Statistics for Business
House Price Prediction

A. Introduction & Background

House price prediction is a significant task in the real estate industry, where accurate predictions can help homeowners, buyers, and sellers to make informed decisions. Linear regression is a widely used statistical technique that helps to estimate the relationship between a dependent variable (in this case, house prices) and one or more independent variables (such as the size of the house, location, number of rooms, etc.). In this context, linear regression models can be used to predict the house prices based on the available data.

The housing market is one of the most important sectors of the economy, with a significant impact on the overall financial well-being of individuals and society as a whole. Accurately predicting the housing prices can help various stakeholders, including homebuyers, sellers, investors, and real estate agents, to make informed decisions. The traditional approach to house price prediction involves analyzing various factors such as the size of the property, location, the number of rooms, and other amenities to estimate the price. However, with the advent of big data and machine learning techniques, it has become possible to develop more accurate and reliable predictive models.

Linear regression is a widely used statistical method for predictive modeling in various fields, including finance, economics, and real estate. In the context of house price prediction, linear regression models are used to estimate the relationship between the dependent variable (i.e., house prices) and one or more independent variables (such as the size of the property, location, the number of rooms, etc.). The model then uses this relationship to predict the house prices based on the available data. However, developing an accurate linear regression model for house price prediction requires careful analysis of the data, feature engineering, model selection, and tuning, among other things.

The real estate markets, like those in Washington DC, present an interesting opportunity for data analysts to analyze and predict where property prices are moving towards. Prediction of property prices is becoming increasingly important and beneficial. Property prices are a good indicator of both the overall market condition and the economic health of a country. Considering the data provided, we are wrangling a large set of property sales records stored in an unknown format and with unknown data quality issues.

B. Dataset

There are 18 feature on dataset House Prediction Price, with 4.600 record as follows:

- Date : from May 2nd, 2014 to July 10th, 2014
- Price : the price of the house for sale
- bedrooms : total number of bedrooms in a house
- bathrooms : total number of bathrooms in a house

- sqft_living : size of the house in square feet
- sqft_lot : size of the land in square feet
- floors : number of floors in a house
- waterfront : a house bordering water (river, lake, etc)
- view : a house with an attractive view
- condition : the condition of the building (scale 1-5)
- sqft_above : size of housing above basement in square feet
- sqft_basement : size of the basement in square feet
- yr_built : year the house was built
- yr_renovated : year the house was last renovated
- street : address of the house
- city : city where the house is located
- statezip : zip code
- country : country

C. Setting Up Problem

1. Project Goal

To develop a linear regression model that accurately predicts house prices based on various features such as location, square footage, number of bedrooms and bathrooms, and other relevant factors.

Objectives:

- Collect and clean a large dataset of house prices and related features.
- Explore the data to identify any correlations and patterns.
- Choose appropriate features and build a linear regression model using the dataset.
- Evaluate the model's performance using appropriate metrics such as mean squared error (MSE) and coefficient of determination (R-squared).
- Fine-tune the model and re-evaluate its performance until a satisfactory level of accuracy is achieved.
- Use the model to predict the prices of new houses and validate its accuracy using real-world data.
- Document the project and share the findings with stakeholders, including recommendations for improving the model's accuracy and potential applications for the insights gained.

Deliverables:

- A clean dataset of house prices and relevant features.
- A Jupyter notebook or Python script that contains the code used to build and evaluate the linear regression model.
- A report that summarizes the findings of the project, including the model's accuracy, insights gained, and potential applications.
- A presentation that highlights the key findings and recommendations for stakeholders.

What we want to know:

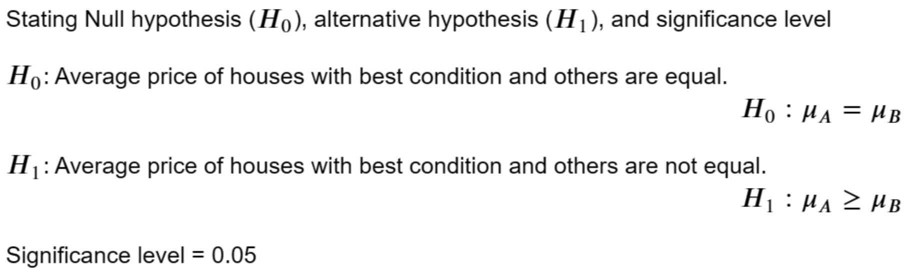
Assume we want to make report about the data to inform these:

- Average, maximum, minimum price of house
- Which transaction that have the highest price purchased?
- Distribution of price, sqft_living, sqft_lot, sqft_above, and sqft_basement
- Association between price and condition
- Association between price and yr built/yr renovated
- Association between price and sqft_living, sqft_lot, sqft_above, and sqft_basement
- Association between price and bedrooms, bathrooms, floors, waterfont, and view
- What are the most frequently purchased houses?
- Which city that give highest and lowest price?

We can answer these question by doing exploration in the data

2. Define Statistical Test

a. Is average of price in highest city price house are same with other?



Stating Null hypothesis (H_0), alternative hypothesis (H_1), and significance level

H_0: Average price of houses with best condition and others are equal.

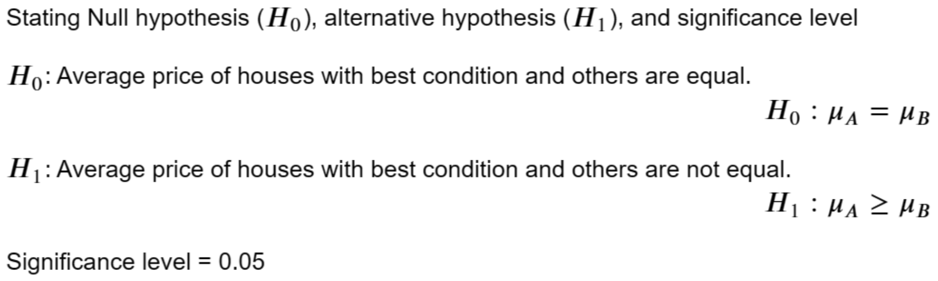
$$H_0 : \mu_A = \mu_B$$

H_1: Average price of houses with best condition and others are not equal.

$$H_1 : \mu_A \geq \mu_B$$

Significance level = 0.05

b. Is average of price in best condition (5) price house are same with other?



Stating Null hypothesis (H_0), alternative hypothesis (H_1), and significance level

H_0: Average price of houses with best condition and others are equal.

$$H_0 : \mu_A = \mu_B$$

H_1: Average price of houses with best condition and others are not equal.

$$H_1 : \mu_A \geq \mu_B$$

Significance level = 0.05

3. Define Design of Regression Model

For design regression model, there are several things need to know:

a. Outcome variable :
Because the target of this regression is to find out the prediction of price, so main target of the variable is price as outcome

b. Predictor variable :
To get know what variable that would be assign as predictors, we should done correlation analysis first to get which variable have most correlation with the outcome

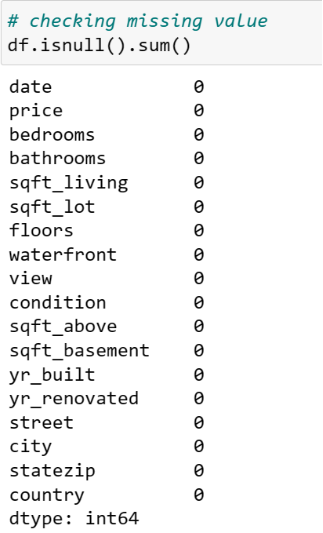
c. Design of Regression Model :

After determine the outcome and predictors variable, we can design regression model. First, do the basics OLS to see the plot regression by single predictor, and do the analysis of cofficient, standard error, prediction interval, and R-Square. Then, add second predictor to increase performance of the model and do analysis as before. Then, try to add interaction between those predictor. We can do analysis of the residual error and do some analysis about homoscedasticity. If there any heteroscedasticity, we can do some transformation (such as log transform, one over square, and etc) to make the variance more uniform. And after that, we can do some transformation using z transformation, scaling data, etc to make better interpretation about the outcome regression model equation.

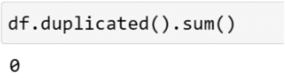
D. Statistical Test

1. Validity Data and Reability Data

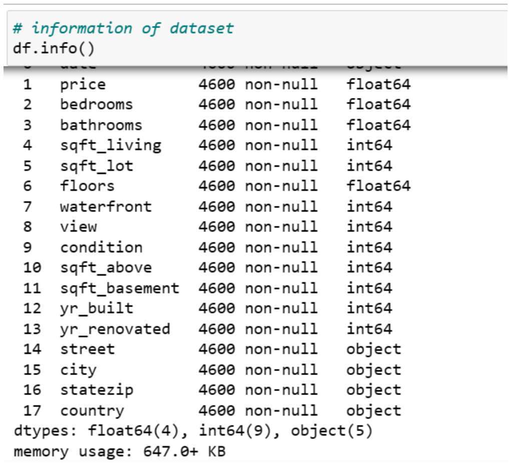
For unsure the validity of data, we do some data checking to make sure there are not null data available and there are no data duplicated. The results as figure below:



```
# checking missing value
df.isnull().sum()

date             0
price            0
bedrooms         0
bathrooms        0
sqft_living      0
sqft_lot         0
floors           0
waterfront       0
view             0
condition        0
sqft_above       0
sqft_basement    0
yr_built         0
yr_renovated     0
street           0
city             0
statezip         0
country          0
dtype: int64
```



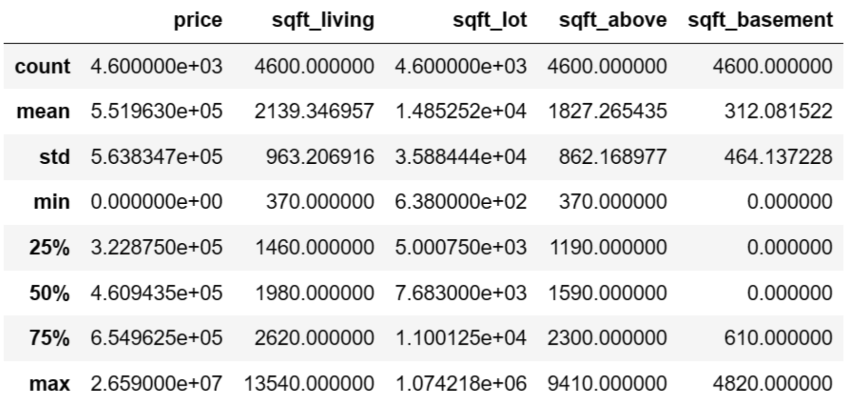
```
df.duplicated().sum()

0
```



```
# information of dataset
df.info()
 1   price          4600 non-null   float64
 2   bedrooms       4600 non-null   float64
 3   bathrooms      4600 non-null   float64
 4   sqft_living    4600 non-null   int64
 5   sqft_lot       4600 non-null   int64
 6   floors         4600 non-null   float64
 7   waterfront     4600 non-null   int64
 8   view           4600 non-null   int64
 9   condition      4600 non-null   int64
 10  sqft_above     4600 non-null   int64
 11  sqft_basement  4600 non-null   int64
 12  yr_built       4600 non-null   int64
 13  yr_renovated   4600 non-null   int64
 14  street         4600 non-null   object
 15  city           4600 non-null   object
 16  statezip       4600 non-null   object
 17  country        4600 non-null   object
dtypes: float64(4), int64(9), object(5)
memory usage: 647.0+ KB
```

2. Statistical Descriptive Test

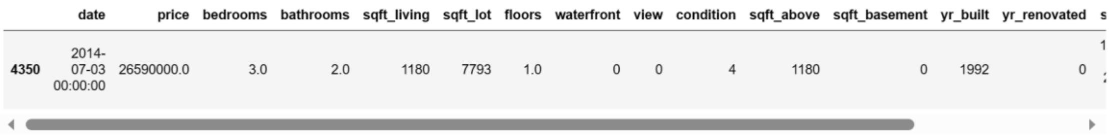
For numerical variables, we can find statistics description such as mean, standar deviation, quartile, etc of each variables



	price	sqft_living	sqft_lot	sqft_above	sqft_basement
count	4.600000e+03	4600.000000	4.600000e+03	4600.000000	4600.000000
mean	5.519630e+05	2139.346957	1.485252e+04	1827.265435	312.081522
std	5.638347e+05	963.206916	3.588444e+04	862.168977	464.137228
min	0.000000e+00	370.000000	6.380000e+02	370.000000	0.000000
25%	3.228750e+05	1460.000000	5.000750e+03	1190.000000	0.000000
50%	4.609435e+05	1980.000000	7.683000e+03	1590.000000	0.000000
75%	6.549625e+05	2620.000000	1.100125e+04	2300.000000	610.000000
max	2.659000e+07	13540.000000	1.074218e+06	9410.000000	4820.000000

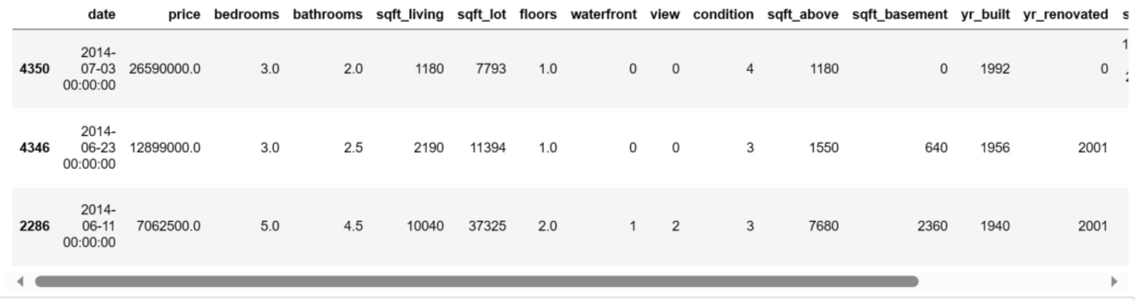
We want to know:

- Average, maximum, minimum price of house
  The average total money per transaction in 2019 is `$551.963` with the maximum total money `$26.590.000` and the minimum total money `$0`
- Which transaction that have the highest price purchased?
  Let's see which transaction that has the highest price product purchase



	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basement	yr_built	yr_renovated	s
4350	2014-07-03 00:00:00	26590000.0	3.0	2.0	1180	7793	1.0	0	0	4	1180	0	1992	0	1

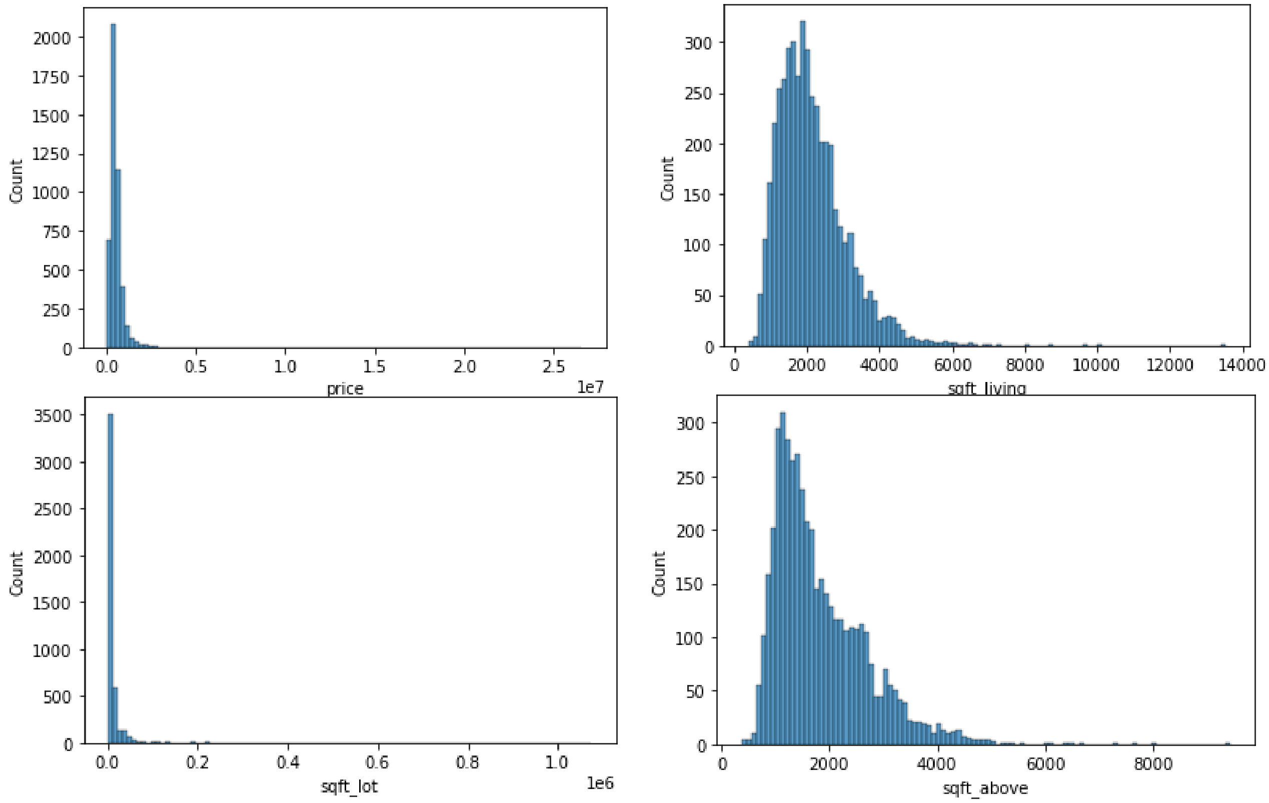
There is 1 house for the maximum price. They are purchase from Kent



	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basement	yr_built	yr_renovated	s
4350	2014-07-03 00:00:00	26590000.0	3.0	2.0	1180	7793	1.0	0	0	4	1180	0	1992	0	1
4346	2014-06-23 00:00:00	12899000.0	3.0	2.5	2190	11394	1.0	0	0	3	1550	640	1956	2001	
2286	2014-06-11 00:00:00	7062500.0	5.0	4.5	10040	37325	2.0	1	2	3	7680	2360	1940	2001	

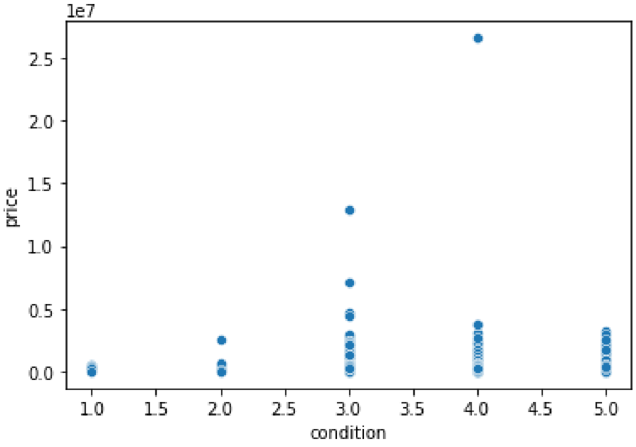
There are 3 price of house that above usual, so we can infer that as outliers.

- Distribution of price, sqft_living, sqft_lot, sqft_above, and sqft_basement
  Next, we perform histogram visualization for each numeric variable

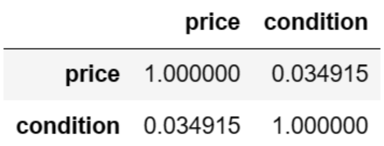


The price variable tends to show skewness to the right, due to the large number of houses that price less than 5.000.000. The second most highest price purchased, is 1 items which almost reached 12.899.000, then the third, is 7.062.500. The histogram for this sqft living shows that many house skew on right side, there are several outliers. The histogram for this sqft lot shows that many house skew on right side, there are several outliers. The histogram for this sqft above shows that many house skew on right side, there are several outliers. The histogram for this sqft basement shows that many house skew on right side, there are several outliers

- Association between price and condition

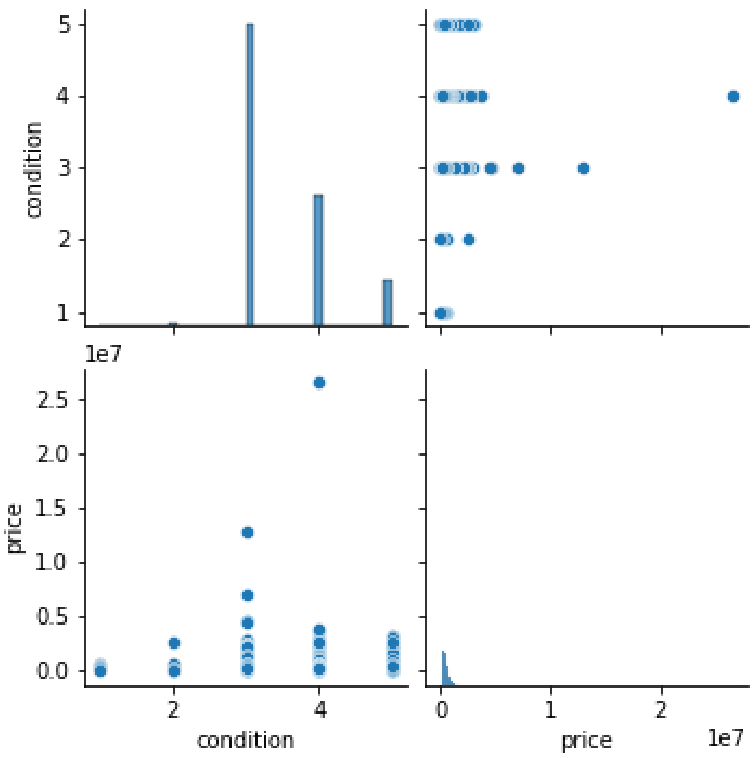


There is almost no visible relationship between price and condition. what about the correlation value between the two?

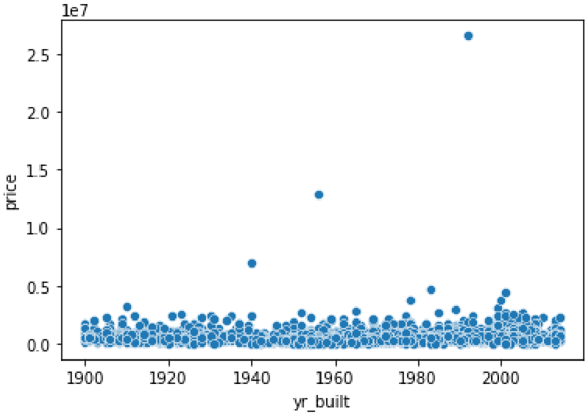


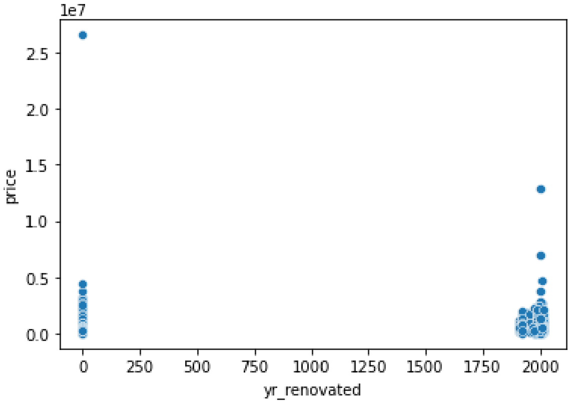
	price	condition
price	1.000000	0.034915
condition	0.034915	1.000000

The correlation value is 0.035, there is a positive relationship between price and condition, meaning that the good condition tend to sold in highest price and vice versa. The value of 0.035 means that the two variables have a weak relationship, meaning that there is not too much visible linear pattern in the two variables.

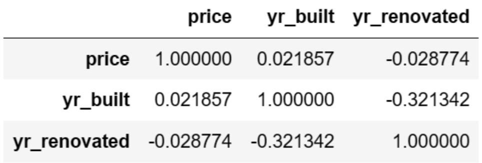


- Association between price and yr built/yr renovated



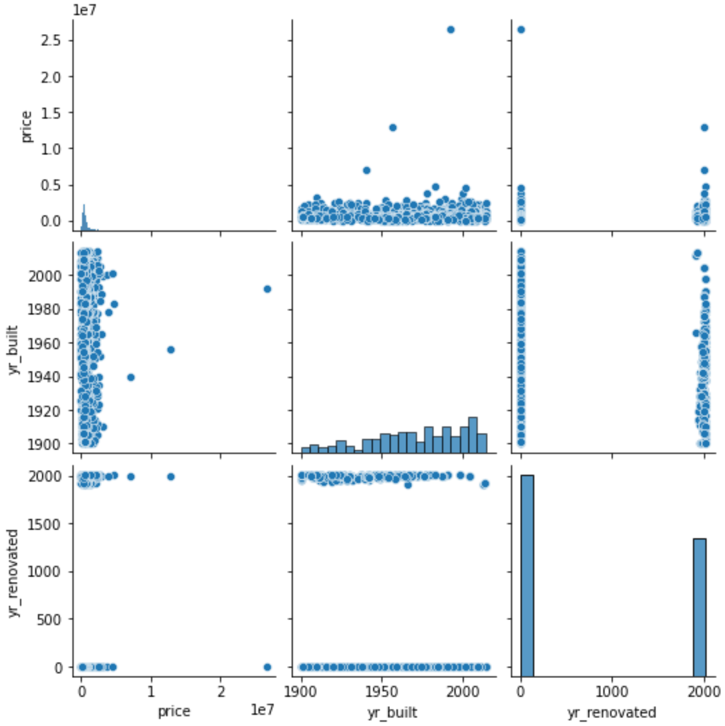


There is almost no visible relationship between price and year built / year renovated. What about the correlation value between the three?

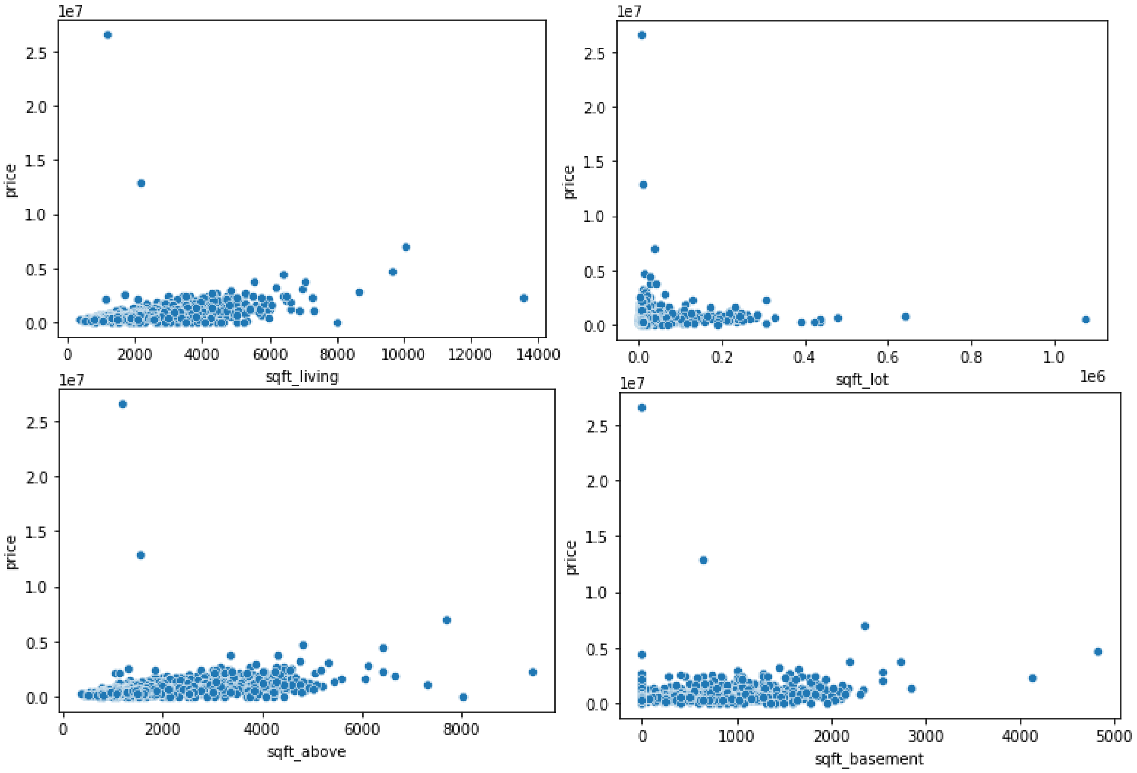


	price	yr_built	yr_renovated
price	1.000000	0.021857	-0.028774
yr_built	0.021857	1.000000	-0.321342
yr_renovated	-0.028774	-0.321342	1.000000

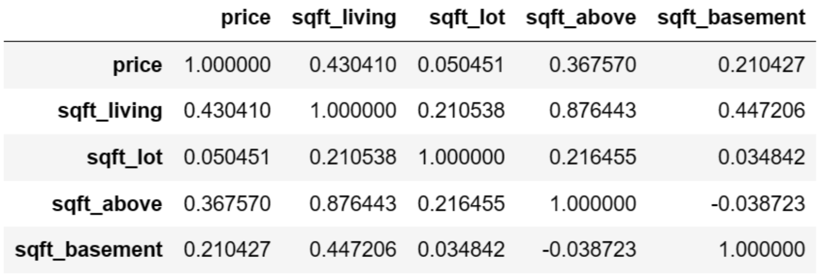
The correlation value is 0.022, there is a positive relationship between price and year built, meaning that the newer year built tend to sold in highest price and vice versa. The correlation value is -0.029, there is a negativ relationship between price and year renovated, meaning that the newer year renovated tend to sold in lowest price and vice versa. The correlation value is -0.321, there is a negative relationship between year built and year renovated, meaning that the newer year built tend to less have renovated and vice versa. The value of 0.022, -0.029, -0.321 means that the three variables have a weak relationship, meaning that there is not too much visible linear pattern in the three variables.



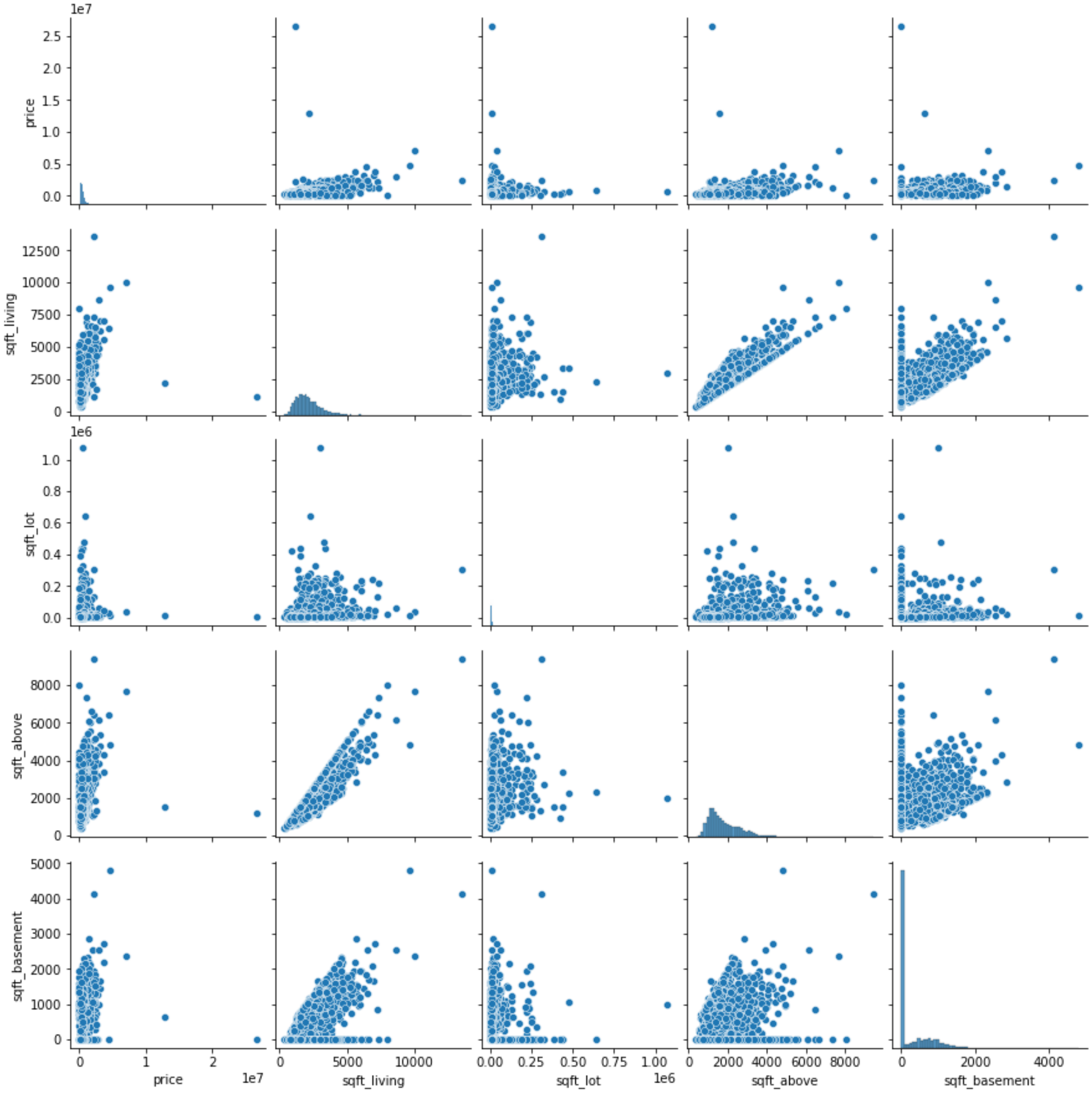
- Association between price and sqft_living, sqft_lot, sqft_above, and sqft_basement



There is almost no visible relationship between price and all sqft feature. what about the correlation value between those?

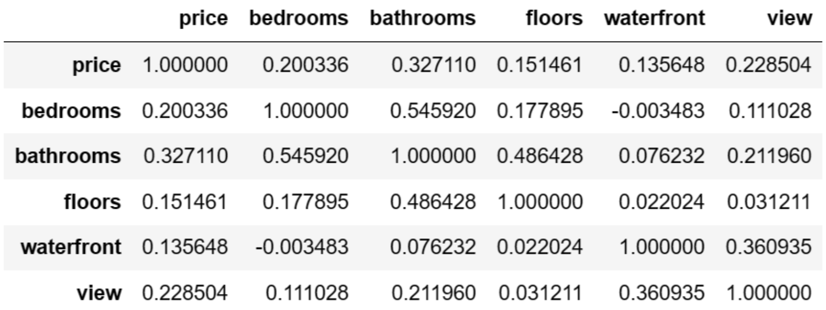


	price	sqft_living	sqft_lot	sqft_above	sqft_basement
price	1.000000	0.430410	0.050451	0.367570	0.210427
sqft_living	0.430410	1.000000	0.210538	0.876443	0.447206
sqft_lot	0.050451	0.210538	1.000000	0.216455	0.034842
sqft_above	0.367570	0.876443	0.216455	1.000000	-0.038723
sqft_basement	0.210427	0.447206	0.034842	-0.038723	1.000000

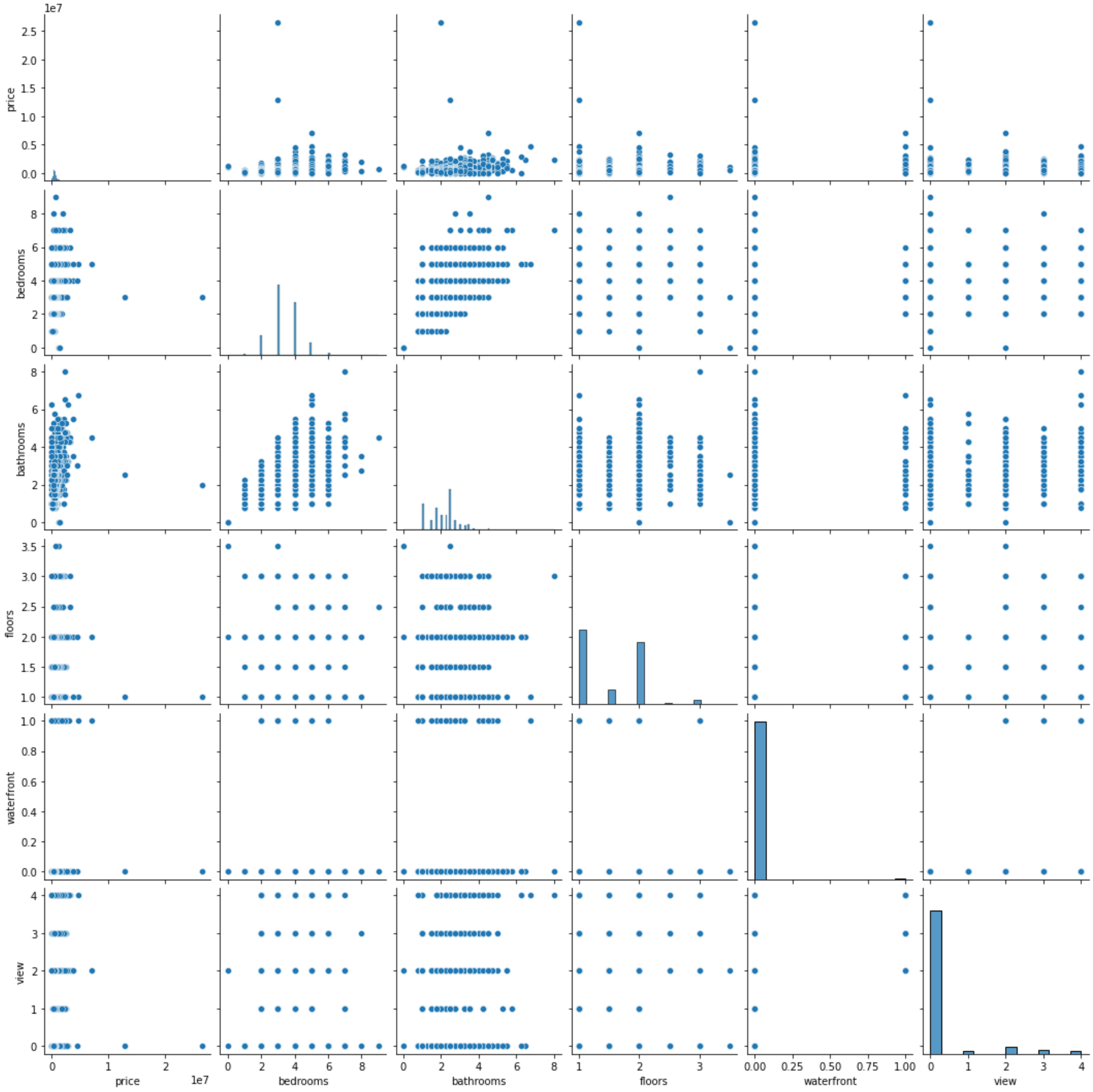


for categorical type data, we understand better if it is in the form of visualization. We can use barplot.

- Association between price and bedrooms, bathrooms, floors, waterfont, and view

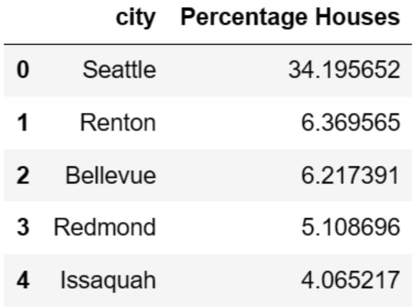


	price	bedrooms	bathrooms	floors	waterfront	view
price	1.000000	0.200336	0.327110	0.151461	0.135648	0.228504
bedrooms	0.200336	1.000000	0.545920	0.177895	-0.003483	0.111028
bathrooms	0.327110	0.545920	1.000000	0.486428	0.076232	0.211960
floors	0.151461	0.177895	0.486428	1.000000	0.022024	0.031211
waterfront	0.135648	-0.003483	0.076232	0.022024	1.000000	0.360935
view	0.228504	0.111028	0.211960	0.031211	0.360935	1.000000



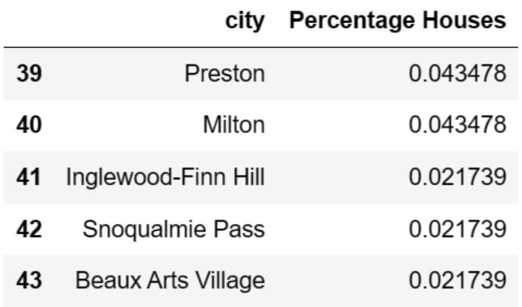
- Which city that give highest and lowest price?

Highest Houses Price City

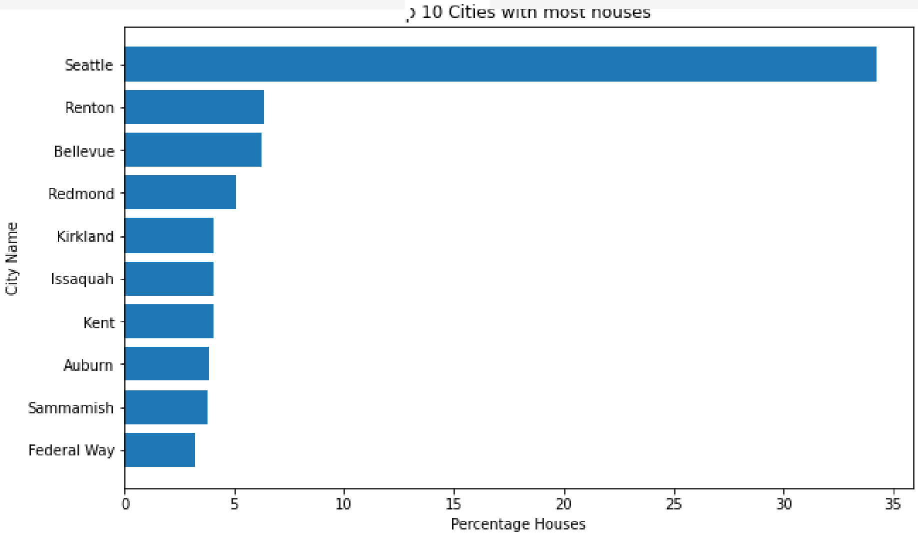


	city	Percentage Houses
0	Seattle	34.195652
1	Renton	6.369565
2	Bellevue	6.217391
3	Redmond	5.108696
4	Issaquah	4.065217

Lowest Houses Price City

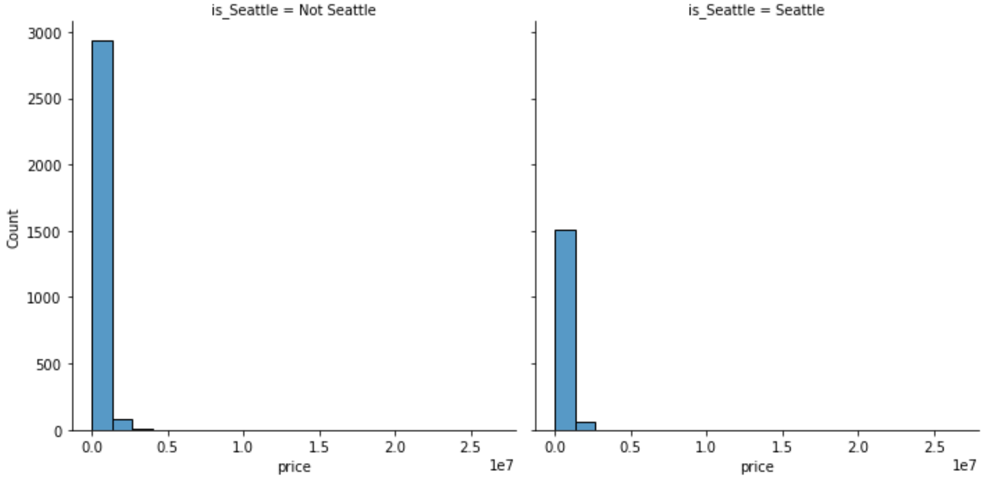


	city	Percentage Houses
39	Preston	0.043478
40	Milton	0.043478
41	Inglewood-Finn Hill	0.021739
42	Snoqualmie Pass	0.021739
43	Beaux Arts Village	0.021739

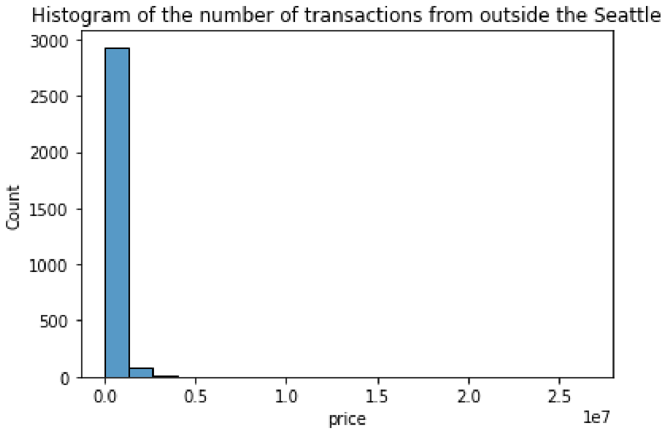


Top 10 city with most pricy houses

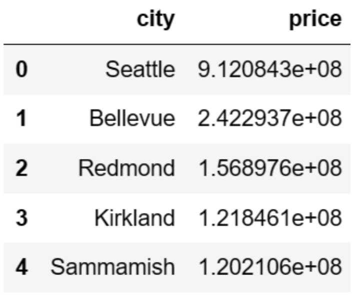
We can see a comparison of histograms from the Seattle and other cities from the figure below



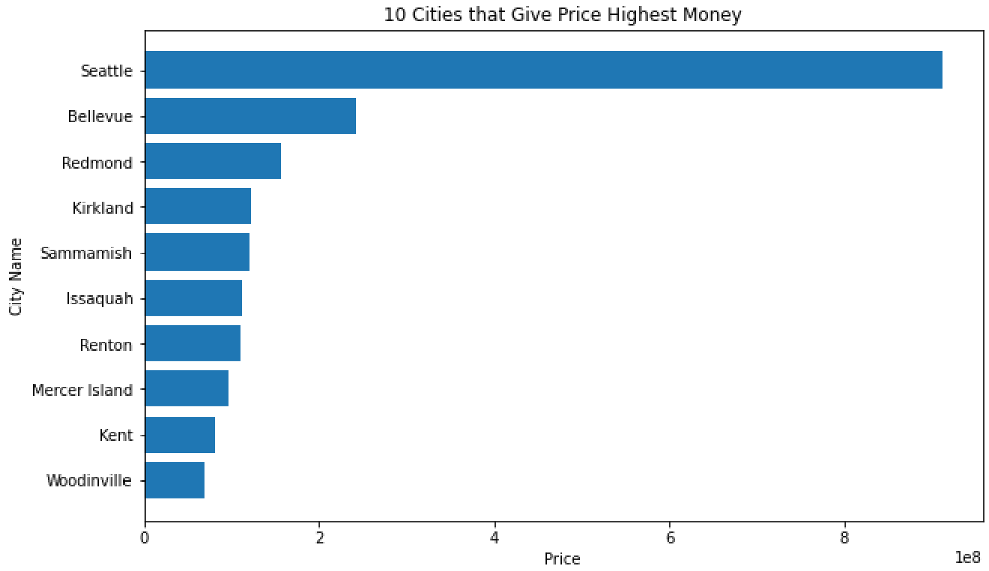
For clearer plot results, the histogram for outside Seattle can be seen as follows

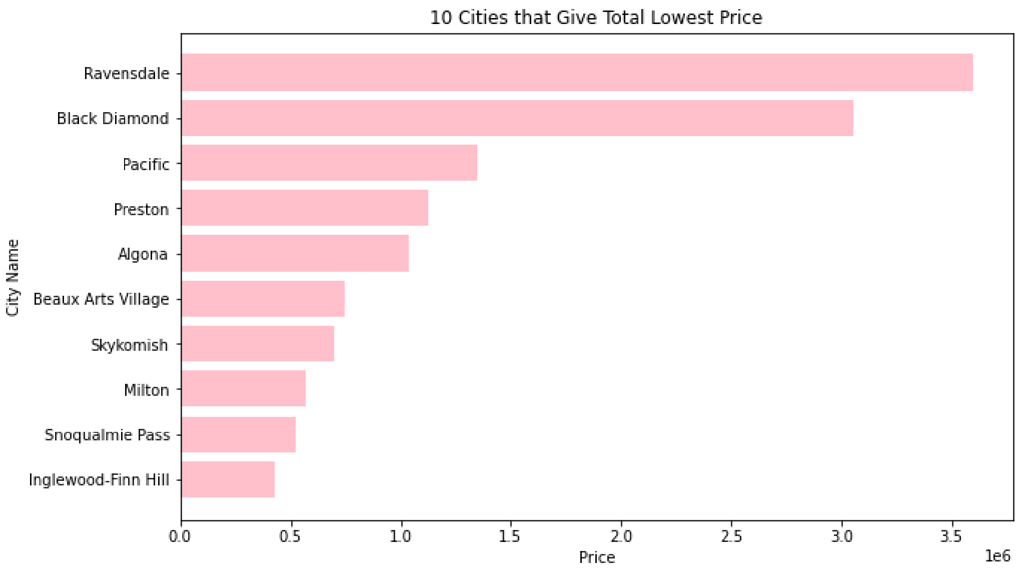


Next, we sum the total money from each city to get the 10 cities with the most price



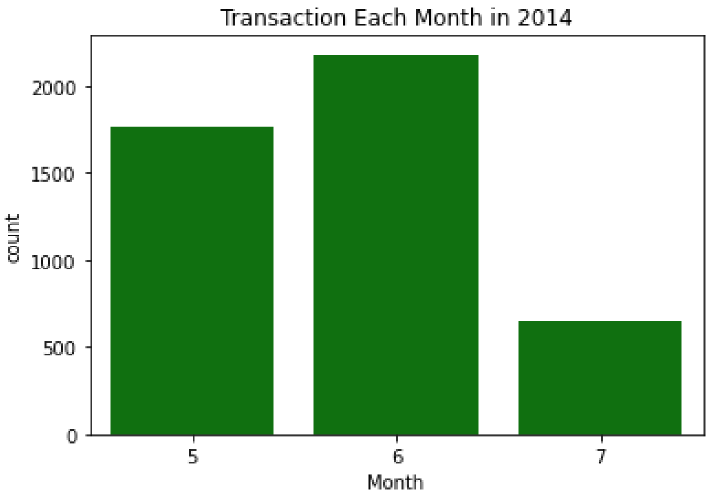
	city	price
0	Seattle	9.120843e+08
1	Bellevue	2.422937e+08
2	Redmond	1.568976e+08
3	Kirkland	1.218461e+08
4	Sammamish	1.202106e+08





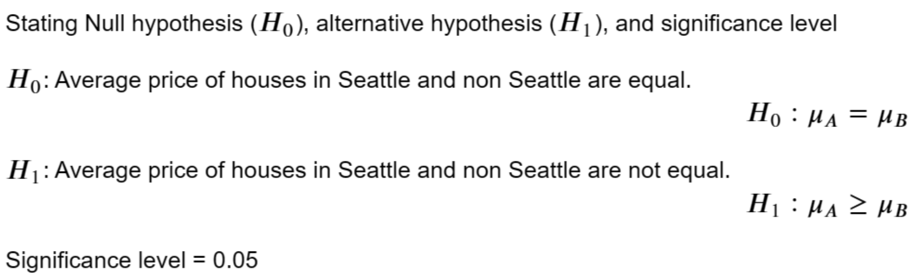
According to the chart "Top 10 Cities That Brings The Highest Price" in 2014 we can see that the Seattle is where we generate the majority of prices. We can take advantage of this by expanding there and offering promotions that are only available there. Additional study should be conducted to determine why sales are low in cities like Ravensdale and Inglendwood-Finn Hill, once we have that information, relevant solutions can be developed.

- How was the price trend over the months?



Most transactions occurred in June. The fewest transactions occurred in July

3. **Performing Statistical Inference**
   a. **Is average of price in highest city price house are same with other?**



Stating Null hypothesis (H_0), alternative hypothesis (H_1), and significance level

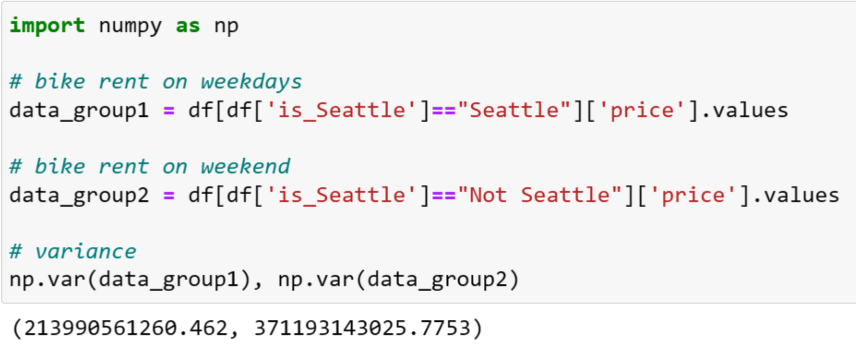
H_0: Average price of houses in Seattle and non Seattle are equal.

$$H_0 : \mu_A = \mu_B$$

H_1: Average price of houses in Seattle and non Seattle are not equal.

$$H_1 : \mu_A \geq \mu_B$$

Significance level = 0.05

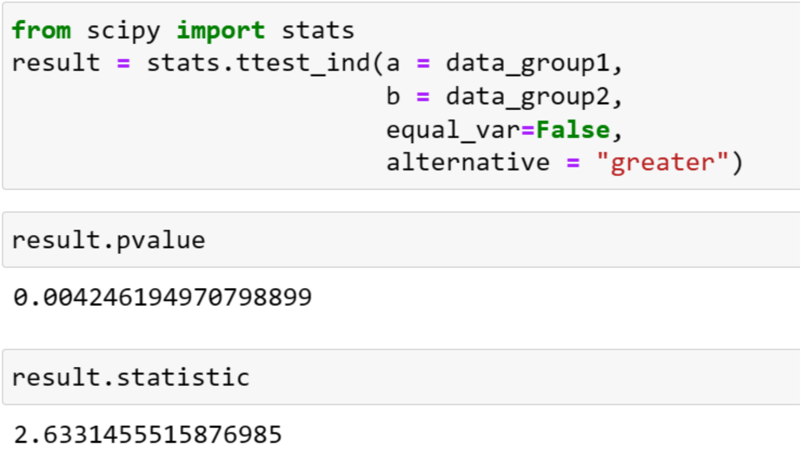


```
import numpy as np

# bike rent on weekdays
data_group1 = df[df['is_Seattle']=="Seattle"]['price'].values

# bike rent on weekend
data_group2 = df[df['is_Seattle']=="Not Seattle"]['price'].values

# variance
np.var(data_group1), np.var(data_group2)
```

```
(213990561260.462, 371193143025.7753)
```

Based on the result, we can see that the variance is not equal for both the samples. Afterward, we can calculate statistics test and p-value using spicy library. To calculate two sample proportion z test, we can use stats.ttest_ind

- Import library
  from scipy import stats
- Use function `stats.ttest_ind(a=...., b=...., equal_var=True/False)`
  `a`: First data group
  `b`: Second data group
  `equal_var = True` : The standard independent two sample t-test will be conducted by taking into consideration the equal population variances.
  `equal_var = False` : The Welch's t-test will be conducted by not taking into consideration the equal population variances.
- The function will be able to return 2 output, namely statistic test and p_value.



```
from scipy import stats
result = stats.ttest_ind(a = data_group1,
                         b = data_group2,
                         equal_var=False,
                         alternative = "greater")
```

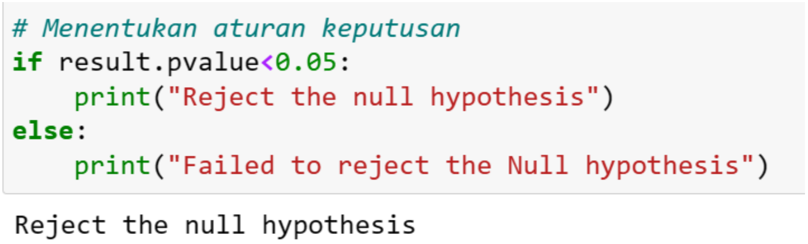
```
result.pvalue
```

```
0.004246194970798899
```

```
result.statistic
```

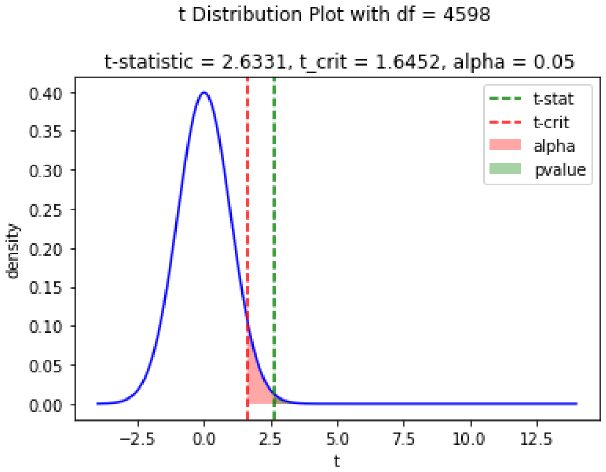
```
2.6331455515876985
```

The next step is to make decision rules by which we can know our hypothesis will be either rejected or fail to reject.



```
# Menentukan aturan keputusan
if result.pvalue<0.05:
    print("Reject the null hypothesis")
else:
    print("Failed to reject the Null hypothesis")
```

```
Reject the null hypothesis
```



t Distribution Plot with df = 4598

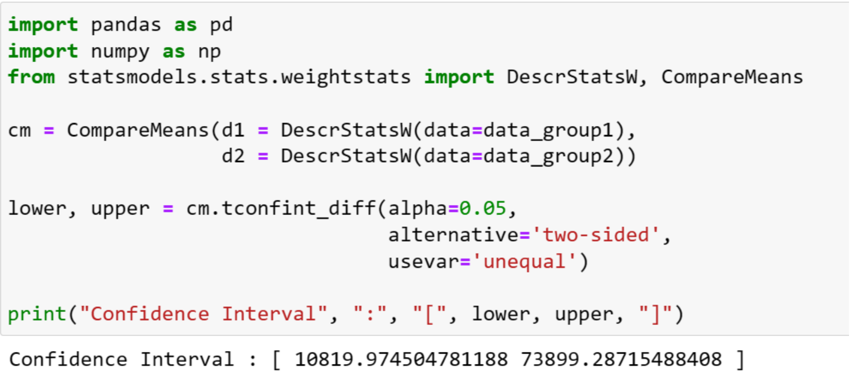
t-statistic = 2.6331, t_crit = 1.6452, alpha = 0.05

After that, we will compute confidence level for the difference in means

To calculate confidence interval for the difference in means, we can use CompareMeans

- import library
  from statsmodels.stats.weightstats import DescrStatsW,[CompareMeans]
- Use function `CompareMeans.tconfint_diff(alpha=0.05, alternative='two-sided', usevar='pooled')`

  `alpha` = significance level for the confidence interval
  `alternative` = It depends on the alternative hypothesis for the test
  - if H1 is not equal to value, we use `two-sided`
  - if H1 is larger than value, we use `larger`
  - if H1 is smaller than value, we use `smaller`
  `usevar` = 'pooled' or 'unequal'
  'pooled` indicates that the standard deviation of the samples is assumed to be the same.
  `unequal` shows that Welch ttest with Satterthwait degrees of freedom is used.
- The function will return lower and upper limits of the confidence interval

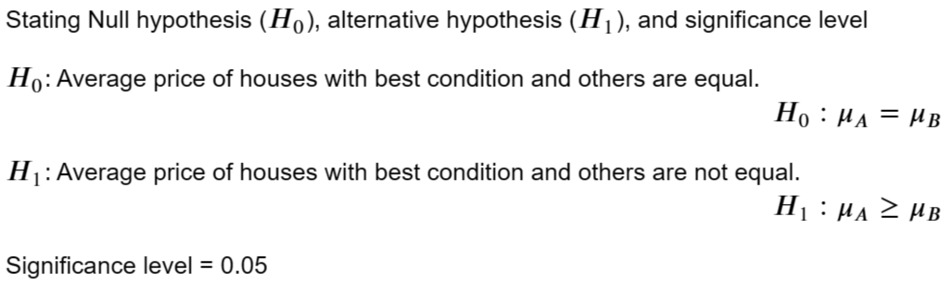


```
import pandas as pd
import numpy as np
from statsmodels.stats.weightstats import DescrStatsW, CompareMeans

cm = CompareMeans(d1 = DescrStatsW(data=data_group1),
                  d2 = DescrStatsW(data=data_group2))

lower, upper = cm.tconfint_diff(alpha=0.05,
                                alternative='two-sided',
                                usevar='unequal')

print("Confidence Interval", ":", "[", lower, upper, "]")
```

```
Confidence Interval : [ 10819.974504781188 73899.28715488408 ]
```

Based on the result, we can 95% confident that the average difference the price in Seattle and non Seattle lies between 10.819,97 and 73.899,28. As the p value < alpha(0.05) , we reject H_0. Therefore, we can say that average price in both Seattle and non-Seattle is not equal. As confidence interval lies between 10.819,97 and 73.899,28, there is significant different between average price in Seattle then other days.

b. Is average of price in best condition (5) price house are same with other?



Stating Null hypothesis (H_0), alternative hypothesis (H_1), and significance level

H_0: Average price of houses with best condition and others are equal.

$$H_0 : \mu_A = \mu_B$$

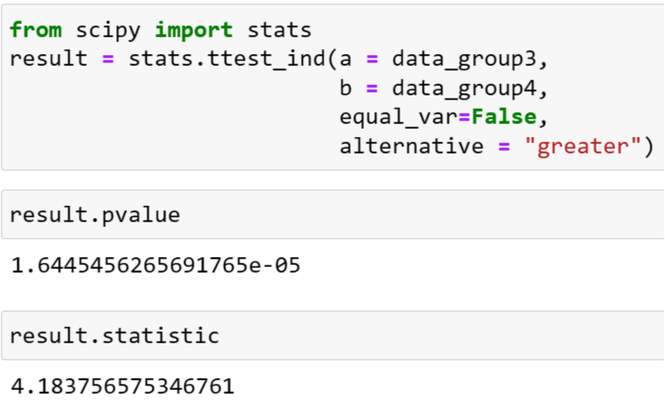
H_1: Average price of houses with best condition and others are not equal.

$$H_1 : \mu_A \geq \mu_B$$

Significance level = 0.05

Based on the result, we can see that the variance is not equal for both the samples. Afterward, we can calculate statistics test and p-value using spicy library. To calculate two sample proportion z test, we can use stats.ttest_ind

- Import library
  from scipy import stats
- Use function `stats.ttest_ind(a=...., b=...., equal_var=True/False)`
  `a`: First data group
  `b`: Second data group
  `equal_var = True` : The standard independent two sample t-test will be conducted by taking into consideration the equal population variances.
  `equal_var = False` : The Welch's t-test will be conducted by not taking into consideration the equal population variances.
- The function will be able to return 2 output, namely statistic test and p_value.



```
from scipy import stats
result = stats.ttest_ind(a = data_group3,
                         b = data_group4,
                         equal_var=False,
                         alternative = "greater")
```

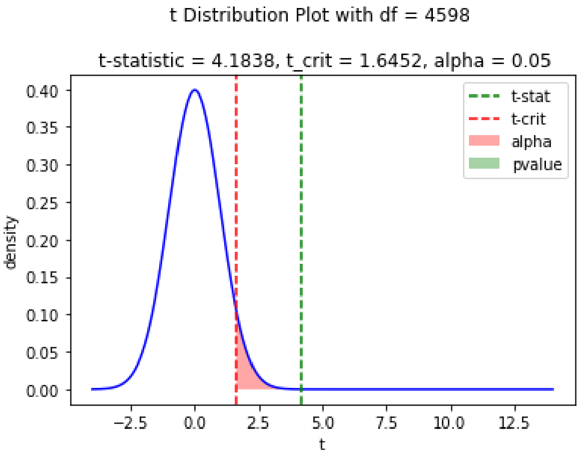
```
result.pvalue
```

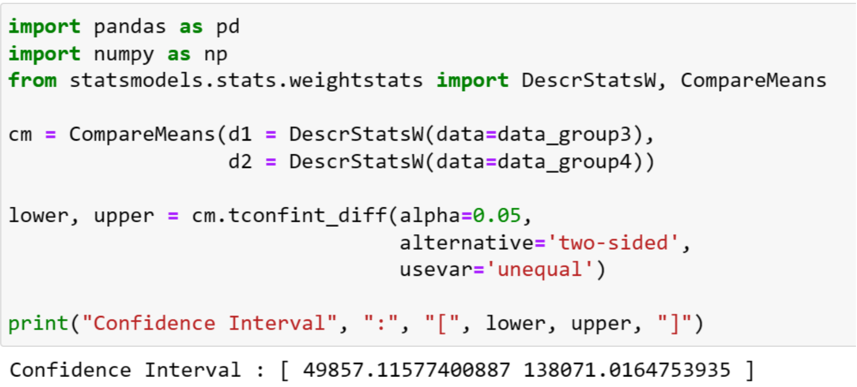
```
1.6445456265691765e-05
```

```
result.statistic
```

```
4.183756575346761
```

The next step is to make decision rules by which we can know our hypothesis will be either rejected or fail to reject.





```
import pandas as pd
import numpy as np
from statsmodels.stats.weightstats import DescrStatsW, CompareMeans

cm = CompareMeans(d1 = DescrStatsW(data=data_group3),
                  d2 = DescrStatsW(data=data_group4))

lower, upper = cm.tconfint_diff(alpha=0.05,
                                alternative='two-sided',
                                usevar='unequal')

print("Confidence Interval", ":", "[", lower, upper, "]")
```

```
Confidence Interval : [ 49857.11577400887 138071.0164753935 ]
```

Based on the result, we can 95% confident that the average difference the price in good condition and other condition houses lies between 49.857,11 and 138.071,01. As the p value < alpha(0.05) , we reject H0. Therefore, we can say that average price in both good condition and other is not equal. As confidence interval lies between 49.857,11 and 138.071,01, there is significant different between average price in both good condition and other

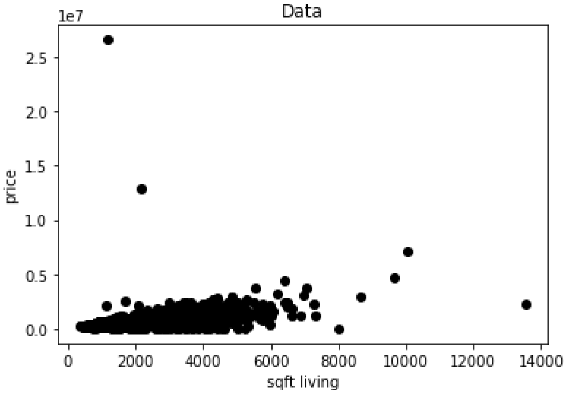
E. Regression Model

Based on statistical test, we conclude there are some feature that have most biggest correlation to price, such as:

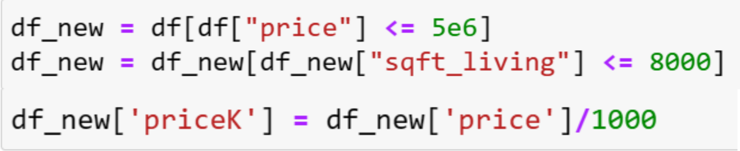
- sqft_living
- sqft_above

1. Single Predictor Regression Model

a. Using sqft living as predictors

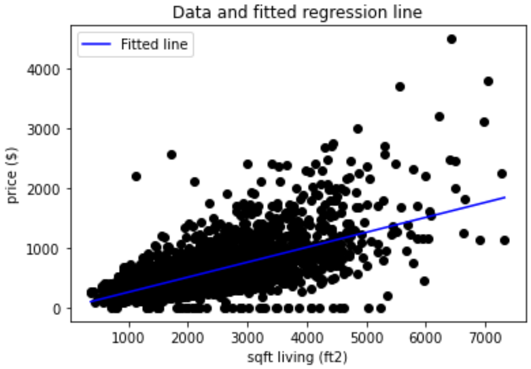


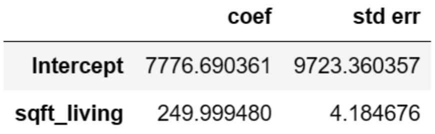
We can directly calculation the mean with usual calculation. The mean is 551962.98. To make better regression model, we drop some data than become outliers because it can distract the regression model and reduce the performance of model and scaling price to thousand.



```
df_new = df[df["price"] <= 5e6]
df_new = df_new[df_new["sqft_living"] <= 8000]

df_new['priceK'] = df_new['price']/1000
```

Under regression model you can also get this value by asssigning the constant term (no predictor) as predictor.



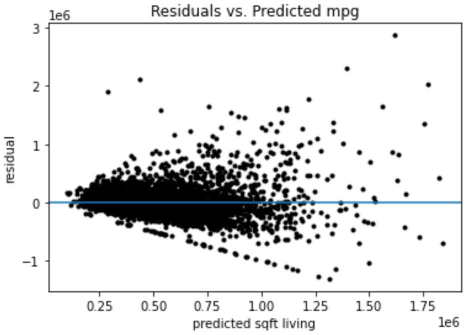


	coef	std err
Intercept	7776.690361	9723.360357
sqft_living	249.999480	4.184676

$$\text{price} = 7776.69 + 249.99\text{sqft\_living}$$

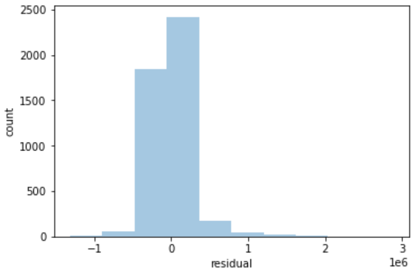
The average of the price from 0 sqft living is `$7776.69`. The difference between average price of it sqft living house is `$249.99`, with the house that has bigger sqft living have the higher price. And r-squared that given by experiment is 0.4373

- Residual Plot and Explained Variance



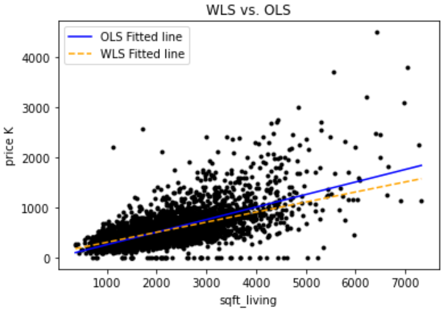
The residuals produce a noticeable pattern make the lack of fit more apperent eventhough the fitted line explain more than two-thirds (44%) of variation in sqft_living. The residual we get, have unconstant pattern around the zero. Lets try use weighted regression to address this issue, and see what happening. Give the lower weight in the data that have high variance, so set the weight inverse proportional of the predictor as we see that, the variance increase as the sqrt living increase.

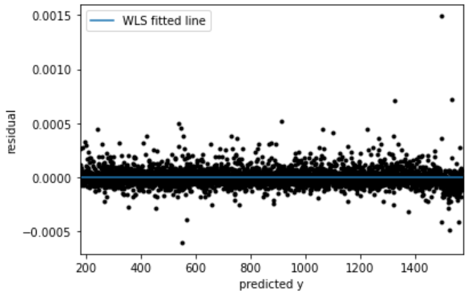
- Normality of error assumption



Using some transformations, here the outcome R-Squared that generated by experiment,

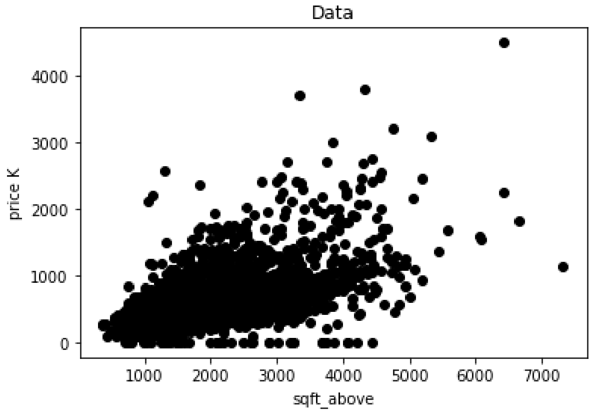
a. Using One Over Square on sqft_living Variable (R-Squared = 0.1392)
b. Using Log Transform on sqft_living Variable (R-Squared = 0.3575)
c. Using Reciprocal Transform on sqft_living Variable (R-Squared = 0.2471)
d. Using Weighted Least Square on sqft_living Variable (R-Squared = 0.3589)



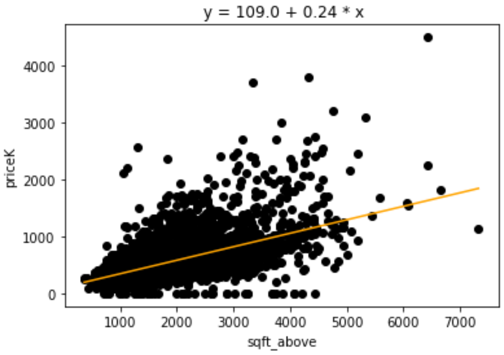


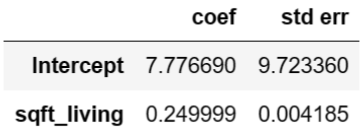
By those transformations, we see that all transformations reduce value of R-Squared, so we using OLS without transformations.

b. Using sqft above as predictors



Under regression model you can also get this value by asssigning the constant term (no predictor) as predictor.



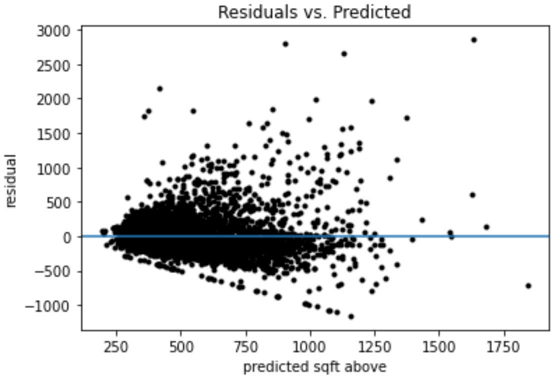


	coef	std err
Intercept	7.776690	9.723360
sqft_living	0.249999	0.004185

$$y = 7.77 + 0.25x$$

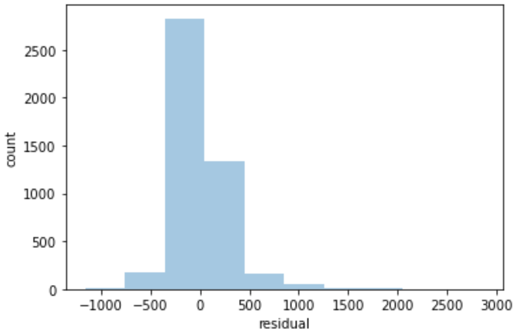
The average of the price from 0 sqft above is `$7.77`. The difference between average price of it sqft above is `$0.25`, with the house that has bigger sqft above have the higher price. And r-squared that given by experiment is 0.3243

- Residual Plot and Explained Variance



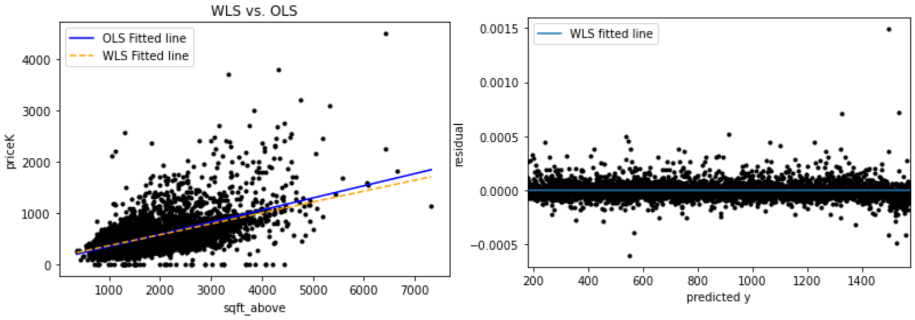
The residuals produce a noticeable pattern make the lack of fit more apperent eventhough the fitted line explain more than two-thirds (34%) of variation in sqft_above. The residual we get, have unconstant pattern around the zero. Lets try use weighted regression to address this issue, and see what happening. Give the lower weight in the data that have high variance, so set the weight inverse proportional of the predictor as we see that, the variance increase as the sqrt above increase.

- Normality of error assumption



Using some transformations, here the outcome R-Squared that generated by experiment,

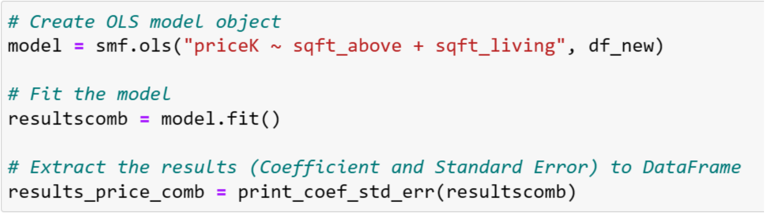
e. Using One Over Square on sqft_living Variable (R-Squared = 0.1413)
f. Using Log Transform on sqft_living Variable (R-Squared = 0.2798)
g. Using Reciprocal Transform on sqft_living Variable (R-Squared = 0.2145)
h. Using Weighted Least Square on sqft_living Variable (R-Squared = 0.2042)



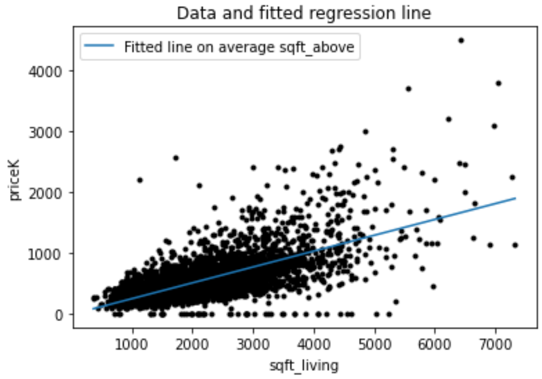
By those transformations, we see that all transformations reduce value of R-Squared, so we using OLS without transformations.

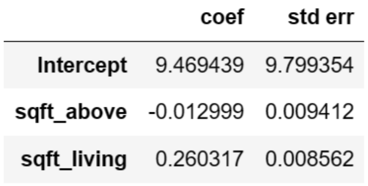
2. Fit Linear Regression - Include Both Variables

Now, We want to predict price from both sqft above and sqft living. Use `+` to add another predictors in the model



```
# Create OLS model object
model = smf.ols("priceK ~ sqft_above + sqft_living", df_new)

# Fit the model
resultscomb = model.fit()

# Extract the results (Coefficient and Standard Error) to DataFrame
results_price_comb = print_coef_std_err(resultscomb)
```





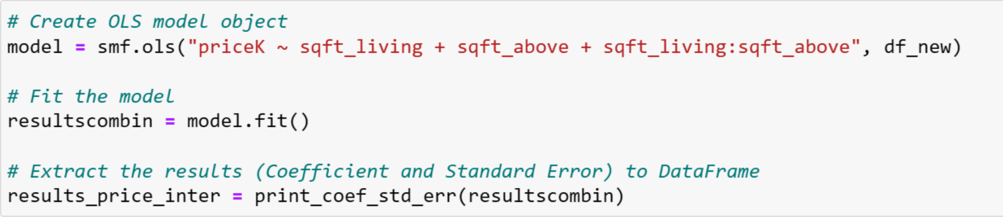
	coef	std err
Intercept	9.469439	9.799354
sqft_above	-0.012999	0.009412
sqft_living	0.260317	0.008562

$$\text{price} = 9.47 + -0.01\text{sqft\_above} + 0.26\text{sqft\_living}$$

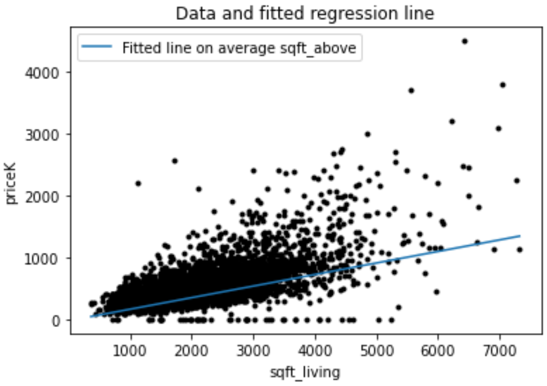
The average of the price from 0 sqft above and 0 sqft living is `$9.47`. The difference between average price of it sqft above is `$-0.013` and the difference between average price of it sqft living is `$-0.26`, with the house that has bigger sqft above have the higher price. And r-squared that given by experiment is 0.4376

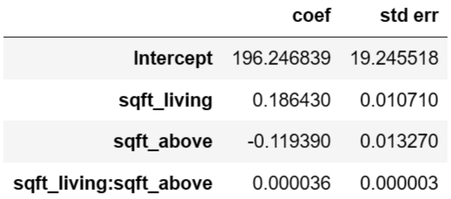
3. **Fit Linear Regression - Include an Interaction**

   In this model, we calculate the interaction between sqrt living and sqrt above to become additional predictor as code below:



```
# Create OLS model object
model = smf.ols("priceK ~ sqft_living + sqft_above + sqft_living:sqft_above", df_new)

# Fit the model
resultscombin = model.fit()

# Extract the results (Coefficient and Standard Error) to DataFrame
results_price_inter = print_coef_std_err(resultscombin)
```





	coef	std err
Intercept	196.246839	19.245518
sqft_living	0.186430	0.010710
sqft_above	-0.119390	0.013270
sqft_living:sqft_above	0.000036	0.000003

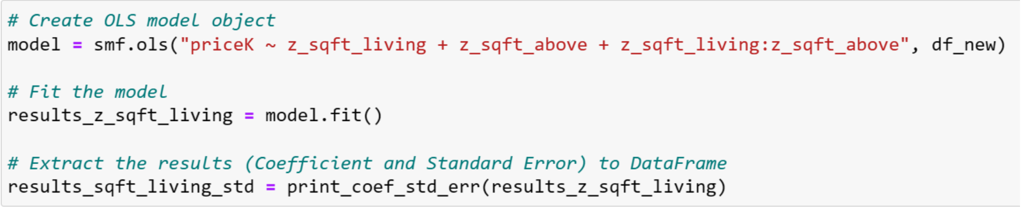
$$\text{price} = 196.25 + 0.186\text{sqft\_living} - 0.119\text{sqft\_above} + 0.000036\text{sqft\_living x sqft\_above}$$

The intercept represents the average price for houses which has 0 sqft above and 0 sqft living is—not a meaningful scenario, we can discuss later to centering the predictor to interpret this better The coefficient of sqft living, 0.186 the difference between the predicted price for houses which is sqft above is 0, and this coefficient is not easily interpretable. The coefficient of sqft_living, 1 The comparison of average prices across houses which house has sqft above 0, but differ by 1 point in sqft above.. The coefficient on the interaction term, 0.000036 represents the difference in the slope for sqft_living, comparing sqft_above.

4. **Transformation Feature**

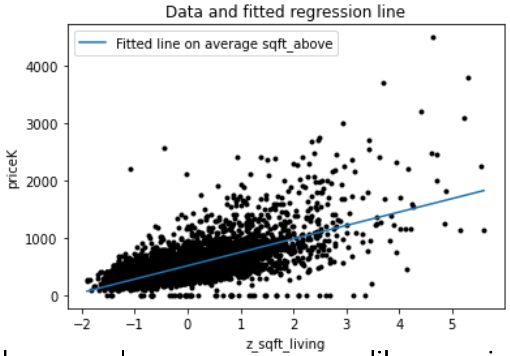
   a. Improve Coefficient Interpretation by Standardization

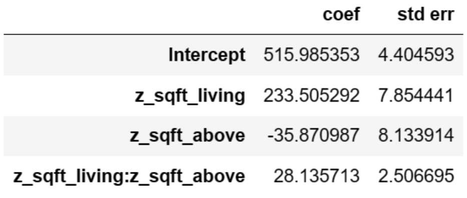
   To improve coefficient interpretation, we using standard z to sqft living and sqft above predictors set. With the z equation, we get the value of standard deviation sqft living is 926.25 and sqft above is 842.63. Then we fit the feature that transformed before to linear regression.



```
# Create OLS model object
model = smf.ols("priceK ~ z_sqft_living + z_sqft_above + z_sqft_living:z_sqft_above", df_new)

# Fit the model
results_z_sqft_living = model.fit()

# Extract the results (Coefficient and Standard Error) to DataFrame
results_sqft_living_std = print_coef_std_err(results_z_sqft_living)
```

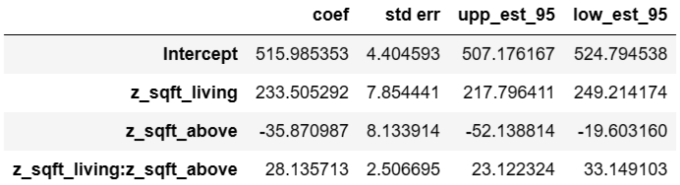
Here the result of z transformation to sqft living and sqft above.





	coef	std err
Intercept	515.985353	4.404593
z_sqft_living	233.505292	7.854441
z_sqft_above	-35.870987	8.133914
z_sqft_living:z_sqft_above	28.135713	2.506695

From these, we have same r-square like experiment before (0.4526). If the model correct, we can get the interval of estimated coefficients under normal distribution assumption using standard error from the model



	coef	std err	upp_est_95	low_est_95
Intercept	515.985353	4.404593	507.176167	524.794538
z_sqft_living	233.505292	7.854441	217.796411	249.214174
z_sqft_above	-35.870987	8.133914	-52.138814	-19.603160
z_sqft_living:z_sqft_above	28.135713	2.506695	23.122324	33.149103

This model has the intercept value of 515.98 and z_sqft_living coefficient of 233.51. We interpret these values based on the specified standardization parameters, which in this case is mean and standard deviation of height, globally.

F. Conclusion and Recommendations

- The conclusions of previous analysis
  - Based on the result, we can 95% confident that the average difference the price in Seattle and non Seattle lies between 10.819,97 and 73.899,28.
  - Based on the result, we can 95% confident that the average difference the price in good condition and other condition houses lies between 49.857,11 and 138.071,01.
  - The two variables that have highest relation to price houses is sqft living and sqft above variables.

- Based on our data, house have average sqft living and average sqft above has price `$ 515.98 K` on average
- Comparing houses who have the same condition 1 sqft living, the biggest have `$233.50` higher price in average than the lowest one
- The model that have highest variance sqft living and sqft above variables, it explained 45% of variance of house prices

- Recommendations for the business

  Since the higher the quality of a houses, the higher the premium will be. One of the indicator of higher quality is higher prices history. From the model we have sqft living and sqft above variables as two indicator that contribute to expected house prices

  The broker houses company can give higher prices to houses who has bigger sqft living and bigger sqft above since they might have higher quality other houses

- Recommendations for the next experiment

  The analysis taking into account only house prices and demographics variables, broker houses company may also consider broader factors such as market trends, regulatory requirements, and economic conditions when setting premium values.

G. Reference

https://www.kaggle.com/datasets/shree1992/housedata
https://www.forbes.com/advisor/mortgages/real-estate/housing-market-predictions/