STATISTICAL ANALYSIS REPORT on HOUSING PRICE

By Jason Adika Tanuwijaya

ID ~ BS23DSY031

Contents

1. Introduction

Analyzing house prices using data is essential for both potential homebuyers and real estate experts. By utilizing statistical models, machine learning, and extensive datasets covering aspects such as location, size, amenities, and market trends, analysts can uncover the complex patterns influencing property values. Through thorough scrutiny and interpretation of this data, valuable insights emerge, exposing the key factors driving changes in housing markets. This process enables informed decision-making for buyers, sellers, and investors as they navigate the dynamic real estate landscape.

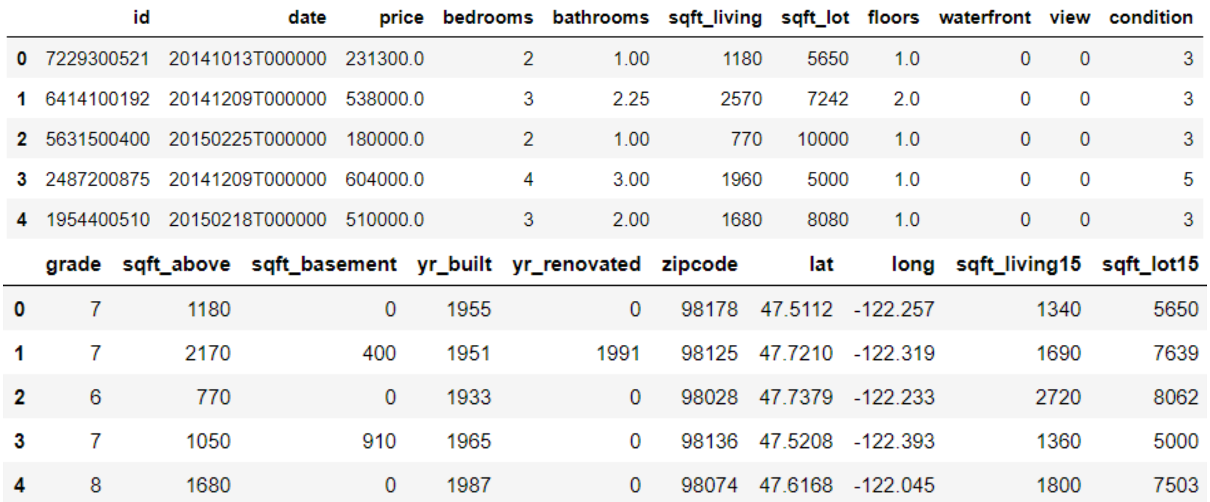
The objective of this is to **predict the house price** based on as the number of bedrooms, bathrooms, living area size, lot size, and location details. Moreover, the dataset can serve as a foundation for developing recommendation systems for homebuyers, guiding them towards properties that align with their preferences and requirements. Overall, the housing price dataset offers a wealth of insights and opportunities for leveraging data-driven approaches to understand and navigate the housing market effectively.

Below are the X variable that we are going to use for predicting the Y dependent variable which is the house pricing:

- Bedrooms
- Bathrooms
- sqft_living
- sqft_lot
- floors
- Waterfront
- View
- Condition
- Grade
- sqft_above
- yr_built
- lat
- long
- sqft_living15
- sqft_lot15

2. Data Overview

2.1. Sample Data

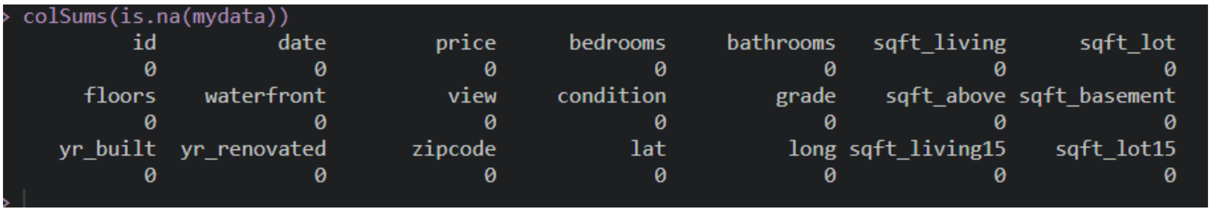


	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition
0	7229300521	20141013T000000	231300.0	2	1.00	1180	5650	1.0	0	0	3
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	3
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	3
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	5
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	3

	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
0	7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
1	7	2170	400	1951	1991	98125	47.7210	-122.319	1690	7639
2	6	770	0	1933	0	98028	47.7379	-122.233	2720	8062
3	7	1050	910	1965	0	98136	47.5208	-122.393	1360	5000
4	8	1680	0	1987	0	98074	47.6168	-122.045	1800	7503

The Dataset shown that the top 5 data in the table, we can observe that

- The data set contains 21613 obervations
- The data of the house is unbiased beacuse it was selected from multiple real-estate website
- There are no Null or NaN value in the dataset after doing testing

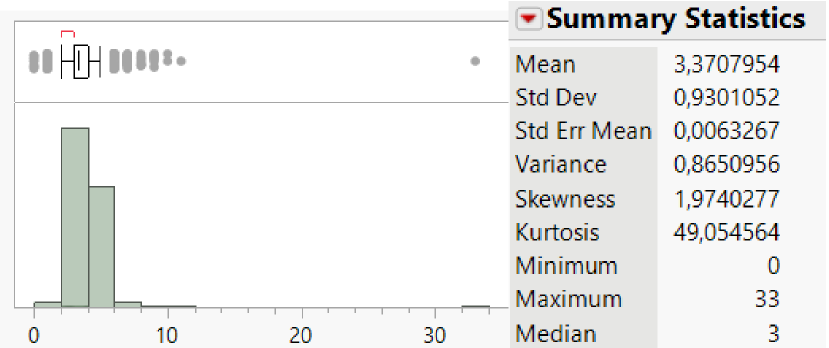


```
> colSums(is.na(mydata))
           id          date         price      bedrooms     bathrooms   sqft_living      sqft_lot 
            0             0             0             0             0             0             0 
       floors    waterfront          view     condition         grade    sqft_above sqft_basement 
            0             0             0             0             0             0             0 
     yr_built  yr_renovated       zipcode           lat          long sqft_living15    sqft_lot15 
            0             0             0             0             0             0             0 
> 
```

- From the data we will exclude “id”, “date”, “yr_renovated”, and “zipcode” from using it in making the model as is it not use for the regression

2.2. Data Description

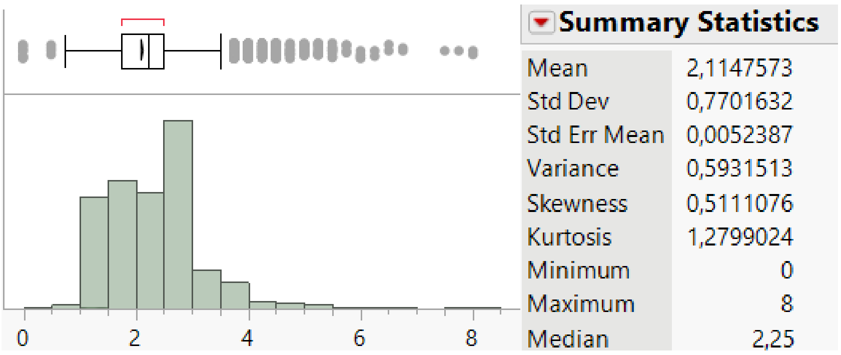
2.2.1. bedrooms



Summary Statistics

Mean	3,3707954
Std Dev	0,9301052
Std Err Mean	0,0063267
Variance	0,8650956
Skewness	1,9740277
Kurtosis	49,054564
Minimum	0
Maximum	33
Median	3

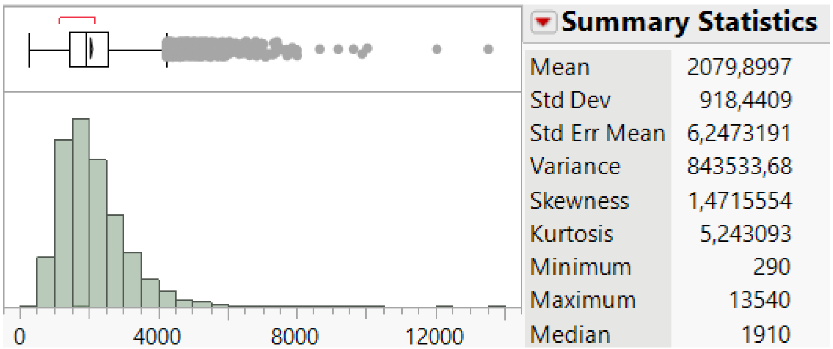
2.2.2. bathrooms



Summary Statistics

Mean	2,1147573
Std Dev	0,7701632
Std Err Mean	0,0052387
Variance	0,5931513
Skewness	0,5111076
Kurtosis	1,2799024
Minimum	0
Maximum	8
Median	2,25

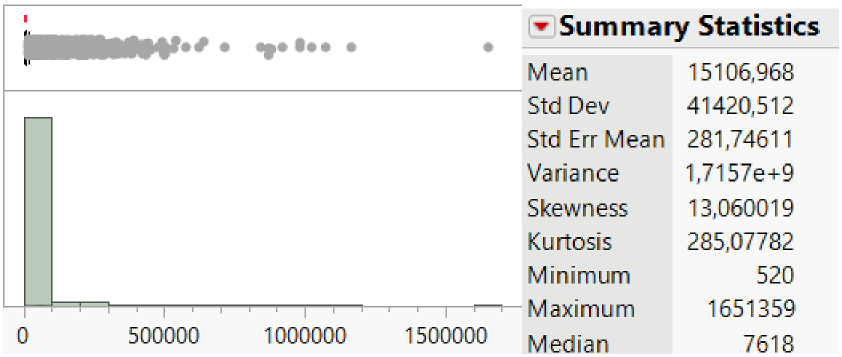
2.2.3. *sqft_living*



Summary Statistics

Mean	2079,8997
Std Dev	918,4409
Std Err Mean	6,2473191
Variance	843533,68
Skewness	1,4715554
Kurtosis	5,243093
Minimum	290
Maximum	13540
Median	1910

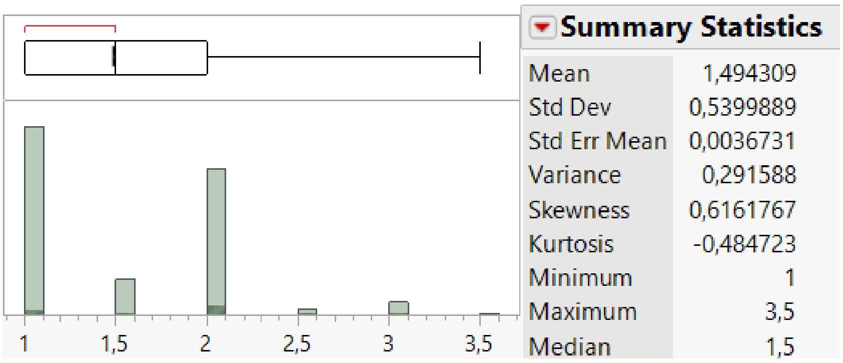
2.2.4. *sqft_lot*



Summary Statistics

Mean	15106,968
Std Dev	41420,512
Std Err Mean	281,74611
Variance	1,7157e+9
Skewness	13,060019
Kurtosis	285,07782
Minimum	520
Maximum	1651359
Median	7618

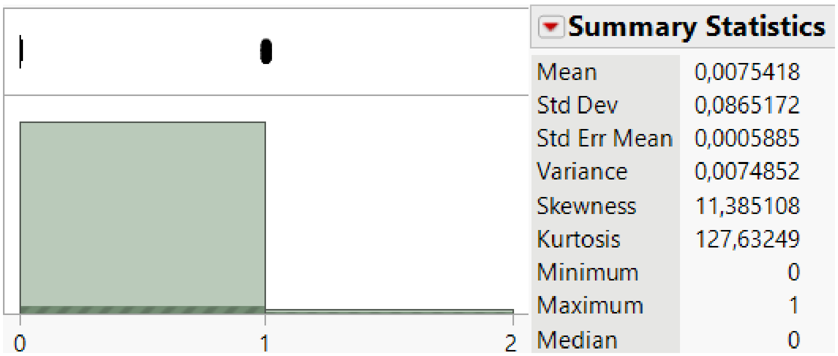
2.2.5. *floors*



Summary Statistics

Mean	1,494309
Std Dev	0,5399889
Std Err Mean	0,0036731
Variance	0,291588
Skewness	0,6161767
Kurtosis	-0,484723
Minimum	1
Maximum	3,5
Median	1,5

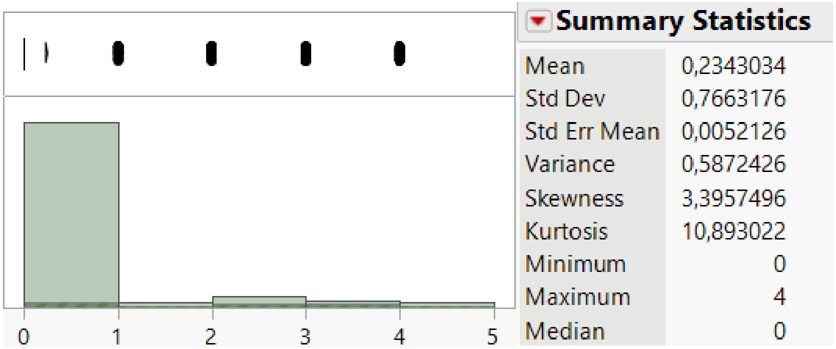
2.2.6. *waterfront*



Summary Statistics

Mean	0,0075418
Std Dev	0,0865172
Std Err Mean	0,0005885
Variance	0,0074852
Skewness	11,385108
Kurtosis	127,63249
Minimum	0
Maximum	1
Median	0

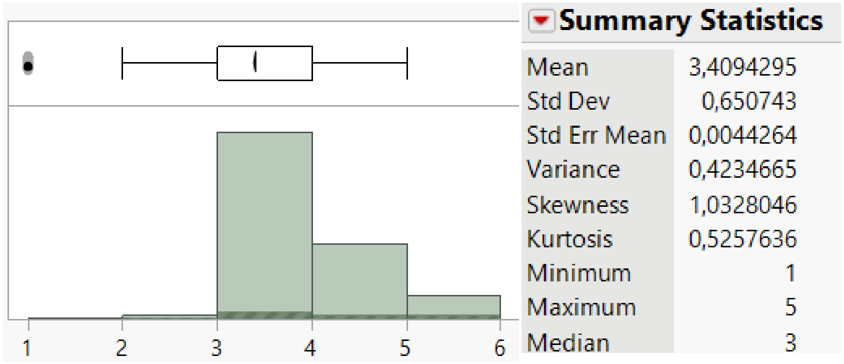
2.2.7. *view*



Summary Statistics

Mean	0,2343034
Std Dev	0,7663176
Std Err Mean	0,0052126
Variance	0,5872426
Skewness	3,3957496
Kurtosis	10,893022
Minimum	0
Maximum	4
Median	0

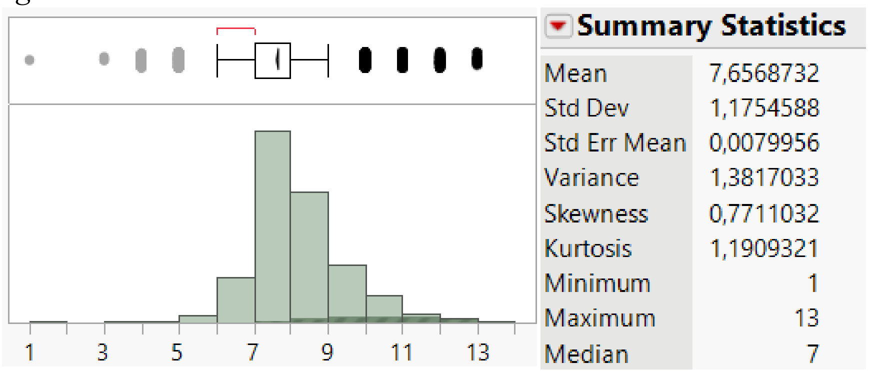
2.2.8. *condition*



Summary Statistics

Mean	3,4094295
Std Dev	0,650743
Std Err Mean	0,0044264
Variance	0,4234665
Skewness	1,0328046
Kurtosis	0,5257636
Minimum	1
Maximum	5
Median	3

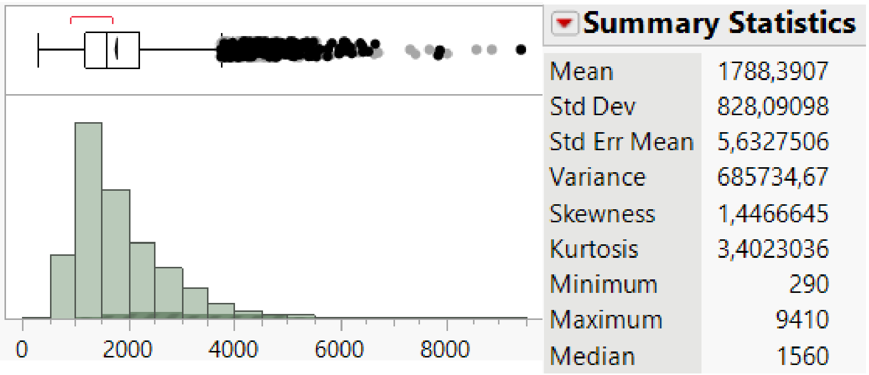
2.2.9. *grade*



Summary Statistics

Mean	7,6568732
Std Dev	1,1754588
Std Err Mean	0,0079956
Variance	1,3817033
Skewness	0,7711032
Kurtosis	1,1909321
Minimum	1
Maximum	13
Median	7

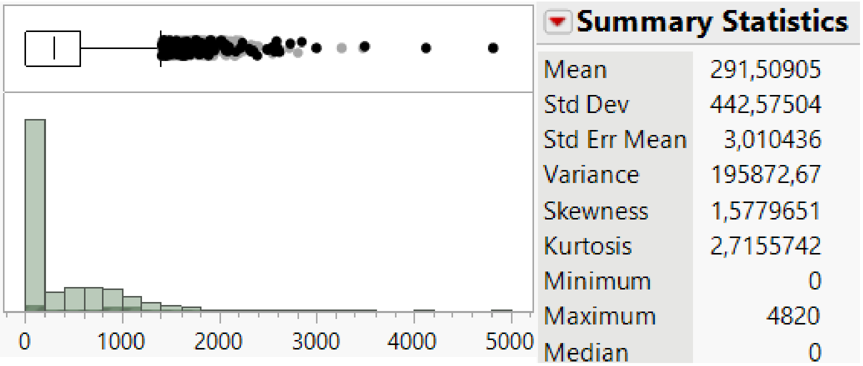
2.2.10. *sqft_above*



Summary Statistics

Mean	1788,3907
Std Dev	828,09098
Std Err Mean	5,6327506
Variance	685734,67
Skewness	1,4466645
Kurtosis	3,4023036
Minimum	290
Maximum	9410
Median	1560

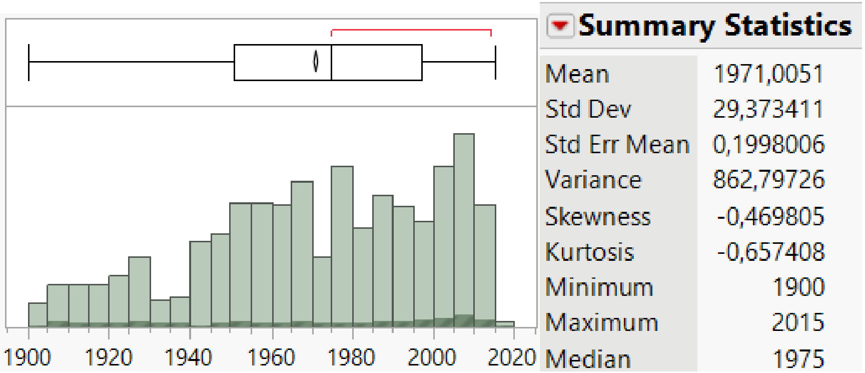
2.2.11. *sqft_basement*



Summary Statistics

Mean	291,50905
Std Dev	442,57504
Std Err Mean	3,010436
Variance	195872,67
Skewness	1,5779651
Kurtosis	2,7155742
Minimum	0
Maximum	4820
Median	0

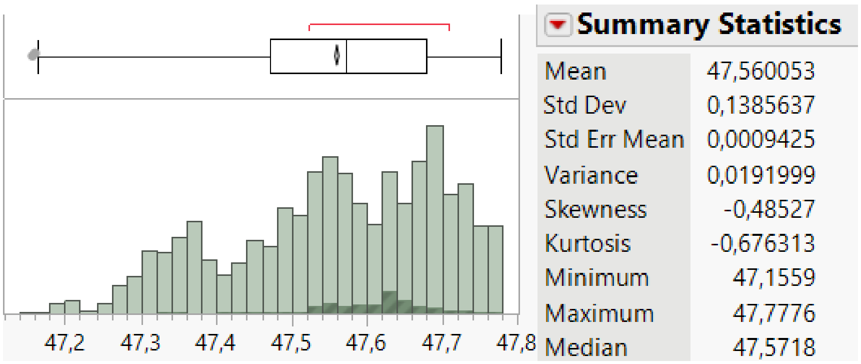
2.2.12. *yr_built*



Summary Statistics

Mean	1971,0051
Std Dev	29,373411
Std Err Mean	0,1998006
Variance	862,79726
Skewness	-0,469805
Kurtosis	-0,657408
Minimum	1900
Maximum	2015
Median	1975

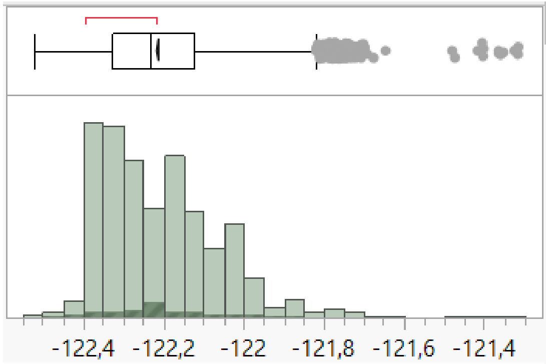
2.2.13. *lat*

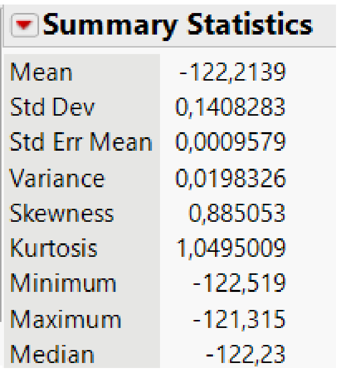


Summary Statistics

Mean	47,560053
Std Dev	0,1385637
Std Err Mean	0,0009425
Variance	0,0191999
Skewness	-0,48527
Kurtosis	-0,676313
Minimum	47,1559
Maximum	47,7776
Median	47,5718

2.2.14. Long

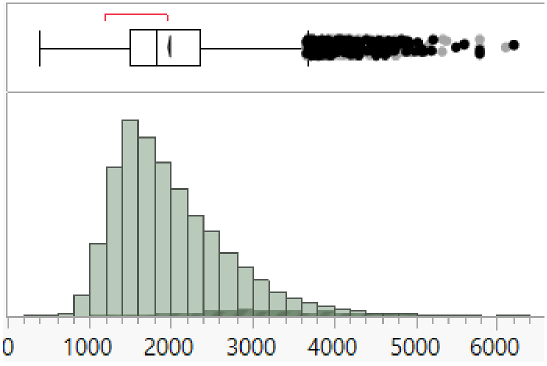


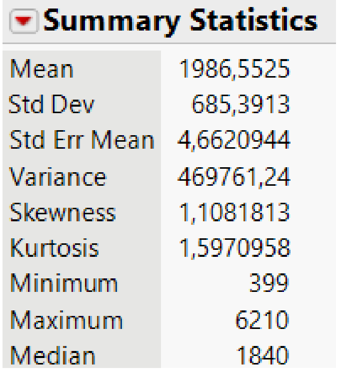


Summary Statistics

Mean	-122,2139
Std Dev	0,1408283
Std Err Mean	0,0009579
Variance	0,0198326
Skewness	0,885053
Kurtosis	1,0495009
Minimum	-122,519
Maximum	-121,315
Median	-122,23

2.2.15. sqft_living15

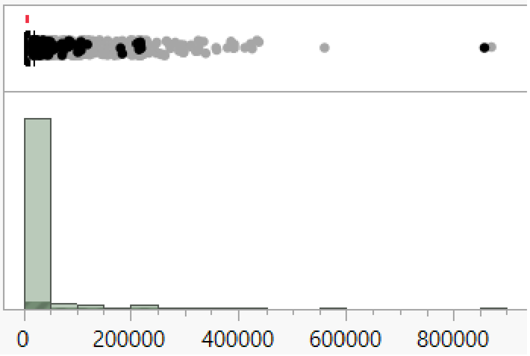


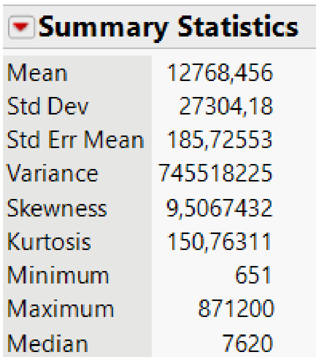


Summary Statistics

Mean	1986,5525
Std Dev	685,3913
Std Err Mean	4,6620944
Variance	469761,24
Skewness	1,1081813
Kurtosis	1,5970958
Minimum	399
Maximum	6210
Median	1840

2.2.16. sqft_lot15

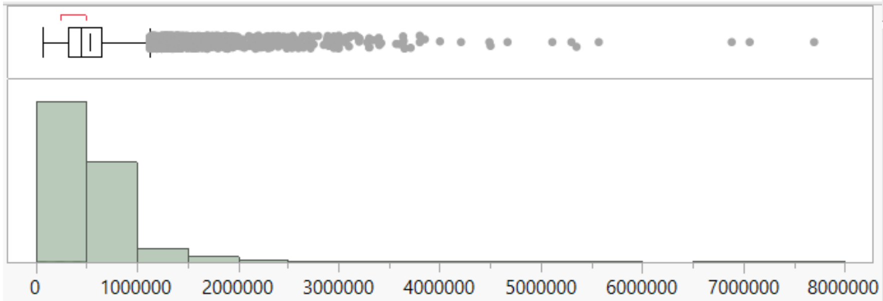


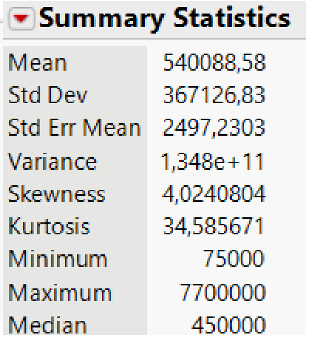


Summary Statistics

Mean	12768,456
Std Dev	27304,18
Std Err Mean	185,72553
Variance	745518225
Skewness	9,5067432
Kurtosis	150,76311
Minimum	651
Maximum	871200
Median	7620

2.2.17. Price (dependent variable)





Summary Statistics

Mean	540088,58
Std Dev	367126,83
Std Err Mean	2497,2303
Variance	1,348e+11
Skewness	4,0240804
Kurtosis	34,585671
Minimum	75000
Maximum	7700000
Median	450000

Observation about data summary above:

- From the price data we can see that the skewness is positively skew (4,02) and have positive kurtosis (34,58). This is also show that the median (450000) is smaller than the mean (540088) of the price
- There are many others variable too with outliers, noticeably “sqft_lot15”, “sqft_living15”, “sqft_above”, ”sqft_basement”,”sqft_living”, “view”, and “bathrooms”
- Also the variable that have right tail outliers in their distribution, have positive skewness following the price variable

2.3. Correlation

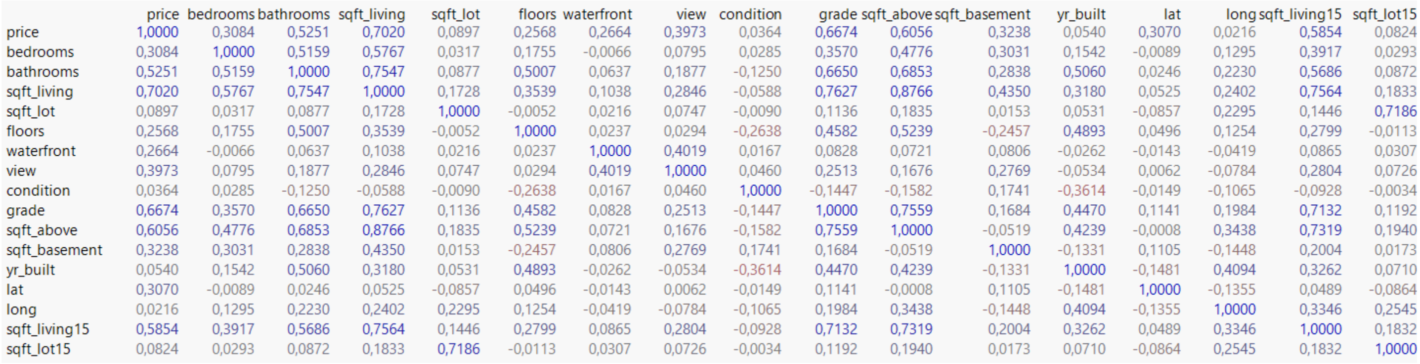
Correlation coefficient between two r.v.s X and Y, usually denoted by $r(X,Y)$ *or* r_{XY} is a numerical measure of linear relationship between them and is defined as:

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- r_{XY} provided a measure of linear relationship between X and Y.
- It is a measure of degree of relationship.

2.3.1. Correlation Matrix

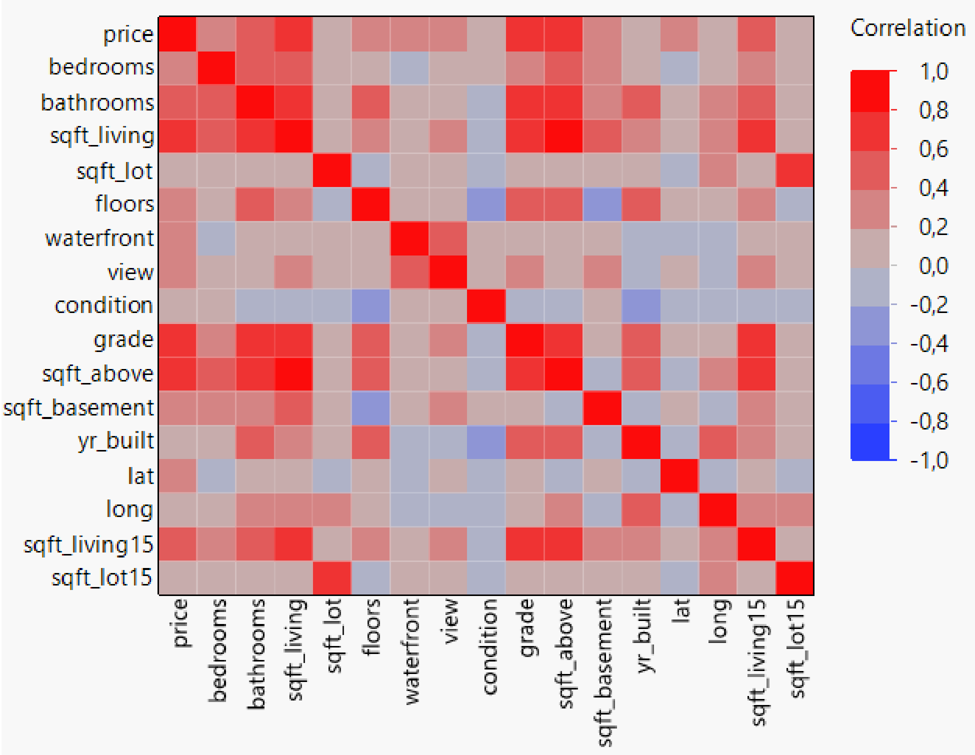
A correlation matrix is a statistical technique used to evaluate the relationship between two variables in a data set. The matrix is a table in which every cell contains a correlation coefficient, where 1 is considered a strong relationship between variables, 0 a neutral relationship and -1 a not strong relationship.



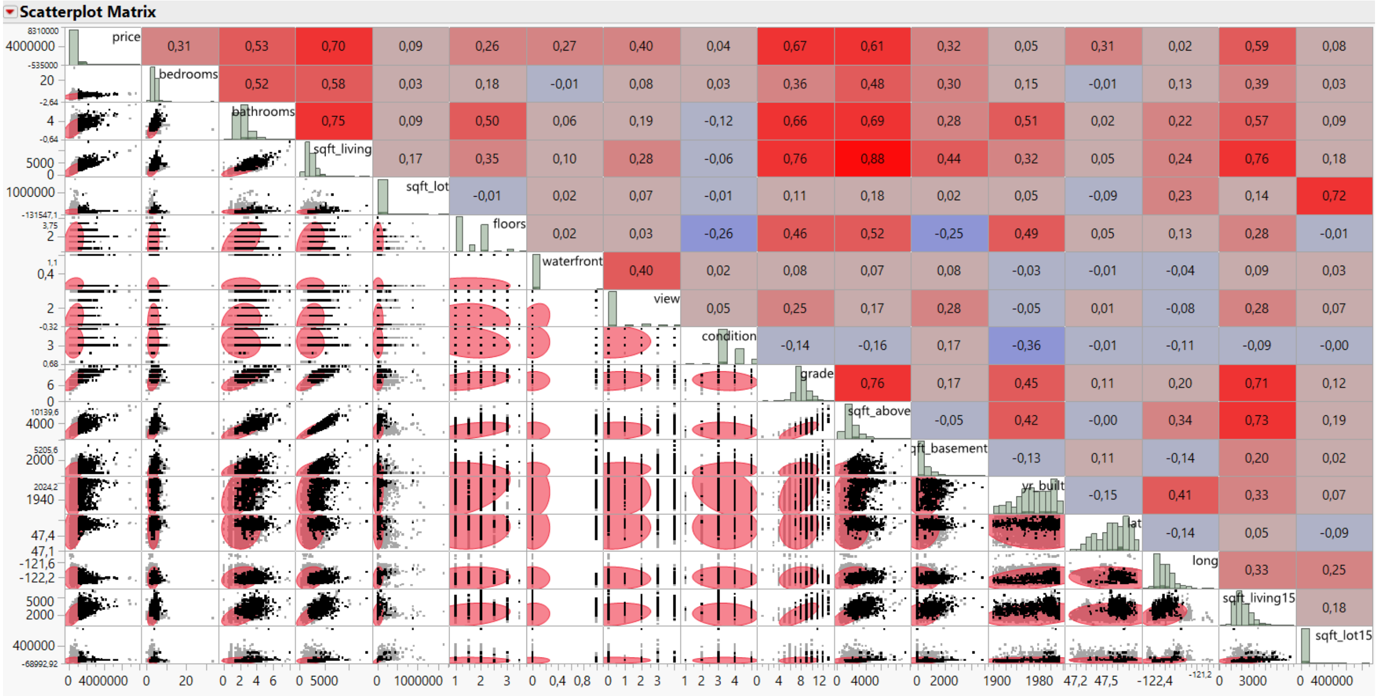
	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	lat	long	sqft_living15	sqft_lot15
price	1,0000	0,3084	0,5251	0,7020	0,0897	0,2568	0,2664	0,3973	0,0364	0,6674	0,6056	0,3238	0,0540	0,3070	0,0216	0,5854	0,0824
bedrooms	0,3084	1,0000	0,5159	0,5767	0,0317	0,1755	-0,0066	0,0795	0,0285	0,3570	0,4776	0,3031	0,1542	-0,0089	0,1295	0,3917	0,0293
bathrooms	0,5251	0,5159	1,0000	0,7547	0,0877	0,5007	0,0637	0,1877	-0,1250	0,6650	0,6853	0,2838	0,5060	0,0246	0,2230	0,5686	0,0872
sqft_living	0,7020	0,5767	0,7547	1,0000	0,1728	0,3539	0,1038	0,2846	-0,0588	0,7627	0,8766	0,4350	0,3180	0,0525	0,2402	0,7564	0,1833
sqft_lot	0,0897	0,0317	0,0877	0,1728	1,0000	-0,0052	0,0216	0,0747	-0,0090	0,1136	0,1835	0,0153	0,0531	-0,0857	0,2295	0,1446	0,7186
floors	0,2568	0,1755	0,5007	0,3539	-0,0052	1,0000	0,0237	0,0294	-0,2638	0,4582	0,5239	-0,2457	0,4893	0,0496	0,1254	0,2799	-0,0113
waterfront	0,2664	-0,0066	0,0637	0,1038	0,0216	0,0237	1,0000	0,4019	0,0167	0,0828	0,0721	0,0806	-0,0262	-0,0143	-0,0419	0,0865	0,0307
view	0,3973	0,0795	0,1877	0,2846	0,0747	0,0294	0,4019	1,0000	0,0460	0,2513	0,1676	0,2769	-0,0534	0,0062	-0,0784	0,2804	0,0726
condition	0,0364	0,0285	-0,1250	-0,0588	-0,0090	-0,2638	0,0167	0,0460	1,0000	-0,1447	-0,1582	0,1741	-0,3614	-0,0149	-0,1065	-0,0928	-0,0034
grade	0,6674	0,3570	0,6650	0,7627	0,1136	0,4582	0,0828	0,2513	-0,1447	1,0000	0,7559	0,1684	0,4470	0,1141	0,1984	0,7132	0,1192
sqft_above	0,6056	0,4776	0,6853	0,8766	0,1835	0,5239	0,0721	0,1676	-0,1582	0,7559	1,0000	-0,0519	0,4239	-0,0008	0,3438	0,7319	0,1940
sqft_basement	0,3238	0,3031	0,2838	0,4350	0,0153	-0,2457	0,0806	0,2769	0,1741	0,1684	-0,0519	1,0000	-0,1331	0,1105	-0,1448	0,2004	0,0173
yr_built	0,0540	0,1542	0,5060	0,3180	0,0531	0,4893	-0,0262	-0,0534	-0,3614	0,4470	0,4239	-0,1331	1,0000	-0,1481	0,4094	0,3262	0,0710
lat	0,3070	-0,0089	0,0246	0,0525	-0,0857	0,0496	-0,0143	0,0062	-0,0149	0,1141	-0,0008	0,1105	-0,1481	1,0000	-0,1355	0,0489	-0,0864
long	0,0216	0,1295	0,2230	0,2402	0,2295	0,1254	-0,0419	-0,0784	-0,1065	0,1984	0,3438	-0,1448	0,4094	-0,1355	1,0000	0,3346	0,2545
sqft_living15	0,5854	0,3917	0,5686	0,7564	0,1446	0,2799	0,0865	0,2804	-0,0928	0,7132	0,7319	0,2004	0,3262	0,0489	0,3346	1,0000	0,1832
sqft_lot15	0,0824	0,0293	0,0872	0,1833	0,7186	-0,0113	0,0307	0,0726	-0,0034	0,1192	0,1940	0,0173	0,0710	-0,0864	0,2545	0,1832	1,0000

2.3.2. *Correlation Colour Map*

A colour map highlights the sign and magnitude of coefficients in a patterned matrix. Red indicate negative values, blue indicates positive values. Intensity of the colour represents the magnitude of the value, the darker more extreme.



2.3.3. *Correlation Scatterplot*

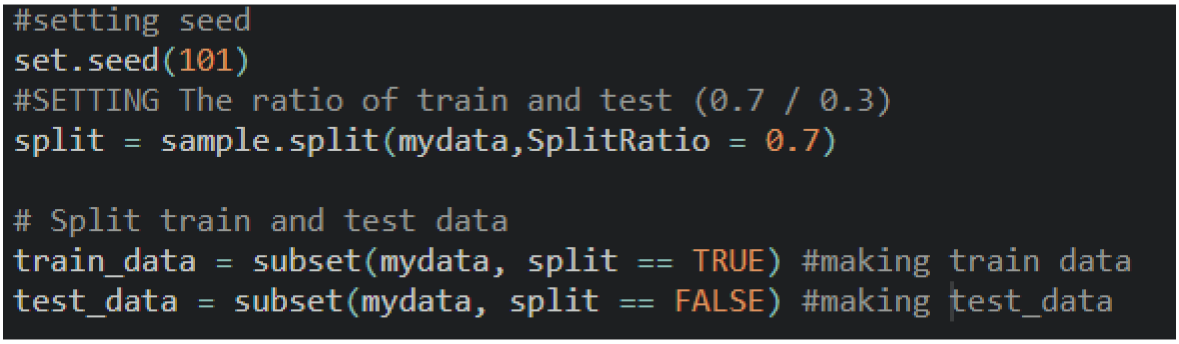


Analysis of Correlation:

- Shaded elipse has α=0.01
- Generally the correlation relationship show that most of the variable doesn't have strong relationship with other variable
- Sqft_living has the highest correlation with price (0,75), follow by grade (0,67) and sqft_above (0,61)
- On the otherhand, sqft_lot, condition, yr_built, long, and sqft_lot15 almost has no relation with price (< 0,1)

3. Splitting Data

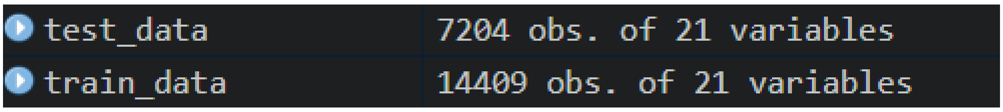
Below are the command in R for splitting the data into 2 dataset which are **train_data** and **test_data** with ratio 7:3 respectively.



```
#setting seed
set.seed(101)
#SETTING The ratio of train and test (0.7 / 0.3)
split = sample.split(mydata,SplitRatio = 0.7)

# Split train and test data
train_data = subset(mydata, split == TRUE) #making train data
test_data = subset(mydata, split == FALSE) #making test_data
```

- Train data has 70% of the original data which are 7024 entries
- Test data has 30% of the original data which are 14409 entries



```
test_data        7204 obs. of 21 variables
train_data       14409 obs. of 21 variables
```

- Train_data will be used for creating the regression model
- Test_data will be used for testing the accuracy of our model

Note that after testing with the train data "sqft_basement" won't be used in the training model because the sample that we got are all zero's value

4. Regression Analysis

4.1. Theory

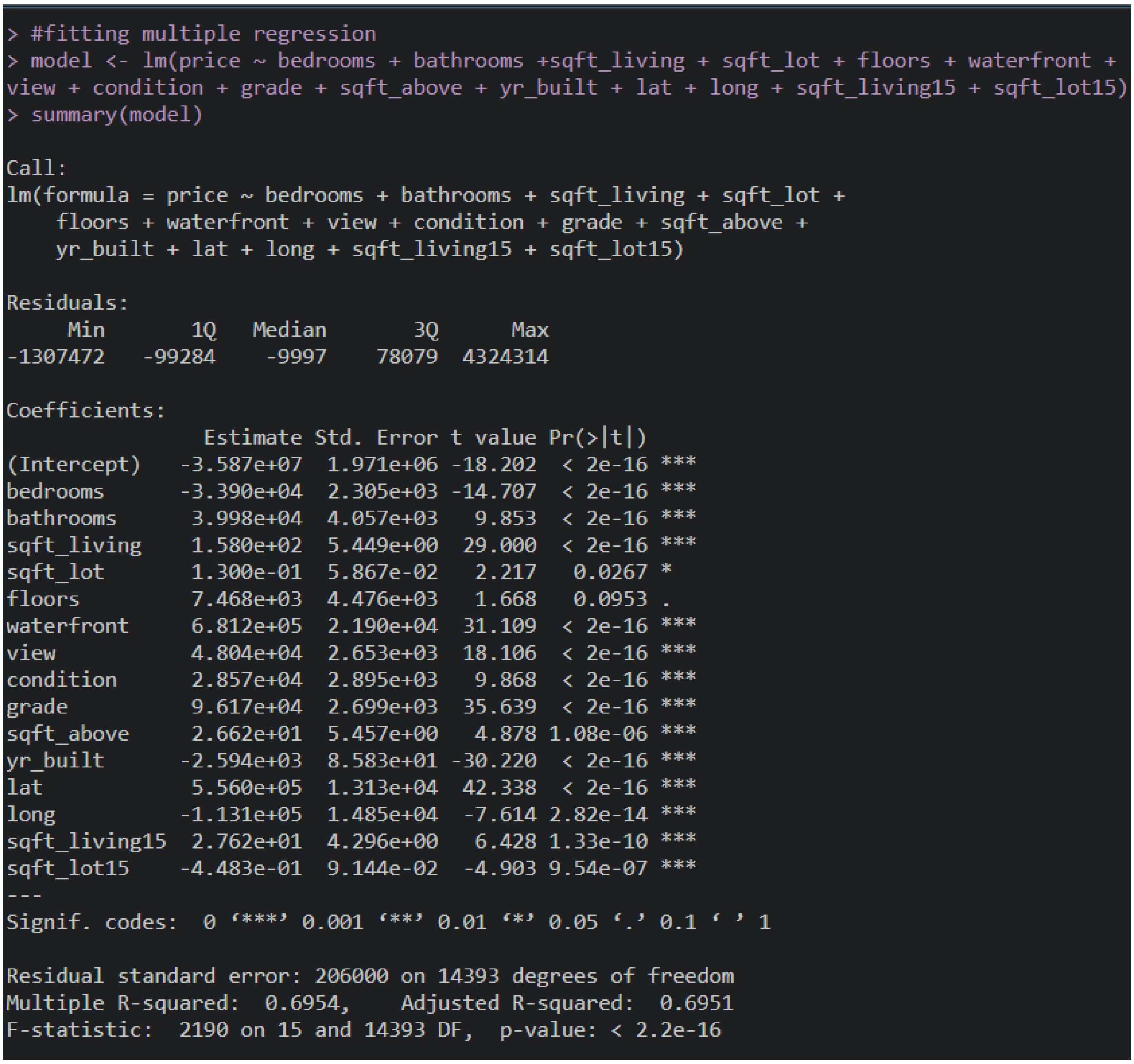
After we visualize the data we can now do the regression analysis. The regression that we will used for our regression model is **multiple linear regression model**.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \varepsilon$$

- X variables β_1, β_2, ... , β_{11} in the test are fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, pH, sulphates, and alcohol
- Y Variable in the data is quality
- we also assume that ε is normally distributed, hence The value of $\Sigma\varepsilon = 0$ for Hypothesis testing and the setting of confidence limits

4.2. First Regression Model

4.2.1. Creating First Model



```
> #fitting multiple regression
> model <- lm(price ~ bedrooms + bathrooms +sqft_living + sqft_lot + floors + waterfront +
view + condition + grade + sqft_above + yr_built + lat + long + sqft_living15 + sqft_lot15)
> summary(model)

Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot +
    floors + waterfront + view + condition + grade + sqft_above +
    yr_built + lat + long + sqft_living15 + sqft_lot15)

Residuals:
     Min       1Q   Median       3Q      Max
-1307472   -99284    -9997    78079  4324314

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.587e+07  1.971e+06 -18.202  < 2e-16 ***
bedrooms      -3.390e+04  2.305e+03 -14.707  < 2e-16 ***
bathrooms      3.998e+04  4.057e+03   9.853  < 2e-16 ***
sqft_living    1.580e+02  5.449e+00  29.000  < 2e-16 ***
sqft_lot       1.300e-01  5.867e-02   2.217   0.0267 *
floors         7.468e+03  4.476e+03   1.668   0.0953 .
waterfront     6.812e+05  2.190e+04  31.109  < 2e-16 ***
view           4.804e+04  2.653e+03  18.106  < 2e-16 ***
condition      2.857e+04  2.895e+03   9.868  < 2e-16 ***
grade          9.617e+04  2.699e+03  35.639  < 2e-16 ***
sqft_above     2.662e+01  5.457e+00   4.878 1.08e-06 ***
yr_built      -2.594e+03  8.583e+01 -30.220  < 2e-16 ***
lat            5.560e+05  1.313e+04  42.338  < 2e-16 ***
long          -1.131e+05  1.485e+04  -7.614 2.82e-14 ***
sqft_living15  2.762e+01  4.296e+00   6.428 1.33e-10 ***
sqft_lot15    -4.483e-01  9.144e-02  -4.903 9.54e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 206000 on 14393 degrees of freedom
Multiple R-squared:  0.6954,	Adjusted R-squared:  0.6951
F-statistic:  2190 on 15 and 14393 DF,  p-value: < 2.2e-16
```

4.2.2. First Model Analysis

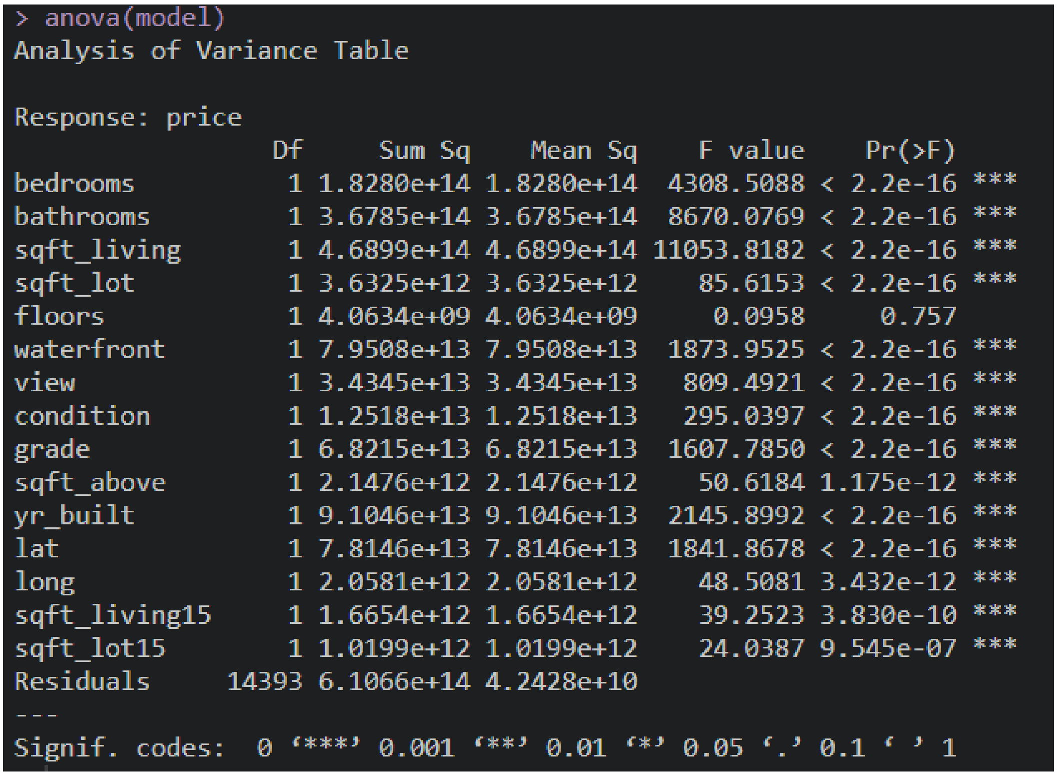
From the Summary above we know that the data :

- The model has R-squared of 0.6952 (in criteria of good model)
- Multiple Regression Equation for Y :

Y = -3.587e+07 -3.390e+04* bedrooms + 3.998e+04 * bathrooms + 1.580e+02 * sqft_living + 1.300e-01* sqft_lot + 7.468e+03* floors + 6.812e+05 * waterfront + 4.804e+04 * view - 2.857e+04 * condition + 9.617e+04 * grade + 2.662e+01 * sqft_above - 2.594e+03 * yr_built + 5.560e+05 * lat - 1.131e+05 * long + 2.762e+01 * sqft_living15 - 4.483e- 01 * sqft_lot15

As we can see that the equation above isn’t effective because there are many variable to calculate but we can perform a p-test to reduce the dimension of equation. This method can be done by creating an anova table for the model

4.2.3. ANOVA analysis – First Model



```
> anova(model)
Analysis of Variance Table

Response: price
                 Df     Sum Sq    Mean Sq    F value    Pr(>F)    
bedrooms          1 1.8280e+14 1.8280e+14  4308.5088 < 2.2e-16 ***
bathrooms         1 3.6785e+14 3.6785e+14  8670.0769 < 2.2e-16 ***
sqft_living       1 4.6899e+14 4.6899e+14 11053.8182 < 2.2e-16 ***
sqft_lot          1 3.6325e+12 3.6325e+12    85.6153 < 2.2e-16 ***
floors            1 4.0634e+09 4.0634e+09     0.0958     0.757    
waterfront        1 7.9508e+13 7.9508e+13  1873.9525 < 2.2e-16 ***
view              1 3.4345e+13 3.4345e+13   809.4921 < 2.2e-16 ***
condition         1 1.2518e+13 1.2518e+13   295.0397 < 2.2e-16 ***
grade             1 6.8215e+13 6.8215e+13  1607.7850 < 2.2e-16 ***
sqft_above        1 2.1476e+12 2.1476e+12    50.6184 1.175e-12 ***
yr_built          1 9.1046e+13 9.1046e+13  2145.8992 < 2.2e-16 ***
lat               1 7.8146e+13 7.8146e+13  1841.8678 < 2.2e-16 ***
long              1 2.0581e+12 2.0581e+12    48.5081 3.432e-12 ***
sqft_living15     1 1.6654e+12 1.6654e+12    39.2523 3.830e-10 ***
sqft_lot15        1 1.0199e+12 1.0199e+12    24.0387 9.545e-07 ***
Residuals     14393 6.1066e+14 4.2428e+10                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

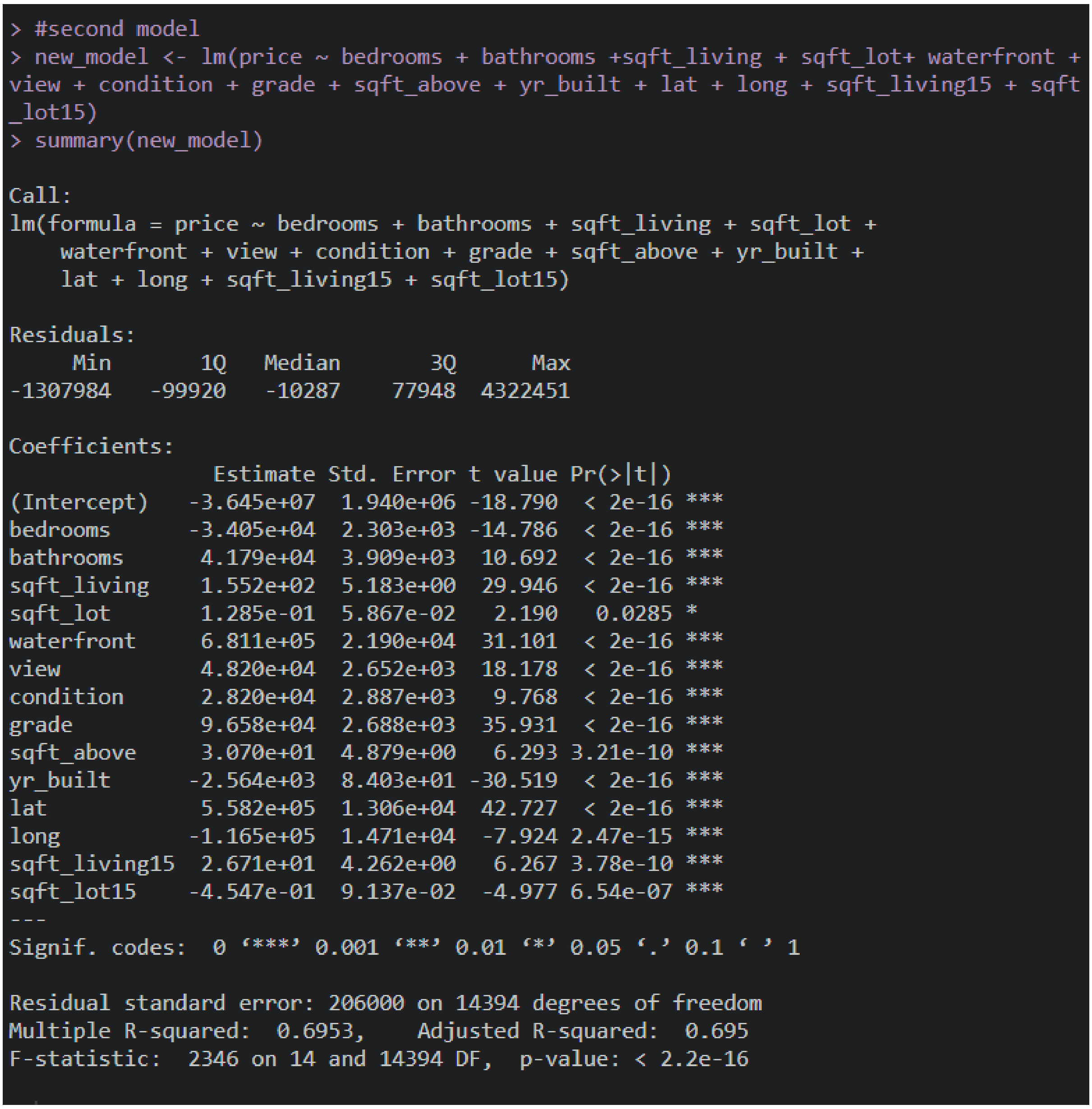
We can dot he p-test using the ANOVA table. For this test, the significant level are set to α = 0.01. The variable that are good for predicting must be satisfied the value P < 0.01. From the data above the value that are satisfied the requirement will notified as “**” (minimum) in the R analysis.

Therefore we can remove “floors” variable for us to maket he second model

4.3. Second Regression Model

4.3.1. Creating Second Model

Creating the second model will be the same as the first one but we exclude variable “floors”.



```
> #second model
> new_model <- lm(price ~ bedrooms + bathrooms +sqft_living + sqft_lot+ waterfront +
view + condition + grade + sqft_above + yr_built + lat + long + sqft_living15 + sqft
_lot15)
> summary(new_model)

Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot +
    waterfront + view + condition + grade + sqft_above + yr_built +
    lat + long + sqft_living15 + sqft_lot15)

Residuals:
     Min       1Q   Median       3Q      Max
-1307984   -99920   -10287    77948  4322451

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.645e+07  1.940e+06 -18.790  < 2e-16 ***
bedrooms      -3.405e+04  2.303e+03 -14.786  < 2e-16 ***
bathrooms      4.179e+04  3.909e+03  10.692  < 2e-16 ***
sqft_living    1.552e+02  5.183e+00  29.946  < 2e-16 ***
sqft_lot       1.285e-01  5.867e-02   2.190   0.0285 *
waterfront     6.811e+05  2.190e+04  31.101  < 2e-16 ***
view           4.820e+04  2.652e+03  18.178  < 2e-16 ***
condition      2.820e+04  2.887e+03   9.768  < 2e-16 ***
grade          9.658e+04  2.688e+03  35.931  < 2e-16 ***
sqft_above     3.070e+01  4.879e+00   6.293 3.21e-10 ***
yr_built      -2.564e+03  8.403e+01 -30.519  < 2e-16 ***
lat            5.582e+05  1.306e+04  42.727  < 2e-16 ***
long          -1.165e+05  1.471e+04  -7.924 2.47e-15 ***
sqft_living15  2.671e+01  4.262e+00   6.267 3.78e-10 ***
sqft_lot15    -4.547e-01  9.137e-02  -4.977 6.54e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 206000 on 14394 degrees of freedom
Multiple R-squared:  0.6953,    Adjusted R-squared:  0.695
F-statistic:  2346 on 14 and 14394 DF,  p-value: < 2.2e-16
```

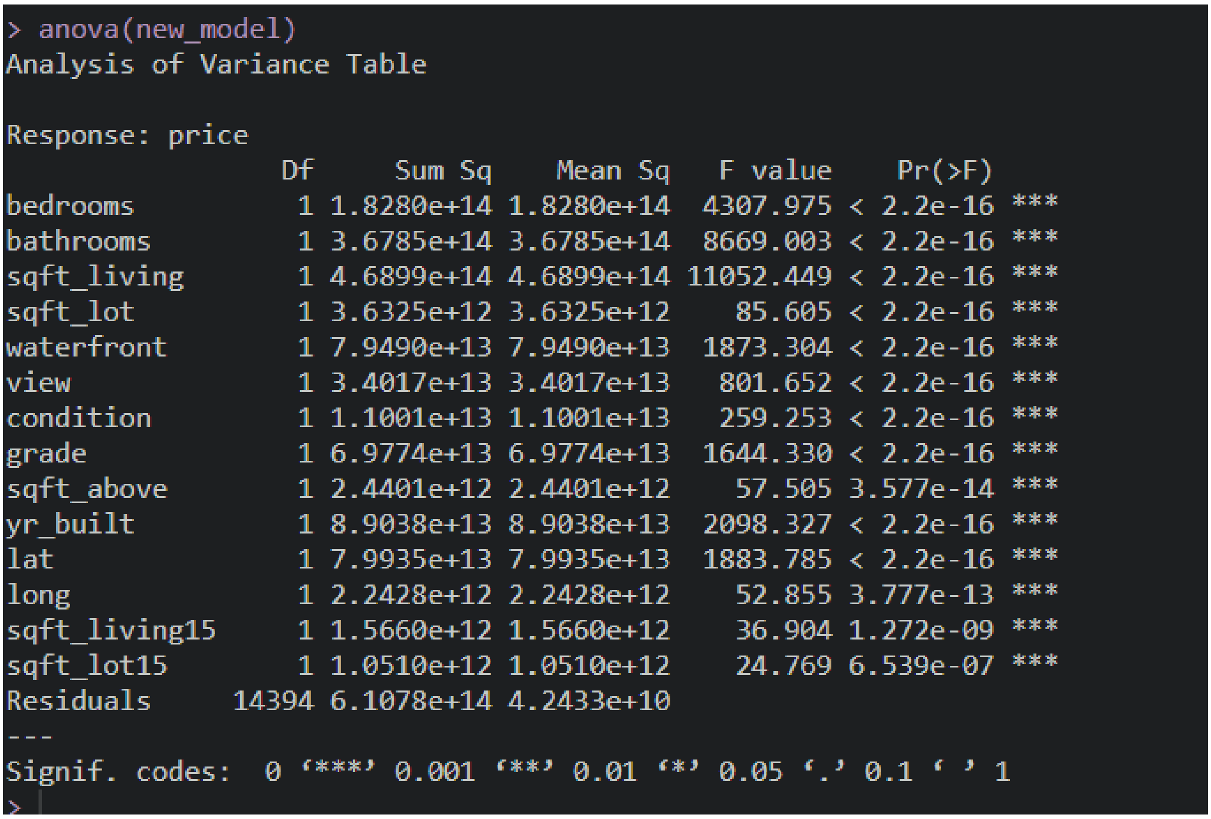
4.3.2. Second Model Analysis

From the Summary above we know that the data :

- The model has R-squared of 0.6952 which doesn’t differ much from the R squared in previous model
- Multiple Regression Equation for the new Y:

Y = -3.645e+07 - -3.405e+04* bedrooms + 4.179e+04* bathrooms + 1.552e+02 * sqft_living + 1.285e-01 * sqft_lot + 6.811e+05 * waterfront + 4.820e+04 * view - 2.820e+04 * condition + 9.658e+04 * grade + 3.070e+01 * sqft_above - 2.564e+03 * yr_built + 5.582e+05 * lat - 1.165e+05 * long + 2.671e+01 * sqft_living15 – 4.547e-01 * sqft_lot15

4.3.3. ANOVA analysis



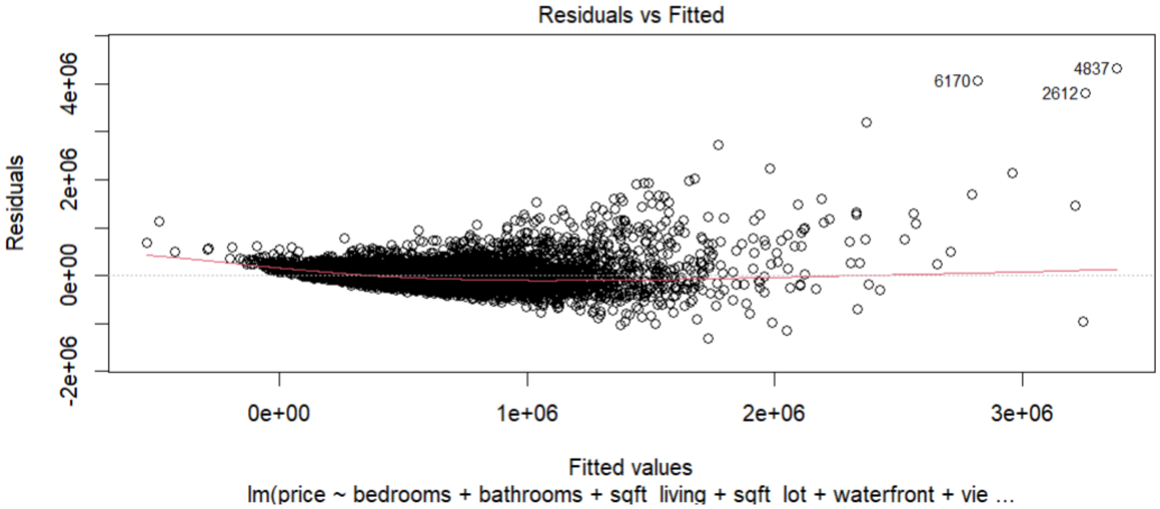
```
> anova(new_model)
Analysis of Variance Table

Response: price
                 Df     Sum Sq    Mean Sq   F value    Pr(>F)    
bedrooms          1 1.8280e+14 1.8280e+14  4307.975 < 2.2e-16 ***
bathrooms         1 3.6785e+14 3.6785e+14  8669.003 < 2.2e-16 ***
sqft_living       1 4.6899e+14 4.6899e+14 11052.449 < 2.2e-16 ***
sqft_lot          1 3.6325e+12 3.6325e+12    85.605 < 2.2e-16 ***
waterfront        1 7.9490e+13 7.9490e+13  1873.304 < 2.2e-16 ***
view              1 3.4017e+13 3.4017e+13   801.652 < 2.2e-16 ***
condition         1 1.1001e+13 1.1001e+13   259.253 < 2.2e-16 ***
grade             1 6.9774e+13 6.9774e+13  1644.330 < 2.2e-16 ***
sqft_above        1 2.4401e+12 2.4401e+12    57.505 3.577e-14 ***
yr_built          1 8.9038e+13 8.9038e+13  2098.327 < 2.2e-16 ***
lat               1 7.9935e+13 7.9935e+13  1883.785 < 2.2e-16 ***
long              1 2.2428e+12 2.2428e+12    52.855 3.777e-13 ***
sqft_living15     1 1.5660e+12 1.5660e+12    36.904 1.272e-09 ***
sqft_lot15        1 1.0510e+12 1.0510e+12    24.769 6.539e-07 ***
Residuals     14394 6.1078e+14 4.2433e+10                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

- All the Variable has “***” which mean that all variable are good for predicting the model because the P value < 0.01
- Hence, we know that all the variable used (P value) is significant for the model

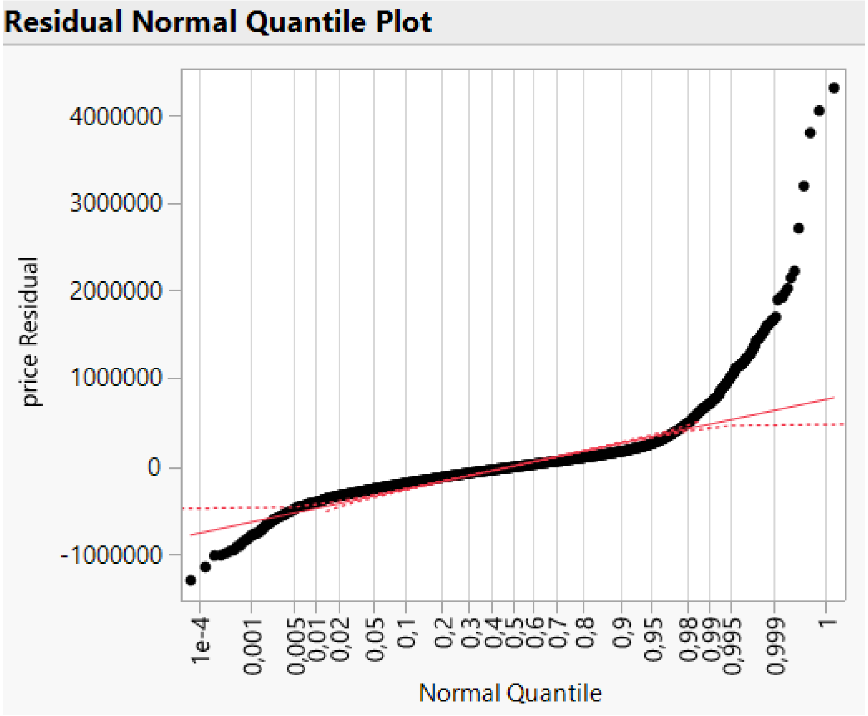
5. Residual Analysis

5.1. Residual Plot



- From the observation the residual points is expanding away from the line when moving towards the right. This mean there are a lot of outlier in the right tail

5.2. Q-Q Plot



- The right side of the data show that the point is going away from the line which mean that the distribution of the data is positively skew

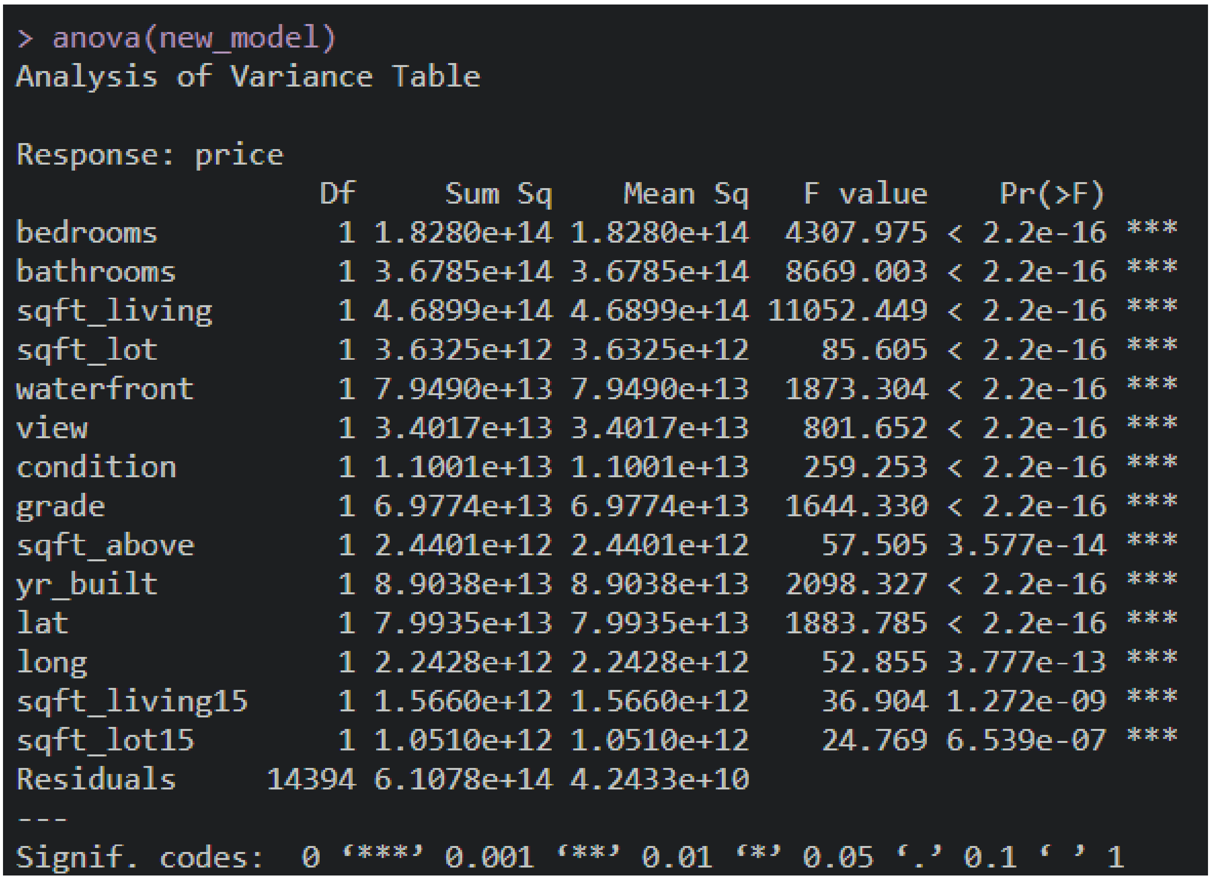
6. Hypothesis Testing

6.1. Null Hypothesis

$$H_0: \beta_0 = \beta_1 = \cdots = \beta_{k-1} = 0$$

$$H_1: \beta_j \neq 0, for\ atleast\ one\ j$$

6.2. Output ANOVA for Hypothesis



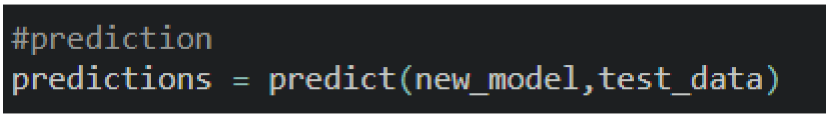
```
> anova(new_model)
Analysis of Variance Table

Response: price
                 Df     Sum Sq    Mean Sq   F value    Pr(>F)    
bedrooms          1 1.8280e+14 1.8280e+14  4307.975 < 2.2e-16 ***
bathrooms         1 3.6785e+14 3.6785e+14  8669.003 < 2.2e-16 ***
sqft_living       1 4.6899e+14 4.6899e+14 11052.449 < 2.2e-16 ***
sqft_lot          1 3.6325e+12 3.6325e+12    85.605 < 2.2e-16 ***
waterfront        1 7.9490e+13 7.9490e+13  1873.304 < 2.2e-16 ***
view              1 3.4017e+13 3.4017e+13   801.652 < 2.2e-16 ***
condition         1 1.1001e+13 1.1001e+13   259.253 < 2.2e-16 ***
grade             1 6.9774e+13 6.9774e+13  1644.330 < 2.2e-16 ***
sqft_above        1 2.4401e+12 2.4401e+12    57.505 3.577e-14 ***
yr_built          1 8.9038e+13 8.9038e+13  2098.327 < 2.2e-16 ***
lat               1 7.9935e+13 7.9935e+13  1883.785 < 2.2e-16 ***
long              1 2.2428e+12 2.2428e+12    52.855 3.777e-13 ***
sqft_living15     1 1.5660e+12 1.5660e+12    36.904 1.272e-09 ***
sqft_lot15        1 1.0510e+12 1.0510e+12    24.769 6.539e-07 ***
Residuals     14394 6.1078e+14 4.2433e+10                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As the above results show all the p values are significant. We can reject the NULL hypothesis. Therefore, “Model can be used for prediction”.

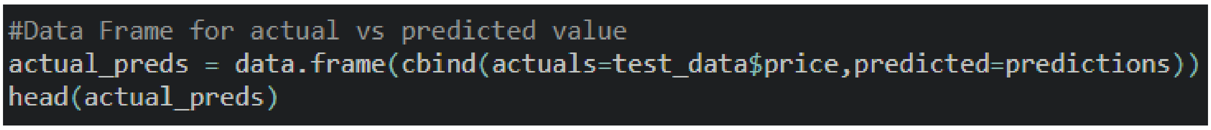
7. Prediction and Accuracy for Final Model

7.1. Prediction using test data



```
#prediction
predictions = predict(new_model,test_data)
```

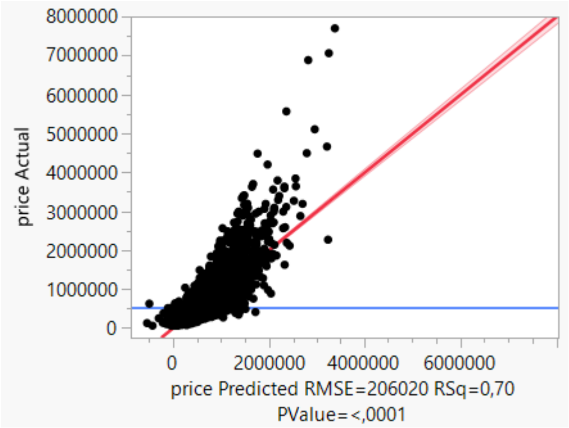
The function above predict the Y value from test_data (obtain from splitting data) using the new_model



```
#Data Frame for actual vs predicted value
actual_preds = data.frame(cbind(actuals=test_data$price,predicted=predictions))
head(actual_preds)
```

After we get the predicted value for test_data from the model then we compare it with the actual price value in the test_data and put it in a dataframe.table

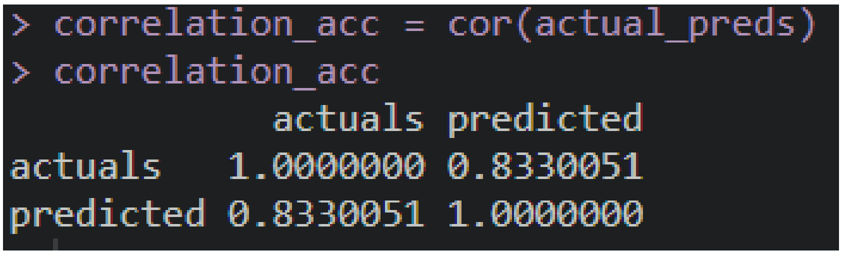
7.2. Actual vs Predicted



A lot of concentration in the bottom left but it started to spread out. Showing that the accuracy is good for most of the data but there are some outliers that the model have some inaccuracy

7.3. Model Accuracy

The code below show how to make a confusion matrix from correlation between the actual values and the predicted values that we already make a data frame



```
> correlation_acc = cor(actual_preds)
> correlation_acc
            actuals predicted
actuals   1.0000000 0.8330051
predicted 0.8330051 1.0000000
```

The Matrix shows that the accuracy of the model is **83,3%**

8. Conclusion

From all the analysis, the distribution of the housing price do not follow normal distribution. It has a high positive skewness and a handful right tail outliers in the data. Even though the data isn’t normally distributed but we can still predict the model. In conclusion the model can be used for prediction because it has the accuracy of 83,3%.

However, it's important to acknowledge the limitations of our study. While our regression model performs well within the parameters of our dataset, it may encounter challenges when applied to different geographical regions or time periods. Additionally, factors such as economic fluctuations, policy changes, and unforeseen events can impact housing markets in unpredictable ways, which may affect the model's accuracy over time. Also there maybe some inaccuracy in analyzing the data.

9. Reference

CC0: Public Domain. (2024) *Housing Price Dataset.*

https://www.kaggle.com/datasets/sukhmandeepsinghbrar/housing-price-dataset/data

Qualitrics. *Interpreting Residual Plots to Improve Your Regression*

https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

JMP. *User Community. https://community.jmp.com/t5/Learn-JMP/ct-p/learn-jmp*

R is use to make the model for multivariate linear regression

R and SAS JMP are used to create plot, tables, and graph