**House Price Prediction**

Team Members -

*Sai Sri Harsha Chakravarthula - 11547144*

*Mallikarjun Narra - 11656443*

*vamshi soleti - 11637574*

*Sindhu Kandimalla - 11642743*

**Overview**:

The Goal of this Project is Estimate

The prices of the Houses so that people can have the general idea about the price of the house that they are going to Buy.

**Introduction:**

The allocation of housing services on the housing market is governed by the laws of supply and demand. The inelasticity of the housing supply is one of the aspects of the housing market that sets it apart from markets for goods and services. One of the most expensive home expenses is housing services. Because they affect socioeconomic conditions and then have an impact on the state of the national economy, changing housing prices have been a source of concern for both individuals and governments. The demand for homes would rise as a result of expectations of capital gains from real estate investments, which would lead to excessive price volatility.To become familiar with the data and discover patterns that can direct the creation of a predictive model, we are conducting an exploratory data analysis (EDA). As part of the EDA process, data exploration is carried out using statistics, correlation analysis, data visualization, and data cleansing. Before creating a model, it is crucial to conduct the process of identifying missing values, outliers, and other issues to be fixed. We will learn more about the variables that affect home prices by looking at the correlations between these variables and the dependent variable (the price of the property). The ultimate goal of EDA is to identify any relevant characteristics and patterns that can predict and explain changes in housing prices.

**Project Flow:**

With the aim that our model is reliable to the

People we gathered the relevant dataset for the

Model building and analysis.

In the Second stage we analyzed the dataset that we got .Here we got an idea how the each attribute is contributing to the target attribute Price.

Then we have done Feature Engineering.Here

We we dropped the attributes that are not contributing for the target attribute and created some attributes by using the existing attributes.

In the Final Part we developed the models like

Logistic Regression and decision tree .so,the people can use our models and make predictions.

**Brainstorming**: The main motivation of the project is to predict the prices of Houses by considering certain Factors like no of bed rooms,size..etc.First we have taken data from the kaggle and the analyze the data by considering the factors that are contributing to the price.

Then We have done the feature engineering

Dropped some attributes that are not contributing value to project and created the attributes that are most important to the project.

Then we developed the machine learning models and make the models available to the public.

**Critics**:

We have analyzed the data and done the feature engineering and get the most valuable attributes that are good for prediction and then we developed the two machine learning models like logistic regression and decision tree.

**Goal:** Learning more about the elements that influence home values, such as location, square footage, the number of bedrooms and bathrooms, the age of the property, and extras like a swimming pool or a garage, The main goal of EDA is to analyze data and knowing about the attributes and do feature engineering

To know about the features that are contributing to project and build machine learning models.

**Specification:** We are going to utilize a machine learning models-based architecture and CNN based architecture because this challenge involves excel data, and if we are not happy with the accuracy to forecast the price of the property based on the input we provide, such as the number of bedrooms, square feet area, etc.

**Responsibilities:**

Harsha:Data Preprocessing,Feature selection,Documentation,Statistical analysis

Mallikarjun:Statistical analysis,Data Preprocessing,Documentation,Model Building

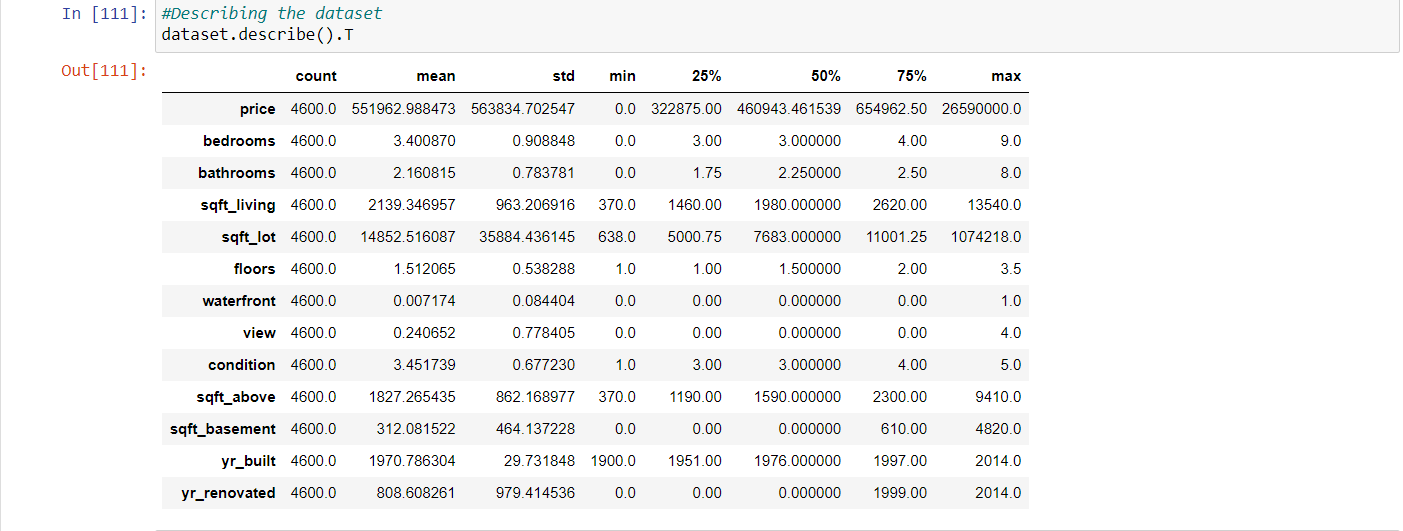
Vamshi:Data Preprocessing ,Documentation,

Model Building ,statistical analysis

Sindui:Data Preprocessing ,Documentation,

Model Building ,statistical analysis

**Output:**



This graphic displays the data description values for each column of the data frame, including the mean, mode, median, minimum value for that column, standard deviation values, and Q1, Q2, and Q3 values.

**Parameter estimation :**

The top 3 correlated columns for the price attribute are sqft\_living, sqft\_above, and bedrooms.

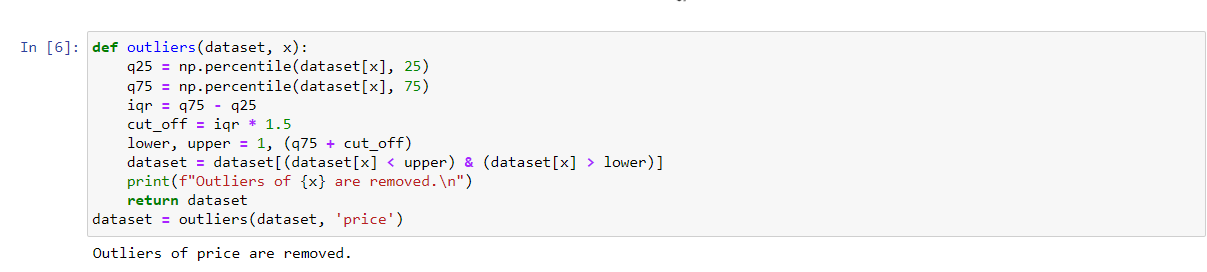
The price median is 551962.988473, the price standard deviation (SD) is 563834.702547, the price column's 25% (Q1) value is 322875.00, the price column's 50% (Q2) value is 460943.461539, and the price column's 75% (Q3) value is 654962.50.

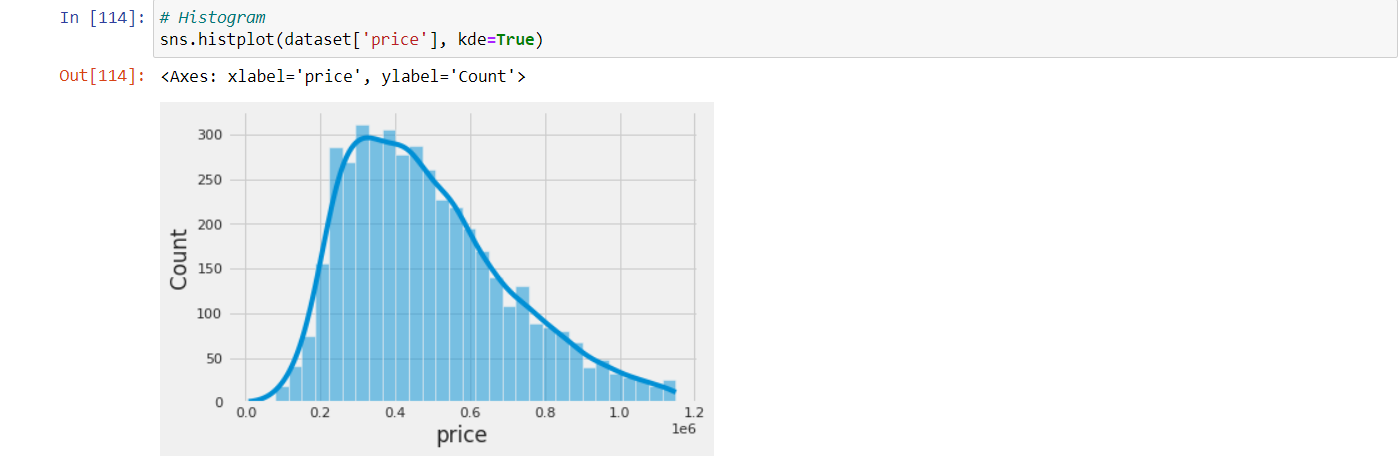
The average cost of living in square feet is 2139.346957; the standard deviation (SD) is 963.206916; 25% (Q1) is 1460.00; 50% (Q2) is 1980.000000; and 75% (Q3) is 2620.00.

The mean for sqft\_above is 1827.265435, the standard deviation (SD) is 862.168977, and the 25% (Q1), 50% (Q2), and 75% (Q3) of sqft\_above are each calculated as follows.

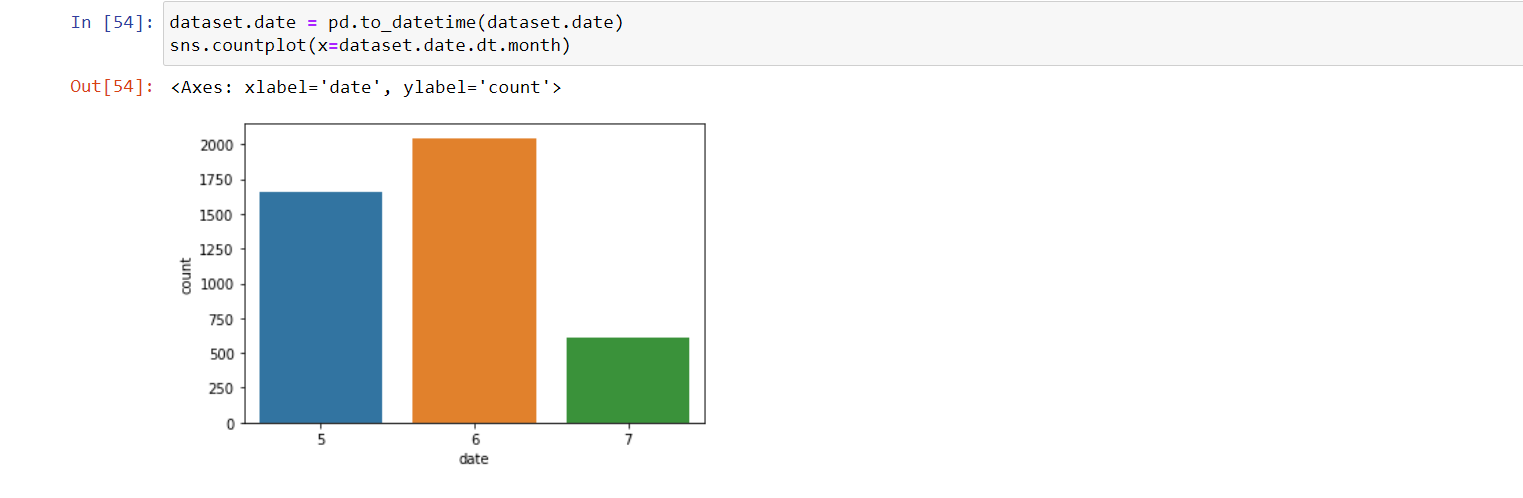
The average bathroom has a value of 2.160815, the standard deviation (SD) is 0.783781, 25% of bathrooms have a value of 1.75, 50% have a value of 2.250000, and 75% have a value of 2.50.

After that, we removed any and all outliers from the data sets, for which we removed the top 25 records and bottom 75 records for removing outliers for prices based on the IQR (interquartile range), indicating that we only considered the in-between values. We also removed all low ranged values and zero values from the data sets. to create data that is reliable and to achieve good correlation.

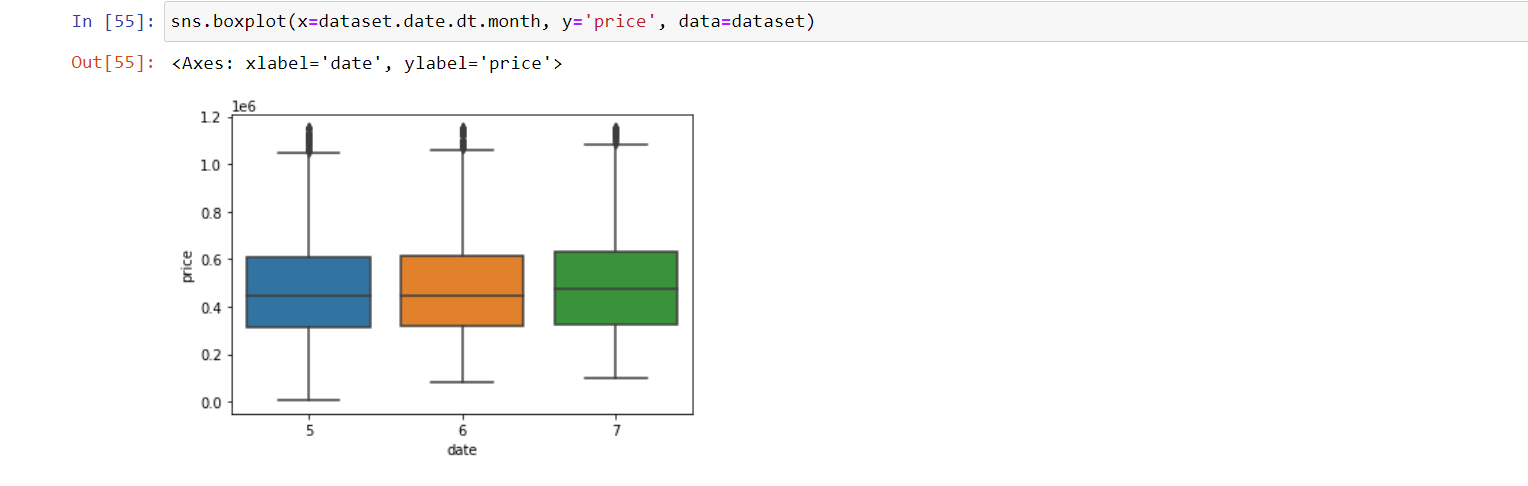




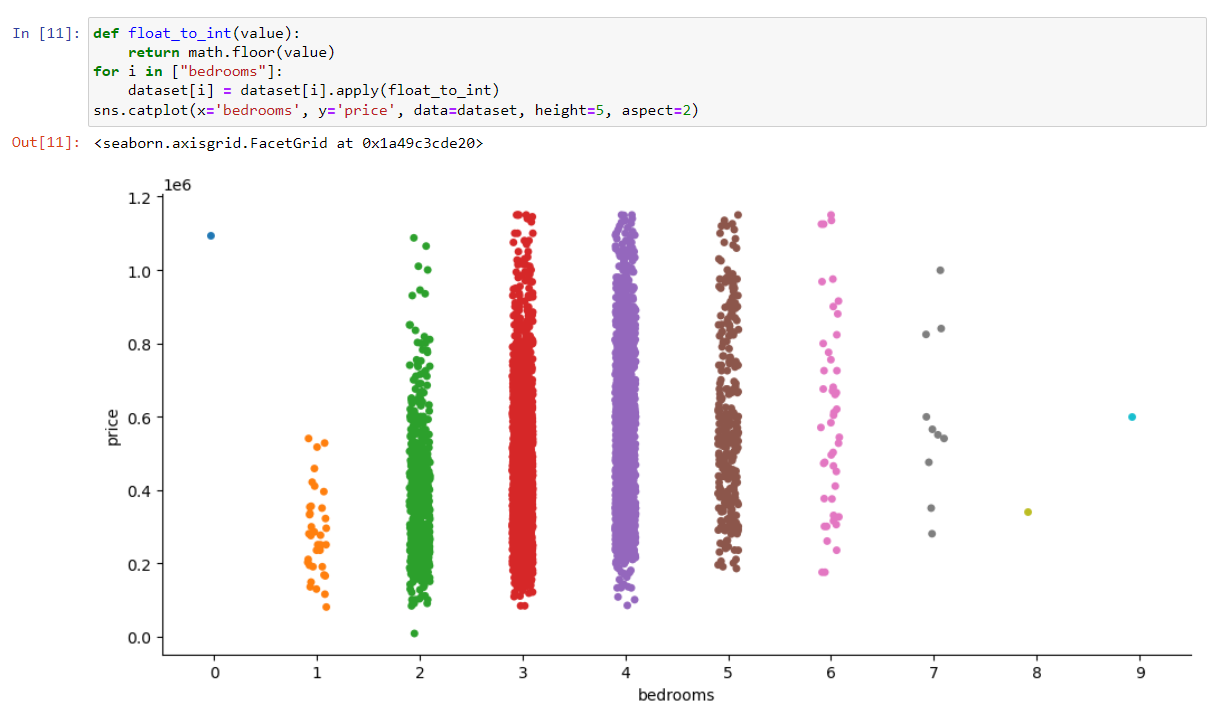
The price of the house is plotted against the number of houses with those pricings in this histogram. The number of homes in that certain price range is shown in this graph. For instance, there are around 75 residences totaling $1,250,000.

**

The analysis of the number of houses sold month by month is shown in this bar plot, which shows the count vs. months. For instance, the number of houses available in the seventh month was 600, which is fewer than in the sixth month.



This box plot shows pricing month by month, and we may infer the minimum, mean, and highest prices from this graph. For instance, we might state that the minimum cost of a house sold in the fifth month is 1.3 million dollars.



This graph illustrates the amount of bedrooms we can have based on the price range and displays the relationship between pricing and bedrooms. For instance

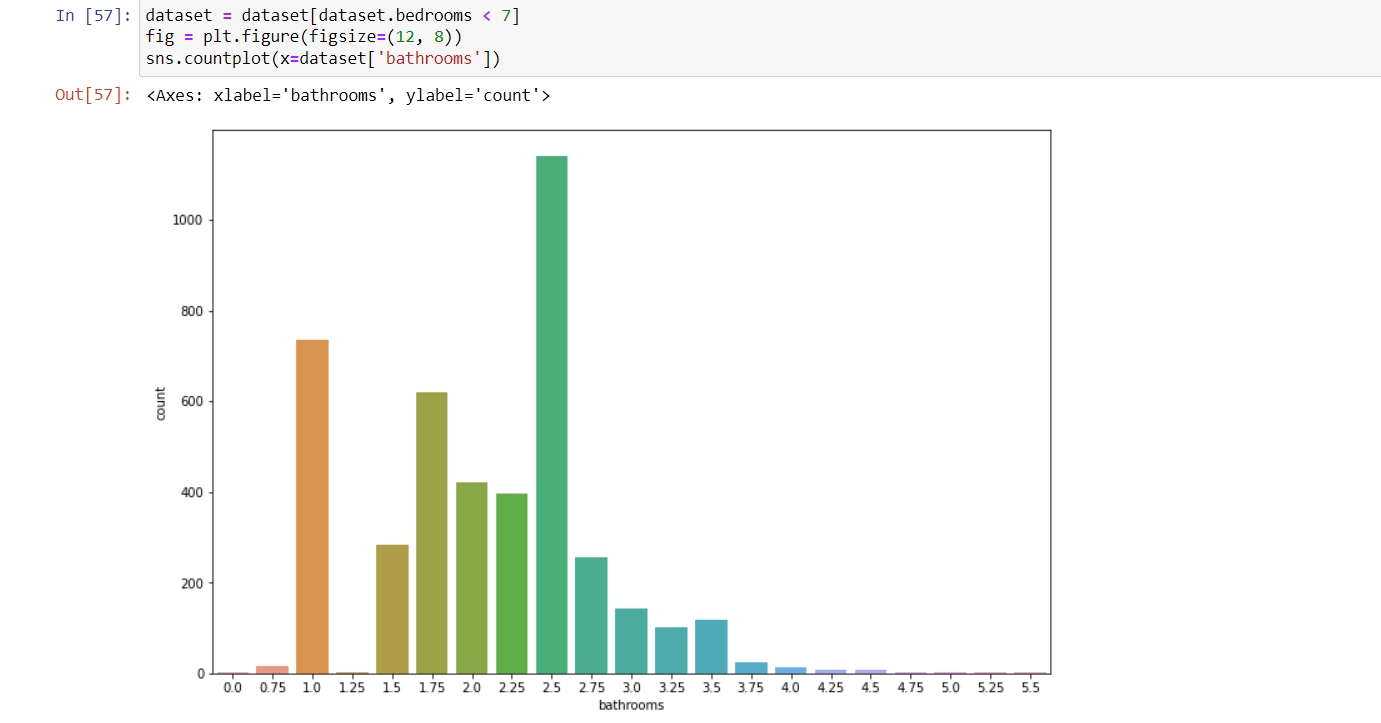


Fig. 1

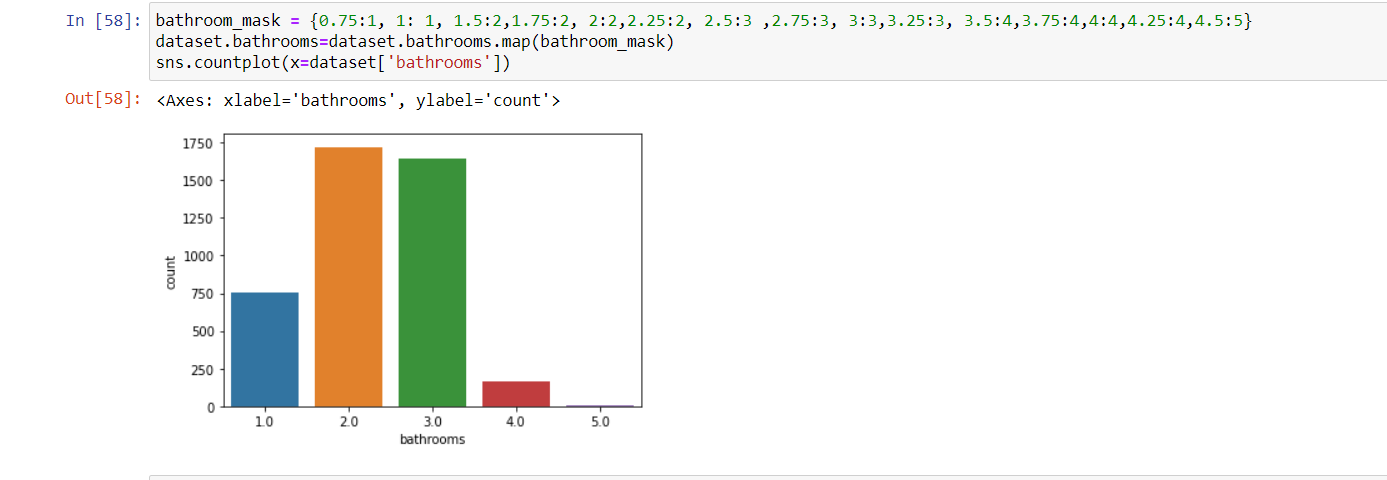


Fig. 2

Fig. 1 displays the count of bathrooms without rounding the values, whereas Fig. 2 shows the count with rounding the numbers. Both graphs, Fig. 1 and Fig. 2, indicate the number of households and bathrooms in them. One bathroom homes, for instance, are available in around 750 homes.

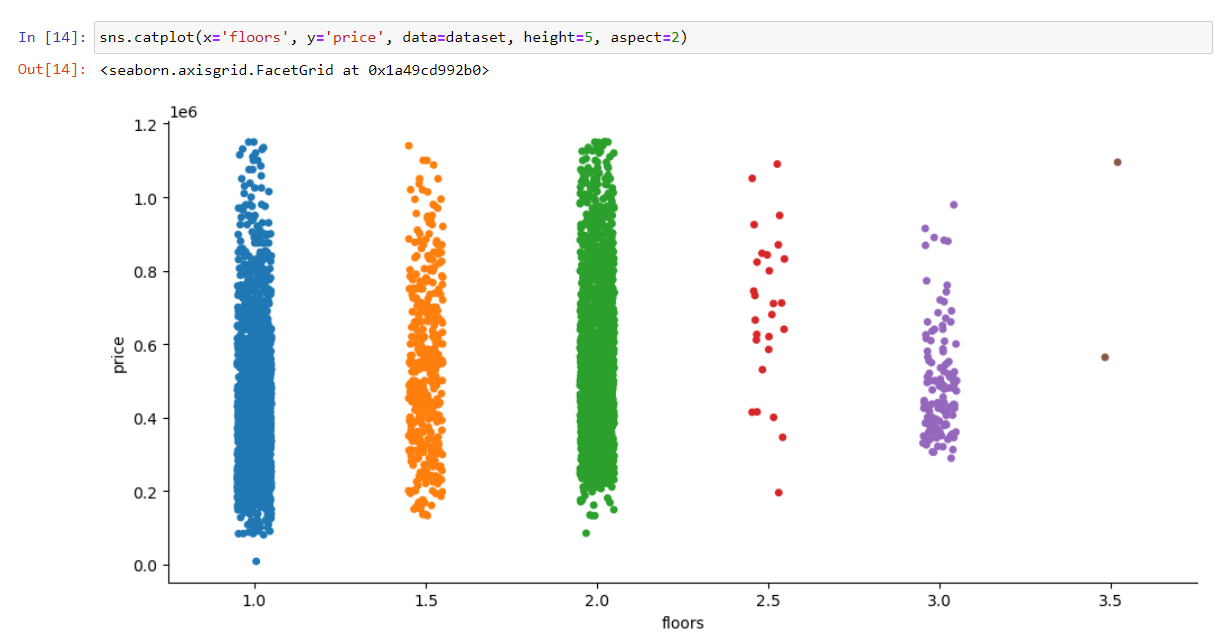
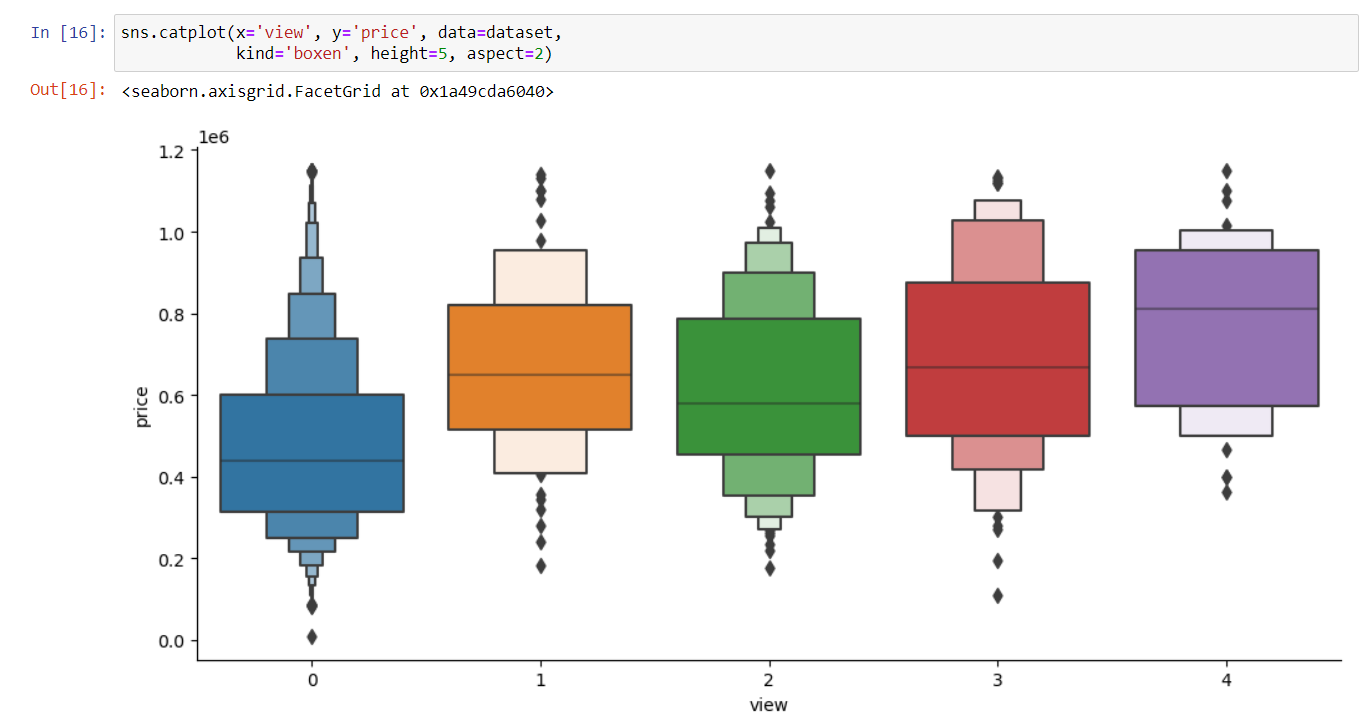


Fig.1



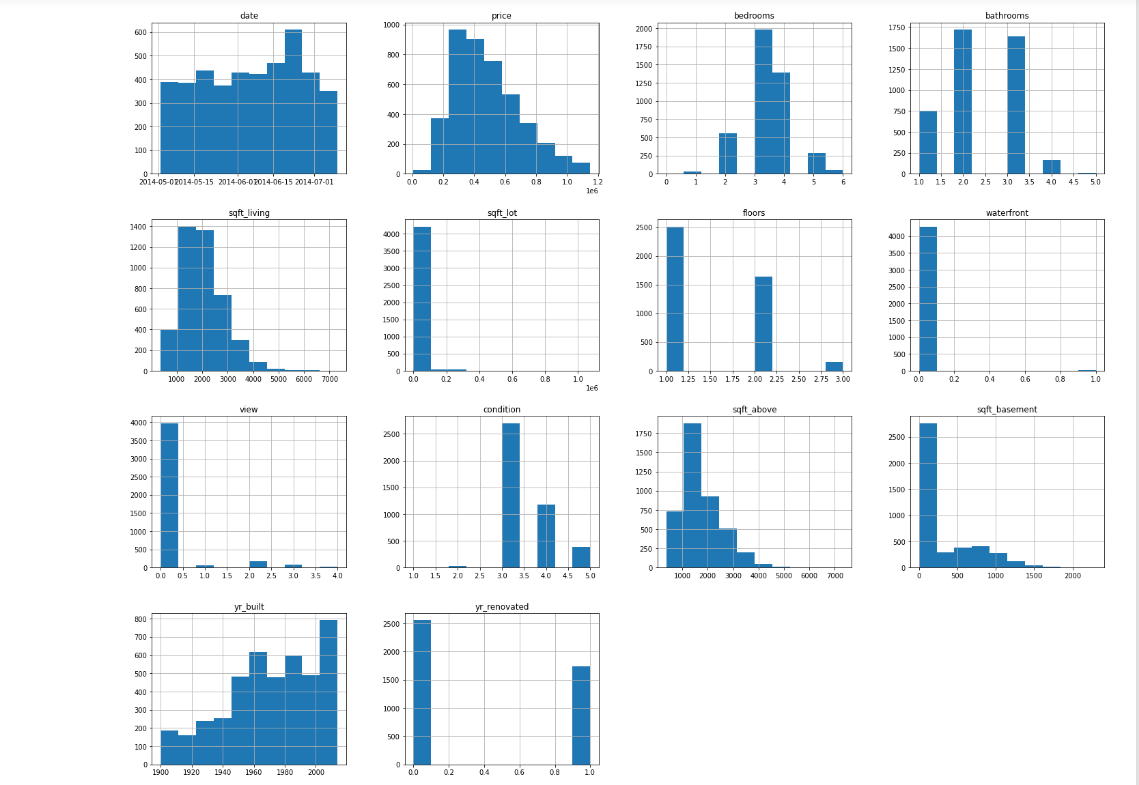
Fig. 2

Both Figs. 1 and 2 above show the price in relation to the number of floors. Fig. 1 shows the price vs. the number of floors without rounding the values, while Fig. 2 shows the price vs. the number of floors with rounding the values. For instance, if we look at the last green dot in the three-floor design, we can estimate that its cost may range from 1.2 million to 1.4 million.

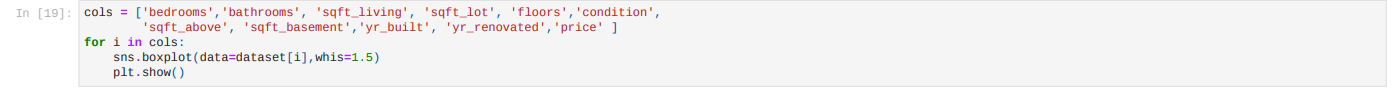


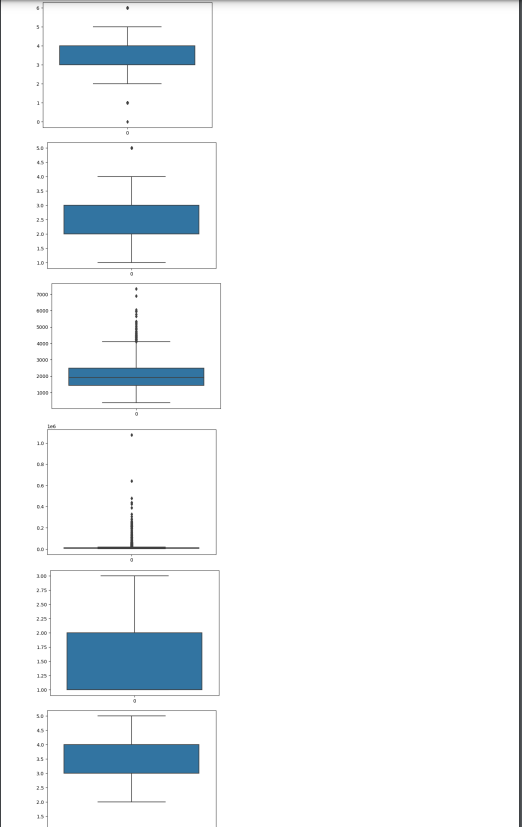
This graph shows the link between price and review. It shows the price in relation to the review. For instance, based on this graph, we might infer that homes with favorable evaluations typically have higher average values than similar properties.

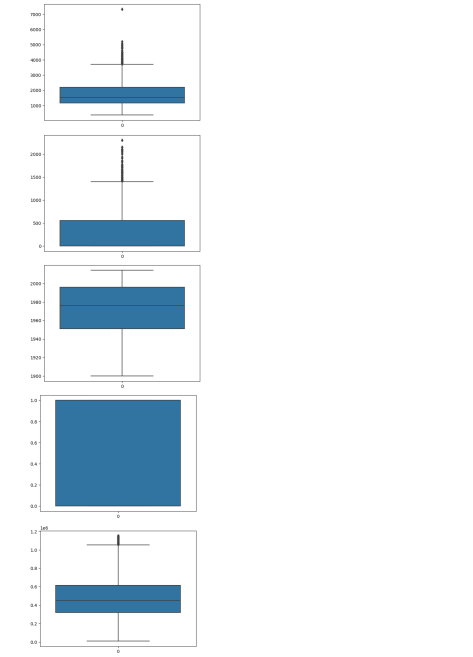
The price versus refurbished status is seen in this graph. The association between a home's price and whether it has been renovated or not may be objectively determined.

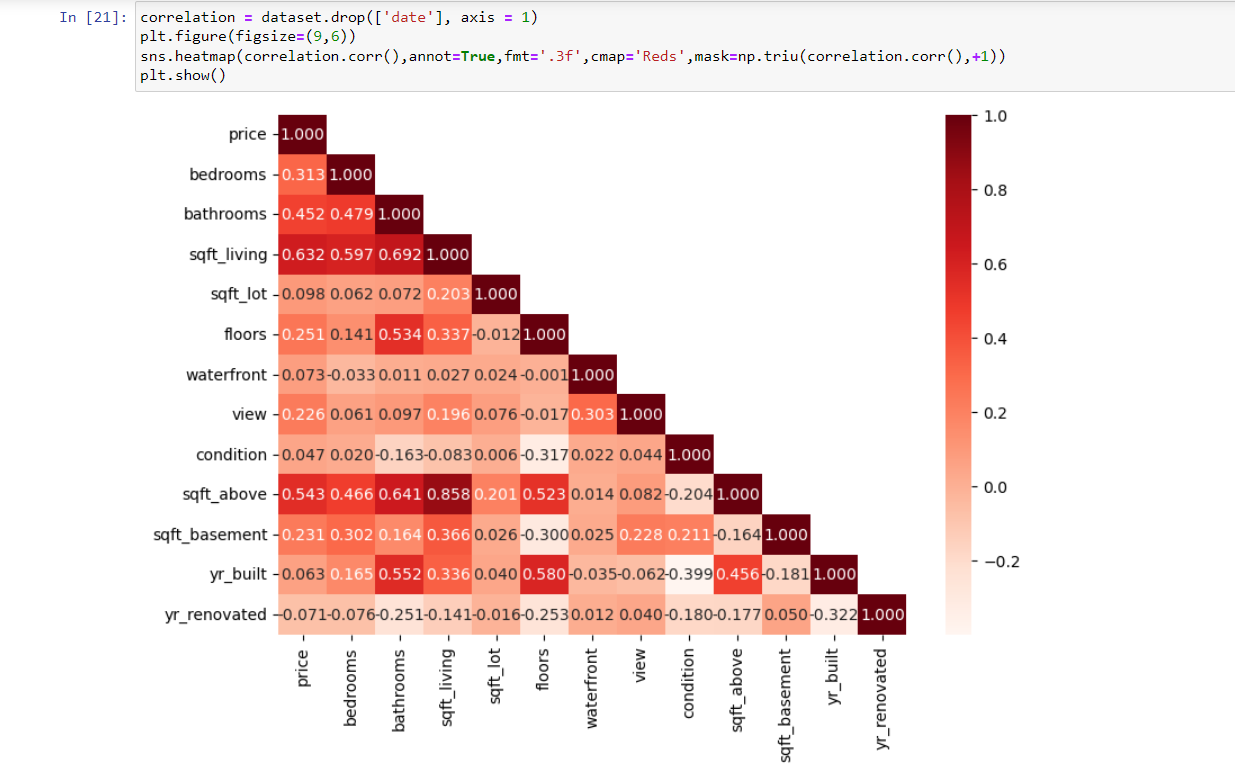


These graphs up top are just ordinary histograms without the EDA, with the title of the graph as the x-axis and the number of houses on the y-axis. From these graphs, we can only objectify a few characteristics of homes for reference: graph 1 shows the number of homes sold month over month, and graph 2 shows the number of homes sold in that price range.





These graphs are just standard box plots without an EDA, and the x-axis is only one column from an ordered list of the number of bedrooms, bathrooms, square feet of living space, square footage of the lot, floors, condition, square footage above and below grade, and price of the graph. The y-axis represents the number of dwellings. These graphs allow us to make inferences for our own self-analysis. For instance, graph 1 shows the number of bedrooms for each house, whereas graph 2 shows the number of bathrooms per house.



This graph shows the correlation between the columns in the various data sets that were used. For instance, the correlation shows how strongly the values of the various columns are related; if the correlation is close to +1, then there is a good correlation between those columns, and if it is moving towards -1, then there is a bad correlation. Since price and sqft\_living have a good correlation, we can infer from the first column, for instance, that price and sqft\_above have a good correlation after these values.

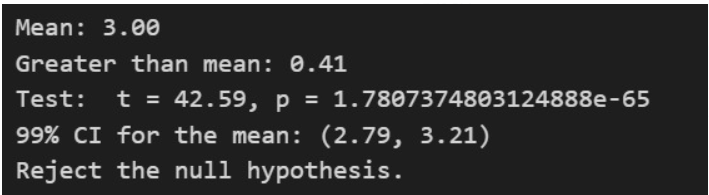
**Conclusion 1:**

**Observation 1:**

When we considered 100 random samples with null hypothesis harsha has considered samples from

['sqft\_living','renovated','bedrooms','yr\_built','price']

Where i have considered null hypothesis as the overall number of bedrooms are less than the mean of bedrooms in the samples.



Here from the above output as p<0.05 and 99% confident interval of mean is between (2.79, 3.21).

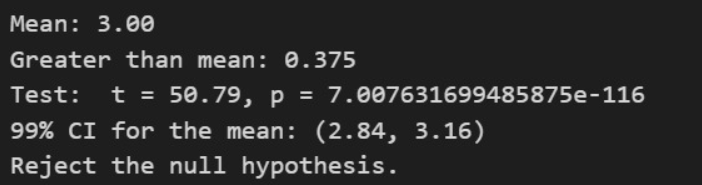
So,we can reject the null hypothesis and considered that overall number of bedrooms is >3 or mean.

**Observation 2:**

When we considered 200 random samples with null hypothesis harsha has considered samples from

['sqft\_living','renovated','bedrooms','yr\_built','price']

Where i have considered null hypothesis as the overall number of bedrooms are less than the mean of bedrooms in the samples.



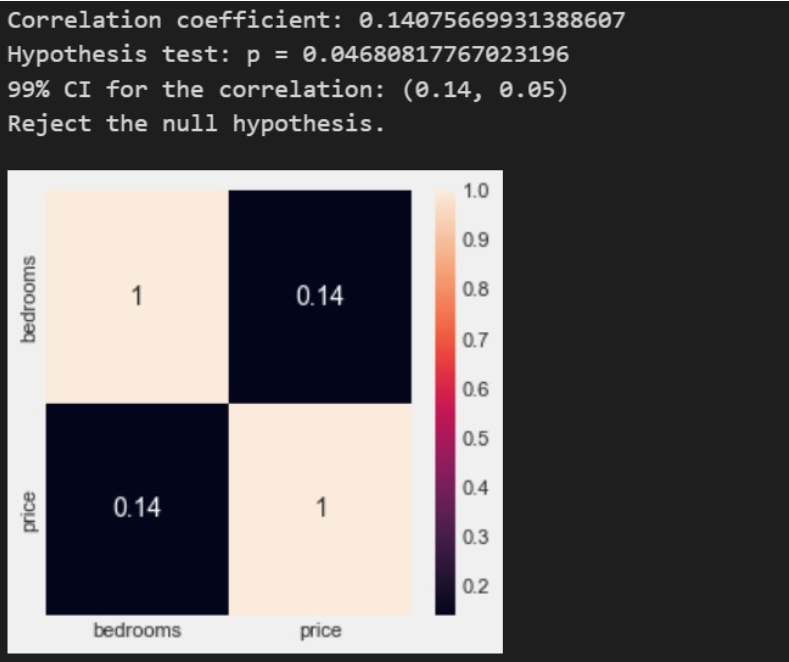
Here from the above output as p<0.05 and 99% confident interval of mean is between (2.84, 3.16).

So,we can reject the null hypothesis and considered that overall number of bedrooms is >3 or mean.

**Observation 3**

When considering the correlation of 200 samples the correlation between the number of bedrooms and the price.i.e.,selected variable and target variable is

I have hypothesised that both the variables are not corelated



Correlation coefficient: 0.14075669931388607

Hypothesis test: p = 0.04680817767023196

99% CI for the correlation: (0.14, 0.05)

Hence ,we interpretation that we can consider that both the variables are correlated with Correlation coefficient: 0.14 and

Reject the null hypothesis.

**Observation 4:**

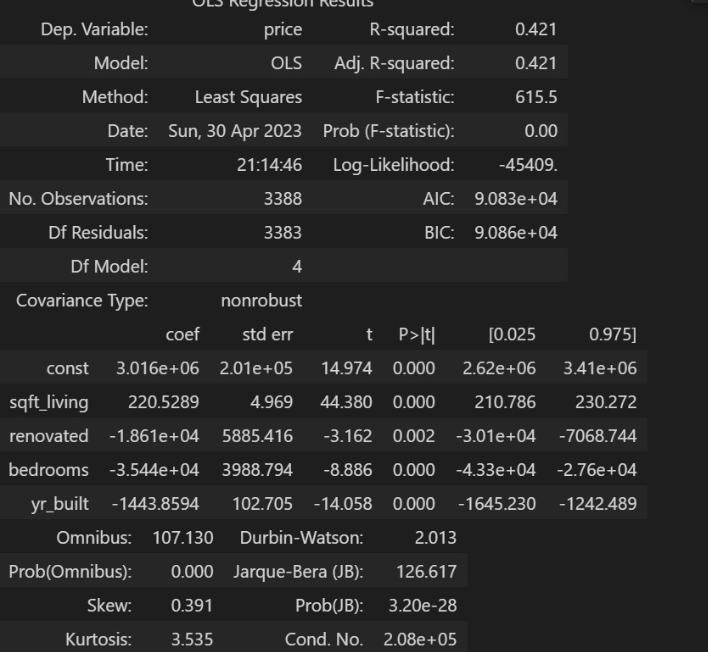
When we considered all the bedrooms as the variable and predicted the price using OLS Regression the R2 score is: 0.16962236102363082



**Observation 5**

When we consider all the variables in observation 1 we got adjusted R2 score is 0.4205257635079638

And result to be



**Observation 6**

The Multi Regression equation is

The multi regression equation is : Y = 3.016e+06 + 220.5289 \* (sqft\_living) + -1.861e+04 \* (renovated) + -3.544e+04 \* (bedrooms) + -1443.8594 \* (yr\_built).

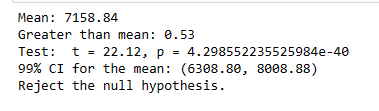
**Conclusion 2:**

**Observation 1:**

When we considered 100 random samples with null hypothesis. sindhu has considered samples from

[‘sqft\_lot’,'bathrooms','waterfront','condition','price']

Where I have considered null hypothesis as the overall number of sqft\_lot are less than the mean of sqft\_lot in the samples.



Here from the above output as p<0.05 and 99% confident interval of mean is between (6308.80, 8008.88).

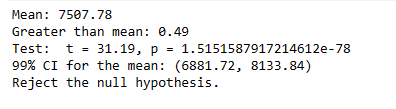
So,we can reject the null hypothesis and considered that overall number of sqft\_lot is greater than the mean.

**Observation 2:**

When we considered 200 random samples with null hypothesis sindhu has considered samples from

[‘sqft\_lot’,'bathrooms','waterfront','condition','price']

Where i have considered null hypothesis as the overall number of sqft\_lot are less than the mean of sqft\_lot in the samples.

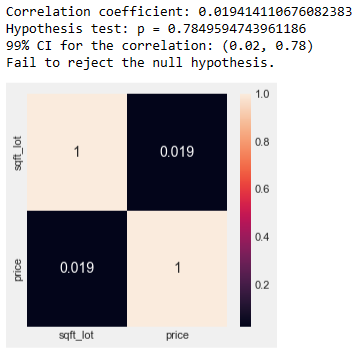
****

Here from the above output as p<0.05 and 99% confident interval of mean is between (6881.72, 8133.84).

So,we can reject the null hypothesis and considered that overall number of sqft\_lot is greater than the mean.

**Observation 3**

When considering the correlation of 200 samples the non correlation between the number of sqft\_lot and the price.i.e.,selected variable and target variable is I have performed hypothesis test and can conclude that both the variables are not correlated.



Correlation coefficient: 0.019414110676082383

Hypothesis test: p = 0.7849594743961186

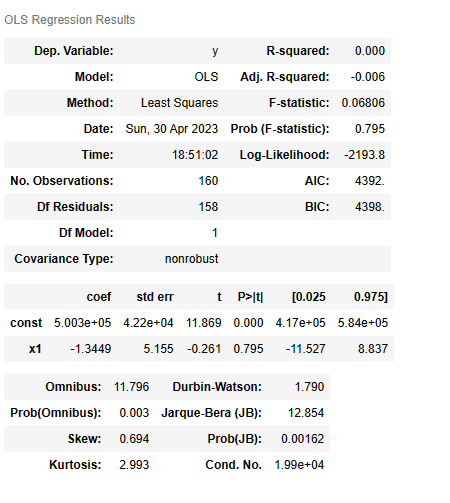
99% CI for the correlation: (0.02, 0.78)

Hence ,we interpretation that we can consider that both the variables are not correlated as there is no evidence with Correlation coefficient: 0.019 and hence We fail to reject the null hypothesis.

**Observation 4:**

When we considered all the sqft\_lot as the variable and predicted the price using OLS Regression the R2 score is:

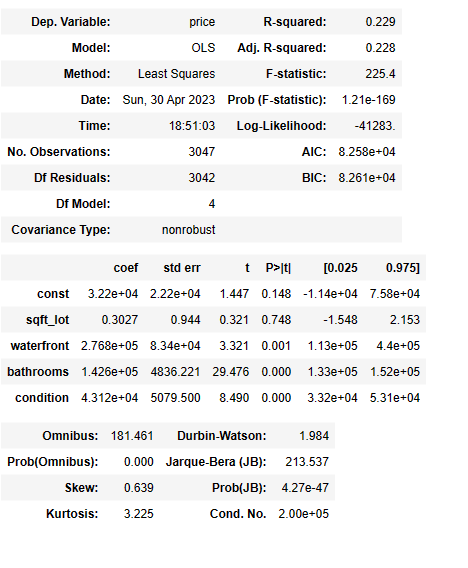
-0.05572964783255974



**Observation 5**

When we consider all the variables in observation 1 we got adjusted R2 score is 0.22759786323666764

And result to be



**Observation 6**

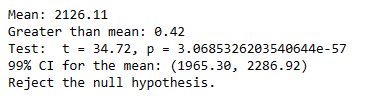
The multi regression equation is : Y = 3.22e+04 + 0.3027 \* (sqft\_lot) + 2.768e+05 \* (waterfront) + 1.426e+05 \* (bathrooms) + 4.312e+04 \* (condition)

**Conclusion 3:**

**Observation 1:**

When we considered 100 random samples with null hypothesis Mallikarjun has considered samples from

['sqft\_living','sqft\_above','bathrooms','sqft\_basement','price']

Where i have considered null hypothesis as the overall number of sqft\_living are less than the mean of 

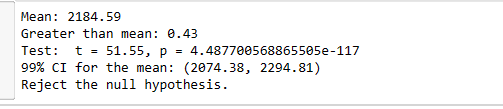
Here from the above output as p<0.05 and 99% confident interval of mean is between (1965.30, 2286.92). So,we can reject the null hypothesis and considered that overall number of sqft\_living is >3 or mean.

**Observation 2:**

When we considered 200 random samples with null hypothesis Mallikarjun has considered samples from

['sqft\_living','sqft\_above','bathrooms','sqft\_basement','price']

Where i have considered null hypothesis as the overall number of 'sqft\_living' are less than the mean of 'sqft\_living' in the samples.

****

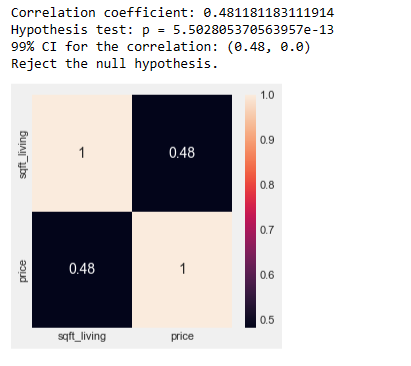
Here from the above output as p<0.05 and 99% confident interval of mean is between (2074.38, 2294.81)

So,we can reject the null hypothesis and considered that overall number of 'sqft\_living' is >3 or mean.

**Observation 3**

When considering the correlation of 200 samples the correlation between the number of 'sqft\_living' and the price.i.e.,selected variable and target variable is

I have hypothesised that both the variables are not corelated



Correlation coefficient: 0.481181183111914

Hypothesis test: p = 5.502805370563957e-13

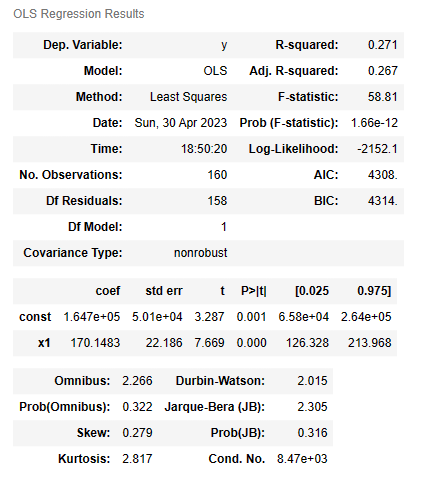
99% CI for the correlation: (0.48, 0.0)

Hence ,we interpretation that we can consider that both the variables are correlated with Correlation coefficient: 0.48 and

Reject the null hypothesis.

**Observation 4:**

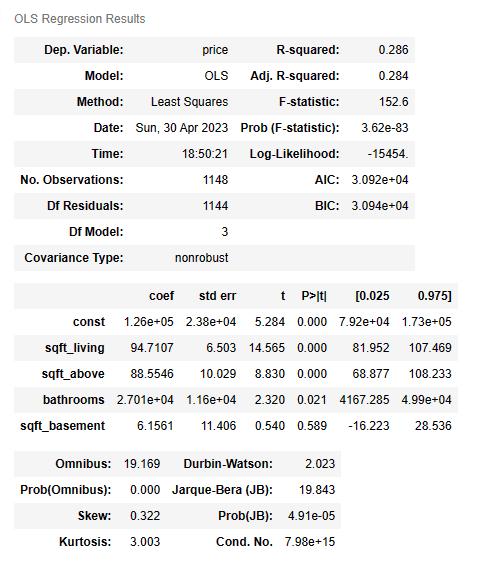
When we considered all the sqft\_living as the variable and predicted the price using OLS Regression the R2 score is: 0.004314193221751195



**Observation 5**

When we consider all the variables in observation 1 we got adjusted R2 score is 0.28388874514549867

And result to be



**Observation 6**

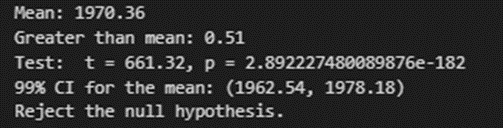
The Multi Regression equation is

The multi regression equation is : Y = 1.26e+05+ 94.7107\* (sqft\_living) + 88.5546\* (sqft\_above) + 2.701e+04 \* (bathrooms) + 6.1561\* (sqrt\_basement).

**Conclusion 4:**

**OBSERVATION 1:**

**When we considered 100 random samples with null hypothesis vamshi has considered samples from ['sqft\_living','renovated','bedrooms','yr\_built','price'] Where i have considered null hypothesis as the overall number of bedrooms are less than the mean of bedrooms in the samples.**

****

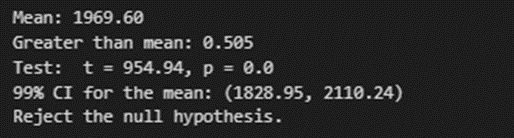
**Here from the above output as p<0.05 and 99% confident interval of mean is between (1962.54, 1978.18).**

**So,we can reject the null hypothesis and considered that overall**

**number of bedrooms is >3 or mean.**

**OBSERVATION 2:**

When we considered 200 random samples with null hypothesis vamshi has considered samples from ['sqft\_living','renovated','bedrooms','yr\_built','price'] Where i have considered null hypothesis as the overall number of bedrooms are less than the mean of bedrooms in the samples.

****

Here from the above output as p<0.05 and 99% confident interval of mean is between (1828.95, 2110.24).

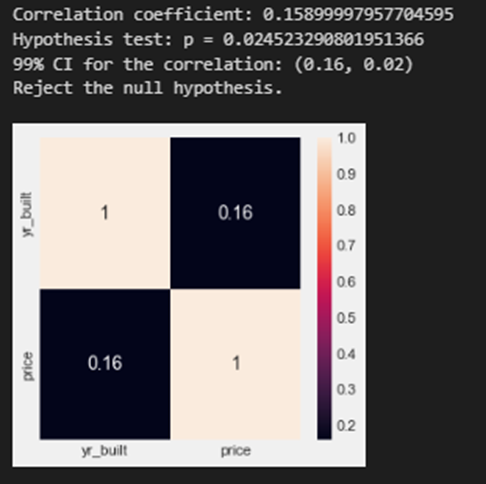
So,we can reject the null hypothesis and considered that overall

number of bedrooms is >3 or mean.

**OBSERVATION 3:**

When considering the correlation of 200 samples the correlation between the number of bedrooms and the price.i.e.,selected variable and target variable is

I have hypothesised that both the variables are not corelated

****

Correlation coefficient: 0.15899997957704595

Hypothesis test: p = 0.024523290801951366

99% CI for the correlation: (0.16, 0.02)

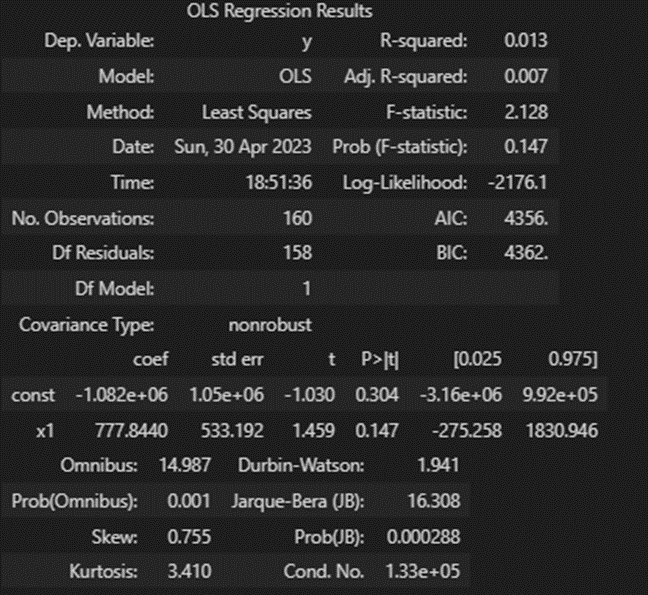
Hence ,we interpretation that we can consider that both the variables are correlated with Correlation coefficient: 0.14 and

Reject the null hypothesis.

**Observation 4:**

When we considered all the bedrooms as the variable and predicted

the price using OLS Regression the R2 score is: 0.0015962283355622775

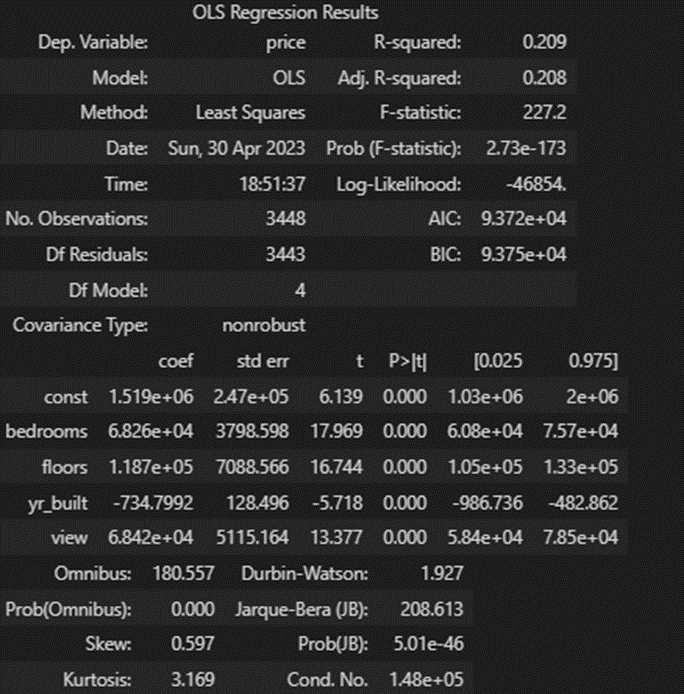
****

**Observation 5**

When we consider all the variables in

observation 1 we got R2 score is 0.4205257635079638

And result to be

****

**Observation 6**

The Multi Regression equation is

The multi regression equation is : Y =

1.519e+06 + 6.826e+04 \* (bedrooms) + 1.187e+05 \* (floors) + -734.7992 \*

(yr\_built) + 6.842e+04 \* (view)

**References** :

* <https://www.sciencedirect.com/science/article/abs/pii/S0304393206001279?casa_token=rxkEyVnDILEAAAAA:-PAT1yegiflHpCfcGhoeffHHcVg8LVmtuHDB3DpBKgakbamXPrkz8VSSHOeUVKr1Hbpo7Cn7>
* <https://www.aeaweb.org/articles?id=10.1257/aer.101.5.2108>
* <https://onlinelibrary.wiley.com/doi/abs/10.1111/jmcb.12011?casa_token=FYjWzmV_vskAAAAA:TSRb81iCe-NrkW7h6L85jsWhmtJzUbqbJXh76WW6-Q3dstOcMo14H-JNPw2PY1sgBwjgrTNZ7mPnvQ>
* <https://www.sciencedirect.com/science/article/abs/pii/S0378426609002088?casa_token=-INSzlJWZfEAAAAA:TSjTL3G1gHQOChwJ2rCBCDxr7RHAgJTE3b_UFLq_sYpZpDCVx8DkXwu6GuxgN97_vToSuNQZ>
* <https://www.sciencedirect.com/science/article/abs/pii/S1051137796900182>