HOUSE PRICE PREDICTION

CAPSTONE PROJECT – TEAM 4

MILESTONE 1 REPORT

Contents

- Problem Statement
- Data Information
- Data Preprocessing: Features datatype, Compare Result
- Descriptive Statistics
- Exploratory Data Analysis
- Analytics approaches that may see fit to be applied to the problem.

PROBLEM STATEMENT

Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad.

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 neighbourhood and based on gathered data you will try to assess your house price.

PROBLEM STATEMENT

Objective

The objective of this problem statement is to predict the housing prices of a town, or a suburb based on the features of the locality provided and identify the most important features to consider while predicting the prices. The goal is to build machine learning models that accurately predicts house prices and recommend the best model. This will assist both buyers and sellers to understand the value of a house. This objective has been broken down into milestones to mark progress and help ensure that the project is on track.

Milestone 1

Under milestone 1, the whole exercise is to explore the dataset. Check the number of rows, columns, variable information, and descriptive statistics. Perform necessary data cleaning such as removing unwanted variables, handling missing values, outliers, and creating new variables. Do the exploratory data analysis using visualization to understand the distribution and spread for continuous and categorical variables , and the relationships between different variables.

Potential business and social opportunities that addressing this problem may offer:

- Buyers can make more informed decisions about purchasing a house by having a reliable estimate of its price. This helps them avoid overpaying or missing out on good deals
- Sellers can benefit by setting optimal prices for their houses. Knowing the key factors that influence house prices will help them to competitively position their homes in the market.
- Financial Institutions can use the recommended model to assess houses values for loan approvals. This can also help them to manage credit risk

DATA INFORMATION

The dataset consists of 21613 houses located at Seattle of Washington DC of USA. The houses have 23 different features will help determine optimal prices for individual house. The target variable is the price of the house. The data collection process for this dataset is likely to be involved in a blend of real estate transactions, property assessments, public records, and possibly information provided by house owners.

FEATURES	DESCRIPTIONS
cid	a notation for a house
dayhours	Date house was sold
price	Price is prediction target (in $)
room_bed	Number of Bedrooms per house
room_bath	Number of bathrooms per bedrooms
living_measure	square footage of the home
lot_measure	square footage of the lot
ceil	Total floors (levels) in house
coast	House which has a view to a waterfront (0 - No, 1 - Yes)
sight	Has been viewed
condition	How good the condition is (Overall out of 5)
quality	grade given to the housing unit, based on grading system
ceil_measure	square footage of house apart from basement
basement	square footage of the basement
yr_built	Built Year
yr_renovated	Year when house was renovated
zipcode	zip code
lat	Latitude coordinate
long	Longitude coordinate
living_measure15	Living room area in 2015 (implies-- some renovations) This might or might not have affected the lot size area
lot_measure15	lotSize area in 2015 (implies-- some renovations)
furnished	Based on the quality of room (0 - No, 1 - Yes)
total_area	Measure of both living and lot

DATA RE – PROCESSING FOR EDA

- There were no null and duplicate values in the dataset.

- Columns ceil, coast, condition, yr_built, long, and total_area had '$' as part of their values, which distorted their data type. These '$' were removed and replaced with NaN. Following this replacement, the affected columns now exhibit missing values.

- There were few missing values in 17 columns. We imputed the missing values using the median and the mean.
  - Columns with outliers, the impute strategy was median
  - Columns with no outliers, the impute strategy was the mean.

Two new columns were created:

- yr_before_sold column – This was created from Dayhour and yr_built. This indicate the numbers of between the year a house was built and the year a house was sold or bought. In this case, I can drop Dayhour and yr_built columns in linear regression model as they might not be significant.

- Renovation_status column – Created from yr_renovated. Zero values were categorized as 'notrenovated' and Non-Zero values (i.e. rows with years) were as 'renovated'. yr_renovated column could be dropped when building linear regression model.

- Houses with no bedrooms and/or no bathrooms were dropped. This reduced the houses in the dataset from 21613 to 21597 for further analysis. 16 houses were dropped. The number of bathrooms were in decimals (float), so they were rounded up for easy use for analysis.

- There were few outliers in the dataset. However, they were not treated because they are all proper values. Also, In the real world many outliers exist, and these outliers should be incorporated in the model building for optimal predictions.

FEATURES(COLUMNS) DATATYPE

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 23 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   cid                 21613 non-null  int64  
 1   dayhours            21613 non-null  object 
 2   price               21613 non-null  int64  
 3   room_bed            21505 non-null  float64
 4   room_bath           21505 non-null  float64
 5   living_measure      21596 non-null  float64
 6   lot_measure         21571 non-null  float64
 7   ceil                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: float64(18), int64(4), object(1)
memory usage: 3.8+ MB
```

- After fixing the columns, we have 1 object (dayhours), and the rest (22) are numeric variables

Compare results - Missing values (Original Date) and Missing Values (After removing $)

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

- Features with Zero (o), have no missing values

DESCRIPTIVE STATISTICS

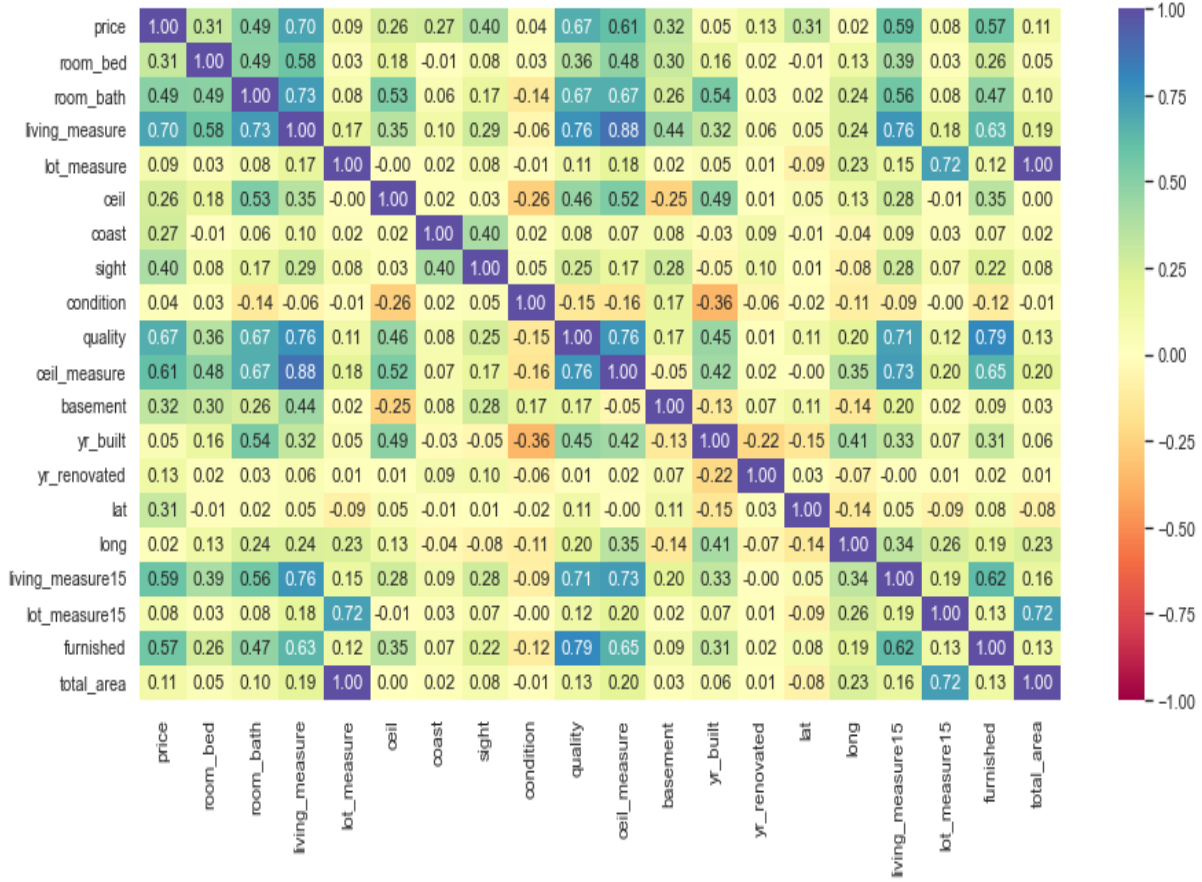
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
cid	21613.000	NaN	NaN	NaN	4580301520.865	2876565571.312	1000102.000	2123049194.000	3904930410.000	7308900445.000	9900000190.000
dayhours	21613	372	20140623T000000	142	NaN	NaN	NaN	NaN	NaN	NaN	NaN
price	21613.000	NaN	NaN	NaN	540182.159	367362.232	75000.000	321950.000	450000.000	645000.000	7700000.000
room_bed	21505.000	NaN	NaN	NaN	3.371	0.930	0.000	3.000	3.000	4.000	33.000
room_bath	21505.000	NaN	NaN	NaN	2.115	0.770	0.000	1.750	2.250	2.500	8.000
living_measure	21596.000	NaN	NaN	NaN	2079.861	918.496	290.000	1429.250	1910.000	2550.000	13540.000
lot_measure	21571.000	NaN	NaN	NaN	15104.583	41423.619	520.000	5040.000	7618.000	10684.500	1651359.000
ceil	21541.000	NaN	NaN	NaN	1.494	0.540	1.000	1.000	1.500	2.000	3.500
coast	21582.000	NaN	NaN	NaN	0.007	0.086	0.000	0.000	0.000	0.000	1.000
sight	21556.000	NaN	NaN	NaN	0.234	0.766	0.000	0.000	0.000	0.000	4.000
condition	21528.000	NaN	NaN	NaN	3.409	0.651	1.000	3.000	3.000	4.000	5.000
quality	21612.000	NaN	NaN	NaN	7.657	1.175	1.000	7.000	7.000	8.000	13.000
ceil_measure	21612.000	NaN	NaN	NaN	1788.367	828.103	290.000	1190.000	1560.000	2210.000	9410.000
basement	21612.000	NaN	NaN	NaN	291.523	442.581	0.000	0.000	0.000	560.000	4820.000
yr_built	21598.000	NaN	NaN	NaN	1971.009	29.373	1900.000	1951.000	1975.000	1997.000	2015.000
yr_renovated	21613.000	NaN	NaN	NaN	84.402	401.679	0.000	0.000	0.000	0.000	2015.000
zipcode	21613.000	NaN	NaN	NaN	98077.940	53.505	98001.000	98033.000	98065.000	98118.000	98199.000
lat	21613.000	NaN	NaN	NaN	47.560	0.139	47.156	47.471	47.572	47.678	47.778
long	21579.000	NaN	NaN	NaN	-122.214	0.141	-122.519	-122.328	-122.230	-122.125	-121.315
living_measure15	21447.000	NaN	NaN	NaN	1987.066	685.520	399.000	1490.000	1840.000	2360.000	6210.000
lot_measure15	21584.000	NaN	NaN	NaN	12766.543	27286.987	651.000	5100.000	7620.000	10087.000	871200.000
furnished	21584.000	NaN	NaN	NaN	0.197	0.398	0.000	0.000	0.000	0.000	1.000
total_area	21545.000	NaN	NaN	NaN	17192.042	41628.688	1423.000	7032.000	9575.000	13000.000	1652659.000

Descriptive statistics help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The recognized types of descriptive statistics for this problem are measures of centre and other parameters listed: the mean, median, mode, min, max, unique, 75% quartiles and 25% quartiles. NaN shows that the values cannot be calculated for that variables.

Inference

- The highest price of the houses is USD 7,700,000.00, and 75% of the houses were priced for USD 645,000.00
- Some of the houses have either no bedrooms or bathrooms or both. The maximum number of bedrooms for a house is 33 and the maximum number of bathrooms for a house is 8
- The total_area column is the sum of the living_measure and the lot_measure. The minimum and maximum house area in the dataset is 1423 square feet and 1,652,659 square feet respectively
- Ceil - The houses floors range from 1 to 4. 75% of the houses have 1 or 2 floors.
- The number of times the houses were viewed by potential buyers range from 0 to 4. No view could indicate low interest. 75% of the houses were not viewed for consideration.
- conditions of the houses graded from 1 to 5. low grade indicate poor state of the house and high grade (4 or 5) indicate good conditions.
- The quality of the houses graded from 1 to 13. A house graded as 13 indicate that that might have been built with quality materials and with top-notch design.
- ceil_measure - The houses floor (apart from basement) sizes range from 290 to 9,410 square feet.
- Basement - about 50% or more houses have not basement. For the houses with basement, the largest area is 4,820 square feet.
- yr_renovated - variables in this column comprises: Zeros and Years. Houses with Zeros could suggest that those houses were not renovated and houses with years may indicate renovation.
- Lat and long - all the houses can be located on the world map within geometry coordinates (47.78, -122.52)
- furnished - 75% or more houses in the dataset were not furnished.

EXPLORATORY DATA ANALYSIS RESULTS



Correlation Plot:

- Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated, respectively. Correlation values near to 0 are not correlated to each other
- Observing the plot, numerous variables display minimal correlation or no correlation with each other
- The target variable (price) is correlated to the following variables: living measure, quality, ceil_measure, living measure15 and furnished.
- The highest correlation exist lot_measure and total_area in the plot. Possibly, total_area will be dropped when building our linear regression model as it will affect the performance of the model – collinearity.
- Lat and long variables have no correlation with other different variables.
- Furnished is highly correlated to quality. This indicates that the quality of a house determines its level of furnishing.

EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

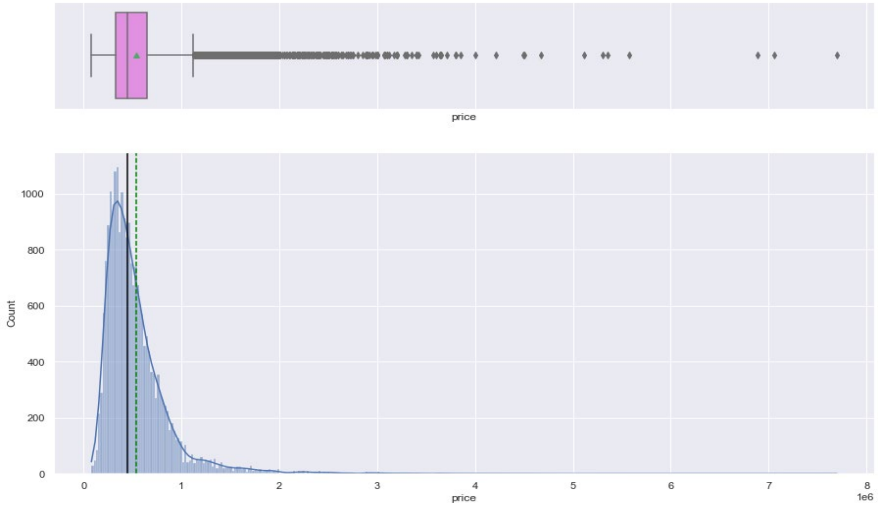
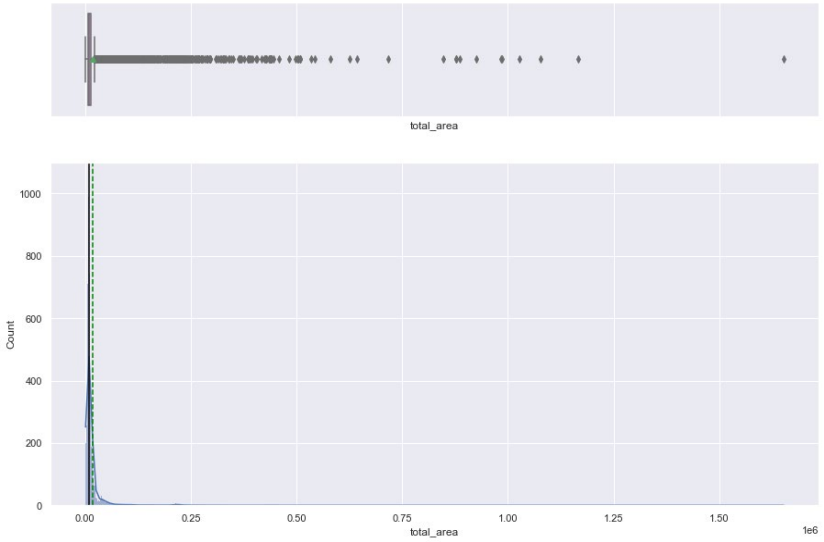


Fig – Prices of Houses

- The price distribution is skewed to the right with many high outliers. This suggests that more than 50% of the houses have lower prices.
- About 5.4% of houses (1,158) are identified as having high prices, exceeding $1,129,500.00.

- The total_area distribution is skewed to the right with many high outliers. This suggests that more than 50% of the houses have low square footage areas.
- About 11% of houses (2,406) are identified as having large area, exceeding 21,949.5 square footage area.
- The total area has no correlation with house price, lets check the distribution for lot_measure and living_measure

Fig – Total area of the house in square footage



EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

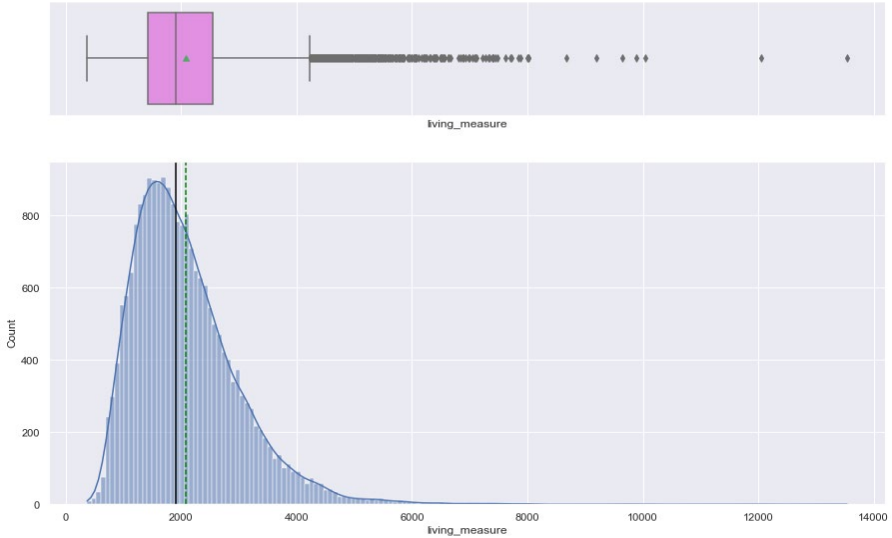
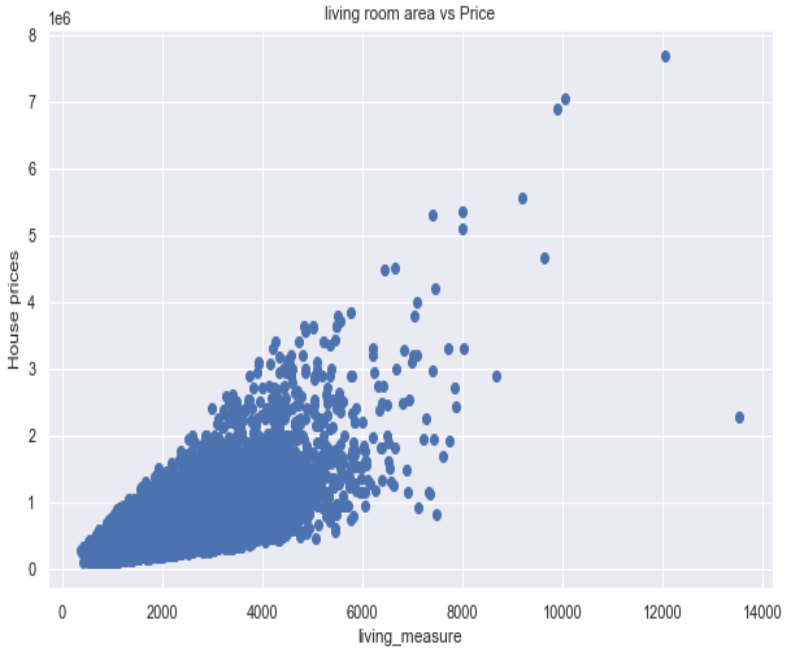


Fig – living_measure

- The distribution of the house living area is skewed to the right with few high outliers.
- 50% and slightly more houses have lower living area.

- There is a high correlation concentration upto a 6000 square footage area, and the plot spreads beyond 6000 sqft area

Fig – correlation – living_measure vs price



EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

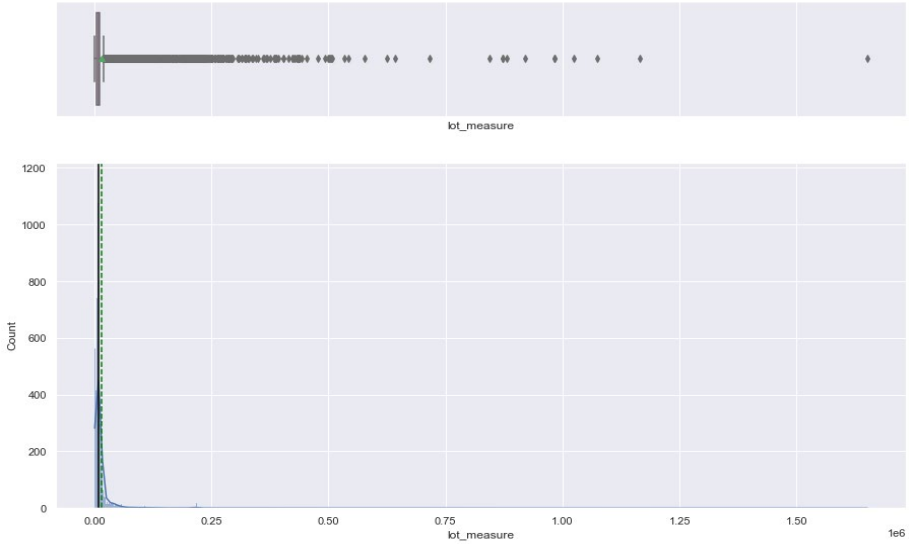
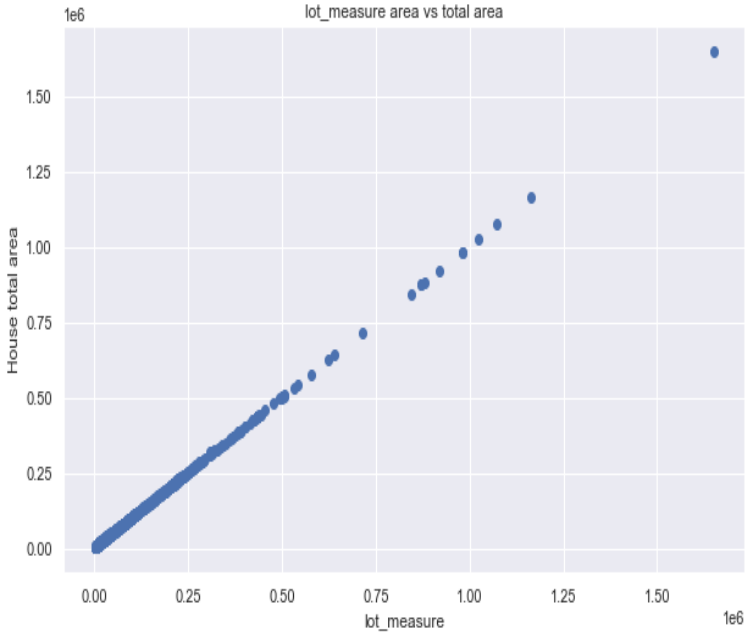


Fig – lot_measure

- total_area and lot_measure appear to have the same distribution - that is, right skewed distribution with high outliers.
- This suggest a perfect correlation between them as shown in the correlation heatmap - There is collinearity

- All the points are on the straight line, indicating a perfect correlation - presence of collinearity

Fig – correlation – lot_measure vs Total_area



EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

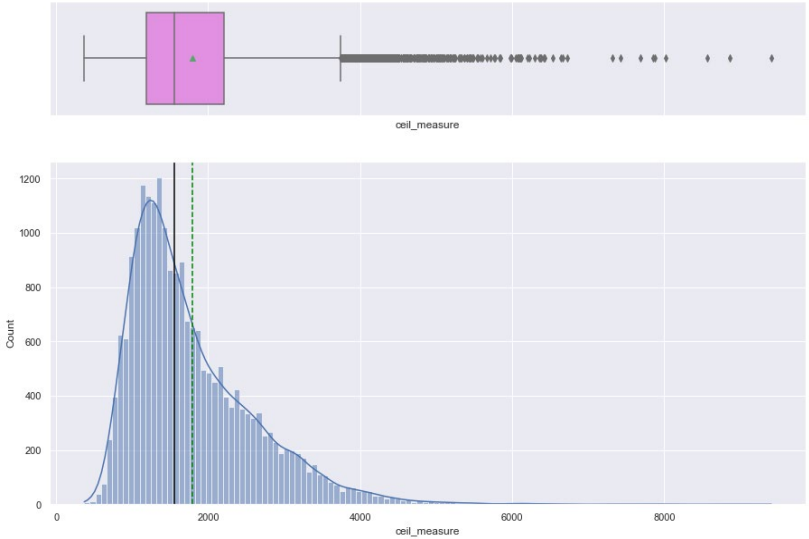
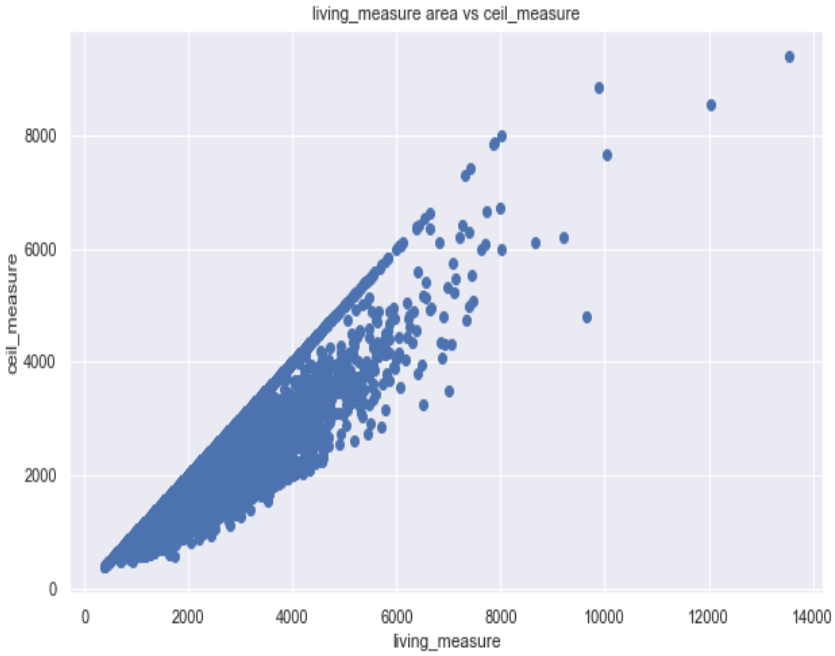


Fig – ceil_measure

- Out of 21597, 610 houses have large square footage areas apart from basement
- comparatively, about 97% of the houses have quite low square footage area.
- living_measure and ceil_measure appear to have similar distribution, suggesting collinearity

- Living_measure and ceil_measure are correlated

Fig – correlation – ceil_measure vs living_measure



EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

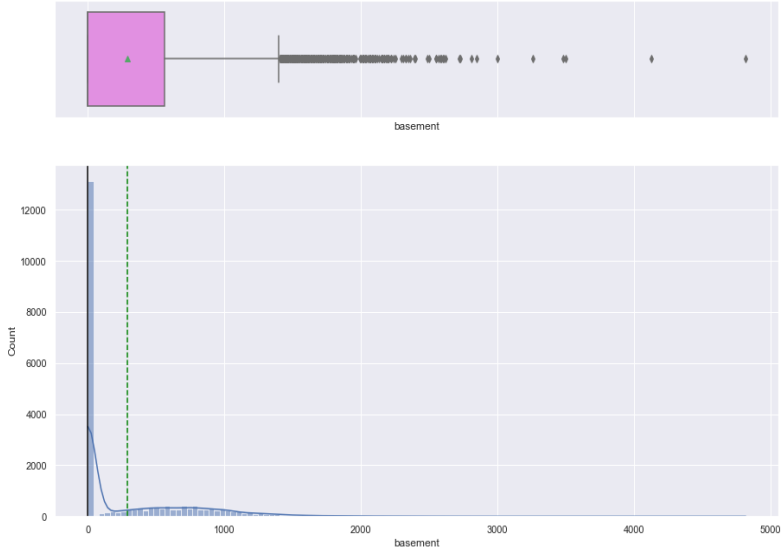
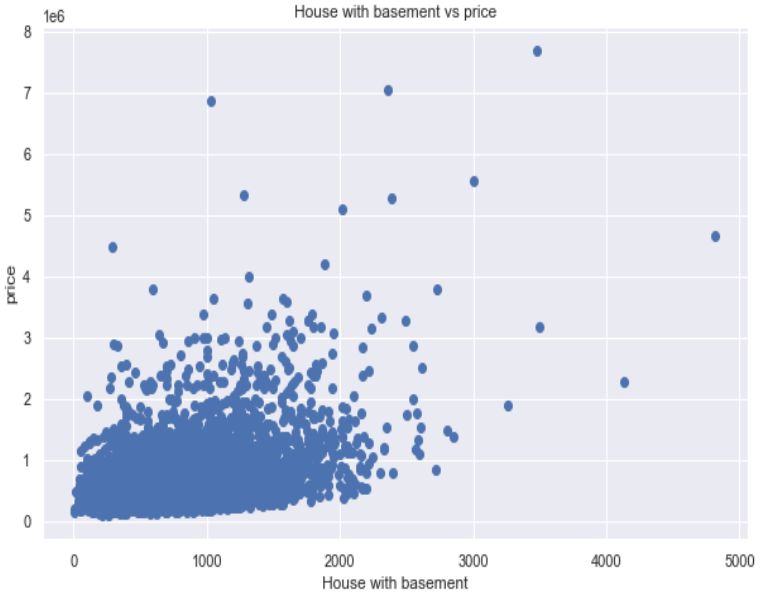


Fig – basement

- More that 50% of the houses have no basement. However, there are few houses with large square footage area of basement.

- There is no correlation.

Fig – correlation – price vs houses with basement



EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

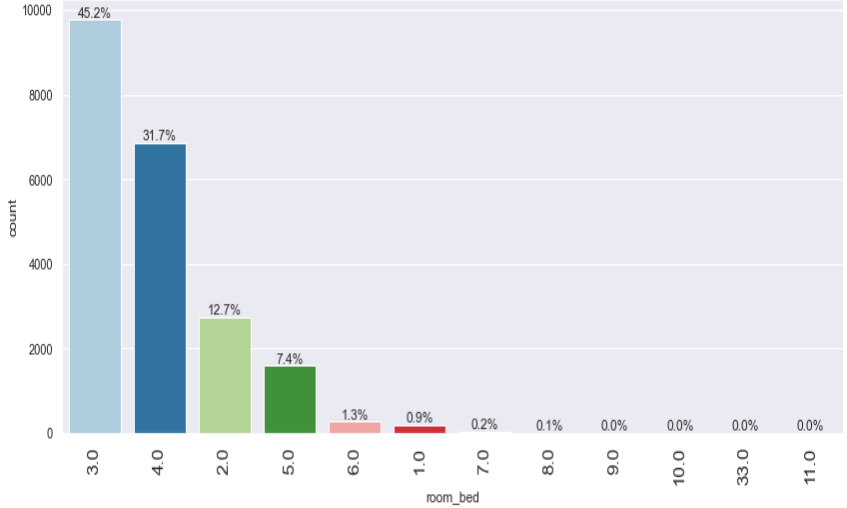
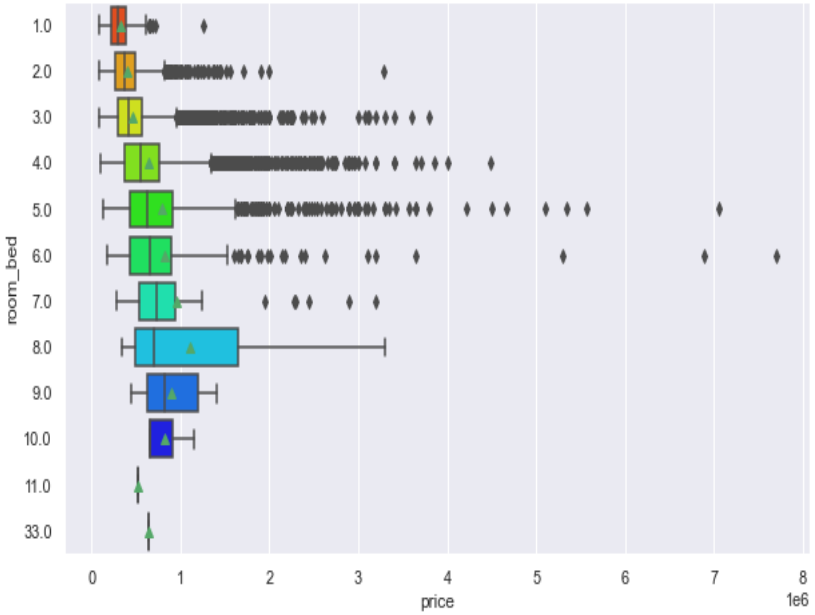


Fig – count or % of bedroom per house

- Three (3) bedroom houses have the highest frequency in the dataset follow by 4-bedroom houses
- 11, 33, 10, 9, 8, and 7-bedroom houses appear to be few in the dataset.

- Due to the presence of outliers, the general price for the number of bedrooms per house is set at the median price
- Generally, Houses with more bedrooms have higher prices that houses with few bedrooms

Fig – how prices vary across different bedroom houses



EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

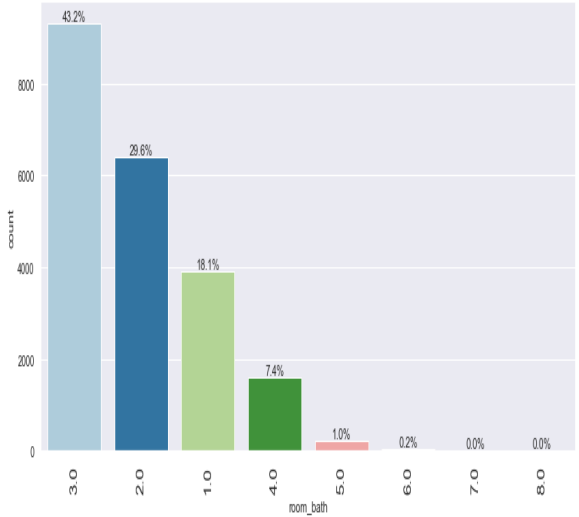


Fig – count or % of bathroom per house

- Houses with three (3) bathrooms have the highest frequency in the dataset follow by houses with 2 bathrooms
- Houses with 8, 7, 6, and 5 bathrooms appear to be few in the dataset.

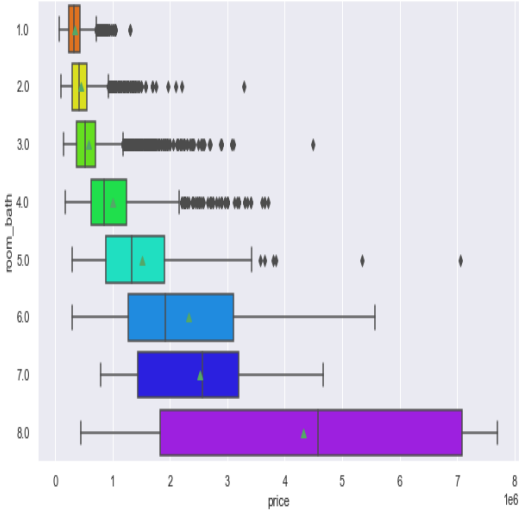


Fig – how prices vary across different bathroom houses

- Generally, houses with more bathrooms appear to have higher prices than houses with few bathrooms.

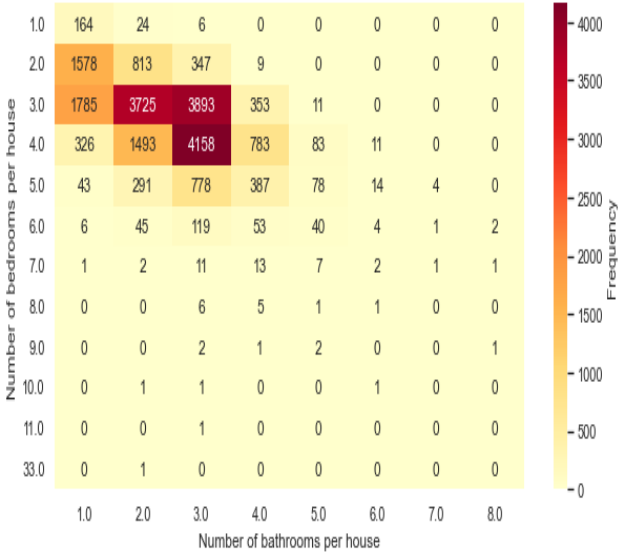


Fig – distribution of bedrooms per house against their bathrooms

- 4-bedroom houses with 3 bathrooms appear to have the highest frequency in the dataset followed by 3 bedroom and 3-bathroom houses, then 3-bedroom houses with 2 bathrooms.
- Prices of houses with few bedrooms and bathrooms have lower prices than houses with plenty bedrooms and bathrooms.

EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

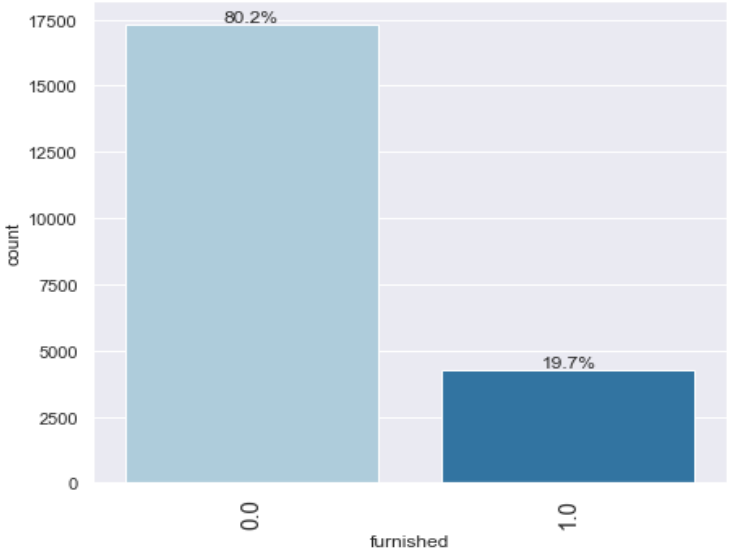
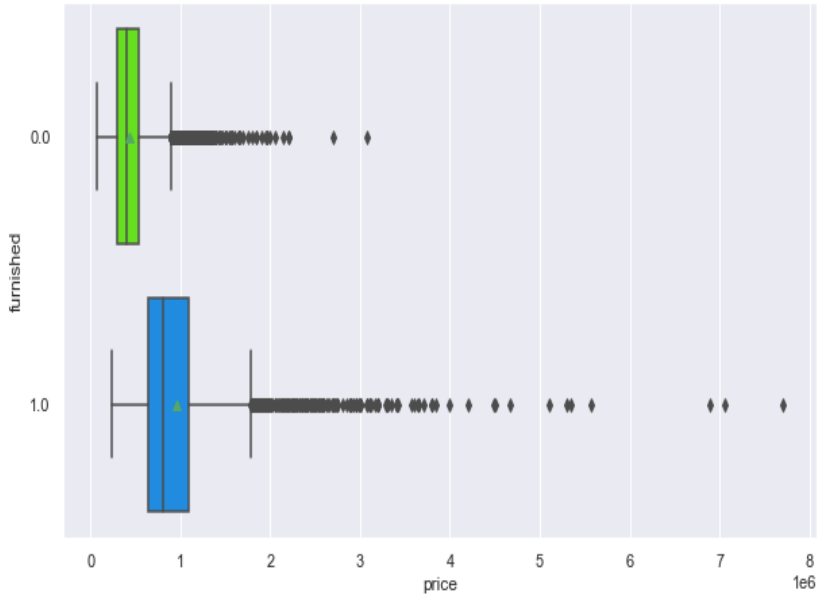


Fig – count or % of furnished and not-furnished houses

- About 80% of the houses are not furnished

- Generally, furnished houses have higher prices than unfurnished houses

Fig – how prices vary across furnished and not-furnished houses



EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

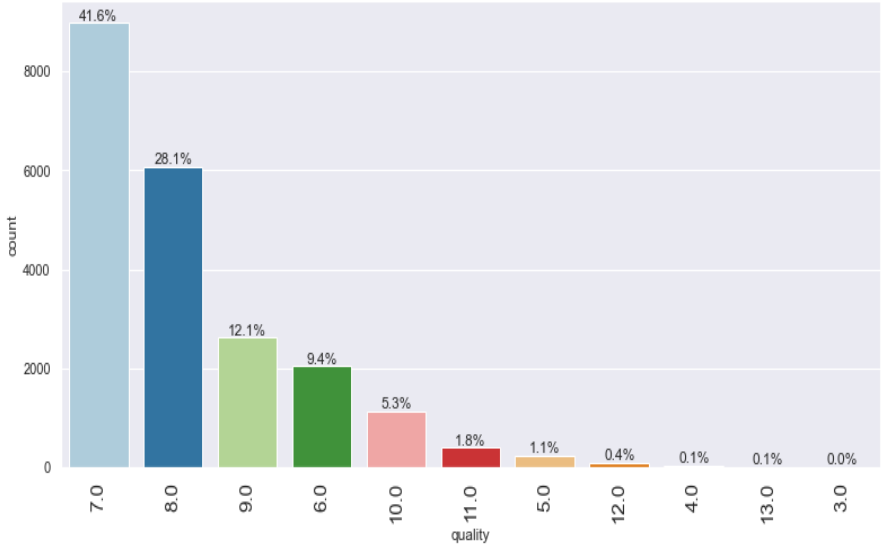
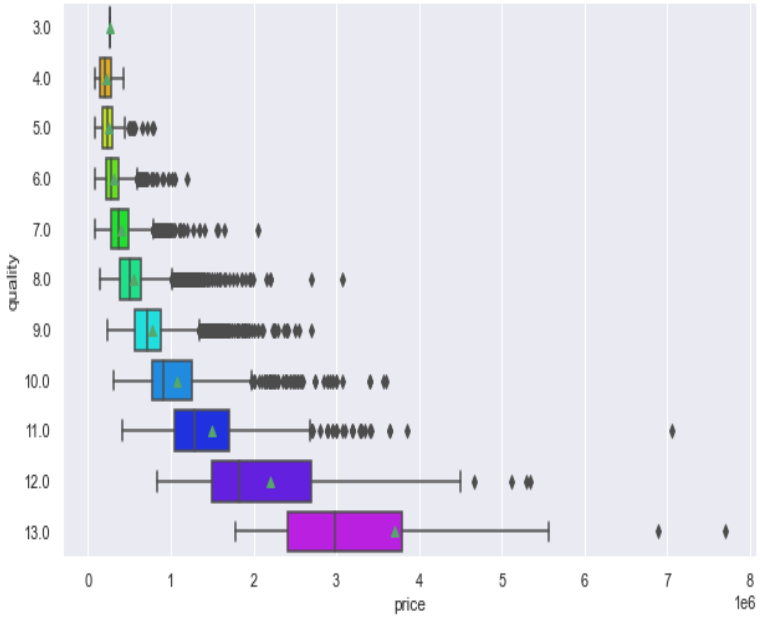


Fig – count or % of quality

- About 42% of the house were graded 7 and 28% were graded 8 moderate quality.
- Houses graded 3, 4, 5, and 6 could be suggested a low-quality houses and house graded 9, 10, 11, 12 and 13 could also be suggested as high-quality houses.
- High-quality houses and low-quality houses are very few in the dataset

- Generally high-quality houses have higher prices than lower quality houses.
- comparatively, moderate quality houses have moderate prices.

Fig – how a house quality impact price



EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

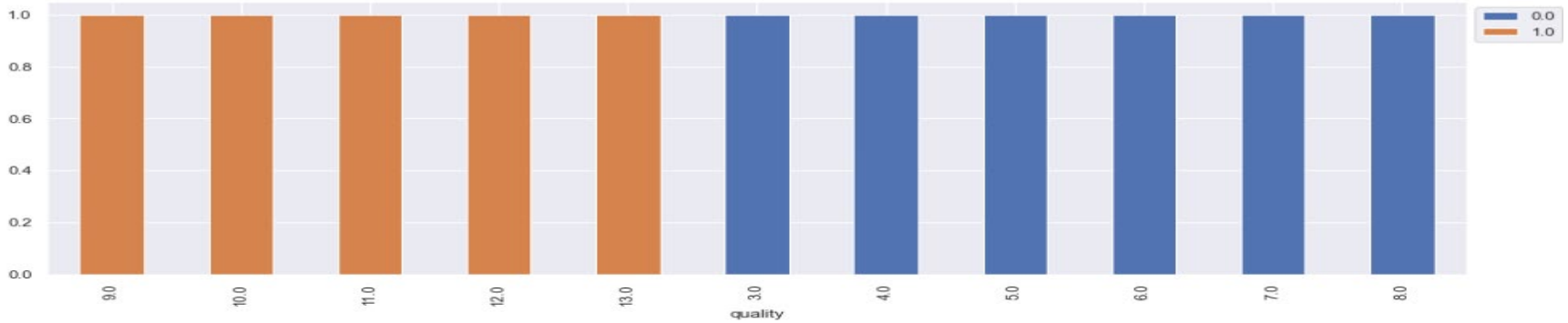


Fig – Houses furnishing are based on quality

furnished quality	0.0	1.0	All
All	17322	4245	21567
9.0	0	2612	2612
10.0	0	1132	1132
11.0	0	399	399
12.0	0	89	89
13.0	0	13	13
3.0	1	0	1
4.0	27	0	27
5.0	241	0	241
6.0	2035	0	2035
7.0	8965	0	8965
8.0	6053	0	6053

- High quality houses (graded 9 - 13) are furnished house while low and moderate quality(graded 3- 8) houses are non-furnished houses.

EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

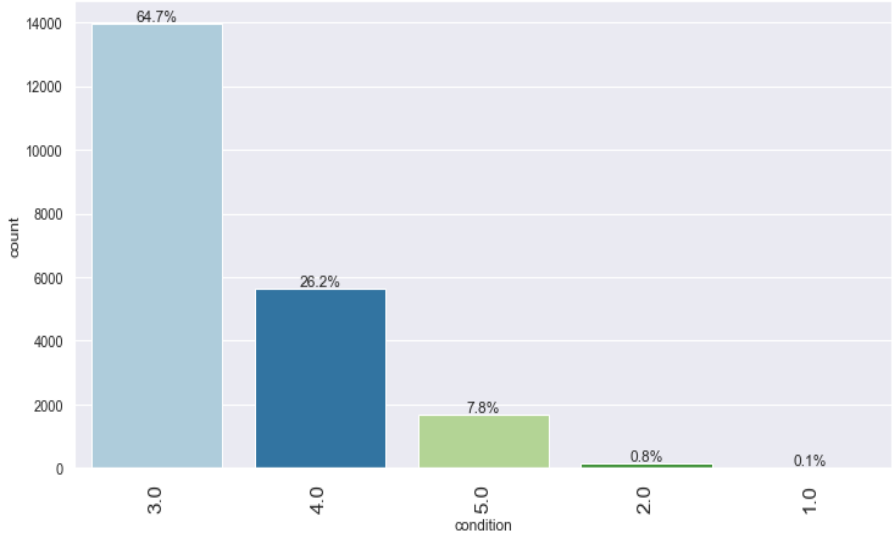
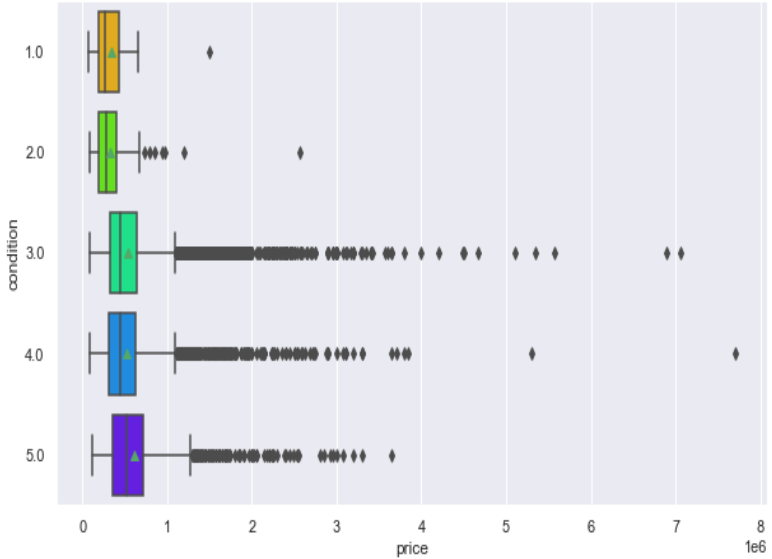


Fig – count or % of conditions

- Houses with moderate condition state appear to be more in the dataset – 64.7%
- Houses in good condition state are just 7.8% and poor state is 0.1%

- Generally, houses with good conditions have high prices than prices houses with bad conditions

Fig – how price vary across house condition



EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

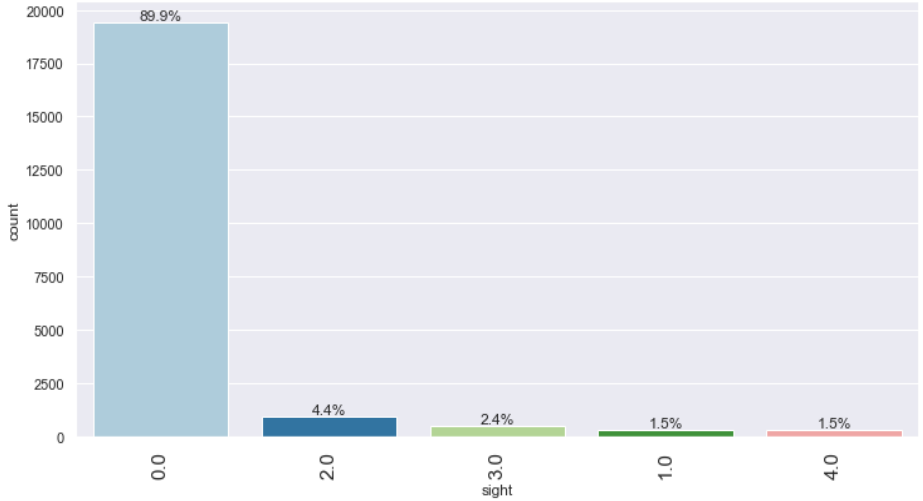
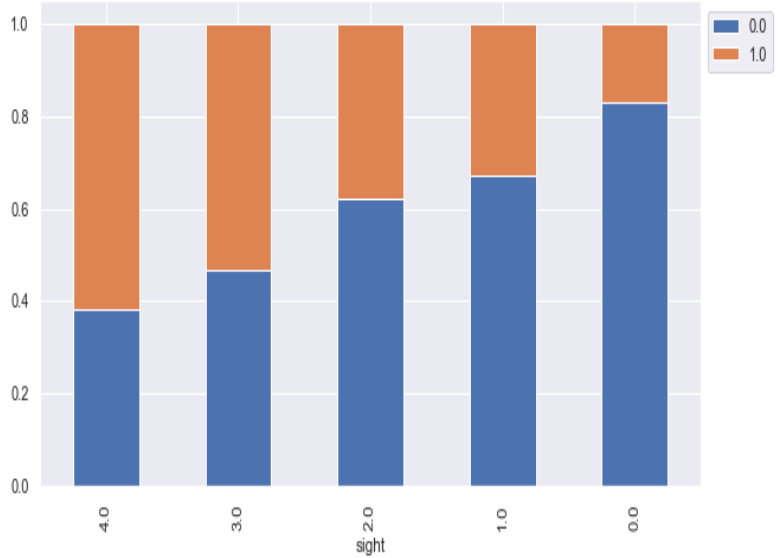


Fig – count or % of sight

- About 90% of the houses were not viewed. This could suggest that most of the houses were not attractive
- Just 1.5% of the houses appeared to be viewed many times.

- Furnished houses attract a lot of views than non furnished houses.

Fig – Do furnished houses attract a lot of viewers



EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

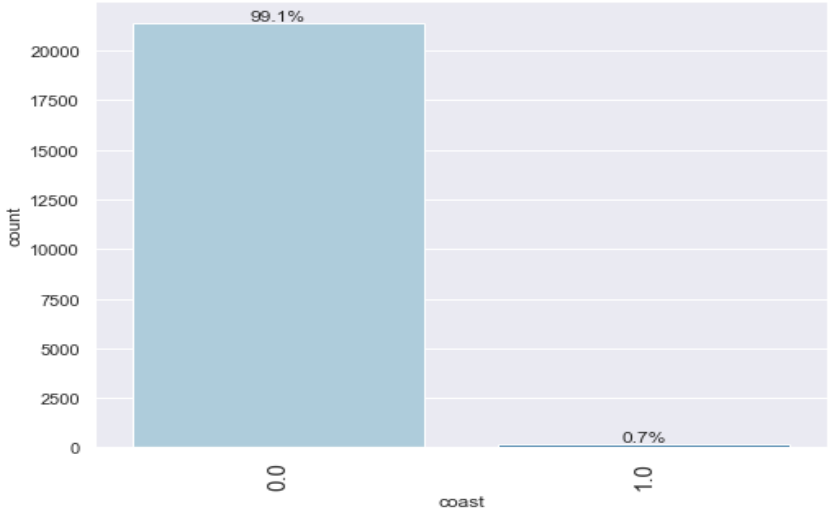
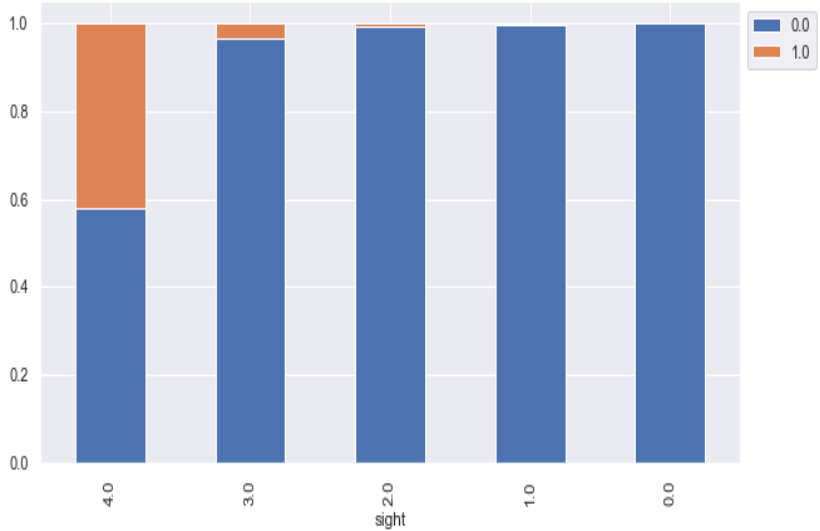


Fig – count or % of houses with waterfront

- About 99% of the houses have waterfront

- Houses with waterfront attract quite number of views.

Fig – Do houses with waterfront attract a lot of viewers



EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

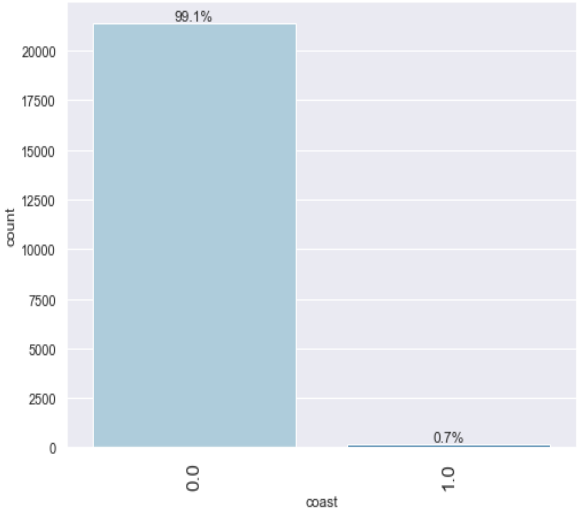


Fig – count or % of houses with waterfront

- About 99% of the houses have waterfront

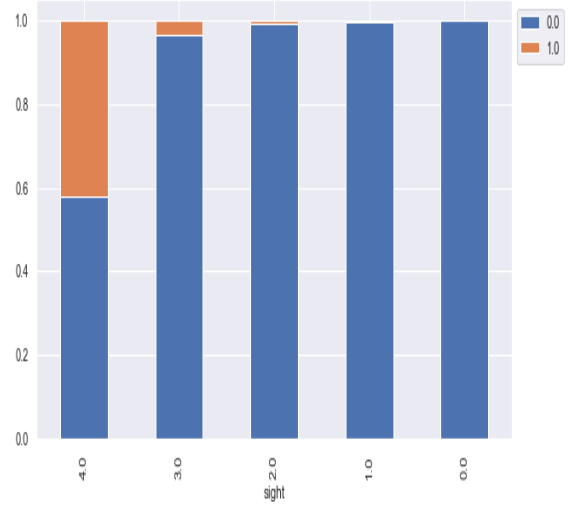


Fig – Do houses with waterfront attract a lot of viewers

- Houses with waterfront attract quite number of views.

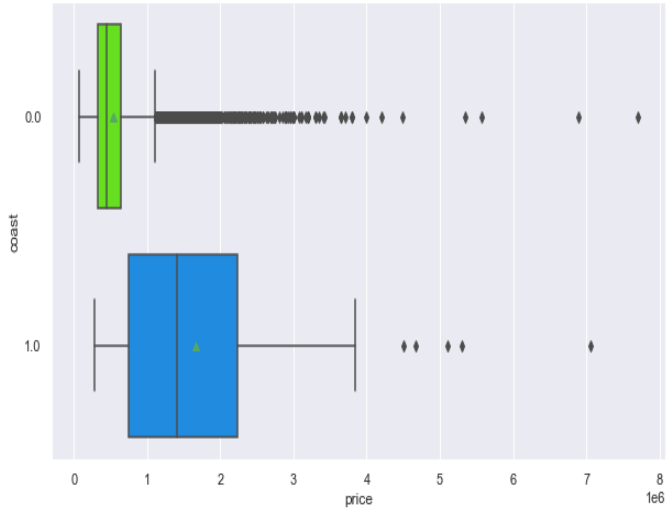


Fig – Do houses with waterfront attract higher prices

- Generally, houses with waterfront attract higher prices.

EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

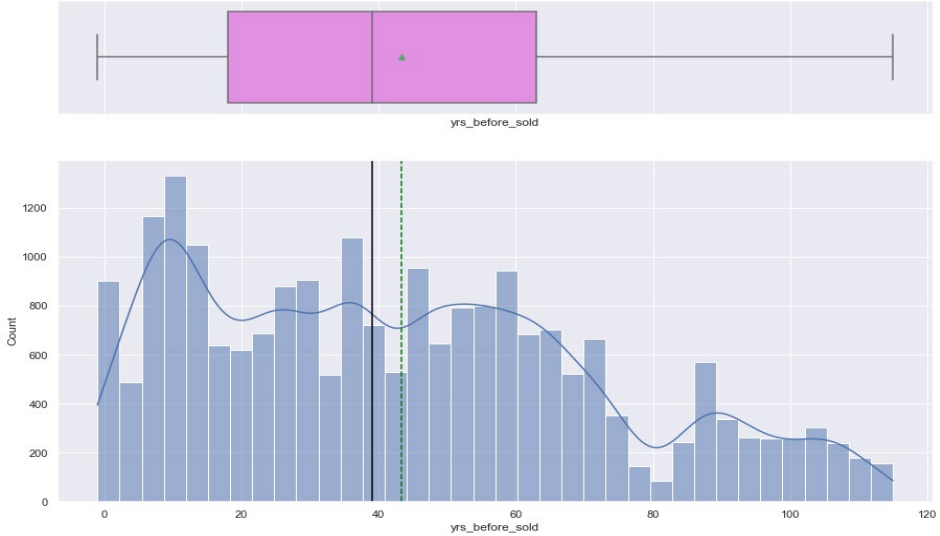
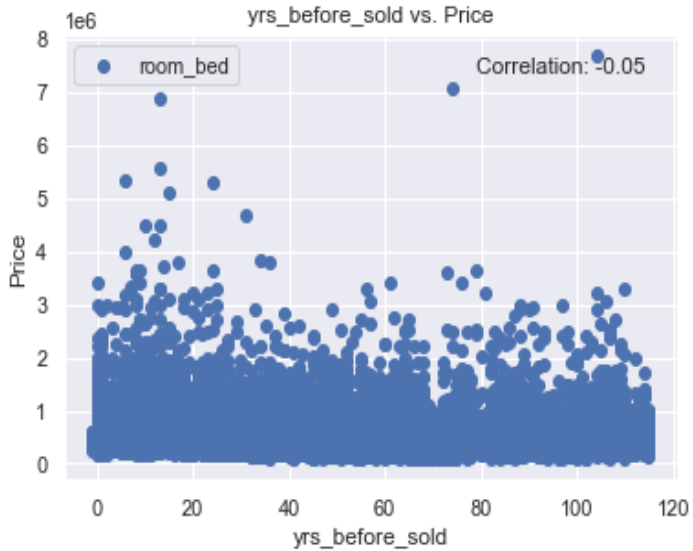


Fig – distribution of the years before houses were sold

- The distribution is slightly skewed to the right. This indicates that most of the houses have few years between year built and the sales years

- There is no correlation between price and yr_before_sold. This indicates that years of a house can not determine it price

Fig – correlation – price vs yr_before_sold



EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

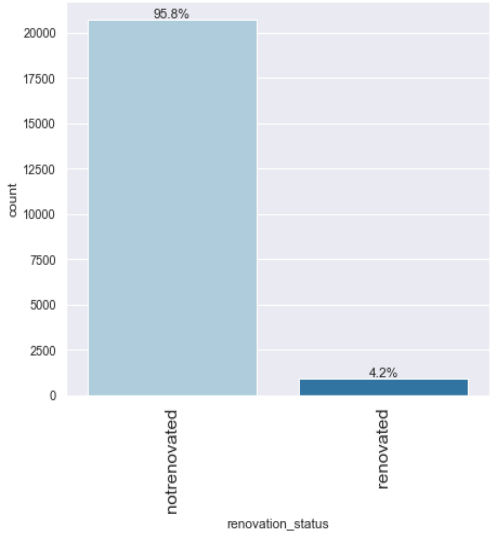


Fig – count or % of renovated houses

- Just 4% of the houses are renovated.
- Almost all the houses are not renovated.

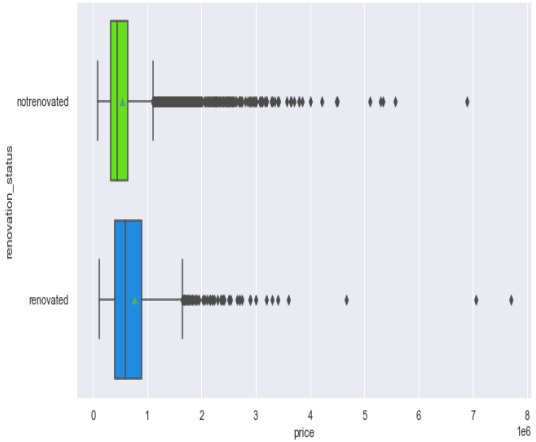


Fig – Do renovated houses have high prices

- Generally, renovated houses have higher prices than houses that are not renovated.

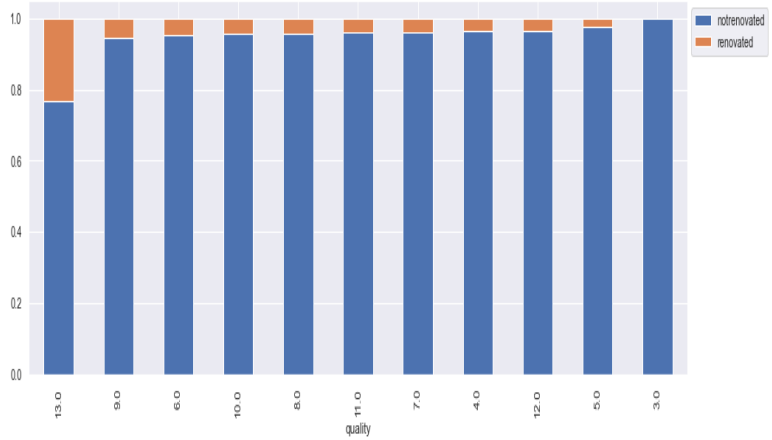


Fig – houses quality are based on their renovations

- Somehow renovation slightly improves the houses quality and furnishing.

EXPLORATORY DISTRIBUTION ANALYSIS RESULTS

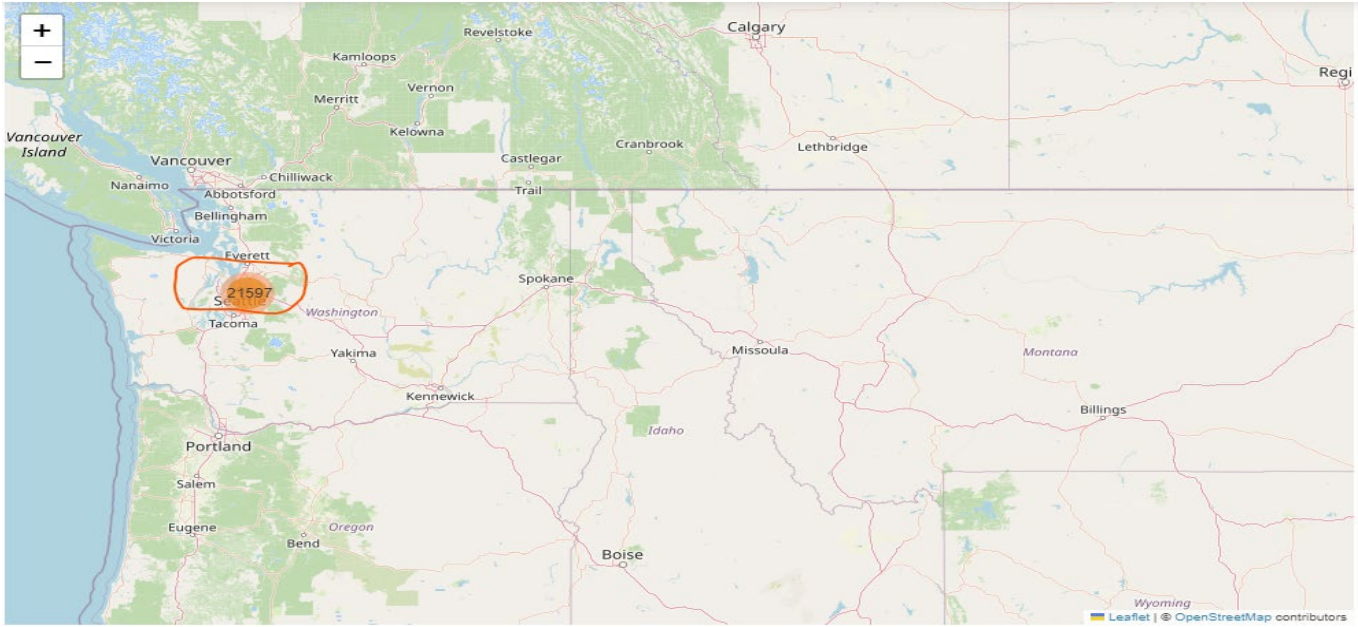


Fig – Exact locations of the houses in the dataset with long and lat

- All the houses can be located at Seattle in Washington, USA

Alternative analytical approaches that you may see fit to be applied to the problem

Given the problem statement and the goal of assessing house prices by considering various features, several analytical approaches can be applied. Here are some relevant approaches:

Feature Importance Analysis:

- Assess the importance of different features in influencing house prices. The EDA analysis, using the correlation plot, can help identify the key factors that contribute significantly to the value or prices of a house. Features such as the square footage of the house, the quality and the furnishing of the house, and the floor area of the house have a significant role in pricing a house. As these variables increases, prices will also increase. House buyers and Seller can study these variables to make informed home price decisions.

Location Analysis:

- Location can have impact on house prices. Proximity to good schools, parks, public transportation, and other excellent amenities can influence the price of a house. With the longitude and latitude or zipcode figures of the houses, Geospatial analysis can be employed to visualize and understand the precise location's influence on pricing.

Condition Assessment:

- Evaluate the condition of the house, considering factors such as the grading system, renovations, and overall maintenance. Buyers and sellers are often interested in the current state of houses and any recent improvements.

END OF REPORT