UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF INFORMATION SYSTEMS

____

PROJECT REPORT REPORT

DATA ANALYSIS WITH R/PYTHON COURSE

TOPIC: USING LINEAR REGRESSION TO FORECAST HOME VALUES IN KING COUNTY, USA

Lecturer: Nguyen Phat Dat,
MA. Group: 08

Ho Chi Minh City, May 25, 2022

Table of Contents

GROUP MEMBERS

Order	Name	Student ID	Contribution
1	Phạm Thành Đạt	K194111601	20%
2	Huỳnh Nhật Hào	K19411603	20%
3	Trần Đức Duy	K194050694	20%
4	Lương Trường Phước	K194111624	20%
5	Trương Thành Sang	K184060799	20%

LIST OF ACRONYMS
NOTE ACRONYMS

Order	Acronyms	Explain
1	MSE	Mean Squared Error
2	RMSE	Root Mean Square Error
3	EDA	Exploratory Data Analysis
4	IQR	Interquartile Range
5	VIF	Variance Inflation Factor
6	RSS	Residual Sum of Squares
7	AIC	The Akaike Information Criterion
8	BIC	the Bayesian Information Criterion
9	Adj.R-square	Adjusted R-squared

LIST IMAGE

CHAPTER 1: INTRODUCTION

1.1 Requirement & Objective

There are many people who are concerned about the factors that affect house prices to ensure their interests. In fact, the house price is dependent on various factors such as bedrooms, bathrooms, floors, acreage…. The requirements are to find those factors with linear regression and to assess which of those is the most suitable to give the most effective forecasting results. Besides, we have to identify at least 3 regression models with more than 5 features for each model.

To meet the requirements, we define some objectives before research.

1. Find out the correlation between salary and other features.
2. Get an efficient model for the good forecasting results.

1.2 Questions of research

After specifying objectives, we have a big question: “Which factors are involved in a CEO's salary?”. Along with it, we divide it into three sub-questions as the following.

- Which factors are involved in a CEO's salary?
- What affects CEO’s salary the most?
- How does the field that CEO is working on correlate with his/her salary?

Question	Sub-Question	Variables	Evaluating Metrics
What factors affect house prices?	Which factor(s) affects house price the most?	Bedrooms, bathrooms, floors, year built, year renovated, acreage, waterfront.	Correlation coefficient, R-square, adj. R-square
	How does the correlation of these factors affect house price?	Bedrooms, bathrooms, floors, year built, year renovated, acreage, waterfront.	RSS, AIC, BIC, Adj.R-square

1.3 Research process

We have 7 steps to do the research

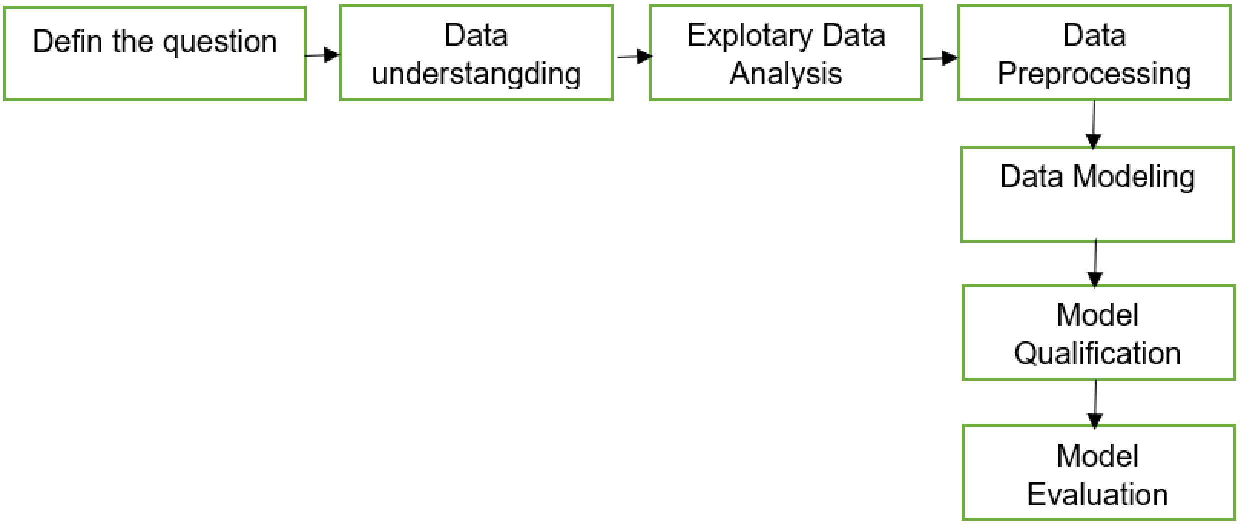


Image 1. 1 Research process

And training model details process is below:

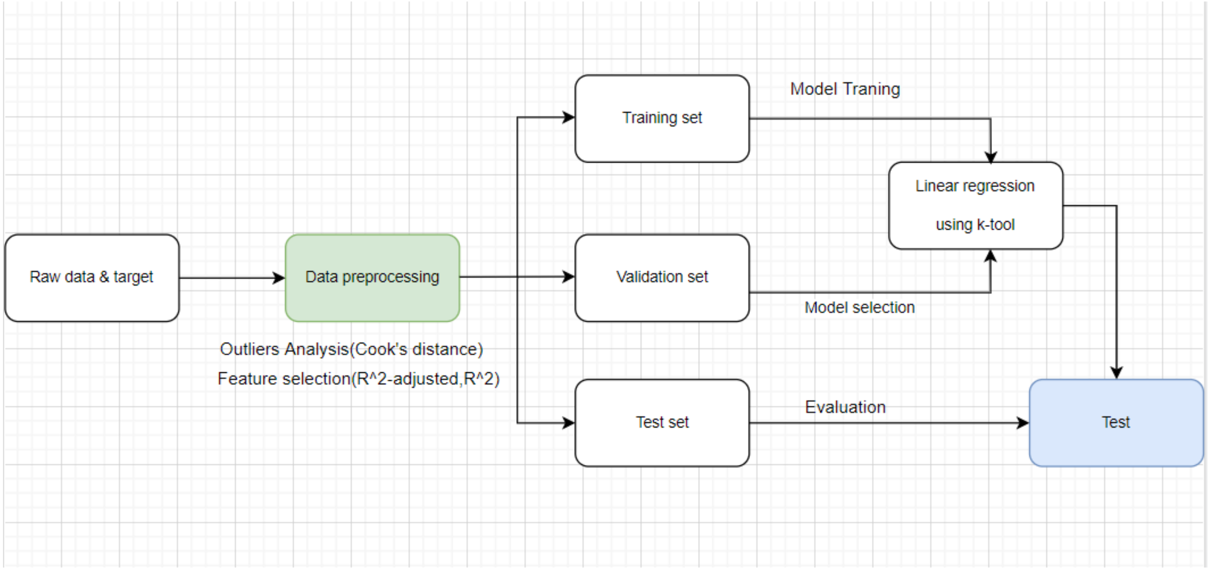


Image 1. 2 Training model process

1.3.1 Define the questions

In this first step, we will identify the problem we are facing. Furthermore, we need to clearly understand the requirements, the goal of this research.

1.3.2 Data understanding

After identifying what we need to answer, we explore the data set to have a better understanding about it. We find out that this data set contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

These are descriptions of the data set:

Price: house price, thousands $

Bedrooms: Number of rooms

Bathrooms: Number of rooms

Sqft_living: square feet living

Sqft_ Lot: square feet lot

Floor: Number of floors

Waterfront: Yes or No

View: Number of views

Condition

Grade

Sqft_above:

Sqft basement: (Sqft_living - Sqft_above)

Yr_built: Year built

Yr_renovated: Year renovated

1.3.3 Exploratory data analysis (EDA)

Exploratory data analysis (EDA) step is used to initially investigate the data set so that we can discover patterns, to spot anomalies, to test hypotheses and to check assumptions with summary statistics and the visualization of data.

1.3.4 Data preprocessing

The main task in this step is to clean the data set, transform it from raw data to the useful format. The data may have missing data or noisy data. If it has missing data, we can fill in the suitable value. In case of handling noisy data, Clustering will be used for finding the outliers and also in grouping the data. We need to select a feature selection.

1.3.5 Data modeling

After preprocessing the data, we perform data modeling on it. We find a modeling technique, generate test design, build a model and assess the model built. The data model is built to analyze relationships between various selected objects in the data. Test cases are built for assessing the model and model is tested and implemented on the data in this phase.

1.3.6 Qualification model

In this step, we will test the model to check its performance and interpret the results.

1.3.7 Evaluation model

Finally, we evaluate the results of the test cases and review the scope of errors in this step. We check if the results meet requirements of the research or not. We identify the problem and evaluate its performance, and find a method to improve it in the future

1.4 Tools used

In this study, we use the Python programming language, the main algorithm is Linear Regression and the libraries that support all our work are statsmodels, SKlearn. Initially, we have to use EDA to get the insight and find out the problems of the given dataset. Next step, outlier detection and collinearity problems are the main purposes at data preprocessing, after that we select important features through validating Adjusted R-squared. Then, the model is built based on the Linear Regression using K-fold Cross Validation. Following this, we evaluate the model based on Adjusted R-squared, MSE and RMSE score

CHAPTER 2: APPROACH

2.1 Cook’s distance

There are many techniques to remove outliers from a dataset. One method that is often used in regression settings is Cook’s Distance. Cook’s Distance is an estimate of the influence of a data point. It takes into account both the leverage and residual of each observation. Cook’s Distance is a summary of how much a regression model changes when the ith observation is removed. The concept was introduced by an American statistician named R. DennisCook, hence it was called after him. Formula for Cook’s distance:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{(p+1)\hat{\sigma}^2}$$

Where:

$\hat{Y}_j$: the jth fitted response value.

$\hat{Y}_{j(i)}$: the jth fitted response value, where the fit does not include observation i. p: the number of regression coefficients.

$\hat{\sigma}^2$: the estimated variance from the fit, based on all observations, i.e… Mean Squared

2.2 Feature Selection

Feature selection, one of the main components of feature engineering, is the process of selecting the most important features to input in machine learning algorithms.

Feature selection techniques are employed to reduce the number of input variables by eliminating redundant or irrelevant features and narrowing down the set of features to those most relevant to the machine learning model.

The main benefits of performing feature selection in advance, rather than letting the machine learning model figure out which features are most important, include:

1. Simpler models: simple models are easy to explain - a model that is too complex and unexplainable is not valuable.

2. Shorter training times: a more precise subset of features decreases the amount of time needed to train a model.

3. Variance reduction: increase the precision of the estimates that can be obtained for a given simulation.

4. Avoid the curse of high dimensionality: dimensionally cursed phenomena states that, as dimensionality and the number of features increases, the volume of space increases so fast

that the available data become limited - PCA feature selection may be used to reduce dimensionality.

2.3 R-Squared

R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. R-squared explains to what extent the variance of one variable explains the variance of the second variable. R-Squared is also called the coefficient of determination. It lies between 0% and 100%. A r-squared value of 100% means the model explains all the variation of the target variable. And a value of 0% measures zero predictive power of the model. Higher Rsquared value, better the model.

Formula for R-Squared:

$$\text{R-Squared} = \frac{TSS-RSS}{TSS} = 1 - \frac{Unexplained\ variation}{Total\ variation}$$

The actual calculation of R-squared requires several steps. This includes taking the data points (observations) of dependent and independent variables and finding the line of best fit, often from a regression model. From there we would calculate predicted values, subtract actual values and square the results. This yields a list of errors squared, which is then summed and equals the unexplained variance.

To calculate the total variance, we would subtract the average actual value from each of the actual values, square the results and sum them. From there, divide the first sum of errors (explained variance) by the second sum (total variance), subtract the result from one, and get the result R-squared

In fact, it suffers from a major flaw. Its value never decreases no matter the number of variables we add to our regression model. That is, even if we are adding redundant variables to the data, the value of R-squared does not decrease. It either remains the same or increases with the addition of new independent variables. This clearly does not make sense because some of the independent variables might not be useful in determining the target variable.

2.4 Adjusted R-squared

The Adjusted R-squared measures the proportion of variation explained by onl those independent variables that really help in explaining the dependent variable. In a portfolio model that has more independent variables, adjusted R-squared will help determine how much of the correlation with the index is due to the addition of those variables. The adjusted R-squared compensates for the addition of variables and only increases if the new predictor enhances the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance.

Formula for adjusted R-squared:

$$\text{Adj.}R^2 = 1 - \frac{(1-R^2)(n-1)}{(n-k-1)}$$

Where:

n: the number of data points in dataset
k: the number of independent variables
R^2: the R-squared values determined by the model.

2.5 Cross validation

2.5.1 Definition and techniques:

Cross-validation is a technique for evaluating a machine learning model and testing its performance. Cross-validation is commonly used in applied machine learning tasks. It helps to compare and select an appropriate model for the specific predictive modeling problem.

Cross-validation is easy to understand, easy to implement, and it tends to have a lower bias than other methods used to count the model's efficiency scores. All this makes cross-validation a powerful tool for selecting the best model for the specific task.

There are a lot of different techniques that may be used to cross-validate a model. Still, all of them have a similar algorithm:

1. Divide the dataset into two parts: one for training, other for testing

2. Train the model on the training set

3. Validate the model on the test set

4. Repeat 1-3 steps a couple of times. This number depends on the cross-validation method that you are using

There are plenty of cross-validation techniques. Some of them are commonly used, others work only in theory.

- Hold-out

- K-fold

- Leave-one-out

- Leave-p-out

- Stratified K-folds

- Repeated K-folds

- Nested K-folds

- Complete

In this report, our group focuses on using k-fold cross validation.

2.5.2. K-fold cross validation

K-fold cross validation is a technique that minimizes the disadvantages of the hold-out method. K-fold introduces a new way of splitting the dataset which helps to overcome the "test only once bottleneck".

The algorithm of the k-Fold technique:

1. Pick a number of folds – k. Usually, k is 5 or 10 but you can choose any number which is less than the dataset's length.
2. Split the dataset into k equal (if possible) parts (they are called folds)
3. Choose k – 1 folds as the training set. The remaining fold will be the test set
4. Train the model on the training set. On each iteration of cross-validation, you must train a new model independently of the model trained on the previous iteration
5. Validate on the test set
6. Save the result of the validation
7. Repeat steps 3 – 6 k times. Each time use the remaining fold as the test set. In the end, you should have validated the model on every fold that you have.
8. To get the final score average the results that you got on step 6.

2.6 Mean squared error

In statistics, the mean squared error (MSE) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors - that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

Formula for MSE:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} \quad (y_i - \tilde{y}_i)^2$$

Where:

n: number of items,

Σ: summation notation,

yi: original or observed y-value,

$\tilde{y}i$: y-value from regression.

General steps to calculate the MSE from a set of X and Y values:

1. Find the regression line.

2. Insert your X values into the linear regression equation to find the new Y, values (Y’).

3. Subtract the new Y value from the original to get the error.

4. Square the errors.

5. Add up the errors (the Σ in the formula is summation notation).

6. Find the mean.

2.7 Root Mean Square Error

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

Formula for RMSE:

$$RMSE = \sqrt{(\underline{(f - 0)^2})}$$

Where:

f: forecasts (expected values or unknown results),

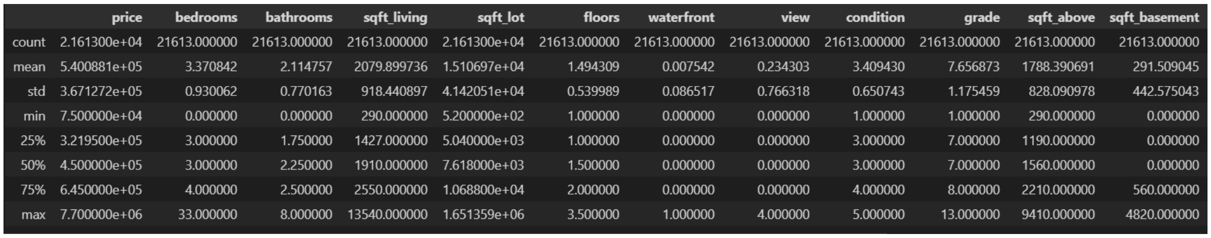
o: observed values (known results).

The bar above the squared differences is the mean (similar to x̄). The same formula can be written with the following, slightly different, notation (Barnston, 1992)

CHAPTER 3: LINEAR REGRESSION MODELING

3.1 EDA

3.1.1 Descriptive analysis

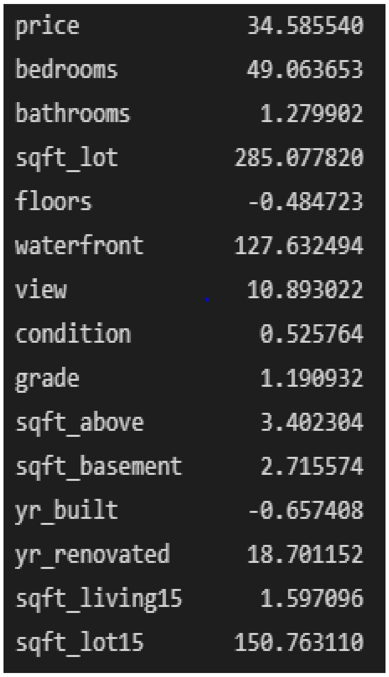


	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement
count	2.161300e+04	21613.000000	21613.000000	21613.000000	2.161300e+04	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000
mean	5.400881e+05	3.370842	2.114757	2079.899736	1.510697e+04	1.494309	0.007542	0.234303	3.409430	7.656873	1788.390691	291.509045
std	3.671272e+05	0.930062	0.770163	918.440897	4.142051e+04	0.539989	0.086517	0.766318	0.650743	1.175459	828.090978	442.575043
min	7.500000e+04	0.000000	0.000000	290.000000	5.200000e+02	1.000000	0.000000	0.000000	1.000000	1.000000	290.000000	0.000000
25%	3.219500e+05	3.000000	1.750000	1427.000000	5.040000e+03	1.000000	0.000000	0.000000	3.000000	7.000000	1190.000000	0.000000
50%	4.500000e+05	3.000000	2.250000	1910.000000	7.618000e+03	1.500000	0.000000	0.000000	3.000000	7.000000	1560.000000	0.000000
75%	6.450000e+05	4.000000	2.500000	2550.000000	1.068800e+04	2.000000	0.000000	0.000000	4.000000	8.000000	2210.000000	560.000000
max	7.700000e+06	33.000000	8.000000	13540.000000	1.651359e+06	3.500000	1.000000	4.000000	5.000000	13.000000	9410.000000	4820.000000

Image 3. 1 Overview description of all variables

- All of the count values of numerical variables are 21613. There is no null value in the data set.
- In this data set, the mean price is $5.400.881, the smallest value is $750.000 and the largest is $77.000.000, there is a big gap between them, it shows that the data set is skewed
- In the data set, the number of houses with waterfront accounts for very little with only 163 units, on average all houses have 1 floor and the number of bedrooms and bathrooms is approximate.
- The average of sqft_living equals the average of sqft_abve plus average of sqft_abve. This is perfectly reasonable because it is the nature of these variables
- The average of sqft_living is 2080 is not larger than sqft_living15 is 1986 and the average of sqft_lot is 15.106 is larger than sqft_lot15 is 12.768. This data set shows that there have been huge renovations with sqft_lot and sqft_living not changed significantly.
- The average of the condition variable is 3,4 and average of the grade variable is 7.65. This proves that the houses in King district have an average condition.

Sample Variance:

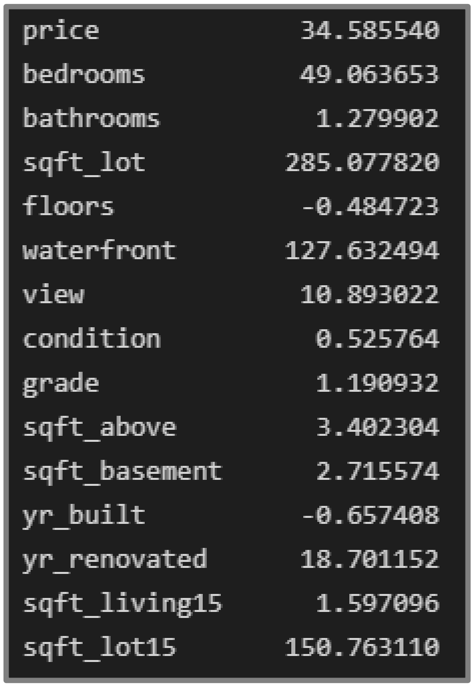


```
price             34.585540
bedrooms          49.063653
bathrooms          1.279902
sqft_lot         285.077820
floors            -0.484723
waterfront       127.632494
view              10.893022
condition          0.525764
grade              1.190932
sqft_above         3.402304
sqft_basement      2.715574
yr_built          -0.657408
yr_renovated      18.701152
sqft_living15      1.597096
sqft_lot15       150.763110
```

Image 3. 2 Sample variance of all the variables

- The variables: sqft_lot, waterfront, sqft_lot15, bedrooms, price, yr_renovated, view which have high sample variance, so the data of these variables has a strong dispersion
- The remaining variables are less dispersed.

Kurtosis:

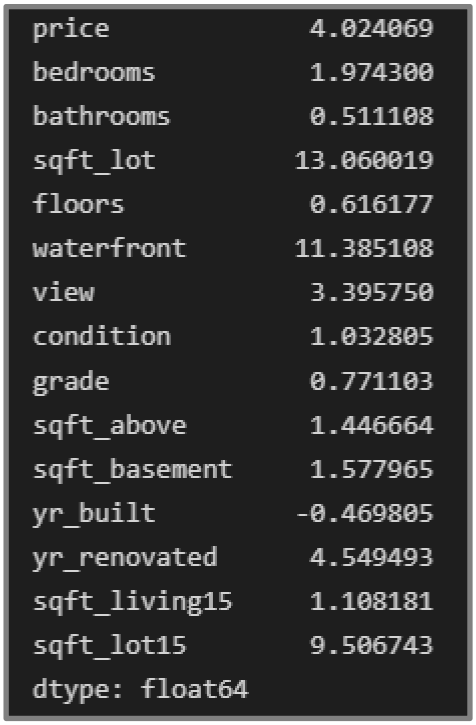


```
price             34.585540
bedrooms          49.063653
bathrooms          1.279902
sqft_lot         285.077820
floors            -0.484723
waterfront       127.632494
view              10.893022
condition          0.525764
grade              1.190932
sqft_above         3.402304
sqft_basement      2.715574
yr_built          -0.657408
yr_renovated      18.701152
sqft_living15      1.597096
sqft_lot15       150.763110
```

Image 3. 3 Kurtosis coefficient of all variables

- In the above statistics, sqft_above, sqft_basement variables have kurtosis approximately 3. Nên two variables have distribution near standard
- In the above statistics, (price, bedrooms, sqft_lot, waterfront, view, yr_renovated, sqft_lot15) have High kurtosis, showing that the impact of outliers is relatively large.
- (Bathrooms, floors, condition, yr_built) have Low kurtosis, showing that data has lack of the outliers

Skewness:



```
price             4.024069
bedrooms          1.974300
bathrooms         0.511108
sqft_lot         13.060019
floors            0.616177
waterfront       11.385108
view              3.395750
condition         1.032805
grade             0.771103
sqft_above        1.446664
sqft_basement     1.577965
yr_built         -0.469805
yr_renovated      4.549493
sqft_living15     1.108181
sqft_lot15        9.506743
dtype: float64
```

Image 3. 4 Skewness coefficient of all variables

- The 'bathrooms' variable and the 'floors' variable has a close distribution graph, while the remaining variables most have a positive distribution graph, except for 'yr_built' and 'lat'

3.1.2 Box plots

- Box plot indicates how the values in the data are spread out with a five-number summary of variables which includes minimum, first quartile (Q1), median, third quartile (Q3), and maximum.

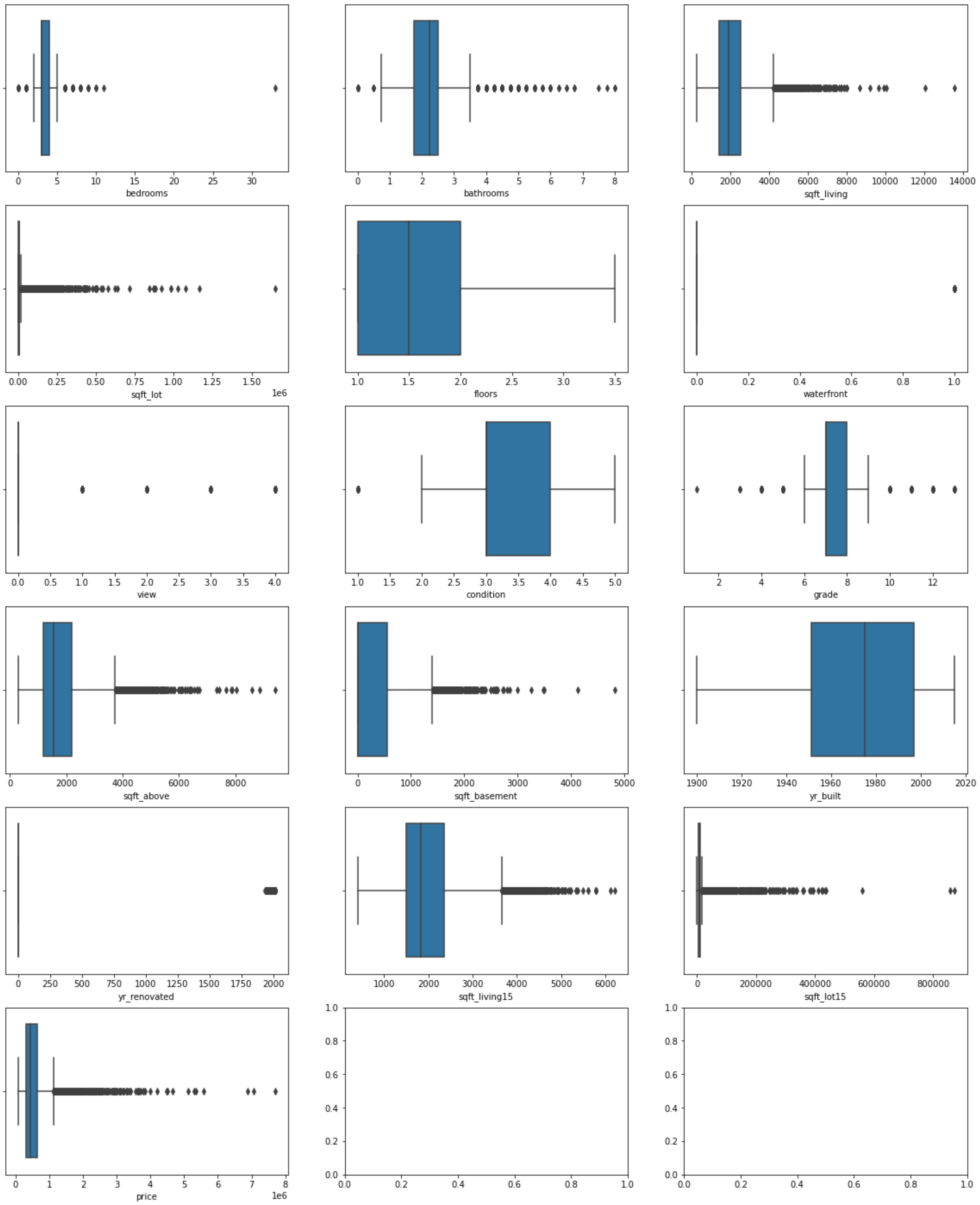


Image 3. 5 Box plots of all features

- All the given variables have outliers, in particular ‘sqft_lot’ with a lot of outliers on the plot. So, ‘price’ also has a lot of outliers. It means the price of houses get an increase or a decrease, which is far different from the others.
- The distribution of ‘price’ is strongly skewed-right with outliers. We will consider how to handle them in the following part.

3.1.3 Scatter Plots

- Scatter Plots is understood as a type of graph that expresses the correlation between the target variable (price) and predict variables.

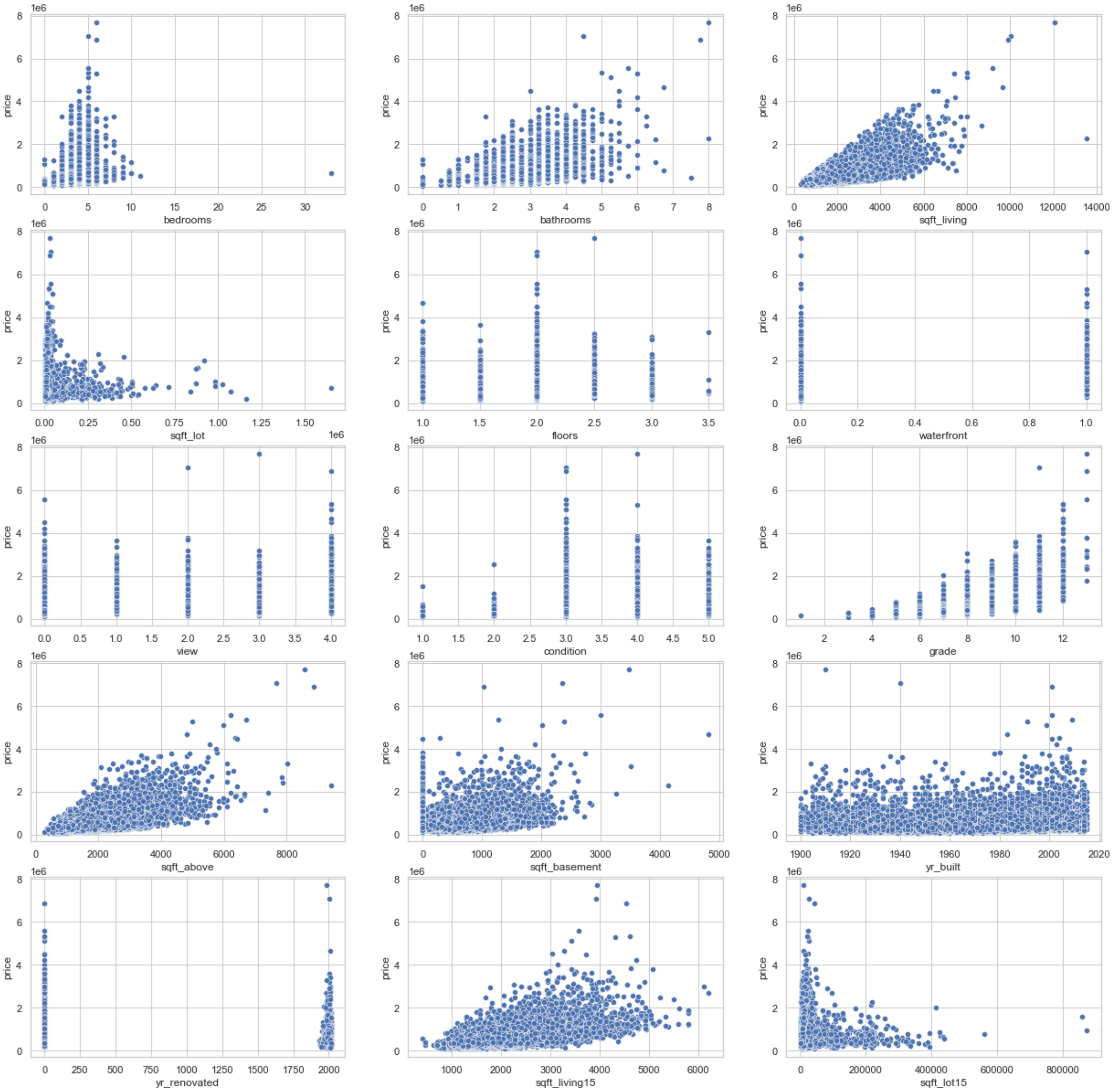


Image 3. 6 Scatter plots of predict variables with price variable

- Based on Scatter plot, we find variables bathroom, sqft_living, sqft_above, sqft_basement, sqft_living15 have positive linear correlation with target variables

3.1.4 Histogram

- Histogram used to monitor the allocation of variables.

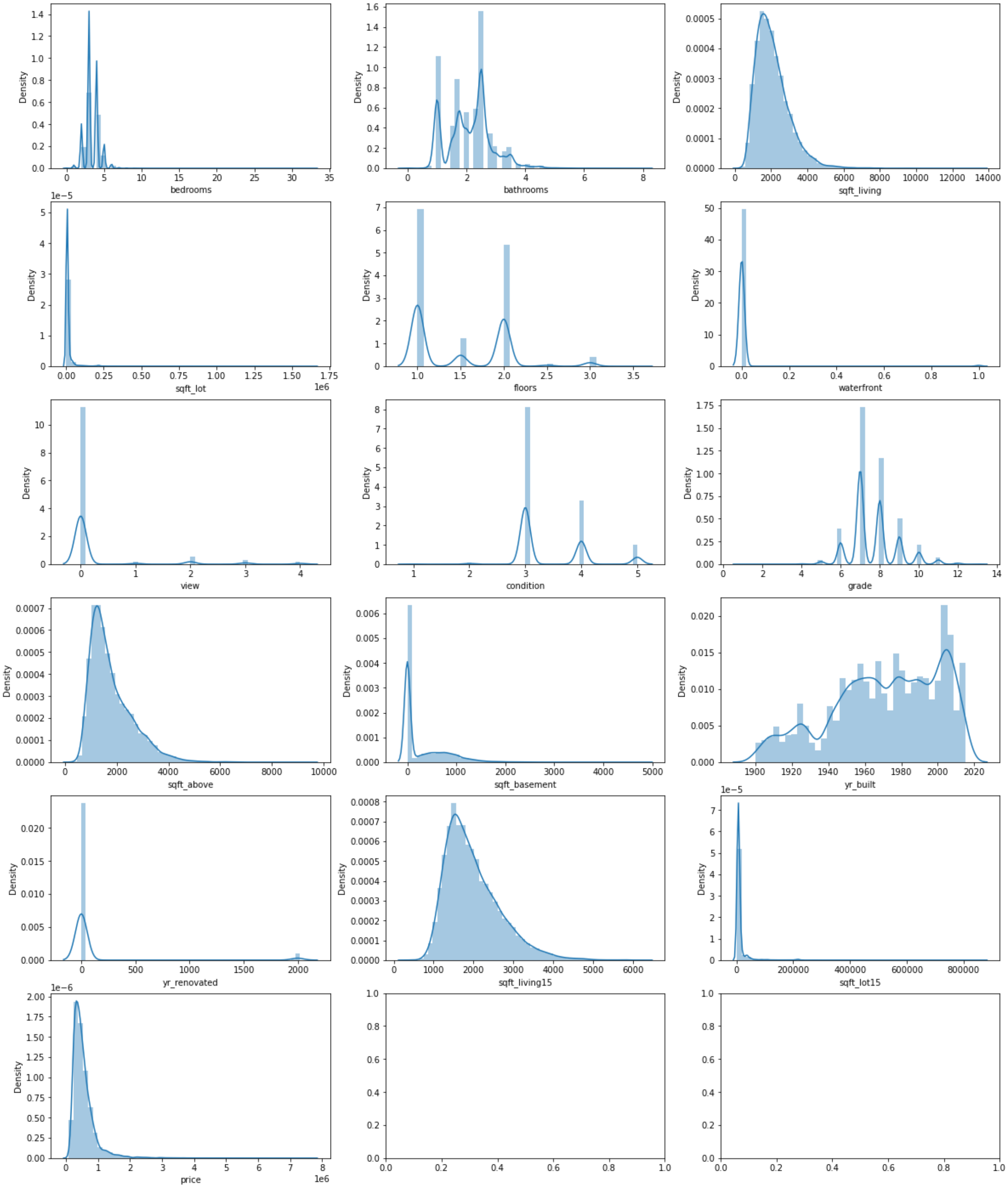


Image 3. 7 Histogram plots of all variables

- There are some variables that do not have normal distribution, such as bedroom, bathroom, sqft_lot... From the histograms, it is clear that those of sqft_living, sqft_above and sqft_living15 have a long tail with right-skewed distribution. They need to be undergoing log transformation to be normalized.

3.1.5 Correlation matrix

- A correlation matrix is simply a table which displays the correlation. It is best used in variables that demonstrate a linear relationship between each other, coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table.

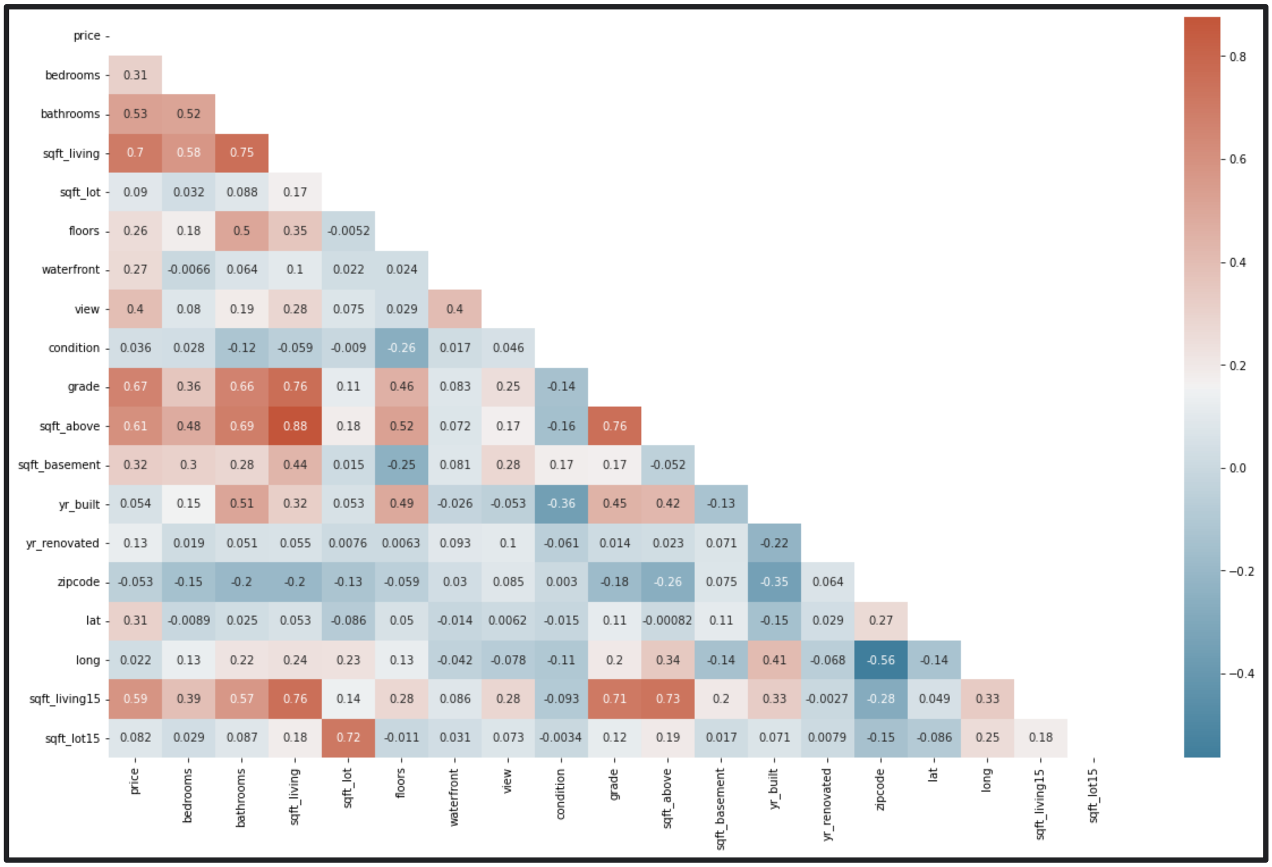


Image 3. 8 Heatmap indicate the correlation coefficient between variable pairs

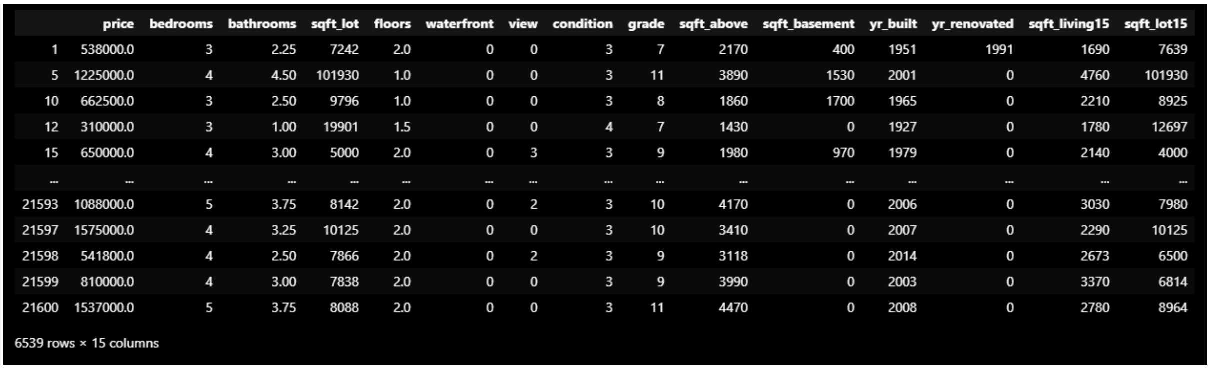
- From an overview, sqft_living & bathroom, sqft_living & sqft_living15, sqft_living & sqft_above, sqft_living & grade, sqft_lot & sqft_lot15, sqft_above & grade, sqft_living15 & grade, sqft_living15 & sqft_above have strong positive correlation with value greater than 0.7, which can indicate the presence of multicollinearity, as the result of logarithm. Therefore, we will check the dataset 's multicollinearity in the following part.
- Looking closer into the correlation of price with others, price has positive correlation with sqft_living, grade, sqft_above, sqft_living15. That is it and these variables have covariance

3.2 Preprocessing

- After using EDA techniques to have the first sight of the dataset and its distribution, we perform data preprocessing with outlier analysis and feature selection. The main goals are to remove outliers, standardize the data and select suitable features.

3.2.1 Outlier analysis

- We use Interquartile Range Rule method



	price	bedrooms	bathrooms	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	sqft_living15	sqft_lot15
1	538000.0	3	2.25	7242	2.0	0	0	3	7	2170	400	1951	1991	1690	7639
5	1225000.0	4	4.50	101930	1.0	0	0	3	11	3890	1530	2001	0	4760	101930
10	662500.0	3	2.50	9796	1.0	0	0	3	8	1860	1700	1965	0	2210	8925
12	310000.0	3	1.00	19901	1.5	0	0	4	7	1430	0	1927	0	1780	12697
15	650000.0	4	3.00	5000	2.0	0	3	3	9	1980	970	1979	0	2140	4000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
21593	1088000.0	5	3.75	8142	2.0	0	2	3	10	4170	0	2006	0	3030	7980
21597	1575000.0	4	3.25	10125	2.0	0	0	3	10	3410	0	2007	0	2290	10125
21598	541800.0	4	2.50	7866	2.0	0	2	3	9	3118	0	2014	0	2673	6500
21599	810000.0	4	3.00	7838	2.0	0	0	3	9	3990	0	2003	0	3370	6814
21600	1537000.0	5	3.75	8088	2.0	0	0	3	11	4470	0	2008	0	2780	8964

6539 rows × 15 columns

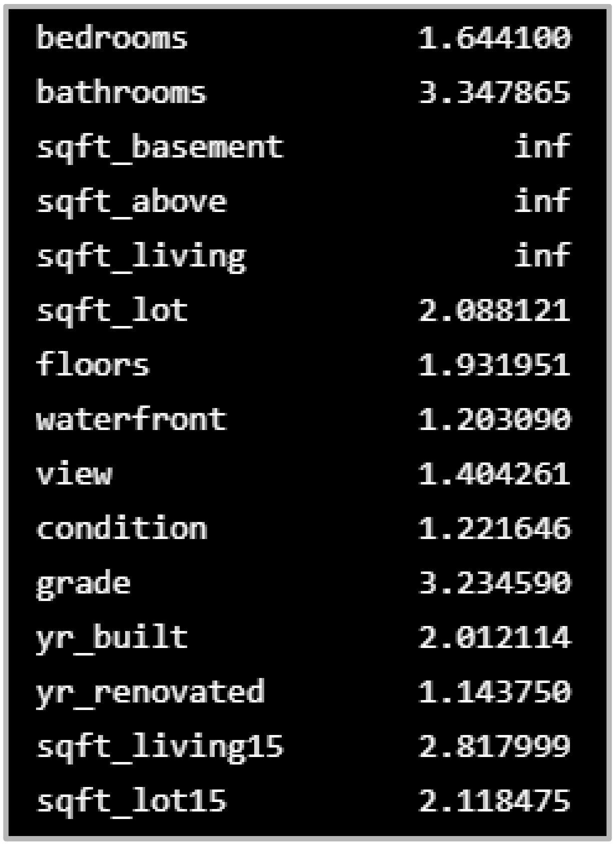
Image 3. 9 Data is thought to be an exception

- This is done using these steps:
    1. Calculate the interquartile range for the data.
    2. Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
    3. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
    4. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.
- We realize that exceptional values are accurate and appropriate values. They are not errors from data processing steps. So, we decided to keep these exceptional values. They play an important part of our model

3.2.2 Multicollinearity problem

- Multicollinearity happens if there are two or more independent variables that are highly correlated with another variable in a regression model. It can be a problem when we can not differentiate individual effects of independent variables from the dependent variable. To detect multicollinearity in the given dataset, we use Variable Inflation Factors (VIF) - the most common methods. VIF is the reciprocal of the tolerance value; small VIF values indicate low correlation among variables under ideal conditions at which VIF value is lower than 3. However, it is acceptable if it is less than 10 according to Vittinghoff E and his partners (2012).

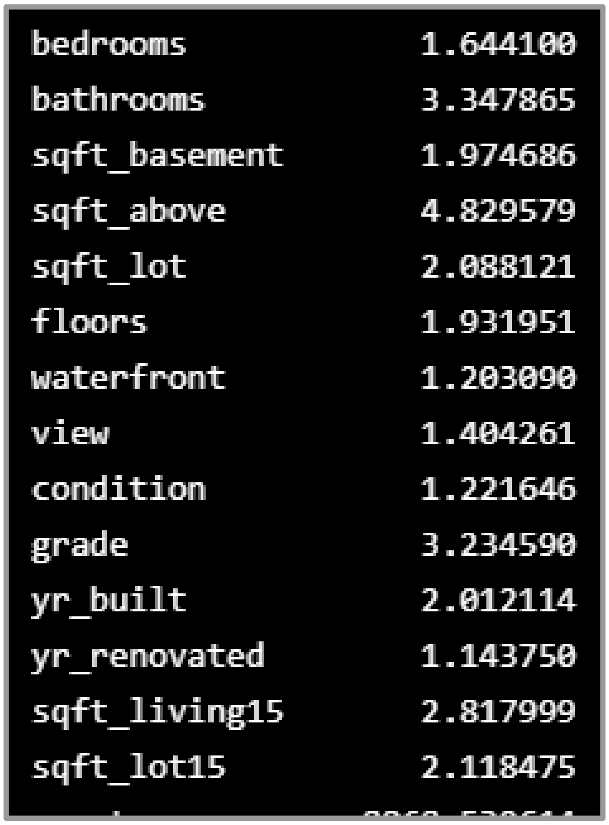
- VIF is used for continuous numerical variables. In this case we check if multicollinearity occurs in numerical ones as below.



```
bedrooms          1.644100
bathrooms         3.347865
sqft_basement          inf
sqft_above             inf
sqft_living            inf
sqft_lot          2.088121
floors            1.931951
waterfront        1.203090
view              1.404261
condition         1.221646
grade             3.234590
yr_built          2.012114
yr_renovated      1.143750
sqft_living15     2.817999
sqft_lot15        2.118475
```

Image 3. 10 VIF of all the features

- From the results calculated, there are three variables (sqft_basement, sqft_above, sqft_living) that have VIF equals infinity. This shows a perfect correlation between three independent variables. This is understandable, because the variable sqft_living in the dataset is equal to the total variable sqft_basement and the variable sqft_above.
- To overcome this, we eliminate A variable and run VIF again. We have results below:



```
bedrooms          1.644100
bathrooms         3.347865
sqft_basement     1.974686
sqft_above        4.829579
sqft_lot          2.088121
floors            1.931951
waterfront        1.203090
view              1.404261
condition         1.221646
grade             3.234590
yr_built          2.012114
yr_renovated      1.143750
sqft_living15     2.817999
sqft_lot15        2.118475
```

Image 3. 11 VIF of the remaining variables after eliminating sqft_living variable

- From the results calculated, all values are lower than 10. Inclusion is that there is no multicollinearity in numerical features

3.2.3 Future selection

- We will also use the Forward selection method to find R2 and Adjusted R2 scores. We use statsmodels.api to find the optimum number of features.
- Forward selection is an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.
- There is almost always an increase in the value of R2 score with the addition of new features.
- The increase in value of Adj R2 is less compared to R2 when the number of features increase from 10 to 14. The Adj R2 remains constant or reduces with the addition of every new feature.

STT	R-squared	Adj. R-squared	Features
1	**0.798**	**0.798**	sqft_above
2	**0.839**	**0.839**	sqft_above, sqft_basement
3	**0.853**	**0.853**	sqft_above, sqft_basement, view
4	**0.858**	**0.858**	sqft_above, sqft_basement, view, waterfront
5	**0.860**	**0.860**	sqft_above, sqft_basement, view, waterfront, bedrooms
6	**0.865**	**0.865**	sqft_above, sqft_basement, view, waterfront, bedrooms, grade
7	**0.875**	**0.875**	sqft_above, sqft_basement, view, waterfront, bedrooms, grade, yr_built
8	**0.878**	**0.878**	sqft_above, sqft_basement, view, waterfront, bedrooms, grade, yr_built, condition
9	**0.880**	**0.879**	sqft_above, sqft_basement, view, waterfront, bedrooms, grade, yr_built, condition, yr_renovated
10	**0.880**	**0.880**	sqft_above, sqft_basement, view, waterfront, bedrooms, grade, yr_built, condition, yr_renovated, sqft_lot15
11	**0.880**	**0.880**	sqft_above, sqft_basement, view, waterfront, bedrooms, grade, yr_built, condition, yr_renovated, sqft_lot15, sqft_living15
12	**0.880**	**0.880**	sqft_above, sqft_basement, view, waterfront, bedrooms, grade, yr_built, condition, yr_renovated, sqft_lot15,

			sqft_living15, bathrooms
13	**0.880**	**0.880**	sqft_above, sqft_basement, view, waterfront, bedrooms, grade, yr_built, condition, yr_renovated, sqft_lot15, sqft_living15, bathrooms, floors
14	**0.880**	**0.880**	sqft_above, sqft_basement, view, waterfront, bedrooms, grade, yr_built, condition, yr_renovated, sqft_lot15, sqft_living15, bathrooms, floors, sqft_lot

- Adj. R-Squared coefficient increases gradually when adding variables and unchanged from the 10th variable, but in the set of variables No. 13.14 new variables 'floors', 'sqft_lot' is added without statistical significance. 'Price' (Target Variables).

- Value of R-squared is 88%, indicating that those variables can explain 88% about the model. In addition, I also consider other indicators to choose the appropriate number of features such as: RSS, AIC, BIC.
- The charts below will assist us in choosing the appropriate number of features

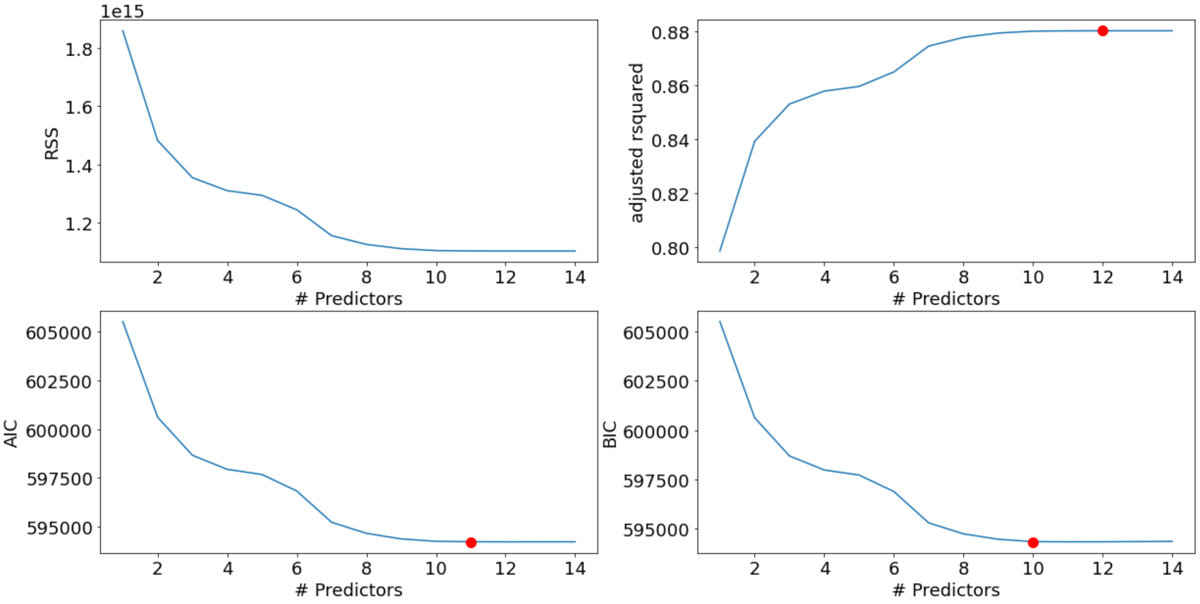


Image 3. 12 Line chart describes the change of RSS, adj. r-Square, AIC, BIC by the

number of features

Figure 3.12 Line chart describes the change of RSS, adj. r-Square, AIC, BIC by the number of features

Base on the above result, we choose 3 models with the following features:
- Multiple linear 10 features: sqft_above, sqft_basement, view, waterfront, bedrooms, grade, yr_built, condition, yr_renovated, sqft_lot15
- Multiple linear 11 features: sqft_above, sqft_basement, view, waterfront, bedrooms, grade, yr_built, condition, yr_renovated, sqft_lot15, sqft_living15
- Multiple linear 12 features: sqft_above, sqft_basement, view, waterfront, bedrooms, grade, yr_built, condition, yr_renovated, sqft_lot15, sqft_living15, bathrooms

4. Modeling

First, we normalize the data using the minmax method from scikit-learn library. The, we applied a linear regression on some features to predict the target and give models:

- With 10 features:

price = 1750825.1199477***sqft_above** + 890681.45437358***sqft_basement** + **187893.55051283*****view** + **537463.06258666*****waterfront** - **1054721.32858896***, **bedrooms** + 1541129.20241037***grade** **-** 346923.18868423***yr_built** + 79605.39267542 * **condition** + 41301.67750076* **yr_renovated** - 559218.53686164* **sqft_lot15** - 392252.7655628517

- With 11 features:

price = 1695099.89878308***sqft_above** + 873510.92434182***sqft_basement** + 183469.41046896***view** + 539777.61241601***waterfront** **-** 1051063.66793404***bedrooms** + 1514338.39491067***grade** **-** 346819.35764972***yr_built** **+** 79756.33060158***condition** + 43037.59079663***yr_renovated** - 571151.01340501***sqft_lot15** + 82309.01135332* **sqft_living15** **-** 389872.07503063534

- With 12 features:

price = 1483424.95743491***sqft_above** + 741673.34740902***sqft_basement** + 179291.02242302***view** + 541257.85049297***waterfront** **-** 1255549.11397273***bedrooms** + 1459108.78914243***grade** **-** 400516.57319204***yr_built** **+** 73226.26598962***condition**

+ 24096.75523225* **yr_renovated -** 522739.37839228* **sqft_lot15** + 107874.14689939* **sqft_living15** +435996.56575025 *** bathrooms -** 380346.27274930326

5. Model validation

- In the previous, we found 4 sets of important features that make the model better. Now, we build a model with K-Fold cross validation whose number of folds is 10. The model is built based on Statsmodel.OLS, sklearn model_selection library then using some metrics to evaluate the model via most optimal mean and standard value of model scores, and Linear Regression Score.
- The model is implemented to train with each of feature sets, and collected results then following:

Numbers of feature	The mean of the folds	The standard deviation of the folds	Linear Regression Score
10 features	0.6479882226789826	0.018563376651076963	0.6494838188397358
11 features	0.6481415217894172	0.018117876921146053	0.6497449778048168
12 features	0.6520923767332547	0.018235955237121577	0.6540432636215237

- The values of model validation of each metric are equivalent to each other but from the data in the table above, it shows that multiple linear regression (12 features) model is better than others with the small standard deviation value (0.0182) and highest mean value (0.65209) and Linear Regression Score value (0.6540432636215237)

6. Use data to experiment

We evaluate the experimental model selected by drawing the lines plot that indicate the predictable price and the actual price of the test data

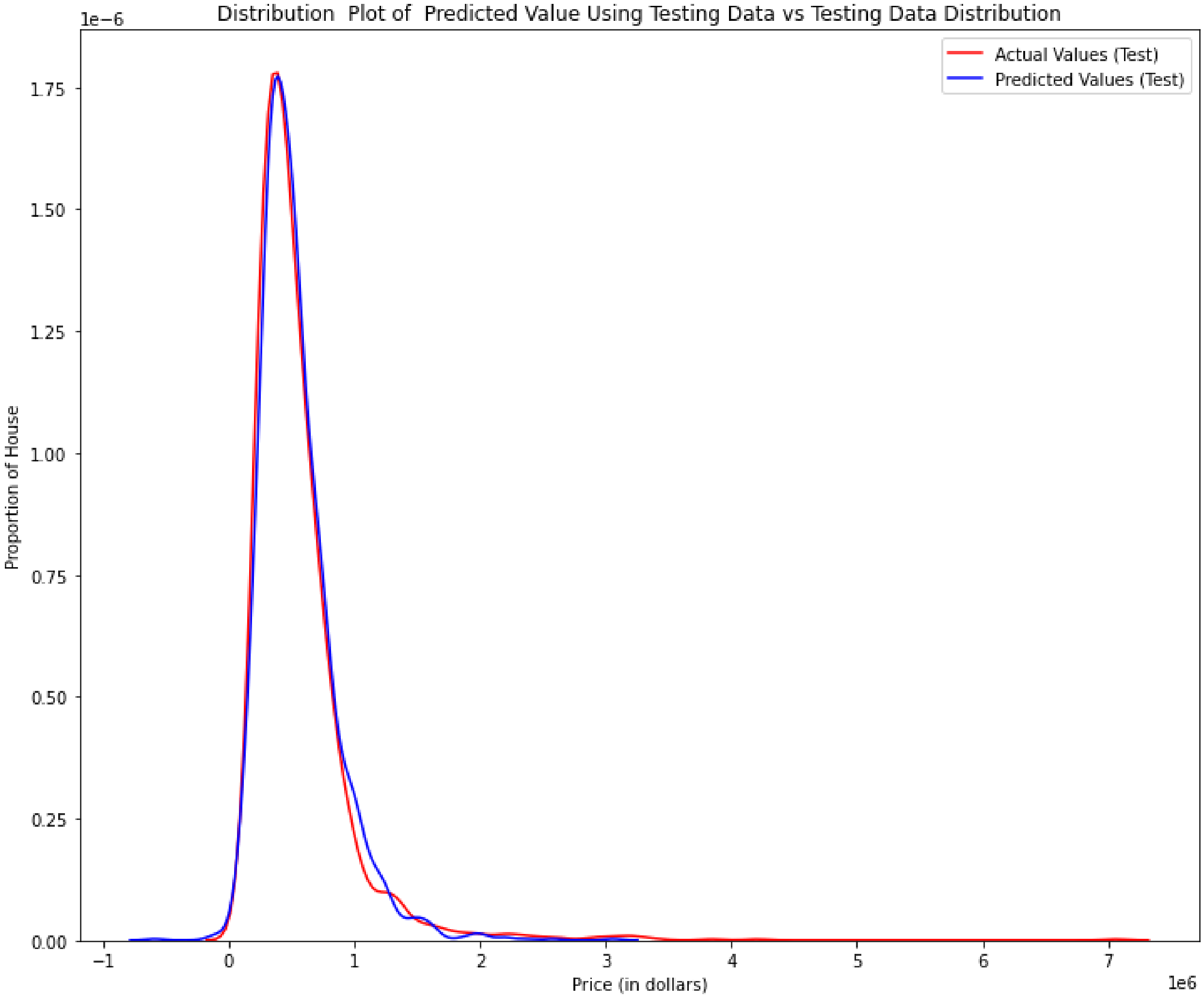


Image 3. 13 Distribution Plot of Predicted Value Using Testing Data vs Testing Data Distribution

- After building and evaluating and selecting the most optimal model, we will apply the model to solve a specific case. Because this is the data collected from 2014-2015, we will be set from May 2015.
- A person wants to buy a house in King district, USA. He wants his new home, has 4 bedrooms for 4 people, 2 bathrooms, (6*25) house area equals 150 square meters equivalent to 1615 square feet, of which 323 square feet basement area, has a view

to a waterfront, not yet seen, overall grade is 9/13, the condition at 4/5, built in 2014, has not been renovated and has not changed the house and land lot from the time of construction.

- From the above requirement, the model predicted the price he had to pay to get the desired house as:

price = 1483424.95743491***1292** **+** 741673.34740902***323** **+** 179291.02242302***0** + 541257.85049297***1** **-** 1255549.11397273***4** + 1459108.78914243***9** **-** 400516.57319204***2014** + 73226.26598962***4** + 24096.75523225***0** **-** 522739.37839228***1615** + 107874.14689939* **1615 +** 435996.56575025*** 2** **-**380346.27274930326 = **688933401.3688245(USD)**

CHAPTER 4: CONCLUSION

After building and evaluating the model, we have determined that a linear regression model with 12 features is the most appropriate model.

The model shows that the variables sqft_above, grade, bedrooms have the strongest impact on the price target variable. This is the answer to the question we posed.

On the other hand, our study still has many limitations. Because the model only explains 65% of the target variable. There are other variables that affect the target variable that the data does not have.

The size of the data set is quite small, we have difficulty in dividing the data set into training sets and validation sets. With a small set, it does not have many samples, and if it does so, the number of observations in validation data may be too few to give meaningful performance estimates.

In the feature selection step, because of the limited facilities, we have to use the Forward selection method. This method cannot scan through all subset combinations. Because a variable can affect the whole when there is the presence or absence of one or more other variables.

The data set was collected in 2015 so it is not highly applicable at present. Based on this analysis, we will implement related analysis projects on new datasets to increase realism.

REFERENCES

[1] Jeffrey M. Wooldridge (2012). Introductory econometrics a modern approach (5th edition).

[2] Prasad Patil (2018). What is Exploratory Data Analysis? Retrieved from: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[3] Sadhvi Anunaya (2021). Data Preprocessing in Data Mining -A HandsOn Guide. Retrieved from: https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/

[4] Data Analytics Tutorial for Beginners – From Beginner to Pro in 10 Mins! (2018). Retrieved from: https://data-flair.training/blogs/data-analytics-tutorial/

[5] Problems of Small Data and How to Handle Them (2016). Retrieved from: https://www.edupristine.com/blog/managing-small-data

[6] Akshita Chugh (2020). MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? Retrieved from: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

[7] Jason Fernando (2021). R-squared. Retrieved from: https://www.investopedia.com/terms/r/r-squared.asp

[8] Hussain Mujtaba (2021). What is Cross Validation in Machine learning? Types of Cross Validation. Retrieved from: https://www.mygreatlearning.com/blog/cross-validation

[9] NIST/SEMATECH e-Handbook of Statistical Methods, Section 3 http://www.itl.nist.gov/div898/handbook/ , date 16/11/2021

[10] Christian Thieme (2021). Identifying Outliers in Linear Regression — Cook’s Distance. Retrieved from: https://towardsdatascience.com/identifying-outliers-in-linear-regression-cooks- distance-9e212e9136a