**Washington DC House Price Prediction**

Submitted towards partial fulfillment of the criteria for the award of

**PGPDSE** by **Great Lakes Institute of Management**

**Submitted by:**

**Group No.3** of **Batch: June 2019**

**Group Members:**

Aravind Goli

Hitesh

Hijam Gyaneswar

Asfia

Mahe Jabeen

**Mentor:**

Animesh Tiwari



Contents

[I. INTRODUCTION 3](#_Toc25406105)

[1. Problem statement: 3](#_Toc25406106)

[2. Data-set: 3](#_Toc25406107)

[3. Shape: 3](#_Toc25406108)

[II. LITERATURE 3](#_Toc25406109)

[III. DATA CLEANSING 4](#_Toc25406110)

[1. Data-formats: 4](#_Toc25406111)

[2. Dealing with Categorical Data: 4](#_Toc25406112)

[IV. EXPLORATORY DATA ANALYSIS 5](#_Toc25406113)

[V. BASELINE MODEL: 29](#_Toc25406114)

[VI. DATA PREPARATION: 29](#_Toc25406115)

[Outlier Treatment: 29](#_Toc25406116)

[MICE 29](#_Toc25406117)

[VII. MODELLING: 31](#_Toc25406118)

[Feature Selection: 32](#_Toc25406119)

[VIII.RECOMMENDATIONS 32](#_Toc25406120)

[1. living\_measure: 32](#_Toc25406121)

[2. Quality: 32](#_Toc25406122)

[3. Sight: 32](#_Toc25406123)

[XI. REFERENCES: 33](#_Toc25406124)

# INTRODUCTION

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 the aspects that give a house its value. For example, you want to sell a house and you don’t know the price which you can take — it can’t be too low or too high. To find a 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. You can find a house price by adding a number of relevant characteristics to the market value such as location, location, square footage, family size & age.

## Problem statement:

The aim here is to predict the price a house using the given features.

## Data-set:

Features of a house and the price of the houses sold in Washington DC by a firm from May 2014 to May 2015.

## Shape:

21613 rows and 23 columns

# LITERATURE

**Columns:**

* 1. **cid:** A notation for a house
  2. **dayhours:** Date house was sold
  3. **price:** Price is the prediction target
  4. **room\_bed:** Number of Bedrooms/House

#### **room\_bath:**

This feature is calculated by calculating each of the bathroom equipment Shower, Sink, Toilet and Bath tub as a value of one-fourth, so as to give better representation of the bathrooms with more bathroom equipment.

#### **living\_measure:**

This is the total carpet area of the house including the basement of the house. This is the feature that buyers ask to estimate the size and price of the house.

#### **lot\_measure:**

This is the total land area of the house. This is not related to the living measure hence does not influence on the price of the house.

* 1. **ceil:** Total floors (levels) in house
  2. **coast:** House which has a view to a waterfront
  3. **sight:** Times the house has been viewed
  4. **condition:** How good the condition is (Overall)
  5. **quality:** grade given to the housing unit, based on a grading system
  6. **ceil\_measure:** square footage of house apart from the basement
  7. **basement\_measure:** square footage of the basement
  8. **yr\_built:** Built Year
  9. **yr\_renovated:** Year when the house was renovated
  10. **zipcode:** zip
  11. **lat:** Latitude coordinate
  12. **long:** Longitude coordinate
  13. **living\_measure15:** Living room area in 2015(implies-- some renovations). This might or might not have affected the lot size area.
  14. **lot\_measure15:** lot size area in 2015(implies-- some renovations)
  15. **furnished:** Binary feature showing if the house is furnished or not.
  16. **total\_area:** Measure of both living and lot measures.

# DATA CLEANSING

The following changes have been done for better analysis, visualization and model building. The changes done for the required columns are as below:

## Data-formats:

**dayhours**: The original format of data in this column is 20141107T000000 with data type as string, where the first 4 characters represent the year, next 2 represent the month, next 2 represent the day and the rest of the characters doesn't represent any information. This format has been changed to date-time using **datetime.strptime** function, after slicing the string. Then this is applied to the whole column using .**apply** function. This also changed the data type of the column to date-time.

The columns **coast**, **sight** and **furnished** are converted to object data type to make them categorical. And the columns **room\_bed**, **room\_bath**, **ceil**, **sight**, **condition**, **quality**, **yr\_built**, **yr\_renovated** and **dayhours** are to be decided between continuous and categorical. Exploratory Data Analysis will solve this problem.

## Dealing with Categorical Data:

In the dataset there are 9 categorical variables. Some of the categories have ordinal data and can be dealed by using label encoding.

# EXPLORATORY DATA ANALYSIS

EDA and Visualization helped us have better insights into data. Univariate analysis is done for each column and then a bivariate analysis of each column with price and multivariate analysis of the columns that are needed to be assessed. Upon the univariate and bivariate analysis of the columns, the columns that could be either categorical or continuous are decided what type of data type they are. So following are the steps in EDA and the insights we found for column-wise:

1. **cid:** The number of houses sold in the data is 21613, while 21436 of them are unique, of out which 1 house was sold thrice, 175 sold twice and the rest sold only once. This is the only insight it gave out and it can be dropped for modelling.
2. **price:** This is clearly a continuous variable and the dependent variable that we need to regress.

Uni-variate analysis:

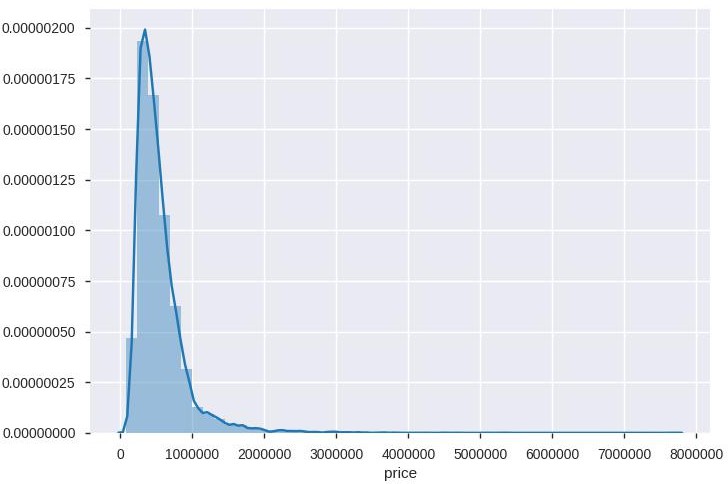
Mean:540182.158793

Std: 367362.231718

Five-point summary:

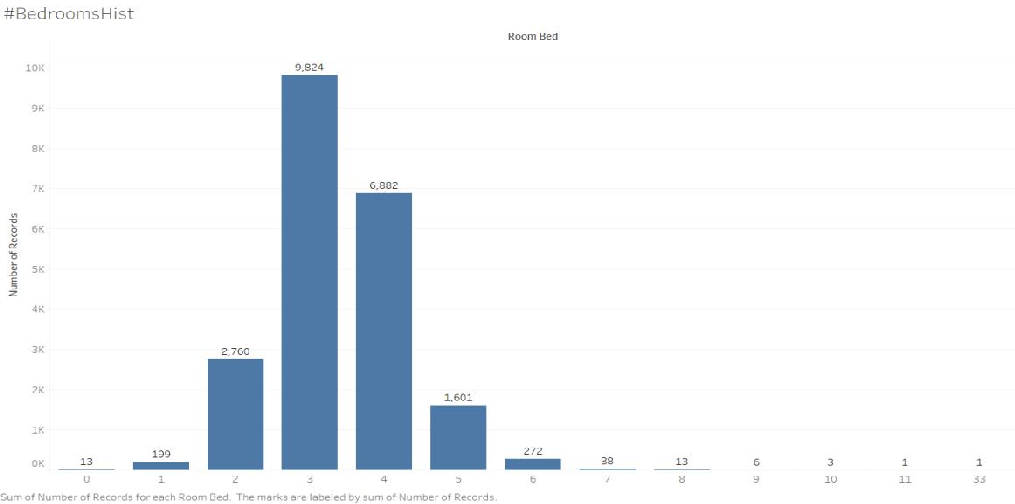
|  |  |
| --- | --- |
| min | 75000.000000 |
| 25% | 321950.000000 |
| 50% | 450000.000000 |
| 75% | 645000.000000 |
| max | 7700000.000000 |

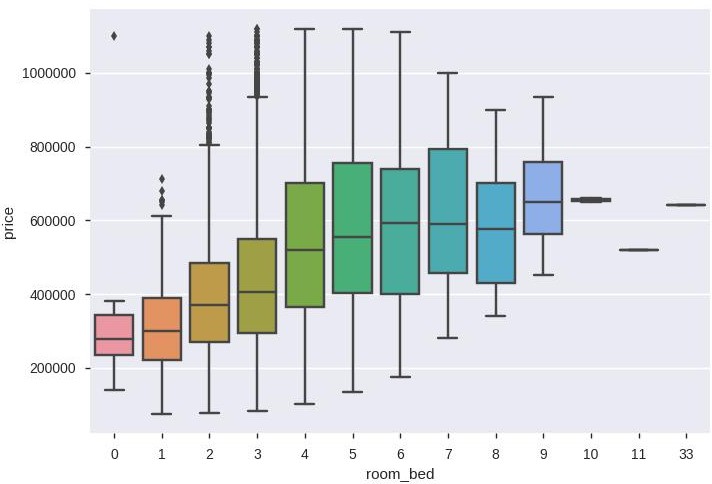
Mean price greater than the median price indicating right-skewed data.



The above graph is the distribution plot of the price. The y-axis here indicates the probability of the price and the x-axis being the price, clearly indicating that the data is right- skewed and the same can be found through the 5 point summary. 5.87% of the data are outliers are have to be treated.

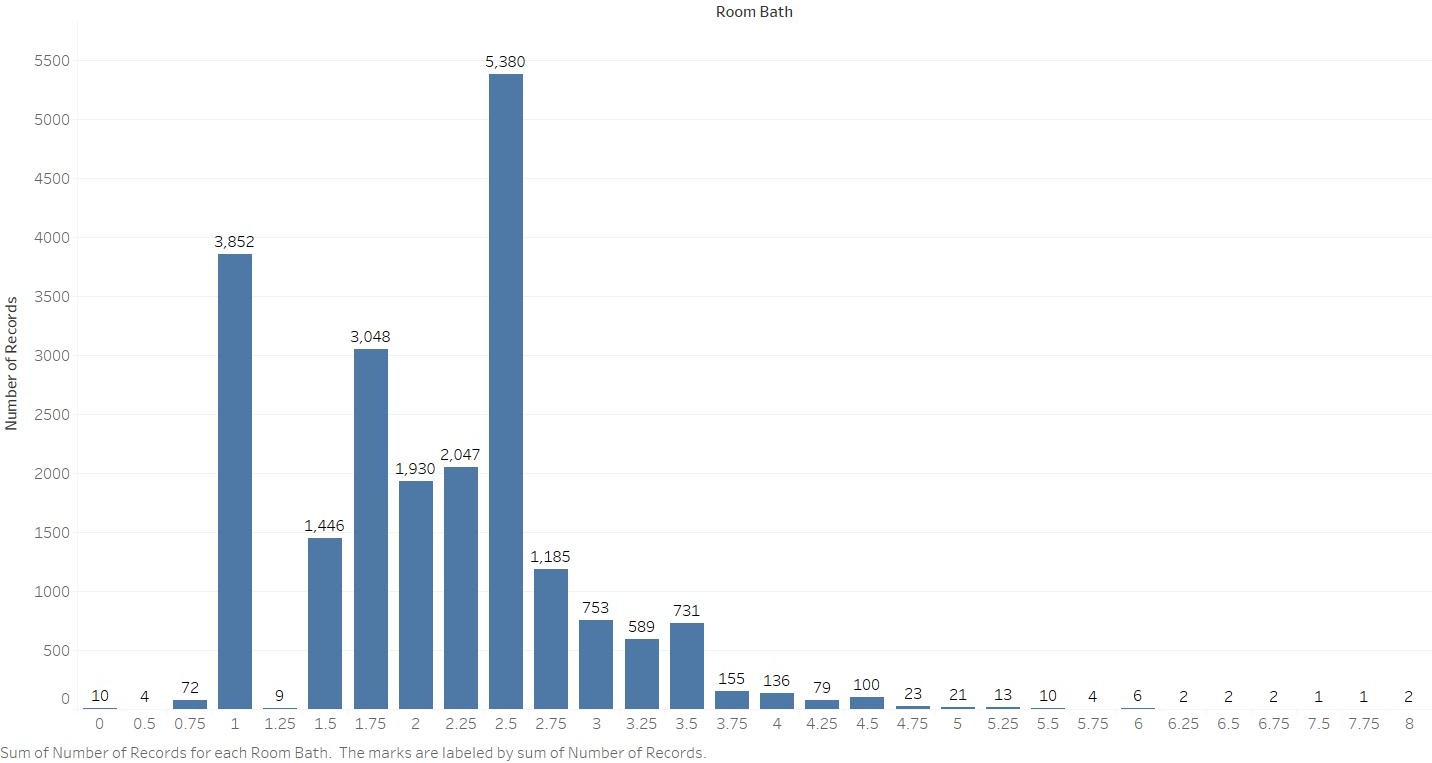
1. **room\_bed:** Intuition is that price increases with the number of bedrooms. It can be seen from the below count plot that houses with 3 bedrooms are highest in number. And we can see that almost all of the houses have the number of bedrooms in between 1 and 6.

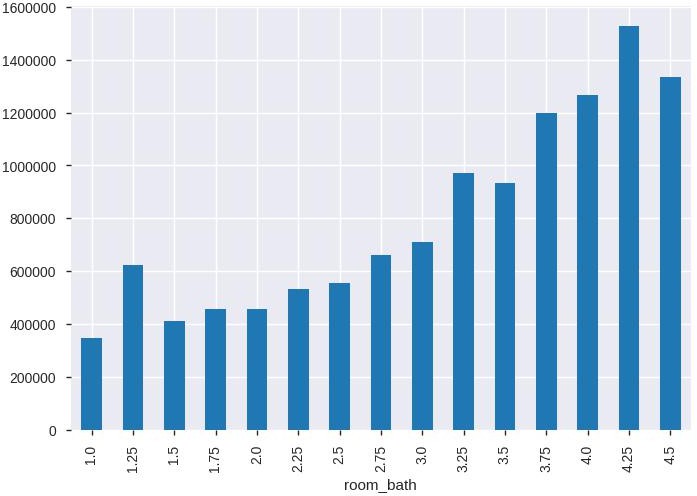




The boxplot between the price in the y-axis and the number of bedrooms in the x-axis is shown above. The above boxplot for the bedrooms in between 1 and 6 clearly shows us that the price is increasing with the number of bedrooms. This confirms that room\_bed is categorical.

1. **room\_bath:** Intuition is that price increases with the bathrooms. It can be seen from the below count plot that houses with 2.5 bathrooms are highest in number. And we can see that almost all of the houses have the number of bathrooms in between 1 and 4.5.





This above graph has the mean price for the given number of bathrooms in the y-axis and the number of bathrooms in the x-axis. It shows that the average price is increasing with the increase in the number of bathrooms. room\_bath is categorical and a significant feature for price prediction.

1. **living\_measure:** Intuition is that price increases with the living\_measure. This is the measure that buyers look at first to buy a house out in the market.

Uni-variate analysis:

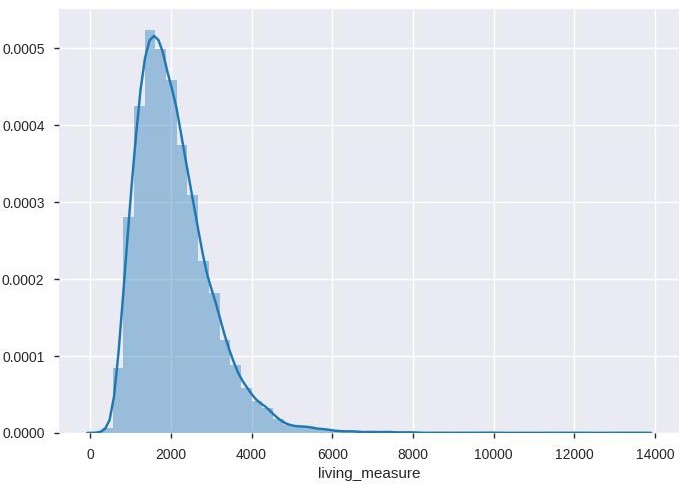
mean 2079.899736

std 918.440897

Five-point summary:

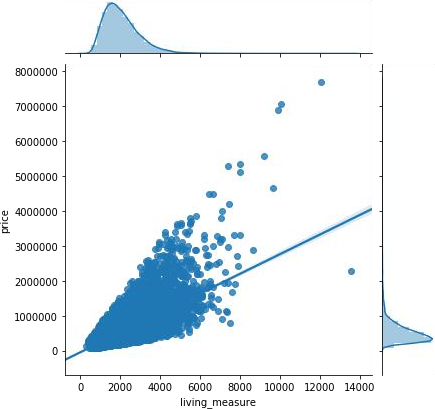
|  |  |
| --- | --- |
| min | 290.000000 |
| 25% | 1427.000000 |
| 50% | 1910.000000 |
| 75% | 2550.000000 |
| max | 13540.000000 |

Mean price greater than the median price indicating right-skew.



X-axis: living\_measure; Y-axis: probability density of living measure

The graph above clearly shows that the measure is right-skewed. 2.65% of the data are outliers and needs to be treated accordingly.



This graph above is the joint plot of living measure with the price. It is clear with the regression line that the price increases with the living\_measure.

1. **lot\_measure:** Intuition is that price increases with the lot\_measure.

Univariate analysis:

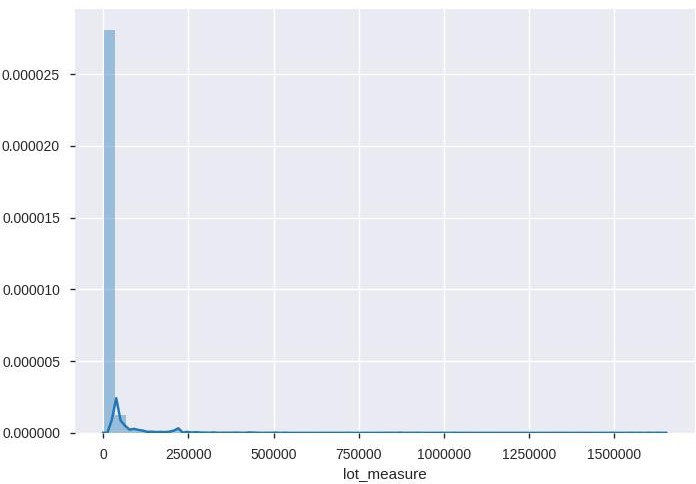
mean 15106.967566

std 41420.511515

Five-point summary:

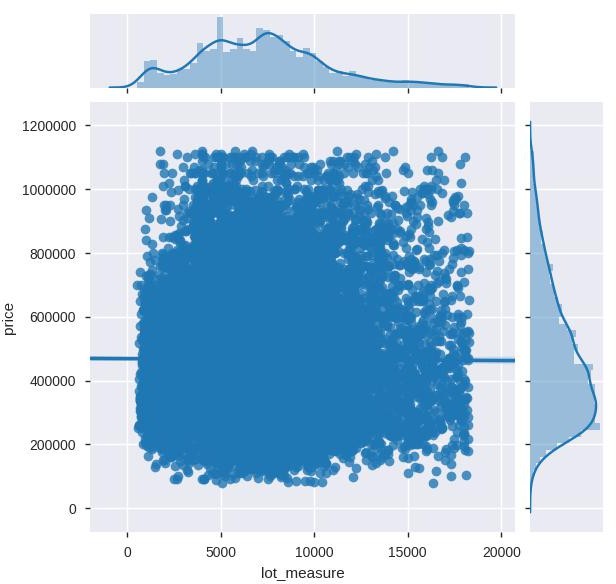
|  |  |
| --- | --- |
| min | 520.000000 |
| 25% | 5040.000000 |
| 50% | 7618.000000 |
| 75% | 10688.000000 |
| max | 1651359.000000 |

Mean is greater than the median indicating right-skew.



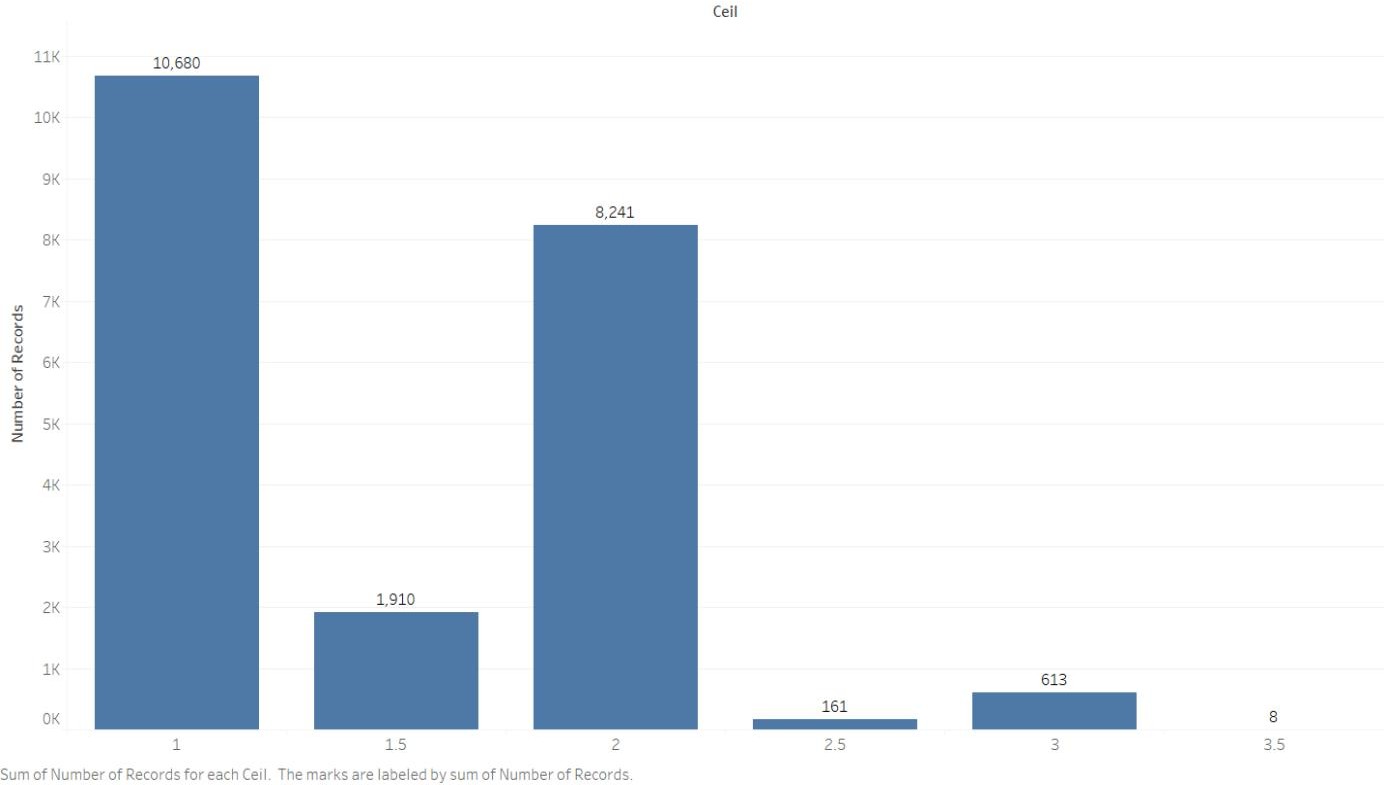
X-axis: lot\_measure; Y-axis: probability density

The above probability density graph shows the lot\_measure is not nearly a normal distribution, and also very highly right-skewed. The dependency of price in lot\_measure is almost zero as we can see from the graph below. This is not what the initial intuition was before analysis. The regression line is plain flat and hence lot measure is not influencing the price.



1. **ceil:** The initial intuition is that the price increases with ceil.

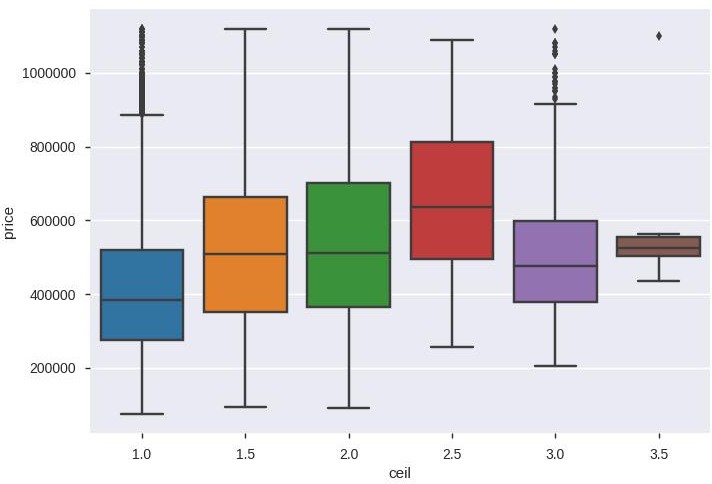
Value counts:

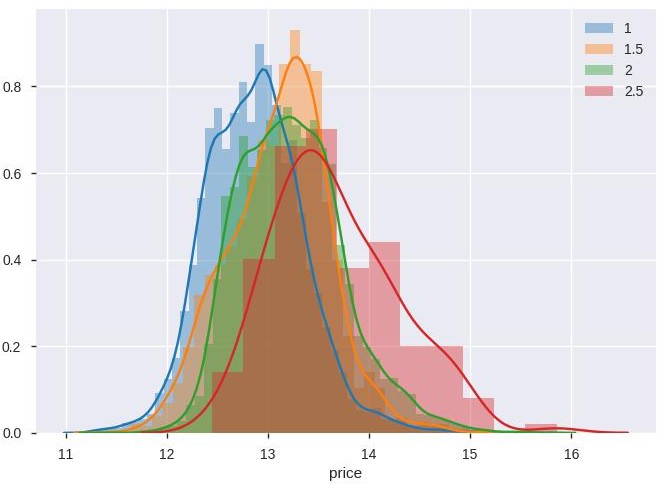


X-axis: ceil; Y-axis: count of records

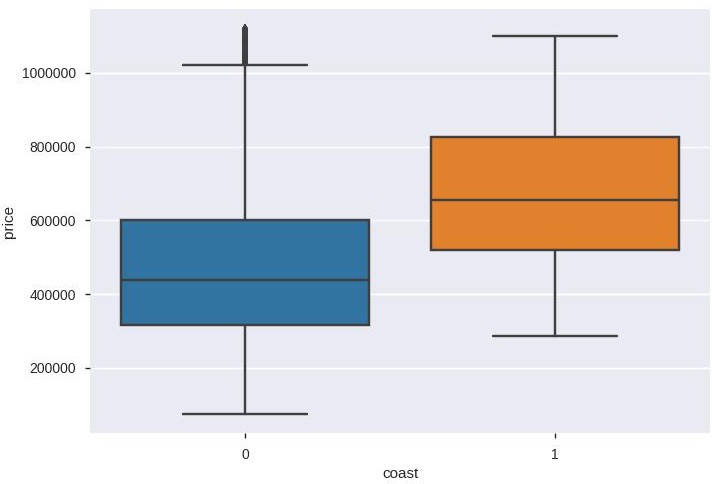
The houses with 1 ceil are the majority class and it shows that almost all the houses have 3 or less than 3 ceils. The houses with 3.5 ceil are very low in number.

The below box plot for price vs ceil shows that price increases with ceil. But that is not true for ceils 3 and 3.5, but they can be neglected as the number of records for them are very low. Another observation from the above plot is that 1.5 ceil and 2 ceil are almost the price distribution, which implies penthouses have the same importance as a normal floor. The same can be inferred from the graph below showing the distribution of price for various ceils.

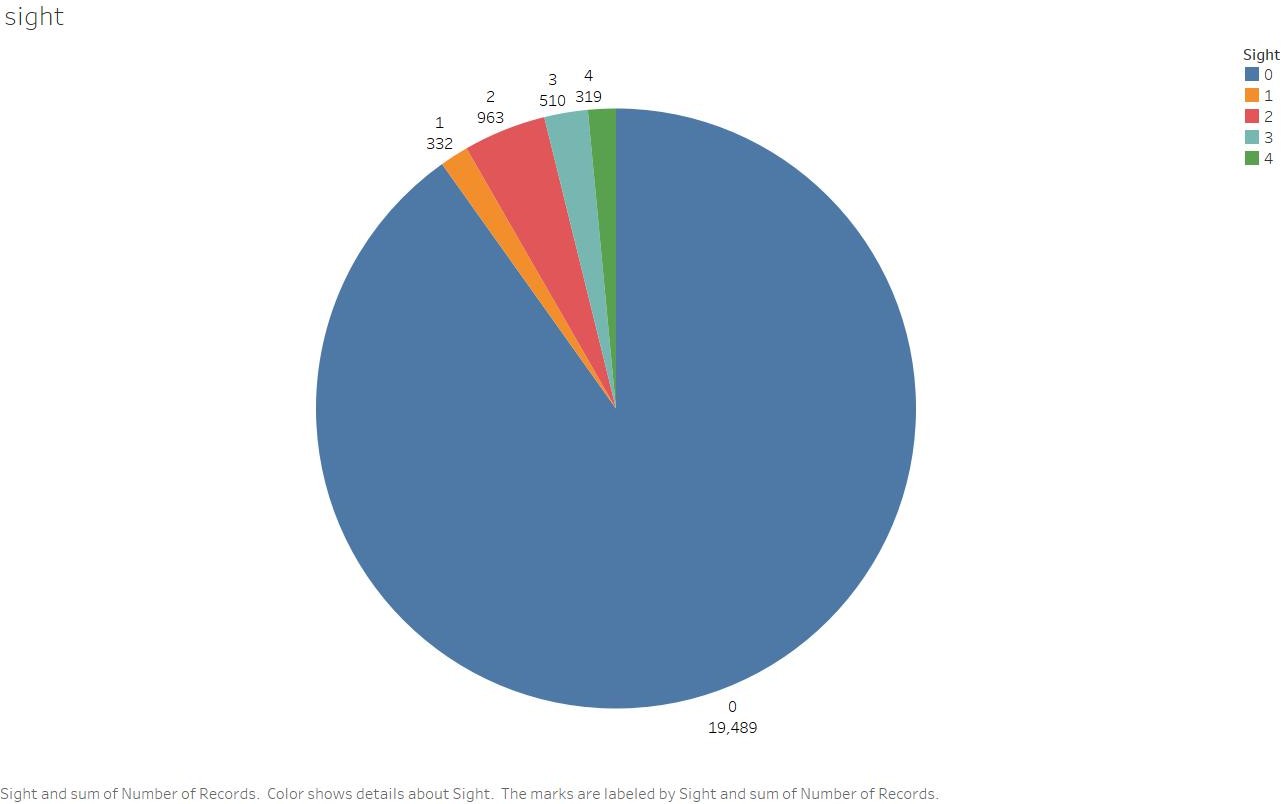




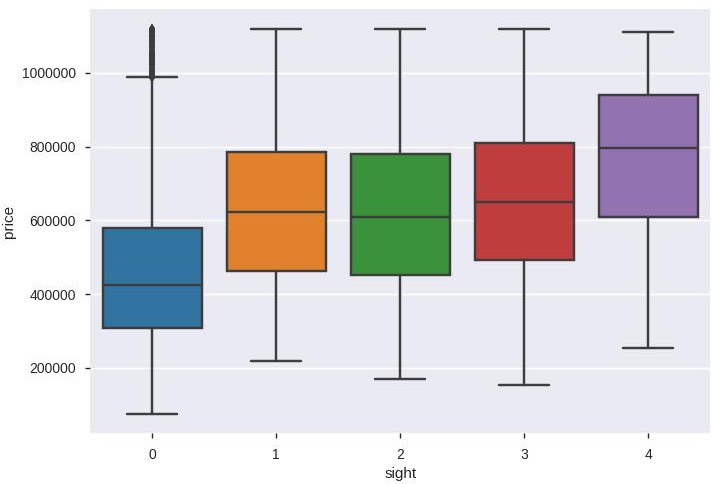
1. **coast:** Intuition is that houses by the coast are high priced. Coast is considered categorical. Only 163 out of 21613 houses are by the coast. From the boxplot of price vs coast below, we can see that the prices of the houses by the coast are much higher than those are not.



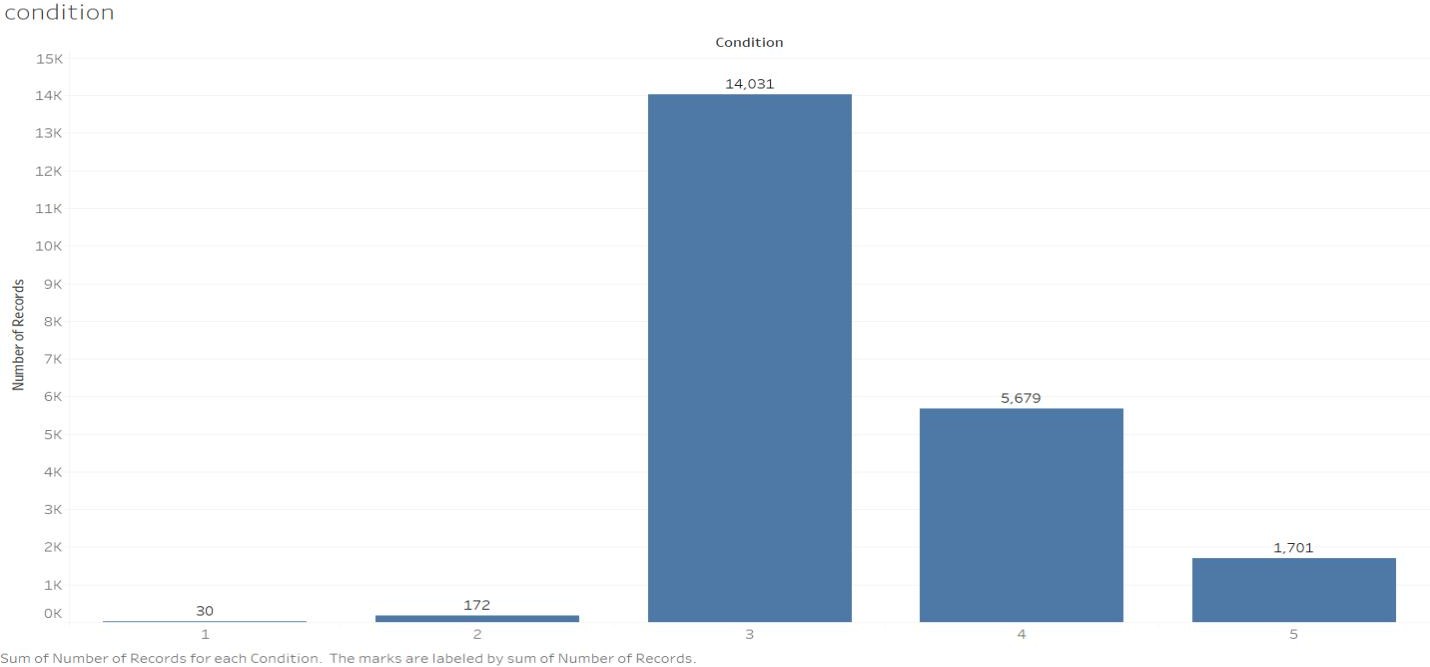
1. **sight:**

univariate analysis:

From the graph below we can infer that the mean price of sight 1, 2 and 3 are higher than that of sight 1 and also sight 4 houses are higher priced.



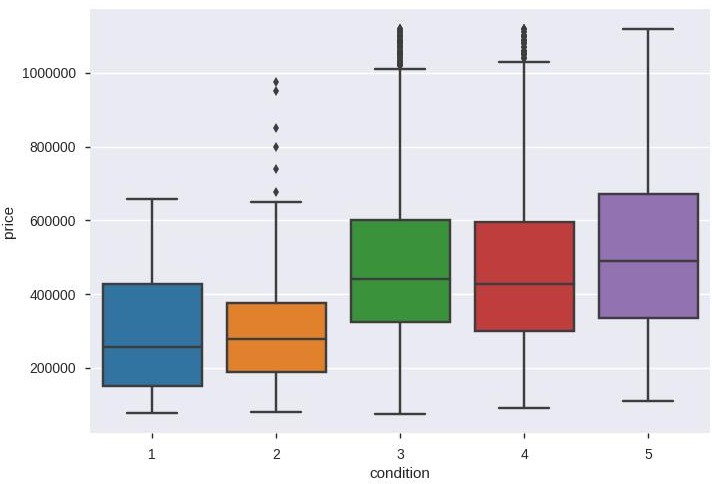
1. **condition:** Intuition is that price increases with the condition. This is a categorical variable. Univariate analysis:



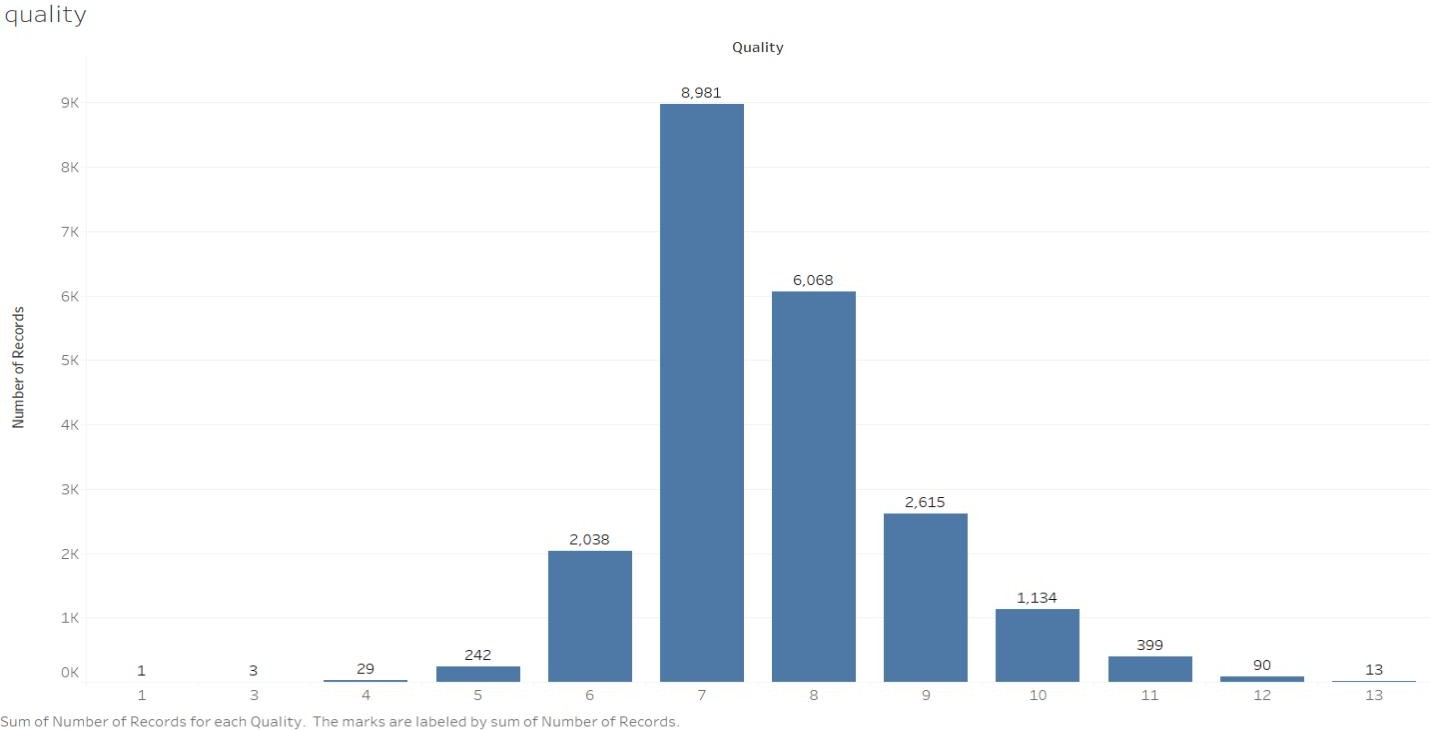
As can be seen above almost all the houses have condition 3 to 5 and condition 1 and 2

are negligible. The graph below shows how the condition affects the price of the house. It is

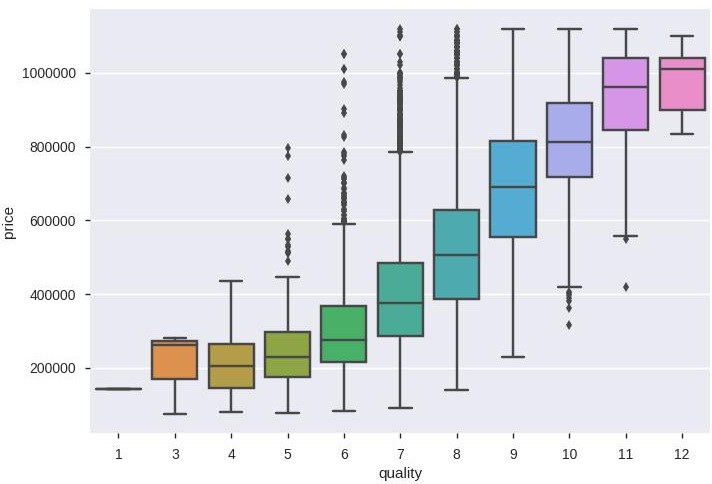
observed that the condition doesn't really affect the price unless it is low as q or 2 but such houses are rare, rest the condition 5 house is slightly high priced.



1. **quality:** intuition is that the price increases with increase in quality. Univariate analysis:



It can be seen that quality is concentrated at 7 decreasing as we move either side. Most of the values are in between 5 and 11. From the boxplot of price vs quality below we can observe that quality is a very significant feature in price prediction. The increase in price is high with an increase in quality.



1. **ceil\_measure:** It increases with living measure, so it is expected to have a positive regression line in the graph with the price.

Univariate analysis:

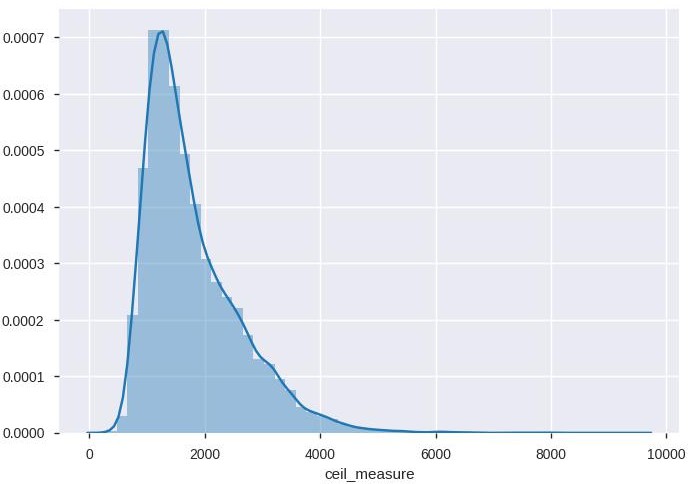
mean 1788.390691

std 828.090978

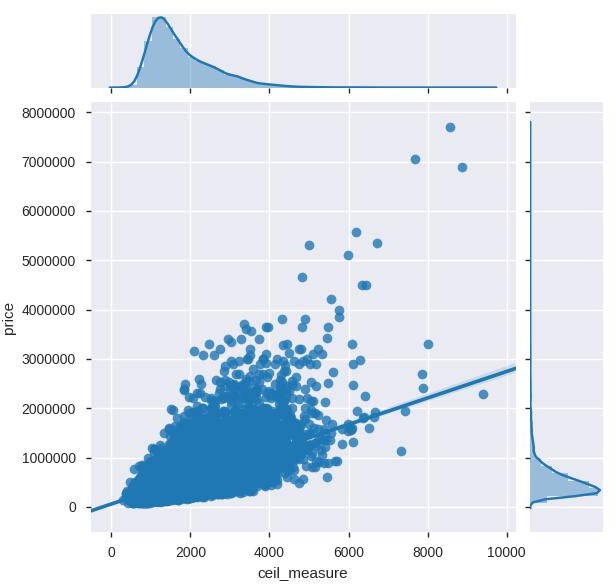
Five-point summary:

|  |  |
| --- | --- |
| min | 290.000000 |
| 25% | 1190.000000 |
| 50% | 1560.000000 |
| 75% | 2210.000000 |
| max | 9410.000000 |

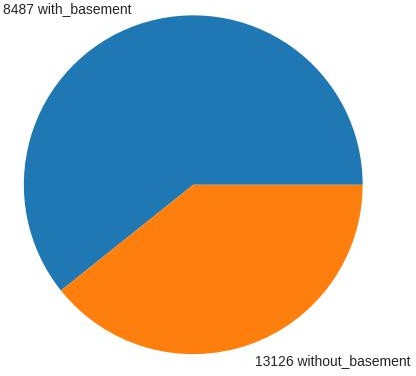
The below graph indicates the probability density in the y-axis and ceil\_measure in the x- axis. It is observed to be right-skewed. Outliers have to be treated accordingly.



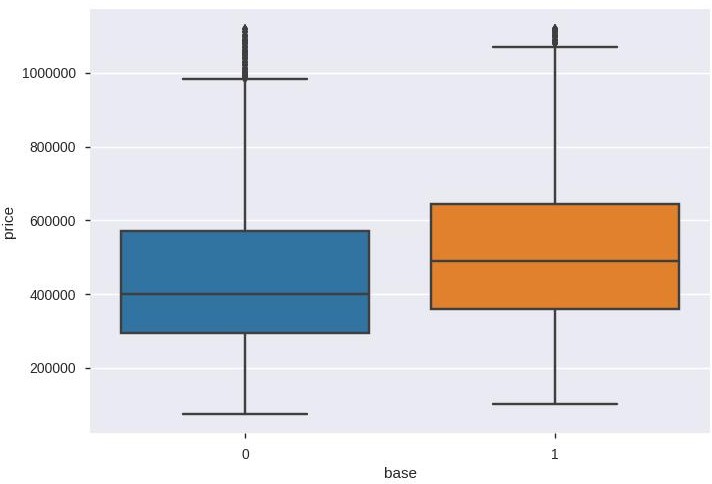
The joint plot for the ceil\_measure and the price is in a similar fashion to that of the price vs living\_measure.

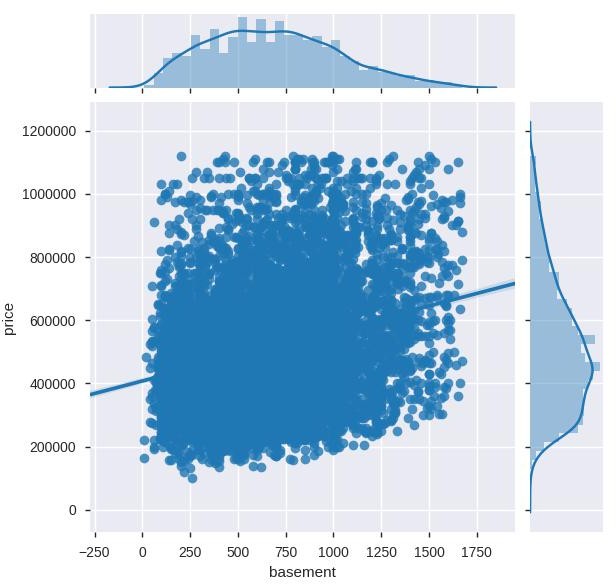


1. **basement:** Out of 21613 houses 13126 houses have no basement.



And the price difference between the houses with and without a basement is not so high but significant. The next observation is the dependency of the basement measure if a house has one on the price



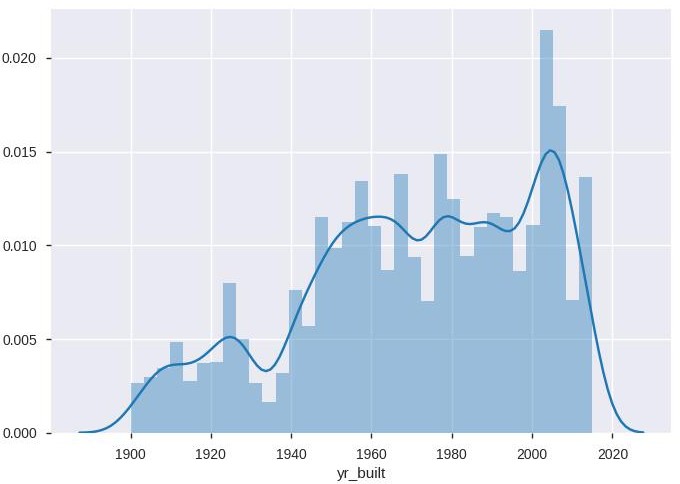


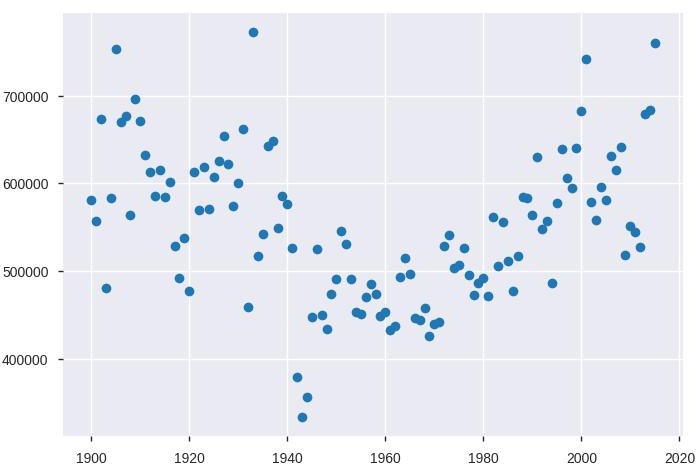
This joint plot above shows us using the regression line that the price is increasing with the increasing basement area. But this might be also due to the increase in the living\_measure as the basement area is part of the living\_measure.

1. **yr\_built:**

Univariate analysis:

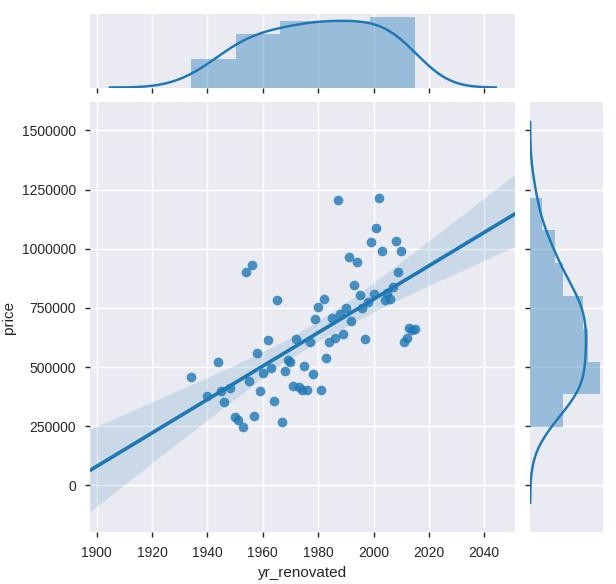
The y-axis representing the probability density of the year built and the x-axis representing the year built. It is observed that there is an increase in the number of houses built with the increase in the year built. We can see the effect of The Great Depression during the 1930s in the United States.

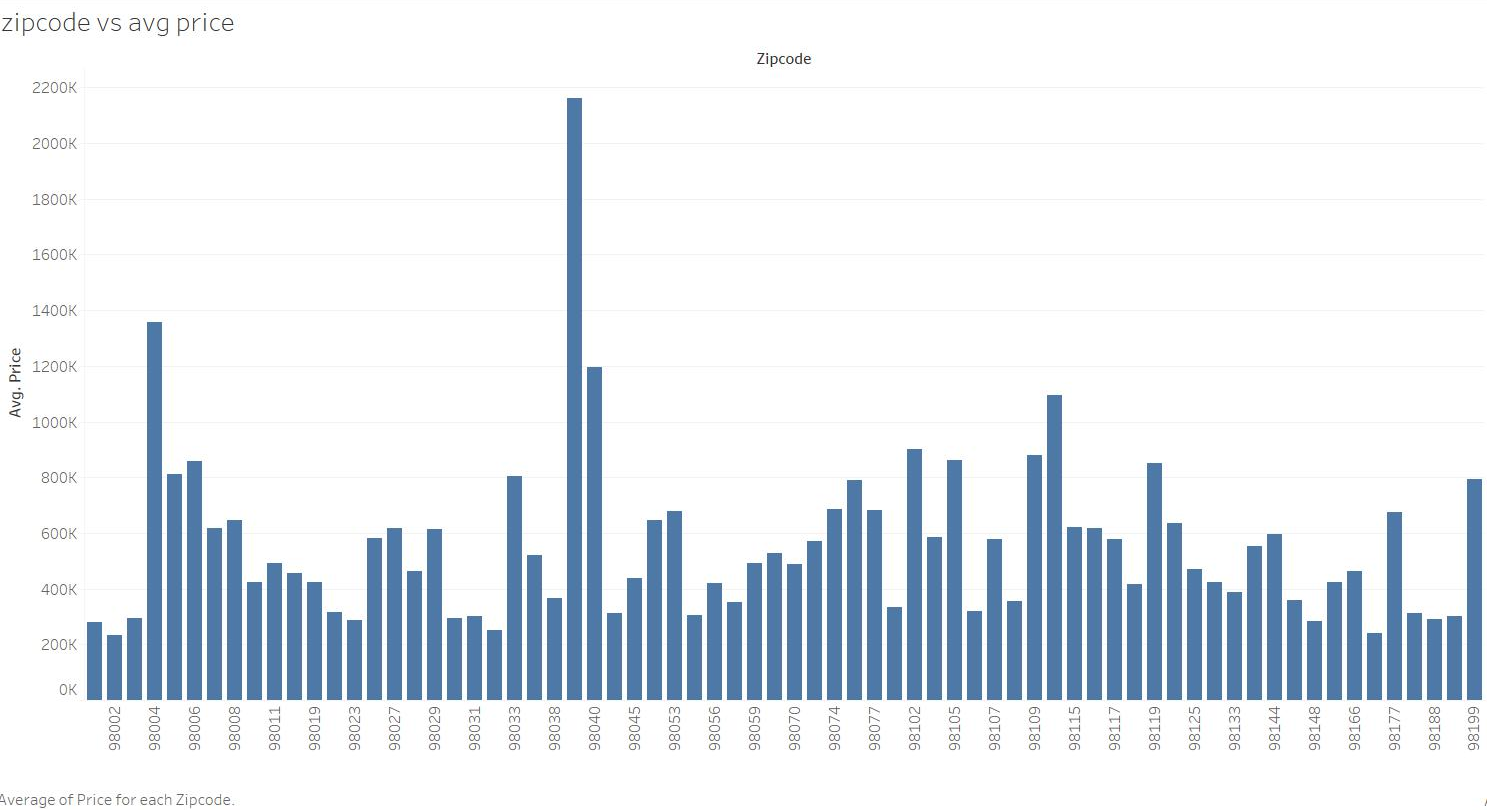


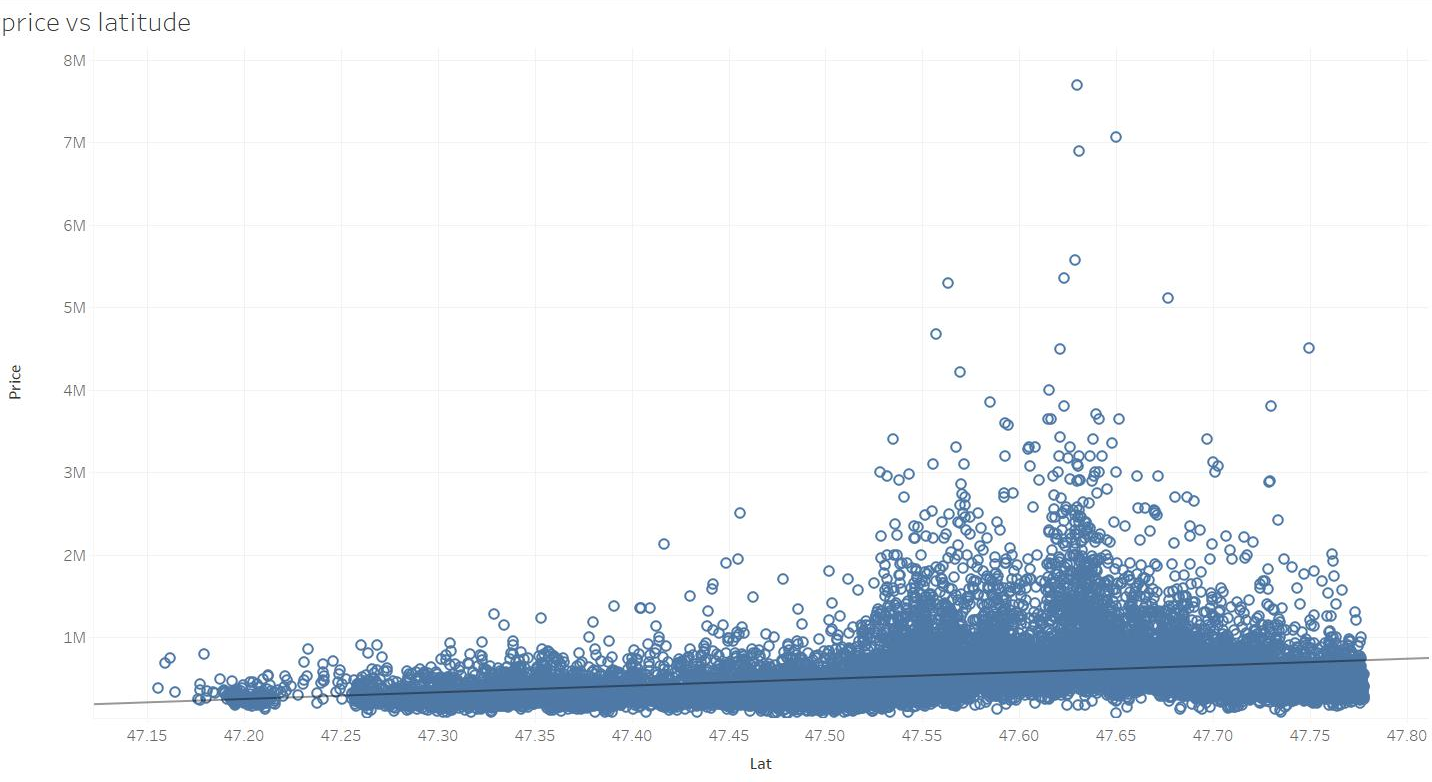


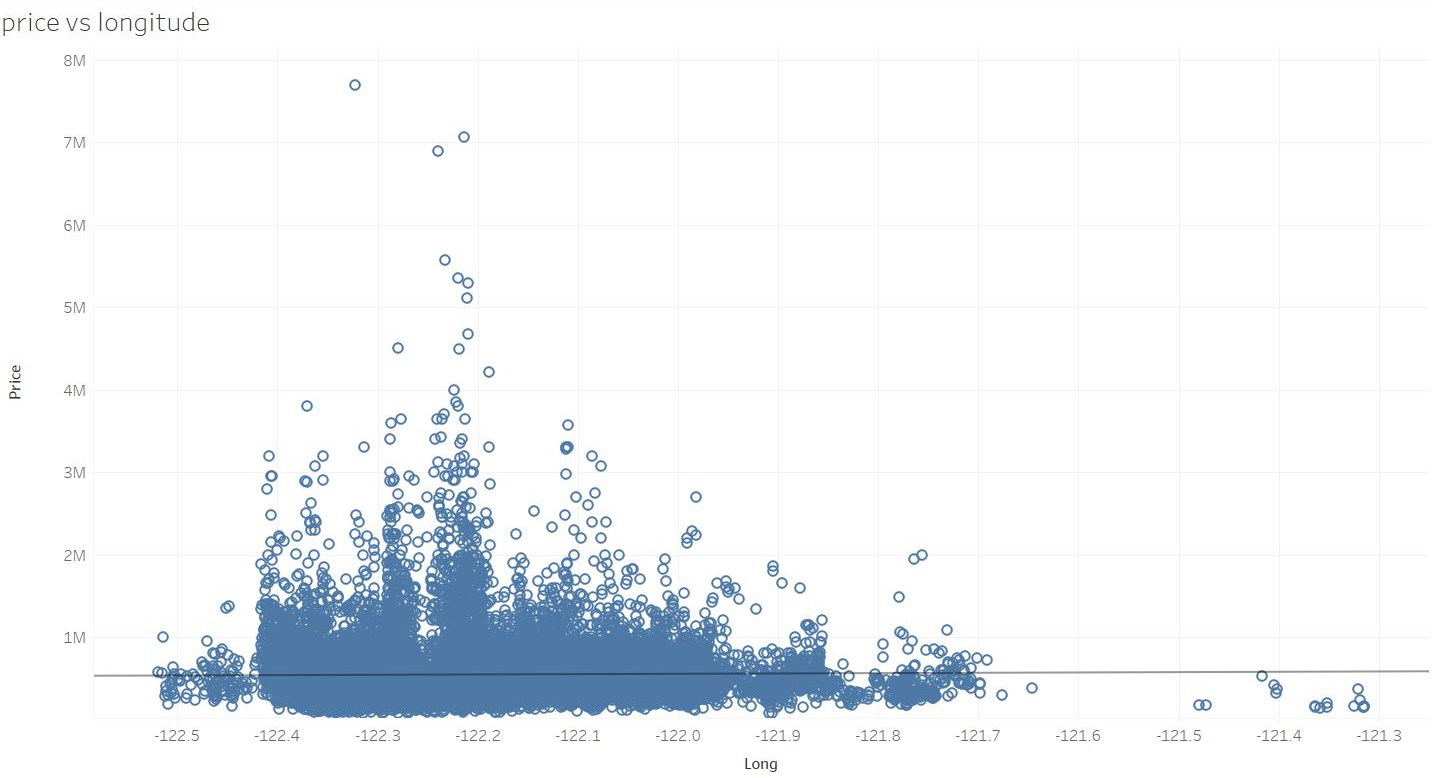
The above graph represents the average house price on the y-axis and the year built on the x-axis. We can see 3 distinct dots near 1940 which are a result of the world war 2. And also the average prices of the houses built around 1960 are relatively low for some reason.

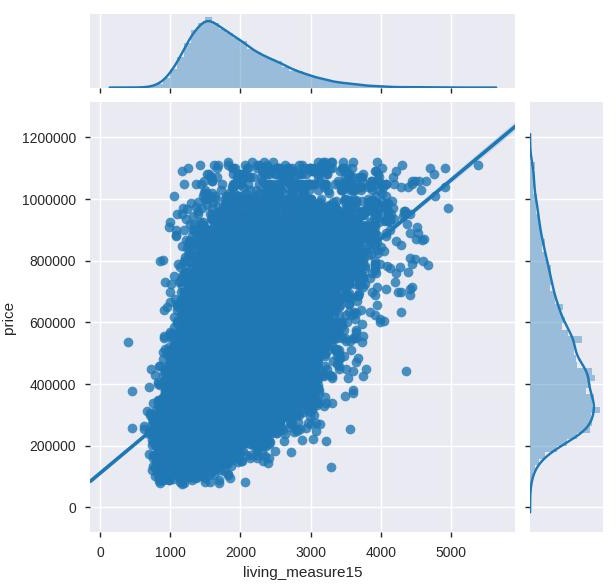
1. **yr\_renovated:** Price increasing with the recency of year of renovation. The below graph has the average price on the y-axis and the year of renovation x-axis. The regression line clearly indicates the relation.



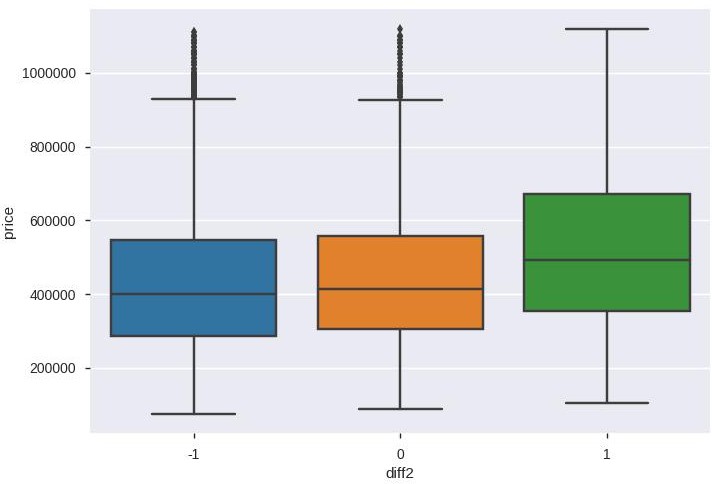
1. **zipcode:**
2. **lat:**

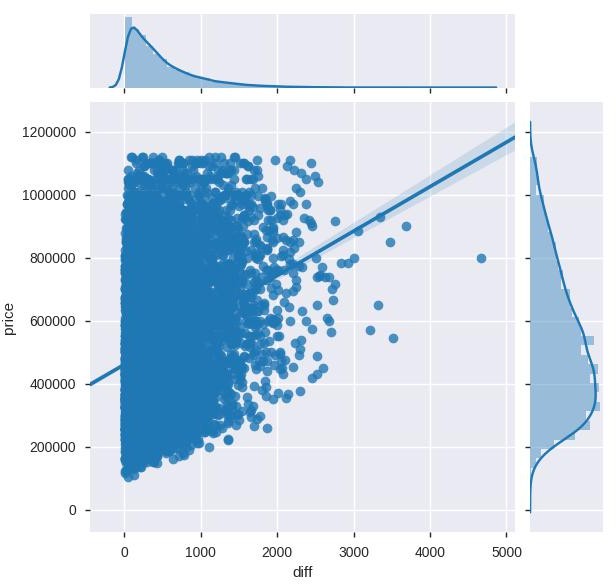


1. **long:**
2. **living\_measure15:** same properties as of living\_measure. Increase in the living\_measure15 resulted in the increase in the price.



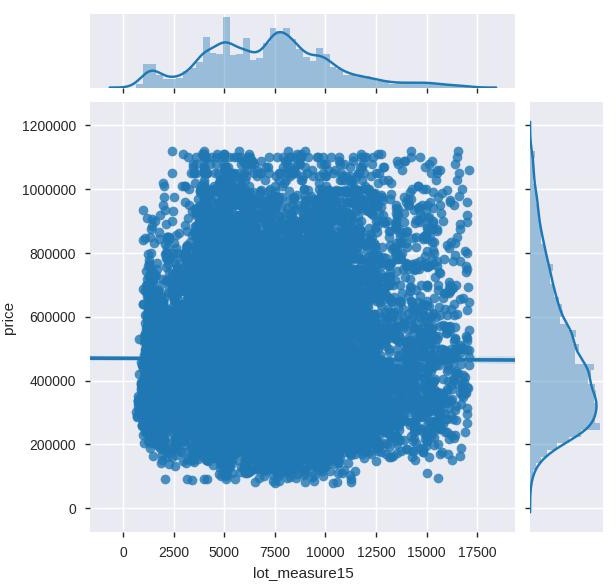
A new column created for determining whether the living measure increased (+1) or decreased (-1) or remained the same(0) compared with the initial living measure is created.

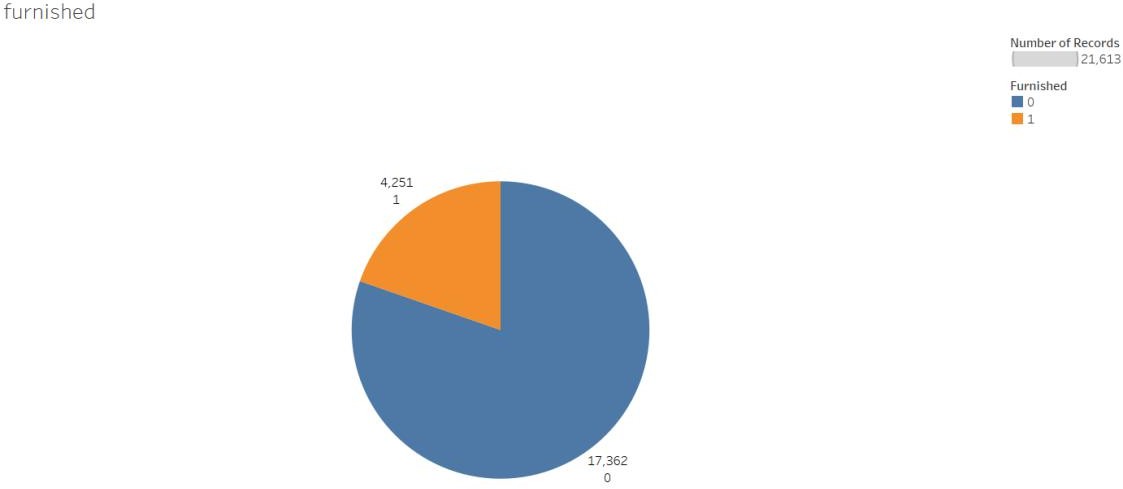
This graph below shows how the price is affected by this column. The price increases where there was an increase in the living measure and the price remained the same for the other two categories.

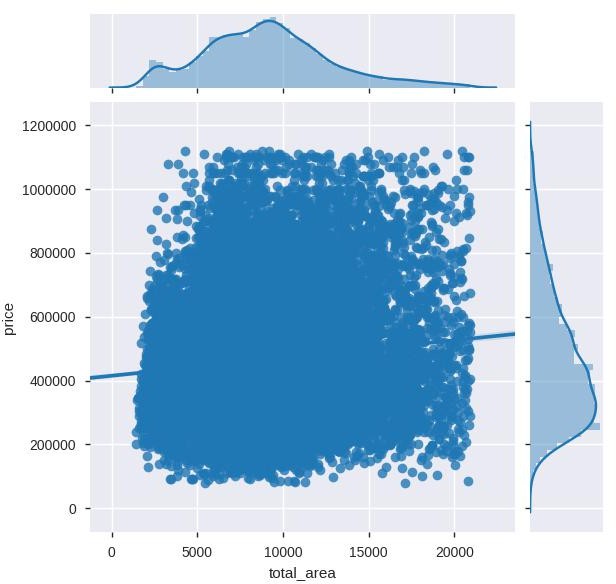


This above graph shows how the price is affected by the amount of increase in the living measure from the initial living measure to the living measure in 2015. Regression line with a positive slope indicating a positive relationship

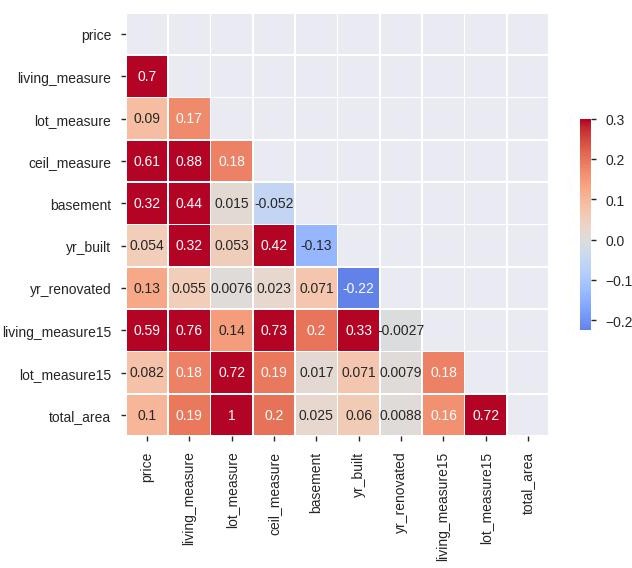
1. **lot\_measure15:** same relationship as that of lot\_measure. The price is not being affected by lot\_measure or lot\_measure15. So they can be dropped for the regression.



1. **furnished:** Houses that are furnished have higher prices than that are not. Univariate analysis:
2. **total\_area:** A large part of the total measure is lot measure which doesn't influence the price but the small part of it which is the living area is the one influencing the price the most. So the total measure is expected to have some influence over the price but that depends on the living measure part of it but not the lot measure as from the previous insights.



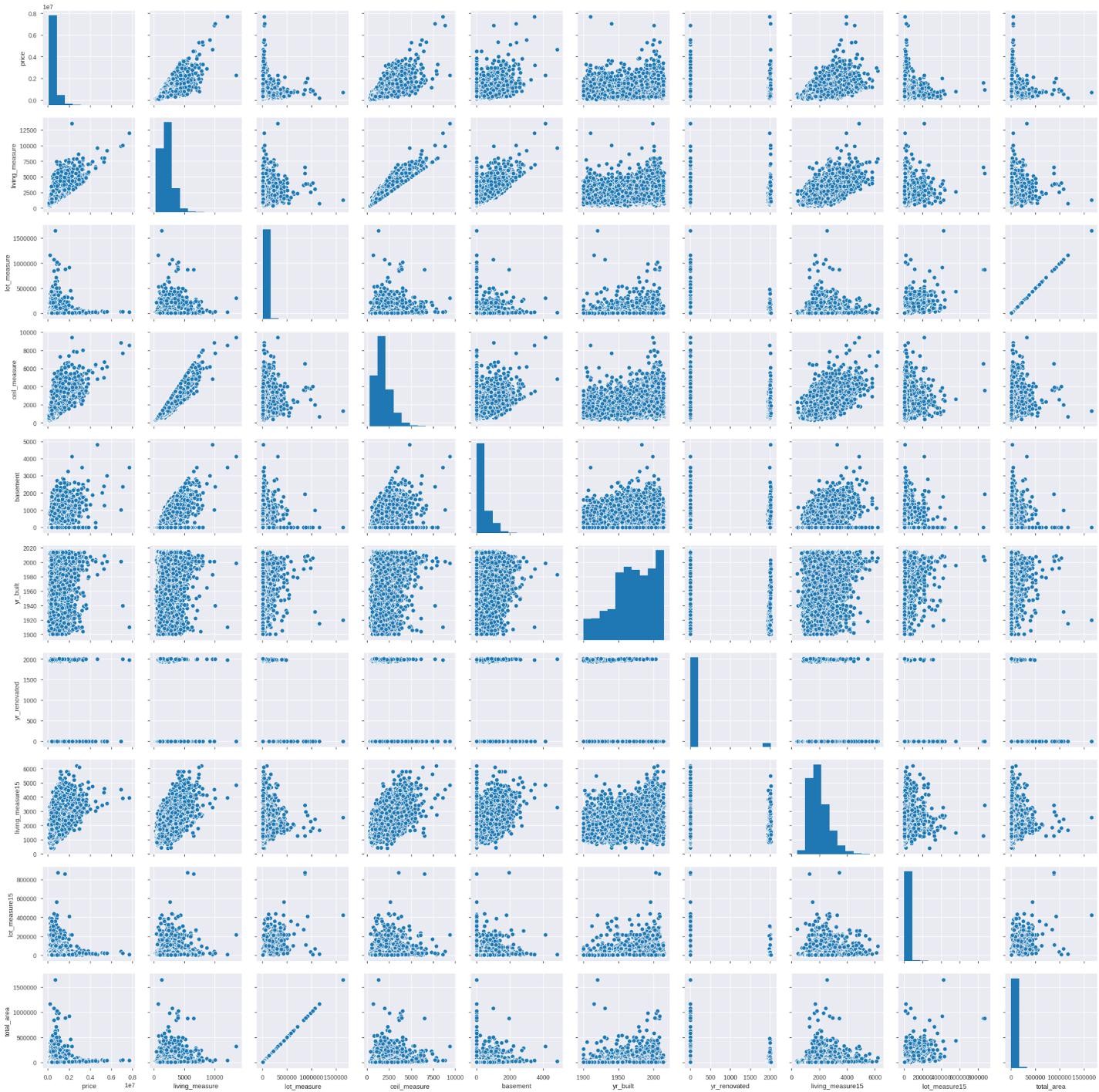
1. **Correlation Matrix:**



From the above correlation matrix we observed that, price is positively correlated with living\_measure, ceil\_measure, basement, living\_measure15. Living\_measure is positively correlated with living\_measure15, ceil\_measure, basement, yr\_built. Lot\_measure is highly correlated with total\_area, positively correlated with lot\_measure15. Ceil\_measure is positively correlated with living\_measure15, yr\_built, highly negatively correlated with basement.

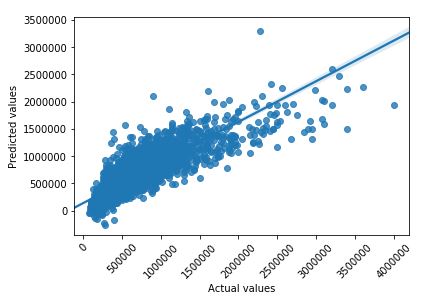
Basement is highly negatively correlated with yr\_built. From the above we can say that Multicollinearity is existing in the features.

1. **Pairplot:**



# BASELINE MODEL:

Linear Regression is used as the baseline model.



It is observed that the R-square value is 0.709. The low value of r2\_score can be due to

heteroscedastic nature of residuals or multicollinearity between the independent variables.

# DATA PREPARATION:

## Outlier Treatment:

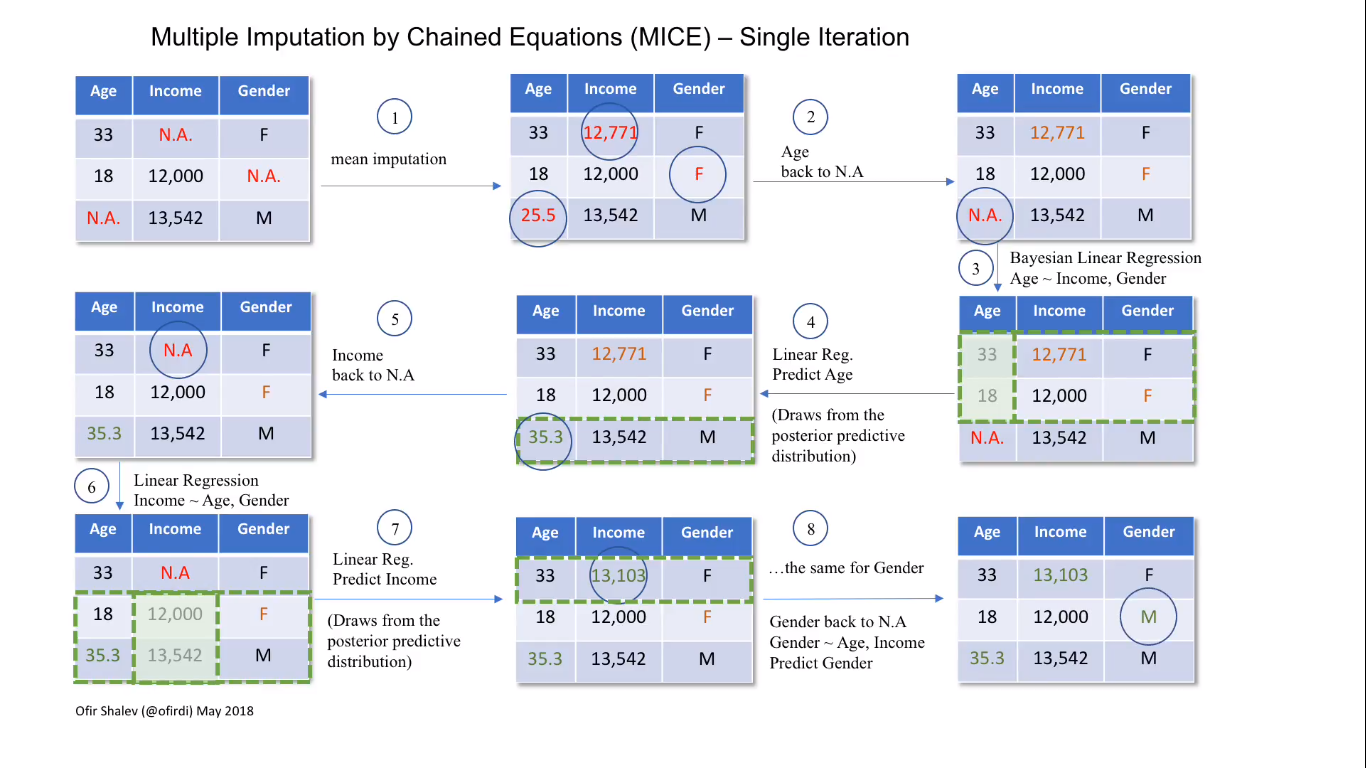
The outliers are lookup row-wise and they accounted for 5.87% of the data, having at least one outlier in each row. As the percentage is significant enough to cause data loss if the rows are removed, they are treated using MICE.

### MICE

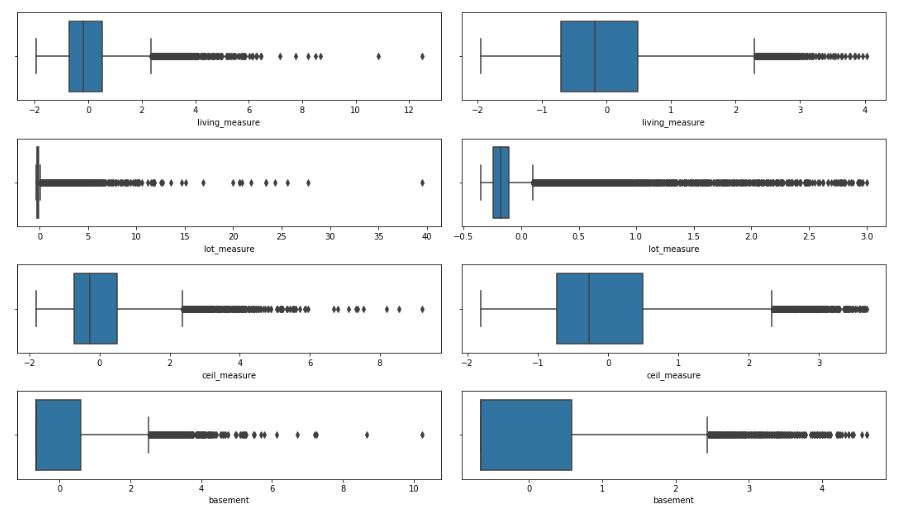
Multivariate Imputation by Chained Equation is one of the most commonly used packages for imputation. Creating multiple imputations as compared to a single imputation (such as mean) takes care of uncertainty in missing values.

The deviations caused by the outliers are minimized by the mice imputed values. As a result there was a significant increase in the R-squared value. And the error terms followed a homoscedastic pattern when visualized in a scatter plot along with the dependent variable.

For example: Suppose we have X1, X2 …. Xk variables. If X1 has missing values, then it will be regressed on other variables X2 to Xk. The missing values in X1 will be then replaced by predictive values obtained. Similarly, if X2 has missing values, then X1, X3 to Xk variables will be used in prediction model as independent variables. Later, missing values will be replaced with predicted values.



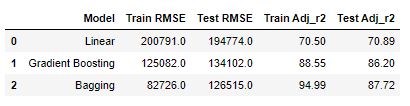
First, the Outliers are converted to null values and then treated with MICE.

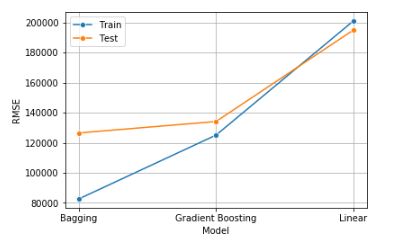


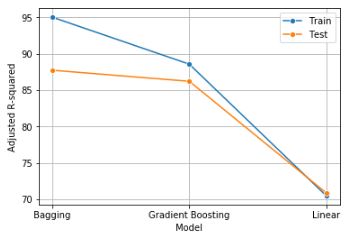
From the above comparison it can be observed that after treating the outliers with MICE, they are converged into approximately 4 standard deviations.

# VII. MODELLING:

As the linear regression model is not giving a good Adj.R-squared score, we tried out decision tree and various ensemble techniques. Among all the models built, Bagging and Gradient Boosting Algorithms were giving good Adj.R-squared scores.







From the above table and plots, it can be observed that though bagging algorithm gives better test score than that from gradient boosting algorithm, the difference in scores of train and test data of bagging algorithm is huge. It implies that model is overfit. So, we chose Gradient Boosting over Bagging for better model.

## Feature Selection:

We have used RFE (Recursive feature engineering) to get the best features from the data. It was observed that we are getting an approximately same adj.r2 score using less number of features.

So, these features can be used to make business recommendations.

# VIII. RECOMMENDATIONS

The final 4 important features are used to build a model and the recommendations to the firm by analysing the coefficients of each feature are as follows:

## 1. living\_measure:

Living measure is the most significant variable and the first impression of the house is given by this measure. The size of the house is also represented by this variable. Most of the regression is explained by this feature alone, so the first thing to compare houses or the first thing to ask for when trying to assess the price of a house is living measure.

## 2. Quality:

This is the 3rd most important feature, and the reason why price increases is very self-explanatory, as with increase in quality the price increases but as observed from the coefficients, the quality of the house is not as significant as the above two features in increasing the price of the house. But if the quality is way higher than average then this is the most significant feature that determines the price of the house. As after the average there is an exponential growth in price with quality.

## 3. Coast:

This is the 2nd most important feature, the prices for houses which are near to coast higher than the not neat to coast.

# XI.REFERENCES:

1. [6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples)](https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779)
2. [Multiple Imputation by Chained Equations (MICE) - YouTube](https://www.youtube.com/watch?v=zX-pacwVyvU)
3. [Variance Inflation Factor (VIF) Explained - Python](https://etav.github.io/python/vif_factor_python.html)