CAPSTONE PROJECT

2022

GREAT LEARNING

Name: Rajat Prakash Singh

Date: 05th JUNE 2022

Table of Contents

Contents

LIST OF FIGURES

Notebook: CapstoneProject_Rajat_prakash_Singh_5th_june_2022.ipynb

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 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.

Problem Definition:

When any person/business wants to sell or buy a house, they always face this kind of issue as they don't know the price which they should offer. Due to this they might be offering too low or high for the property. Therefore, we can analyze the available data of the properties in the area and can predict the price. We need to find how these attributes influence the house prices Right pricing is very important aspect to sell house. It is very important to understand what are the factors and how they influence the house price. Objective is to predict the right price of the house based on the attributes.

Objective:

- Build model which will predict the house price when required features passed to the model. So, we will
- Find out the significant features from the given features dataset which affects the house price the most.
- Build best feasible model to predict the house price with 95% confidence level

Business Reason:

As people don't know the features/aspects which cumulate property price, we can provide them House Buying Selling guiding services in the area so they can buy or sell their property with most suitable price tag and they didn't lose their hard-earned money by offering low price or keep waiting for buyers by putting high prices.

Data Understanding:

First, we load the data from the given excel provide part of the study. After reading the data.

	cid	dayhours	price	room_bed	room_bath	living_measure	lot_measure	ceil	coast	sight
0	3.88E+09	20150427T000000	600000	4	1.75	3050	9440	1	0	0
1	3.15E+09	20150317T000000	190000	2	1	670	3101	1	0	0
2	7.13E+09	20140820T000000	735000	4	2.75	3040	2415	2	1	4
3	7.34E+09	20141010T000000	257000	3	2.5	1740	3721	2	0	0
4	7.95E+09	20150218T000000	450000	2	1	1120	4590	1	0	0

basement	yr_built	yr_renovated	zipcode	lat	long	living_measure15	lot_measure15	furnished	total_area
1250	1966	0	98034	47.7228	-122.183	2020	8660	0	12490
0	1948	0	98118	47.5546	-122.274	1660	4100	0	3771
0	1966	0	98118	47.5188	-122.256	2620	2433	0	5455
0	2009	0	98002	47.3363	-122.213	2030	3794	0	5461
0	1924	0	98118	47.5663	-122.285	1120	5100	0	5710

- Data is loaded successfully as we can see first 5 records from the dataset.
- We have more than 21k records having 23 features.

From the above we can see the different columns we have in dataset.

These columns provide below information

- cid: Notation for a house. Will not of our use. So we will drop this column
- dayhours: Represents Date, when house was sold.
- price: It's our TARGET feature, that we have to predict based on other featues
- room_bed: Represents number of bedrooms in a house
- room_bath: Represents number of bathrooms
- living_measure: Represents square footage of house
- lot_measure: Represents square footage of lot
- ceil: Represents number of floors in house
- coast: Represents whether house has waterfront view. It seems to be a categorical variable. We will see in our further data analysis
- sight: Represents how many times sight has been viewed.
- condition: Represents the overall condition of the house. It's kind of rating given to the house.
- quality: Represents grade given to the house based on grading system
- ceil_measure: Represents square footage of house apart from basement
- basement: Represents square footage of basement
- yr_built: Represents the year when house was built
- yr_renovated: Represents the year when house was last renovated
- zipcode: Represents zipcode as name implies
- lat: Represents Lattitude co-ordniates
- long: Represents Longitude co-ordinates
- living_measure15: Represents square footage of house, when measured in 2015 year as house area may or may not changed after renovation if any happened
- lot_measure15: Represents square footage of lot, when measured in 2015 year as lot area may or may not change after renovation if any done
- furnished: Tells whether house is furnished or not. It seems to be categorical variable as description implies
- total_area: Represents total area i.e. area of both living and lot

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	cid	21613 non-null	float64
1	dayhours	21613 non-null	object
2	price	21613 non-null	float64
3	room_bed	21613 non-null	float64
4	room_bath	21613 non-null	float64
5	living_measure	21613 non-null	float64
6	lot_measure	21613 non-null	float64
7	ceil	21613 non-null	object
8	coast	21613 non-null	float64
9	sight	21613 non-null	float64
10	condition	21613 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	object
15	yr_renovated	21613	non-null	float64
16	zipcode	21613	non-null	float64
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

In the dataset, we have more than 21k records and 23 columns, out of which

- 18 features are of float type
- 5 feature is of object type (we may need to convert this object type to specific datatype)

	count	mean	std	min	25%	50%	75%	max
cid	21613	4.58E+09	2.88E+09	1.00E+06	2.12E+09	3.90E+09	7.31E+09	9.90E+09
price	21613	5.40E+05	3.67E+05	7.50E+04	3.22E+05	4.50E+05	6.45E+05	7.70E+06
room_bed	21613	3.37E+00	9.28E-01	0.00E+00	3.00E+00	3.00E+00	4.00E+00	3.30E+01
room_bath	21613	2.12E+00	7.69E-01	0.00E+00	1.75E+00	2.25E+00	2.50E+00	8.00E+00
living_measure	21613	2.08E+03	9.18E+02	2.90E+02	1.42E+03	1.91E+03	2.55E+03	1.35E+04
lot_measure	21613	1.51E+04	4.14E+04	5.20E+02	5.03E+03	7.61E+03	1.07E+04	1.65E+06
ceil	21613	1.49E+00	5.40E-01	1.00E+00	1.00E+00	1.50E+00	2.00E+00	3.50E+00
coast	21613	7.45E-03	8.60E-02	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.00E+00
sight	21613	2.34E-01	7.66E-01	0.00E+00	0.00E+00	0.00E+00	0.00E+00	4.00E+00
condition	21613	3.41E+00	6.50E-01	1.00E+00	3.00E+00	3.00E+00	4.00E+00	5.00E+00
quality	21613	7.66E+00	1.18E+00	1.00E+00	7.00E+00	7.00E+00	8.00E+00	1.30E+01
ceil_measure	21613	1.79E+03	8.28E+02	2.90E+02	1.19E+03	1.56E+03	2.21E+03	9.41E+03
basement	21613	2.92E+02	4.43E+02	0.00E+00	0.00E+00	0.00E+00	5.60E+02	4.82E+03
yr_built	21613	1.97E+03	2.94E+01	1.90E+03	1.95E+03	1.98E+03	2.00E+03	2.02E+03

yr_renovated	21613	8.44E+01	4.02E+02	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.02E+03
zipcode	21613	9.81E+04	5.35E+01	9.80E+04	9.80E+04	9.81E+04	9.81E+04	9.82E+04
lat	21613	4.76E+01	1.39E-01	4.72E+01	4.75E+01	4.76E+01	4.77E+01	4.78E+01
long	21613	-1.22E+02	1.41E-01	-1.23E+02	-1.22E+02	-1.22E+02	-1.22E+02	-1.21E+02
living_measure15	21613	1.98E+03	6.84E+02	3.99E+02	1.49E+03	1.83E+03	2.36E+03	6.21E+03
lot_measure15	21613	1.28E+04	2.73E+04	6.51E+02	5.10E+03	7.62E+03	1.01E+04	8.71E+05
furnished	21613	1.96E-01	3.97E-01	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.00E+00
total_area	21613	1.72E+04	4.16E+04	1.42E+03	7.03E+03	9.56E+03	1.30E+04	1.65E+06

- CID: House ID/Property ID.Not used for analysis
- Dayhours: 5 factor analysis is reflecting for this column
- price: Our taget column value is in 75k - 7700k range. As Mean > Median, it's Right-Skewed.
- room_bed: Number of bedrooms range from 0 - 33. As Mean slightly > Median, it's slightly Right-Skewed.
- room_bath: Number of bathrooms range from 0 - 8. As Mean slightly < Median, it's slightly Left-Skewed.
- living_measure: Square footage of house range from 290 - 13,540. As Mean > Median, it's Right-Skewed.
- lot_measure: Square footage of lot range from 520 - 16,51,359. As Mean almost double of Median, it's Hightly Right-Skewed.
- ceil: Number of floors range from 1 - 3.5 As Mean ~ Median, it's almost Normal Distributed.
- coast: As this value represent whether house has waterfront view or not. It's categorical column. From above analysis we got know, very few houses has waterfront view.
- sight: Value ranges from 0 - 4. As Mean > Median, it's Right-Skewed
- condition: Represents rating of house which ranges from 1 - 5. As Mean > Median, it's Right-Skewed
- quality: Representign grade given to house which range from 1 - 13. As Mean > Median, it's Right-Skewed.
- ceil_measure: Square footage of house apart from basement ranges in 290 - 9,410. As Mean > Median, it's Right-Skewed.
- basement: Square footage house basement ranges in 0 - 4,820. As Mean highlty > Median, it's Highly Right-Skewed.
- yr_built: House built year ranges from 1900 - 2015. As Mean < Median, it's Left-Skewed.
- yr_renovated: House renovation year only 2015. So this column can be used as Categorical Variable for knowing whether house is renovated or not.

- zipcode: House ZipCode ranges from 98001 - 98199. As Mean > Median, it's Right-Skewed.
- lat: Lattitude ranges from 47.1559 - 47.7776 As Mean < Median, it's Left-Skewed.
- long: Longittude ranges from -122.5190 to -121.315 As Mean > Median, it's Right-Skewed.
- living_measure15: Value ragnes from 399 to 6,210. As Mean > Median, it's Right-Skewed.
- lot_measure15: Value ragnes from 651 to 8,71,200. As Mean highly > Median, it's Highly Right-Skewed.
- furnished: Representing whether house is furnished or not. It's a Categorical Variable
- total_area Total area of house ranges from 1,423 to 16,52,659. As Mean is almost double of Median, it's Highly Right-Skewed.

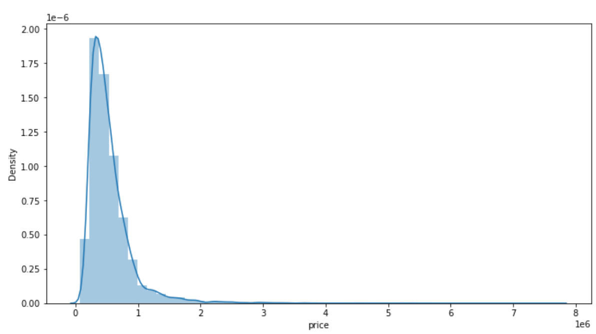
From above analysis we got to know,

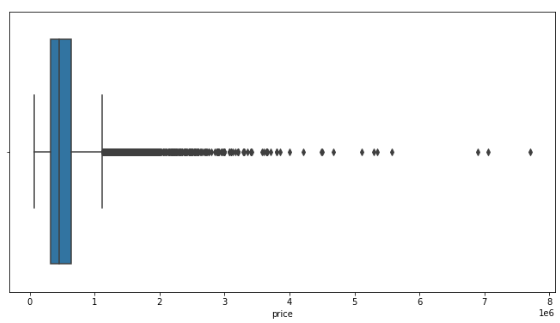
We have columns which are Categorical in nature are -> yr_renovated, furnished

- We have any null or missing values for any of the columns so imputation is done with appropriate value

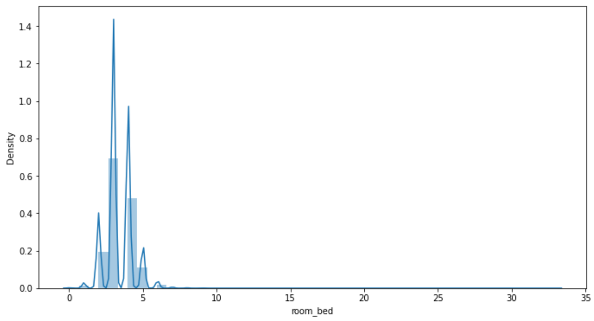
Exploratory Data Analysis:

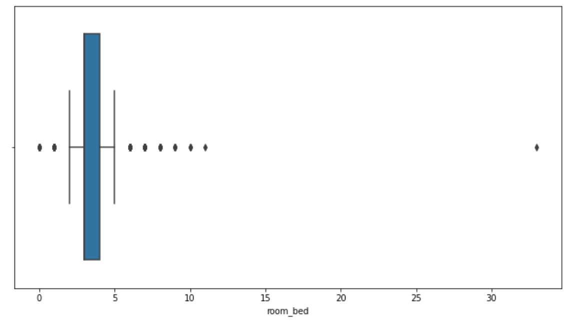
```
price
Skew :  4.022
```



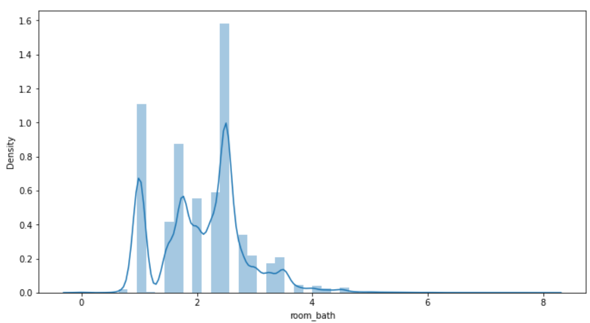


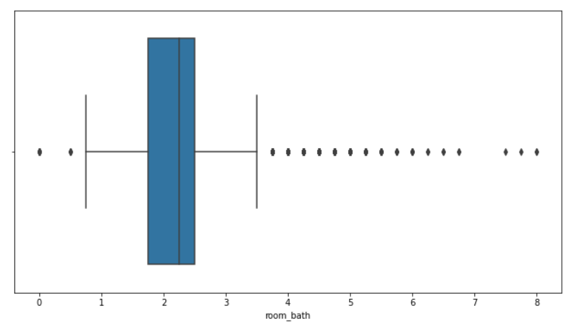

```
room_bed
Skew :  1.989
```



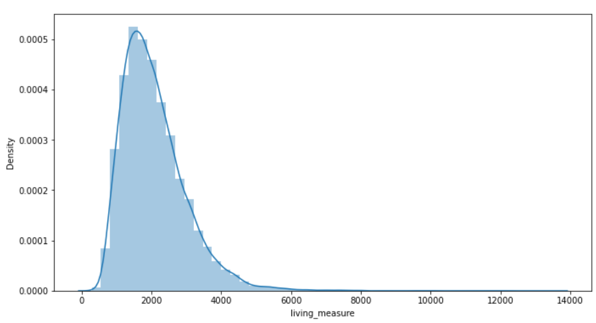


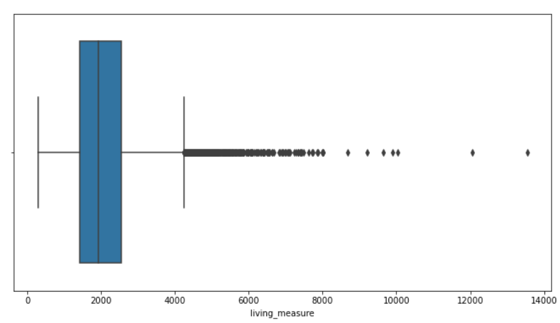
room_bath
Skew : 0.505



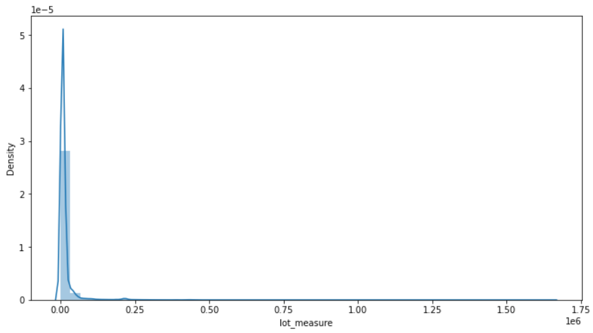


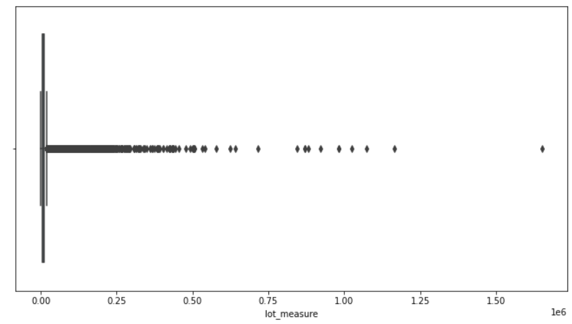
living_measure
Skew : 1.473



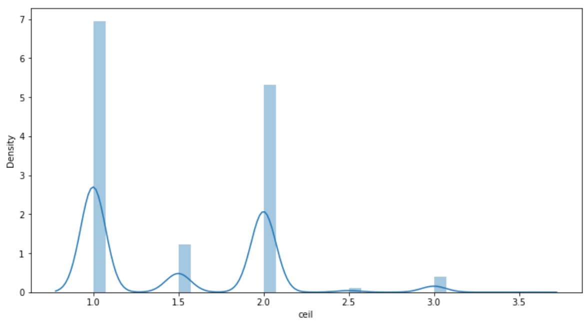


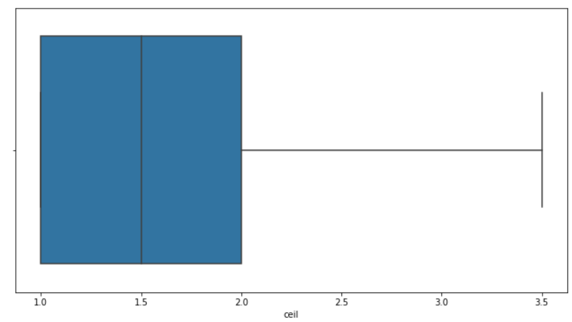
lot_measure
Skew : 13.084



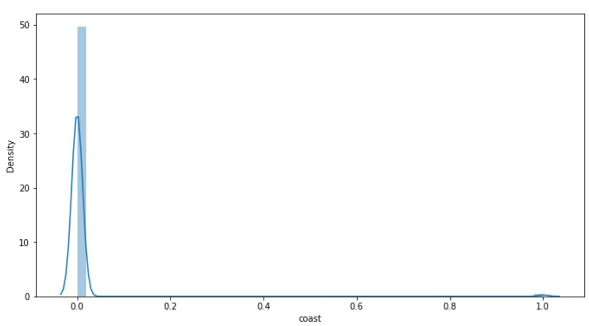


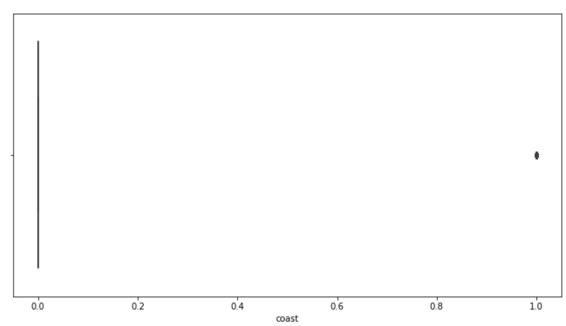
ceil
Skew : 0.622



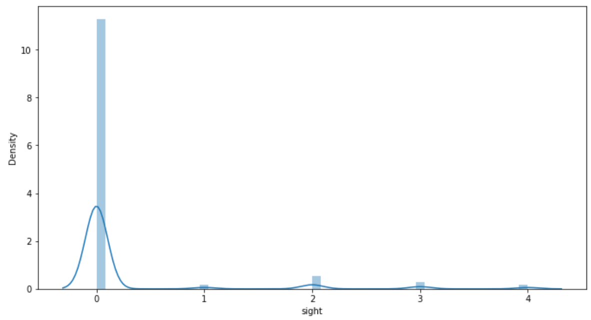


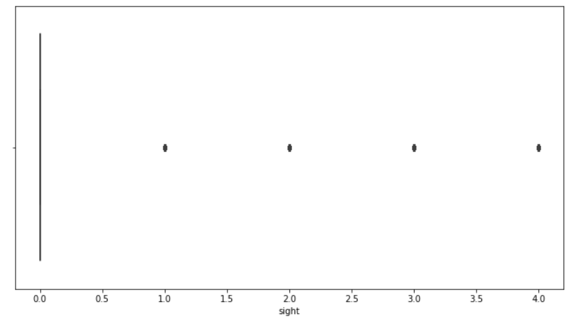
coast
Skew : 11.457



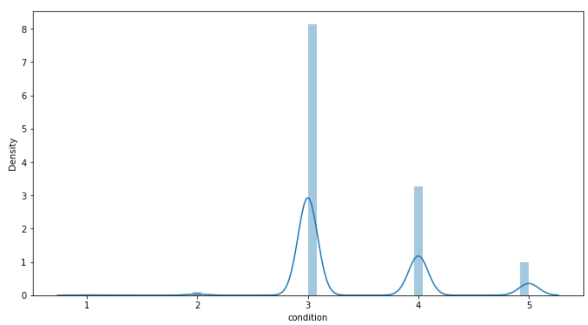


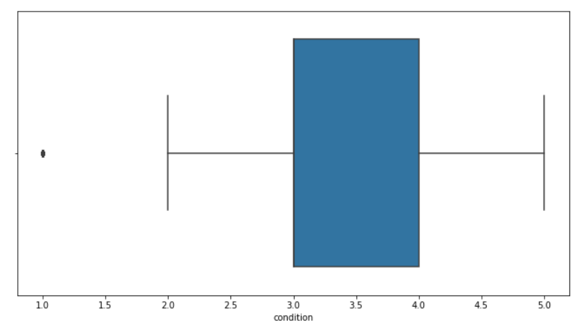
sight
Skew : 3.401



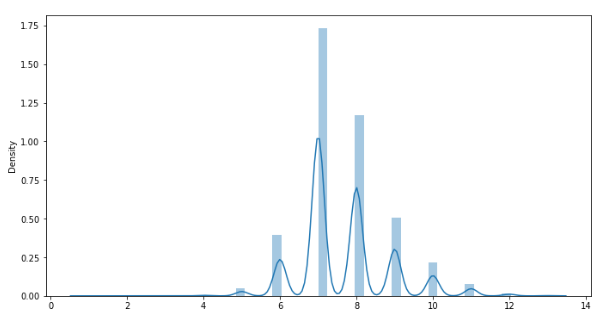


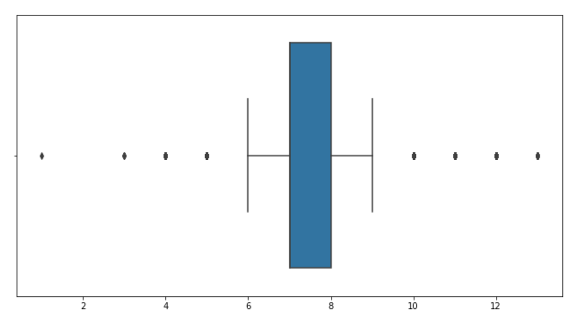
condition
Skew : 1.039



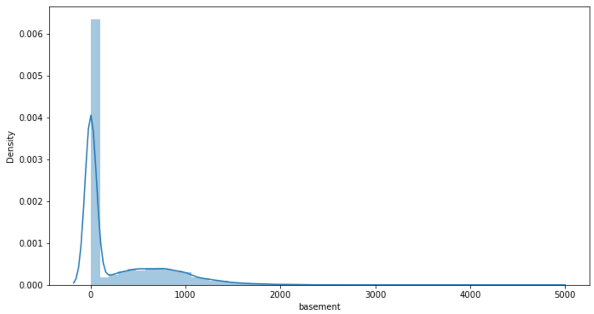


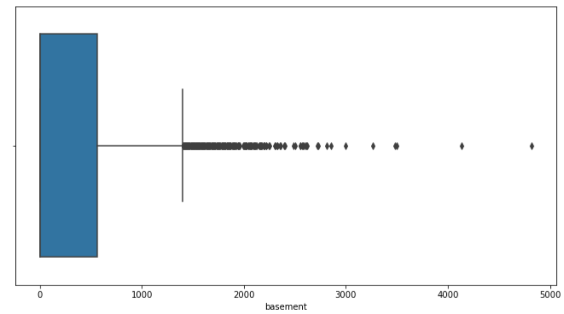
quality
Skew : 0.771



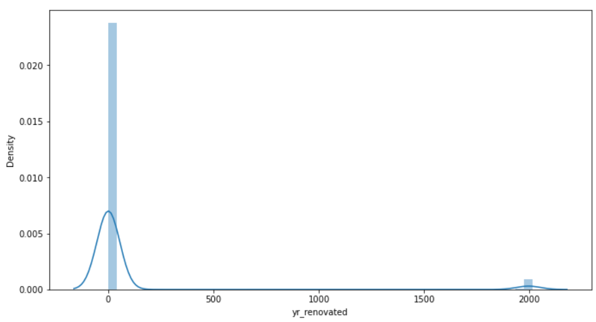


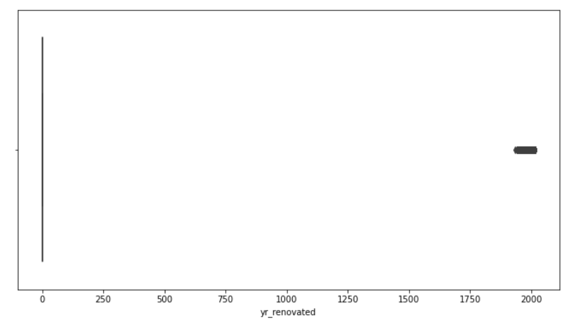
basement
Skew : 1.578

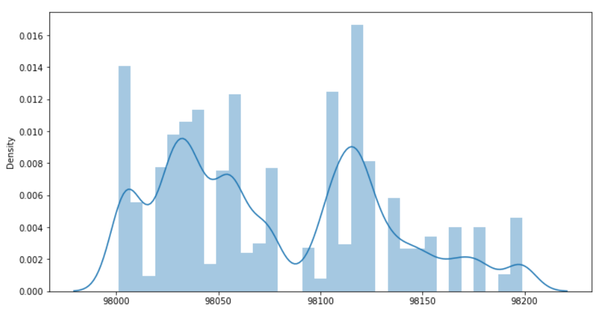


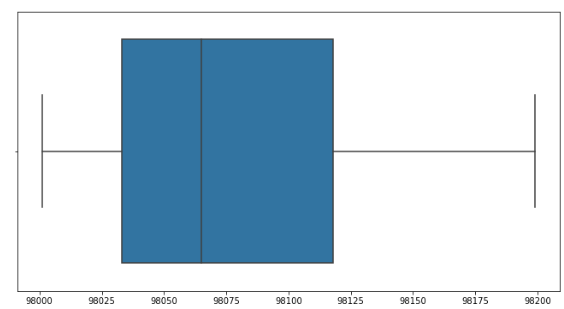


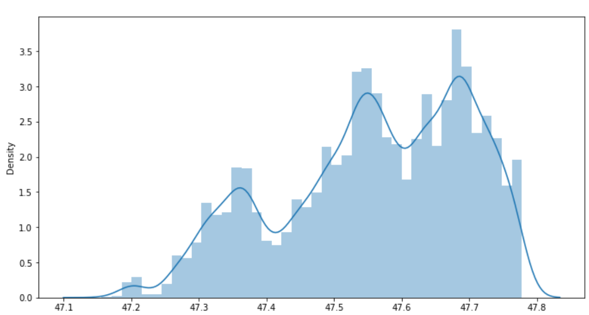
yr_renovated
Skew : 4.549

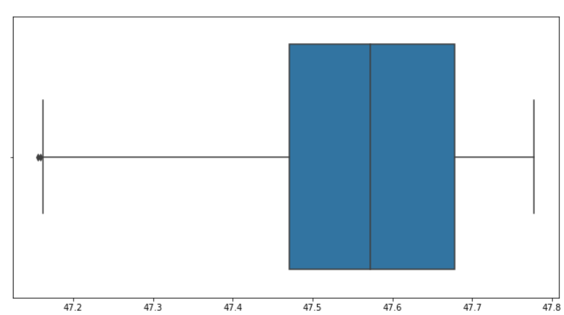


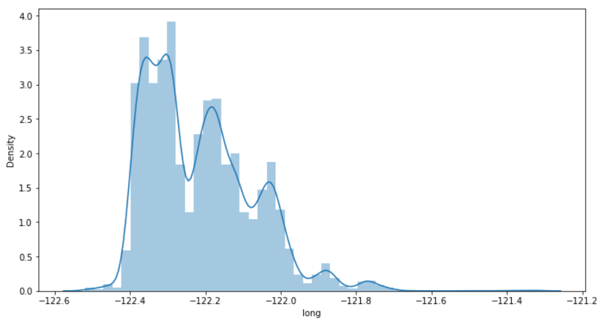


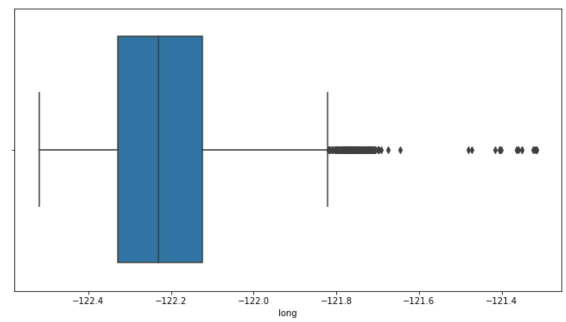


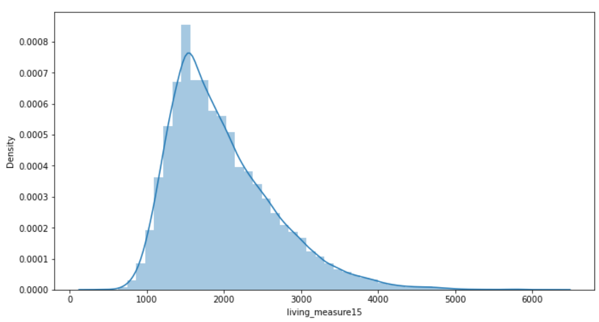


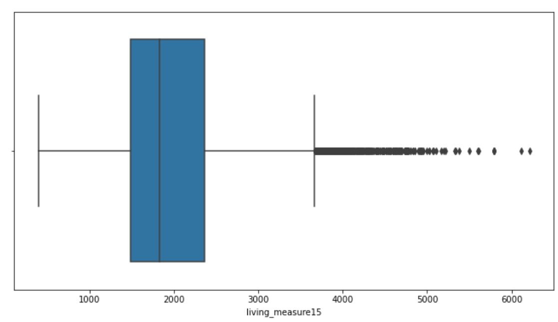


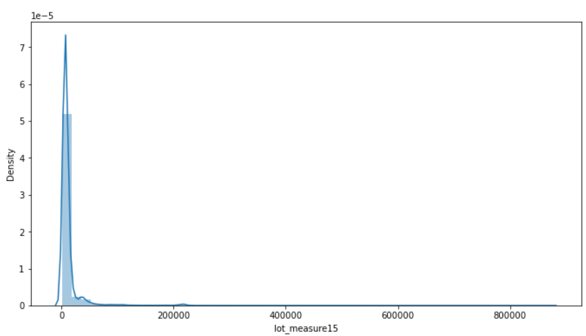


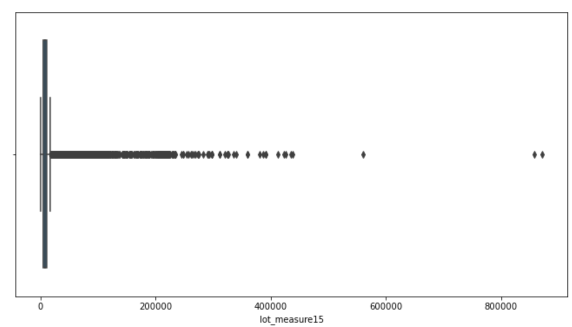

Density
long

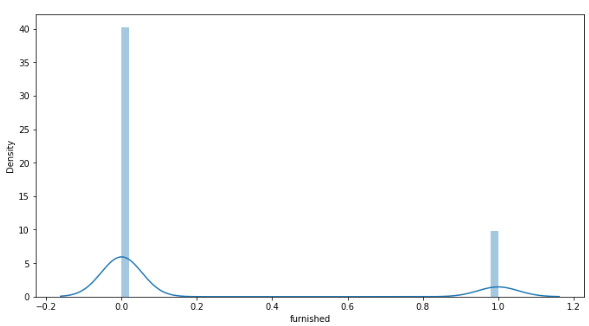

long

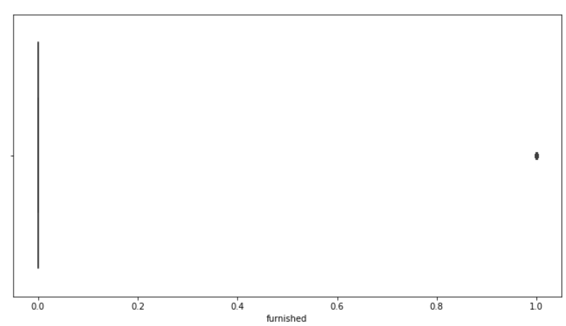

Density
living_measure15

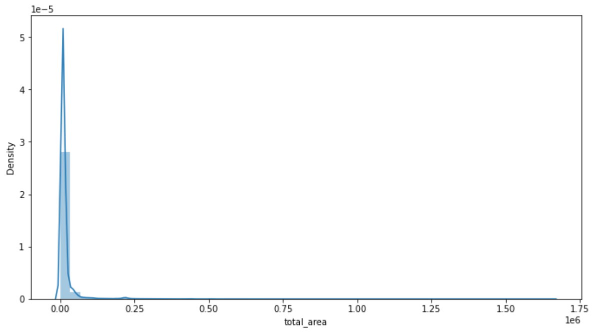

living_measure15

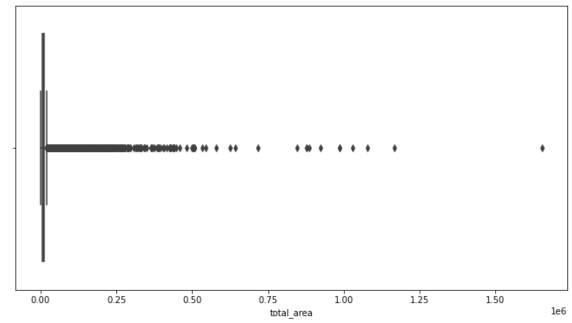

1e−5
Density
lot_measure15


lot_measure15


Density
furnished


furnished





We can see, there are lot of features which have outliers. So, we might need to treat those before building model.

Analyzing Feature: Cid

- **cid - CID is appearing multiple times, it seems data contains house which is sold multiple times**
- **We have 176 properties that were sold more than once in the given data.**

We successfully converted dayhours feature to month_year for better analysis.

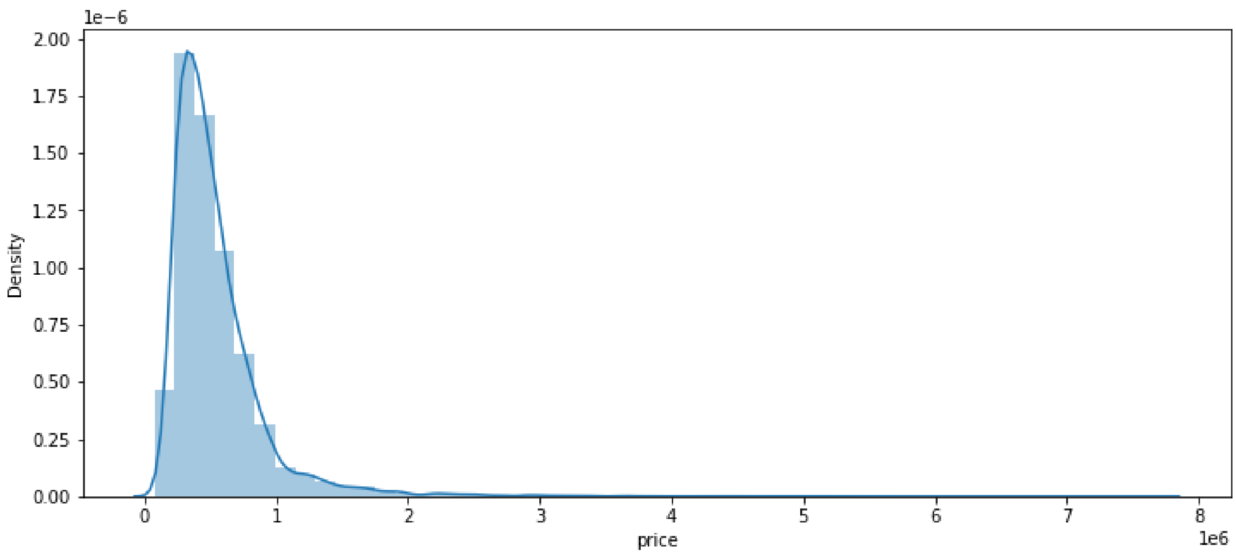
April-2015	2231
July-2014	2211
June-2014	2180
August-2014	1940
October-2014	1878
March-2015	1875
September-2014	1774
May-2014	1768
December-2014	1471
November-2014	1411
February-2015	1250
January-2015	978
May-2015	646

- **We can see, most houses sold in April, July month**

month_year	
April-2015	561933.463021
August-2014	536527.039691
December-2014	524602.893270
February-2015	507919.603200
January-2015	525963.251534
July-2014	544892.161013
June-2014	558123.736239
March-2015	544057.683200
May-2014	548166.600113
May-2015	558193.095975
November-2014	522058.861800
October-2014	539127.477636
September-2014	529315.868095
Name: price, dtype: float64	

- **So, the time line of the sale data of the properties is from May-2014 to May-2015 and April month have the highest mean price.**

Analyzing Feature: Price:

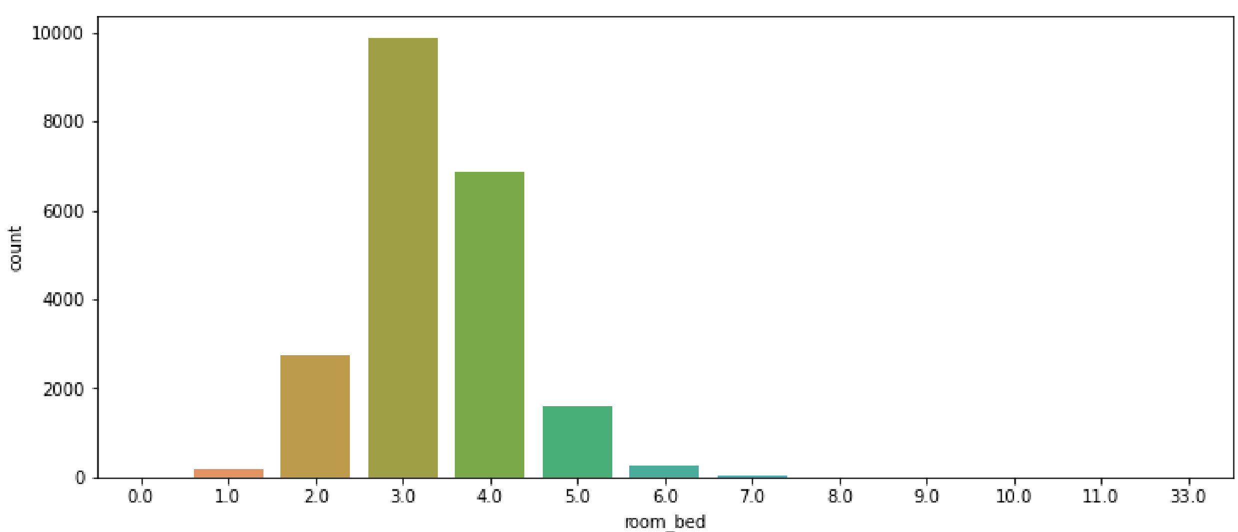


- The Price is ranging from 75,000 to 77,00,000 and distribution is right-skewed.

Analyzing Feature: room_bed:

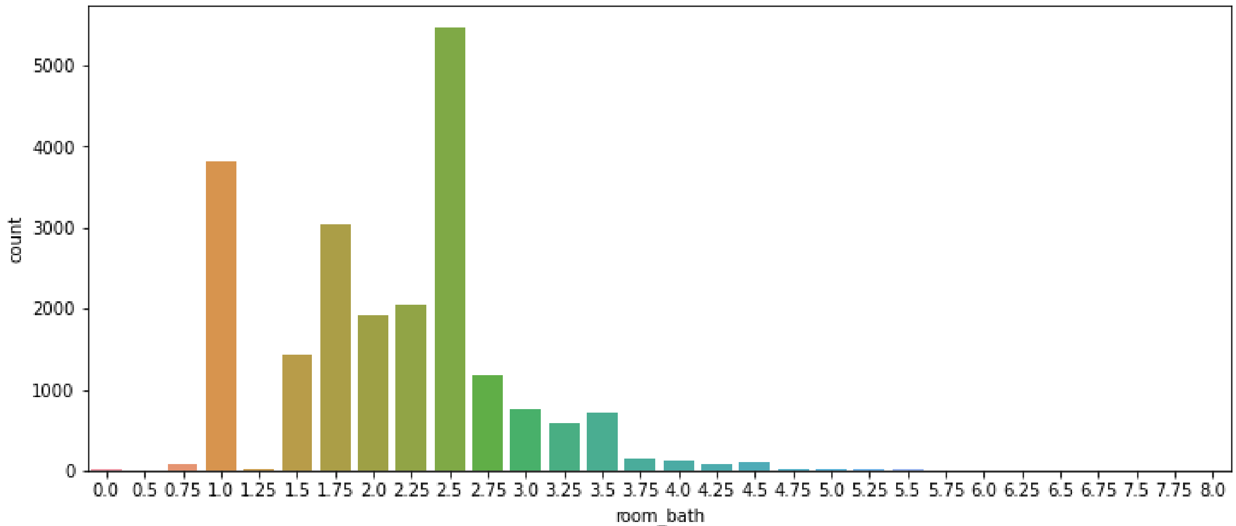
3.0	9875
4.0	6854
2.0	2747
5.0	1595
6.0	270
1.0	197
7.0	38
8.0	13
0.0	13
9.0	6
10.0	3
33.0	1
11.0	1

- The value of 33 seems to be outlier we need to check the data point before imputing the same, Will delete this data point after bivariate analysis as it looks to be an outlier as it has low price for 33 bed room properties

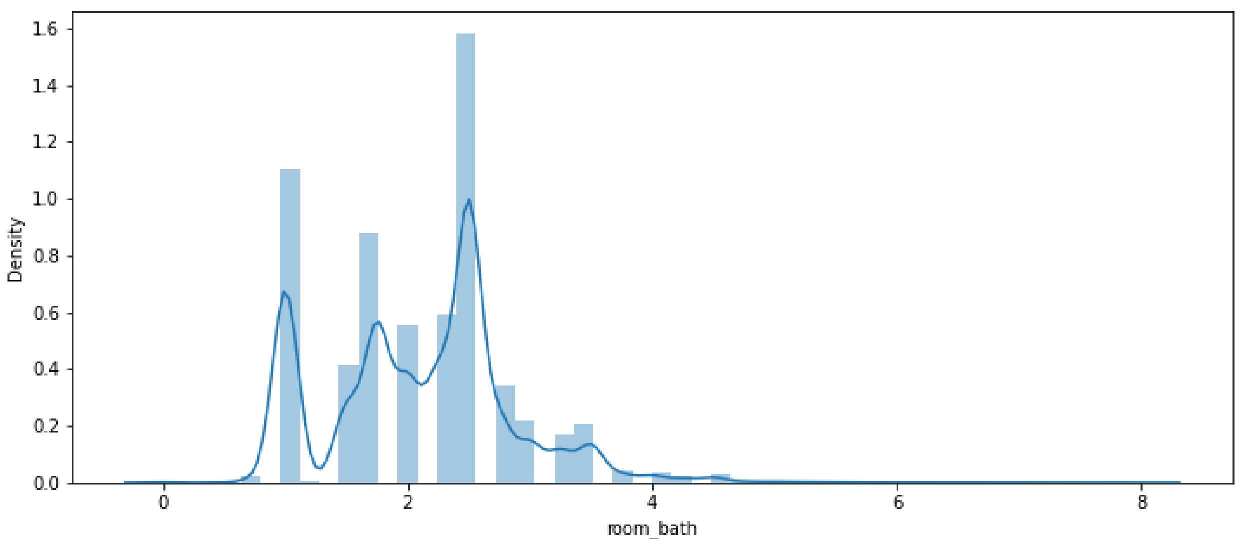


- Most of the houses/properties have 3 or 4 bedrooms

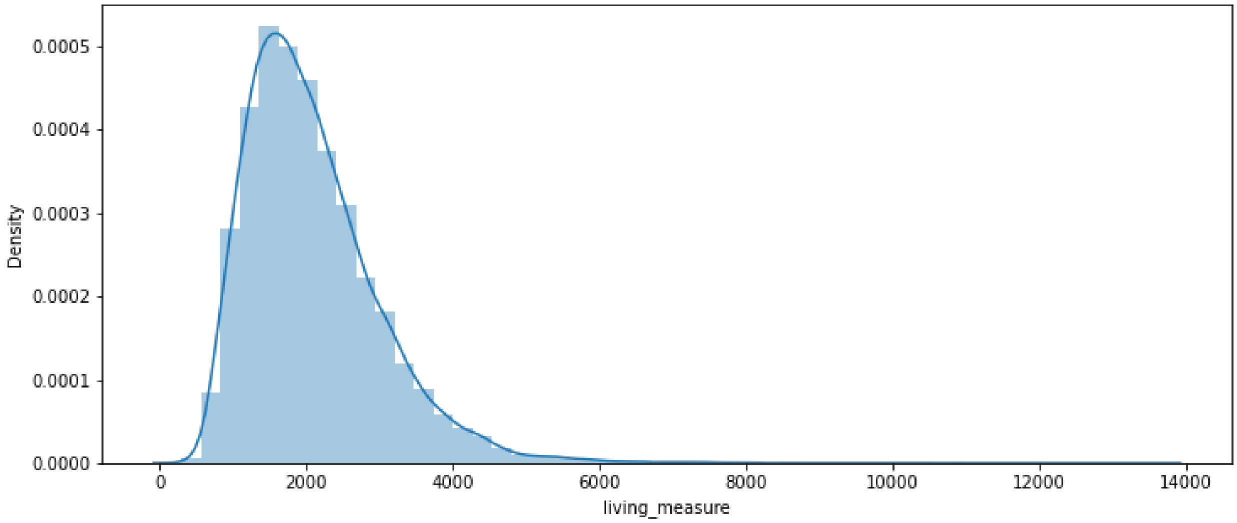
Analyzing Feature: room_bath



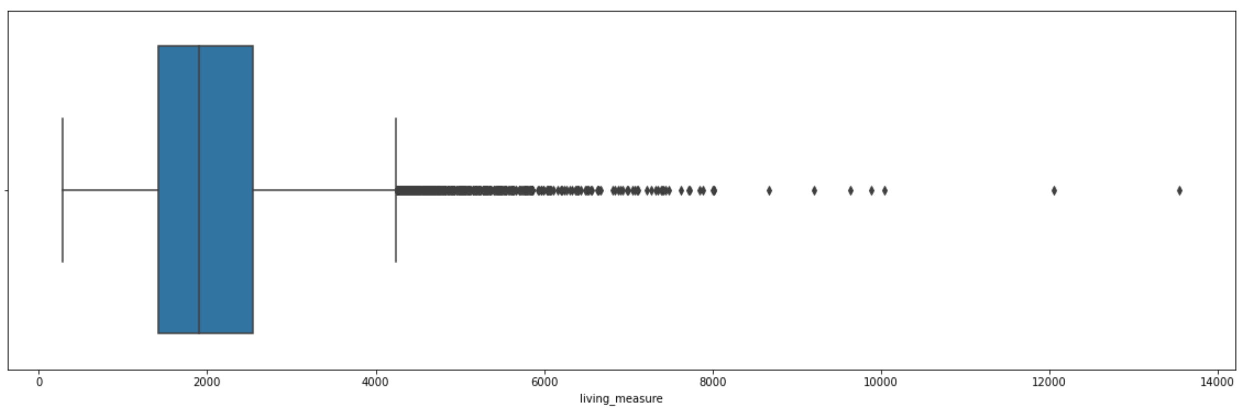
- Majority of the properties have bathroom in the range of 1.0 to 2.5



Analyzing Feature: Living measure



- Data distribution tells us, living_measure is right-skewed.

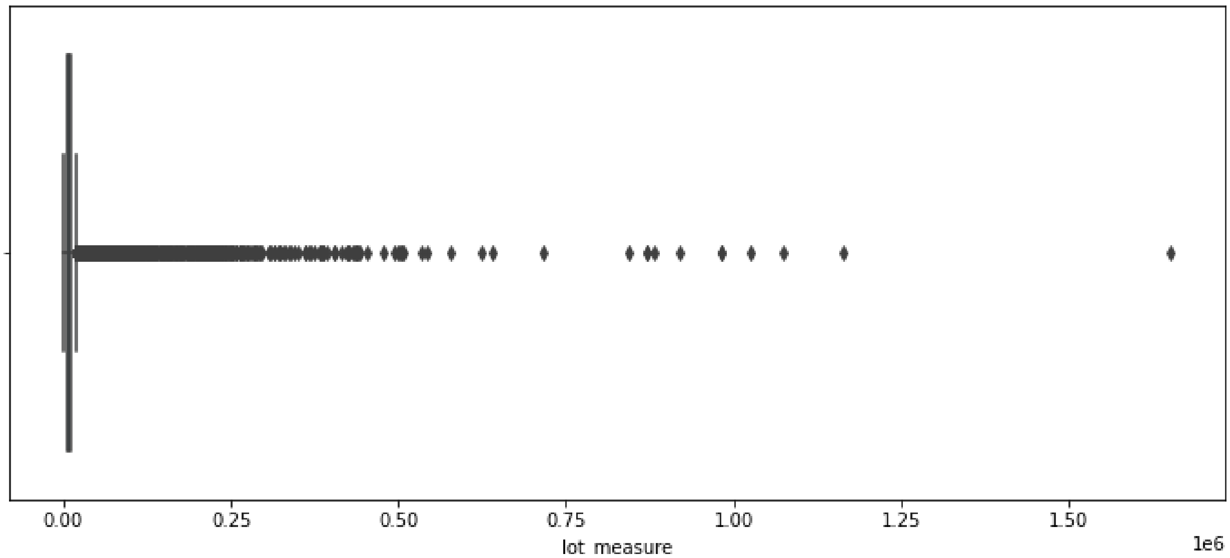


- There are many outliers in living measure. Need to review further to treat the same.

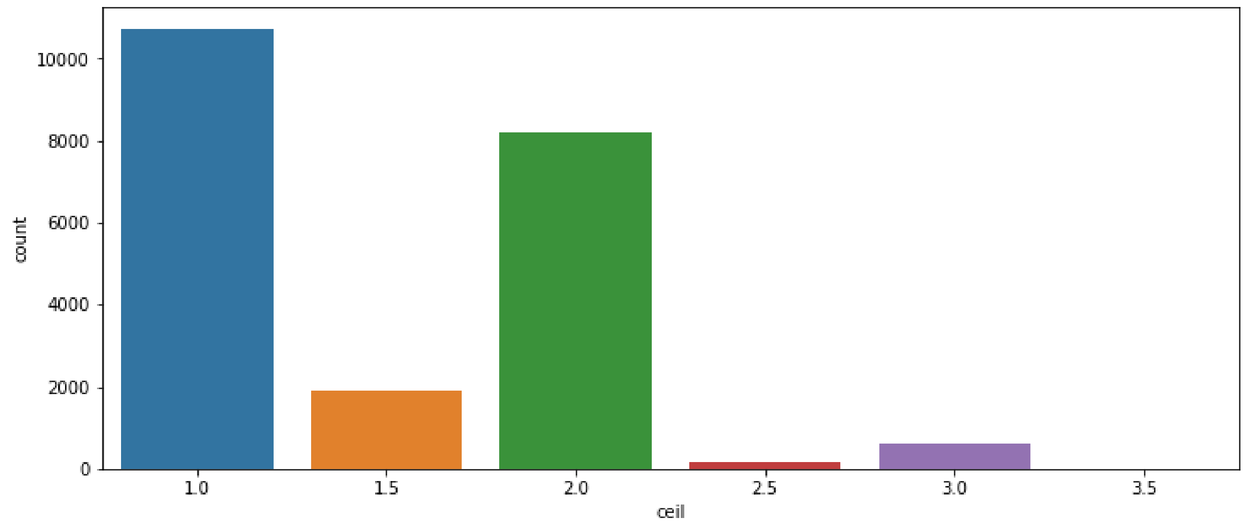
- We have only 9 properties/house which have more than 8k living_measure. So will treat these outliers.

Analyzing Feature: lot_measure

- We have only 1 property with more than 12,50,000 lot_measure. So we need to treat this.

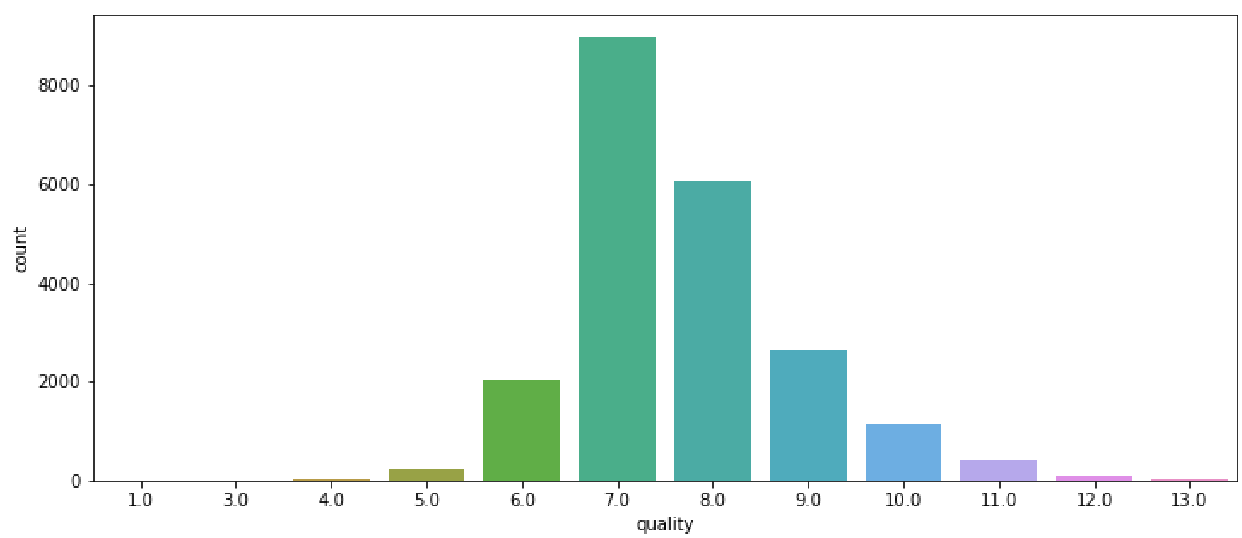


Analyzing Feature: ceil



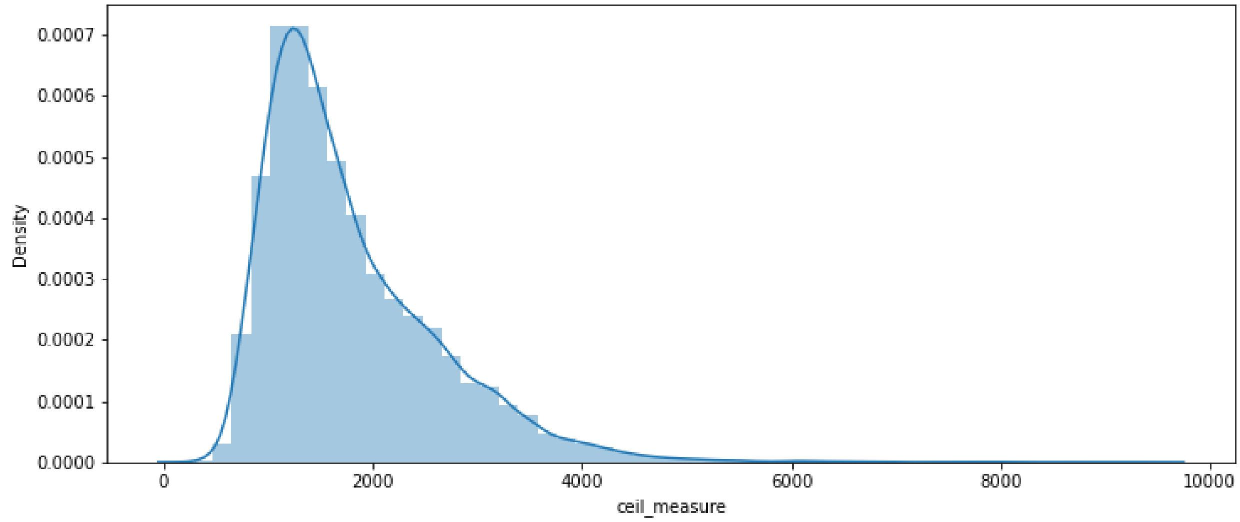
- We can see, most houses have 1 floor

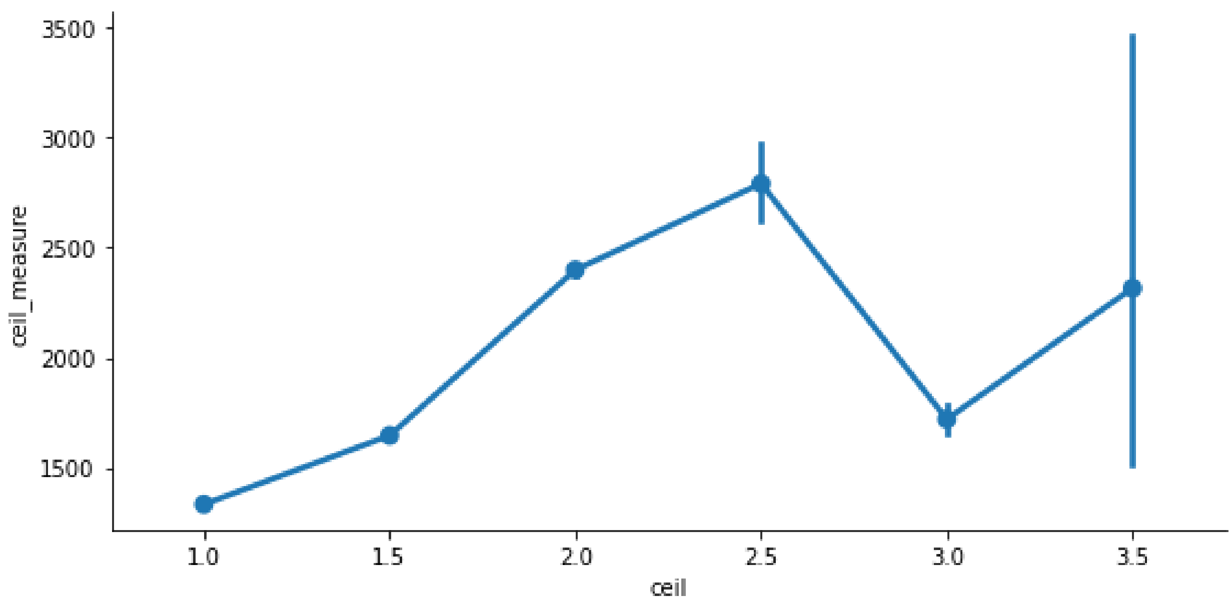
Analyzing Feature: quality



- There are only 13 properties which have the highest quality rating
- Quality - most properties have quality rating between 6 to 10

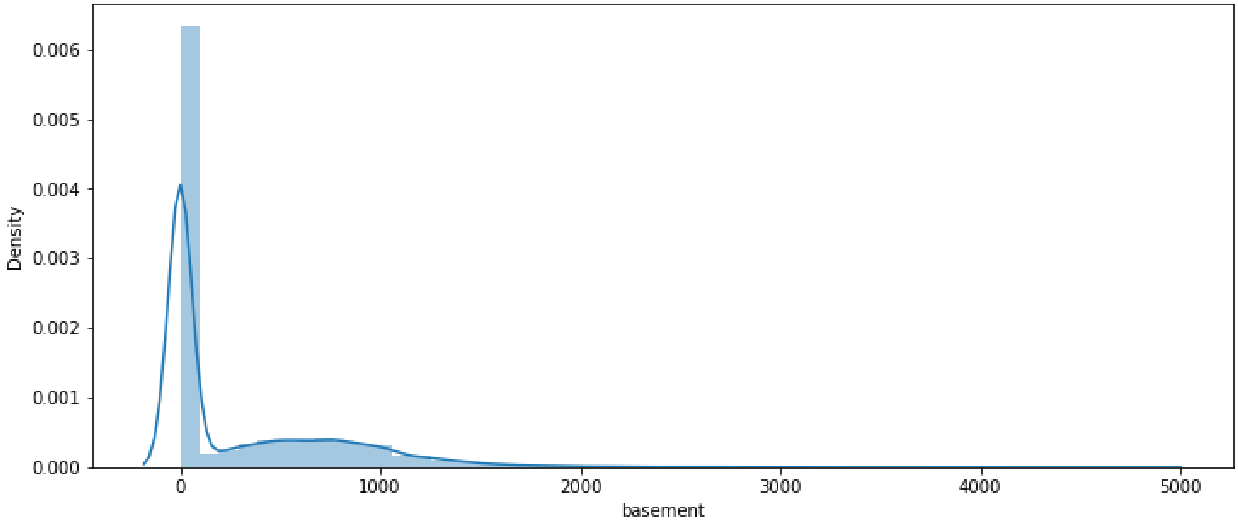
Analyzing Feature: ceil_measure



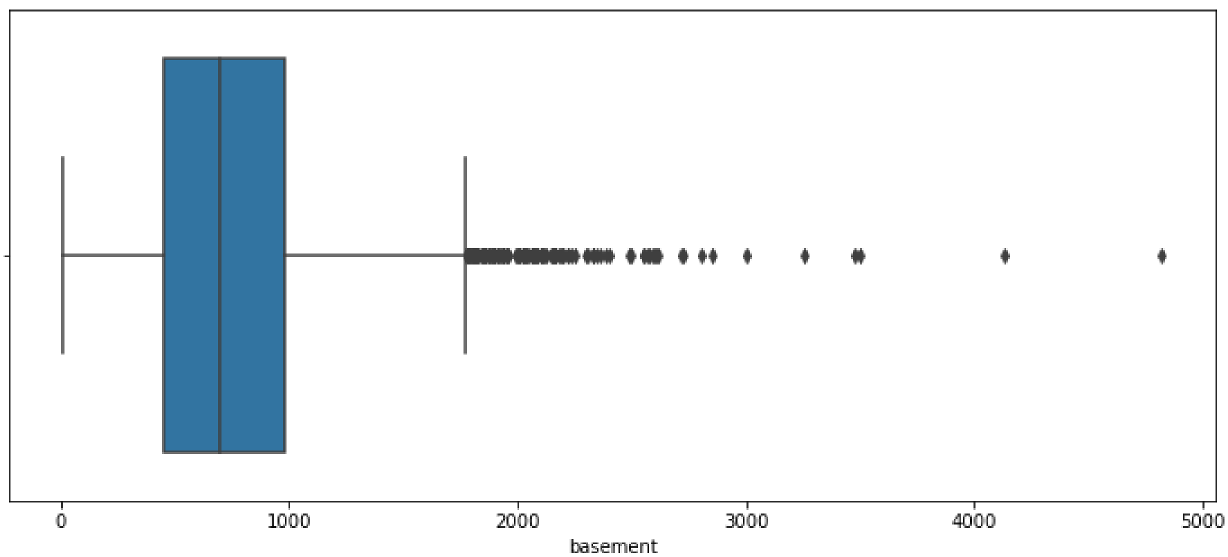


- There is no pattern in Ceil Vs Ceil_measure
- The vertical lines at each point represent the inter quartile range of values at that point.
- ceil_measure - its highly skewed.

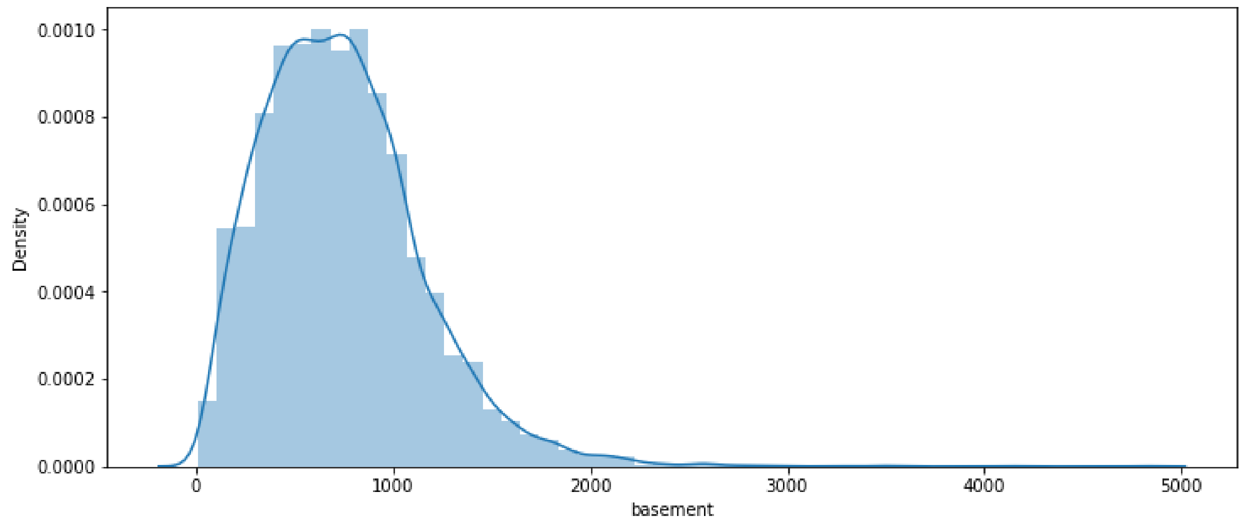
Analyzing Feature: basement



- We can see 2 gaussians, which tells us there are properties which don't have basements and some have the basements
- We have almost 60% of the properties without basement

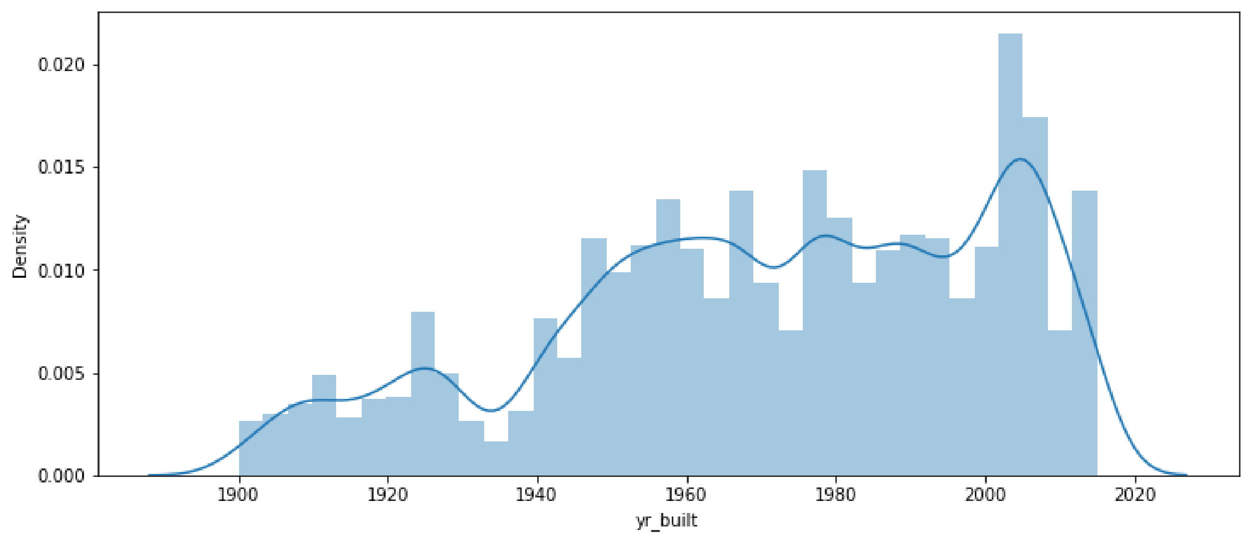


- We can clearly see, there are outliers. We need to treat this before our model.
- We have only 2 properties with more than 4,000 measure basements
- We have only 2 properties with more than 4,000 measure basements



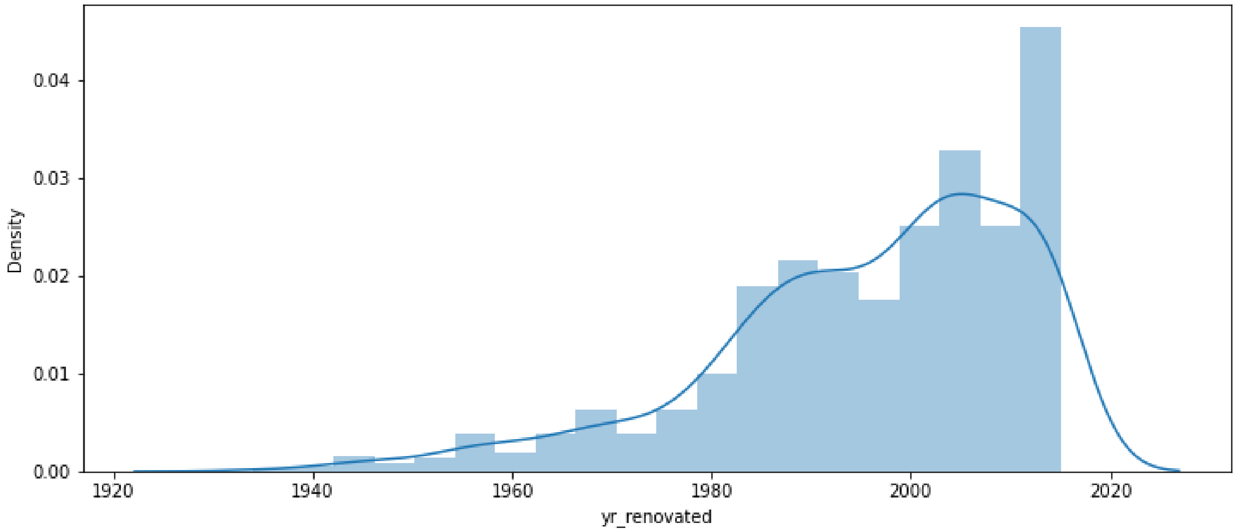
- Distribution having basement is right-skewed

Analyzing Feature: yr_built



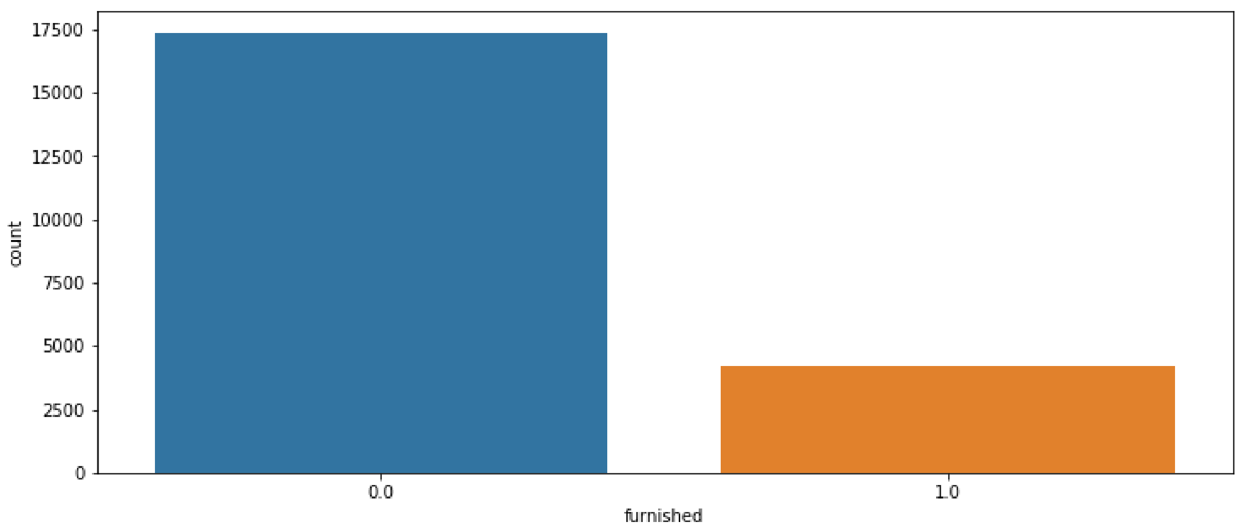
- The built year of the properties range from 1900 to 2014 and we can see upward trend with time

Analyzing Feature: yr_renovated



- Only 914 houses were renovated out of 21613 houses
- Now will create age column from columns : yr_built & yr_renovated

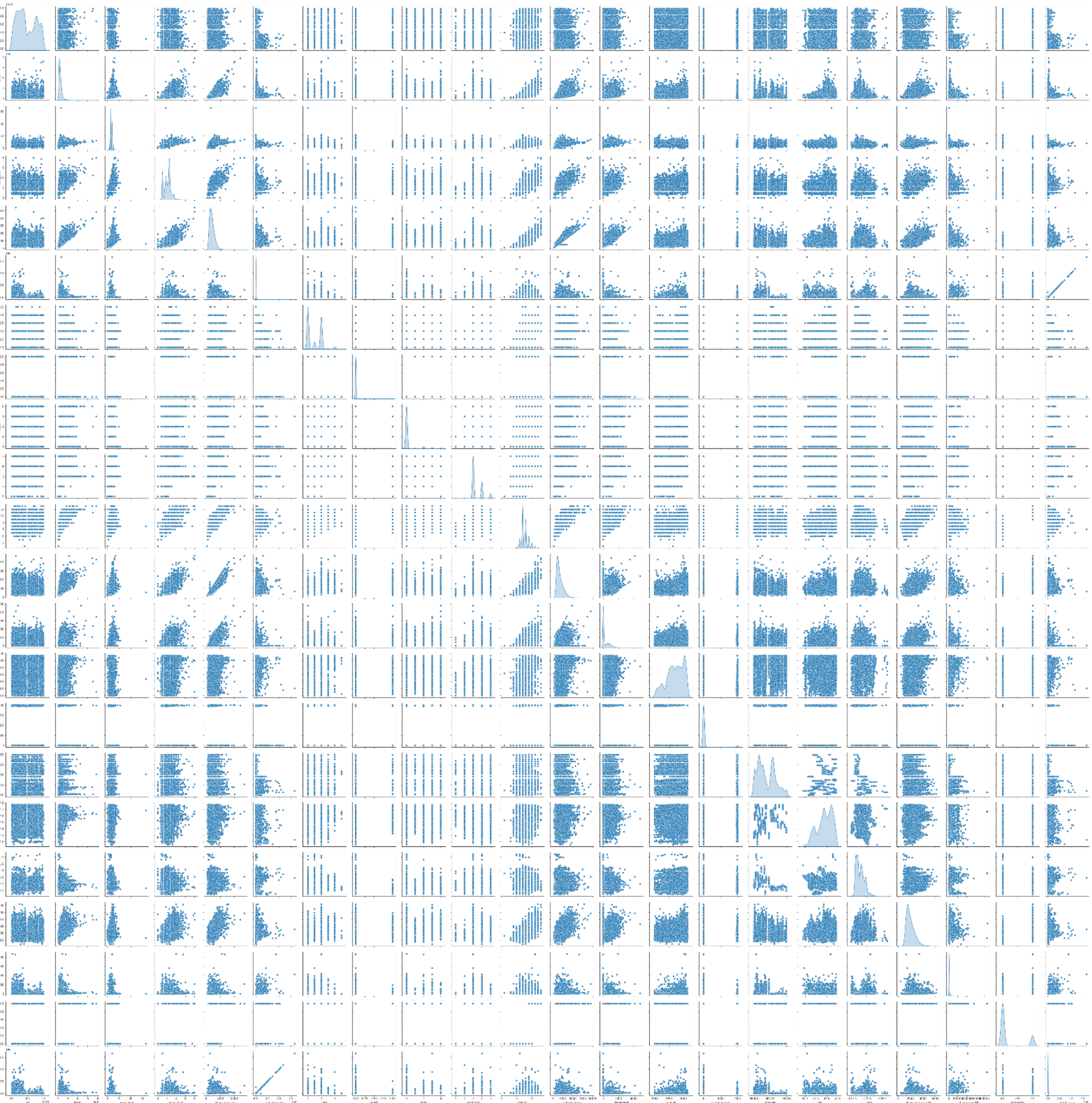
Analyzing Feature: furnished



- Most properties are not furnished. Furnish column need to be converted into categorical column

BIVARIATE ANALYSIS :

Pair Plot:



- *let's plot all the variables and confirm our above deduction with more confidence*

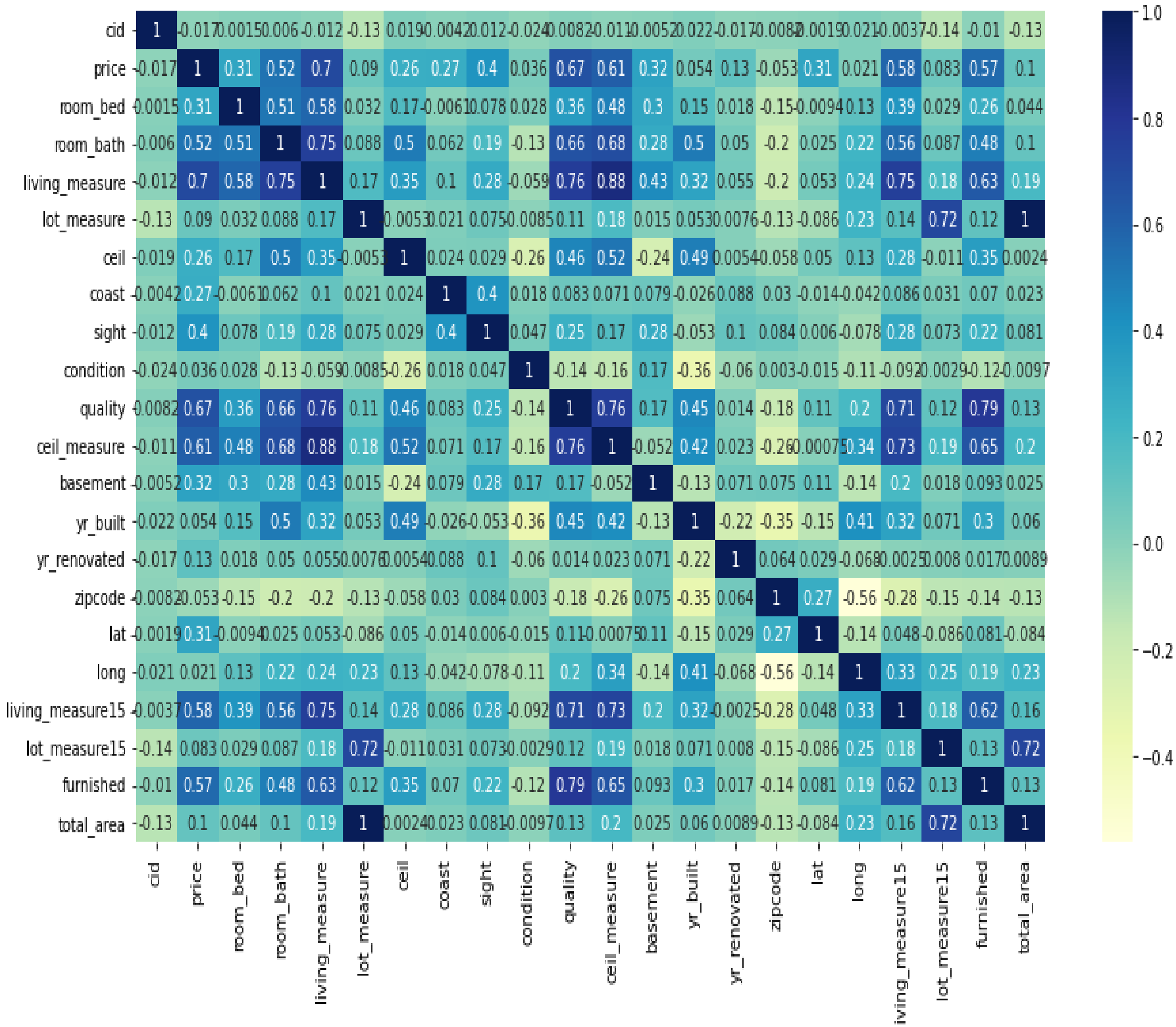
From above pair plot, we observed/deduced below

- price: price distribution is Right-Skewed as we deduced earlier from our 5-factor analysis
- room_bed: our target variable (price) and room_bed plot is not linear. It's distribution have lot of gaussians
- room_bath: It's plot with price has somewhat linear relationship. Distribution has number of gaussians.
- living_measure: Plot against price has strong linear relationship. It also have linear relationship with room_bath variable. So might remove one of these 2. Distribution is Right-Skewed.
- lot_measure: No clear relationship with price.
- ceil: No clear relationship with price. We can see, it's have 6 unique values only. Therefore, we can convert this column into categorical column for values.
- coast: No clear relationship with price. Clearly it's categorical variable with 2 unique values.
- sight: No clear relationship with price. This has 5 unique values. Can be converted to Categorical variable.
- condition: No clear relationship with price. This has 5 unique values. Can be converted to Categorical variable.
- quality: Somewhat linear relationship with price. Has discrete values from 1 - 13. Can be converted to Categorical variable.
- ceil_measure: Strong linear relationship with price. Also with room_bath and living_measure features. Distribution is Right-Skewed.
- basement: No clear relationship with price.
- yr_built: No clear relationship with price.
- yr_renovated: No clear relationship with price. Have 2 unique values. Can be converted to Categorical Variable which tells whether house is renovated or not.
- living_measure15: Somewhat linear relationship with target feature. It's same as living_measure. Therefore we can drop this variable.
- lot_measure15: No clear relationship with price or any other feature.
- furnished: No clear relationship with price or any other feature. 2 unique values so can be converted to Categorical Variable
- total_area: No clear relationship with price. But it has Very Strong linear relationship with lot_measure. So one of it can be dropped.

We have linear relationships in below featues as we got to know from above matrix

- price: room_bath, living_measure, quality, living_measure15, furnished
- living_measure: price, room_bath. So we can consider dropping 'room_bath' variable.
- quality: price, room_bath, living_measure
- ceil_measure: price, room_bath, living_measure, quality
- living_measure15: price, living_measure, quality. So we can consider dropping living_measure15 as well. As it's giving same info as living_measure.
- lot_measure15: lot_measure. Therefore, we can consider dropping lot_measure15, as it's giving same info.
- furnished: quality
- total_area: lot_measure, lot_measure15. Therefore, we can consider dropping total_area feature as well. As it's giving same info as lot_measure.

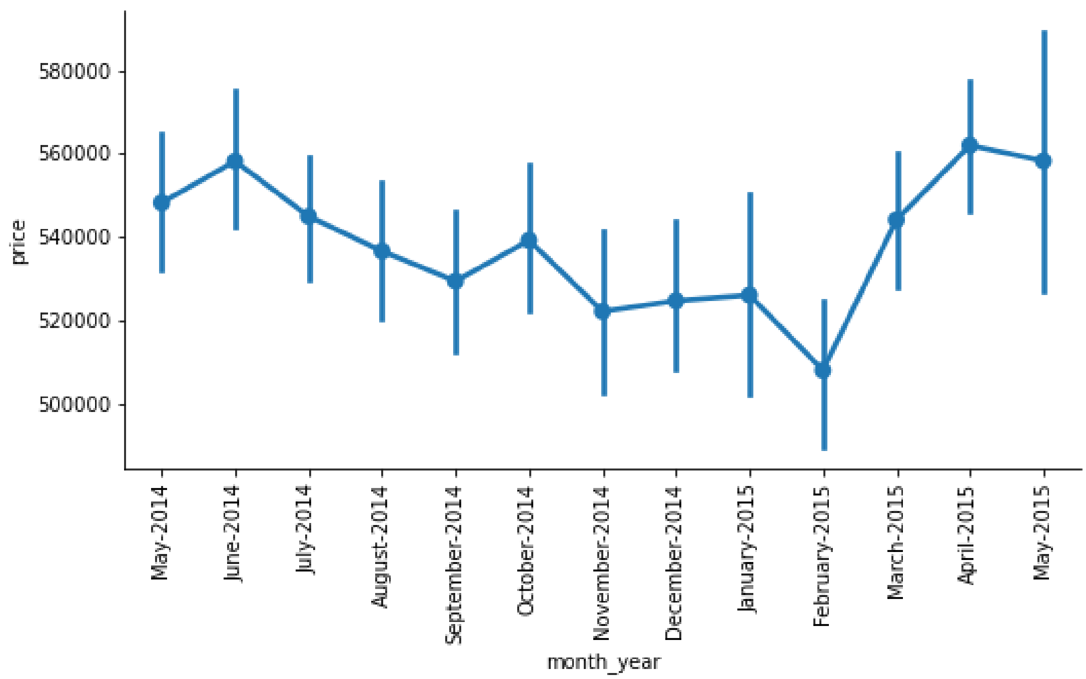
We can plot heatmap and can easily confirm our above findings



Analyzing Bivariate for Feature: month_year:

month_year	mean	median	size
Apr-15	561933.5	476500	2231
Aug-14	536527	442100	1940
Dec-14	524602.9	432500	1471
Feb-15	507919.6	425545	1250
Jan-15	525963.3	438500	978
Jul-14	544892.2	465000	2211

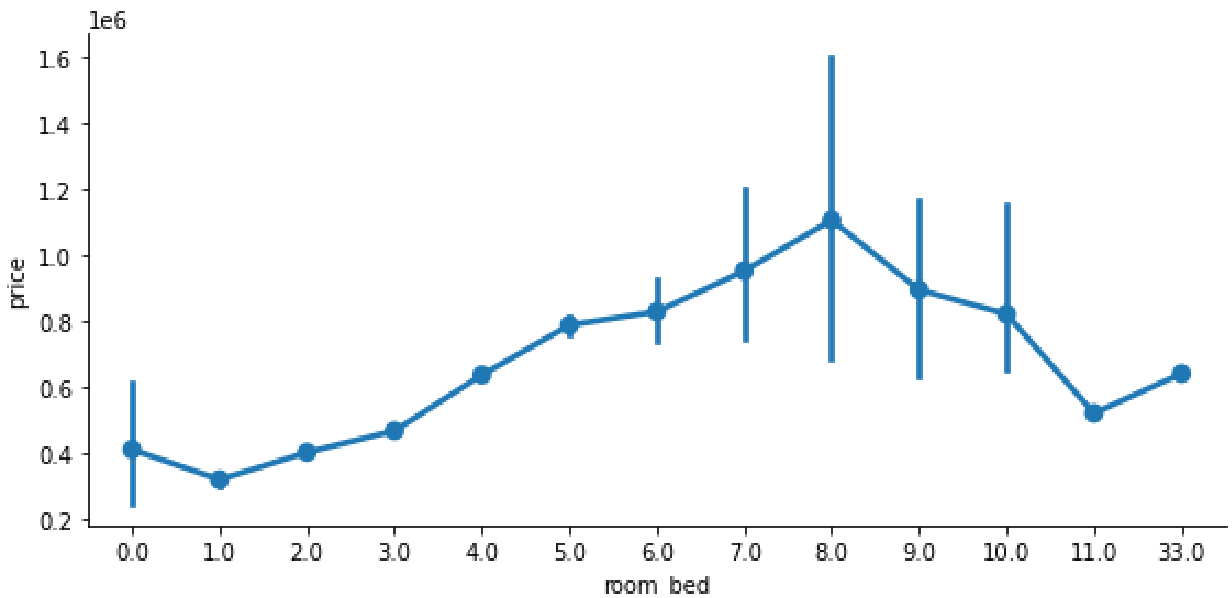
Jun-14	**558123.7**	**465000**	**2180**
Mar-15	**544057.7**	**450000**	**1875**
May-14	**548166.6**	**465000**	**1768**
May-15	**558193.1**	**455000**	**646**
Nov-14	**522058.9**	**435000**	**1411**
Oct-14	**539127.5**	**446900**	**1878**
Sep-14	**529315.9**	**450000**	**1774**



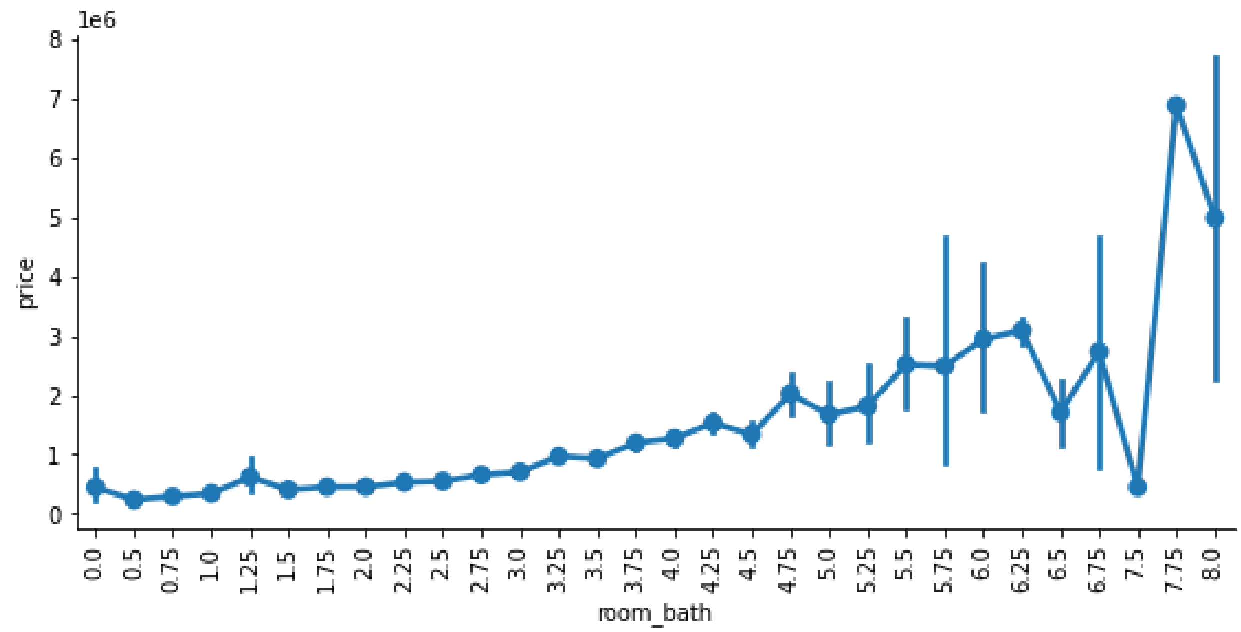
- *month,year in which house is sold. Price is not influenced by it, though there are outliers and can be easily seen.*
- *The mean price of the houses tend to be high during March,April, May as compared to that of September, October, November,December period.*

Analyzing Bivariate for Feature: room_bed

room_bed	mean	median	size
0	4.10E+05	288000	13
1	3.18E+05	299000	199
2	4.01E+05	374000	2760
3	4.66E+05	413000	9824
4	6.36E+05	549997.5	6882
5	7.87E+05	620000	1601
6	8.26E+05	650000	272
7	9.51E+05	728580	38
8	1.11E+06	700000	13
9	8.94E+05	817000	6
10	8.20E+05	660000	3
11	5.20E+05	520000	1
33	6.40E+05	640000	1

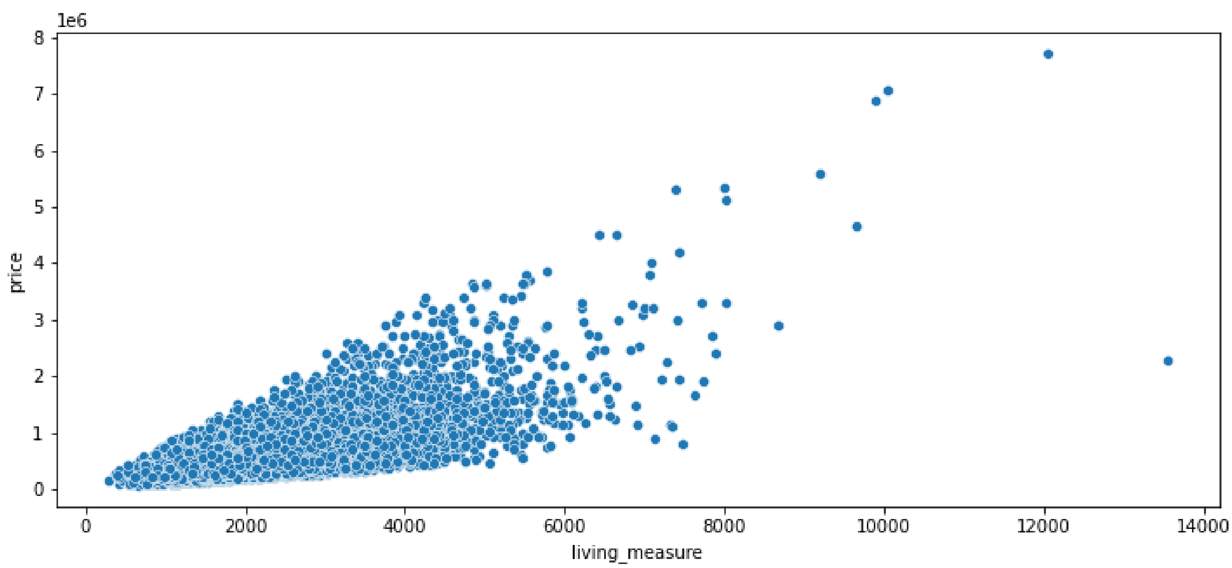


- There is clear increasing trend in price with room_bed



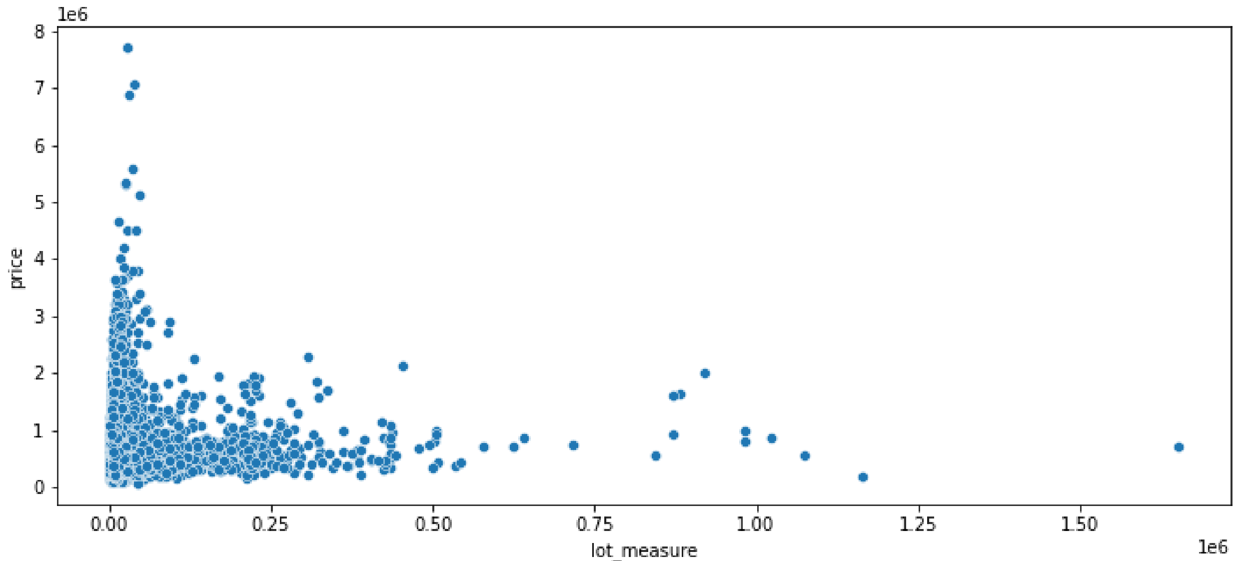
- There is upward trend in price with increase in room_bath

Analyzing Bivariate for Feature: living_measure

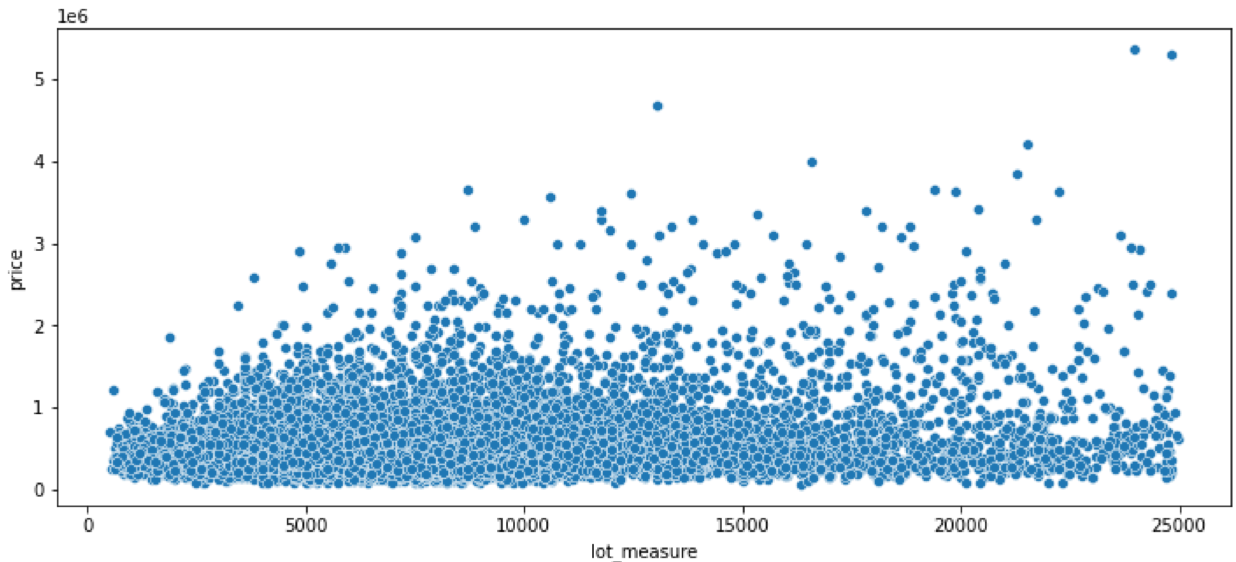


- There is clear increment in price of the property with increment in the living measure But there seems to be one outlier to this trend. Need to evaluate the same

Analyzing Bivariate for Feature: lot_measure



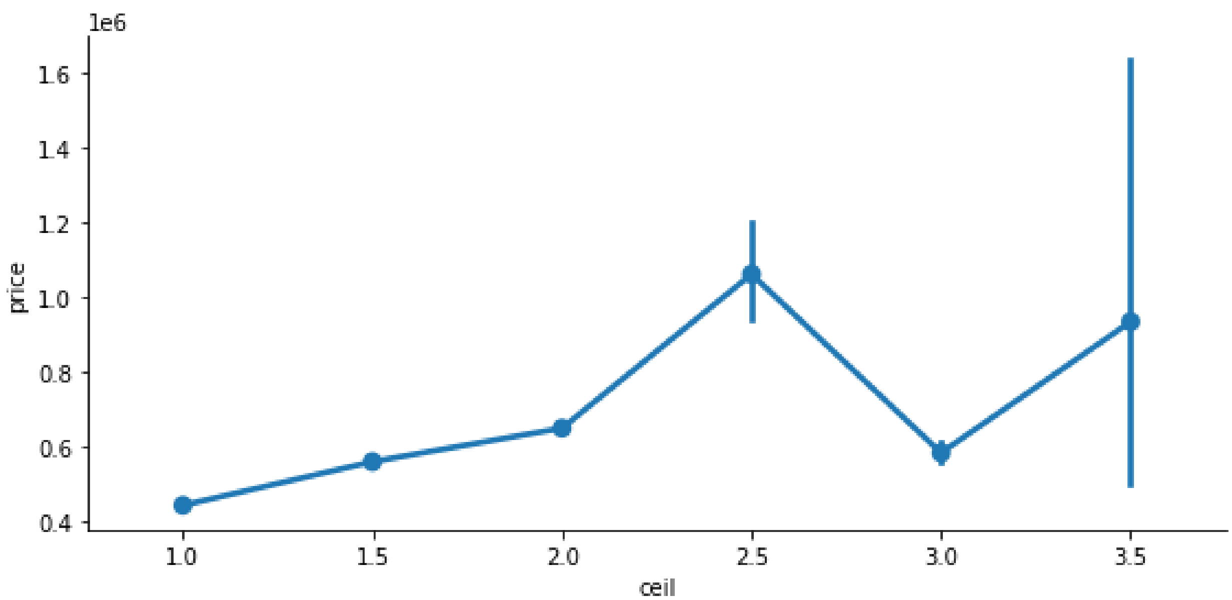
- There doesn't seem to be no relation between lot_measure and price trend



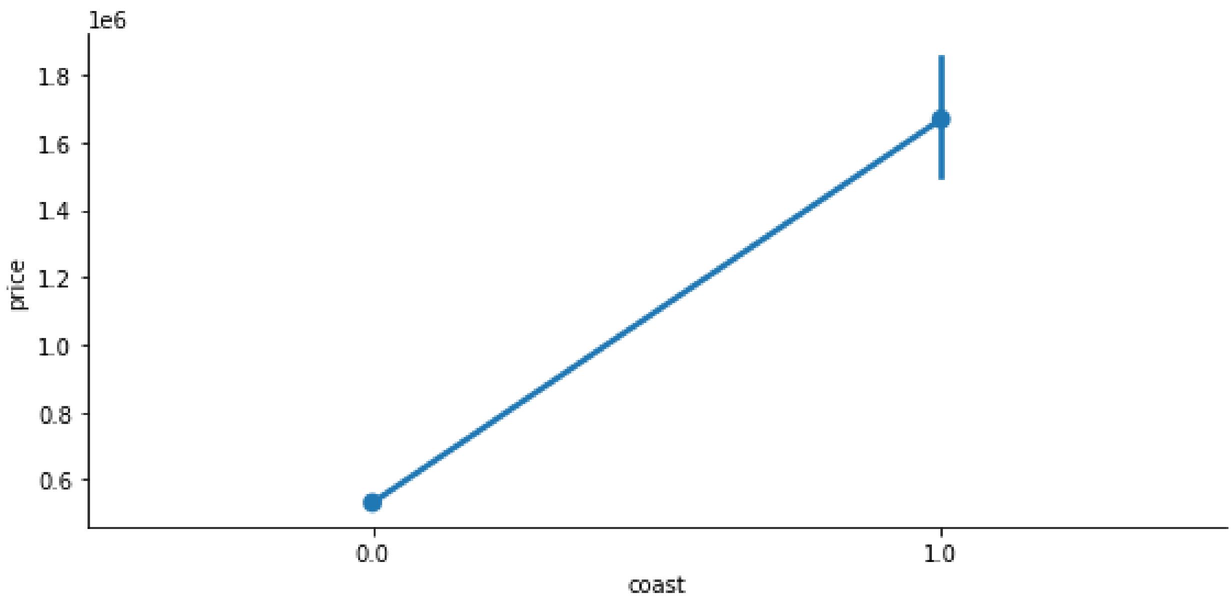
- Almost 95% of the houses have <25000 lot_measure. But there is no clear trend between lot_measure and price

Analyzing Bivariate for Feature: ceil

ceil	mean	median	size
1	4.42E+05	390000	10680
1.5	5.59E+05	524475	1910
2	6.49E+05	542950	8241
2.5	1.06E+06	799200	161
3	5.83E+05	490000	613
3.5	9.34E+05	534500	8



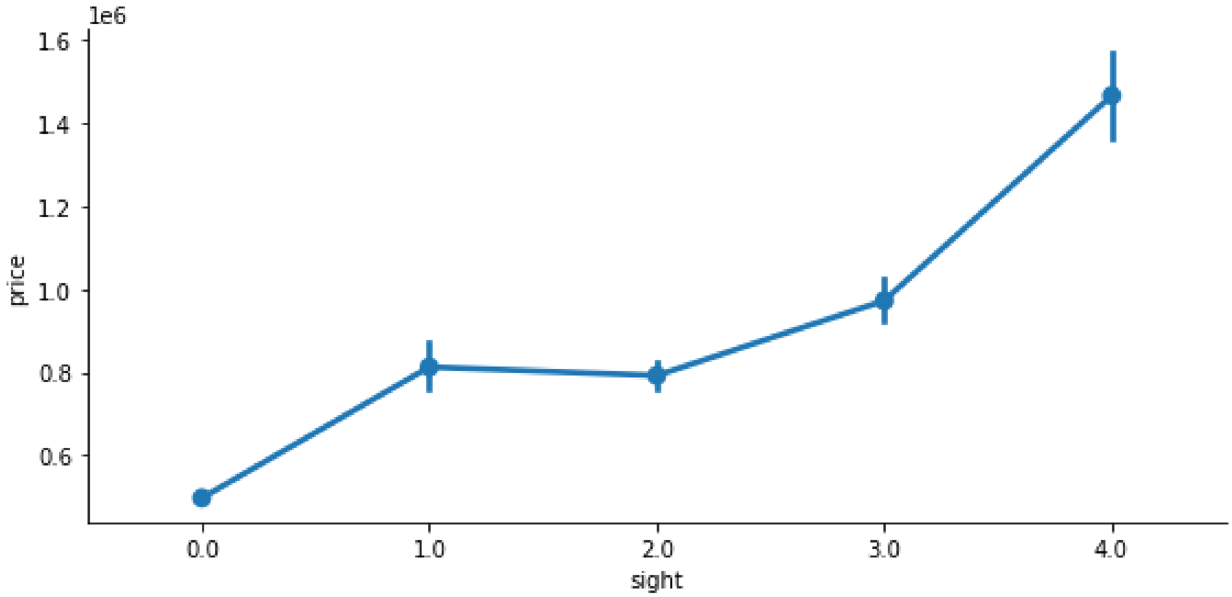
Analyzing Bivariate for Feature: coast



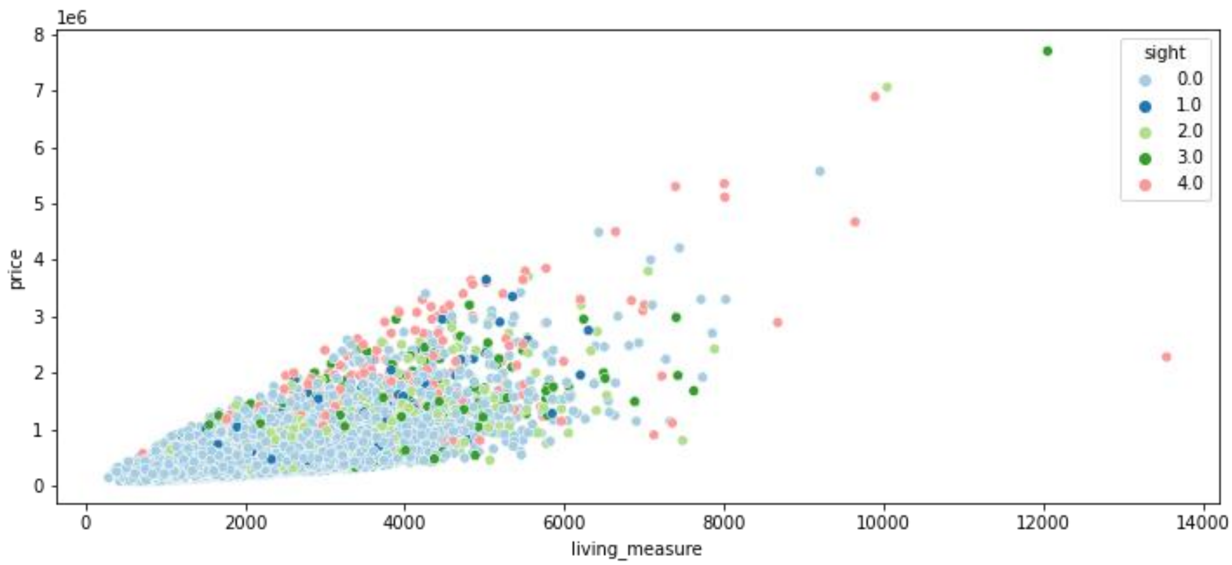
- The house properties with water_front tend to have higher price compared to that of non-water_front properties

Analyzing Bivariate for Feature: sight

price			living_measure			
sight	**mean**	**median**	**size**	**mean**	**median**	**size**
0	**4.97E+05**	**432500**	**19489**	**1997.762**	**1850**	**19489**
1	**8.13E+05**	**690944**	**332**	**2568.961**	**2420**	**332**
2	**7.93E+05**	**675000**	**963**	**2655.258**	**2470**	**963**
3	**9.72E+05**	**802500**	**510**	**3018.565**	**2840**	**510**
4	**1.46E+06**	**1190000**	**319**	**3351.473**	**3050**	**319**

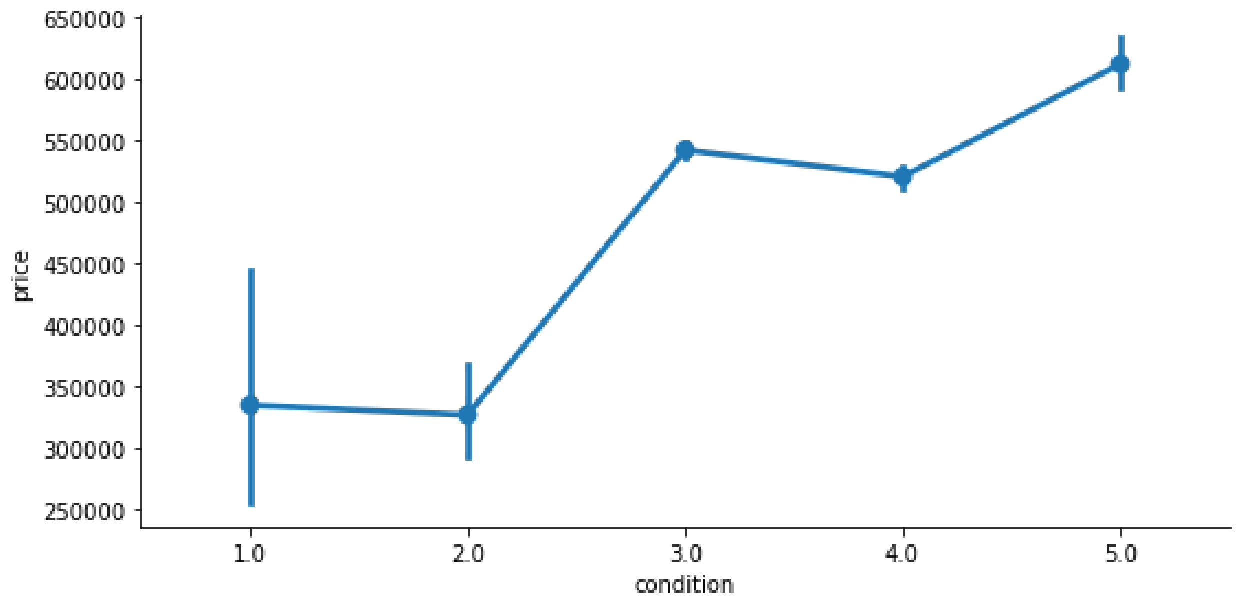


- Properties with higher price have more no.of sights compared to that of houses with lower price

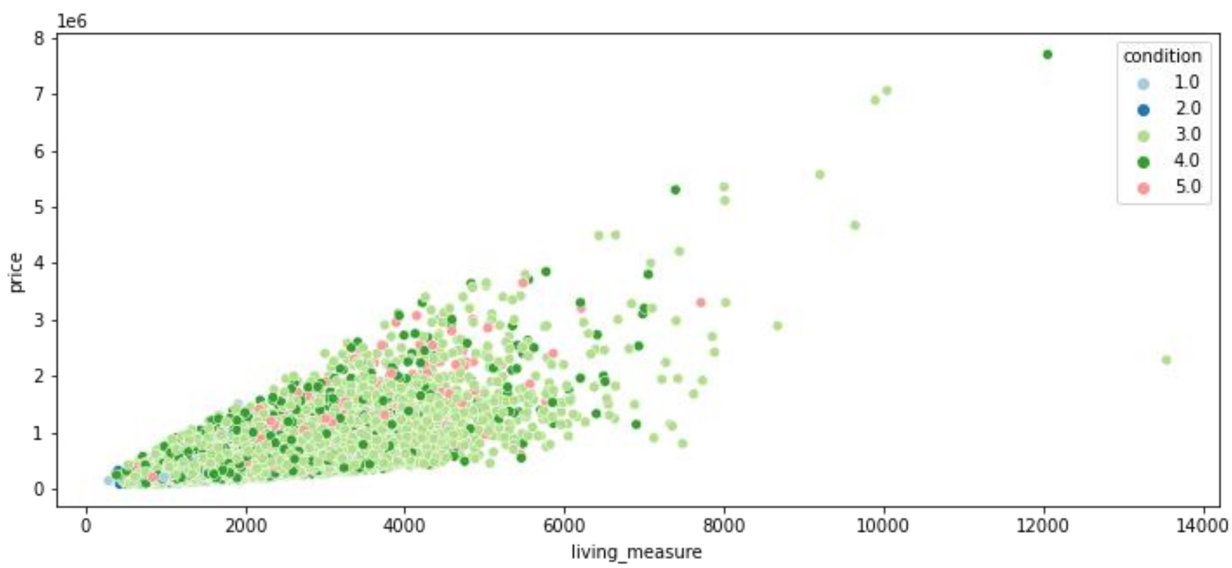


- The above graph also justify that Properties with higher price have more no.of sights compared to that of houses with lower price

Analyzing Bivariate for Feature: condition



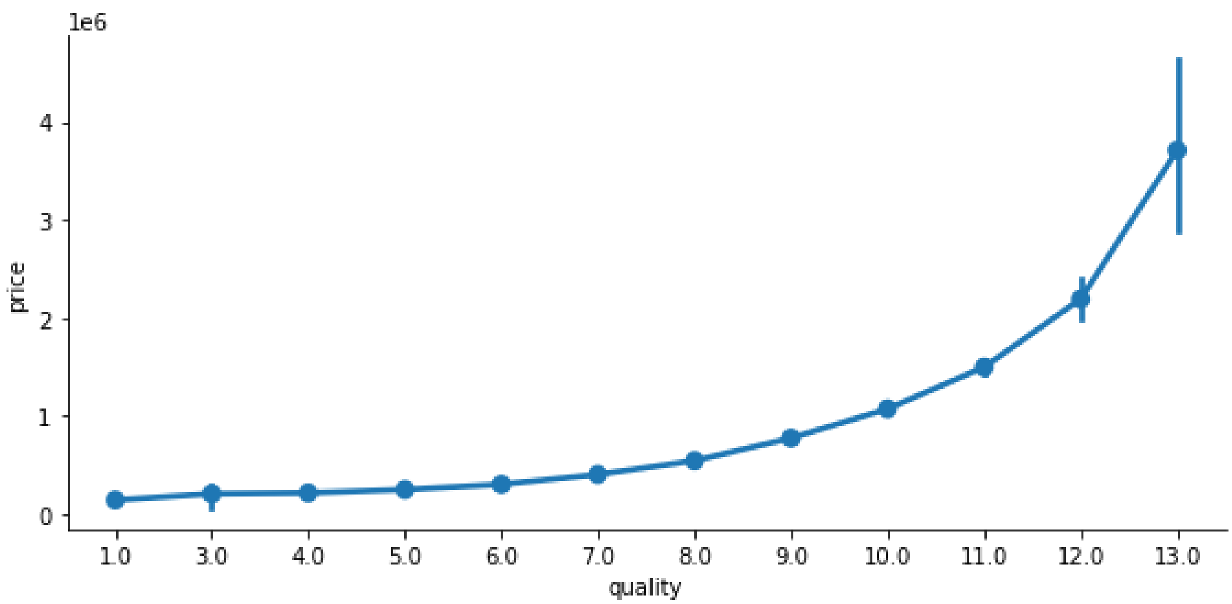
- The price of the house increases with condition rating of the house



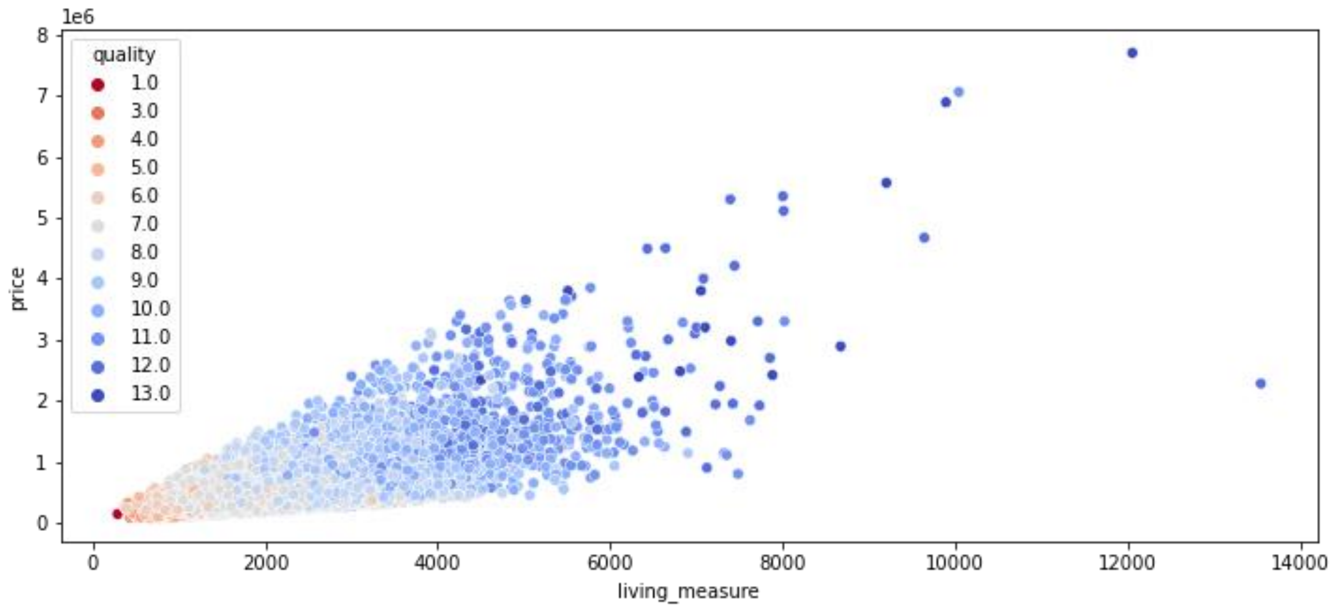
- *So, we found out that smaller houses are in better condition and better condition houses are having higher prices*

Analyzing Bivariate for Feature: quality

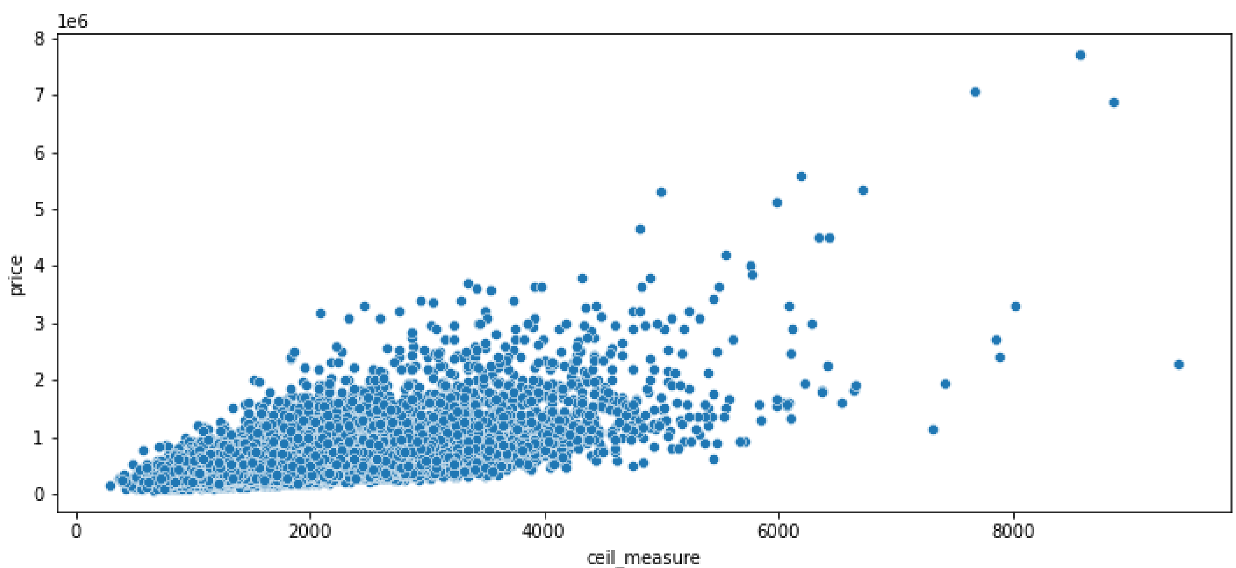
	price			living_measure		
quality	mean	median	size	mean	median	size
1	1.42E+05	142000	1	290	290	1
3	2.06E+05	262000	3	596.6667	600	3
4	2.14E+05	205000	29	660.4828	660	29
5	2.49E+05	228700	242	983.3264	905	242
6	3.02E+05	275276.5	2038	1191.797	1120	2038
7	4.03E+05	375000	8982	1689.456	1630	8982
8	5.43E+05	510000	6067	2183.523	2150	6067
9	7.74E+05	720000	2615	2865.406	2820	2615
10	1.07E+06	914327	1134	3520.3	3450	1134
11	1.50E+06	1280000	399	4395.449	4260	399
12	2.19E+06	1820000	90	5471.589	4965	90
13	3.71E+06	2980000	13	7483.077	7100	13



- There is clear increase in price of the house with higher rating on quality

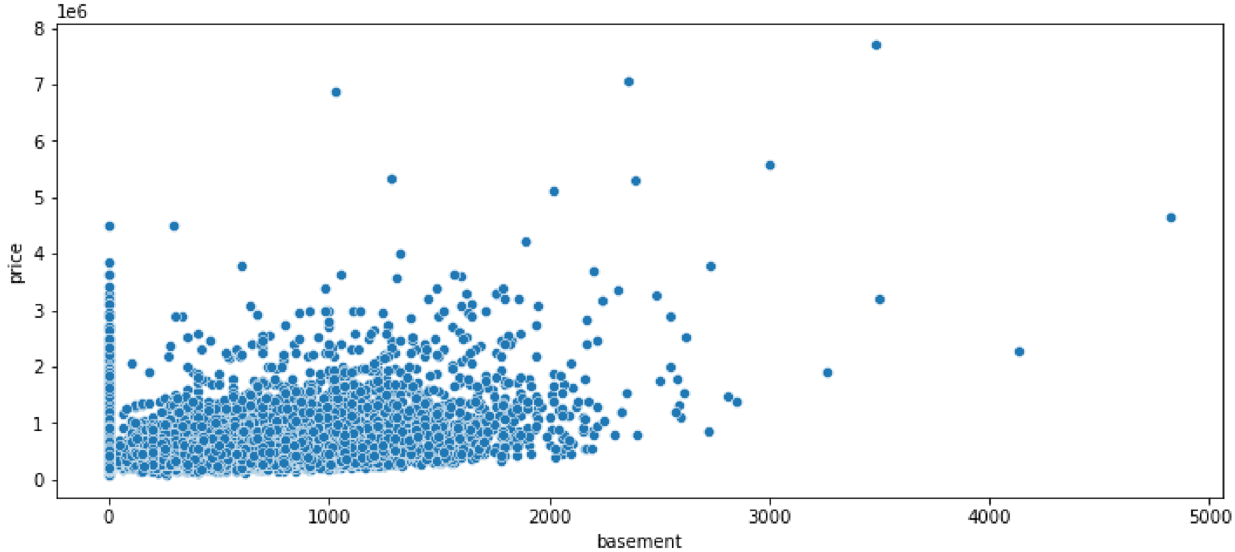


Analyzing Bivariate for Feature: ceil_measure

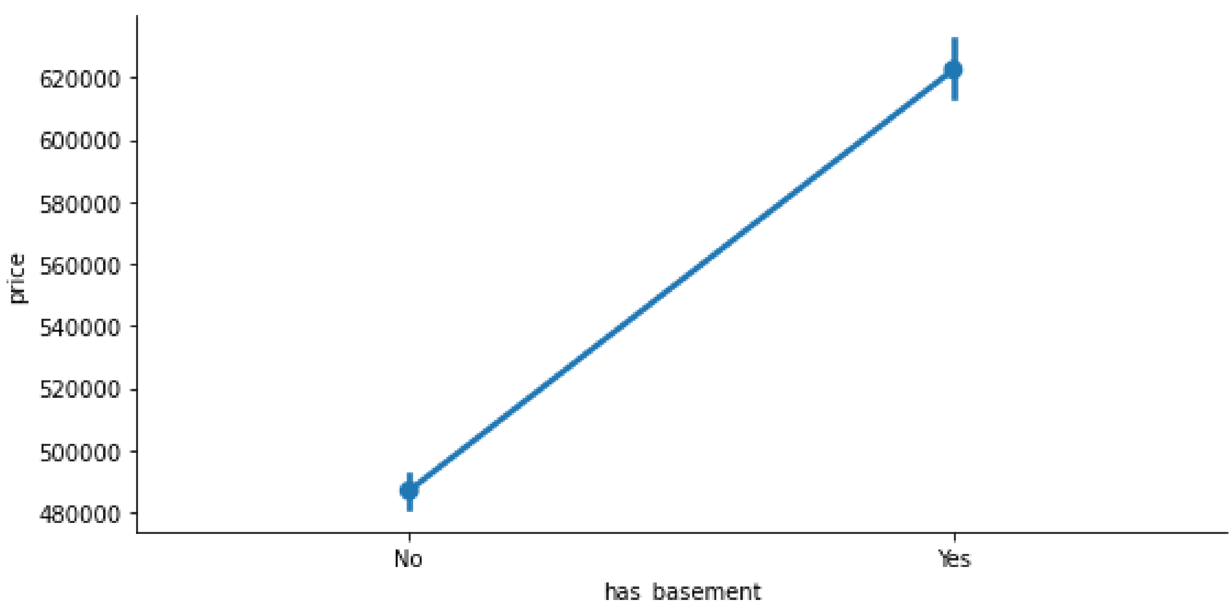


- There is upward trend in price with ceil_measure

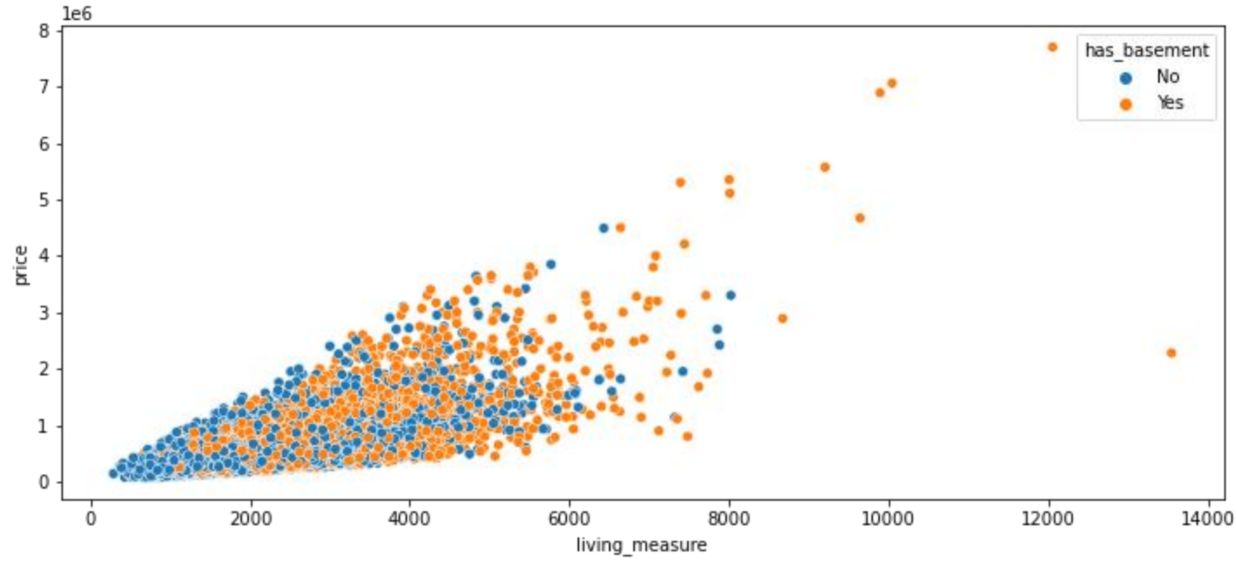
Analyzing Bivariate for Feature: basement



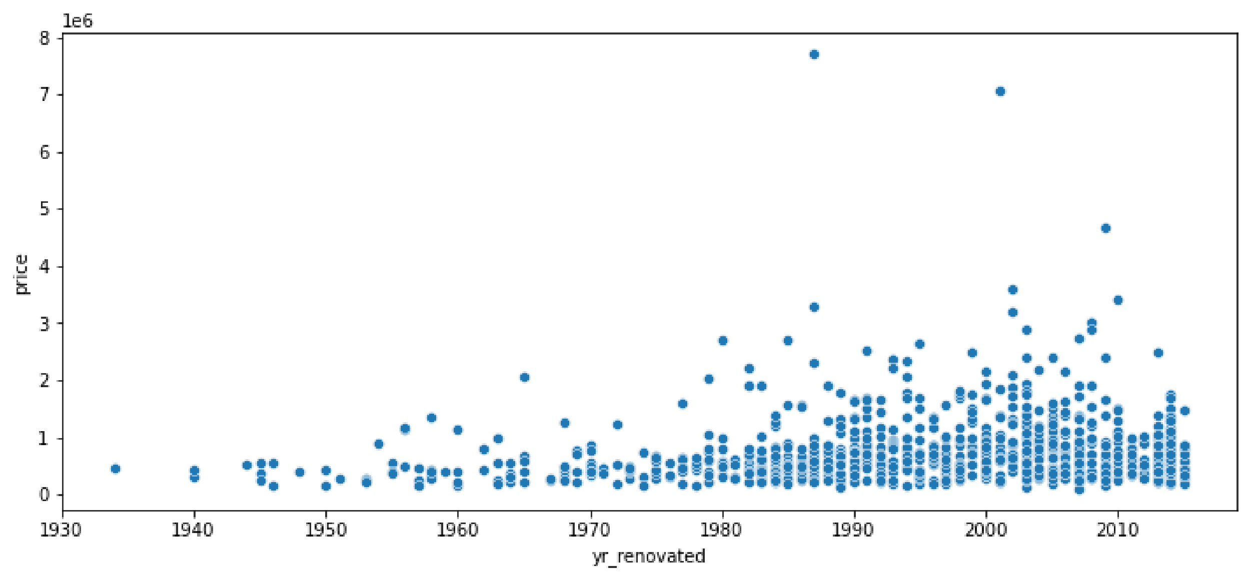
- We will create the categorical variable for basement 'has_basement' for houses with basement and no basement. This categorical variable will be used for further analysis



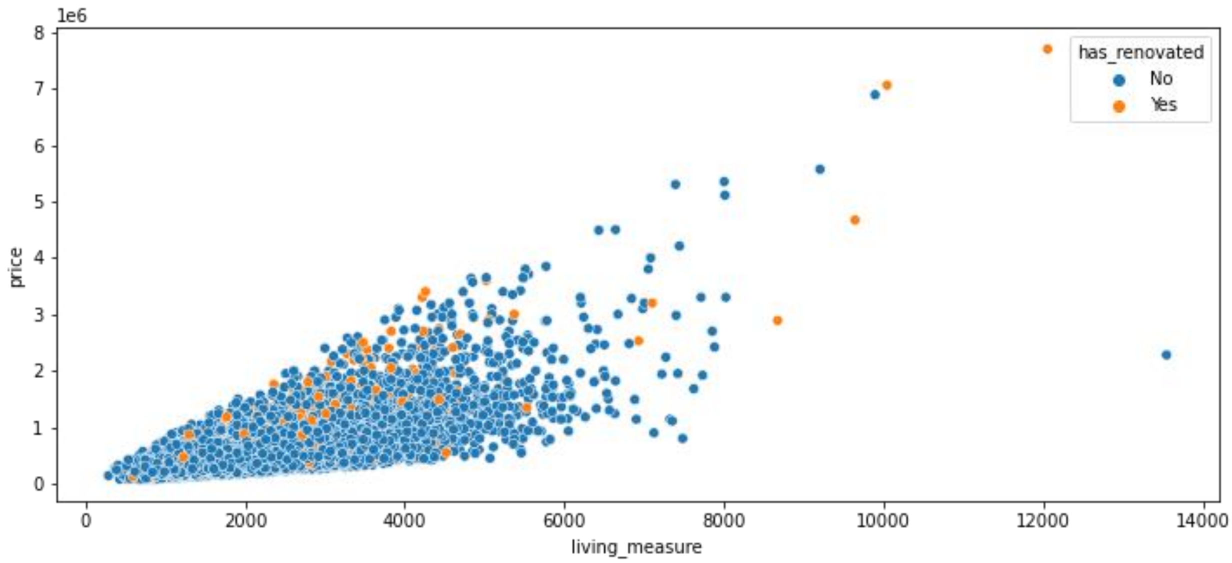
- The houses with basement have better price compared to that of houses without basement



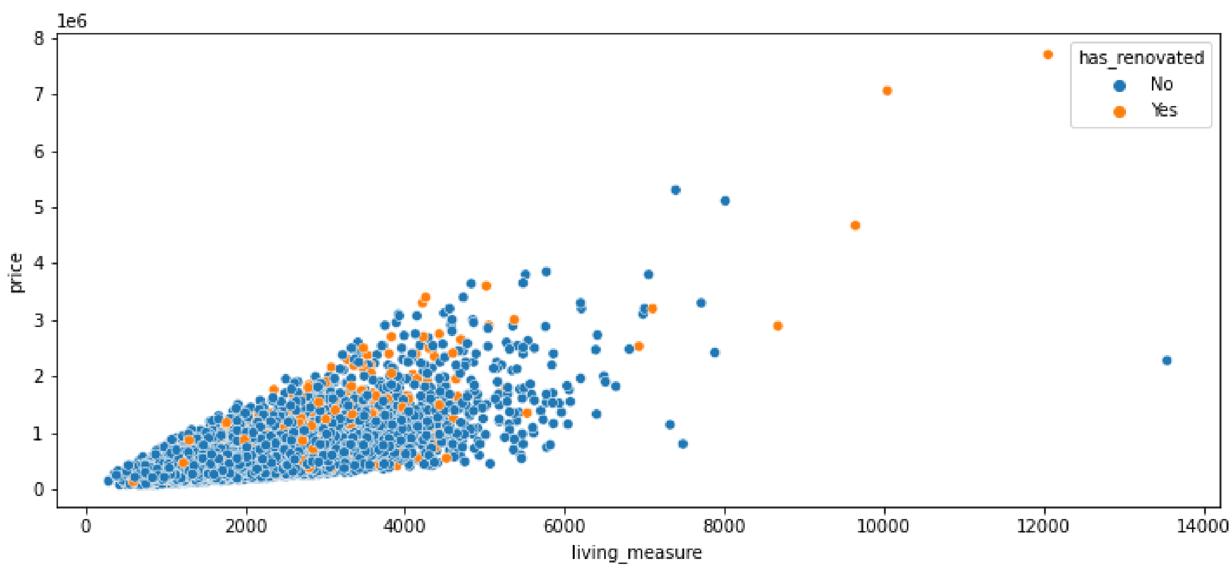
Analyzing Bivariate for Feature: yr_renovated:



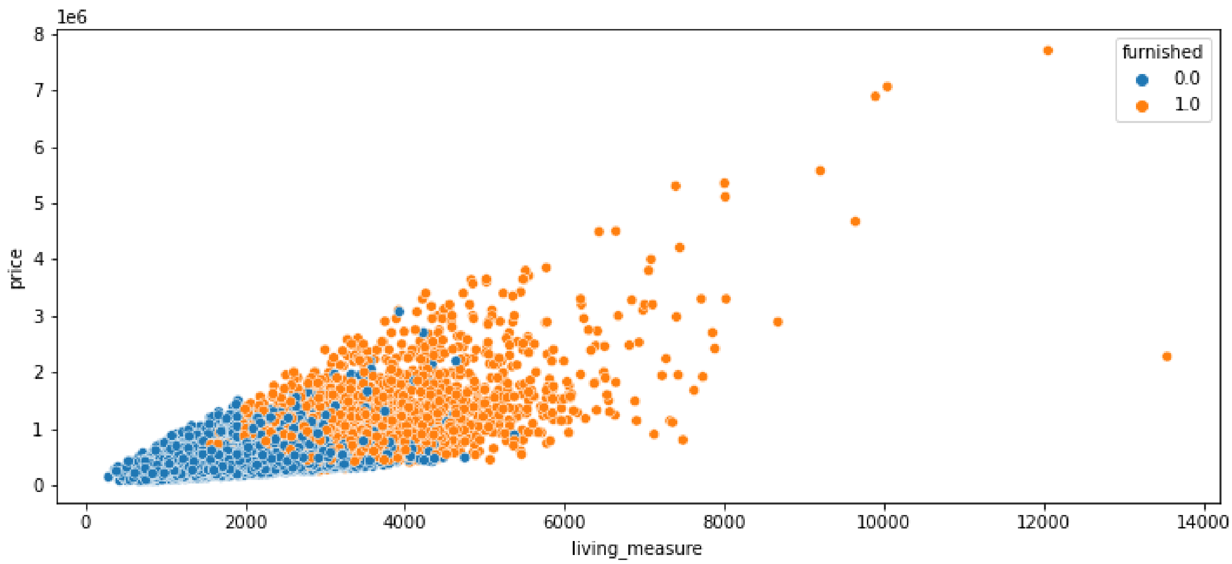
- So, most houses are renovated after 1980's. We will create new categorical variable 'has_renovated' to categorize the property as renovated and non-renovated. For further analysis we will use this categorical variable.



- Renovated properties have higher price than others with same living measure space.



Analyzing Bivariate for Feature: furnished

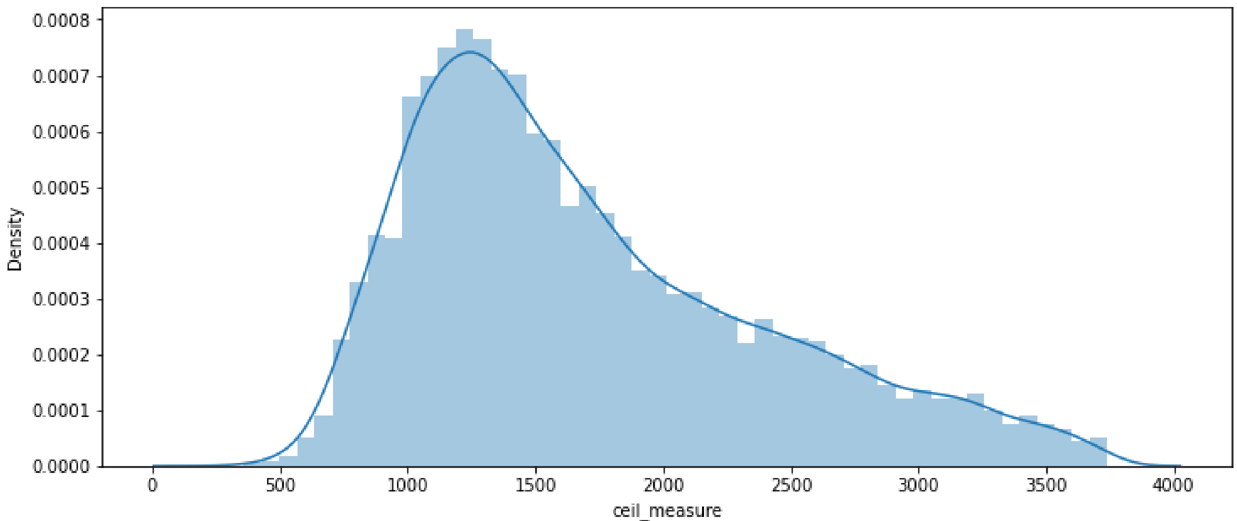


- Furnished houses have higher price than that of the Non-furnished houses

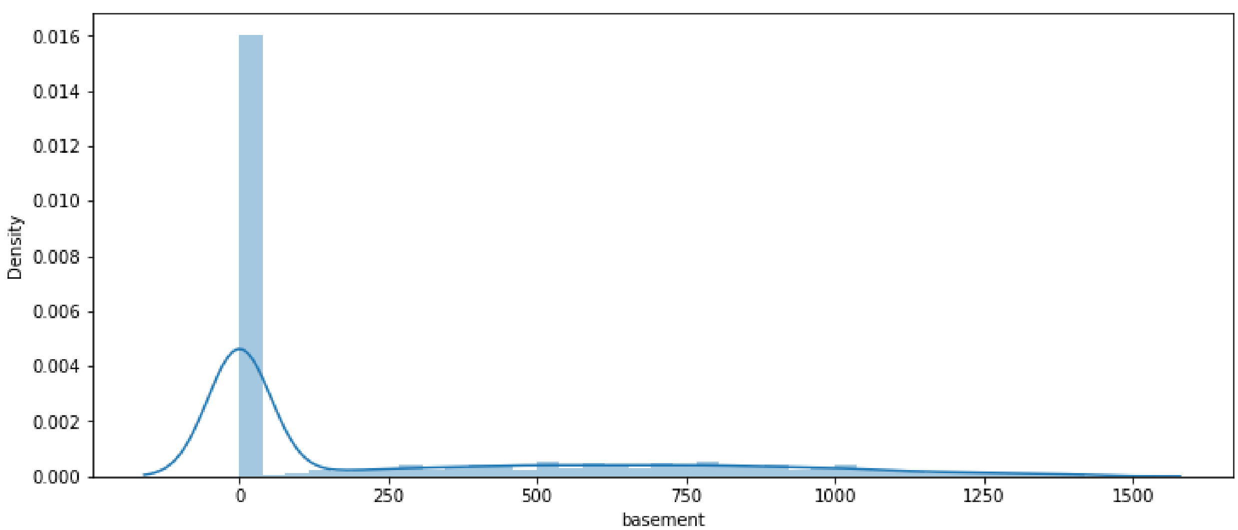
DATA PROCESSING

- **We got 611 records which are outliers**

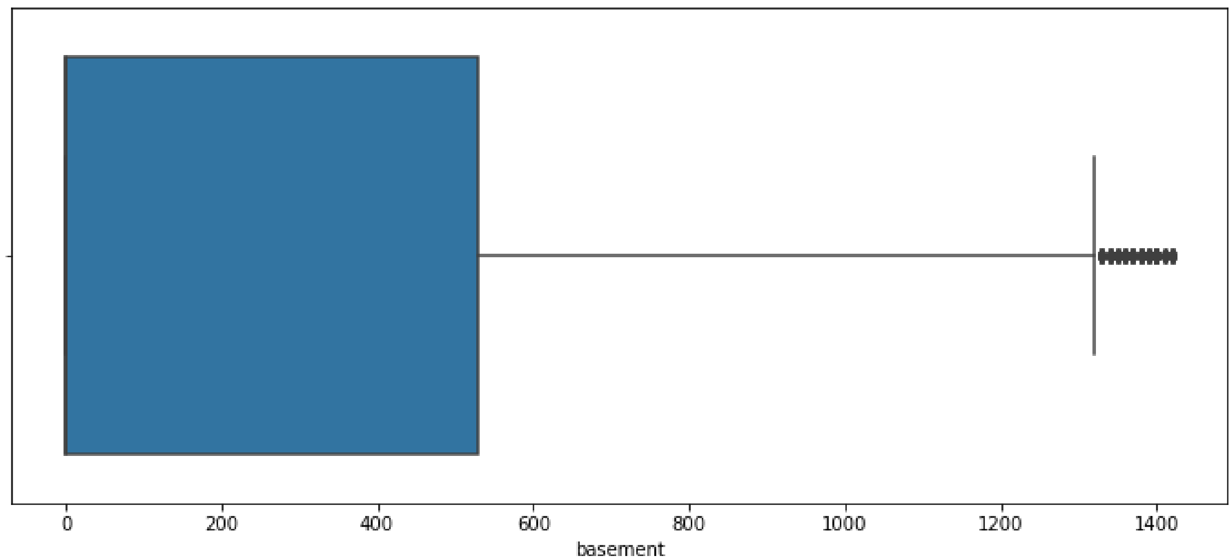
After Treating Outliers:

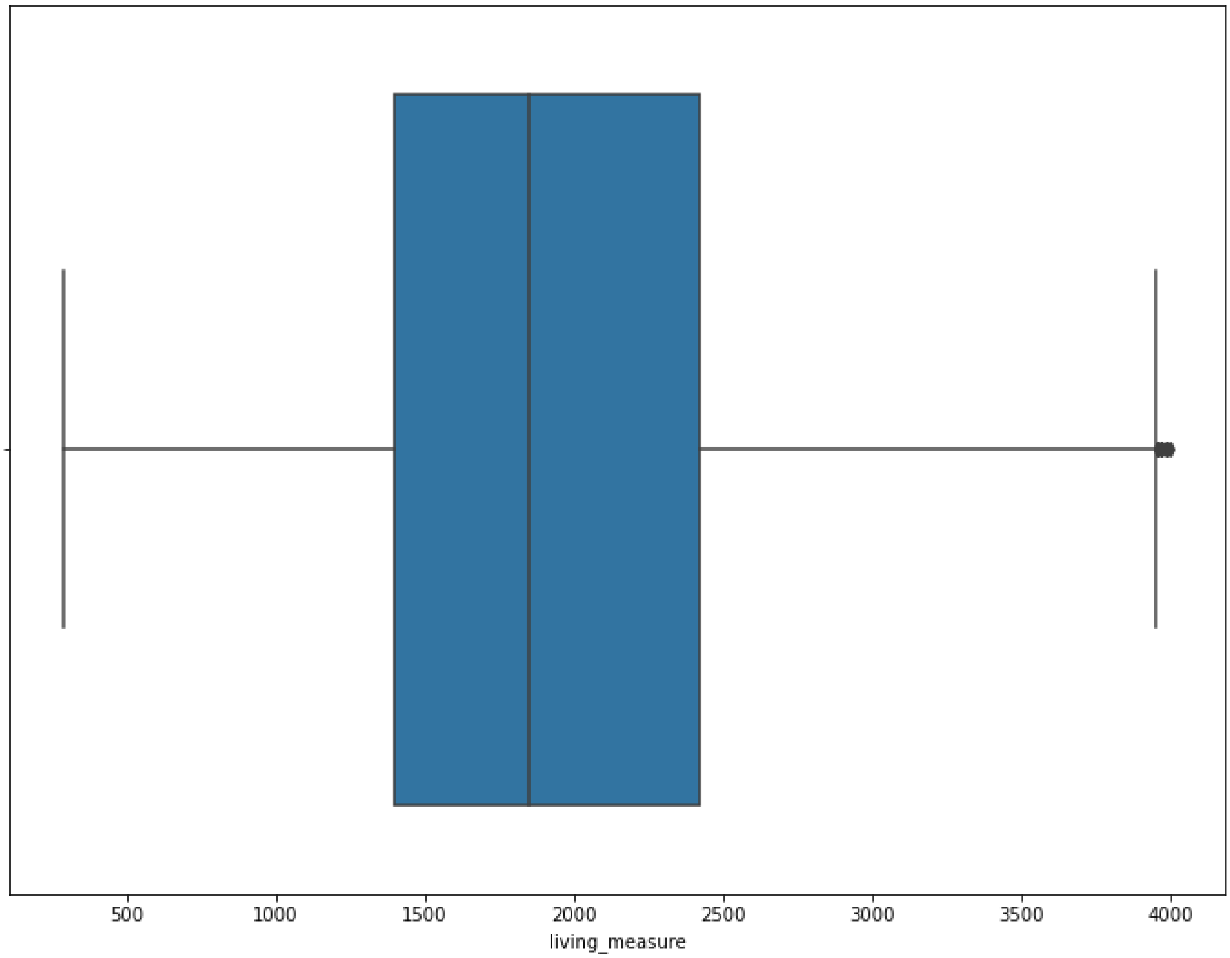


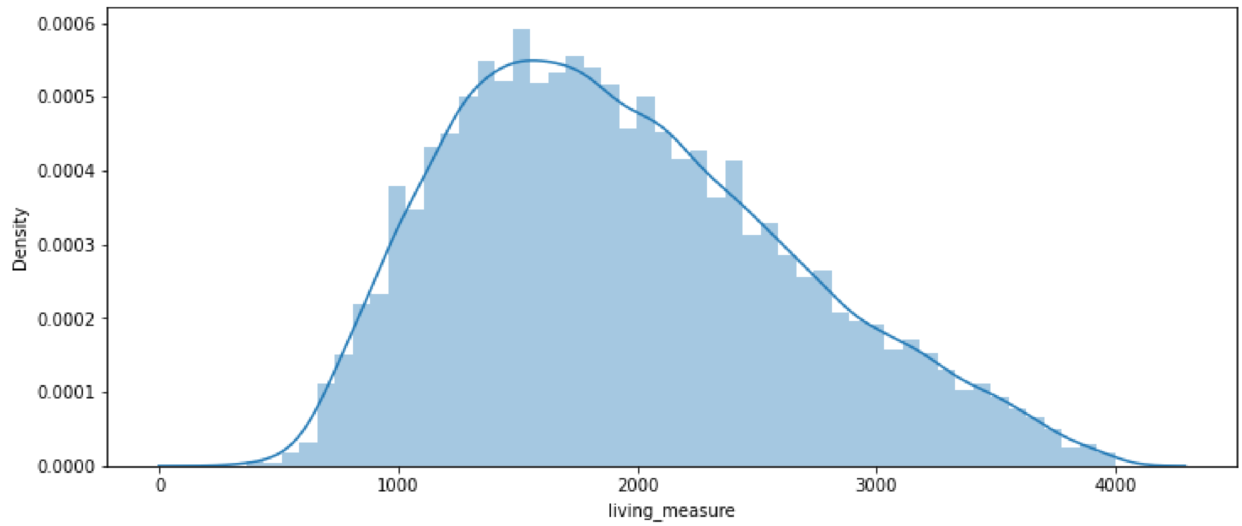
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
Density
0
500
1000
1500
2000
2500
3000
3500
4000
ceil_measure

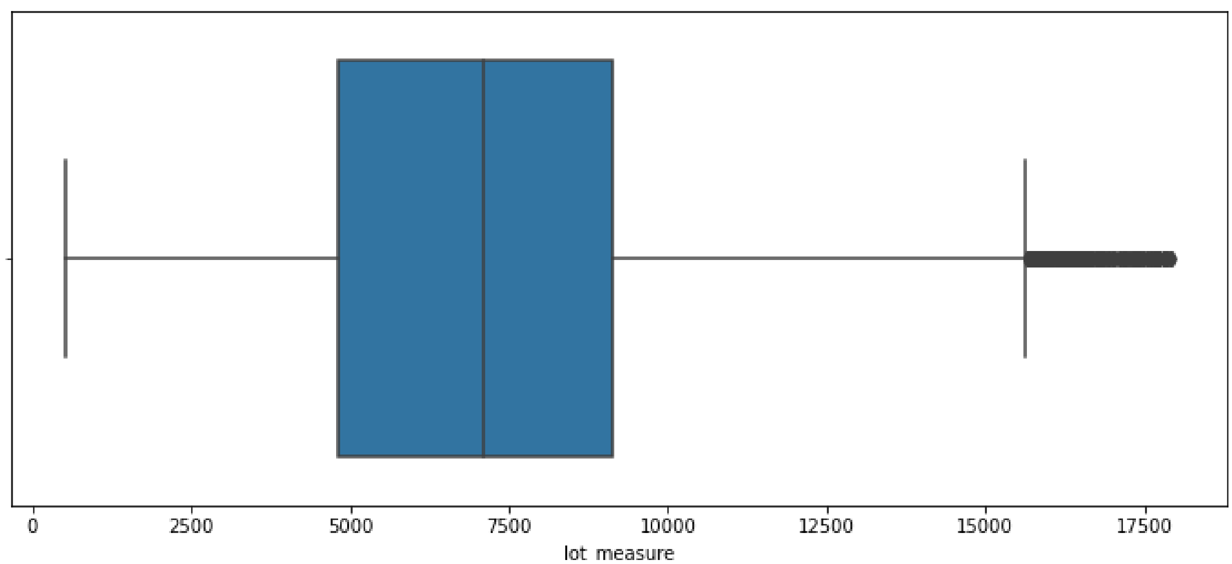


0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
0
250
500
750
1000
1250
1500
basement


0
200
400
600
800
1000
1200
1400
basement


500
1000
1500
2000
2500
3000
3500
4000
living_measure





Business insights:

- After Cleaning Data Is stable for model building.
- most houses are renovated after 1980's. We will create new categorical variable 'has_renovated' to categorize the property as renovated and non-renovated. For further analysis we will use this categorical variable.

- There is upward trend in price with ceil_measure.
- we found out that smaller houses are in better condition and better condition houses are having higher prices
- Properties with higher price have more no.of sights compared to that of houses with lower price