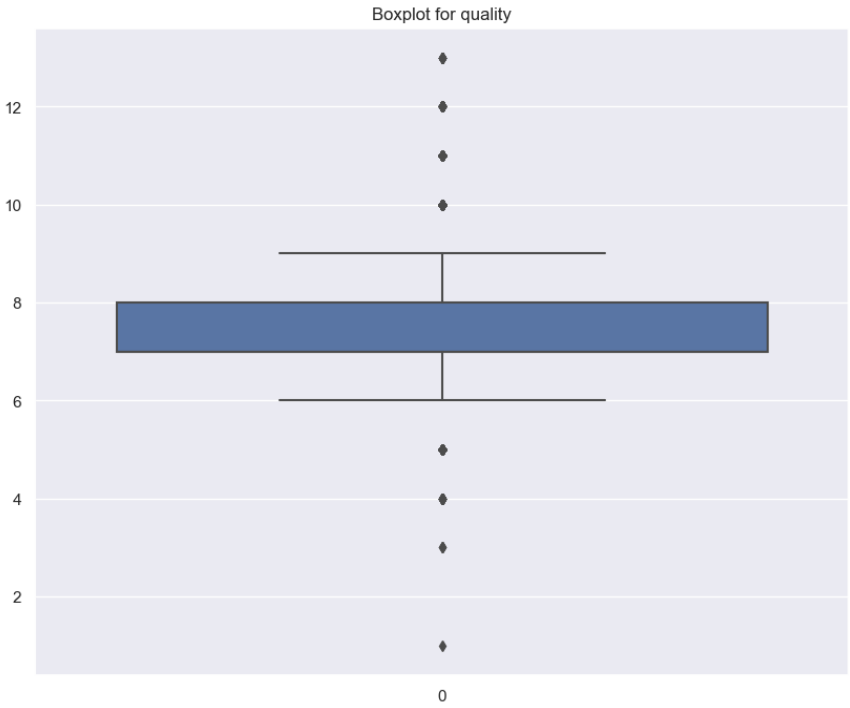


Fig 9 – Boxplot and Histplot for condition

Most of the properties have got an average rating of 3.

28 properties have a rating of 0.

Boxplot and Histplot for quality -



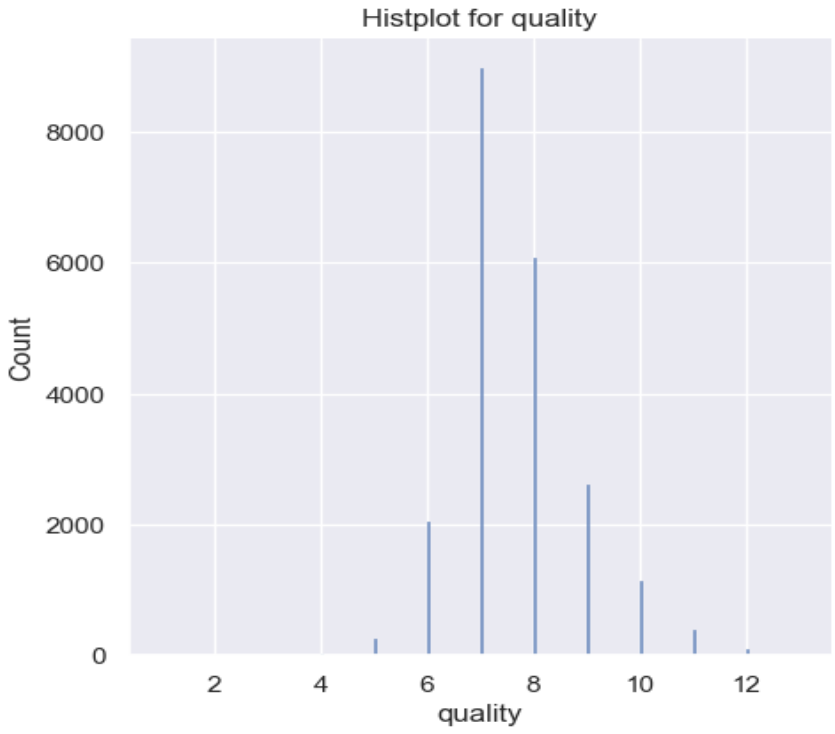


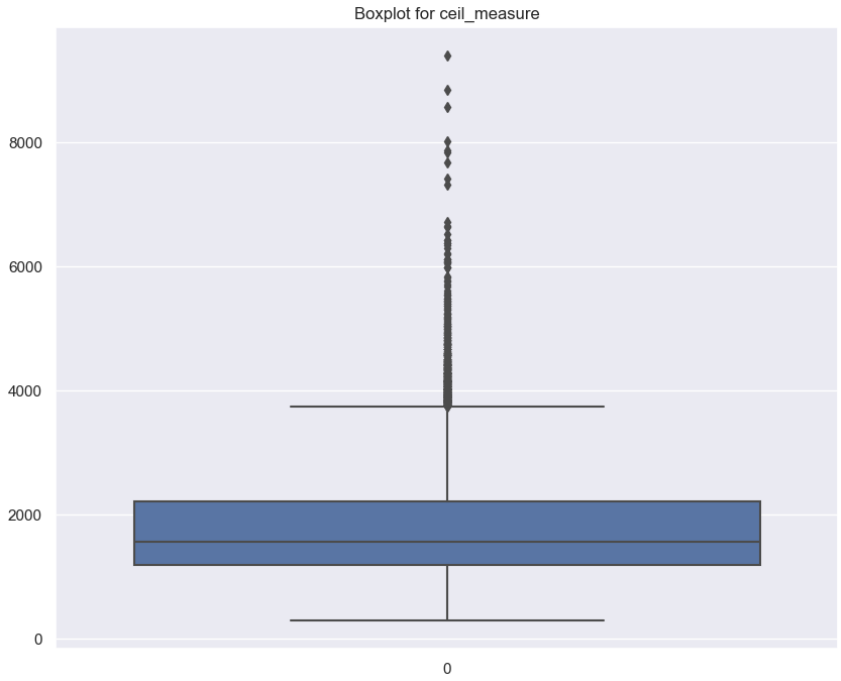
Fig 10 – Boxplot and Histplot for quality

Most of the houses have a rating between 6 and 10.

4 houses have a rating of less than 3.

13 houses have the highest rating of 13.

Boxplot and Histplot for ceil_measure -



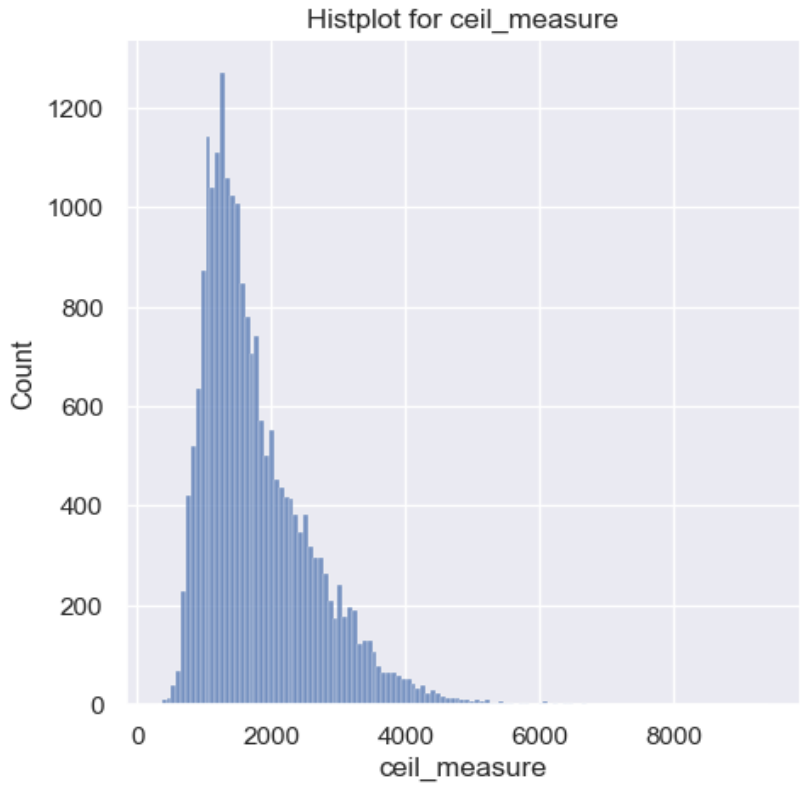


Fig 11 – Boxplot and Histplot for ceil_measure

We can see that the data is right skewed. 4 houses have more than 8000 ceiling measurements.

Most of the houses have 1000 to 2500 ceil measure.

Boxplot and Histplot for basement_measure -

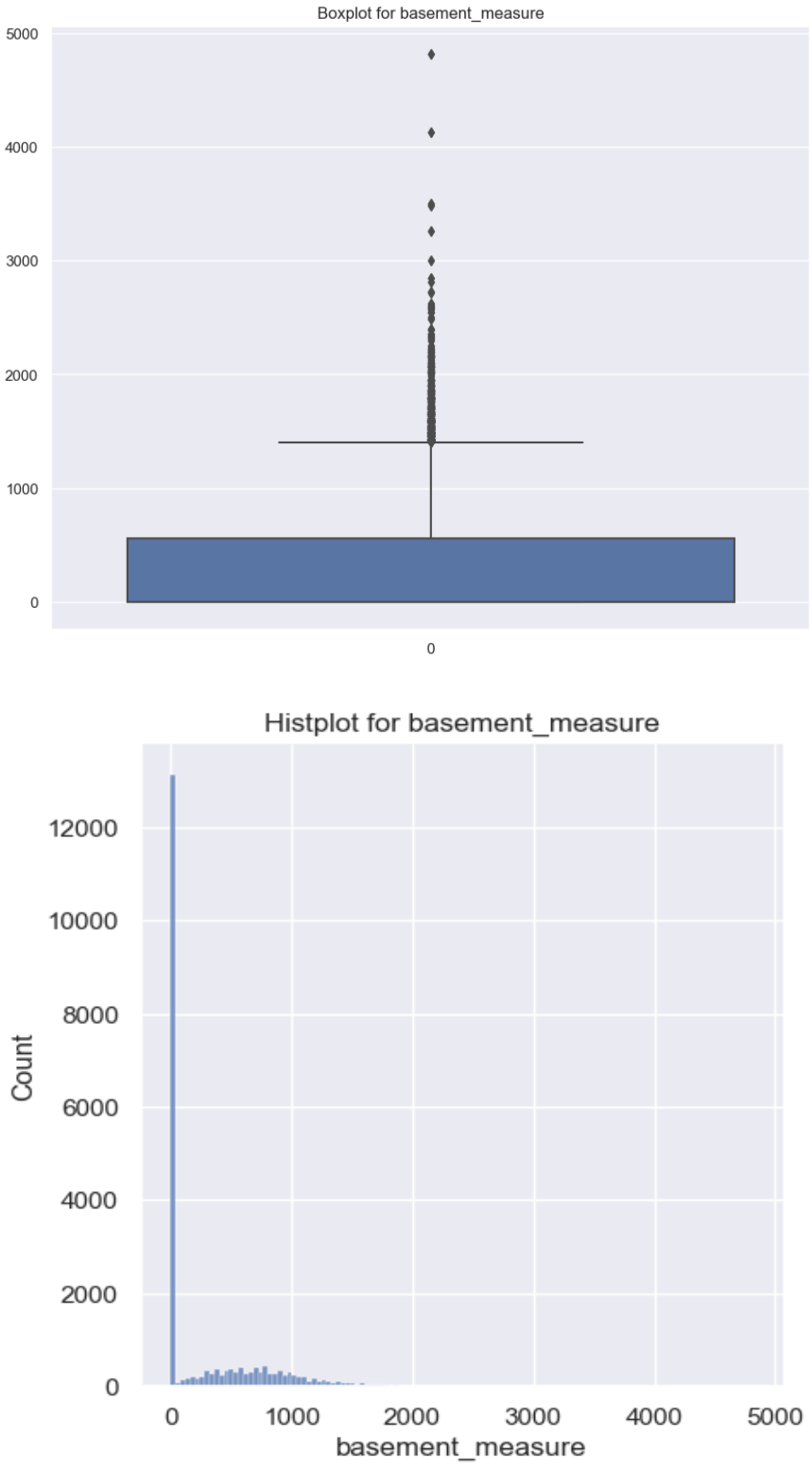
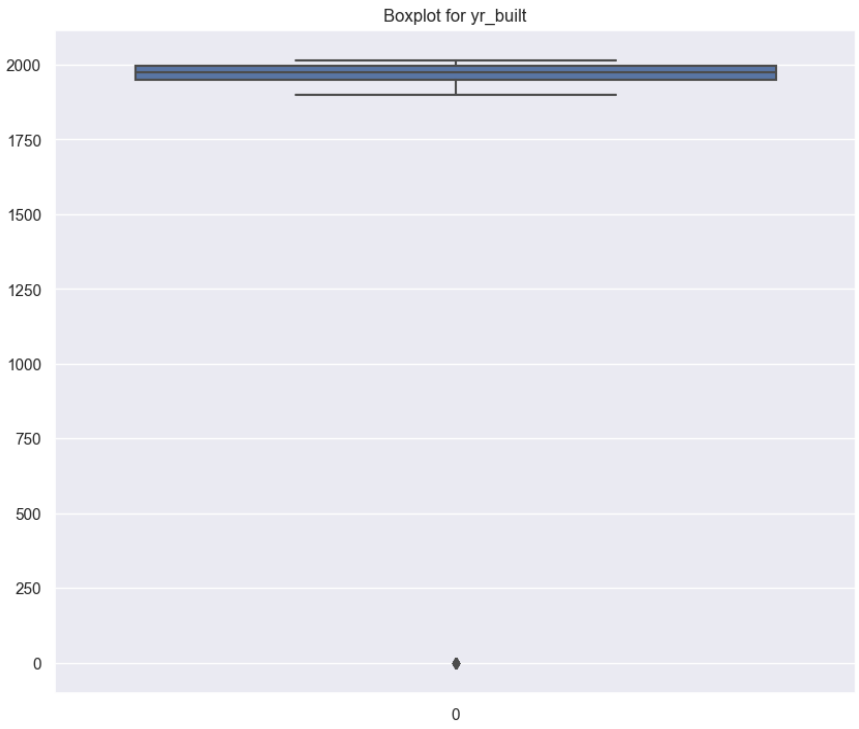


Fig 12 – Boxplot and Histplot for basement

We can see that most of the properties do not have a basement. Only 5 properties have a basement greater than 3000 square footages.

Boxplot and Histplot for yr_built -



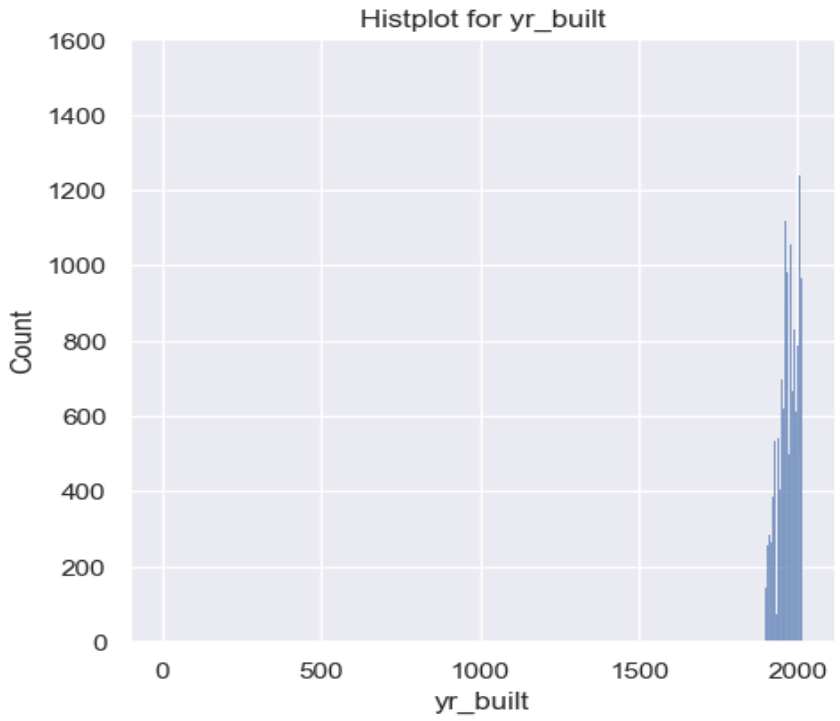
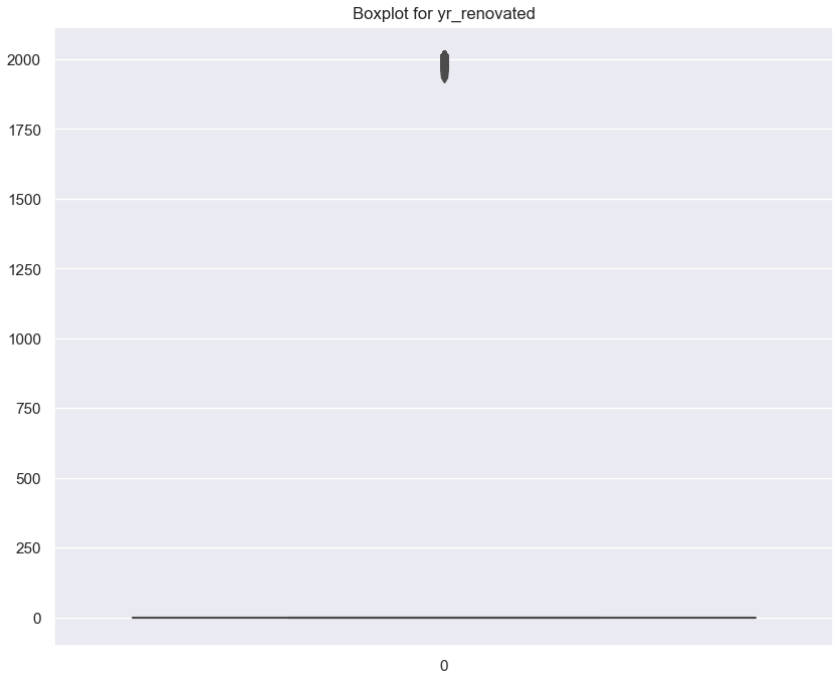


Fig 13 – Boxplot and Histplot for yr_built

Most of the houses were built between 1900 and 2015.

14 houses with null values will be changed.

Boxplot and Histplot for yr_renovated -



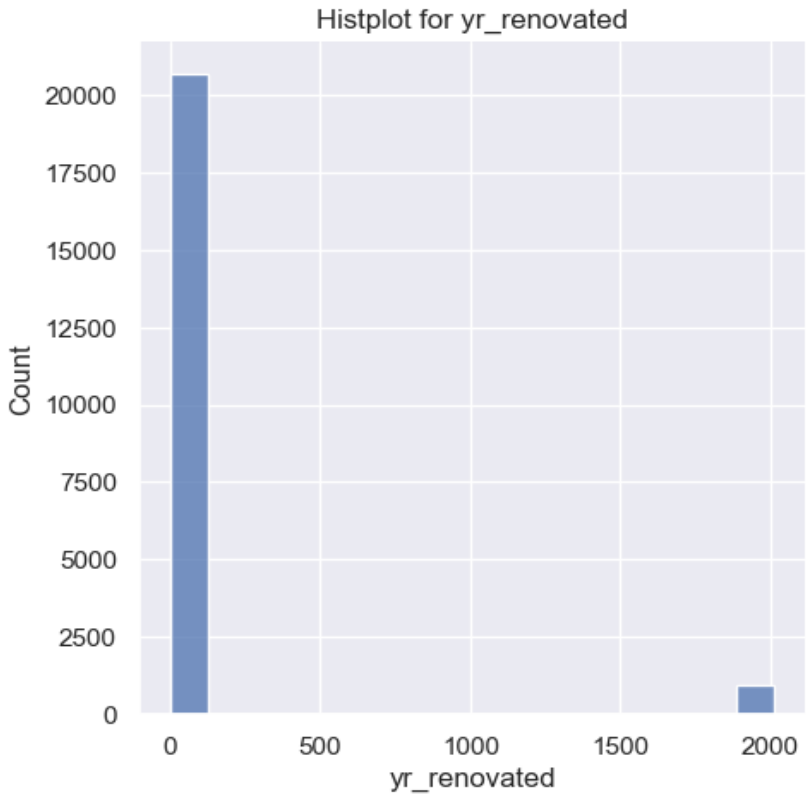
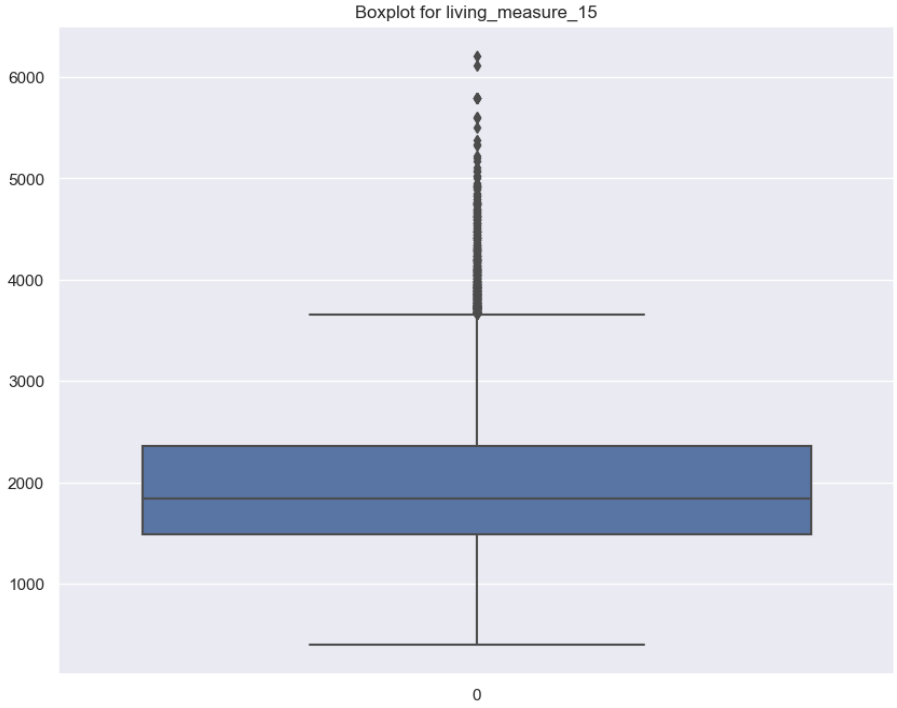


Fig 14 – Boxplot and Histplot for yr_renovated

Almost 20000 houses have not been renovated. Houses have been renovated between 1934 and 2015.

Boxplot and Histplot for living_measure15 -



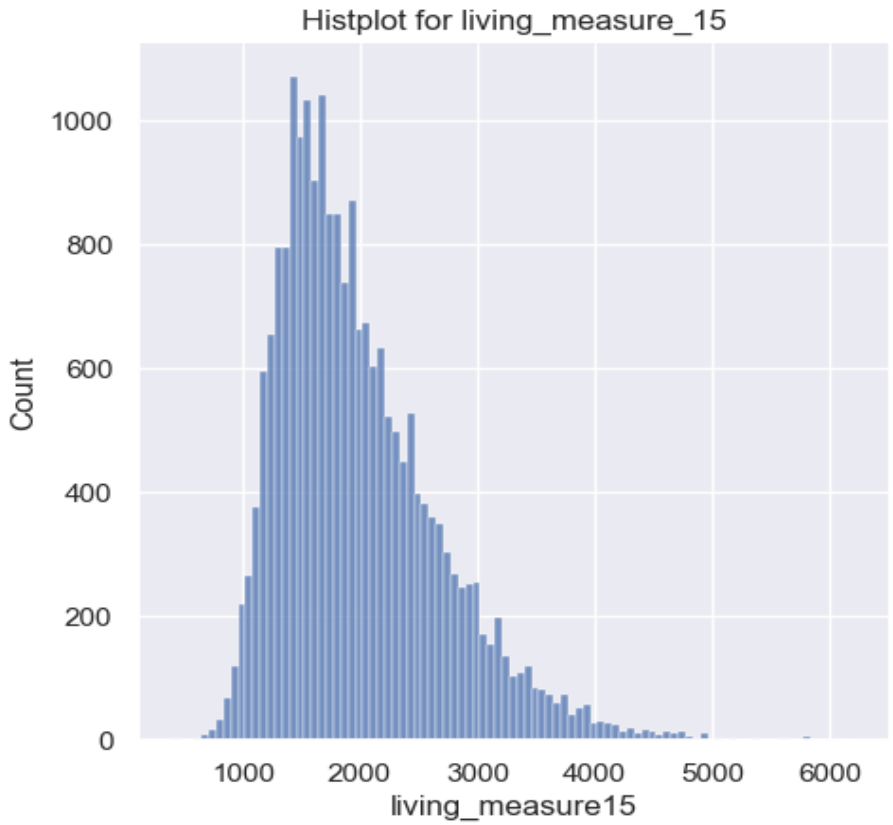
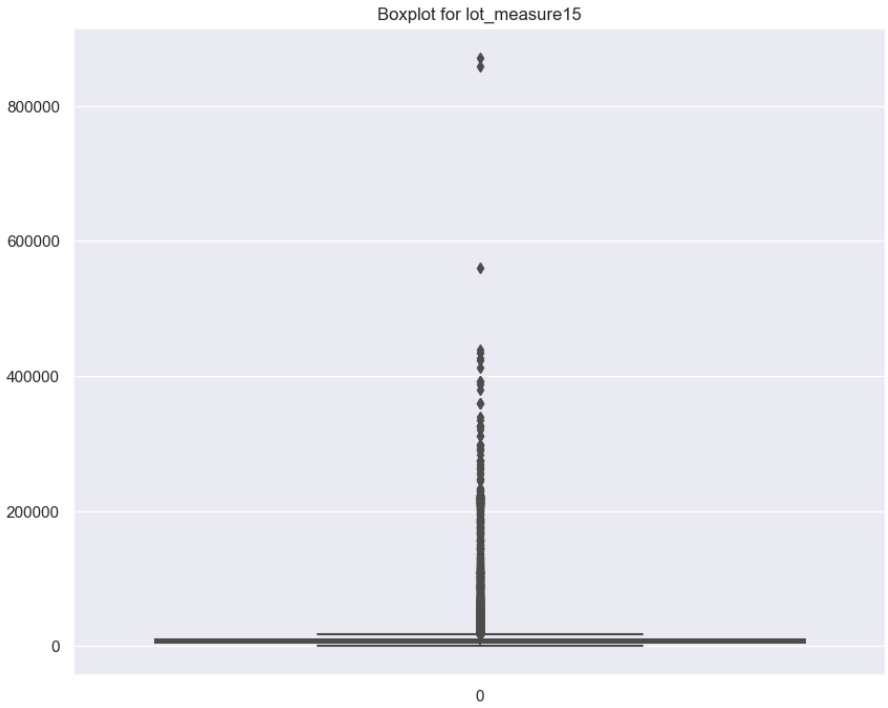


Fig 15 – Boxplot and Histplot for living_measure15

We can see that the data is right skewed. There are 2 houses with more than 6000 square footage of living area. Most of the houses have square footage of around 2000.

Boxplot and Histplot for lot_measure15 -



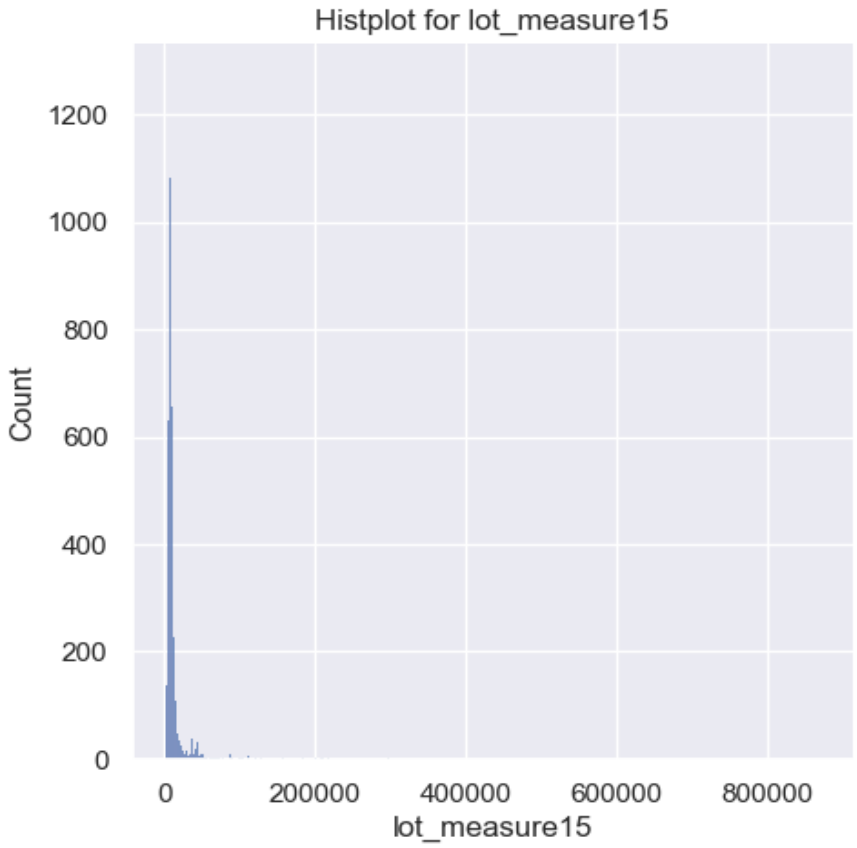
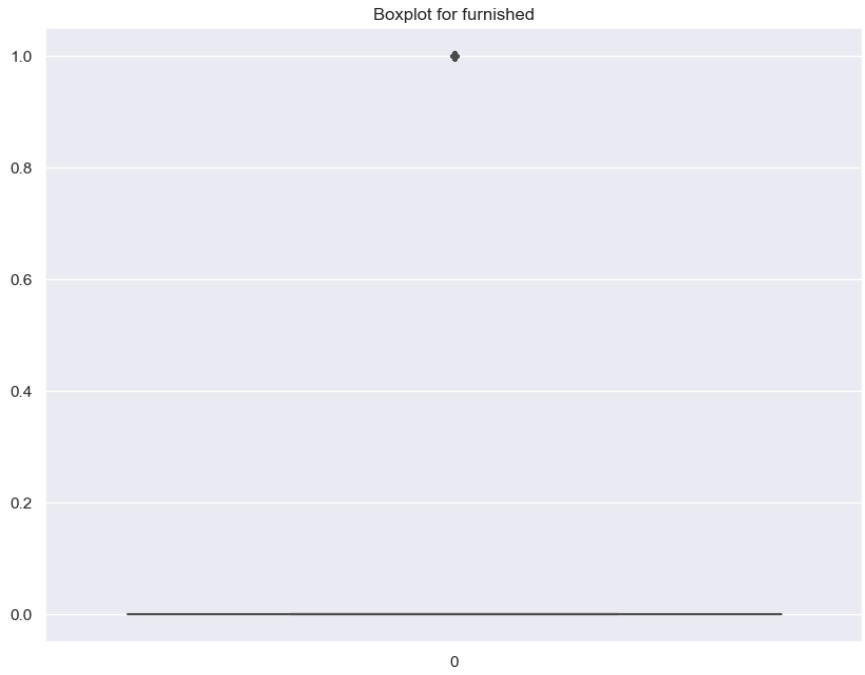


Fig 16 – Boxplot and Histplot for lot_measure15

The data is right skewed. There are 3 houses with more than 500000 square footage of lot measure.

Boxplot and Histplot for furnished -



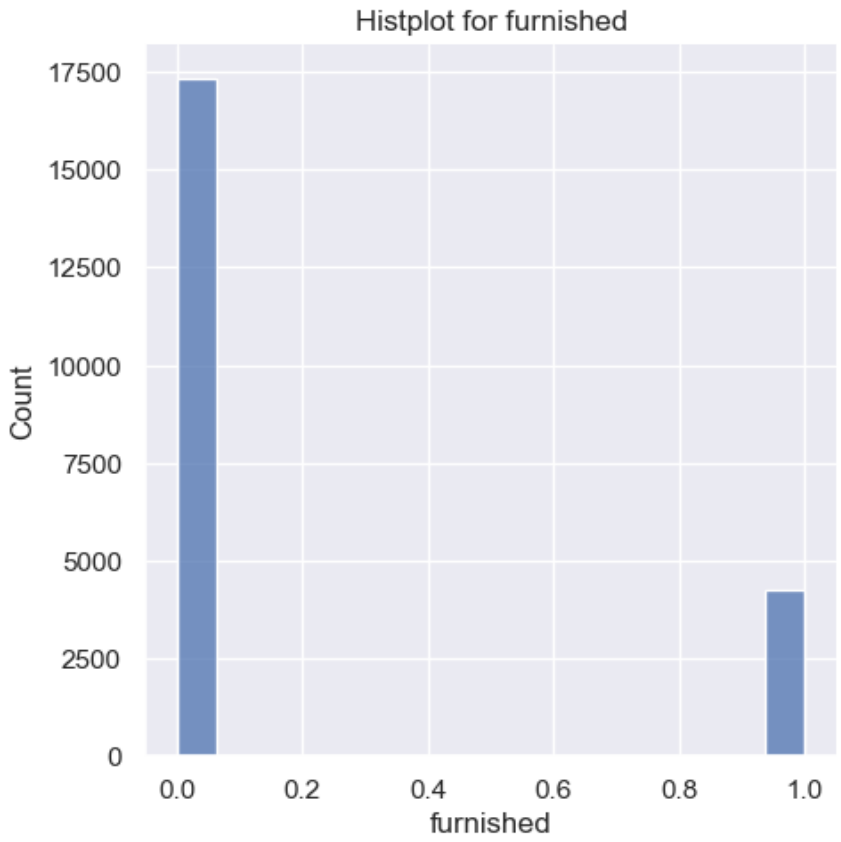


Fig 17 – Boxplot and Histplot for furnished

We can see that 4246 houses have been furnished. More than 50% of the houses have not been furnished.

Boxplot and Histplot for total_area -

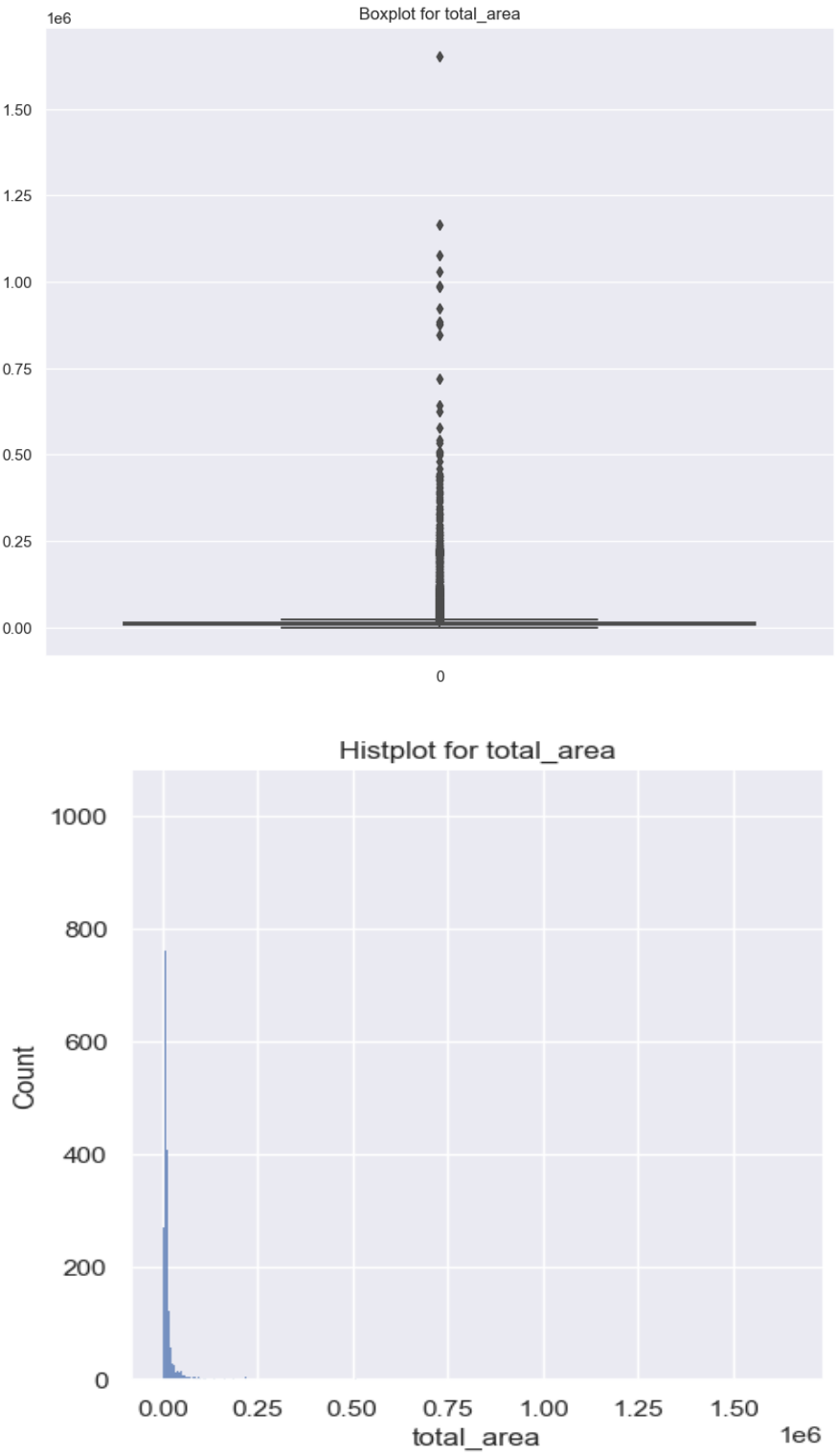


Fig 18 – Boxplot and Histplot for total_area

We can see that the distribution is right skewed. There are 4 houses with more than 1000000

square foot of total area.

Bivariate Analysis -

Pair plot -

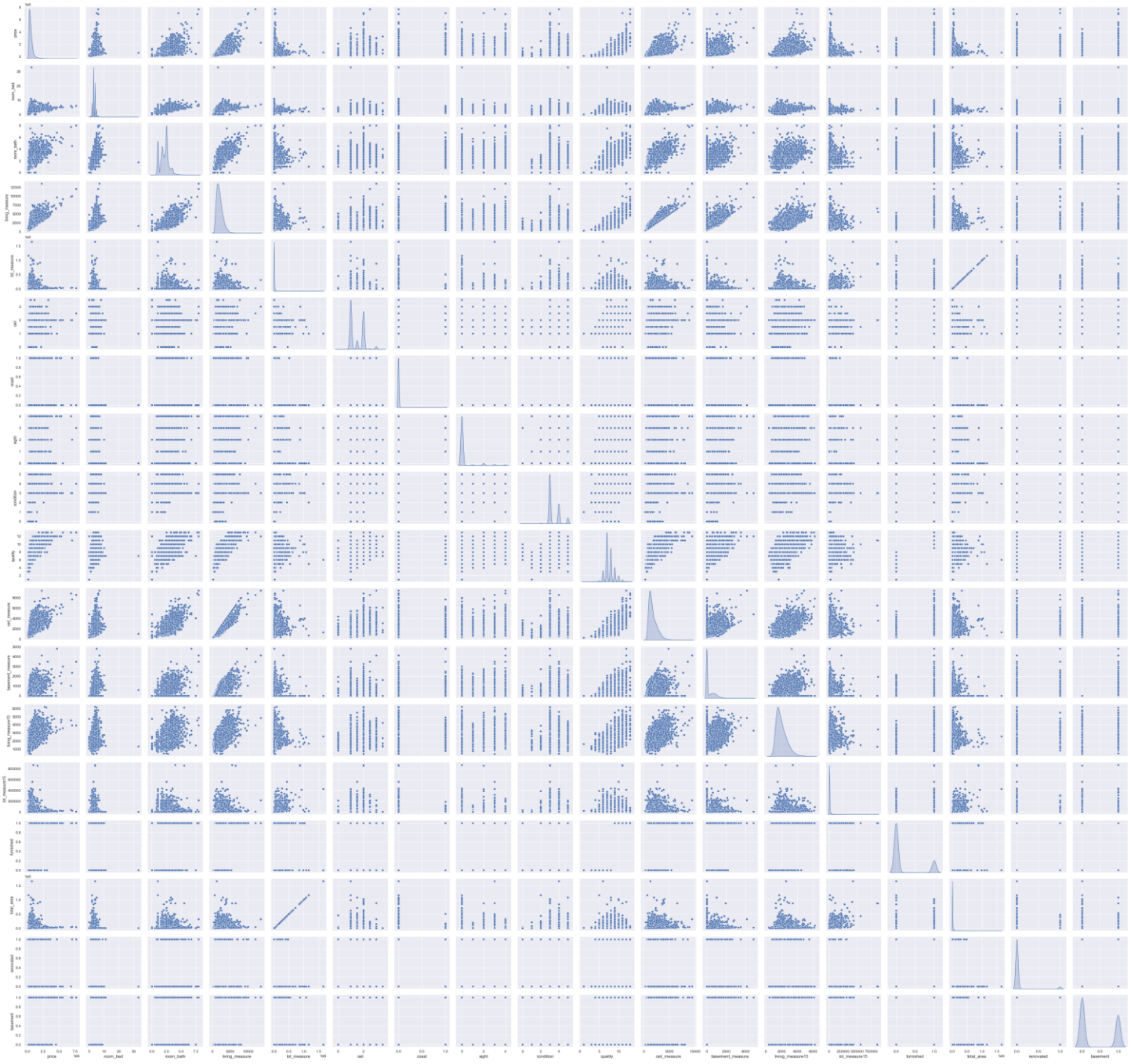


Fig 19 – Pair plot

We can see that price is right skewed. Most of the variables don’t have a clear relationship with price variables.

we will have to convert some of these variables into categorical.

Pearson correlation -

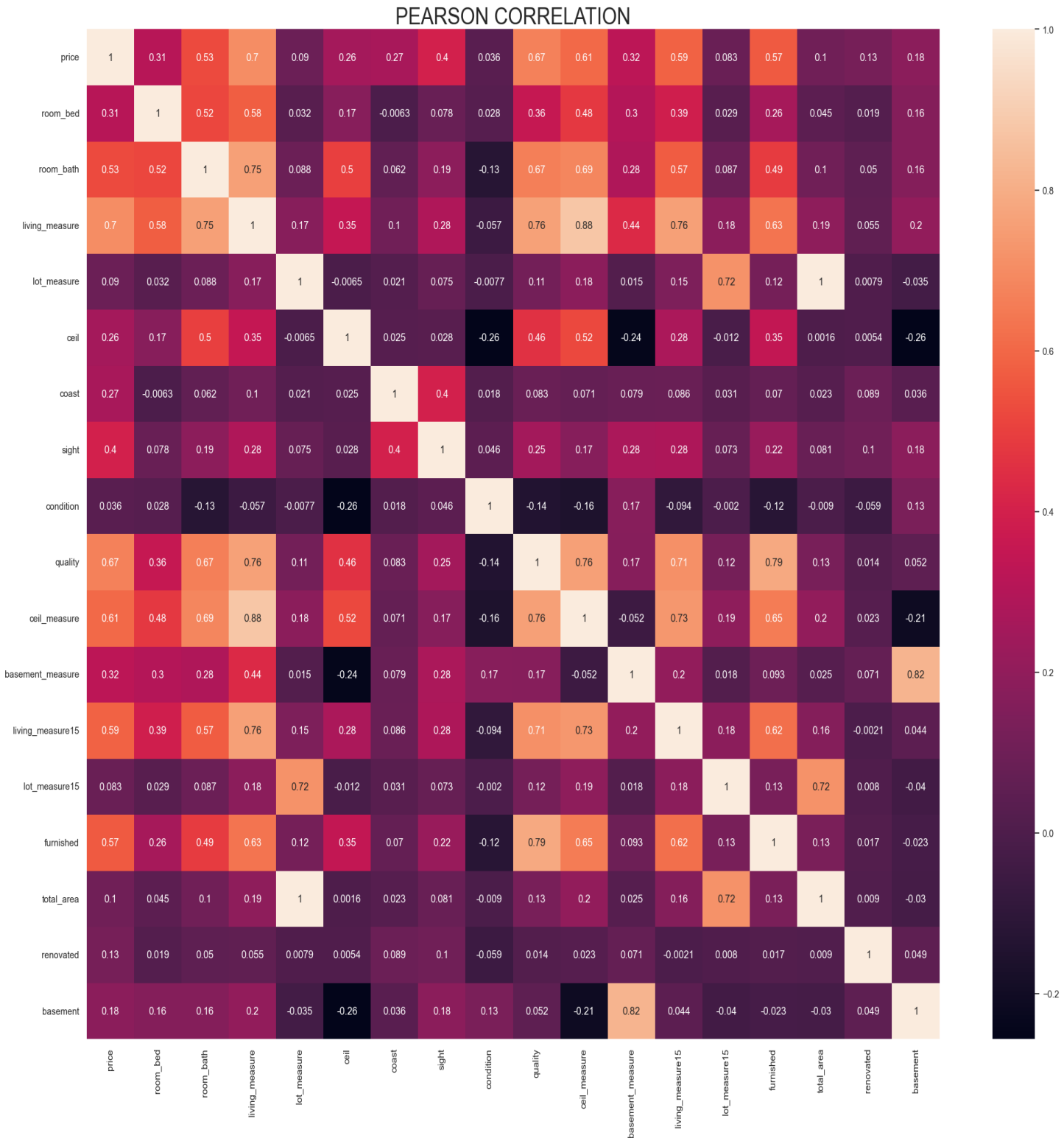


Fig 20 – Pearson Correlation

From the above heatmap we can see that most of the variables are correlated to each other.

Bivariate analysis of price and room_bed -

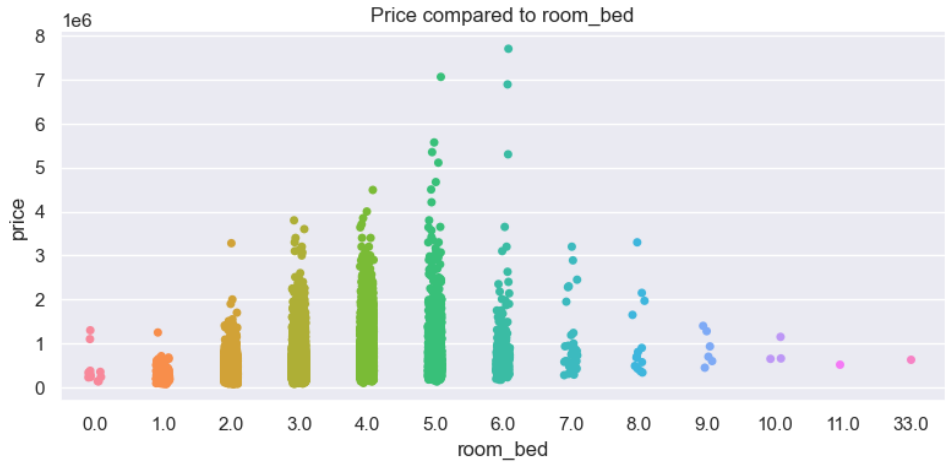


Fig 21 – Price compared to room_bed

We can see that price gives us an increasing trend up to a certain number of rooms.

Bivariate analysis of price and room_bath -

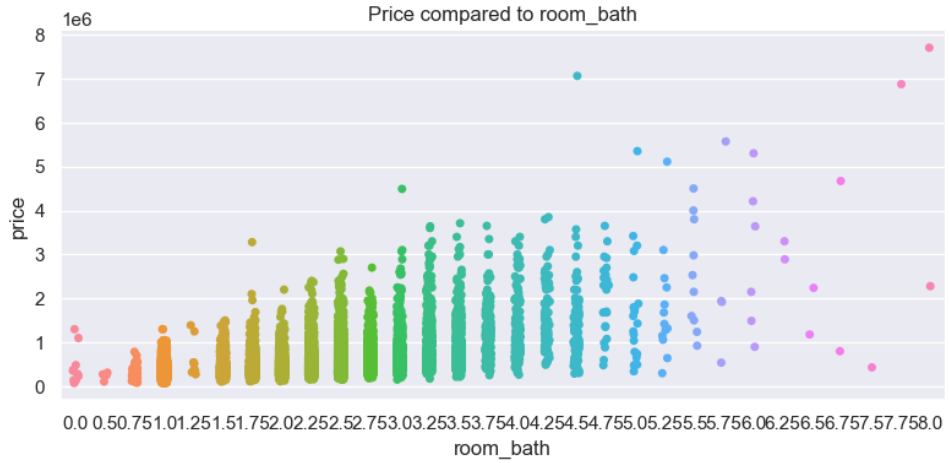


Fig 22 – Price compared to room_bath

We can see that price gives us an increasing trend up on increase in the number of bathrooms.

Bivariate analysis of price and ceil -

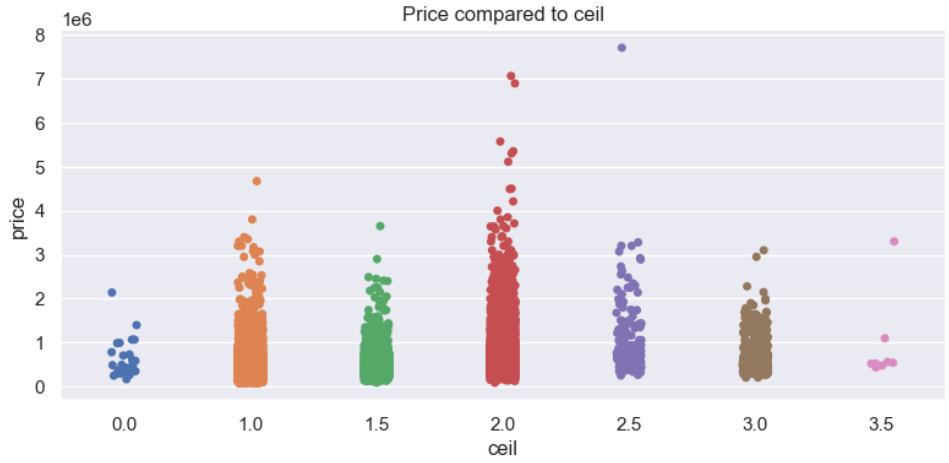


Fig 23 – Price compared to ceil

We can see that there is some upward trend in price upon increase of ceiling levels.

Bivariate analysis of price and coast-

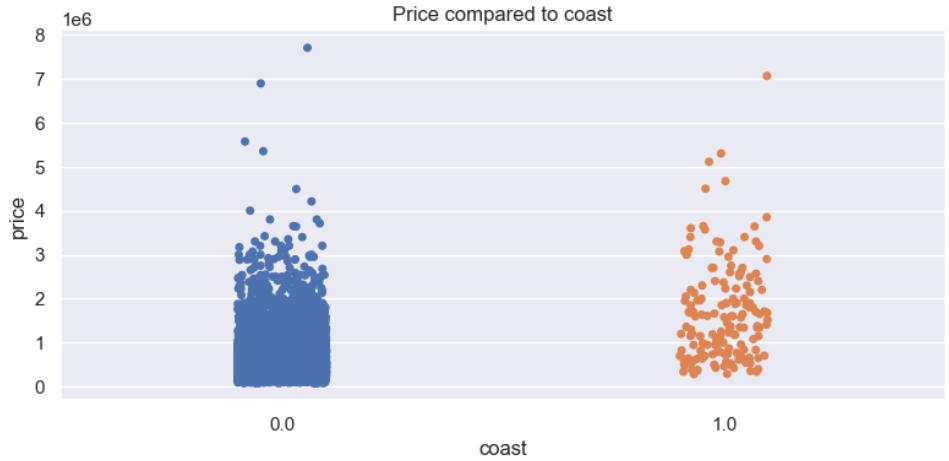


Fig 24 – Price compared to coast

There is a slight increase in price for houses with a waterfront view.

Bivariate analysis of price and quality-

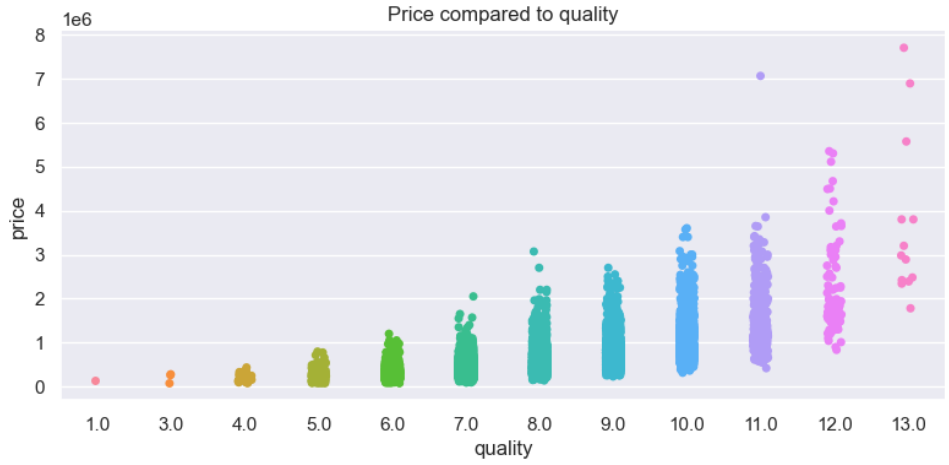


Fig 25 – Price compared to quality

We can see an upward trend in price on better quality rating.

Latitude and Longitude of houses\properties -

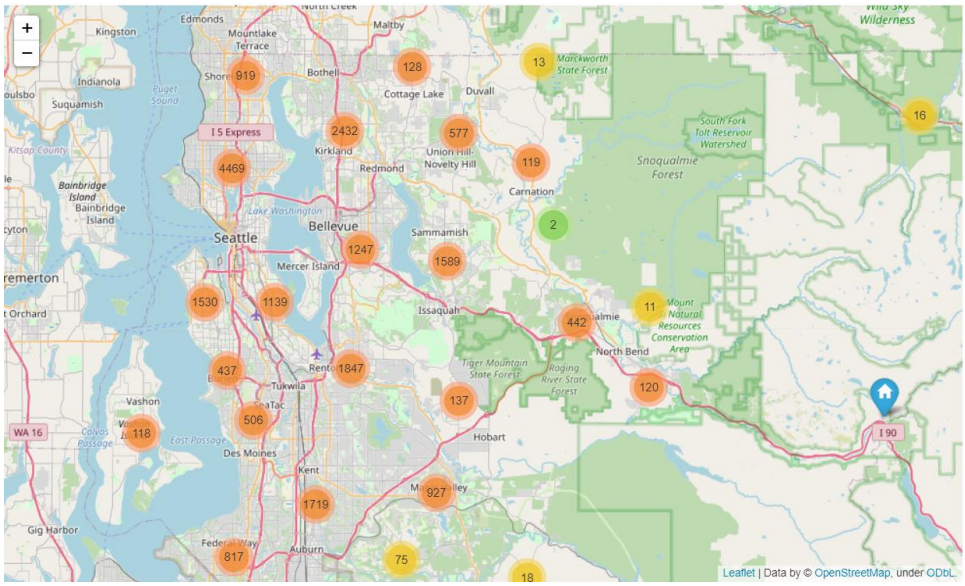


Fig 26 – Lat and Long of House Properties

There are 21579 properties spread across Seattle USA.

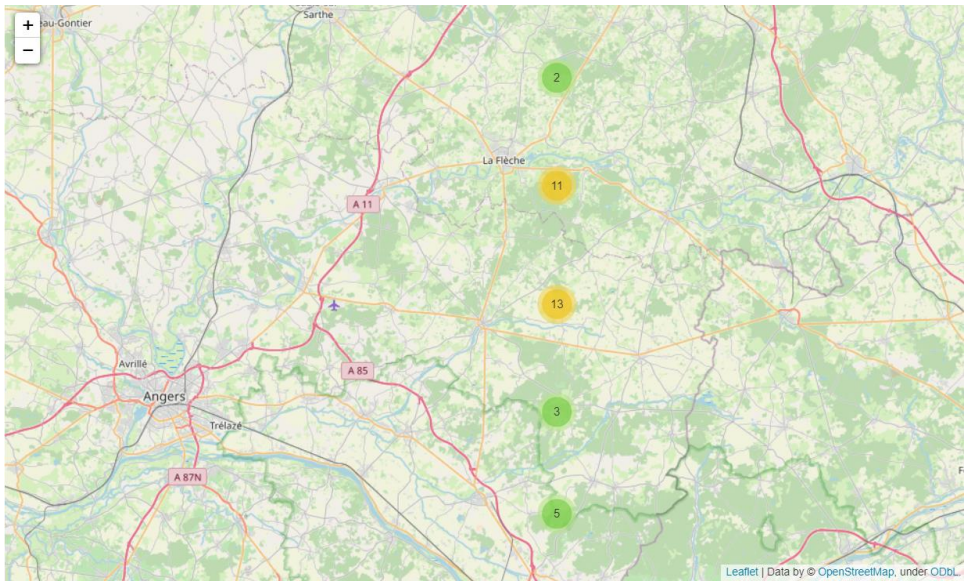


Fig 27 – Lat and Long of House Properties

We can see that there are also 34 properties in France.

Business Implications -

- Target variable price is right skewed with a min value of 75,000 and a max value of 11,29,744.
- Most houses had between 2 to 5 bedrooms.
- Most houses had between 1 to 2.75 bathrooms.
- There were around 10 houses with 0 bedrooms or bathrooms which can be small houses called studio apartments.
- Living square footage is right skewed with 3 houses with more than 10,000 Square footage of living space.
- Lot Square footage is right skewed with 4 houses with more than 10,00,000 Square footage of lot space.
- Most of the houses have 1 to 2 floors.
- Only 161 houses have a waterfront view.
- 19,437 properties have not been viewed even once.
- Most of the houses have a 3.0 overall condition out of 5.0.
- Most houses have got a rating between 6 and 10 out of 13, based on the grading system.
- Only 914 houses have been renovated.
- 4,246 houses are furnished, they have a better quality of rooms.
- The price of houses shows an upward trend on increase in bedrooms and bathrooms.
- The price of houses shows an upward trend on increase grade given to the housing unit, based on grading system.

3) Data Cleaning and Preprocessing -

Null values and Outliers in the data -

There were 403 null values among which some were dropped, and some were replaced.

Most of the variables that had outliers were treated with IQR.

IQR is simply the range of the middle 50% of data values, **it's not affected by extreme outliers**.

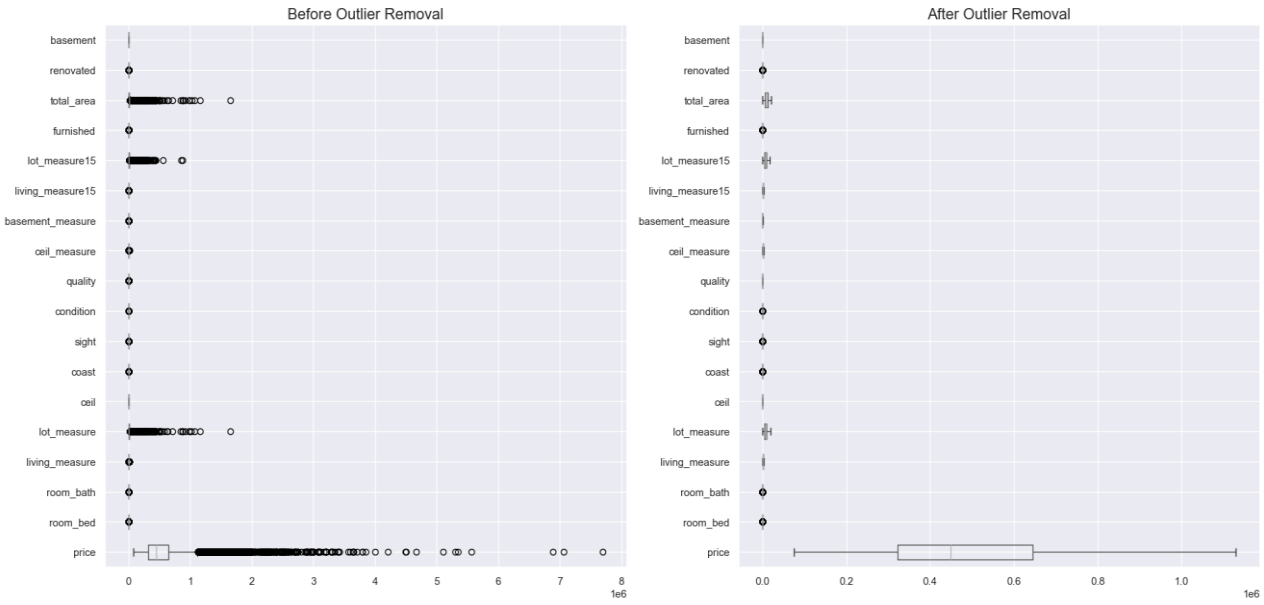


Fig 28 – Outlier Treatment

- Two new variables Basement and Renovated have been created. These tell us whether a house has a basement and whether a house is renovated.
- Seven unwanted variables have been dropped which were not required for our analysis.
- Label encoding has been done on ceil, room_bed and room_bath columns.
- Final data consisted of 21538 rows and 21 columns.

Splitting the data -

The data training and test set is split in 75:25 ratio.

X Train shape is: (16153, 20)
X Test shape is: (5385, 20)

y Train shape is: (16153, 1)
y Test shape is: (5385, 1)

4) Model Building -

We will be using various regression and classification models for this problem. Following are the models used and the accuracies and RMSE values.

1) Linear Regression Model -

- Accuracy on Train Set for Linear Regression Model was 63.2%
- Accuracy on Test Set for Linear Regression Model was 64.4%
- RMSE on Train Set for Linear Regression Model was 151575.04
- RMSE on Test Set for Linear Regression Model was 148767.85

The Linear Regression Model gave us an accuracy of 63.2% on the training set and 64.4% on the test set. The RMSE value shows that the model is underfitting.

2) Lasso Linear Regression Model -

- Accuracy on Train Set for Lasso Regression Model was 63.3%
- Accuracy on Test Set for Lasso Regression Model was 64.4%
- RMSE on Train Set for Lasso Regression Model was 151475.04
- RMSE on Test Set for Lasso Regression Model was 148767.85

The Lasso Linear Regression Model gave us an accuracy of 63.3% on the training set and 64.4% on the test set. The RMSE value shows that the model is underfitting.

3) Ridge Linear Regression Model -

- Accuracy on Train Set for Ridge Regression Model was 63.3%
- Accuracy on Test Set for Ridge Regression Model was 64.4%
- RMSE on Train Set for Ridge Regression Model was 151475.05
- RMSE on Test Set for Ridge Regression Model was 148767.99

The Ridge Linear Regression Model gave us an accuracy of 63.3% on the training set and 64.4% on the test set. The RMSE value shows that the model is underfitting.

Linear Regression Model, Lasso Linear Regression Model and Ridge Linear Regression Model had similar accuracies but had a slight change in RMSE, all three were underfitting. Overall, the three models performed similarly.

4) KNN Regression Model -

- Accuracy on Train Set for KNN Regression Model was 61.5%
- Accuracy on Test Set for KNN Regression Model was 58.1%

- RMSE on Train Set for KNN Regression Model was 155160.69
- RMSE on Test Set for KNN Regression Model was 158318.79

The KNN Regression Model gave us an accuracy of 61.5% on the training set and 58.1% on the test set. The RMSE value for test set is higher than the train set which indicates the model is overfitting.

5) Decision Tree Regression Model -

- Accuracy on Train Set for Decision Tree Regression Model was 99.9%
- Accuracy on Test Set for Decision Tree Regression Model was 46.1%
- RMSE on Train Set for Decision Tree Regression Model was 9439.80
- RMSE on Test Set for Decision Tree Regression Model was 7113.90

The Decision Tree Regression Model gave us an accuracy of 99.9% on the training set and 46.1% on the test set. The RMSE value for the train set is higher than the test set which indicates the model is underfitting. The Decision Tree Regression Model performed well in the train set but underperformed on the test set.

6) Random Forest Regression Model -

- Accuracy on Train Set for Random Forest Regression Model was 79.5%
- Accuracy on Test Set for Random Forest Regression Model was 70.5%
- RMSE on Train Set for Random Forest Regression Model was 113519.97
- RMSE on Test Set for Random Forest Regression Model was 114605.30

The Random Forest Regression Model gave us an accuracy of 79.5% on the training set and 70.5% on the test set. The RMSE value for the test set is higher than the train set which indicates the model is overfitting. This model performed well on the training set and gave us the best test set accuracy yet.

7) Gradient Boost Regression Model -

- Accuracy on Train Set for Gradient Boost Regression Model was 77.6%
- Accuracy on Test Set for Gradient Boost Regression Model was 71.2%
- RMSE on Train Set for Gradient Boost Regression Model was 118224.93
- RMSE on Test Set for Gradient Boost Regression Model was 100205.62

The Gradient Boost Regression Model gave us an accuracy of 77.6% on the training set and 71.2% on the test set. The RMSE value for the train set is higher than the test set which indicates the model is underfitting. This is the best model so far, hyper tuning this model can give us better accuracy.

Model Tuning -

8) Bagging Regression Model -

- Accuracy on Train Set for Bagging Regression Model was 94.2%
- Accuracy on Test Set for Bagging Regression Model was 69.2%
- RMSE on Train Set for Bagging Regression Model was 60206.37
- RMSE on Test Set for Bagging Regression Model was 61361.59

The Bagging Regression Model gave us an accuracy of 94.2% on the training set and 69.2% on the test set. The RMSE value for the test set is higher than the train set which indicates the model is overfitting. The bagging Regression model performed well on the training set but underperformed on the test set.

We will be hyper tuning 2 of the models that performed the best, which are Random Forest Model and Gradient Boost Model.

9) Random Forest Hyper Tune Model -

- Accuracy on Train Set for Random Forest Hyper Tune Model was 80.7%
- Accuracy on Test Set for Random Forest Hyper Tune Model was 71.1%
- RMSE on Train Set for Random Forest Hyper Tune Model was 109854.27
- RMSE on Test Set for Random Forest Hyper Tune Model was 110770.51

Hyper tuning the random forest model with the appropriate measures gave us an accuracy of 80.7% on the training set and 71.1% on the test set. The RMSE value for the test set is higher than the train set which indicates the model is overfitting.

10) Gradient Boost Hyper Tune Model -

- Accuracy on Train Set for Gradient Boost Hyper Tune Model was 84.5%
- Accuracy on Test Set for Gradient Boost Hyper Tune Model was 72.2%
- RMSE on Train Set for Gradient Boost Hyper Tune Model was 98316.62
- RMSE on Test Set for Gradient Boost Hyper Tune Model was 83986.66

Hyper tuning the gradient boost model with the appropriate measures gave us an accuracy of 84.5% on the training set and 72.2% on the test set. The RMSE value for the train set is higher than the test set which indicates the model is underfitting.

Hyper tuning the Gradient Boost Model gave us the best accuracy. The model is also underfitting.

5) Model Validation -

The models were compared based on their accuracies and RMSE values.

Out of all the models Gradient Boost Model performed the best and gave us the highest accuracy.

Hyper Tuning the Gradient boost model with the best parameters gave us an accuracy of 84.5% on the training set and 72.2% accuracy on the test set.

The gradient boost hyper tune model performed the best on both test and train dataset.

Important Features that affect the price variable -

These are the features that affect the price of the house according to the gradient boost hyper tune model.

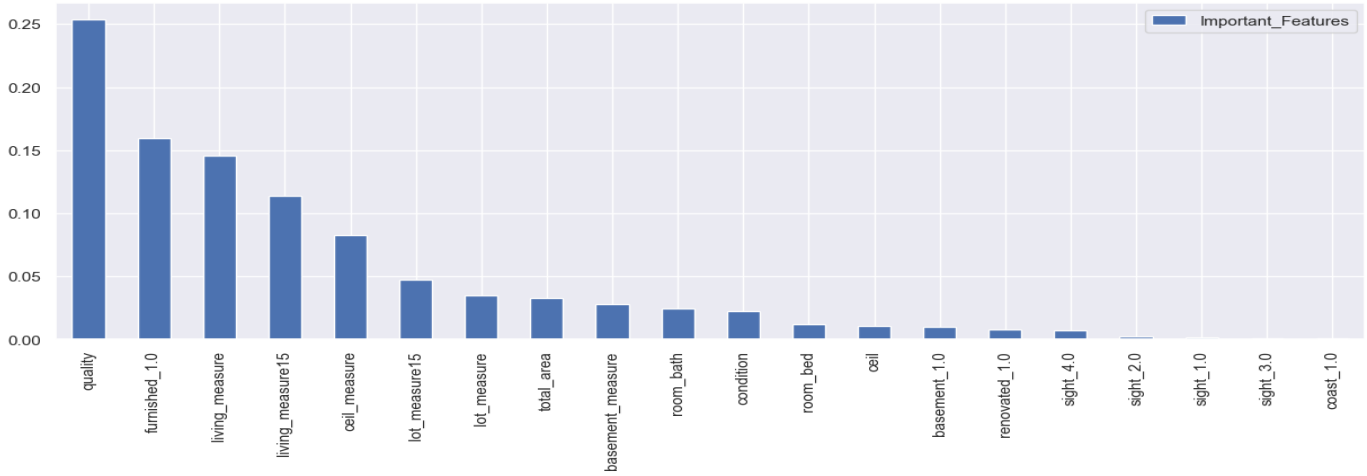


Fig 29 – Important features than affect price

We can see that almost all the features affect the price of the house, quality and furnished_1.0 are the top two features that affect the price of a house.

6) Final Interpretation / Recommendation -

Insights -

- Quality rating is the most important feature that is looked for in a house.
- Having a coast or not doesn't affect much of the price.
- Houses with a 6 – 9.5 quality rating are preferred.
- People prefer furnished houses with good square footage to live in.
- People also prefer houses with 1 to 2 floors.
- Overall, a house with good living space, furnished and with 1-2 floors is what people want to buy.

Recommendations -

- Some of the features that affect the price the most, like quality, furnished and living measure, should be looked for while purchasing or selling a house.

- Most important feature is quality, a house with higher quality rating is priced higher.

- Selling a furnished house with ample living space is easier compared to an unfurnished house with less or excess living space, House with 2000 Sq foot of living space is what people want, a house with less than 1000 Sq foot or more than 3000 Sq foot will be hard to sell.

- More than 50% of the houses are not furnished, furnishing these houses will help sell them as people prefer furnished houses.

- A house with a quality rating higher than 6 is what people prefer, therefore scoring at least a 6 is recommended.

* * * * * * * * * * * *