Final Report of House Price Prediction

Submitted in Partial Fulfilment of requirements for the Award of Certificate of

Post Graduate Program in Data Science and Business Analytics

Capstone Project Report

Submitted to

Institute of Management

Global Mindset – Indian Roots

Submitted By:
Group No.6
Batch: PGP-DSBA-July, 2022
01/07/2023

Group Members
Sangeeth A.
Shruti Tripathi

Under the Guidance of

Abhay Poddar

CERTIFICATE OF COMPLETION SIGNED BY MENTOR

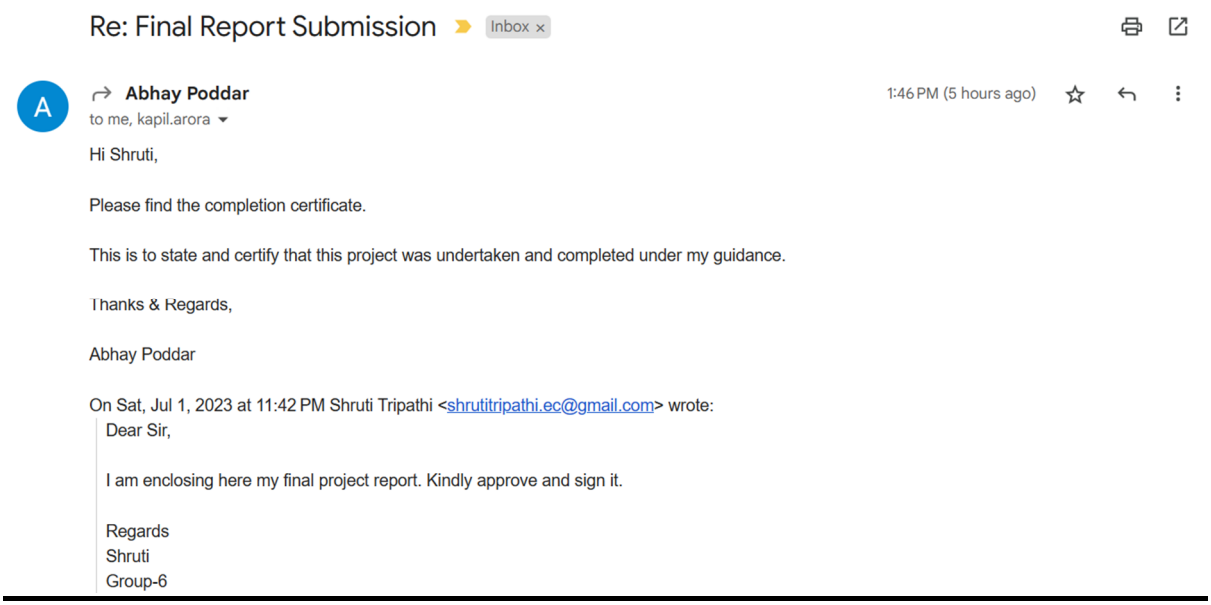
This is to certify that the participants Mr. Sangeeth A. and Ms. Shruti Tripathi who are the students of Great Learning, have successfully completed their project on ***"House Price Prediction"***.

This project is the record of authentic work carried out by them during the academic year 2022-2023.

Mentor: Abhay Poddar

Date: 02 July 2023

Place:



Re: Final Report Submission Inbox ×

A

Abhay Poddar 1:46 PM (5 hours ago)

to me, kapil.arora

Hi Shruti,

Please find the completion certificate.

This is to state and certify that this project was undertaken and completed under my guidance.

Thanks & Regards,

Abhay Poddar

On Sat, Jul 1, 2023 at 11:42 PM Shruti Tripathi <shrutitripathi.ec@gmail.com> wrote:

> Dear Sir,
>
> I am enclosing here my final project report. Kindly approve and sign it.
>
> Regards
> Shruti
> Group-6

Executive Summary

A house is more precious thing for a person who wants to buy or build a home for him/her self and for his/her family because it takes a lot to set that. And value of house is dependent on lots of the components that makes it expensive, luxury, affordable and easy to connect.

Here through this project we are finding all the important components that decides or predict the price of a house.

The data received tells us following:

1. **Problem Statement:** 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 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 gathers data you will try to assess you house price.

2. **Data Description:** Data provided to predict the price (target variable) of a house is dependent on the variables like measurement of house considerably living_measure, ratings in terms of quality, site views, furnished status, number of floors preferably top floor should not be fully constructed or half construction is preference and age of the house.

3. **Main Result:** From the model building XG Boost model best performed model followed by Gradient Boost model. Feature like "quality" is most important feature for deciding the price which straight away affecting the valuation of a house. And most of the houses are in Seattle and Bellevue location which are getting good ratings and both locations are have coast site, which is making it most preferred locations in all aspects.

4. **Recommendations:** From the business point there are other coastal sites as per the data which can be attraction for selecting a location therefore

business can build on those locations by offering people new build houses with preferences like number bed rooms, bathrooms and number of floors as we come to know from model building.

Now from a persons point who want to sell his or her house can predict the price of a house on the basis of selected features given by model.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

Std	Standard Deviation
min	Minimum
Max	Maximum
Freq	Frequency
OLS	Ordinary Least Square
VIF	Variance Inflation Factor
XGB	Extreme Gradient Boosting
KNN	K – Nearest Neighbour
MAE	Mean Absolute Error
MSE	Mean Squares Error
MAPE	Mean Absolute Percentage Error

Introduction to Business Problem

A House value is 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 neighbourhood and based on gathered data you will try to assess your house price.

Problem Statement

On the basis of feature variable available in collected data, prediction of house price.

Scope

To find the optimal model and the variables impacting the target variable (price of house), perform EDA and predictive model analysis using different techniques.

Objective

The objective is to build a model through which price of a house can be predicted, using historical data finding various feature variables available in data which have more impact in terms of pricing.

Dataset Details

Dataset has following details:
The features of the dataset are,

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

Visual Inspection of Data

```
No. of rows 21613
No. of columns 23


Number of duplicate rows = 0
(21613, 23)
```

Understanding of attributes (variable info, renaming if required)

From the below Table No.1 *Information of variable type* we can see that there are 22 columns wherein and 2161 3 rows in the dataset one column [dayhours] is datetime type, 12 columns type [room_bed, room_bath, living_m easure, lot_measure, sight, quality, ceil_measure, basement, lat, living_measure15, lot_measure15, furnished] are float, 4 columns [cid, price, yr_renovated & zipcode] are integer type and 6 columns [ceil, coast, condition, yr_built, long & total_area]are object type.
Further, it also seems that there are null values in the dataset in few columns.

```
<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  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  object
 8   coast             21612 non-null  object
 9   sight             21556 non-null  float64
 10  condition         21556 non-null  object
 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  object
 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  float64
 22  total_area        21584 non-null  object
dtypes: datetime64[ns](1), float64(12), int64(4), object(6)
```

Table No. 1 Information of variables type

Exploratory data analysis

```
Column: lot_measure
count    2.157100e+04
mean     1.510458e+04
std      4.142362e+04
min      5.200000e+02
25%      5.040000e+03
50%      7.618000e+03
75%      1.068450e+04
max      1.651359e+06
Name: lot_measure, dtype: float64

Column: ceil
count     21571
unique        7
top           1
freq      10647
Name: ceil, dtype: int64
```

```
Column: coast
count     21612
unique        3
top           0
freq      21421
Name: coast, dtype: int64

Column: sight
count    21556.000000
mean         0.234366
std          0.766438
min          0.000000
25%          0.000000
50%          0.000000
75%          0.000000
max          4.000000
Name: sight, dtype: float64
```

```
Column: condition
count     21556
unique        6
top           3
freq      13978
Name: condition, dtype: int64

Column: quality
count    21612.000000
mean         7.656857
std          1.175484
min          1.000000
25%          7.000000
50%          7.000000
75%          8.000000
max         13.000000
Name: quality, dtype: float64
```

```
Column: basement
count    21612.000000
mean       291.522534
std        442.580840
min          0.000000
25%          0.000000
50%          0.000000
75%        560.000000
max       4820.000000
Name: basement, dtype: float64

Column: yr_built
count     21612
unique      117
top        2014
freq        559
Name: yr_built, dtype: int64
```

```
Column: yr_renovated
count    21613.000000
mean        84.402258
std        401.679240
min          0.000000
25%          0.000000
50%          0.000000
75%          0.000000
max       2015.000000
Name: yr_renovated, dtype: float64
```

```
Column: furnished
count    21584.000000
mean         0.196720
std          0.397528
min          0.000000
25%          0.000000
50%          0.000000
75%          0.000000
max          1.000000
Name: furnished, dtype: float64

Column: total_area
count     21584
unique    11145
top           $
freq         39
Name: total_area, dtype: object
```

Table No.2 Descriptive Analysis

From the above Table No.2 *Descriptive Analysis* it is clear that most of the columns are right skewed like price, room_bed, room_bath, living_measure, lot_measure, ceil_measure, basement, year_renovated, living_measure15, lot_measur15 and total_area. Subsequently, in variable total_area there are object type value is present and for variable room_bed mean value is 3.3 and maximum value is 33 it means there are outliers as 33 is extreme value for this variable. Similar behaviour is shown by room_bath, living_measure and basement wherein mean is on lower side and maximum are extreme values.

From figure no.1 *Columns with special characters*, it seems that almost every column have special character i.e., $ type values.

```
['cid', 'dayhours', 'price', 'room_bed', 'room_bath', 'living_measure', 'lot_measure', 'ceil', 'coast', 'sight', 'condition',
'quality', 'ceil_measure', 'basement', 'yr_built', 'yr_renovated', 'zipcode', 'lat', 'long', 'living_measure15', 'lot_measure1
5', 'furnished', 'total_area']
```

Figure No. 1 Columns with special characters

```
<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  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              21541 non-null  float64
 8   coast             21582 non-null  float64
 9   sight             21556 non-null  float64
 10  condition         21528 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          21598 non-null  float64
 15  yr_renovated      21613 non-null  int64
 16  zipcode           21613 non-null  int64
 17  lat               21613 non-null  float64
 18  long              21579 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        21545 non-null  float64

dtypes: datetime64[ns](1), float64(18), int64(4)
```

Table No. 3 Dataset information after replacement of special character with null values

From the Table No.3 it seems that after replacement of special character with null values all the object type variables now become float type for example total_area and the null values are also increased.

```
cid                 0
dayhours            0
price               0
room_bed          108
room_bath         108
living_measure     17
lot_measure        42
ceil               72
coast              31
sight              57
condition          85
quality             1
ceil_measure        1
basement            1
yr_built           15
yr_renovated        0
zipcode             0
lat                 0
long               34
living_measure15  166
lot_measure15      29
furnished          29
total_area         68
dtype: int64
```

Figure No. 2 Total Null Values

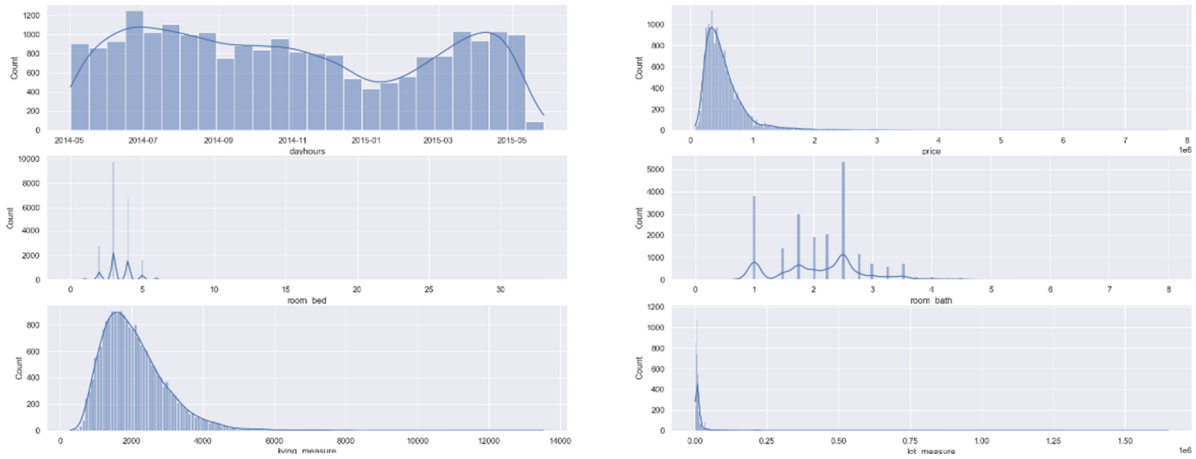
From the above Figure no.2 *Total Null Values,* it is clear that there are null values present in each 22 variables except cid, dayhours, price, yr_renovated, zipcode and lat.

```
cid                0.000000
dayhours           0.000000
price              0.000000
room_bed           0.499699
room_bath          0.499699
living_measure     0.078656
lot_measure        0.194327
ceil               0.333133
coast              0.143432
sight              0.263730
condition          0.393282
quality            0.004627
ceil_measure       0.004627
basement           0.004627
yr_built           0.069403
yr_renovated       0.000000
zipcode            0.000000
lat                0.000000
long               0.157313
living_measure15   0.768056
lot_measure15      0.134179
furnished          0.134179
total_area         0.314625
dtype: float64
```

Figure No. 3 Null value percentage

From the Figure No. 3 *Null value percentage,* it seems null values are in very small quantum in comparison of total number entries for each variable. Therefore, dropping them instead of treating them is better way to deal with null values.

Univariate analysis



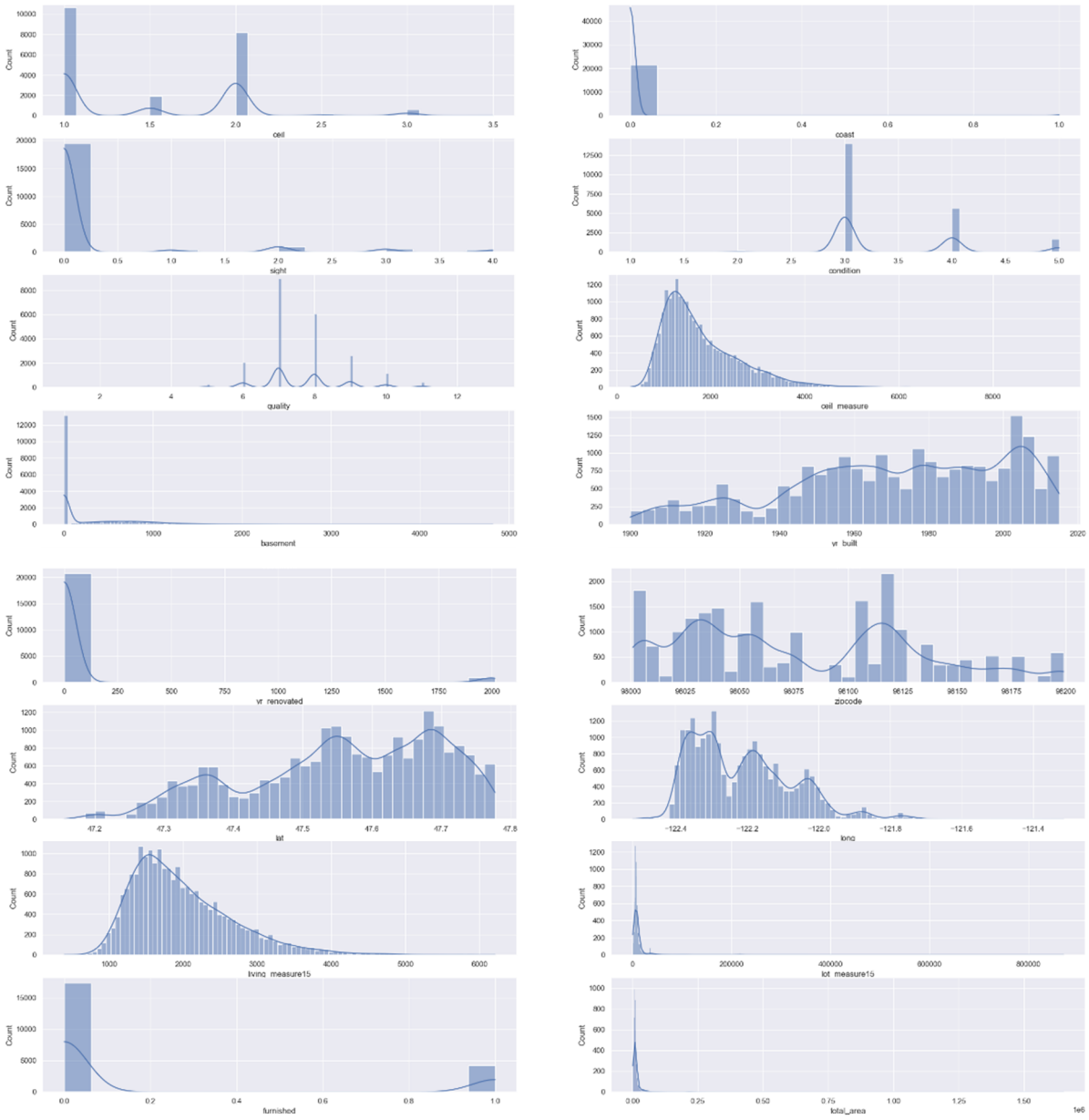
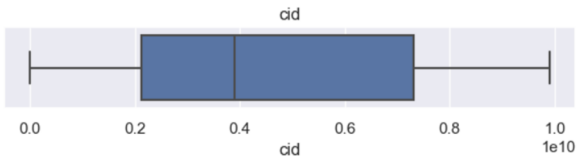
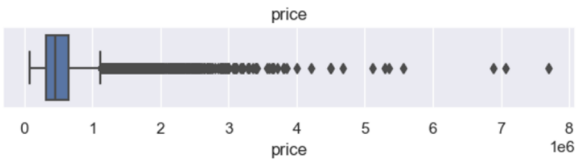


Figure No.4 Histogram of continuous variables

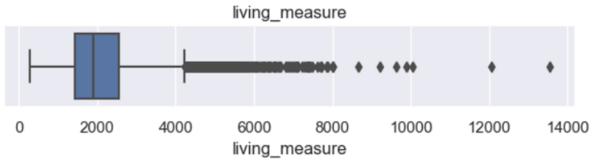
From Figure No.4 Histograms for continuous variable it is clear that data is not normally distributed and right skewed for variables like price, room_bed, room_bath, living_measure, lot_measure, ceil_measure, basement, year_renovated, living_measure15, lot_measur15 and total_area. However, for variable like dayhours(sale date), year built and zipcode data distribution is all over.

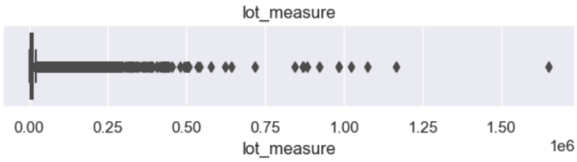

cid
0.0
0.2
0.4
0.6
0.8
1.0
1e10
cid

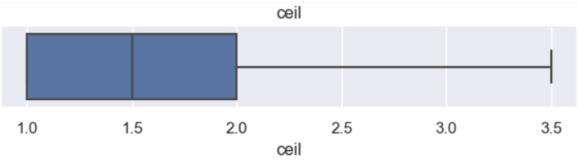

price
0
1
2
3
4
5
6
7
8
1e6
price

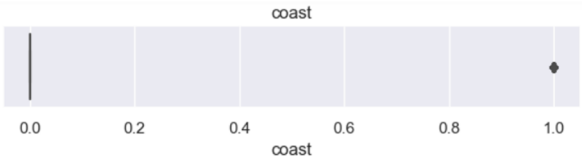

room_bed
0
5
10
15
20
25
30
room_bed


room_bath
0
1
2
3
4
5
6
7
8
room_bath

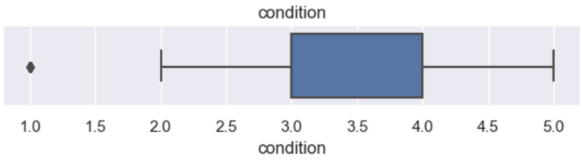

living_measure
0
2000
4000
6000
8000
10000
12000
14000
living_measure

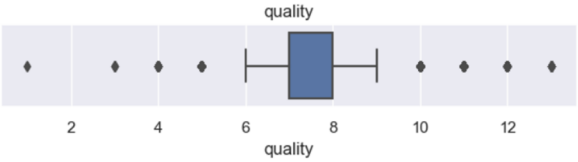

lot_measure
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1e6
lot_measure

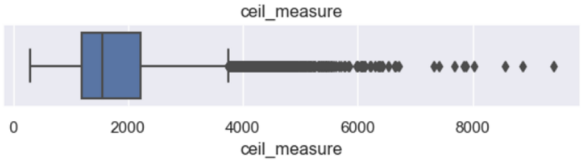

ceil
1.0
1.5
2.0
2.5
3.0
3.5
ceil

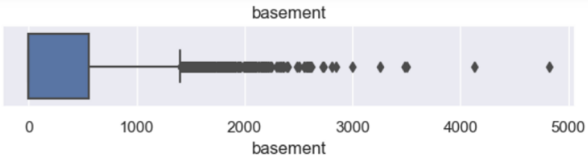

coast
0.0
0.2
0.4
0.6
0.8
1.0
coast

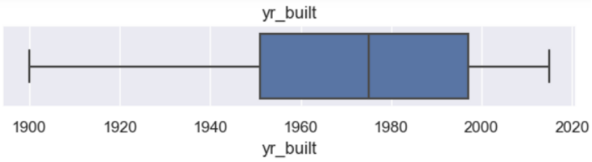

sight
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
sight

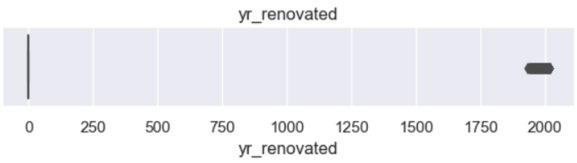

condition
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
condition


quality
2
4
6
8
10
12
quality

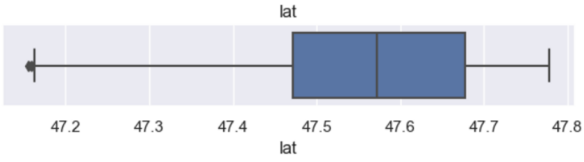

ceil_measure
0
2000
4000
6000
8000
ceil_measure

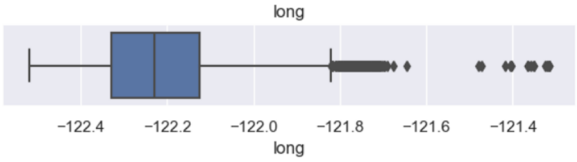

basement
0
1000
2000
3000
4000
5000
basement


yr_built
1900
1920
1940
1960
1980
2000
2020
yr_built


yr_renovated
0
250
500
750
1000
1250
1500
1750
2000
yr_renovated


zipcode
98000
98025
98050
98075
98100
98125
98150
98175
98200
zipcode


lat
47.2
47.3
47.4
47.5
47.6
47.7
47.8
lat


long
−122.4
−122.2
−122.0
−121.8
−121.6
−121.4
long

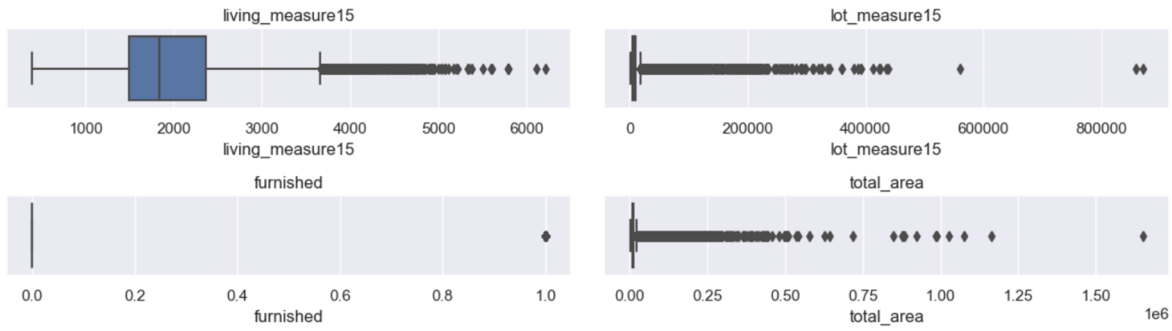


Figure No.5 Boxplot of continuous variables

From the Figure No.5 Boxplot for continuous variables, it seems outliers are present in every variables except ceil, yr_built, and zipcode variables. In variables like room_bed, lot_measure, living_measure, total_area, lot_measure15 have extreme values. Although, outliers are extreme values but for these variables they seems wrongly entered values as the extreme values are at very high distance than other existing outliers.

Bivariate/Multivariate analysis

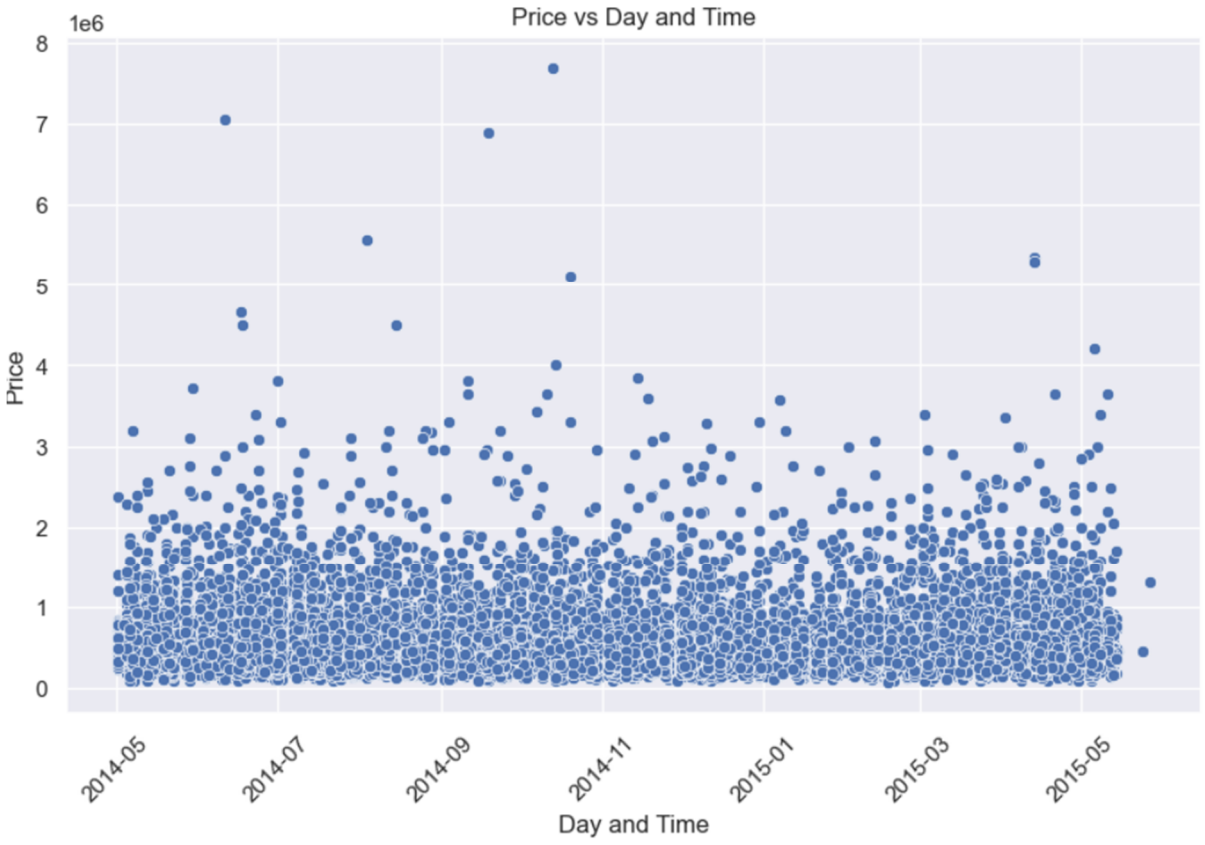


Figure No.6 Scatter Plot between Price and Sale Date

From the Figure No.6 Scatter Plot between price and dayhours (sale date) it seems data is distributed all over. Price is not very much dependent on when it was sell but still it looks like price was high for the period of May 2014 to November 2014.

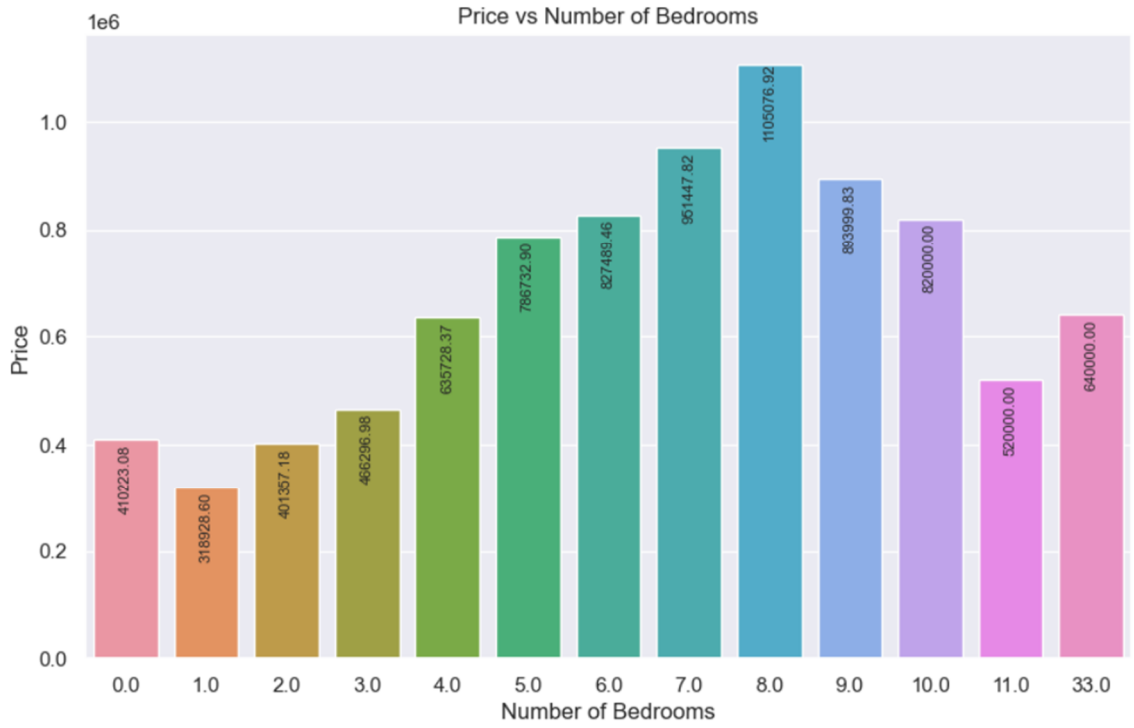


Figure No.7 Bar Plot between Price and Number of Bedrooms

From the Figure No.7 Bar Plot between price and Number of bed rooms it seems that price is high for house which have 8 bedrooms. And the 33 bed rooms looks like extreme value which can be wrongly entered.

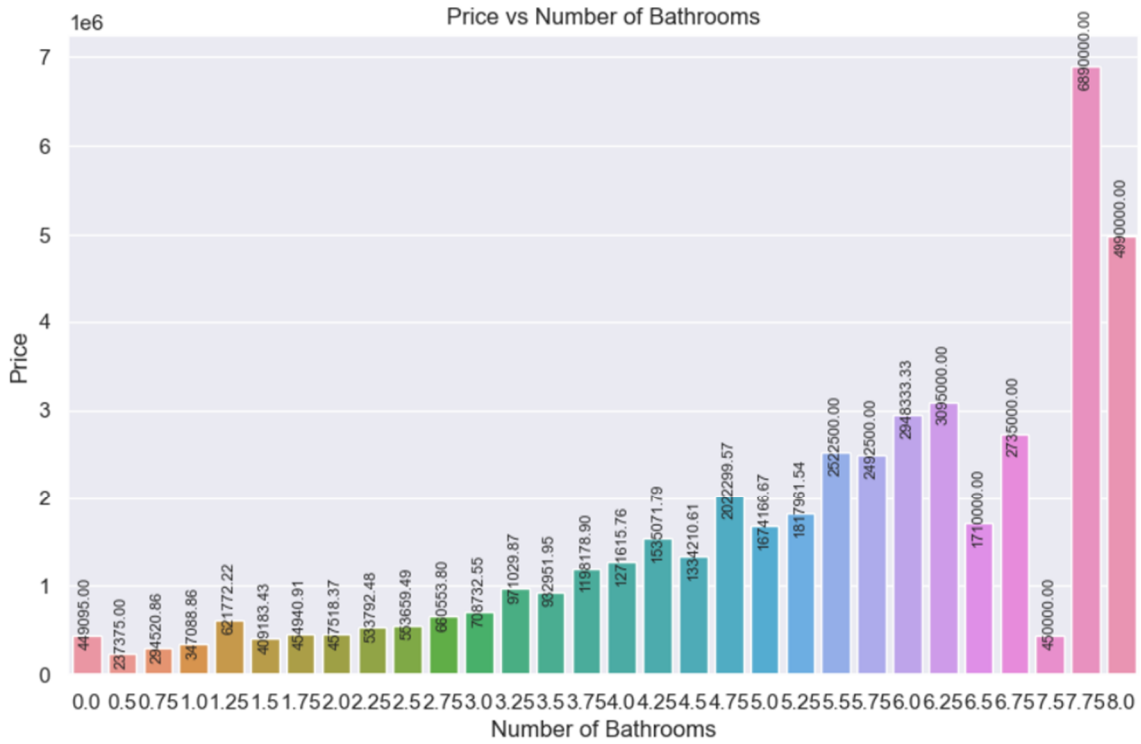


Figure No.8 Bar Plot between Price and Number of Bathrooms

From the Figure No.8 Bar Plot between price and number of bath rooms, it seems that price is high for higher the number of bath rooms.

Figure No.9 Scatter Plot between Price and Living Area

From the Figure No.9 Scatter Plot between price and living measure, it seems price is increasing with the increase in living area measurement. But some values looks like extreme values for this variable.

Figure No.10 Scatter Plot between Price and Lot Area

From the Figure No.10 Scatter Plot between price and Lot measure, it seems that price is high for small lot area more than the bigger lot area measures.

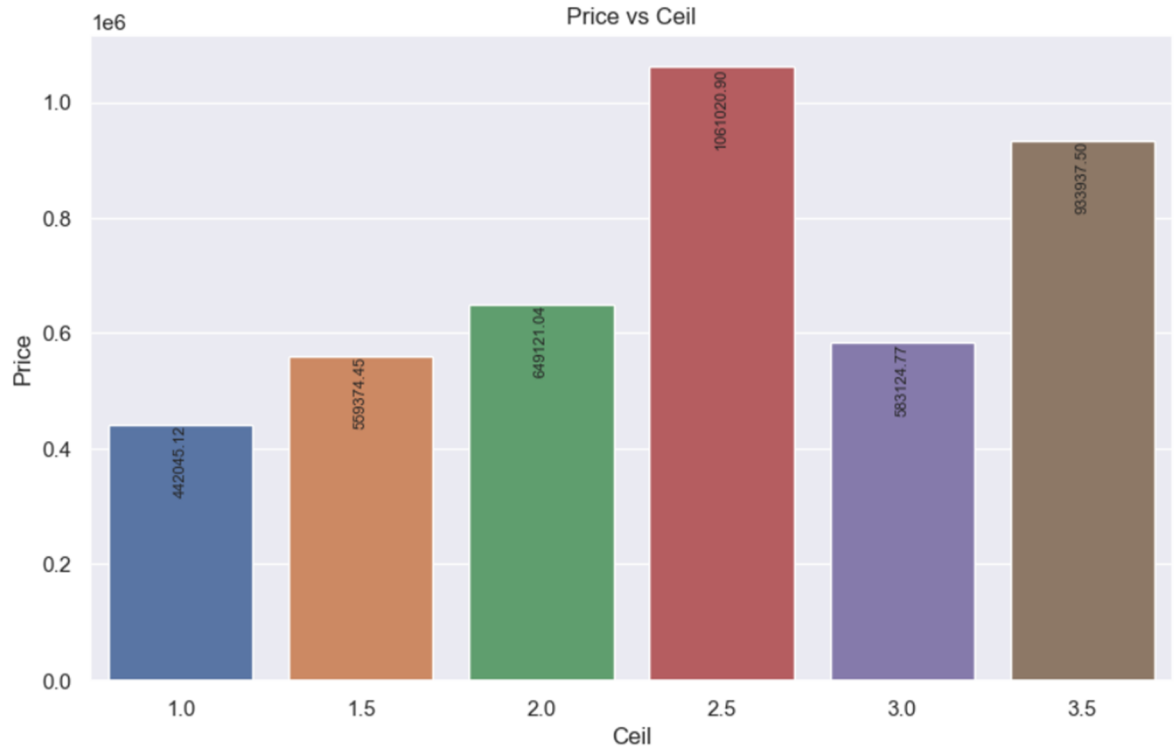


Figure No.11 Bar Plot between Price and Ceil (total no. of floors)

From the Figure No.11 Bar Plot between price and ceil (no. of floors), it seems that price is high for properties which have1.5, 2.5 & 3.5 floors. Complete constructed properties have lower price as people preferred bit open area in their houses.

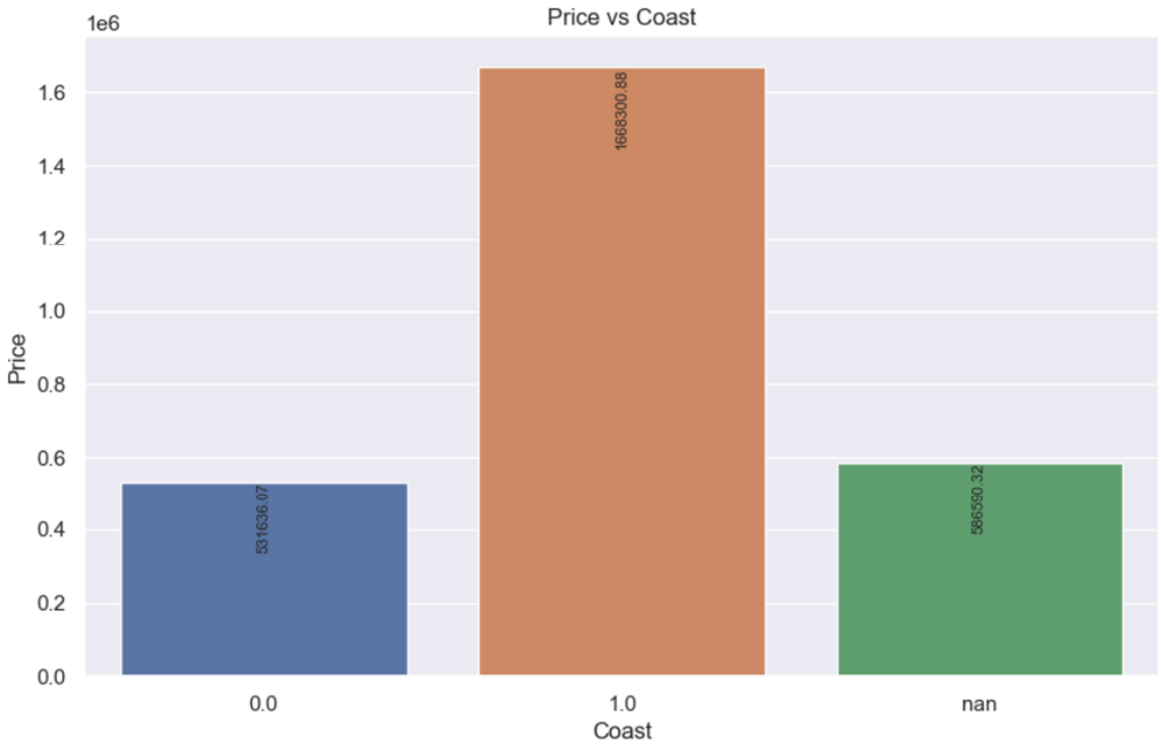


Figure No.12 Bar Plot between Price and Coast

From the Figure No.12 Bar Plot between price and coast, it seems that price is high for properties which have water fronts.

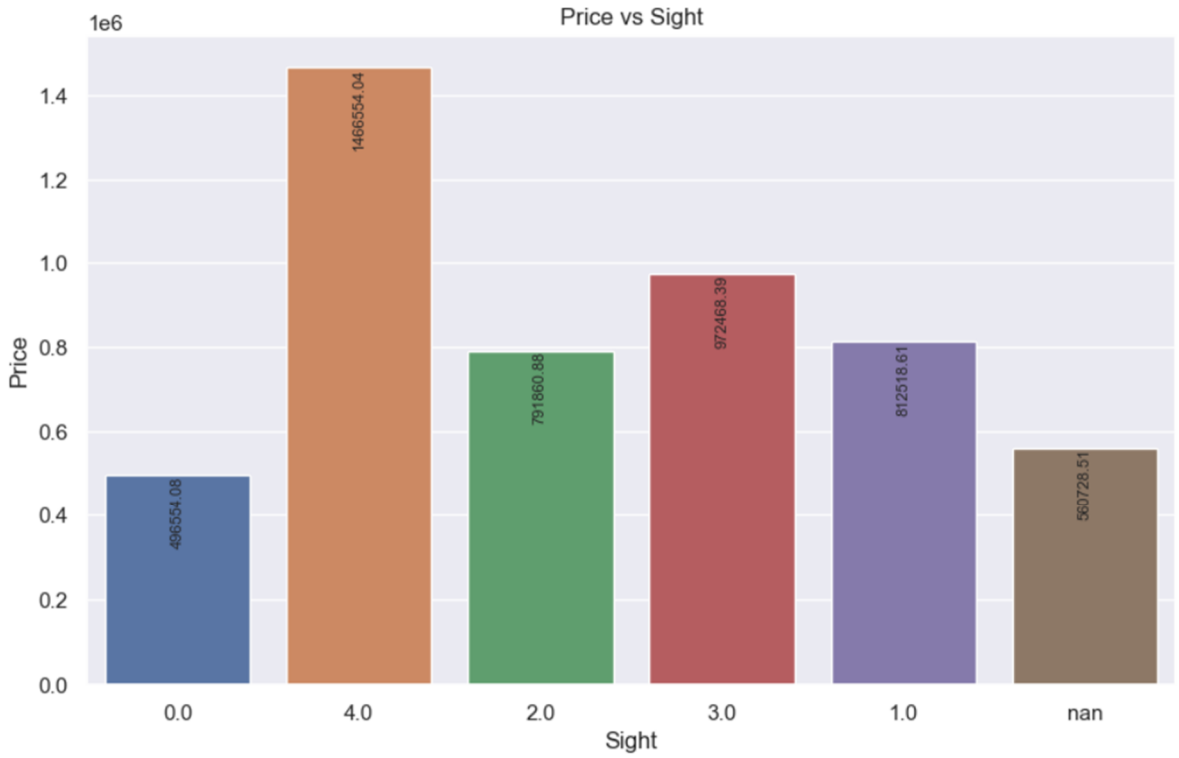


Figure No.13 Bar Plot between Price and Sight

From the Figure No.13 Bar Plot between price and sight (no. of views), it seems that price is high for properties which have higher the number of views.

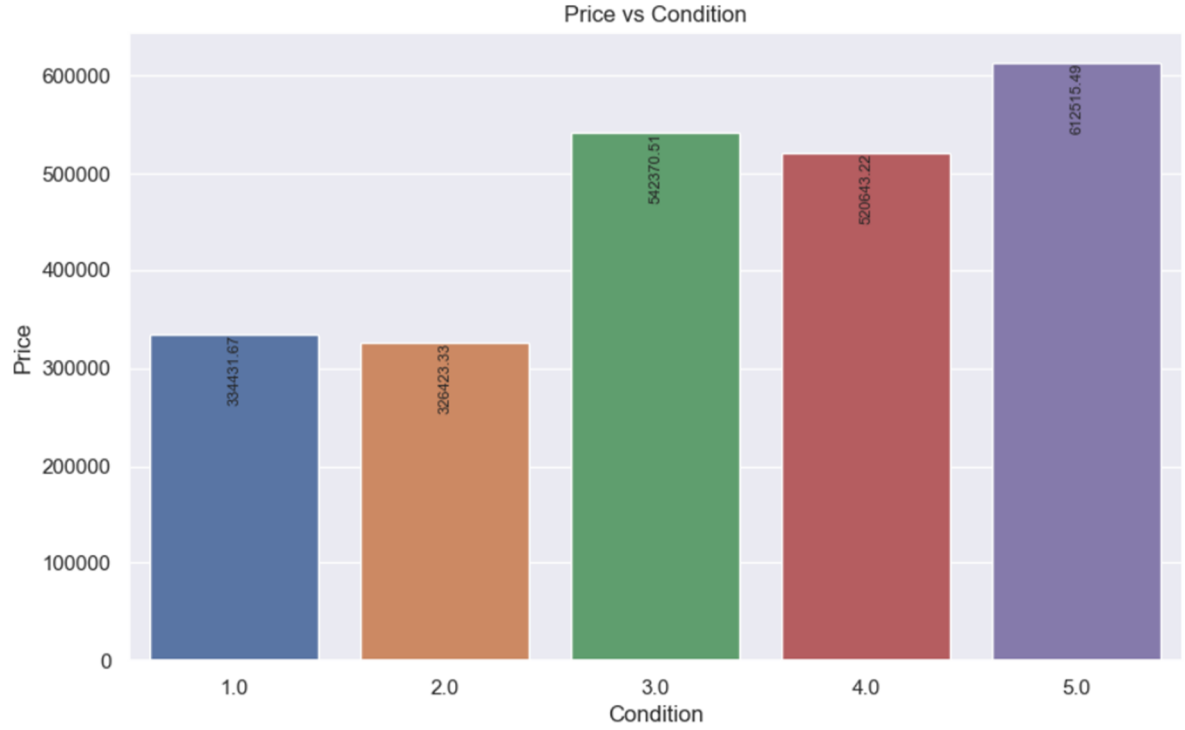


Figure No.14 Bar Plot between Price and Condition

From the Figure No.14 Bar Plot between price and condition, it is clear that price is high for properties are in good condition.

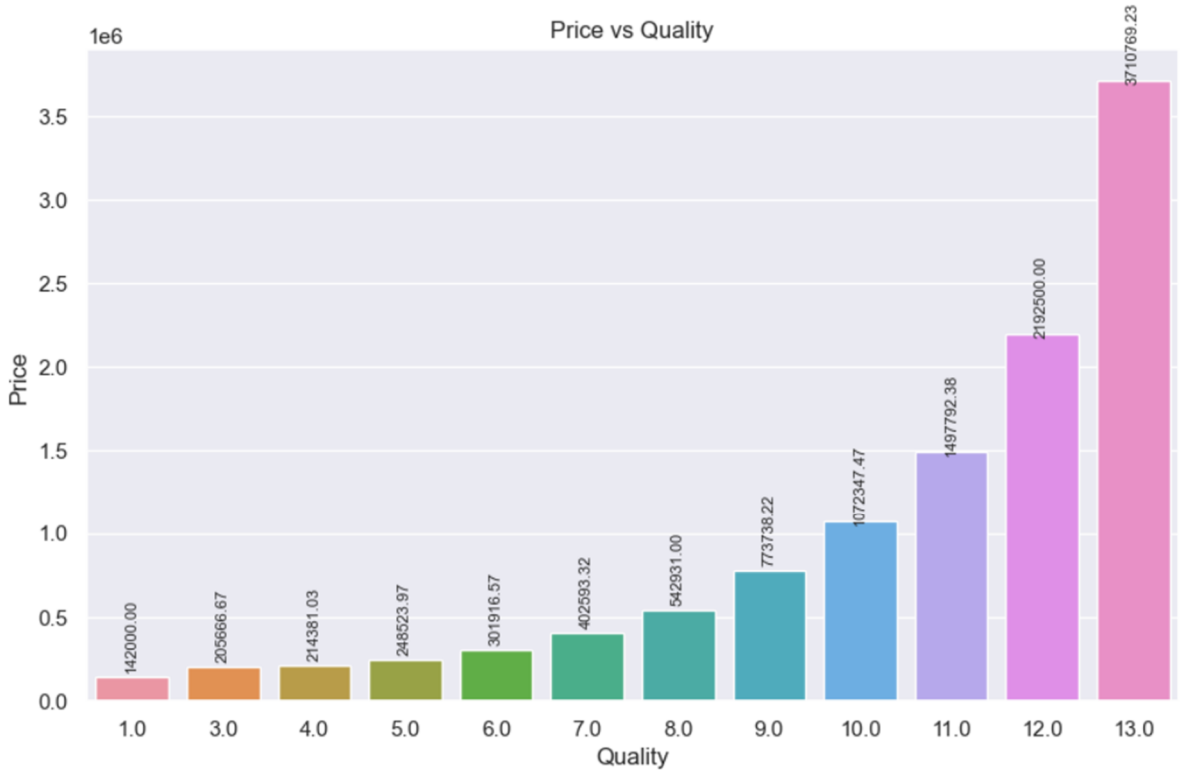


Figure No.15 Bar Plot between Price and Quality

From the Figure No.15 Bar Plot between price and quality (grades), it seems that price is high for properties which are given higher grades as per grading system.

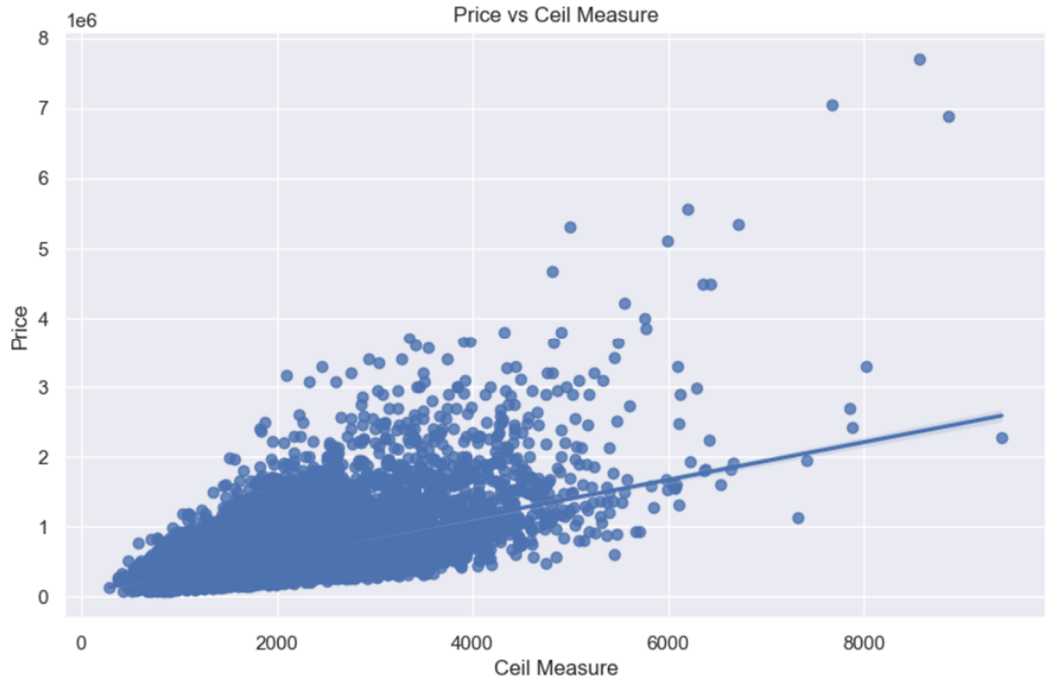


Figure No.16 Reg Plot between Price and Ceil Measure

From the Figure No.16 Reg Plot between price and Ceil_measure (Floor measurements) it is clear that as the ceil measure is increasing price is also increasing.

Figure No.17 Reg Plot between Price and Basement

From the Figure No.17 Reg Plot between price and Basement, it is clear that price is not dependent on the basement_measure as price is not increasing much with basement measure.

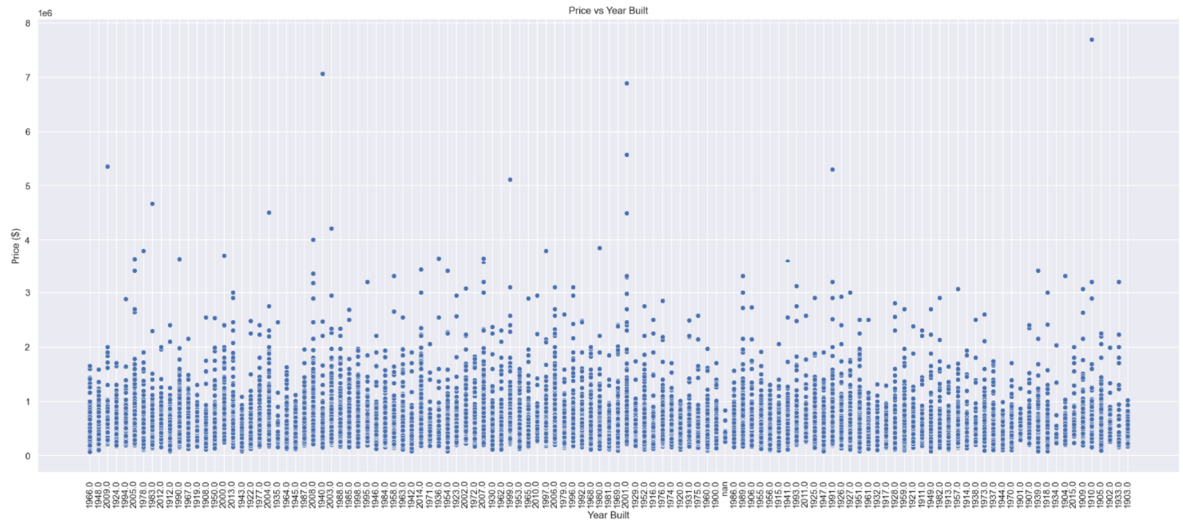


Figure No.18 Scatter Plot between Price and Year Built

From the Figure No.18 Scatter Plot between price and Year Built, it seems that price is not dependent on the year built as price is not increasing with year house was built. And it is also clearly visible that this variable have extreme values.

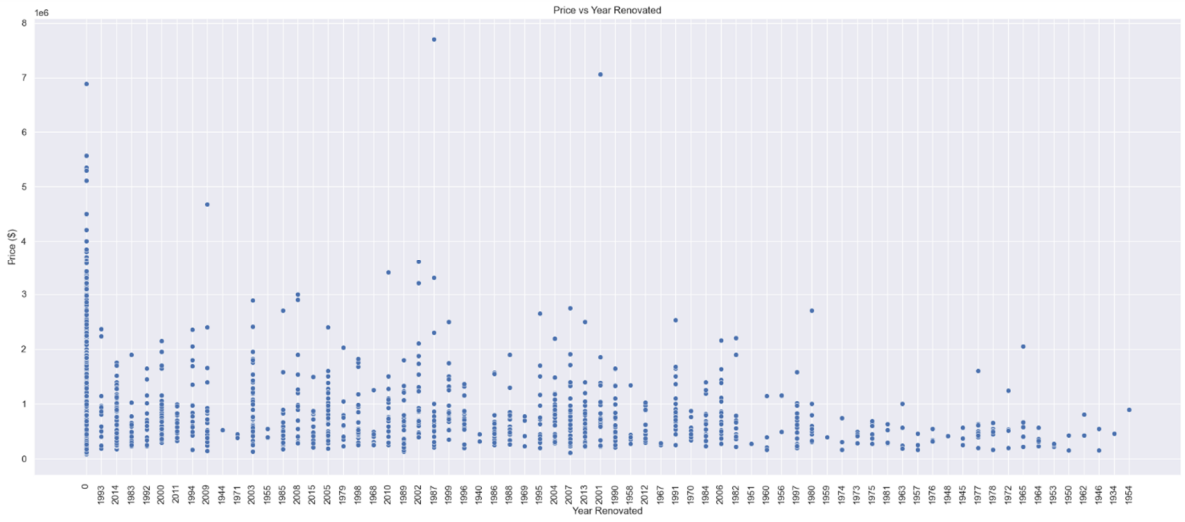


Figure No.19 Scatter Plot between Price and Year Renovated

From the Figure No.19 Scatter Plot between price and Year Renovated, it seems that price is much high for the properties which are not renovated. Further, is also clearly visible that this variable have extreme values.

	cid	dayhours	price	room_bed	room_bath	living_measure	lot_measure	ceil	coast	sight	condition	quality	ceil_measure
0	3876100940	2015-04-27	600000	4.0	1.75	3050.0	9440.0	1.0	0.0	0.0	3.0	8.0	1800.0
1	3145600250	2015-03-17	190000	2.0	1.00	670.0	3101.0	1.0	0.0	0.0	4.0	6.0	670.0
2	7129303070	2014-08-20	735000	4.0	2.75	3040.0	2415.0	2.0	1.0	4.0	3.0	8.0	3040.0
3	7338220280	2014-10-10	257000	3.0	2.50	1740.0	3721.0	2.0	0.0	0.0	3.0	8.0	1740.0
4	7950300670	2015-02-18	450000	2.0	1.00	1120.0	4590.0	1.0	0.0	0.0	3.0	7.0	1120.0

basement	yr_built	yr_renovated	zipcode	lat	long	living_measure15	lot_measure15	furnished	total_area	location
1250.0	1966.0	0	98034	47.7228	-122.183	2020.0	8660.0	0.0	12490.0	Kirkland
0.0	1948.0	0	98118	47.5546	-122.274	1660.0	4100.0	0.0	3771.0	Seattle
0.0	1966.0	0	98118	47.5188	-122.256	2620.0	2433.0	0.0	5455.0	Seattle
0.0	2009.0	0	98002	47.3363	-122.213	2030.0	3794.0	0.0	5461.0	Auburn
0.0	1924.0	0	98118	47.5663	-122.285	1120.0	5100.0	0.0	5710.0	Seattle

Table No.4 Dataset after adding new variable 'Location'

From the Table No. 4 Dataset after adding new variable 'Location', there is new variable added "Location" which is derived with the help of variable "Zipcode". This new variable could be helpful in

the prediction of price of house as location is important factor for evaluating the value of a property / house.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 24 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   cid               21613 non-null  int64
 1   dayhours          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              21541 non-null  float64
 8   coast             21613 non-null  object
 9   sight             21613 non-null  object
 10  condition         21528 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          21613 non-null  object
 15  yr_renovated      21613 non-null  object
 16  zipcode           21613 non-null  object
 17  lat               21613 non-null  float64
 18  long              21579 non-null  float64
 19  living_measure15  21447 non-null  float64
 20  lot_measure15     21584 non-null  float64
 21  furnished         21613 non-null  object
 22  total_area        21545 non-null  float64
 23  location          21613 non-null  object
dtypes: datetime64[ns](1), float64(14), int64(2), object(7)
```

Table No.5 Dataset information after adding new variable 'Location'

From the Table No.5 Dataset information, new variable Location is added in dataset and it has object type values.

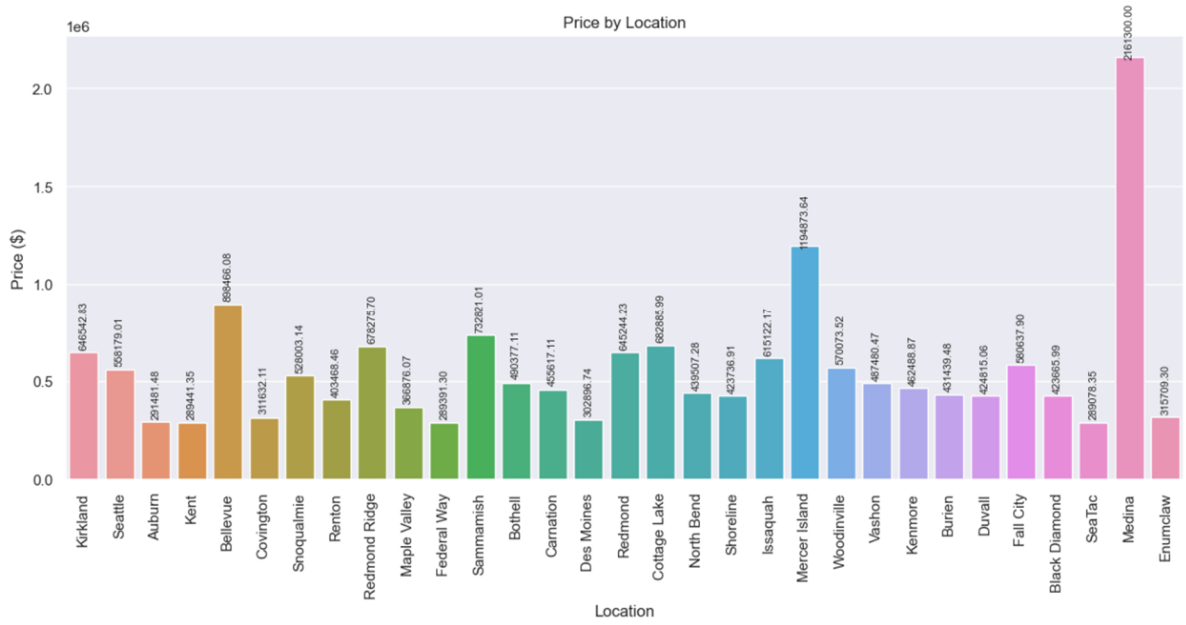


Figure No.20 Bar Plot between Price and Location

From the Figure No.20 Bar Plot between price and Location, it is clearly visible that 'Medina' location has high pricing for their properties / house followed by 'Mercer Island'.

Figure No.21 Scatter Plot between Price and Living Measure15

From the Figure No.21 Scatter Plot between price and Living Measure15, it seems that price is increasing with living measure increase after renovation in 2015.

Figure No.22 Scatter Plot between Price and Lot Measure15

From the Figure No.22 Scatter Plot between price and Lot Measure15, it seems that price is not very much dependent on the Lot Measure even after renovation in 2015 as price is for small lot measures than bigger lot measures. And it is also clear that this variable have extreme values.

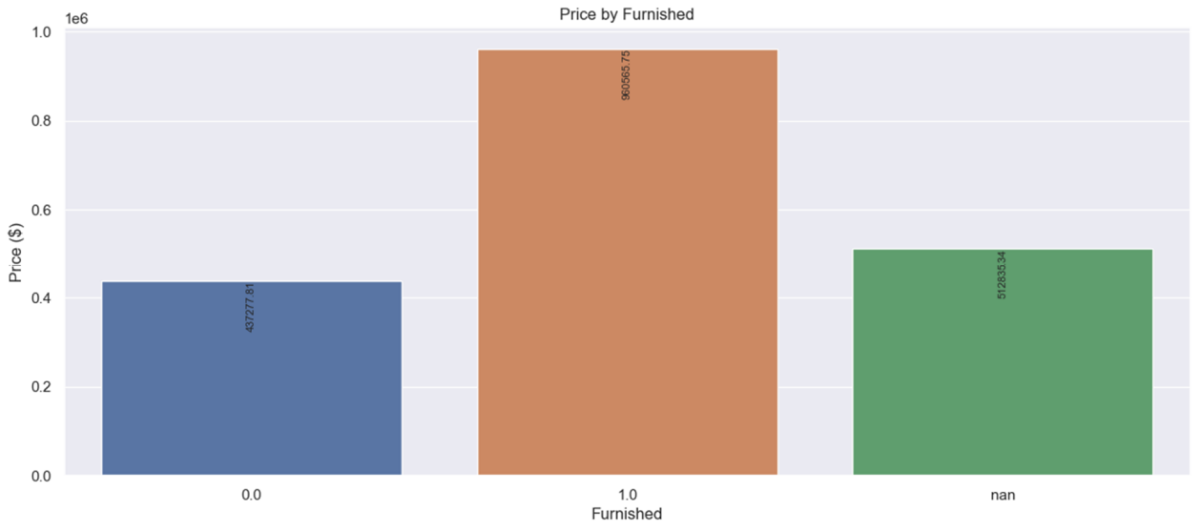


Figure No.23 Bar Plot between Price and Furnished

From the Figure No.23 Bar Plot between price and Furnished, it is clearly visible that price is high for furnished houses.

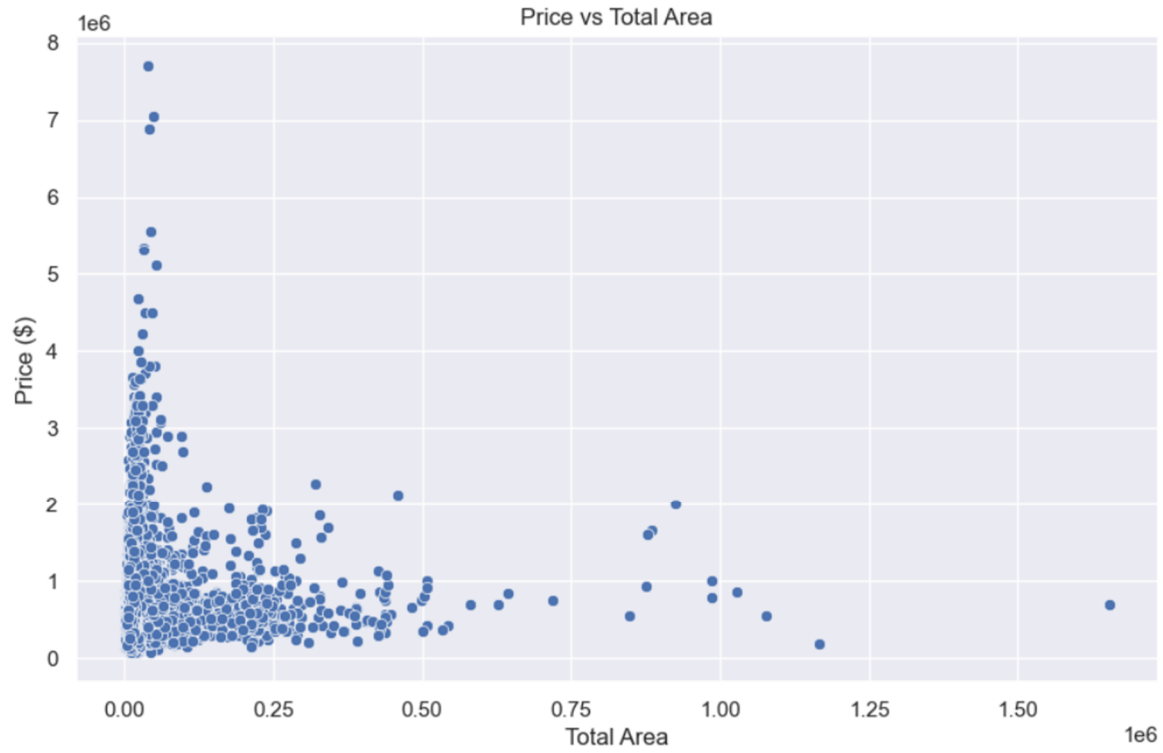


Figure No.24 Scatter Plot between Price and Total Area

From the Figure No.24 Scatter Plot between price and Total area of house, it seems that price is high for small houses compare to bigger ones. Further, it is also clear that this variable have extreme values.

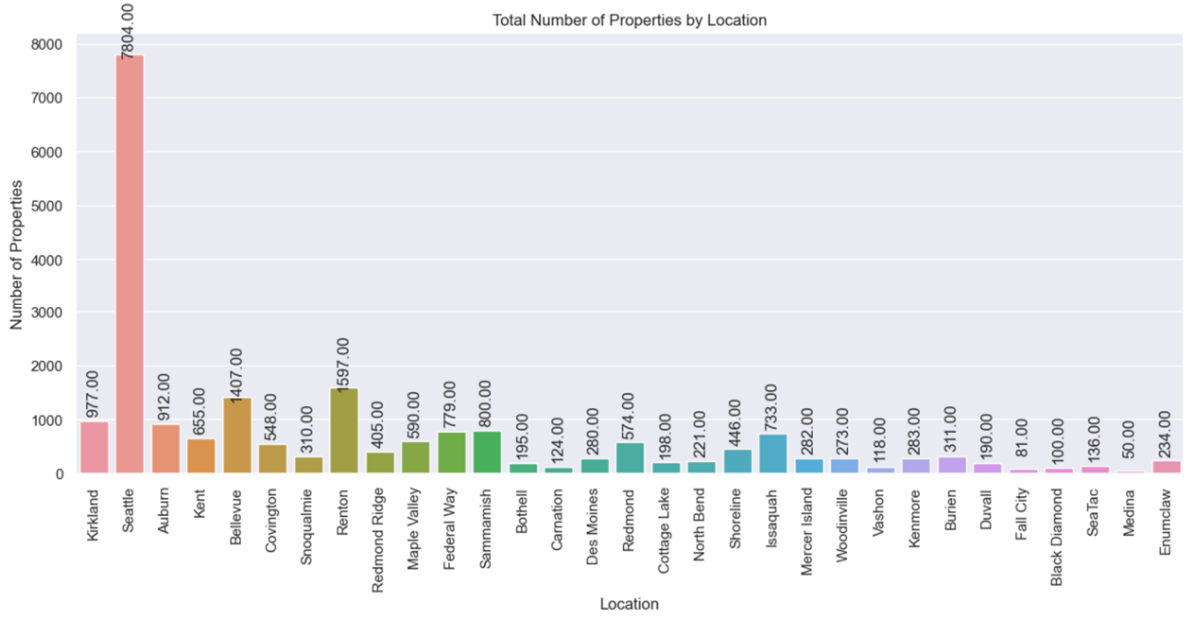


Figure No.25 Bar Plot between Number of Properties Location wise

From the Figure No.25 Scatter Plot between number of properties and location, it seems that maximum houses are in ‘Seattle’ location.

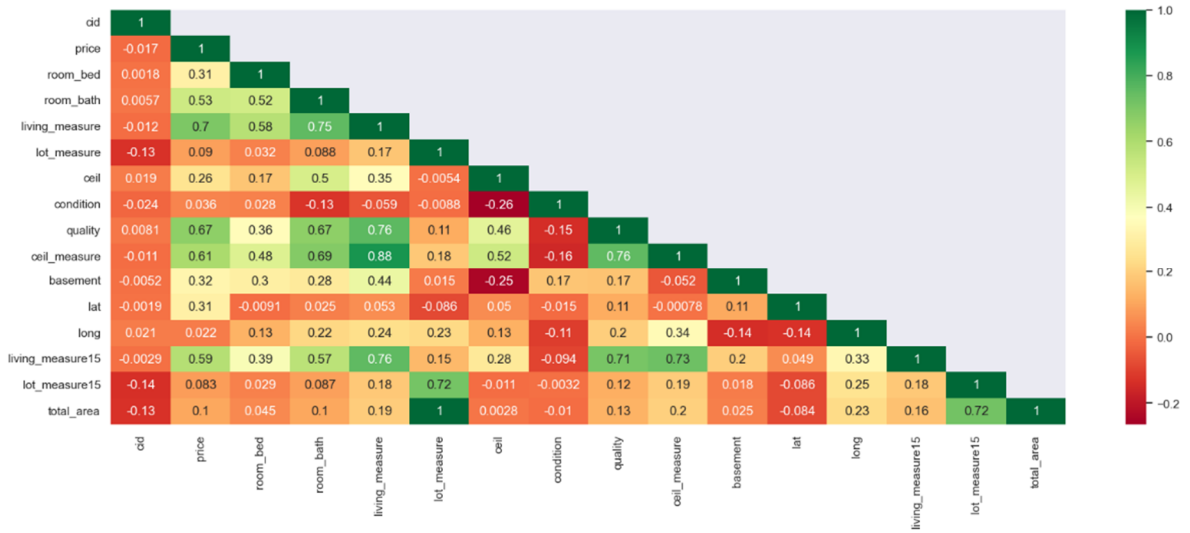


Figure No.26 Heatmap between continuous variables

From the Figure No.26 Heatmap for continuous variables, it seems that some variables like lot_measure with total_area, living_measure with ceil_measure and living_measure with living_measure15 are highly corelated. And correlation of lot_measure with total_area is one therefore dropping lot_measure is required for model building.

Removal of unwanted variables (if applicable)

Missing Value treatment (if applicable)

```
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
location           0
dtype: int64
```

Figure No.27 Null Values presence after treatment

From the Figure No.27 Null values presence after treatment, now after dropping null values it is clear from the above table that there are no values exist in dataset.

```
Column: total_area
count    2.128800e+04
mean     1.718824e+04
std      4.159631e+04
min      1.423000e+03
25%      7.037750e+03
50%      9.580000e+03
75%      1.299600e+04
max      1.652659e+06
Name: total_area, dtype: float64
```

```
Column: location
count       21288
unique         31
top       Seattle
freq          7677
Name: location, dtype: object
```

Figure No.28 Descriptive Analysis for "total_area" after removing null values and special character and Descriptive Analysis for "location"

From the Figure No.28 Descriptive Analysis for "total_area" after removing null values and special character, mean is 171882.4 and maximum value is 16526590.0 that means there are very much extreme values exist in this variable.

From the location descriptive analysis, it seems Seattle is top location with frequency of 7677 out of 21288.

	cid	dayhours	price	room_bed	room_bath	living_measure	lot_measure	ceil	coast	sight	condition	quality	ceil_measure	basement
3155	2954400190	2014-06-24	1300000	0.0	0.0	4810.0	28008.0	2.0	0.0	0.0	3.0	12.0	4810.0	0.0
3405	3980300371	2014-09-26	142000	0.0	0.0	290.0	20875.0	1.0	0.0	0.0	1.0	1.0	290.0	0.0
4238	7849202190	2014-12-23	235000	0.0	0.0	1470.0	4800.0	2.0	0.0	0.0	3.0	7.0	1470.0	0.0
15593	3918400017	2015-02-05	380000	0.0	0.0	1470.0	979.0	3.0	0.0	2.0	3.0	8.0	1470.0	0.0
17704	3374500520	2015-04-29	355000	0.0	0.0	2460.0	8049.0	2.0	0.0	0.0	3.0	8.0	2460.0	0.0
18596	9543000205	2015-04-13	139950	0.0	0.0	844.0	4269.0	1.0	0.0	0.0	4.0	7.0	844.0	0.0
20957	6306400140	2014-06-12	1100000	0.0	0.0	3064.0	4764.0	3.5	0.0	2.0	3.0	7.0	3064.0	0.0

Table No.6 Dataset for zero values in room_bed and room_bath variables

From Table No.6 Dataset for zero values in room_bed and room_bath, it is visible that in these two variable have zero values in few rows and it means how can be a house price predicted without number of rooms and number bathrooms present in it. Moreover, these are very few in counting therefore dropping these values for modelling is good as it is not going to affect over all dataset.

```
Number of rows have zero value in room_bed column (0, 24)

Number of rows have zero value in room_bath column (0, 24)
```

Figure No.29 Details for number of rows having zero value for variable room_bed and room_bath

From Figure No.29 it is clear that after dropping zero value rows from the room_bed and room_variable now there are no zero values exist.

Outlier treatment (if required)

	% OUTLIERS
lot_measure	11.18
total_area	11.16
lot_measure15	10.14
quality	8.87
price	5.37
ceil_measure	2.83
living_measure	2.65
room_bath	2.59
living_measure15	2.52
room_bed	2.44
basement	2.27
long	1.19
condition	0.13
lat	0.01
dayhours	0.00
ceil	0.00
cid	0.00

Figure No.30 Details for outlier percentage present in variables

From Figure No.30 Details of Outlier percentage present in variables, it seems that percentage of outliers in variables are quite small therefore outlier treatment is not required.

```
0.5% properties have a price lower than  135000.00
  1% properties have a price lower than  154413.00
  5% properties have a price lower than  210000.00
 10% properties have a price lower than  245000.00
 90% properties have a price lower than  887245.00
 95% properties have a price lower than  1160000.00
 99% properties have a price lower than  1970000.00
99.5% properties have a price lower than  2456450.00
```

Figure No.31 Distribution of price at different percentile

Figure No.31 Distribution of price at different percentile, it is clear that 10% properties price are between 135000 to 245000 but on and above 90% prices are changing drastically and 1% properties are in very higher price range.

	cid	dayhours	price	room_bed	room_bath	living_measure	lot_measure	ceil	coast	sight	condition	quality	ceil_measure	basement	yr_built	yr_renovated	zipcode	lat	long	living_measure15	lot_measure15	furnished	total_area	location
0	3876100940	2015-04-27	600000	4.0	1.75	3050.0	9440.0	1.0	0.0	0.0	3.0	8.0	1800.0	1250.0	1966.0	NO	98034	47.7228	-122.183	2020.0	8660.0	0.0	12490.0	Kirkland
1	3145600250	2015-03-17	190000	2.0	1.00	670.0	3101.0	1.0	0.0	0.0	4.0	6.0	670.0	0.0	1948.0	NO	98118	47.5546	-122.274	1660.0	4100.0	0.0	3771.0	Seattle
2	7129303070	2014-08-20	735000	4.0	2.75	3040.0	2415.0	2.0	1.0	4.0	3.0	8.0	3040.0	0.0	1966.0	NO	98118	47.5188	-122.256	2620.0	2433.0	0.0	5455.0	Seattle
3	7338220280	2014-10-10	257000	3.0	2.50	1740.0	3721.0	2.0	0.0	0.0	3.0	8.0	1740.0	0.0	2009.0	NO	98002	47.3363	-122.213	2030.0	3794.0	0.0	5461.0	Auburn
4	7950300670	2015-02-18	450000	2.0	1.00	1120.0	4590.0	1.0	0.0	0.0	3.0	7.0	1120.0	0.0	1924.0	NO	98118	47.5663	-122.285	1120.0	5100.0	0.0	5710.0	Seattle

Table No.7 Dataset after converting yr_renovated as binary type variable

From Table No.7 Dataset after converting yr_renovated as binary type variable, it is visible that houses are renovated or not after selling as it was not feasible to sort this variable only looking on given years for renovation and zero values.

Addition of new variables (if required)

	cid	dayhours	price	room_bed	room_bath	living_measure	lot_measure	ceil	coast	sight	condition	quality	ceil_measure	basement	yr_built	yr_renovated	zipcode	lat	long	living_measure15	lot_measure15	furnished	total_area	location	sold year
0	3876100940	2015-04-27	600000	4.0	1.75	3050.0	9440.0	1.0	0.0	0.0	3.0	8.0	1800.0	1250.0	1966.0	NO	98034	47.7228	-122.183	2020.0	8660.0	0.0	12490.0	Kirkland	2015
1	3145600250	2015-03-17	190000	2.0	1.00	670.0	3101.0	1.0	0.0	0.0	4.0	6.0	670.0	0.0	1948.0	NO	98118	47.5546	-122.274	1660.0	4100.0	0.0	3771.0	Seattle	2015
2	7129303070	2014-08-20	735000	4.0	2.75	3040.0	2415.0	2.0	1.0	4.0	3.0	8.0	3040.0	0.0	1966.0	NO	98118	47.5188	-122.256	2620.0	2433.0	0.0	5455.0	Seattle	2014
3	7338220280	2014-10-10	257000	3.0	2.50	1740.0	3721.0	2.0	0.0	0.0	3.0	8.0	1740.0	0.0	2009.0	NO	98002	47.3363	-122.213	2030.0	3794.0	0.0	5461.0	Auburn	2014
4	7950300670	2015-02-18	450000	2.0	1.00	1120.0	4590.0	1.0	0.0	0.0	3.0	7.0	1120.0	0.0	1924.0	NO	98118	47.5663	-122.285	1120.0	5100.0	0.0	5710.0	Seattle	2015

Table No.8 Dataset after adding new variable 'Sold Year'

dayhours	price	room_bed	room_bath	living_measure	lot_measure	ceil	coast	sight	condition	quality	ceil_measure	basement	yr_built	yr_renovated	zipcode	lat	long	living_measure15	lot_measure15	furnished	total_area	location	sold year	Age_of_House
2015-04-27	600000	4.0	1.75	3050.0	9440.0	1.0	0.0	0.0	3.0	8.0	1800.0	1250.0	1966.0	NO	98034	47.7228	-122.183	2020.0	8660.0	0.0	12490.0	Kirkland	2015	49.0
2015-03-17	190000	2.0	1.00	670.0	3101.0	1.0	0.0	0.0	4.0	6.0	670.0	0.0	1948.0	NO	98118	47.5546	-122.274	1660.0	4100.0	0.0	3771.0	Seattle	2015	67.0
2014-08-20	735000	4.0	2.75	3040.0	2415.0	2.0	1.0	4.0	3.0	8.0	3040.0	0.0	1966.0	NO	98118	47.5188	-122.256	2620.0	2433.0	0.0	5455.0	Seattle	2014	48.0
2014-10-10	257000	3.0	2.50	1740.0	3721.0	2.0	0.0	0.0	3.0	8.0	1740.0	0.0	2009.0	NO	98002	47.3363	-122.213	2030.0	3794.0	0.0	5461.0	Auburn	2014	5.0
2015-02-18	450000	2.0	1.00	1120.0	4590.0	1.0	0.0	0.0	3.0	7.0	1120.0	0.0	1924.0	NO	98118	47.5663	-122.285	1120.0	5100.0	0.0	5710.0	Seattle	2015	91.0

Table No.9 Dataset after adding new variable 'Age of House'

From Table No.8 Dataset after adding new variable 'Sold Year' and Table No.9 Dataset after adding new variable 'Age of House', it is clear that to know about the property's age adding these variable are required as age could be major factor for deciding the price of house.

```
Seattle          7673
Renton           1579
Bellevue         1383
Others           1171
Kirkland          962
Auburn            892
Sammamish         788
Federal Way       769
Issaquah          726
Kent              644
Maple Valley      582
Redmond           566
Covington         540
Shoreline         442
Redmond Ridge     391
Burien            305
Snoqualmie        303
Mercer Island     279
Kenmore           279
Des Moines        278
Woodinville       272
Enumclaw          230
North Bend        218
Name: location, dtype: int64
```

Figure No.32 Property distribution Location wise

From Figure No.32 Property distribution Location wise, it can see that 1171 number of properties are considered as Others as they are present in different locations wherein count for a single location is quite low like below 200.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 21272 entries, 0 to 21612
Data columns (total 11 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   price             21272 non-null  int64
 1   room_bed          21272 non-null  float64
 2   room_bath         21272 non-null  float64
 3   living_measure    21272 non-null  float64
 4   quality           21272 non-null  float64
 5   ceil_measure      21272 non-null  float64
 6   basement          21272 non-null  float64
 7   living_measure15  21272 non-null  float64
 8   lot_measure15     21272 non-null  float64
 9   total_area        21272 non-null  float64
 10  Age_of_House      21272 non-null  float64
dtypes: float64(10), int64(1)
memory usage: 1.9 MB
```

Table No.10 Data info after removing variables

In Table No.10 Data info after removing variables, here is info of varibales after removing object type columns for Clustering i.e., [cid, dayhours, lot_measure, yr_built, zipcode, lat, long, sold year and ceil, coast, sight, condition, yr_renovated, furnished, location]

```
3.0     9672
4.0     6783
2.0     2720
5.0     1577
6.0      266
1.0      193
7.0       38
8.0       12
9.0        6
10.0       3
33.0       1
11.0       1
Name: room_bed, dtype: int64
```

Figure No.33 Count for number of bedroom per house

From Figure No.33 Count for number of bedroom per house, it is clear that above 9 bedrooms per house there are just 4 values therefore, dropping these values for clustering is required otherwise it can affect the values in clusters.

```
(21272, 11)

(21267, 11)
```

Figure No.34 Shape of dataset for House have less than equal to 9 bedrooms per house

Figure No.34 Shape of dataset for House have less than equal to 9 bedrooms per house, it seems that shape (21272 - rows and 11 columns) of dataset before dropping the values on and above 9 bedrooms per house and shape (21267 - rows and 11 columns) of dataset after dropping the values on and above 9 bedrooms per house.

Business insights from EDA

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

The target variable 'price' in the given dataset not distributed uniformly, therefore we can say that the data is unbalance. Regression problems typically do not require rebalancing the data. This is because regression models aim to predict continuous target variables and are not affected by class distribution imbalance.

In the context of house price prediction, a regression model can handle imbalanced data without specific treatments, since the dataset represents a diverse range of house characteristics and prices to obtain reliable and accurate predictions.

Clustering

	price	room_bed	room_bath	living_measure	quality	ceil_measure	basement	living_measure15	lot_measure15	total_area	Age_of_House
0	0.163955	0.700442	-0.476121	1.057304	0.290345	0.013211	2.167309	0.047778	-0.150316	-0.112821	0.194333
1	-0.957441	-1.524546	-1.453721	-1.537620	-1.413039	-1.352771	-0.658227	-0.477271	-0.317796	-0.322450	0.806825
2	0.533196	0.700442	0.827347	1.046401	0.290345	1.512165	-0.658227	0.922859	-0.379022	-0.281962	0.160306
3	-0.774188	-0.412052	0.501480	-0.370995	0.290345	-0.059319	-0.658227	0.062362	-0.329035	-0.281818	-1.302867
4	-0.246312	-1.524546	-1.453721	-1.046983	-0.561347	-0.808796	-0.658227	-1.264844	-0.281068	-0.275831	1.623479

Table No.11 Scaled Dataset

From Table No.11 Scaled Dataset, it is clear that dataset is scaled because all values in the variables are not at same scale which is required for clustering process.

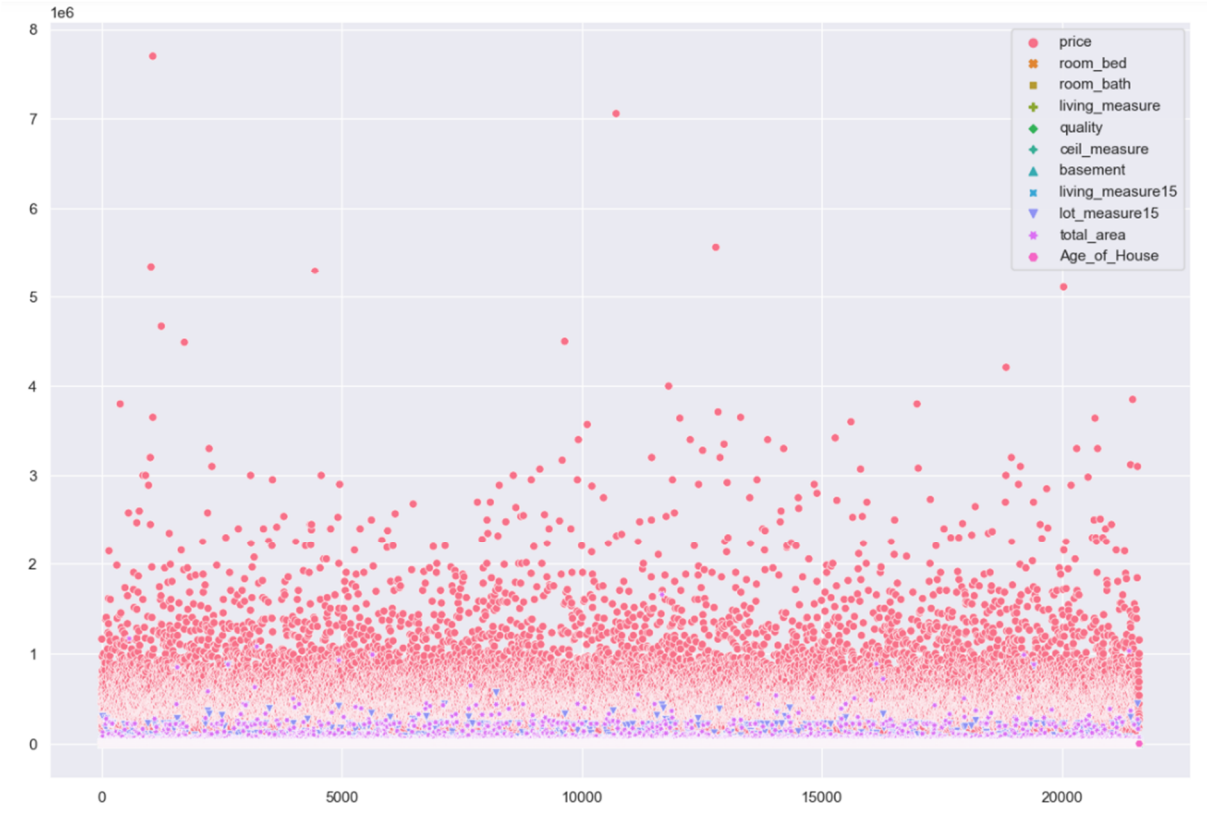


Figure No.35 Scatter Plot of dataset after removing object type variable

From Figure No.35 Scatter Plot of dataset after removing object type variable, distribution of dataset is all over and extreme values are visible clearly.

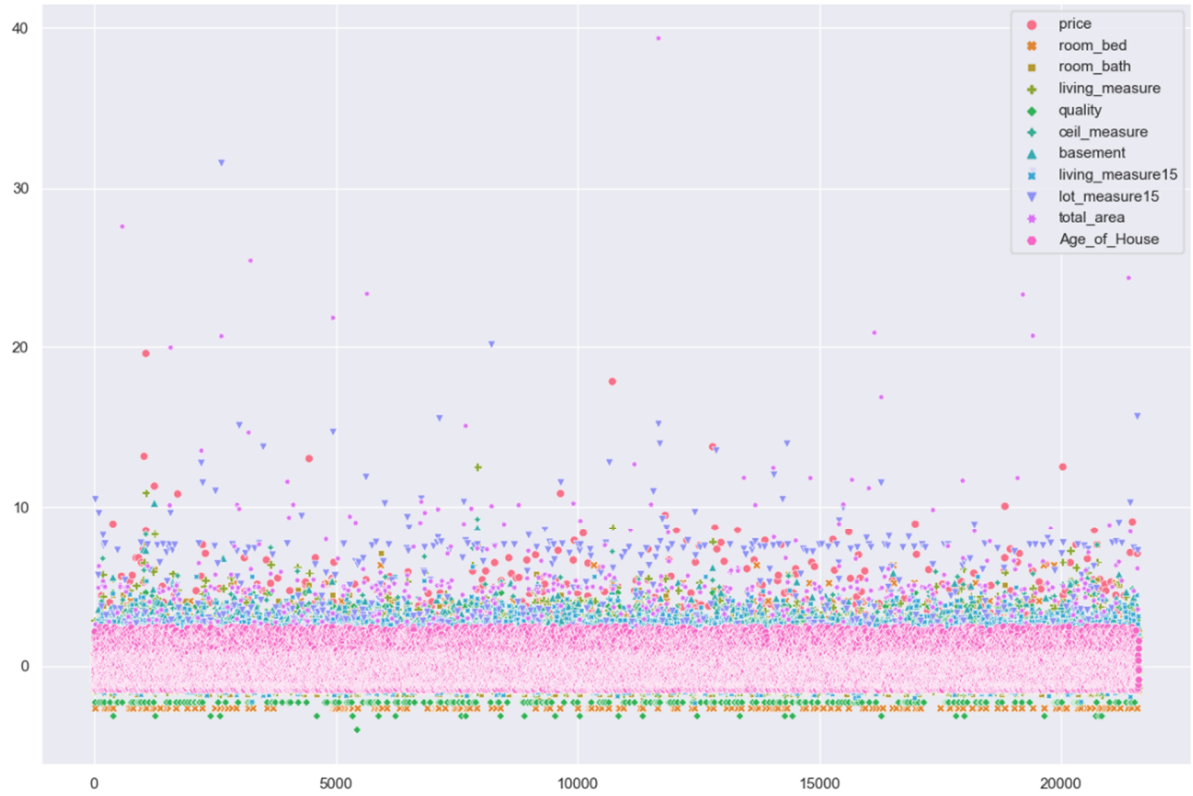


Figure No.36 Scatter Plot of Scaled dataset

From Figure No.36 Scatter Plot of Scaled dataset, data distribution is over all and extreme values have no effect of scaling.

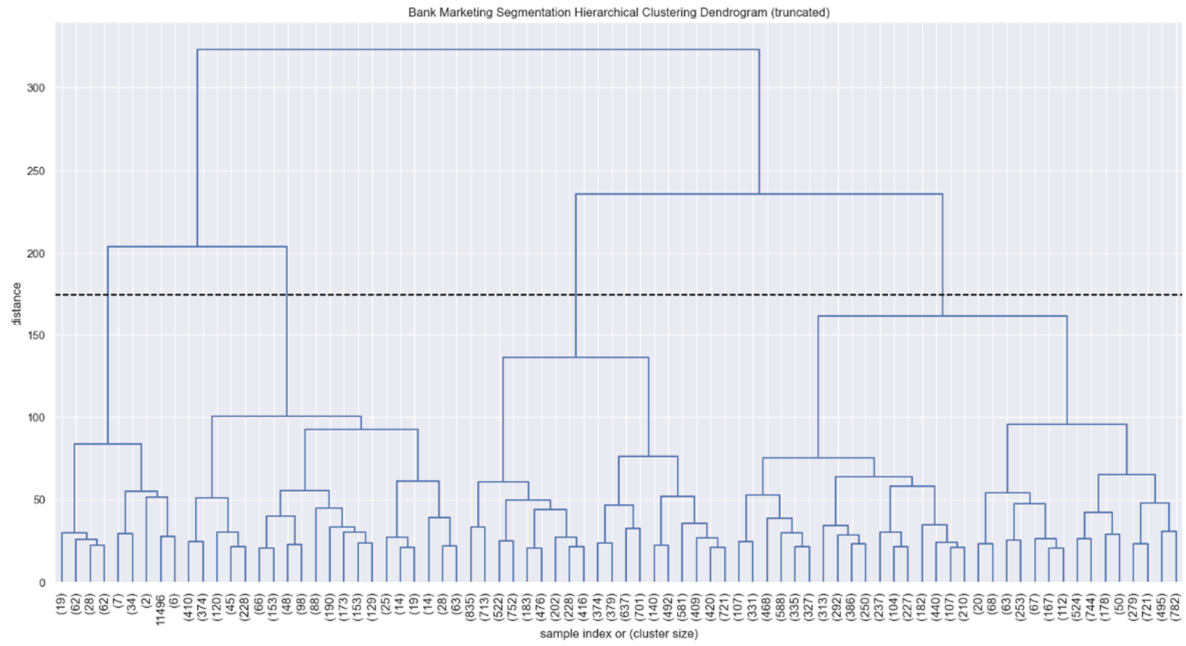


Figure No.37 Dendrogram of Hierarchical Clustering

	price	room_bed	room_bath	living_measure	ceil	coast	sight	condition	quality	ceil_measure	basement	yr_renovated	living_measure15	lot_measure15	furnished	total_area	location	Age_of_House	cluster_1
0	600000	4.0	1.75	3050.0	1.0	0.0	0.0	3.0	8.0	1800.0	1250.0	NO	2020.0	8660.0	0.0	12490.0	Kirkland	49.0	4
1	190000	2.0	1.00	670.0	1.0	0.0	0.0	4.0	6.0	670.0	0.0	NO	1660.0	4100.0	0.0	3771.0	Seattle	67.0	3
2	735000	4.0	2.75	3040.0	2.0	1.0	4.0	3.0	8.0	3040.0	0.0	NO	2620.0	2433.0	0.0	5455.0	Seattle	48.0	4
3	257000	3.0	2.50	1740.0	2.0	0.0	0.0	3.0	8.0	1740.0	0.0	NO	2030.0	3794.0	0.0	5461.0	Auburn	5.0	4
4	450000	2.0	1.00	1120.0	1.0	0.0	0.0	3.0	7.0	1120.0	0.0	NO	1120.0	5100.0	0.0	5710.0	Seattle	91.0	3

Table No.12 Dataset for Number of Clusters

cluster_1	price	room_bed	room_bath	living_measure	quality	ceil_measure	basement	living_measure15	lot_measure15	total_area	Age_of_House	cluster count
1	6.047260e+05	3.262443	2.343891	2618.597285	8.049774	2416.266968	202.330317	2342.809955	220259.583710	272410.022624	31.497738	221
2	1.181114e+06	4.108285	3.140792	3751.379820	9.751846	3212.150123	539.229696	3135.079573	16491.632075	23671.781788	27.194011	2438
3	3.695652e+05	2.797517	1.600588	1378.928330	6.926370	1273.590350	105.337981	1530.984751	6960.728025	8692.708202	51.724322	9181
4	5.387905e+05	3.740002	2.345948	2318.502705	7.822319	1908.362151	410.140554	2126.403204	12561.830593	17789.194441	39.512464	9427

Table No.13 Dataset for Number of Clusters with count of properties per cluster

From the above Table No.13 Dataset for Number of Clusters with count of properties per cluster, it is clear that 9427 number of properties are in Cluster 4 and 9181 properties in Cluster 3. But price is lower for properties which fall in Cluster 3 compare Cluster 4 which is 1.5 times costlier.

Figure No.38 Scatter Plot for distribution of Clusters between Price & Living Measure

From Figure No.41 Scatter Plot for distribution of Clusters between Price & Living Measure, Cluster 2 have higher the living area and price.

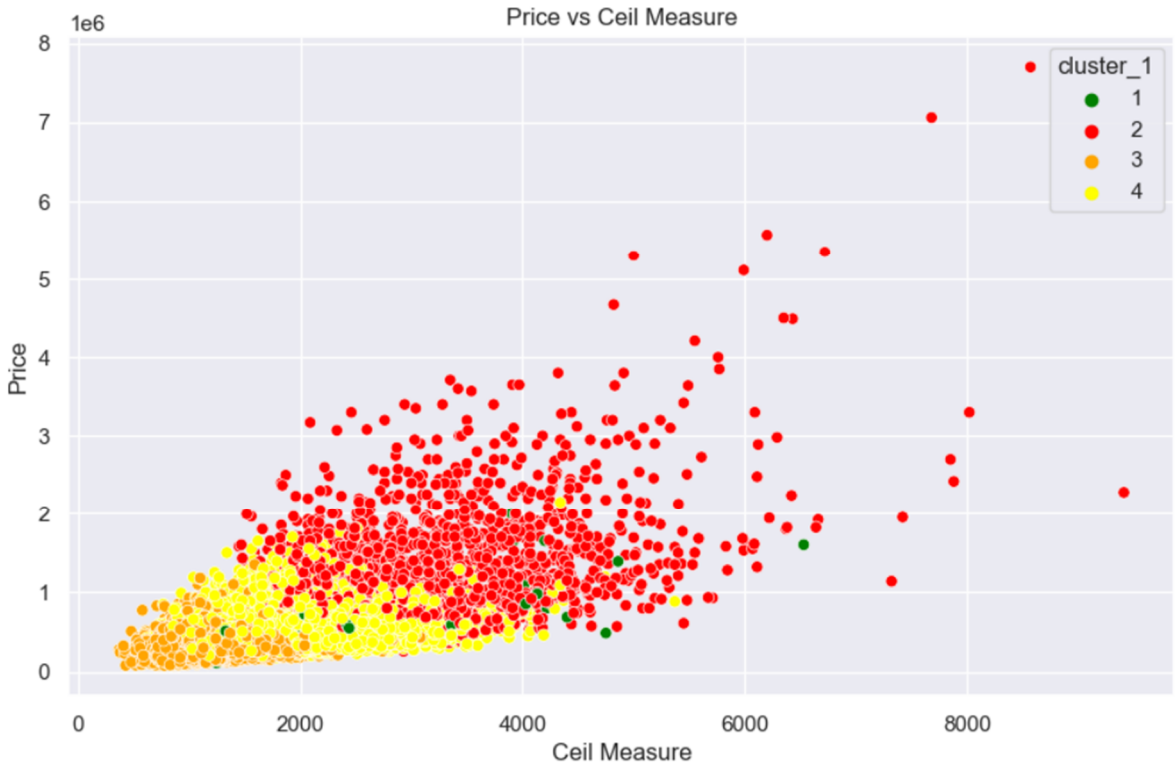


Figure No.39 Scatter Plot for distribution of Clusters between Price & Ceil Measure

From Figure No.42 Scatter Plot for distribution of Clusters between Price & Ceil Measure, it seems Cluster 2 have higher price with the increase in ceil measure size.

Figure No.40 Scatter Plot for distribution of Clusters between Price & Total Measure

From Figure No.43 Scatter Plot for distribution of Clusters between Price & Total Measure, it is clear the price for cluster 2 properties are higher than the price of properties in cluster 4. Although total area for cluster 2 properties / house are small and properties in cluster 4 bigger in size but price is not high.

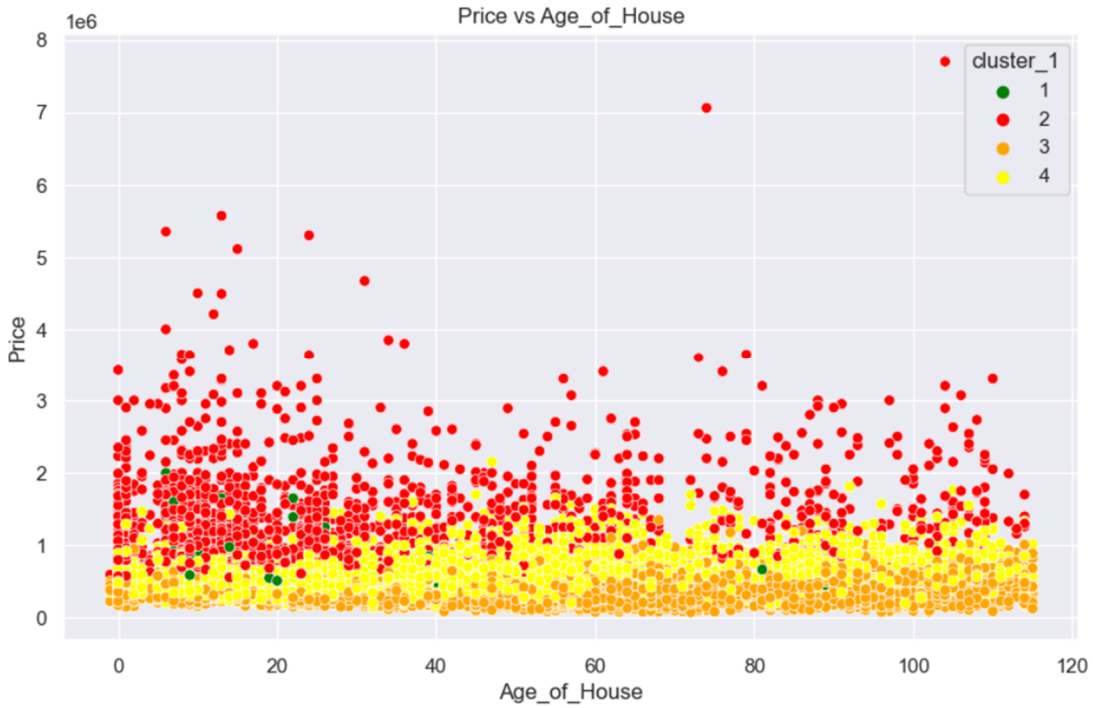


Figure No.41 Scatter Plot for distribution of Clusters between Price & Age of House

From Figure No.44 Scatter Plot for distribution of Clusters between Price & Age of House, it is visible that price of house for every age distributed all over but within age of 0 to 40 years major properties have higher price and here again Cluster 4 is distributed all over in the number of age with lower or medium prices.

Figure No.42 Map of the location of the houses.

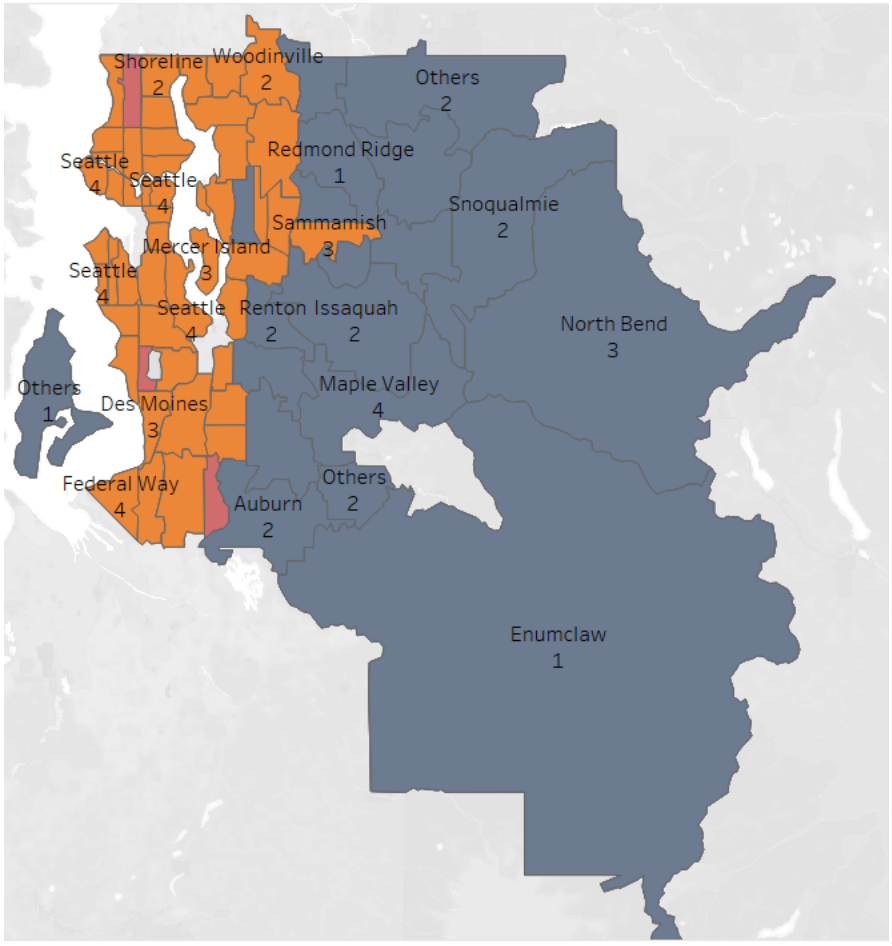


Figure No.43 Map of the location of the houses after clustering.

Business insights From Clustering

1. More sale price is belong to the cluster 1, where very less number of houses are there, this indicates that the properties belong to the cluster 1 is having very high price value and for a common man it is very difficult to afford.
2. The houses belong to cluster 1 have a very high total area, and the houses belong to this area is comparatively new compared to other clusters.
3. In the case of cluster 2, the houses belong to this cluster is having an average number of room of 4, the sale price of this cluster is very low, which indicates people less interested in a house with an average number of room of 4 even though it has a high quality ranking.
4. After cluster 1 more sale is happening is cluster 4 and most number of house belong to this cluster and the age of house is also comparatively less compared to other clusters, so we can conclude that this area is favourable for middle class people with affordable price.
5. The houses belong to cluster 3 has a very high average age of house, most of the house belong to this cluster has an average number of room of 2. The average total area of the houses belong to cluster 3 is very low, even though good sale is happening in this area which indicates the low price of the houses in this area.
6. Most of the houses in the area Seattle, but more sale is happening in Medina. This indicates the reduced living cost of the Seattle area and rich society present in the Medina area.
7. 2 bed room houses have more demands, indicates the increase of nuclear families, so if we concentrate more on these people we can increase the sale.

Model building and interpretation.

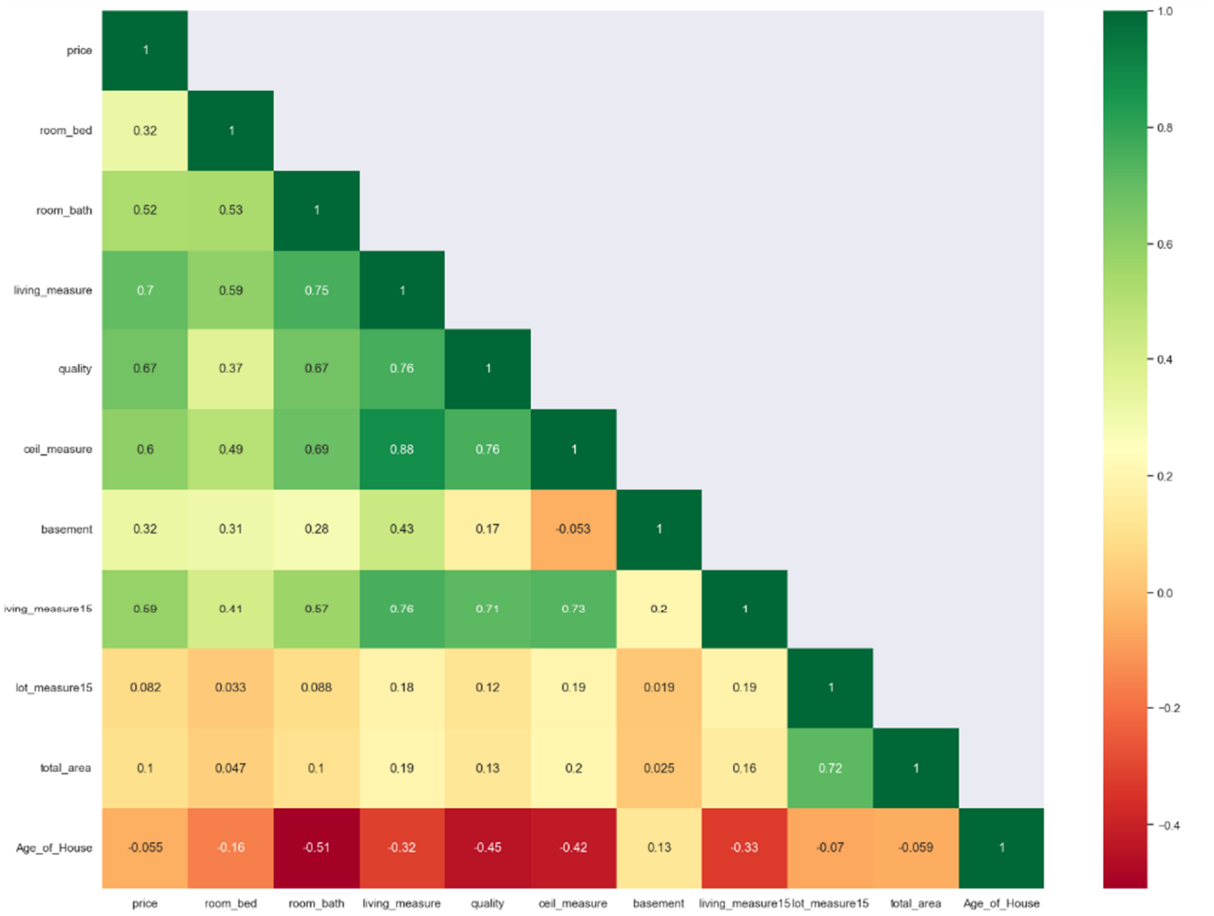


Figure No.44 Heatmap of dataset after removing few variables

From Figure No.45 Heatmap of dataset after removing few variables, it is clear that after dropping lots of variable still high correlation exist between living_measure - room_bath, living_meausre - price, living_measure – quality, living_measure - ceil_measure and living_meaure – living_measure15, so we can remove the column living_measure for model building purpose as it will affect the prediction.

	price	room_bed	room_bath	quality	ceil_measure	basement	living_measure15	lot_measure15	total_area	Age_of_House	ceil_1.5	ceil_2.0	ceil_2.5
0	600000	4.0	1.75	8.0	1800.0	1250.0	2020.0	8660.0	12490.0	49.0	0	0	0
1	190000	2.0	1.00	6.0	670.0	0.0	1660.0	4100.0	3771.0	67.0	0	0	0
2	735000	4.0	2.75	8.0	3040.0	0.0	2620.0	2433.0	5455.0	48.0	0	1	0
3	257000	3.0	2.50	8.0	1740.0	0.0	2030.0	3794.0	5461.0	5.0	0	1	0
4	450000	2.0	1.00	7.0	1120.0	0.0	1120.0	5100.0	5710.0	91.0	0	0	0

ceil_3.0	ceil_3.5	coast_1.0	sight_1.0	sight_2.0	sight_3.0	sight_4.0	condition_2.0	condition_3.0	condition_4.0	condition_5.0	furnished_1.0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0
0	0	1	0	0	0	1	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0

location_Bellevue	location_Burien	location_Covington	location_Des Moines	location_Enumclaw	location_Federal Way	location_Issaquah	location_Kenmore
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

location_Kirkland	location_Maple Valley	location_Mercer Island	location_North Bend	location_Others	location_Redmond	location_Redmond Ridge	location_Renton
1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

location_Sammamish	location_Seattle	location_Shoreline	location_Snoqualmie	location_Woodinville
0	0	0	0	0
0	1	0	0	0
0	1	0	0	0
0	0	0	0	0
0	1	0	0	0

Table No.14 Dataset after adding dummy variable

Type of Measures	MAE		MSE		R-Square		MAPE	
Model Name	Train	Test	Train	Test	Train	Test	Train	Test
Linear Regression	115876.0	120499.1	34,17,13,20,314.07	40,93,53,42,946.28	0.736	0.715	0.24	0.25
Grid Search Cross Validation for Linera Regression	142627.5	146479.4	46,25,29,57,741.51	53,68,77,39,945.38	0.642	0.627	0.29	0.29
Decission Tree Regression	120.5	128788.6	85,06,076.96	54,49,11,67,341.27	1.000	0.621	0.00	0.24
Grid Search Cross Validation for Decision Tree Regression	58619.0	119567.3	10,99,18,31,953.28	44,07,58,72,063.38	0.915	0.693	0.11	0.23
Random Forest	32884.3	94664.5	3,45,46,10,209.81	31,25,93,75,373.45	0.973	0.783	0.06	0.07
Grid Search Cross Validation for Random Forest	33218.2	94661.7	3,39,14,60,361.15	31,39,65,43,205.80	0.974	0.782	0.07	0.18
Gradient Boosting	95083.6	106493.3	20,48,45,53,528.35	31,90,36,57,381.66	0.842	0.778	0.20	0.21
Grid Search Cross Validation for Gradient Boosting	55690.2	90283.4	6,43,17,36,169.23	26,74,10,81,463.41	0.950	0.814	0.12	0.17
Ada Boost	331893.8	337444.2	1,35,04,34,53,336.56	1,44,95,95,02,238.00	0.044	0.008	0.92	0.93
Grid Search Cross Validation for Ada Boost	159398.9	168179.5	49,72,31,58,215.10	62,45,15,64,083.72	0.616	0.566	0.36	0.37
Bagging	37355.6	99371.6	4,96,39,84,721.84	34,36,20,06,479.05	0.962	0.761	0.07	0.19
Grid Search Cross Validation for Bagging	44949.1	94922.1	6,30,36,83,234.85	31,25,54,60,951.02	0.951	0.783	0.09	0.18
Voting Regression	68908.8	109369.8	13,81,45,60,134.25	38,96,59,68,616.73	0.893	0.729	0.13	0.21
Grid Search Cross Validation for Voting Regression	85252.3	109227.8	25,66,50,60,626.84	44,61,78,30,432.61	0.802	0.690	0.16	0.20
XGB Regression	57824.1	93472.6	6,83,18,38,902.58	28,65,51,18,993.02	0.947	0.801	0.13	0.18
Grid Search Cross Validation for XGB Regression	76139.2	94634.7	12,79,63,50,641.04	27,63,56,48,960.60	0.901	0.808	0.16	0.18

Type of Measures	MAE		MSE		R-Square		MAPE	
Model Name	Train	Test	Train	Test	Train	Test	Train	Test
KNN Regression	130658.3	163523.0	44,77,60,01,455.39	76,44,20,76,050.48	0.654	0.468	0.26	0.32
Grid Search Cross Validation for KNN Regression	120.5	156862.1	85,06,076.96	73,73,42,81,878.87	1.000	0.487	0.00	0.31
LASSO Regression	115876.3	120499.8	34,17,13,21,338.64	40,93,54,56,340.18	0.736	0.715	0.24	0.25
Grid Search Cross Validation for LASSO Regression	115848.9	120509.6	34,17,82,32,698.76	40,96,66,01,463.33	0.736	0.715	0.24	0.25
Ridge Regression	115887.3	120532.6	34,17,28,88,715.54	40,94,53,99,579.78	0.736	0.715	0.24	0.25
Grid Search Cross Validation for Ridge Regression	115887.3	120532.6	34,17,28,88,715.54	40,94,53,99,579.78	0.736	0.715	0.24	0.25
Stacking Regression	69644.9	94958.5	9,94,49,49,851.06	27,71,94,12,215.90	0.923	0.807	0.15	0.18
Stacking Regression Meta Model Gradeint Boosting	69644.9	94958.5	9,94,49,49,851.06	27,71,94,12,215.90	0.923	0.807	0.15	0.18
Stacking Regression Meta Model XGB	69644.9	94958.5	9,94,49,49,851.06	27,71,94,12,215.90	0.923	0.807	0.15	0.18
Stacking Regression Final Estimator Gradeint Boosting	65716.6	90506.7	10,73,81,60,305.87	25,68,09,65,764.13	0.917	0.821	0.14	0.17
Stacking Regression Final Estimator XGB	68346.6	93166.7	11,60,57,63,633.32	28,00,75,63,939.45	0.910	0.805	0.14	0.18
Stacking Regression Final Estimator Linear Regression	66468.4	91716.6	9,28,71,56,797.28	26,87,46,07,460.04	0.928	0.813	0.14	0.18

Table No.15 Model Performance Matrix

From the above table we can conclude that the XG boost Regression model is comparatively better model compared to the other models, the difference between the R-square value of XGboost regression is less in train and test model before and after the grid search cross validation. The value of MAPE also less compared to other models in the case of XGB regression. So we can select XGB regression as the final model.

The parameter used in the case of grid search of the XGB regression are n_estimators, learning rate and max depth.

Let us check the feature importance graph of the XGB regression before and after grid search cross validation for further study,

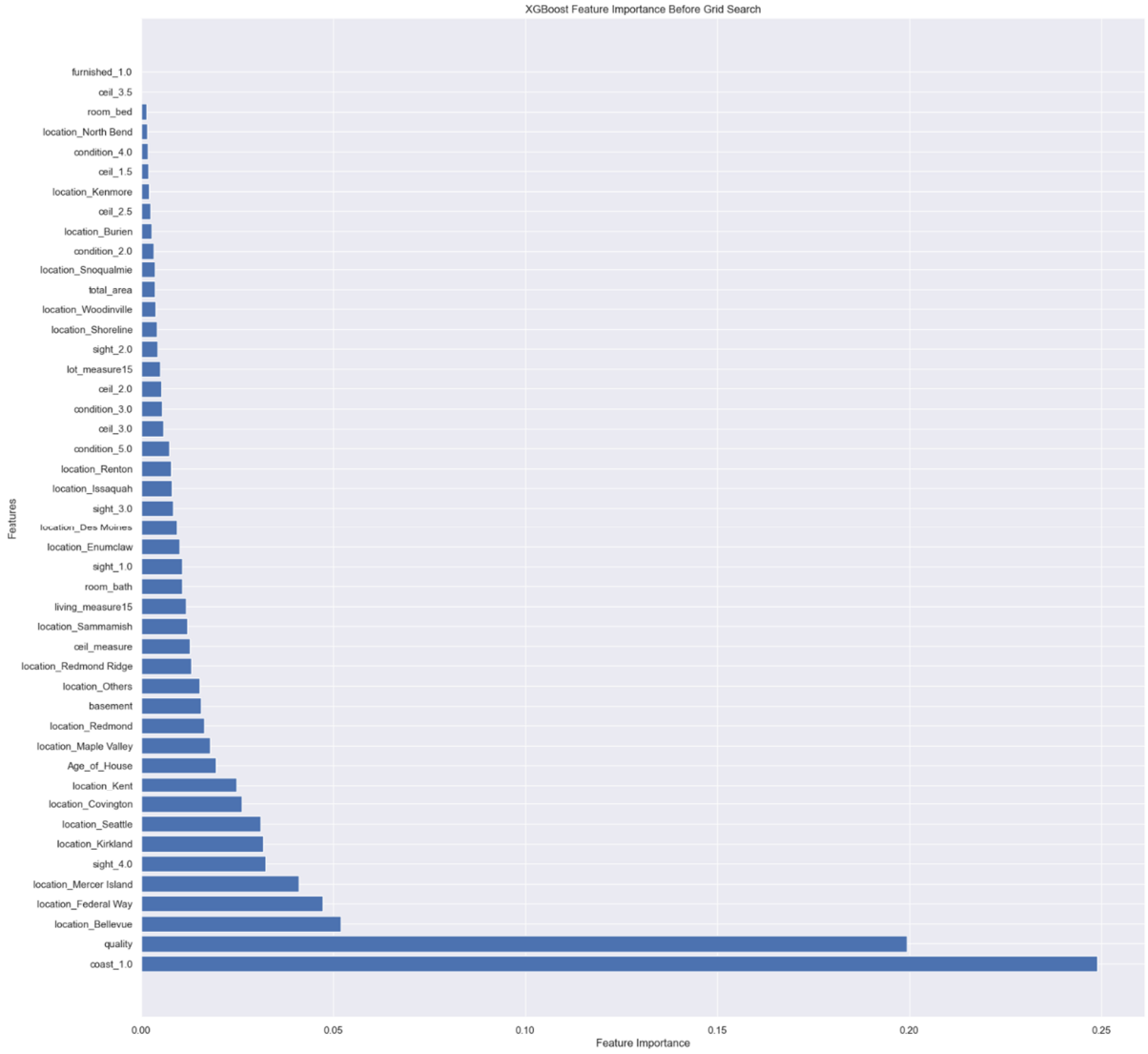


Figure No.45 Bar Plot of Feature Importance for XGBoost before Grid Search Cross Validation

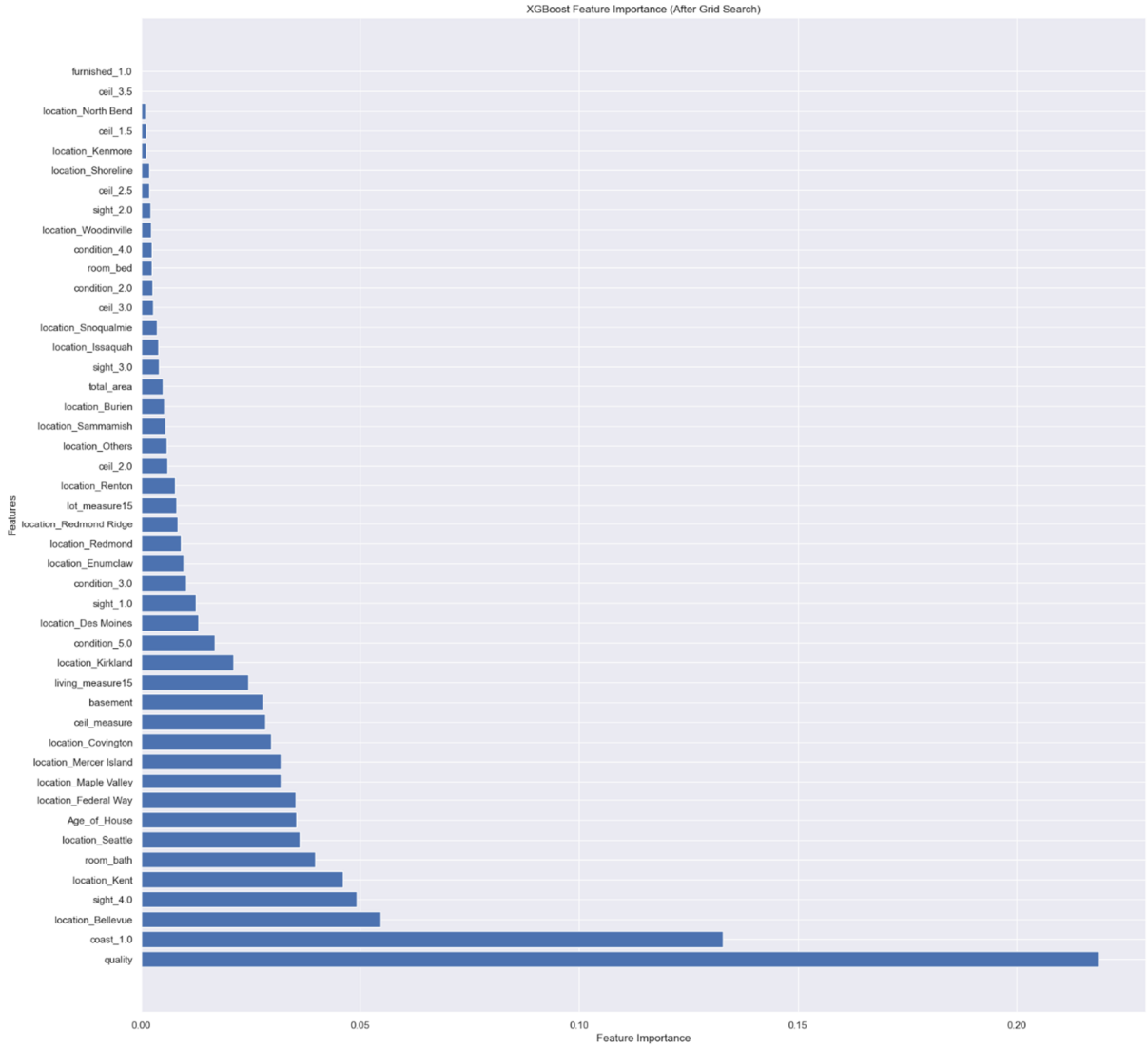


Figure No.46 Bar Plot of Feature Importance for XG Boost after Grid Search Cross Validation

From the above two bar plots for Feature Importance for XG Boost before and after Cross Validation it seems feature like "quality" is most important feature for deciding the price which is straight away affecting the valuation of a house with some other features like "coast_1.0, room_bath, Age_of_House, location_seattle and location_Bellevue, sight_4.0".

Conclusion:

- From the Model Building it is clear that the XG Boost Model is best performed model although other models are also quite well performed in symmetry that indicates the considering the features and removing the features for model building were well selected.
- Now, if we talk on the business front the coastal areas are most preferred location for the population of west US therefore builders can offer new build houses on coastal areas after considering other important features like number of rooms, number of bathrooms and number of floors.

- Persons who wants to purchase house on low budget can prefer Seattle location as we come to know this through clustering, house prices are low at this location compare to other location.
- Moreover, the data set provided is completely a random data just for study purpose and it needs to add more variable to identify the exact idea of price of a house or property.
- Now, the model is based on primarily prices of houses available historically. With the analysis of the clusters, it appears that there are more logical causes applicable that have not been provided in the problem statement - such as political, economical, social, technological, environmental and legal criteria. There's scope of improving the model by first classifying the properties into the PESTEL classes and evolving individual pricing models for each classification.
- The model is developed to provide a price estimation within an acceptable range of variance. However, the problem statement does not divulge the motive for price estimation and the how it is to be utilised. For example, if the pricing estimation is done for a property developer, the larger objective would be to improve the profitability of the business. It would involve utilising other parameters such as cost, taxes and average time to make sale. This would involve developing a multi objective function. The current price estimation model is purely an academic exercise as per our understanding.

……End……..

Click here to go to contents

Appenndix

The features of the dataset are,
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

For python codes and other results please find the attached jupyter notebook