BUSINESS REPORT : CAPSTONE PROJECT

July 3

2022

BIBEK KUMAR GIRI
PGP-DSBA-JULY"21

TABLE OF CONTENTS

APPENDIX – I

List of Figures

The list of Tables

APPENDIX – II

Below is the table showing a sample of the given dataset

	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 Dataset (Inverted)*

APPENDIX – III

The table shows a descriptive analysis of the dataset after the missing value treatment.

	COUNT	MEAN	STD	MIN	25%	50%	75%	MAX
CID	21613	4580301521	2876565571	1000102	2123049194	3904930410	7308900445	9900000190
PRICE	21613	540182.16	367362.23	75000	321950	450000	645000	7700000
ROOM_BED	21613	3.37	0.93	0	3	3	4	33
ROOM_BATH	21613	2.12	0.77	0	1.75	2.25	2.5	8
LIVING_MEASURE	21613	2079.73	918.15	290	1430	1910	2550	13540
LOT_MEASURE	21613	15090.03	41384.66	520	5043	7618	10660	1651359
CEIL	21613	1.49	0.54	1	1	1.5	2	3.5
COAST	21613	0.01	0.09	0	0	0	0	1
SIGHT	21613	0.23	0.77	0	0	0	0	4
CONDITION	21613	3.41	0.65	1	3	3	4	5
QUALITY	21613	7.66	1.18	1	7	7	8	13
CEIL_MEASURE	21613	1788.36	828.08	290	1190	1560	2210	9410
BASEMENT	21613	291.51	442.58	0	0	0	560	4820
YR_BUILT	21613	1971.01	29.36	1900	1951	1975	1997	2015
YR_RENOVATED	21613	84.4	401.68	0	0	0	0	2015
ZIPCODE	21613	98077.94	53.51	98001	98033	98065	98118	98199
LAT	21613	47.56	0.14	47.16	47.47	47.57	47.68	47.78
LONG	21613	-122.21	0.14	-122.52	-122.33	-122.23	-122.12	-121.31
LIVING_MEASURE15	21613	1985.94	683	399	1490	1840	2360	6210
LOT_MEASURE15	21613	12759.64	27269.32	651	5100	7620	10080	871200
FURNISHED	21613	0.2	0.4	0	0	0	0	1
TOTAL_AREA	21613	1184681.76	20781781.52	1423	7040	9589	13058	371090065

Table 2: *Descriptive summary of the Dataset after Missing value treatment*

- *The summary of the dataset after and before the treatment is pretty much same.*

APPENDIX – IV

The table shows the grouping of data on basis of price distribution over other factors.

clust	**0**	**1**	**2**
price	905946.17	657534.48	420702.62
room_bed	3.98	3.39	3.17
room_bath	2.87	2.47	1.87
living_measure	3210.94	2842.47	1701.22
lot_measure	15674.79	258512.83	9432.04
ceil	1.88	1.57	1.37
coast	0.02	0.01	0
sight	0.57	0.52	0.12
condition	3.25	3.28	3.46
quality	9.17	8.24	7.16
ceil_measure	2838.92	2558.01	1435.46
basement	372.82	284.46	265.69
yr_built	1988.51	1983.36	1965.15
yr_renovated	111.9	88.17	75.53
zipcode	98062.06	98043.59	98083.79
living_measure15	2796.8	2393.17	1717.76
lot_measure15	14174.05	168155.36	8815.12
furnished	0.78	0.42	0.01
total_area	18901.91	262069.48	11135
freq	5145	362	16106

Table 3: *Cluster of Data*

- *To understand the data better we divided the given dataset into 03 different clusters by K-means method.*

APPENDIX – V

The given figure is for the Multiple Linear Regression without Outliers method.

The given figure is for the Multiple Linear Regression without Outliers method.

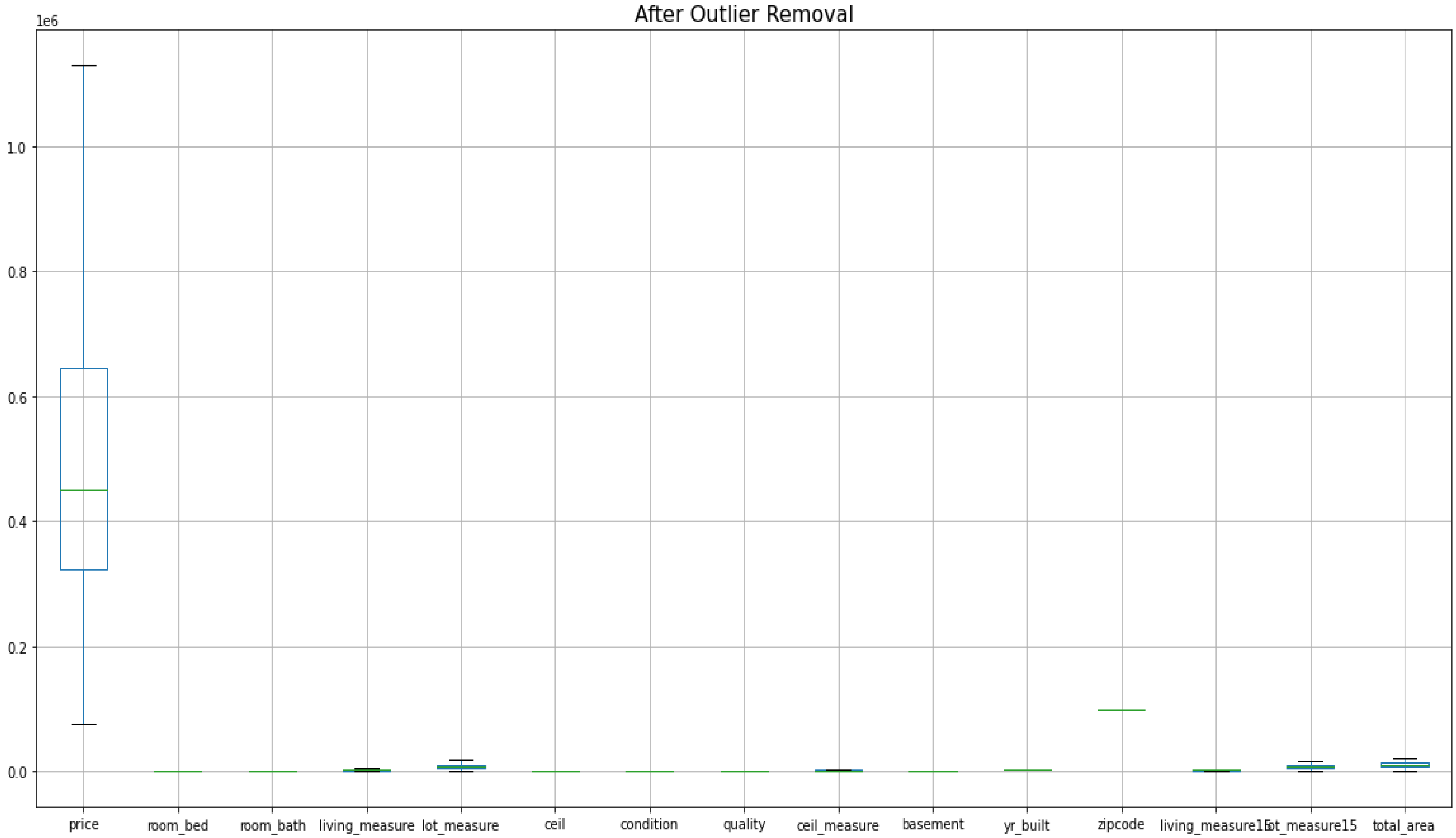


Figure 1: *After outliers treatment*

- *before the treatment there were almost all variable were having outliers but as we treat the data now there are no outliers present*

Problem: House Price Prediction

1. *Introduction of 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 price you usually try to find similar properties in your neighborhood and based on gathered data you will try to assess your house price.

To find the right price for a house with all its basic amenities. This study is done so that people who are looking to buy a new house or the people who are selling their house they shouldn't value their house too low or too high bas ed on social aspects.

1.1 Scope

- *To predict the right price of the house.*
- *By using different algorithms find the best algorithms to achieve aur objective.*

2. *Data Report*

- *A sample data of the given dataset is shown in Appendix-I*

Sr. No	Variable Name	Description
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
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

Table 4: *List of variables*

	COUNT	MEAN	STD	MIN	25%	50%	75%	MAX
CID	21613	4580301521	2876565571	1000102	2123049194	3904930410	7308900445	9900000190
PRICE	21613	540182.16	367362.23	75000	321950	450000	645000	7700000
ROOM_BED	21505	3.37	0.93	0	3	3	4	33
ROOM_BATH	21505	2.12	0.77	0	1.75	2.25	2.5	8
LIVING_MEASURE	21596	2079.86	918.5	290	1429.25	1910	2550	13540
LOT_MEASURE	21571	15104.58	41423.62	520	5040	7618	10684.5	1651359
SIGHT	21556	0.23	0.77	0	0	0	0	4
QUALITY	21612	7.66	1.18	1	7	7	8	13
CEIL_MEASURE	21612	1788.37	828.1	290	1190	1560	2210	9410

BASEMENT	21612	291.52	442.58	0	0	0	560	4820
YR_RENOVATED	21613	84.4	401.68	0	0	0	0	2015
ZIPCODE	21613	98077.94	53.51	98001	98033	98065	98118	98199
LAT	21613	47.56	0.14	47.16	47.47	47.57	47.68	47.78
LIVING_MEASURE15	21447	1987.07	685.52	399	1490	1840	2360	6210
LOT_MEASURE15	21584	12766.54	27286.99	651	5100	7620	10087	871200
FURNISHED	21584	0.2	0.4	0	0	0	0	1

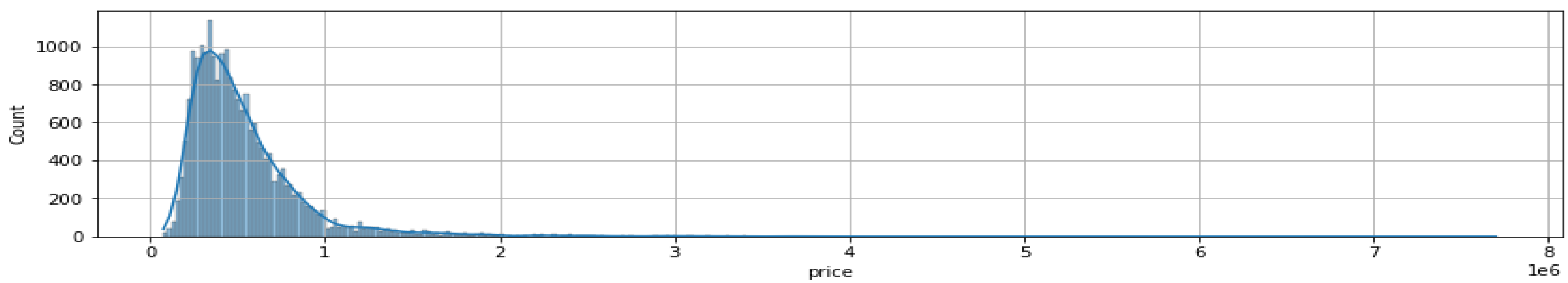
Table 5: *Descriptive summary of the Dataset*

- *The given data has rows of 21613 and columns of 23. Among those 23 columns, float is 12, integer is 4 and object are 7.*
- *When checked the duplicates with the 'cid' we found around 177 rows are duplicated.*
- *We didn't do any kind of treatment with those as a person can have more than one house to sell or maybe the it is an agent who is trying to sell houses.*
- *There are a total of 689 missing values present in dataset.*
- *Columns like 'coast' and 'furnished' treated with mode and other variables or columns are treated with median.*
- *Whereas variable 'total_area' missing values treated with the sum of 'living_measure' and 'lot_measure'.*
- *As mentioned earlier, the summary of the dataset after and before the treatment is pretty much same (**shown in appendix-II**)*
- *There have been few additions of variables as before their datatypes were object later, we need to convert it to integer or float.*

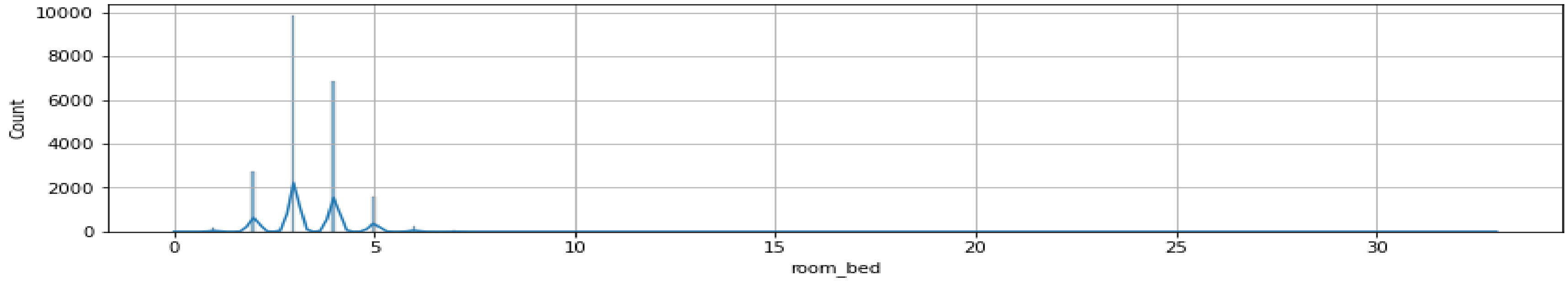
3. *Exploratory Data Analysis*

Histogram of different variables:

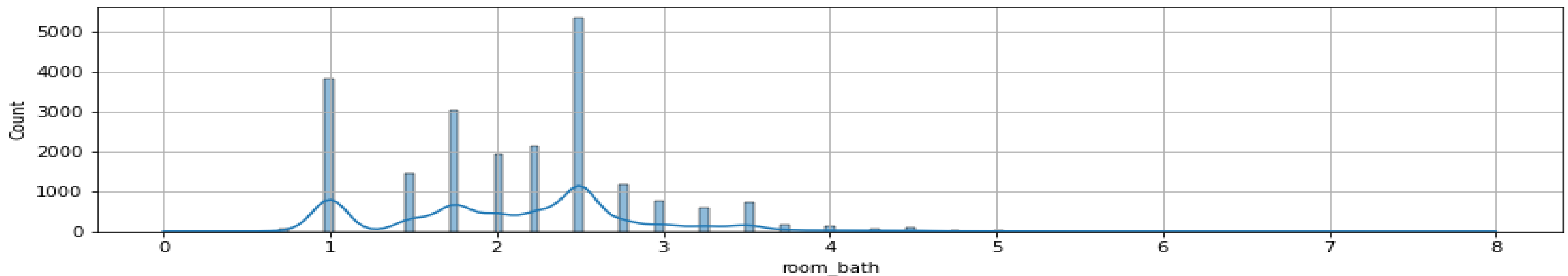
PRICE



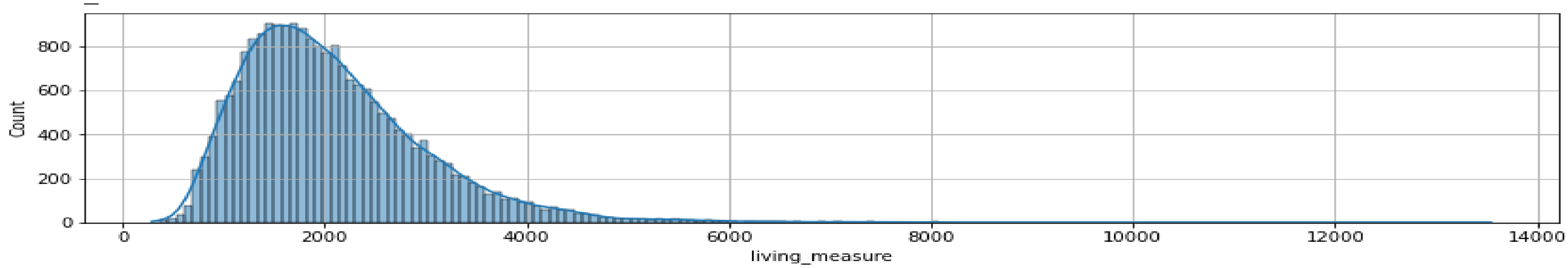
ROOM_BED



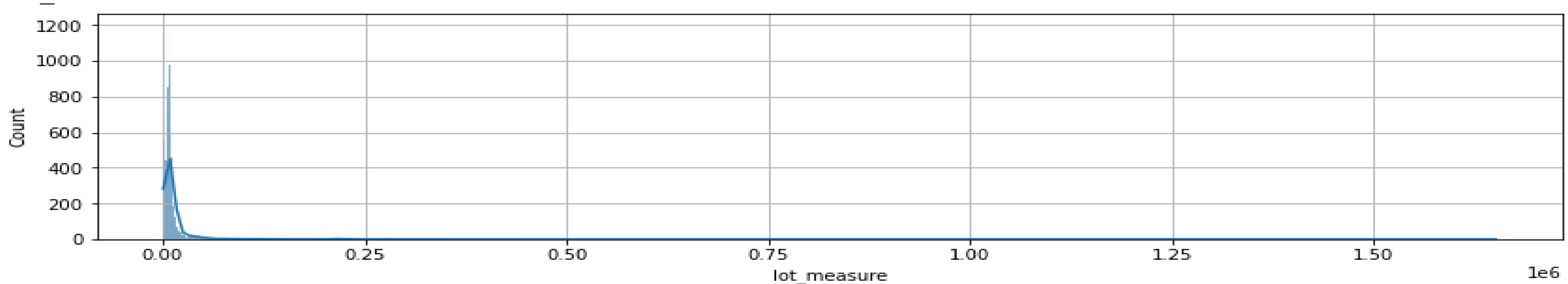
ROOM_BATH



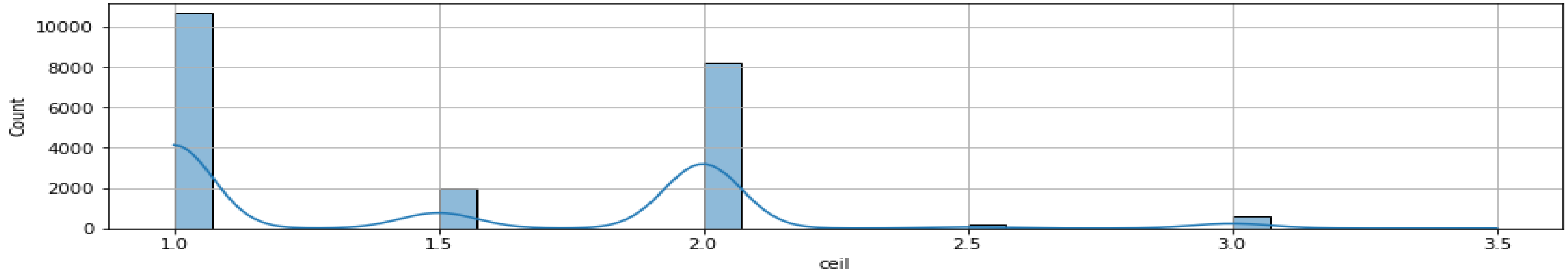
LIVING_MEASURE



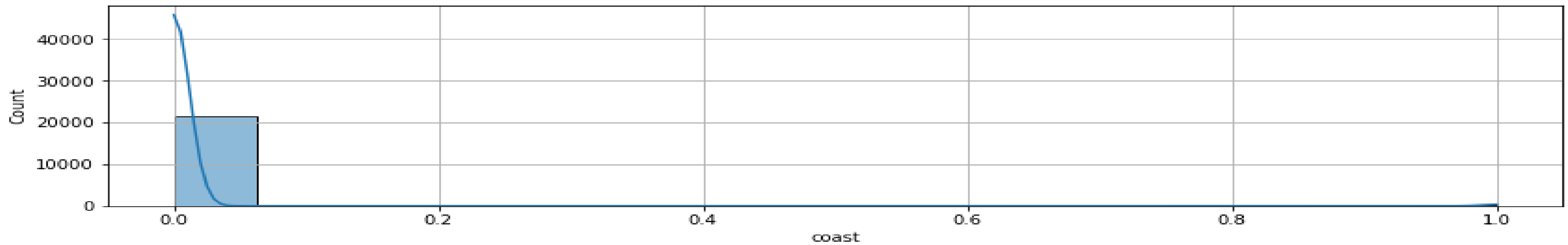
LOT_MEASURE



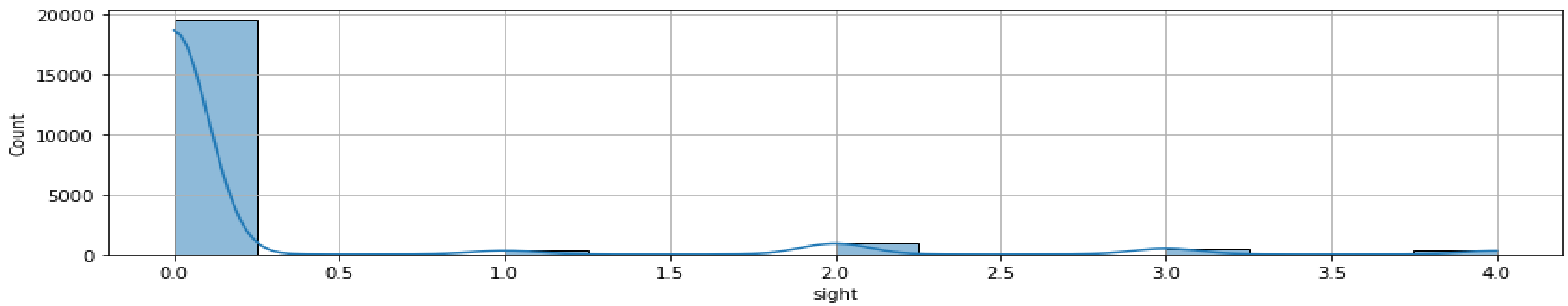
CEIL



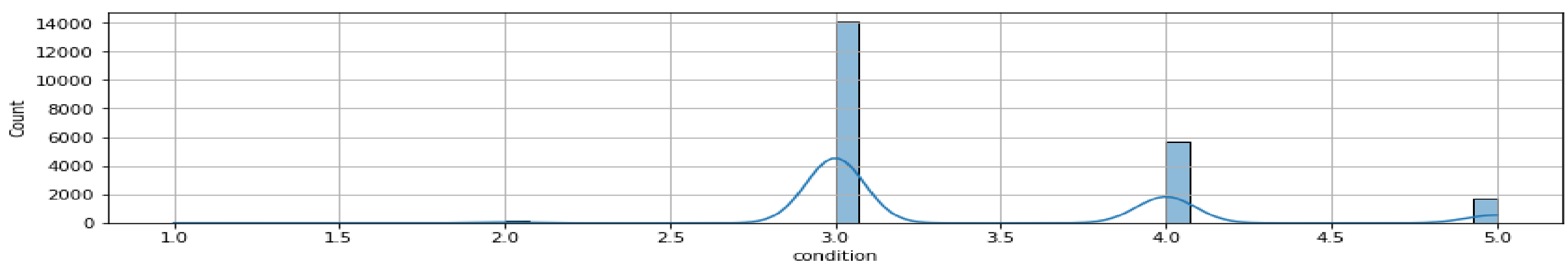
COAST



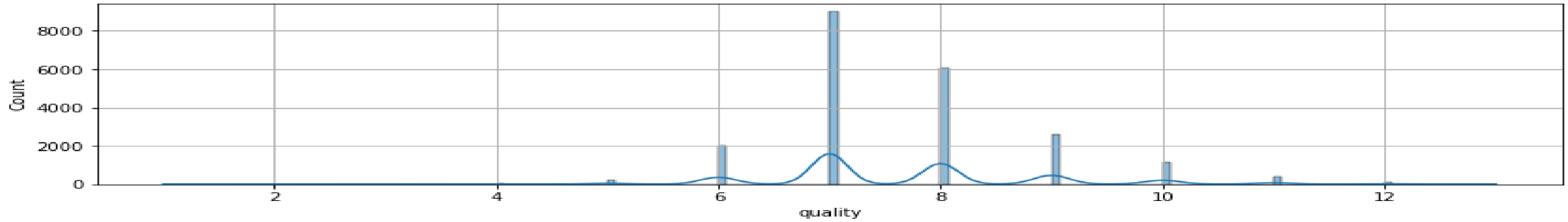
SIGHT



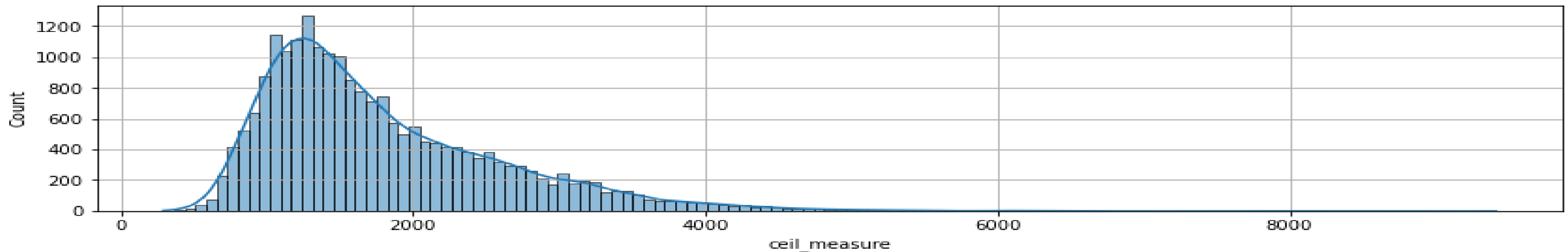
CONDITION



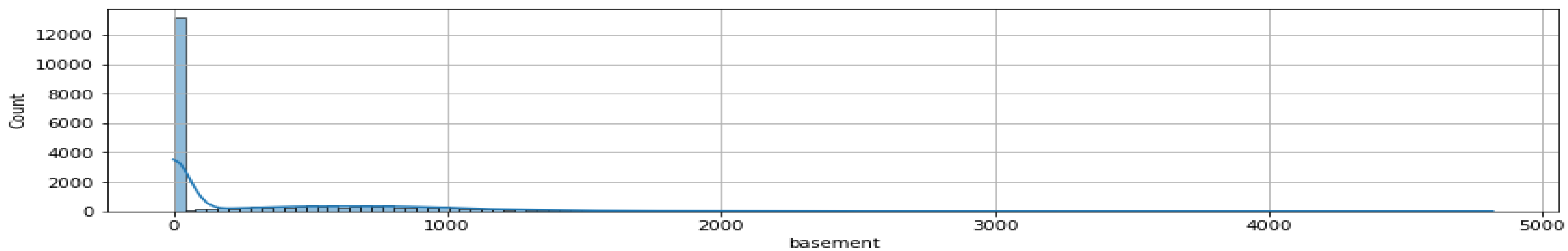
QUALITY



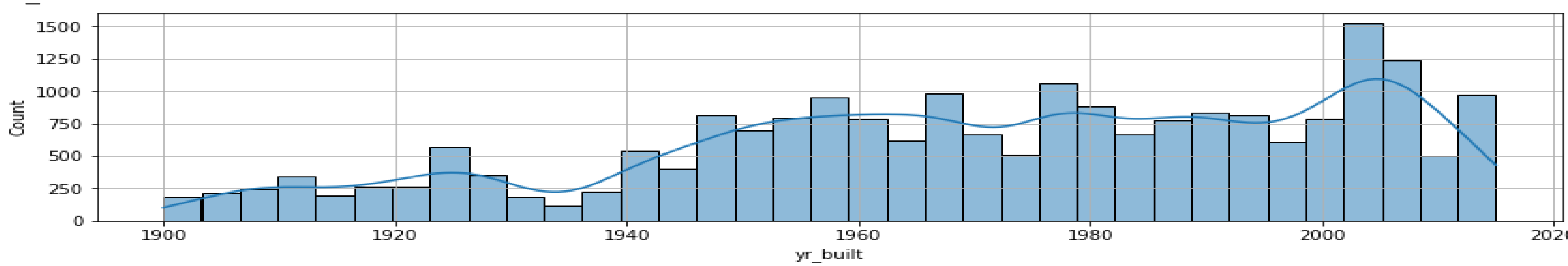
CEIL_MEASURE



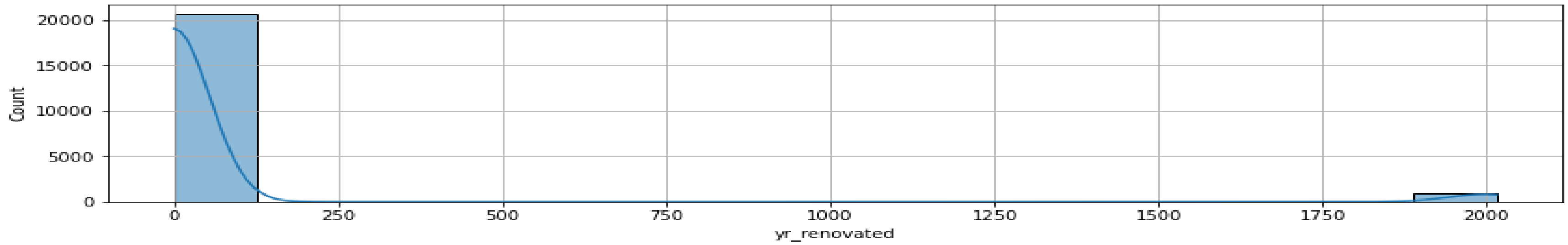
BASEMENT



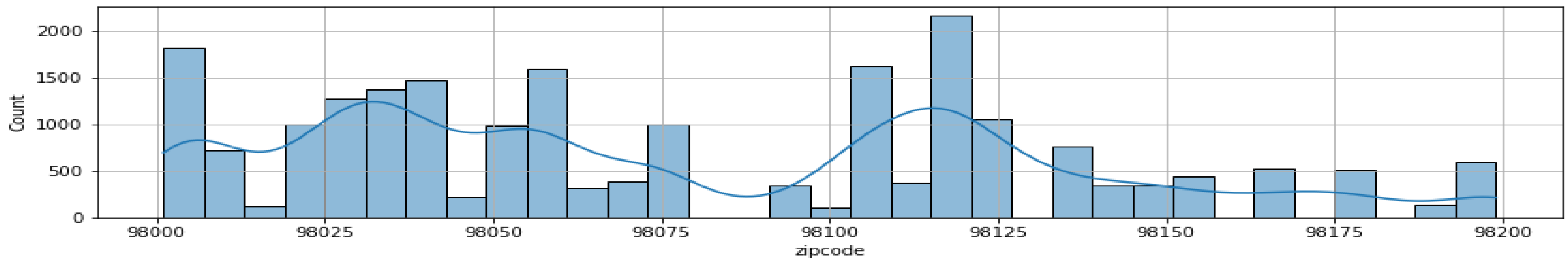
YR_BUILT



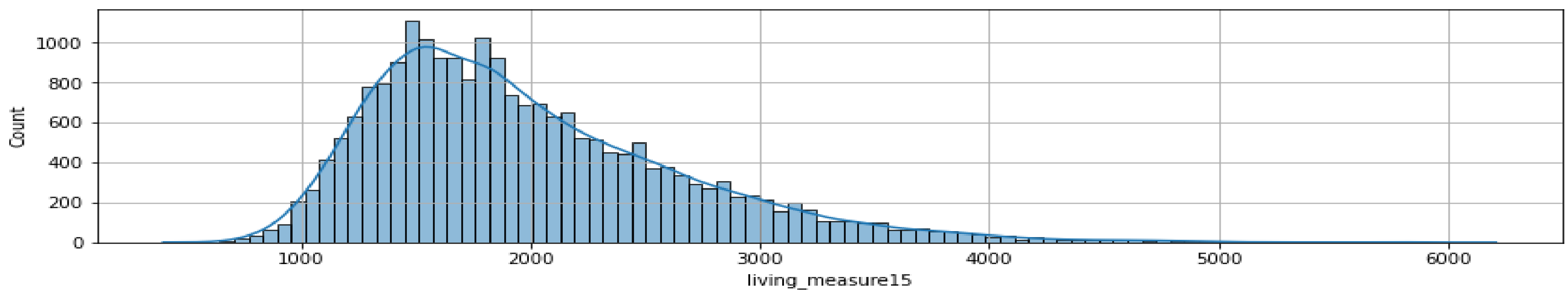
YR_RENOVATED



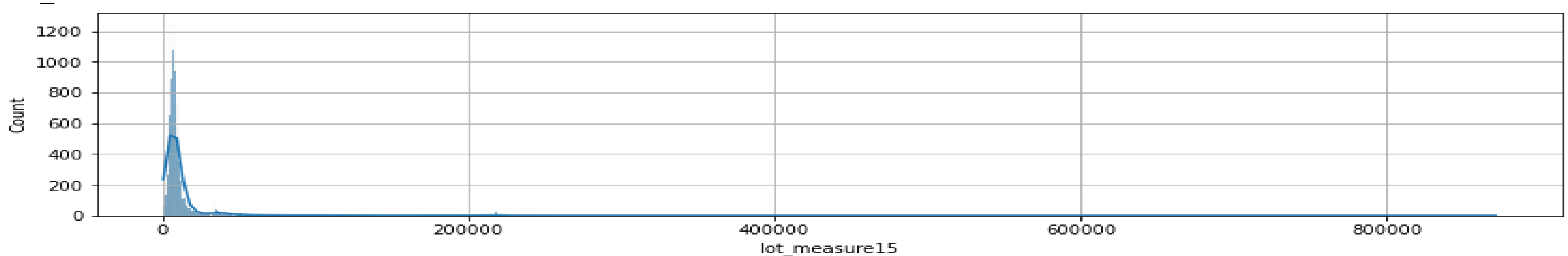
ZIPCODE



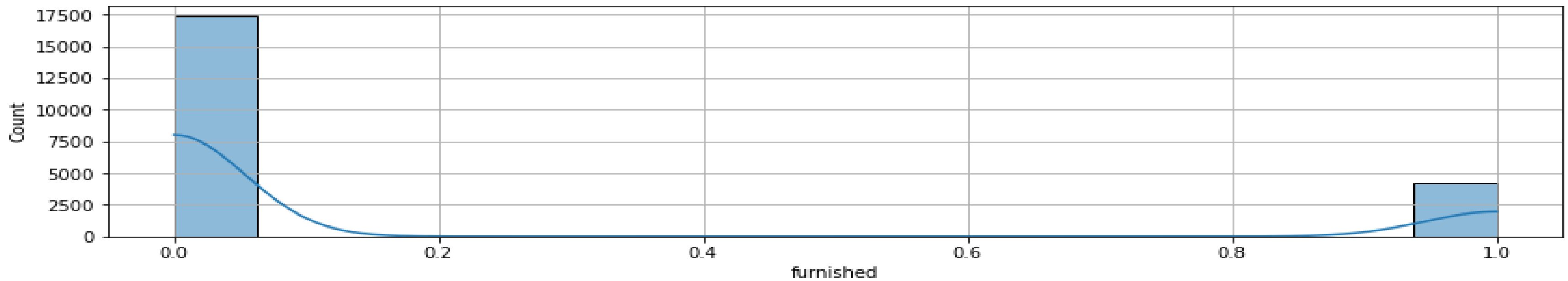
LIVING_MEASURE15



LOT_MEASURE15



FURNISHED



TOTAL_AREA

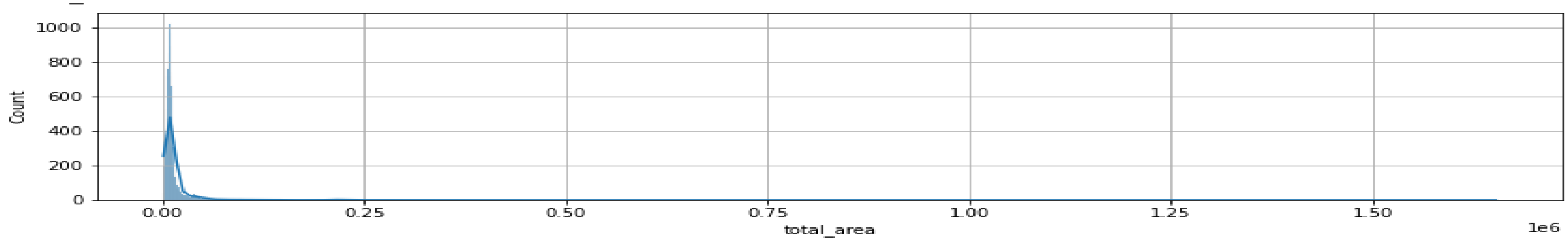
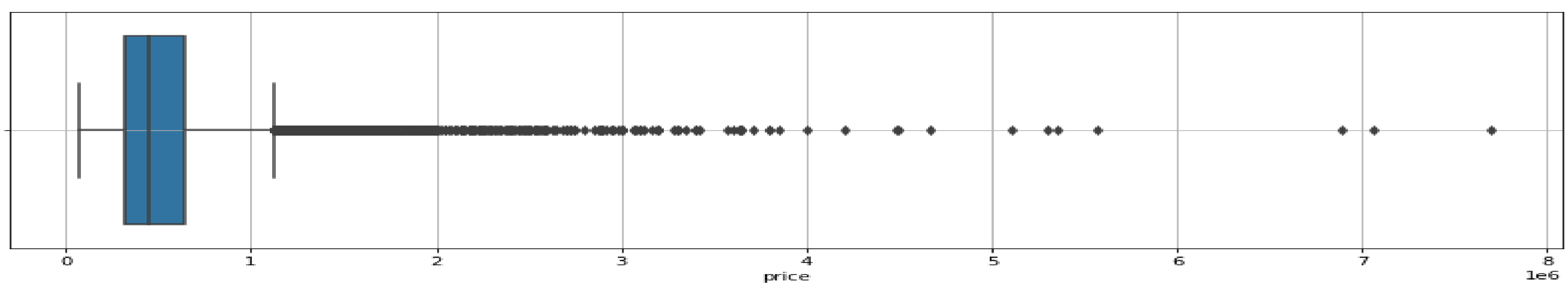


Figure 2: *Histogram representation of all variables*

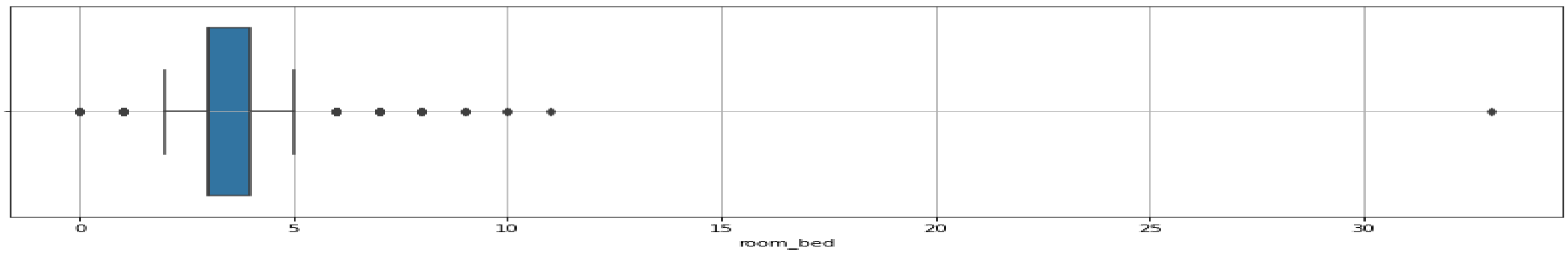
- *All variables are right-skewed except for the year built.*
- *The histogram of lot_measure and total_area is quite similar.*
- *We need to look into different factors as well before coming into any decision.*

Boxplot of different variables:

PRICE

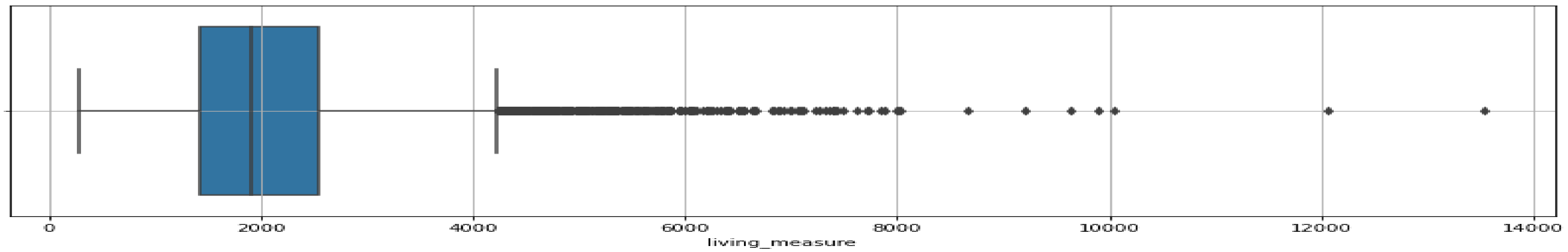


ROOM_BED

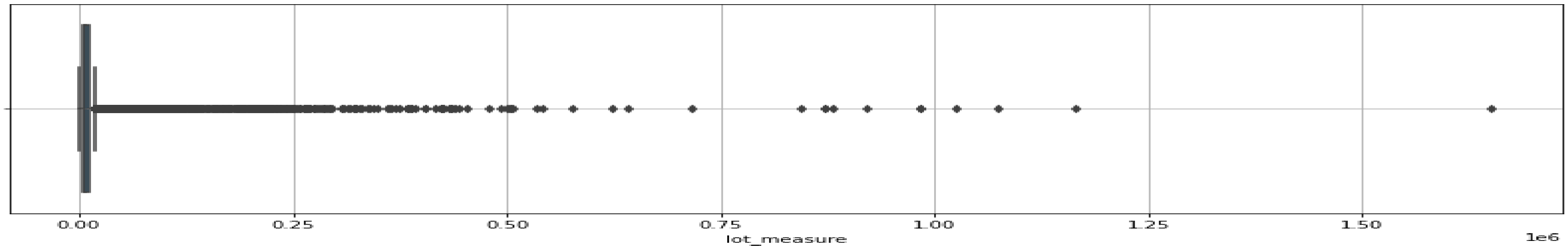


ROOM_BATH

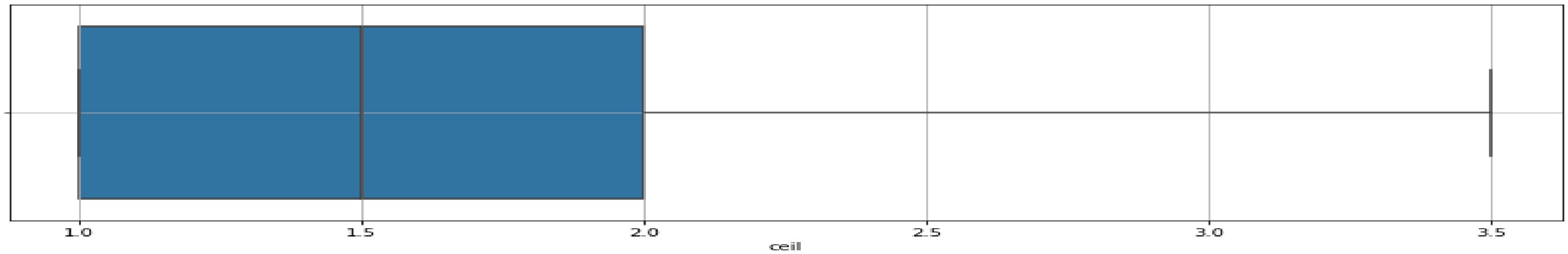
LIVING_MEASURE



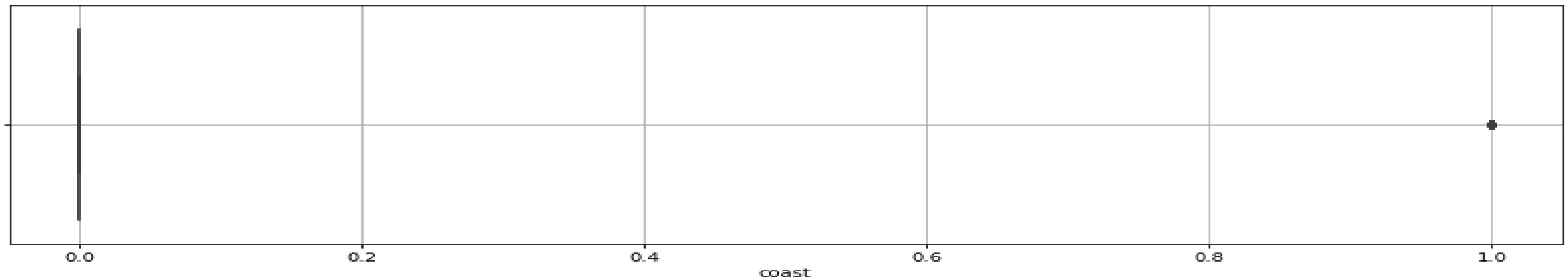
LOT_MEASURE



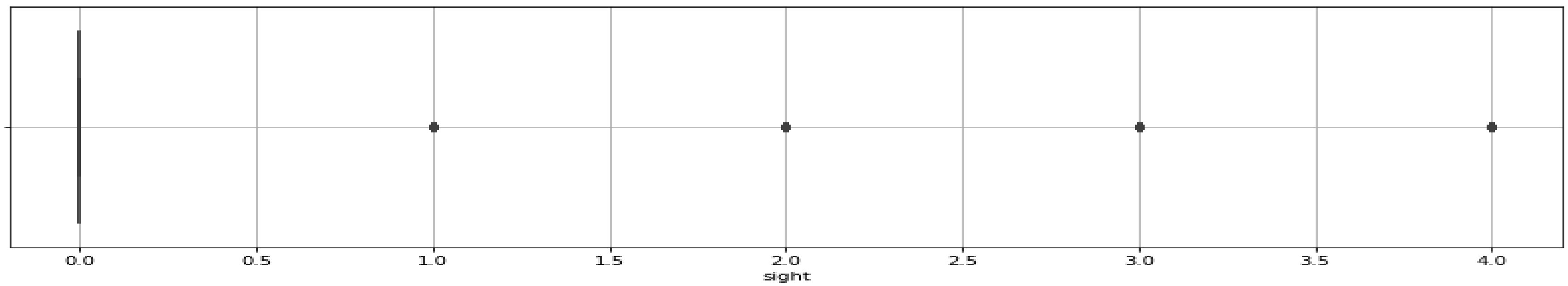
CEIL



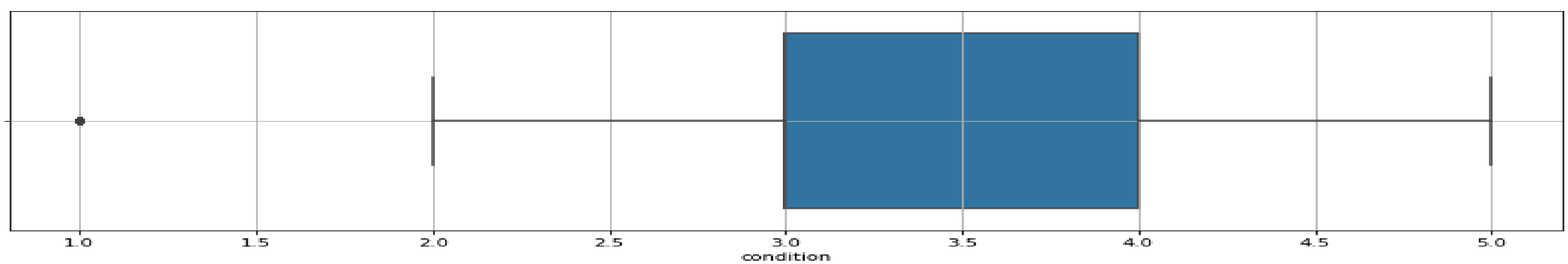
COAST



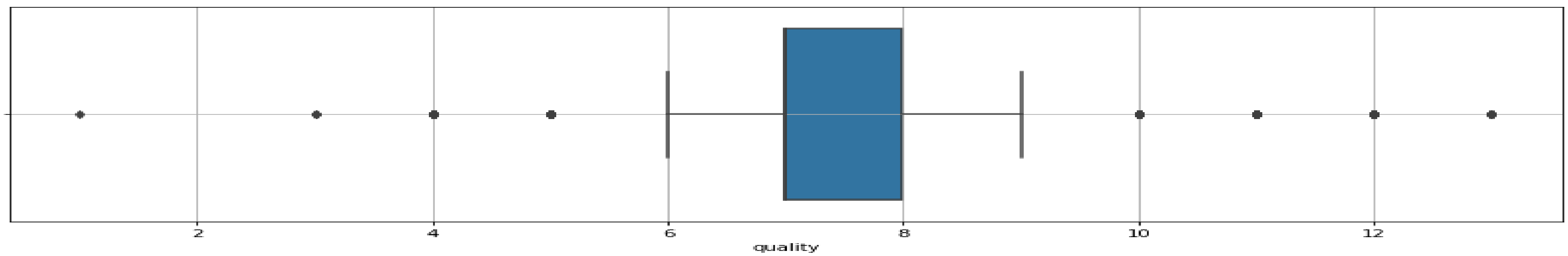
SIGHT



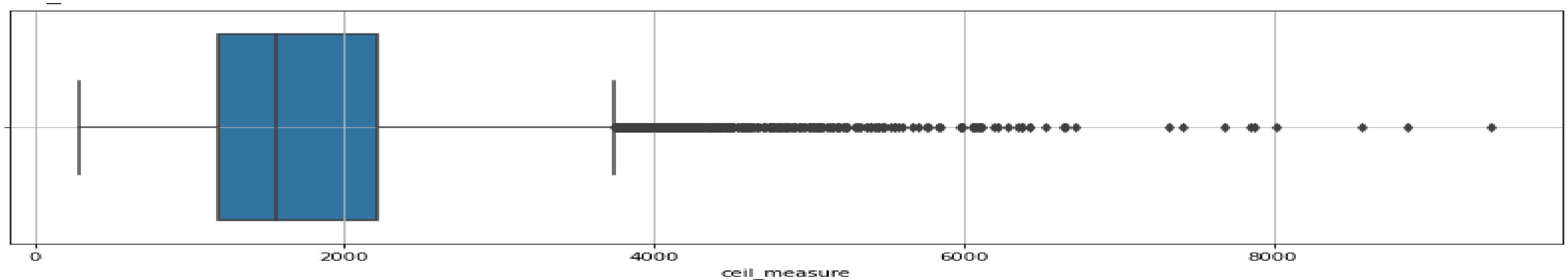
CONDITION



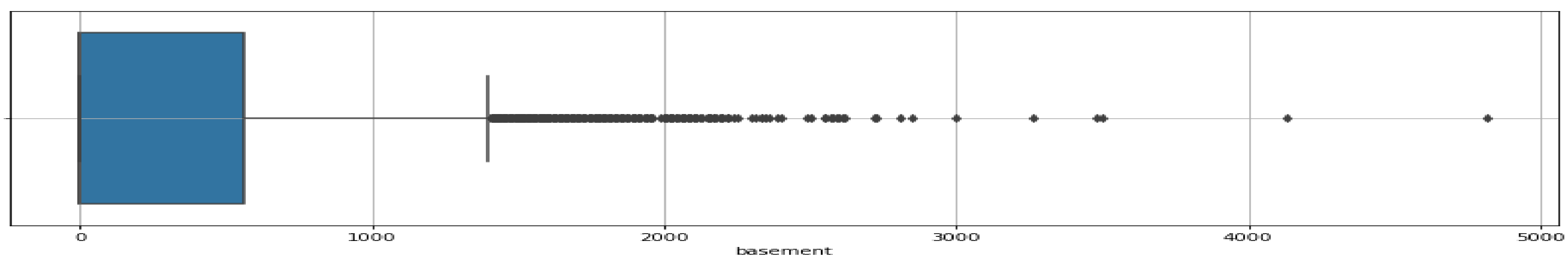
QUALITY



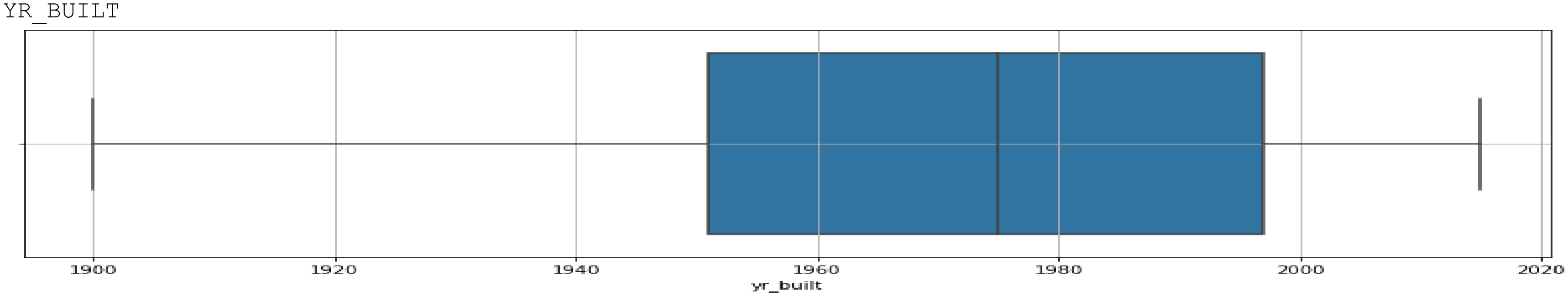
CEIL_MEASURE



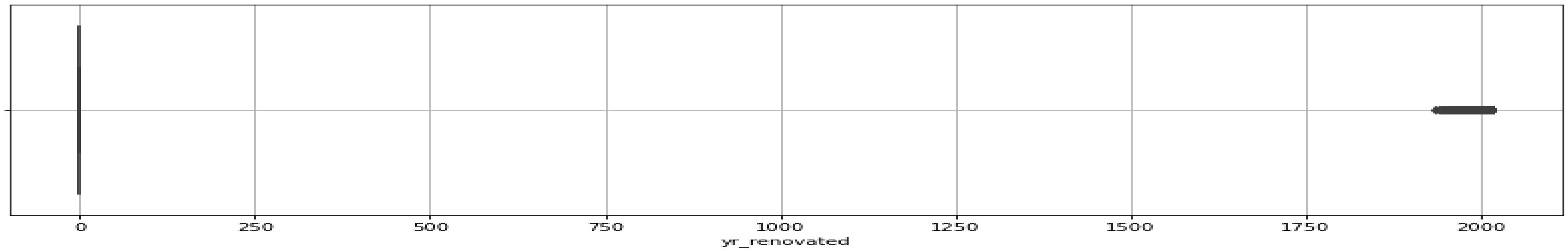
BASEMENT



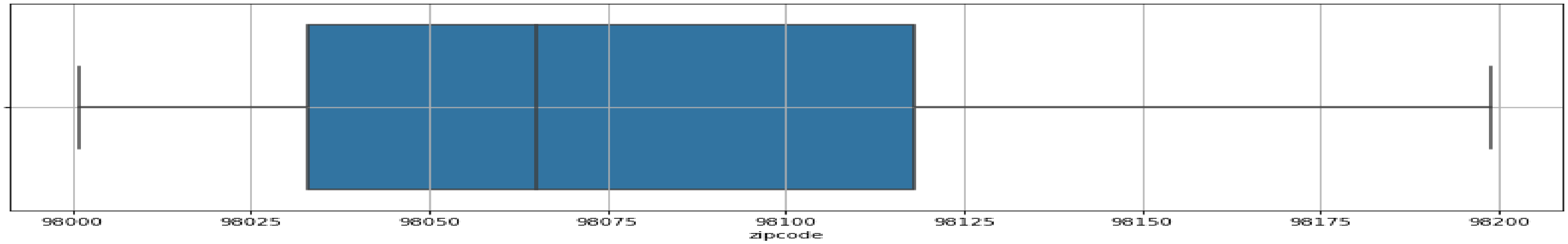
YR_BUILT



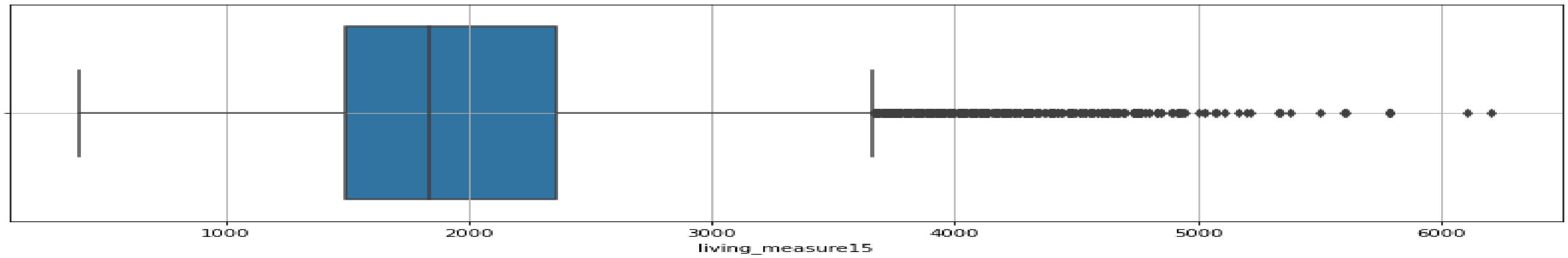
YR_RENOVATED



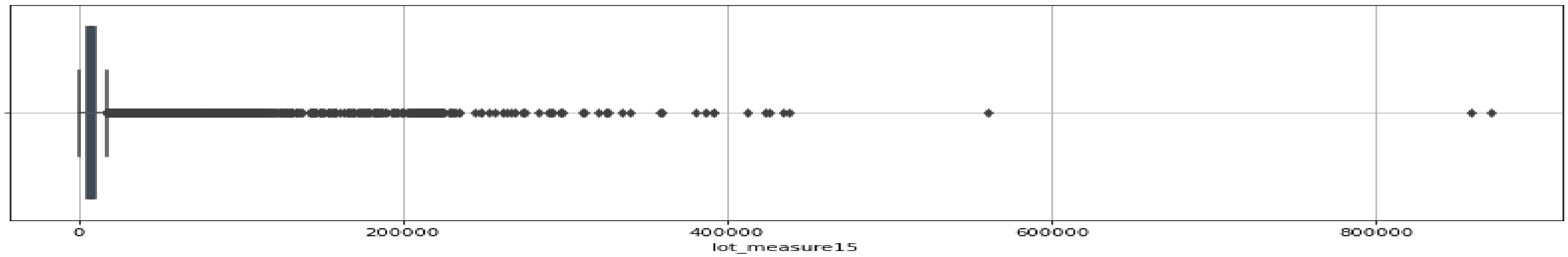
ZIPCODE



LIVING_MEASURE15



LOT_MEASURE15



FURNISHED

TOTAL_AREA

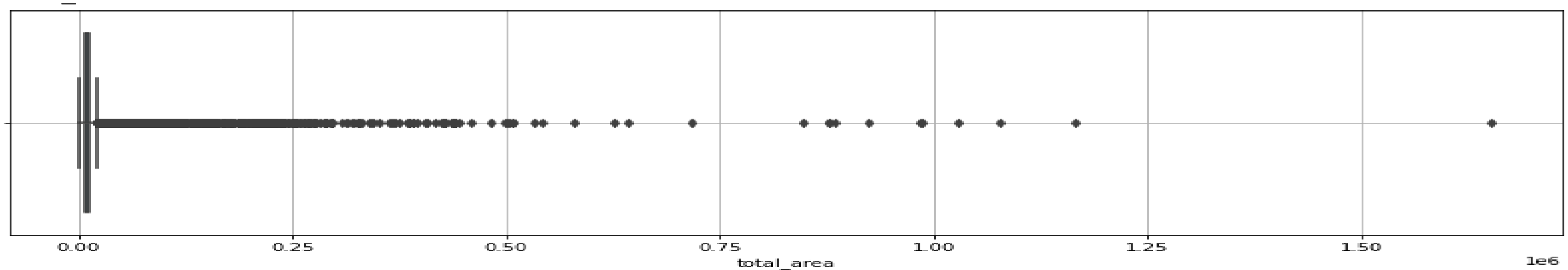


Figure 3: *Boxplot representation of all variables*

- *The variables like cid, ceil and yr_built do not any outliers.*
- *All other variables have outliers.*
- *But the variables like coast, condition, quality and sight are classifiers.*
- *Here also the total_area and lot_measure is similar.*
- *We decided to not to treat outliers as it is common that on basis of location and the condition offered the price of house may Vary.*

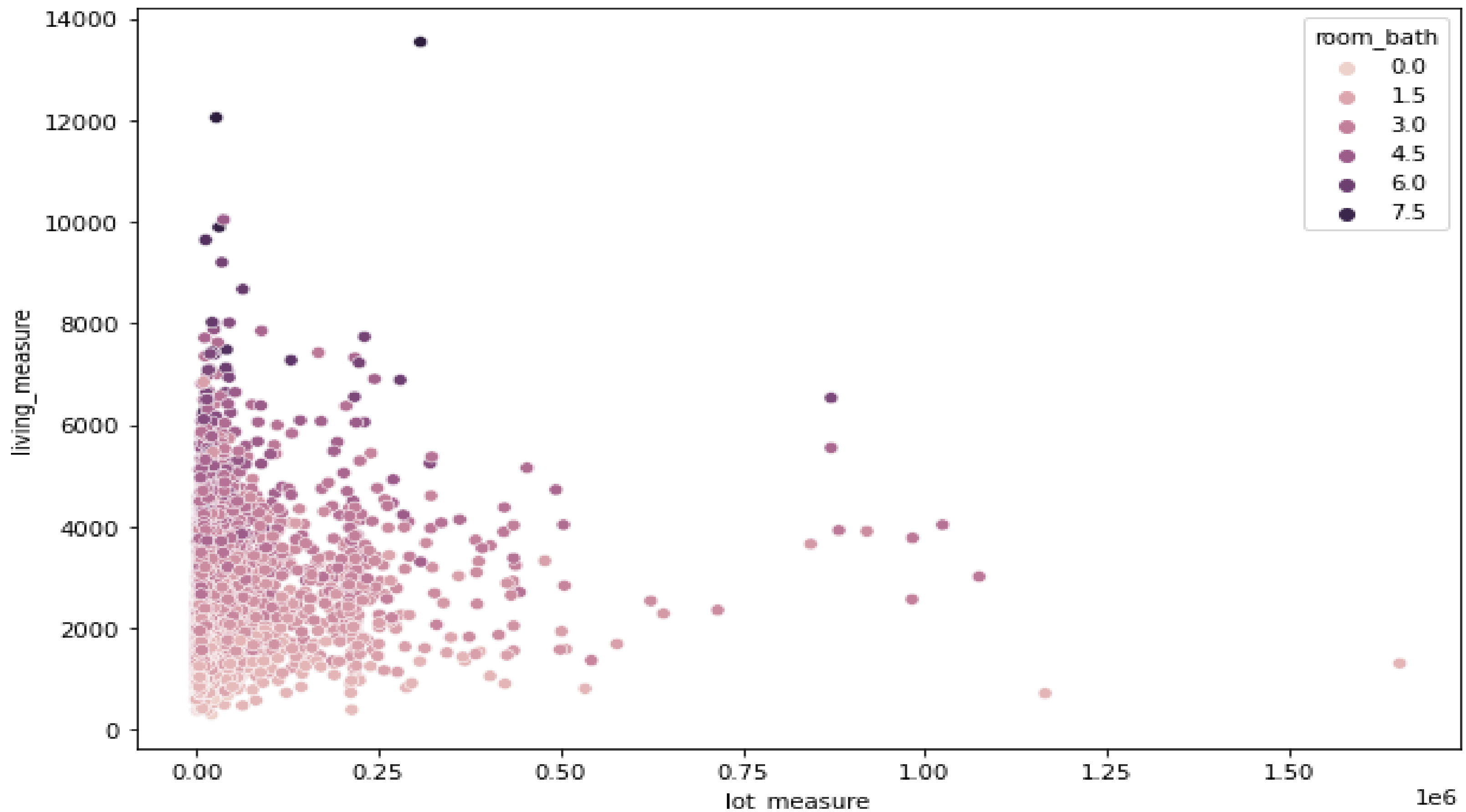


Figure 4: *Scatter plot lot_measure vs living_measure on basis of room_bath*

- *People prefer mostly room_bath of 3 to 4 with living measure of around 2000units more.*

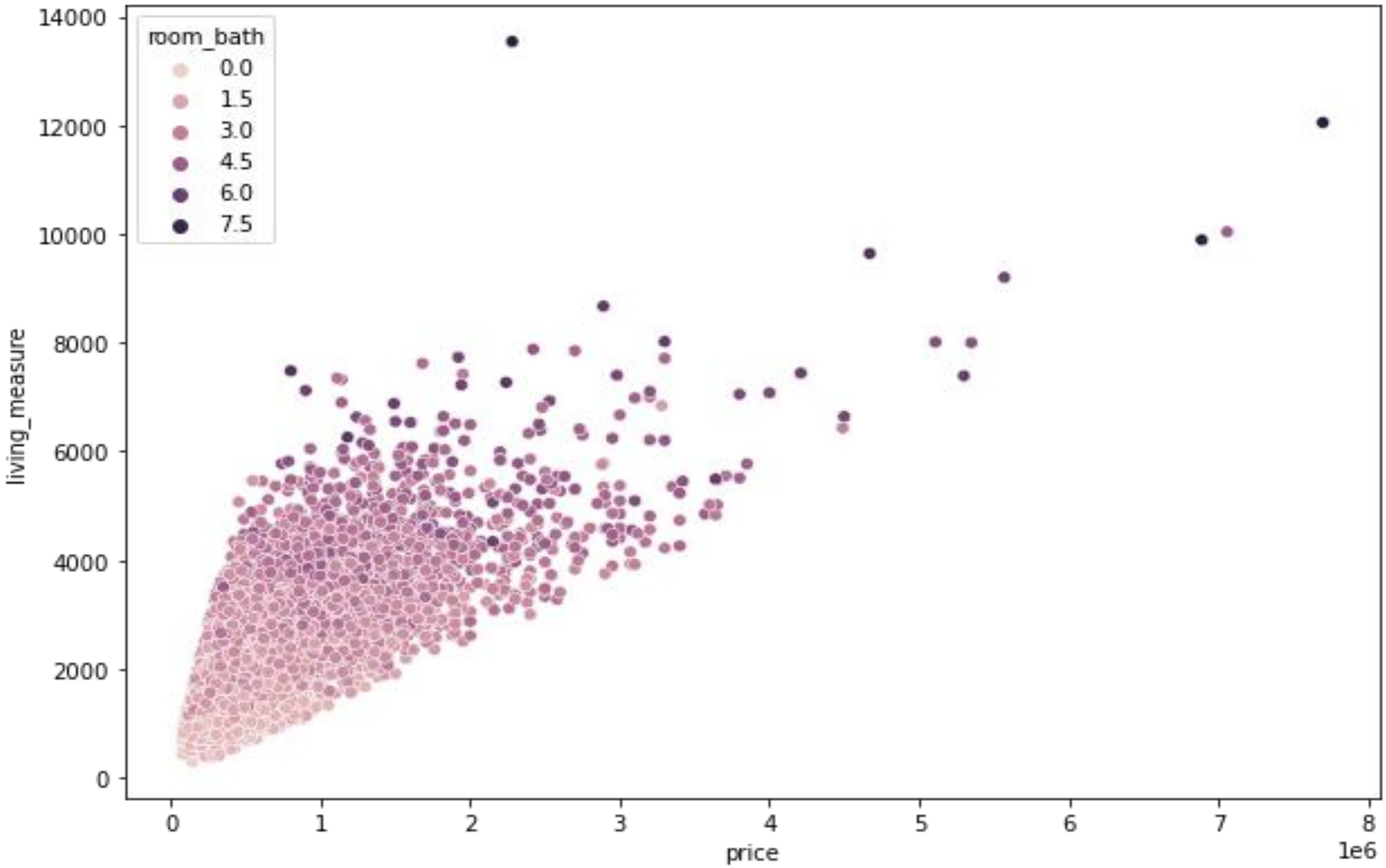


Figure 5: *Scatter plot price vs living_measure on basis of room bath*

- *But from above graph looks like people ready to any price for their choice.*
- *It shows that as the room per bath increases the price also increases with increase in living measure*

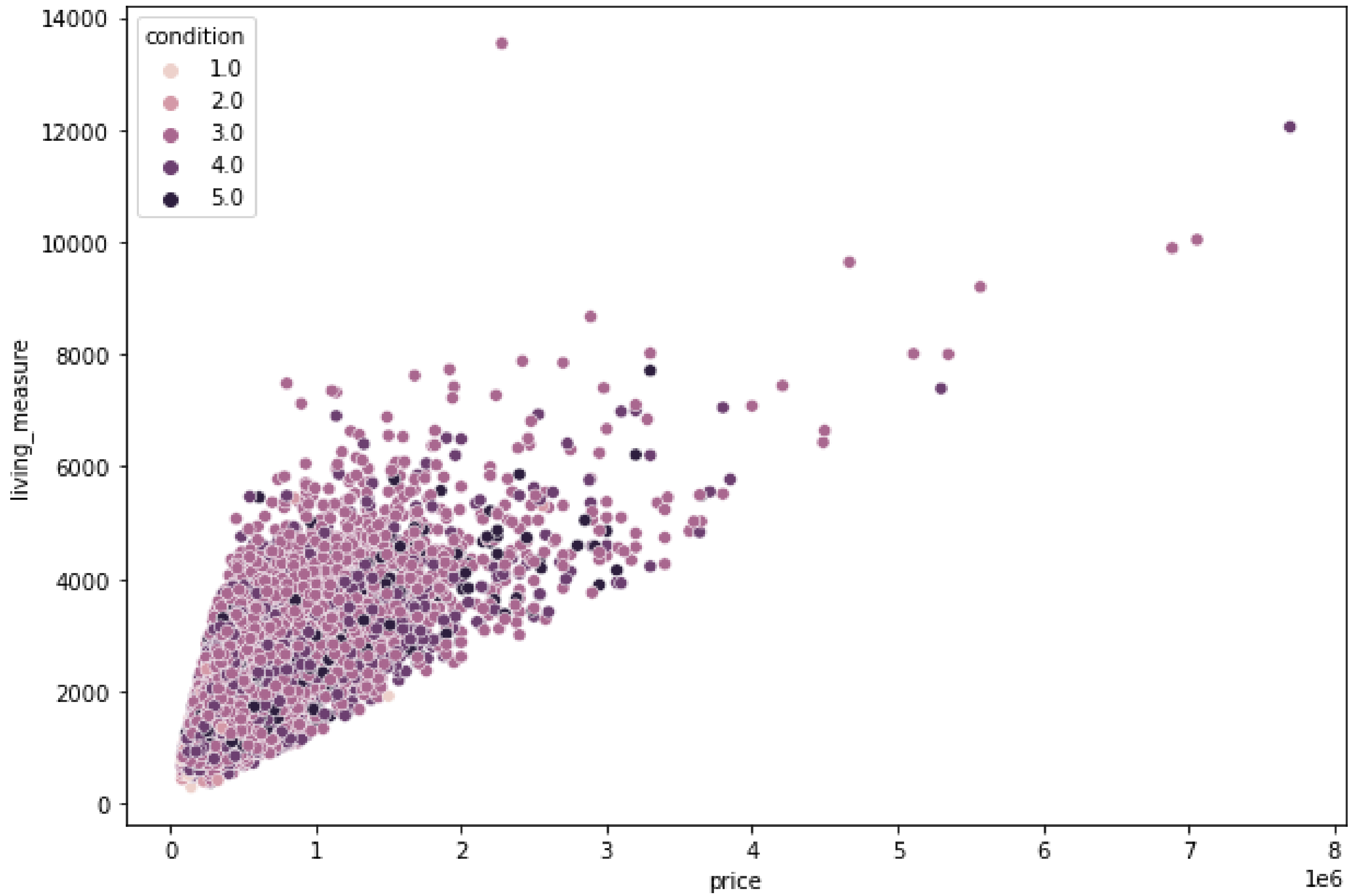


Figure 6: *Scatter plot price vs living_measure on basis of condition*

- *In some cases the with increase in living measure the condition is not increasing but here it constant for a price above 40,00,000 – 50,00,000units*
- *People are more preferring the house with living measure between 4000 – 2000 and with condition of around 3 to 4.*

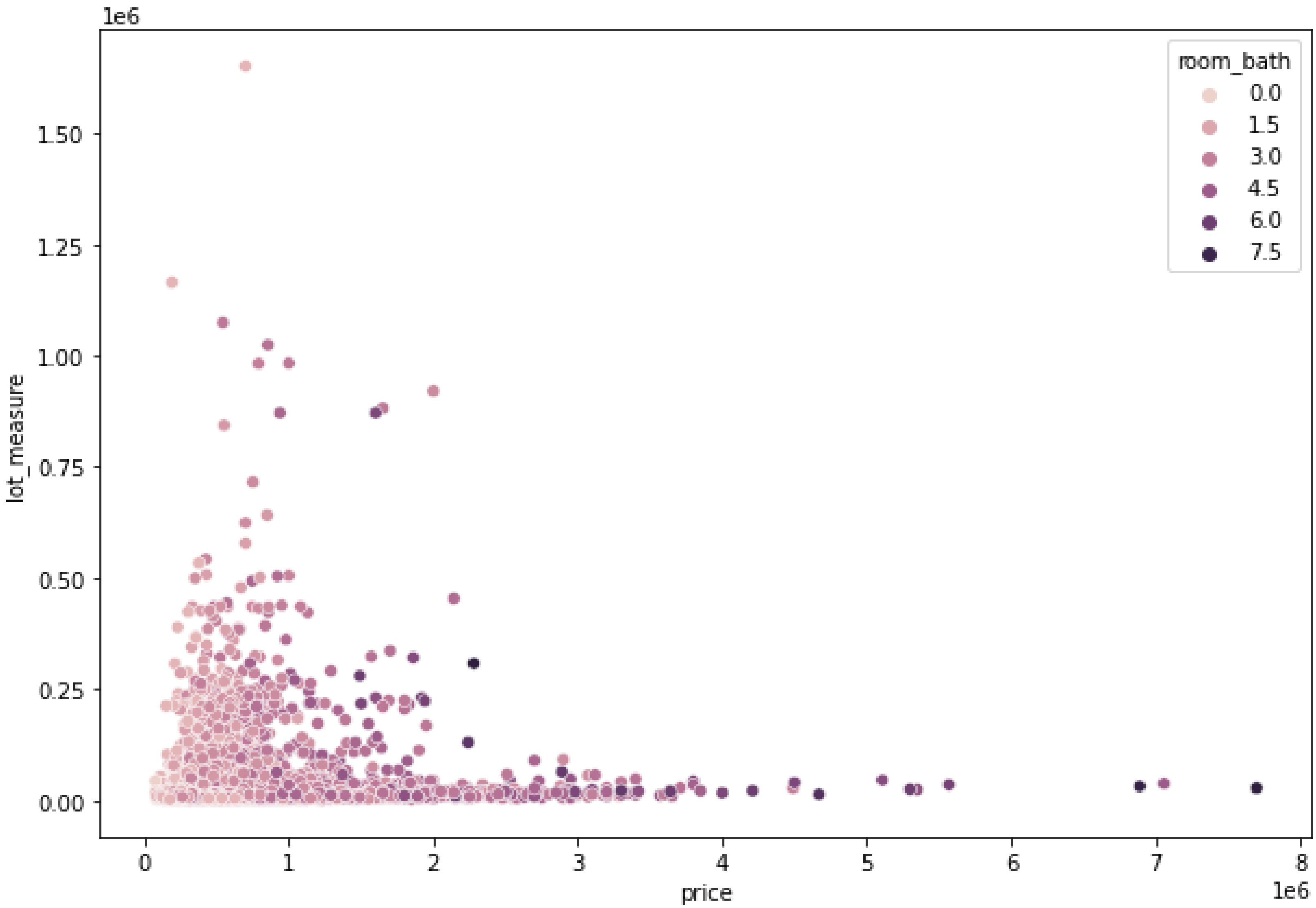


Figure 7: *Scatter plot price vs lot_measure on basis of room_bath*

- *Here for low price the room per bath is less for increase in lot measure.*

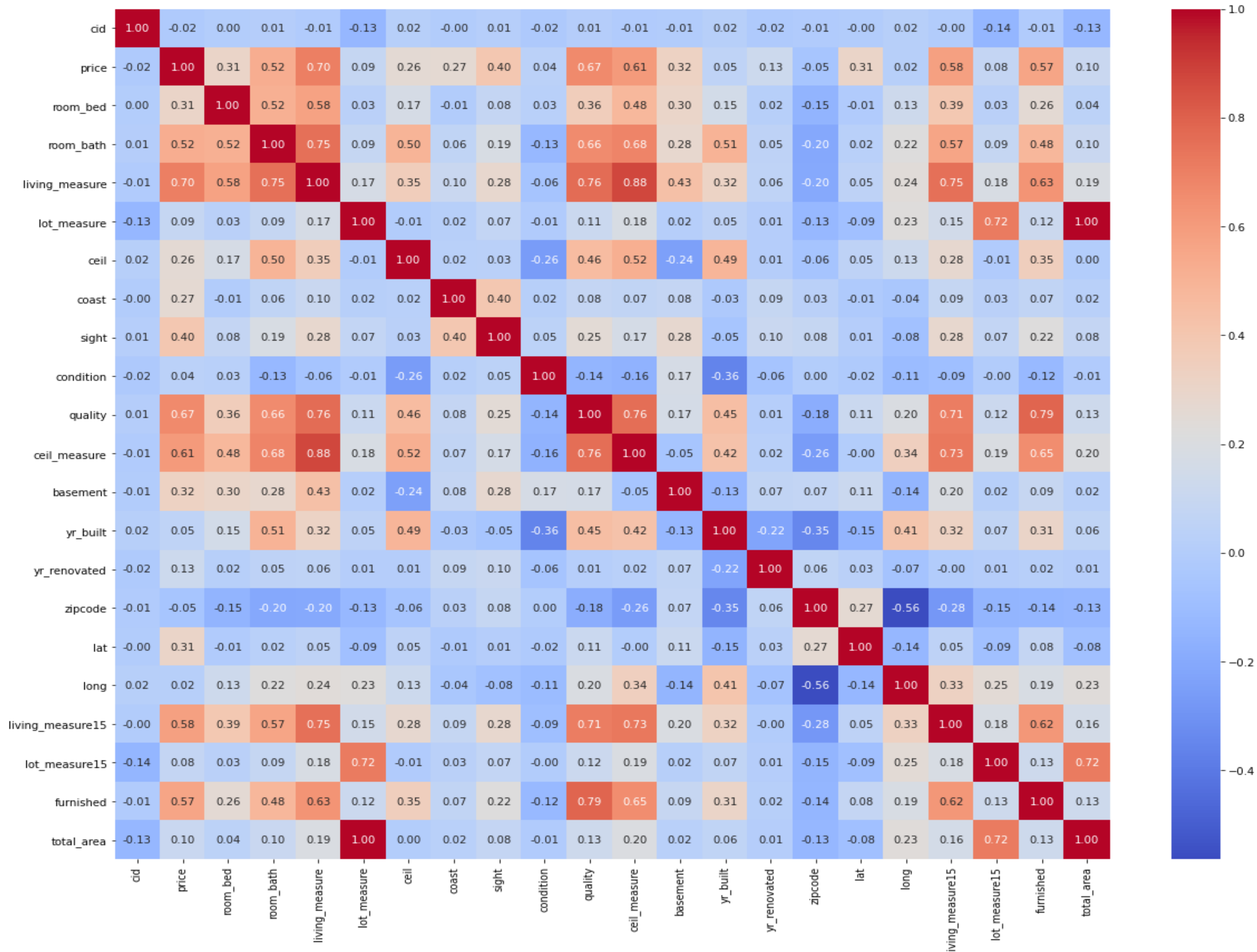


Figure 8: *Heat-map of dataset*

- *From this heat map we identified few relationships but the relation between total_area and lot_measure is quite strong.*

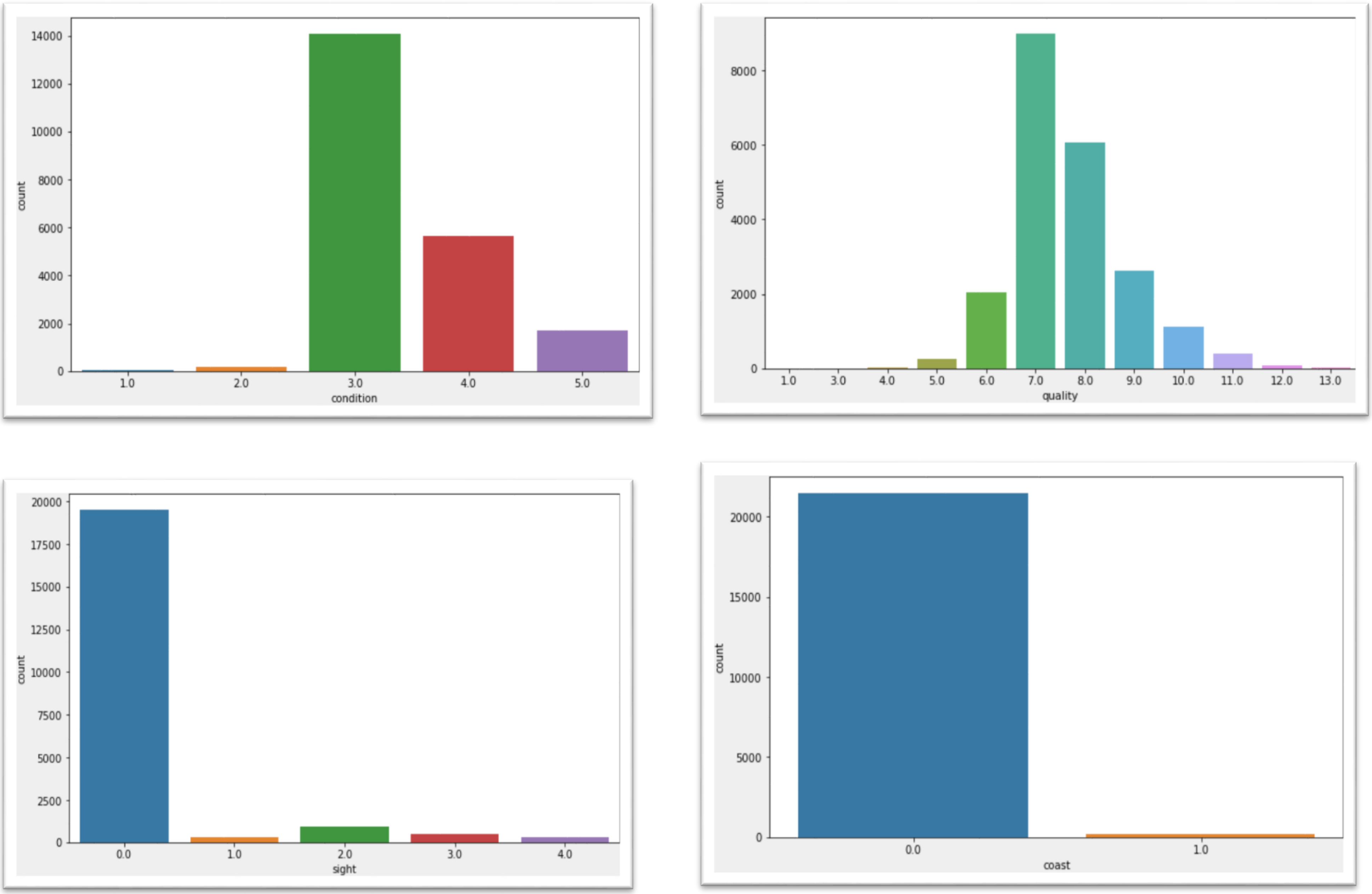
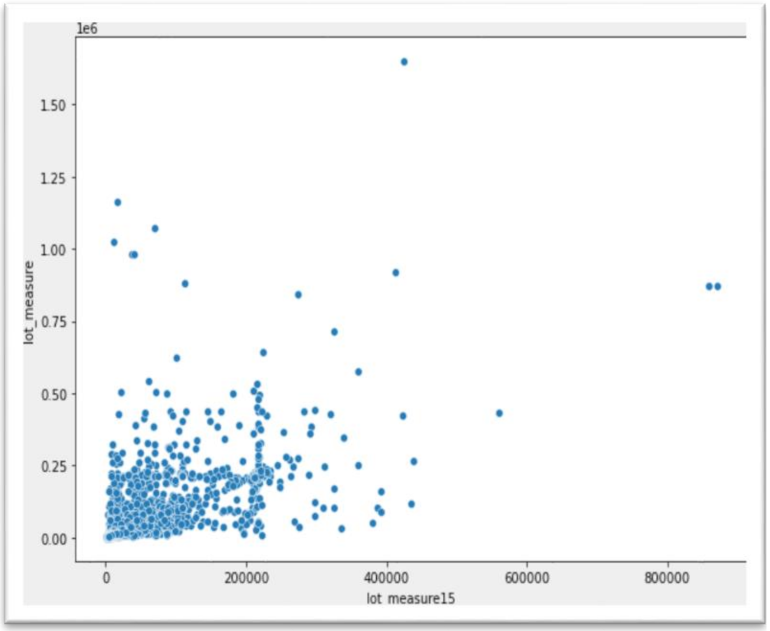
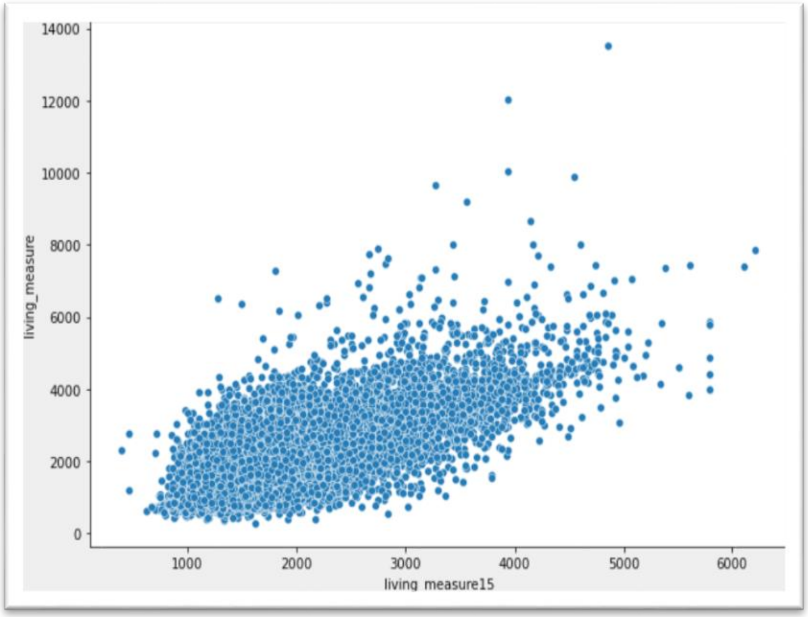


Figure 09: *Count-plot for condition, Quality, Coast and Sight(Clock-Wise)*

- *People don't like go for more comfort or high-class profile house rather they prefer a house minimal need.*
- *This will also keep price in check for them.*
- *Same case with the condition, people want to buy a house with basis requirements and then wants to make changes Accordingly.*
- *For this data it looks like people buy house without even visiting the actual house may be through friend or agent.*
- *Most of the house do not have sea facing house.*

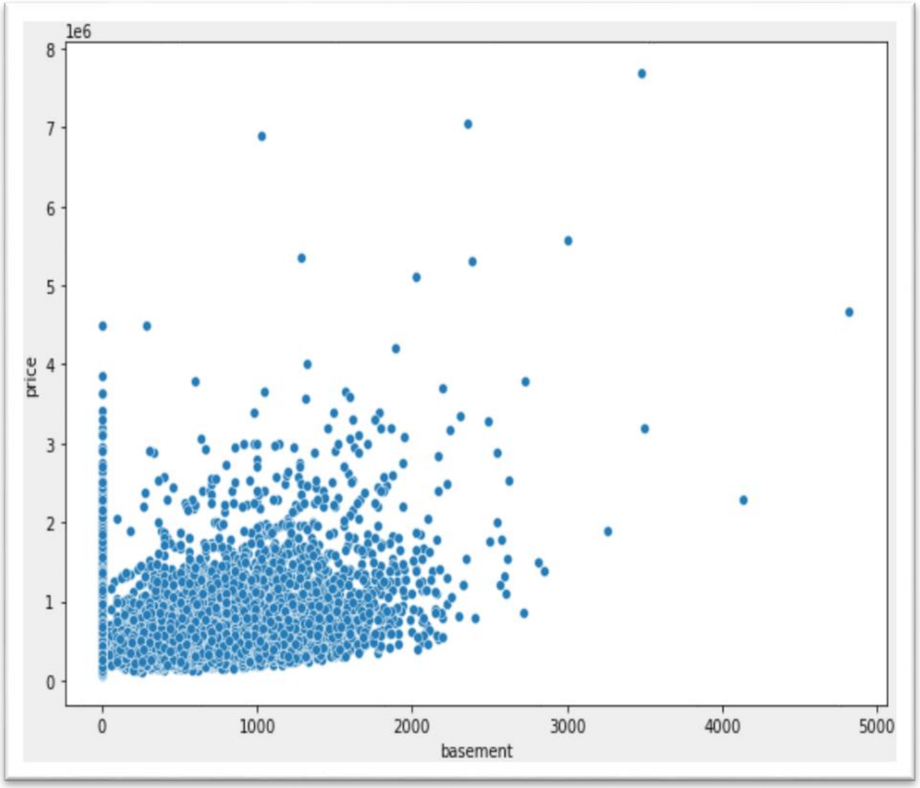


(a)lot_measure15 vs lot_measure

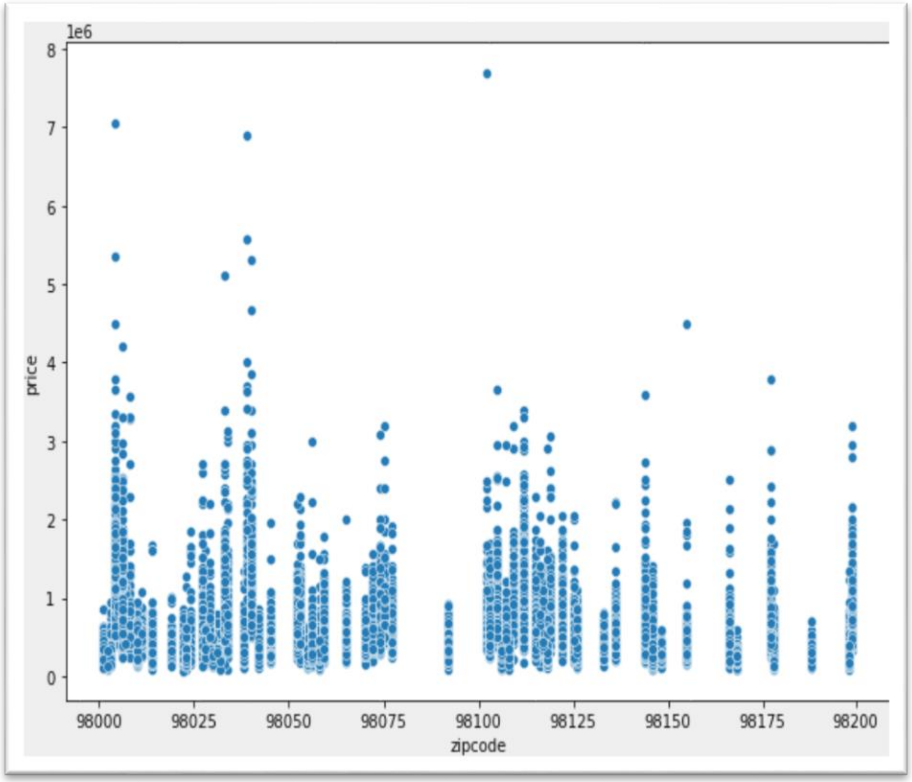


(b)living_measure15 vs living_measure

Figure 10: *Comparison chart*



(a)basement vs price



(b)zipcode vs price

Figure 11: *Comparison of price*

- *Surely the basement is an add on value for the seller for house*
- *As price has gone up by some extent when basement is there.*
- *Most of the houses are renovated in late 90s or after 2000s.*
- *Mostly renovated during the time of sales.*

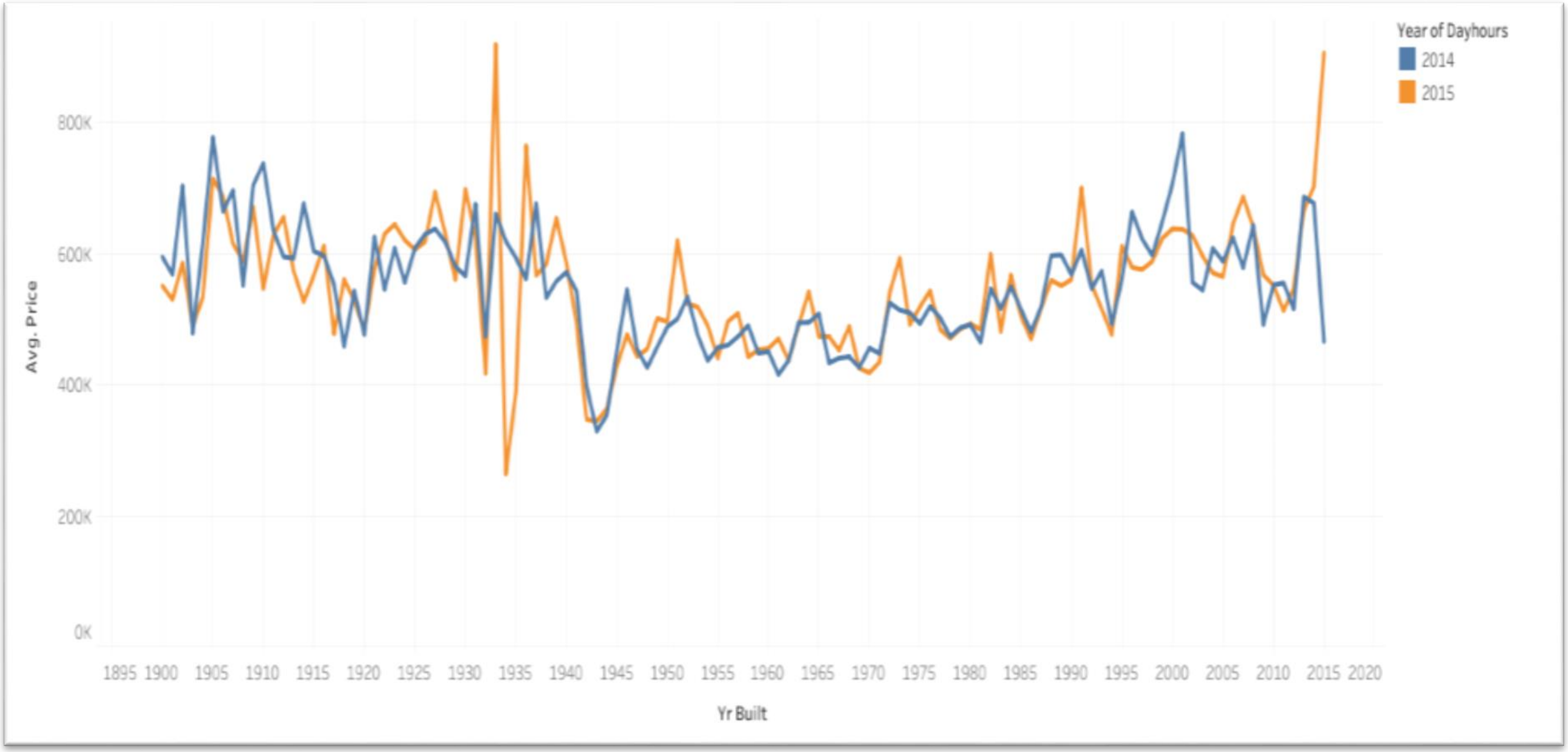


Figure 12: *Avg. price vs yr_built with the year sold*

- *Over the time the average price of house sold is same across year.*
- *But the more price variation is seen in 2015 rather than 2014.*
- *The price of house is more when difference between selling time and year built is less.*

4. Business Insights from Exploratory Data Analysis

To understand the data better we divided the given dataset into 03 different clusters by K-means method.(as shownin Appendix -III).

- *By looking at the table it looks like the data has been cluster into 03 groups of highly price, medium- and low-priced house.*
- *The people who are willing to pay any price they are going for more comfort and more ceiling space.*

- *But looks like data is not equally distributed over the groups.As low-priced house has around 16,000 data and medium priced have 362 data only.*
- *We can also confer that with the basement the price of house increases.*
- *From business perceptive the house category can be classified as Gold, Diamond and Platinum class and according to their needs and budget we can offer the house of their choice.*

5. Model building and Interpretation

<u>Assumptions made:</u>

I. *Data split into train and test set in ratio of 75:25.*
II. The VIF> 6 are considered as highly multicollinear.
III. Level of significance, p=0.05
IV. Hypothesis statements for Linear regression

H0: There is No relationship between Price and the corresponding variable.

H1: There is some relationship between Price and the corresponding variable.

a. Multiple Linear Regression

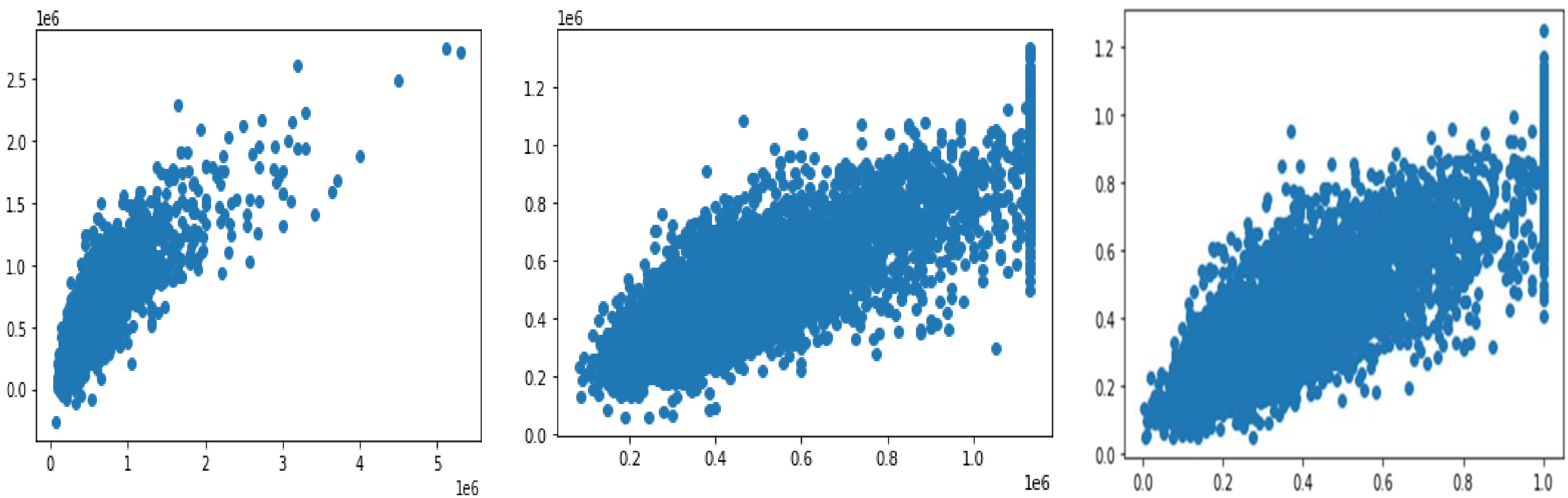
- *We are going to test our data with different models.*
- *Here we test our model with three different models as mentioned below:*
  - *With outliers and No scaling*
  - *Without outliers and No scaling*
    - ✓ *Data treatment before and after is shown in Appendix - IV*
  - *Without Outliers and Scaling*
    - ✓ *For scaling we choose to do the Min-Max Scaler method*

- *The results from all models are as follows:*

METHOD	*R-Squared*	*Adj. R-Squared*	*TRAIN*		*TEST*	
			SCORE	*RMSE*	*SCORE*	*RMSE*
MLP (with outliers and No scaling)	*0.651*	*0.65*	65.12	217223	66.37	142931
MLP (without outliers and No scaling)	*0.672*	*0.672*	67.2	142411	68.18	142931
MLP (without outliers and scaling)	*0.679*	*0.679*	67.9	0.1353	65.96	0.1351

Table 6: *Values of different MLP models*

- *From the given table we can confer that best model the three is the MLP with scaling and no outliers model.*
- *As the score and the adj. R^2 value is higher as compare to others and train score is also higher.*
- *The RMSE value is also the least among the models.*



(a) with outliers and No scaling *(b) without outliers and No scaling* *(c) without outliers and scaling*

Figure 13: *Scatterplot between predicted Y-value vs Actual Y-value*

- *As in the above figure the comparison shows that the predicted values just clutter over a single point. But in the other two figures [b & c] the values trend to follow a line pattern.*
- *The vif value of the best model above 03 models we found that for most of variables vif value is < 30.*

b. Others Model

- *Other than Multiple Linear Regression, We used:*
  - ✓ ***KNN,***
  - ✓ ***Random Forest** and*
  - ✓ ***ANN algorithms** to achieve our goal.*
- *A comparison is shown in below table.*

METHOD	TRAIN		TEST	
	SCORE	RMSE	SCORE	RMSE
MLP (without outliers and scaling)	67.9	0.1353	65.96	0.1351
KNN(N=10)	99.8	0.0082	69.03	0.129
Random Forest	76.58	0.115	72.09	0.122
ANN	62.39	0.145	60.89	0.1451

Table 7: *Values of different models*

- *From the above data it very much clear that Random Forest model has outperformed other models.*
- *It has a score of 76% which much higher than the other models.*
- *Even the RMSE is the least among others.*

6. Model Tuning

- *As the Random Forest model is the best performing model among others. Let's try to improve the model with Bagging and Gradient Boosting technique.*
- *Among the two models Gradient Boosting model is more acceptable as the score of train and test are close to each other whereas the score for Bagging is not matching.*

METHOD	TRAIN		TEST	
	SCORE	RMSE	SCORE	RMSE
BAGGING	93.5	0.1	81.11	0.06
GRADIENT BOOSTING	79.9	0.106	77.1	0.11

Table 8: *Values for Bagging and Gradient Boosting models*

- *Hence on the basis of above table we come to conclusion that The Gradient Boosting model is best fitted for this data.*
- *As the score is better and the error is also minimal.*

7. Model Selection: Evaluation Parameters

The parameters we used to evaluate the model are mentioned below:

- ***Accuracy:*** *Accuracy is defined as the percentage of correct prediction over total values of data.*
- ***MSE(Mean Squared Error):*** *MSE is the average of the squared difference between actual and Predicted value*
- ***RMSE(Root Mean Sq.Error):*** *MSE is the squared root of the average of the squared difference between actual and Predicted value, or the square root of MSE.*
- ***R2:*** *R2 is a statistical measure of how well the regression predictions approximate the real data points. It ranges from 0 to 1*

8. Business Insights and Recommendations

- *From this model building we have found the best model for predicting the house price is given by Gradient Boosting Model as it is able to predict correct house price by almost 80% accuracy and with an error of around 0.1.*
- *And we found hat some important variables are:*

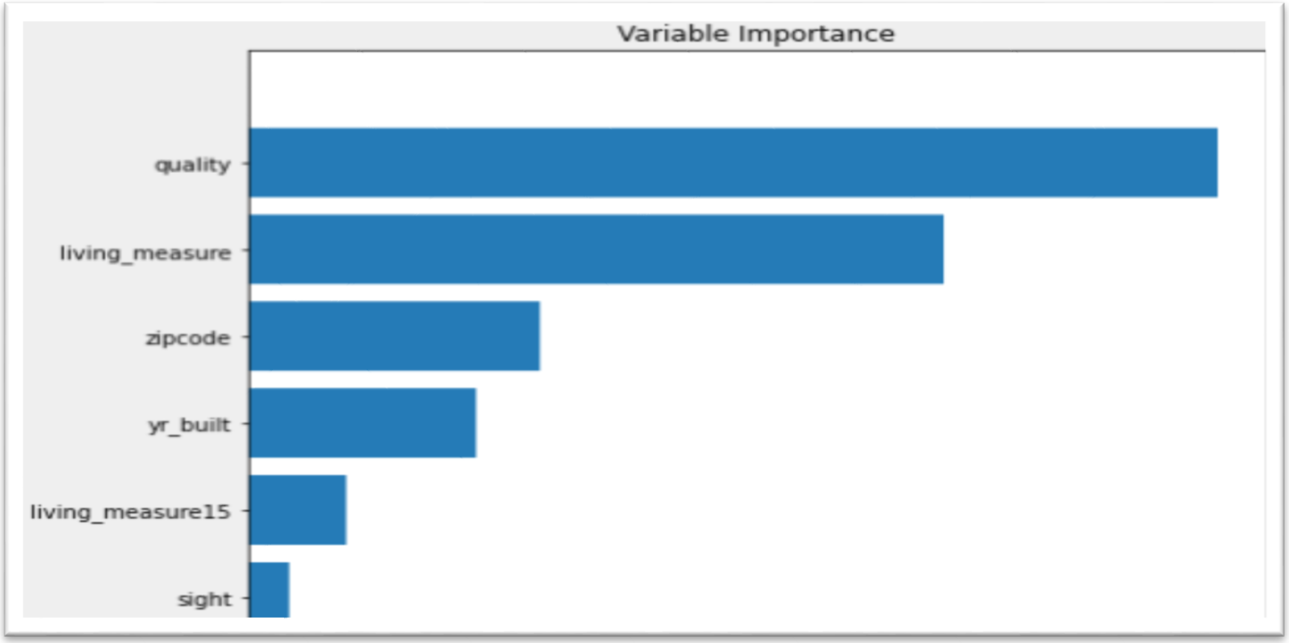


Figure 14: *Important Variables*

- *Need to be thoughtful about the location of the house*
- *Need to keep the quality of the house in check as it is the most important factor*
- *Different price segmentation must be made according to the requirements of the buyers.*
- *For low quality and condition of house renovation must be done for up in house price.*
- *Need to focus on some different aspects while choosing the house which is not mentioned in the dataset like locality, distance from important places, etc.*

_______ *THE END* _______