ST362 Project

Project #: 3

Members: Herteg Kohar, Laith Adi

Student #: 190605160, 170256190

Introduction

The dataset is of housing sales in the King County area located in Washington state. Disregarding the id column there are 19 other features that can be used as predictor variables in order to predict the target variable of price. The problems we wish to solve are discovering which of the given 19 variables are adequate predictors to estimate the sale price of a house within King County given features that resemble the ones given in the dataset. After finding the suitable predictors to estimate the sale price of a house the model which we build will also follow the four assumptions of linear regression which are:

1. The relationship between the chosen predictors and price is linear
2. The residuals of the model are uncorrelated
3. The residuals of the model have a constant variance
4. The residuals of the model are normally distributed

Some other research questions we wish to answer are:

1. Can a logarithmic transformation be used to help satisfy linear regression assumptions?
2. What are the relationships between price and the predictor variables?
3. Is an interactive term between ***sqft_living*** and ***sqft_above*** be effective?

Data Description:

In total there are 21,613 observations within the dataset as well as 21 total columns. A five-point summary was done for each column and the data type and meaning of each column were derived from the dataset.

The first step in building the model was preliminary data analysis. This included analysis of the distributions of each given column in the dataset. Some notable discoveries were uncovering variables that may not provide any use to giving a prediction for the housing prices in King County. For example, the id column of the dataset was just a unique identifier for each house sale and would not have contributed to predicting the house sale price.

Since part of the data was geographical, a graph was made which plotted the latitude and longitude of each sale and colour scaled each point accordingly to the price. Since the scale of pricing was very large and right-skewed (based on the distribution plot) it was decided that a logarithmic transformation would be appropriate to attempt to see less skew in the price. From the plot of the longitude and latitude, there is a concentrated area where the colour is lighter (meaning a larger price based on the colour scale). It can also be seen that there are some outliers where the sales of houses may not be a part of King County, specifically the little cluster of houses eastward in the plot. This simple analysis and plotting gave some more insight into the given data.

Model 1 - Variable/Model Selection

The next step was to check which predictor variables correlated with price and which predictors were correlated with each other. This was done by producing a correlation matrix for the dataset. The threshold used for the correlation analysis was 0.8 to deem two predictors highly correlated predictors and at least a correlation of 0.3 with price for considering predictors to add to the model. From the correlation matrix, it is clear that sqft_living and sqft_above are highly linearly correlated with each other. In this case, it makes sense to only utilize one of the predictors as using both of the predictors may not have an effect in explaining the variance in the data because of the extremely high correlation between the two predictor variables. Moving forward for the next step was to create a baseline model with the chosen variables from the correlated features excluding sqft_above because spft_living had a higher correlation with price. The results of the baseline model were not all too impressive as it resulted in an adjusted R^2 of 0.64 and extremely high values in the AIC and BIC criterion. The upside of choosing the variables based on correlation before was a Durbin-Watson statistic which was very close to 2 meaning almost no auto-correlation in the data. To eliminate variables that don’t seem to add to the predictive power of the model, analysis of plots was done for each predictor chosen in the baseline model. From the plots and further correlation analysis, sqft_living15 was also highly correlated with sqft_living so sqft_living15 was removed from the model. Another reason sqft_living15 was removed was because of the test of significance for the predictor as its p-value = 0.393 which is greater than $\boldsymbol{\alpha}$ = 0.05, so in this case, the null hypothesis of its coefficient being zero cannot be rejected under significance $\boldsymbol{\alpha}$ = 0.05. Bedrooms didn’t seem to have a linear relationship with the price given in the added variable and partial residual plots there didn’t seem to be a linear relationship. Bathrooms did seem to have a linear relationship with the price however, when paired with sqft_living, view, lat and grade there was a minimal change in the adjusted R^2 meaning the amount of variance in price it explains when added to the model is minimal.

As mentioned above in the plot of the longitudinal and latitudinal points of sale the price variable is right-skewed. Applying a log transformation on the price would make the target variable more normalized. This same thing can be said about sqft_living and a log transformation would benefit sqft_living making it more normalized. On top of this residual plot vs. the fitted value and the Q-Q plot of the model with the specified logarithmic transformation appears to present the residuals as normally distributed. A random pattern can be seen in the first plot indicating a constant variance for the residuals. In the Q-Q plot, there seems to be a linear pattern meaning the data may be approximately normally distributed. Utilizing the logarithmic transformation helped satisfy the last three linear regression assumptions regarding the distribution of the error terms.

Model 2 - Variable/Model Selection

For model 2, four key considerations went into model selection. (1) Adjusted R^2, Cp, and BIC. (2) Correlation between all parameters. (3) Statistically significant parameters. (4) Logarithmic transformation. After performing an exhaustive search, the resulting adjusted R^2, Cp, and BIC are 18, 18, and 16, respectively. Mallow's Cp is 18, meaning that the ideal model would need to have about 18 parameters (Cp ≈ number of parameters, including β_0). After taking a look at the Cp graph, out of the 19 original parameters, parameter yr_built is the one that is recommended not to be included in the model. The BIC represents the penalty term for the number of parameters in a model, and in this case, 18 happens to be the smallest penalty term out of all models examined during the exhaustive search. Similar to the adjusted R^2, the selected model contains the highest adjusted R^2 value ensuring that the most information is being covered by the model.

The second consideration was the correlation between all parameters. Taking a look at the correlation matrix, two parameters stand out as highly correlated. As mentioned above, the parameters sqft_above and sqft_living have a correlation coefficient of 0.88. Since this correlation coefficient is higher than the threshold placed of 0.80, it is worth further investigating. An interaction term was worth a try because one, it is fairly easy and quick to test the theory in r, and two, the variables are both dependent on each other. For more detail on the interaction term, refer to this research question.

Third consideration, the statistically insignificant parameters, as they were removed from the model. This may have decreased the adjusted R^2 but the little increase in the adjusted R^2

did not justify having a more complex model. The sqft_basement parameter p-value is 0.586461, clearly greater than 0.05. Therefore, the sqft_basement parameter was removed from the model.

The fourth consideration was the possibility of applying a logarithmic transformation to the parameters. Through trial and error, transforming the sqft_living parameter gave the best results. With that being said, the logarithmic transformation was applied to the target variable in order to satisfy the four assumptions of linear regression. For more detail, refer to this research question.

Analysis of Variance

Taking a look at both of the ANOVA tables for Model 1 and Model 2, Model 2 has a lower sum squares residual value than Model 1 exactly a difference of 188. According to the p-values from each of the ANOVA tables of each of the respective models, predictors are not equal to 0 under significance level $\alpha = 0.05$. The mean square residuals of each model are the same. In Model 2 there are 13 more parameters making the complexity comparison between Model 1 and 2 extremely large. However, adding this complexity did result in a lower residual sum of squares and there seem to be significant predictors added seeing that the p-value < $\alpha = 0.05$.

Can a Logarithmic Transformation be Used to Help Satisfy Linear Regression Assumptions?

From the results in the variable transformation of Model 1, it can be seen that a logarithmic transformation to deter skewness is very useful in helping satisfy some of the assumptions in linear regression regarding the error terms and greatly helps the predictive power of the model. This improved the visual plots of the baseline model as it was clear there was a pattern in the residual plots and not a linear pattern in the Q-Q plot. This contrasts very differently from the baseline model residual plots as there doesn't appear to be a pattern in the residuals vs. fitted values plot and there was a linear pattern in the Q-Q plot of the baseline model 1. The same point can be made for baseline model 2 as there were similar features in both residual plots for each of the respective baseline models and very drastic improvements in both transformed models 1 and 2 seen in their residual plots.

What are the relationships between price and the predictor variables?

From the various added-variable and partial residual plots, the general relationship between price and the given predictor variables can be accessed. By observing linear relationships in the added-variable and partial residual plots it exemplifies the linearity between the predictors and price. The correlation matrix also gives an indication of linear relationships as the correlation value gets closer to one when looking at the price column one can identify highly correlated variables with the target variable of price. Lastly, when plotting price against other variables the polynomial relationship between the predictor and price will emerge. Overall, the relationships between price and predictor variables are observed through visual and mathematical analysis including plots and hypothesis tests as well as correlation matrices.

Is an Interactive Term Between ***sqft_living*** and ***sqft_above*** be Effective?

Earlier, the high correlation between sqft_above and sqft_living was mentioned and it was pointed out that an interaction term is worth investigating. The question is how effective will an interactive term between the two parameters be and would it make sense. A regression model with the interactive term yields a similar adjusted R^2 value around 0.74 as a model that does not include the interactive term. Also, when observing the interactive plot, there is no intersection in the lines indicating that there will be no interaction effect. With no significant change in the model adjusted R^2 and the interactive plot showing no signs of an interaction effect, the interaction term will not be included in the model.

Conclusion

To conclude, the recommended model would be model 2. When observing the summary of model 1 and model 2, not only does the adjusted R^2 for model 2 happen to be greater than model 2, but the residual standard error is also lower. Some recommendations would be to include more observations meaning to broaden the area of research to expand outside of King County. Another recommendation would be to include less predictors which are highly correlated with each other and some information not specifically about the home but surrounding area other than zip code and longitude and latitude.

APPENDIX 1

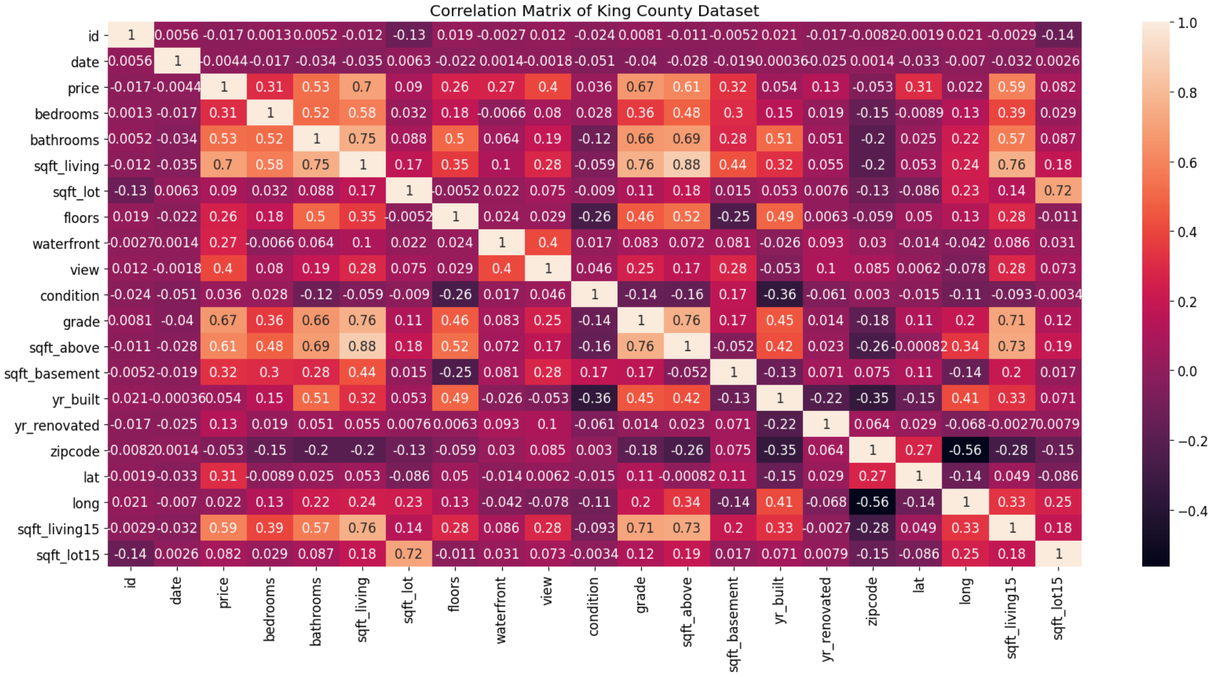
Herteg worked on building model 1. Laith Adi worked on building model 2. Both wrote the whole paper together.

Laith R Code.

Herteg R Code.

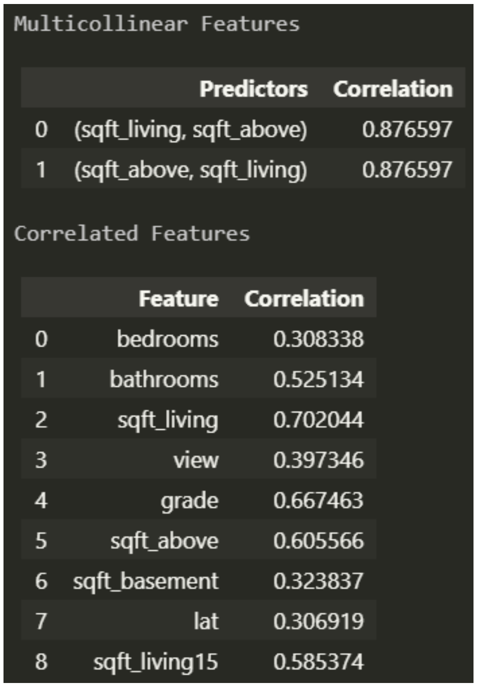
APPENDIX 2

Correlation Matrix of The King County Dataset



	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
id	1	0.0056	-0.017	0.0013	0.0052	-0.012	-0.13	0.019	-0.0027	0.012	-0.024	0.0081	-0.011	-0.0052	0.021	-0.017	-0.0082	-0.0019	0.021	-0.0029	-0.14
date	0.0056	1	-0.0044	-0.017	-0.034	-0.035	0.0063	-0.022	0.0014	-0.0018	-0.051	-0.04	-0.028	-0.019	-0.00036	-0.025	0.0014	-0.033	-0.007	-0.032	0.0026
price	-0.017	-0.0044	1	0.31	0.53	0.7	0.09	0.26	0.27	0.4	0.036	0.67	0.61	0.32	0.054	0.13	-0.053	0.31	0.022	0.59	0.082
bedrooms	0.0013	-0.017	0.31	1	0.52	0.58	0.032	0.18	-0.0066	0.08	0.028	0.36	0.48	0.3	0.15	0.019	-0.15	-0.0089	0.13	0.39	0.029
bathrooms	0.0052	-0.034	0.53	0.52	1	0.75	0.088	0.5	0.064	0.19	-0.12	0.66	0.69	0.28	0.51	0.051	-0.2	0.025	0.22	0.57	0.087
sqft_living	-0.012	-0.035	0.7	0.58	0.75	1	0.17	0.35	0.1	0.28	-0.059	0.76	0.88	0.44	0.32	0.055	-0.2	0.053	0.24	0.76	0.18
sqft_lot	-0.13	0.0063	0.09	0.032	0.088	0.17	1	-0.0052	0.022	0.075	-0.009	0.11	0.18	0.015	0.053	0.0076	-0.13	-0.086	0.23	0.14	0.72
floors	0.019	-0.022	0.26	0.18	0.5	0.35	-0.0052	1	0.024	0.029	-0.26	0.46	0.52	-0.25	0.49	0.0063	-0.059	0.05	0.13	0.28	-0.011
waterfront	-0.0027	0.0014	0.27	-0.0066	0.064	0.1	0.022	0.024	1	0.4	0.017	0.083	0.072	0.081	-0.026	0.093	0.03	-0.014	-0.042	0.086	0.031
view	0.012	-0.0018	0.4	0.08	0.19	0.28	0.075	0.029	0.4	1	0.046	0.25	0.17	0.28	-0.053	0.1	0.085	0.0062	-0.078	0.28	0.073
condition	-0.024	-0.051	0.036	0.028	-0.12	-0.059	-0.009	-0.26	0.017	0.046	1	-0.14	-0.16	0.17	-0.36	-0.061	0.003	-0.015	-0.11	-0.093	-0.0034
grade	0.0081	-0.04	0.67	0.36	0.66	0.76	0.11	0.46	0.083	0.25	-0.14	1	0.76	0.17	0.45	0.014	-0.18	0.11	0.2	0.71	0.12
sqft_above	-0.011	-0.028	0.61	0.48	0.69	0.88	0.18	0.52	0.072	0.17	-0.16	0.76	1	-0.052	0.42	0.023	-0.26	-0.00082	0.34	0.73	0.19
sqft_basement	-0.0052	-0.019	0.32	0.3	0.28	0.44	0.015	-0.25	0.081	0.28	0.17	0.17	-0.052	1	-0.13	0.071	0.075	0.11	-0.14	0.2	0.017
yr_built	0.021	-0.00036	0.054	0.15	0.51	0.32	0.053	0.49	-0.026	-0.053	-0.36	0.45	0.42	-0.13	1	-0.22	-0.35	-0.15	0.41	0.33	0.071
yr_renovated	-0.017	-0.025	0.13	0.019	0.051	0.055	0.0076	0.0063	0.093	0.1	-0.061	0.014	0.023	0.071	-0.22	1	0.064	0.029	-0.068	-0.0027	0.0079
zipcode	-0.0082	0.0014	-0.053	-0.15	-0.2	-0.2	-0.13	-0.059	0.03	0.085	0.003	-0.18	-0.26	0.075	-0.35	0.064	1	0.27	-0.56	-0.28	-0.15
lat	-0.0019	-0.033	0.31	-0.0089	0.025	0.053	-0.086	0.05	-0.014	0.0062	-0.015	0.11	-0.00082	0.11	-0.15	0.029	0.27	1	-0.14	0.049	-0.086
long	0.021	-0.007	0.022	0.13	0.22	0.24	0.23	0.13	-0.042	-0.078	-0.11	0.2	0.34	-0.14	0.41	-0.068	-0.56	-0.14	1	0.33	0.25
sqft_living15	-0.0029	-0.032	0.59	0.39	0.57	0.76	0.14	0.28	0.086	0.28	-0.093	0.71	0.73	0.2	0.33	-0.0027	-0.28	0.049	0.33	1	0.18
sqft_lot15	-0.14	0.0026	0.082	0.029	0.087	0.18	0.72	-0.011	0.031	0.073	-0.0034	0.12	0.19	0.017	0.071	0.0079	-0.15	-0.086	0.25	0.18	1

Multicollinear and Correlated Features to Price in the King County Dataset



```
Multicollinear Features

                  Predictors  Correlation
0  (sqft_living, sqft_above)     0.876597
1  (sqft_above, sqft_living)     0.876597

Correlated Features

         Feature  Correlation
0       bedrooms     0.308338
1      bathrooms     0.525134
2    sqft_living     0.702044
3           view     0.397346
4          grade     0.667463
5     sqft_above     0.605566
6  sqft_basement     0.323837
7            lat     0.306919
8  sqft_living15     0.585374
```

Data Descriptions of King County Dataset

Column	Description	Preliminary Analysis
id	The unique identifier of each sale. (Integer)	Minimum 1000102 Q1 2123049194.0 Median 3904930410.0 Q3 7308900445.0 Maximum 9900000190

date	The date of the observation being recorded. (Integer)	Minimum 735355 Q1 735436.0 Median 735522.0 Q3 735646.0 Maximum 735745
price	The sale price of the given house. (Decimal)	Minimum 75000.0 Q1 321950.0 Median 450000.0 Q3 645000.0 Maximum 7700000.0
bedrooms	The number of bedrooms the home has. (Integer)	Minimum 0 Q1 3.0 Median 3.0 Q3 4.0 Maximum 33
bathrooms	The number of bathrooms the home has. (Decimal)	Minimum 0.0 Q1 1.75 Median 2.25 Q3 2.5 Maximum 8.0
sqft_living	The amount of square	Minimum 290

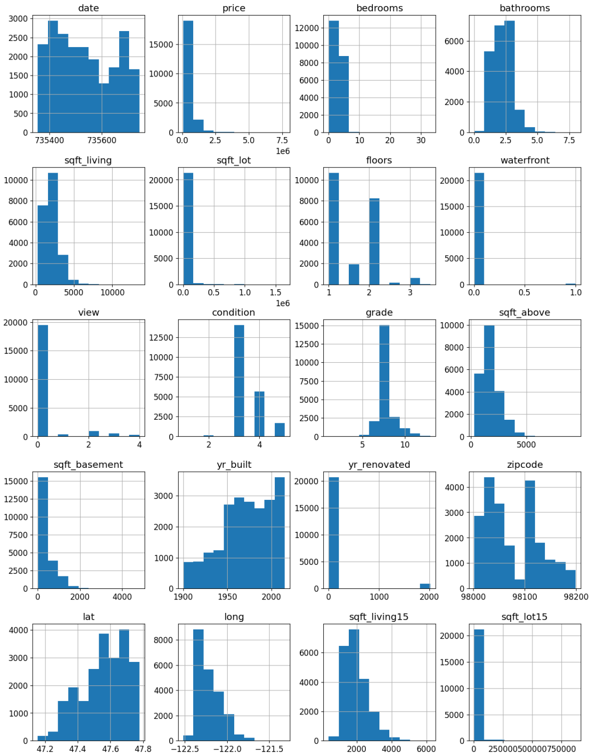
	footage of living space the home has (Integer)	Q1 1427.0 Median 1910.0 Q3 2550.0 Maximum 13540
sqft_lot	The square footage of the total lot the home resides on (Integer)	Minimum 520 Q1 5040.0 Median 7618.0 Q3 10688.0 Maximum 1651359
floors	The total number of storeys the home has. (Decimal)	Minimum 1.0 Q1 1.0 Median 1.5 Q3 2.0 Maximum 3.5
waterfront	A binary variable for whether the home has a waterfront view or not (Integer 0 or 1)	Minimum 0 Q1 0.0 Median 0.0 Q3 0.0 Maximum 1
view	A rating on the view of which the home has (Integer 0-4)	Minimum 0 Q1 0.0

		Median 0.0 Q3 0.0 Maximum 4
condition	A grade on the condition of the home (Integer 1-5)	Minimum 1 Q1 3.0 Median 3.0 Q3 4.0 Maximum 5
grade	A grade on the build of the house (Integer)	Minimum 1 Q1 7.0 Median 7.0 Q3 8.0 Maximum 13
sqft_above	Square footage of the above living space. (Integer)	Minimum 290 Q1 1190.0 Median 1560.0 Q3 2210.0 Maximum 9410
sqft_basement	The square footage of the home’s basement (Integer)	Minimum 0 Q1 0.0

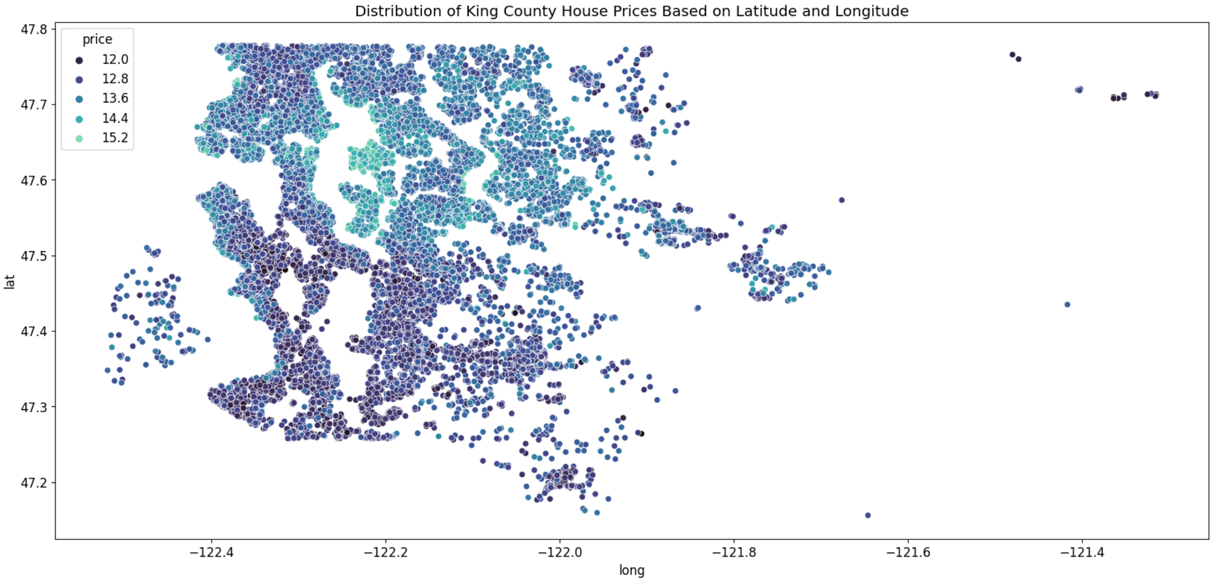
		Median 0.0 Q3 560.0 Maximum 4820
yr_built	The year in which the home was built (Integer)	Minimum 1900 Q1 1951.0 Median 1975.0 Q3 1997.0 Maximum 2015
yr_renovated	If the home has had renovations a year will appear otherwise if there is no renovation 0 will be inputted instead	Minimum 0 Q1 0.0 Median 0.0 Q3 0.0 Maximum 2015
zipcode	The zip code of the given home (Integer)	Minimum 98001 Q1 98033.0 Median 98065.0 Q3 98118.0 Maximum 98199
lat	The latitudinal coordinate of the home's location (Decimal)	Minimum 47.1559 Q1 47.471 Median 47.5718

		Q3 47.678 Maximum 47.7776
long	The longitudinal coordinate of the home’s location (Decimal)	Minimum -122.519 Q1 -122.328 Median -122.23 Q3 -122.125 Maximum -121.315
sqft_living15	The square footage of the house (Integer)	Minimum 399 Q1 1490.0 Median 1840.0 Q3 2360.0 Maximum 6210
sqft_lot15	The square footage of the lot (Integer)	Minimum 651 Q1 5100.0 Median 7620.0 Q3 10083.0 Maximum 871200

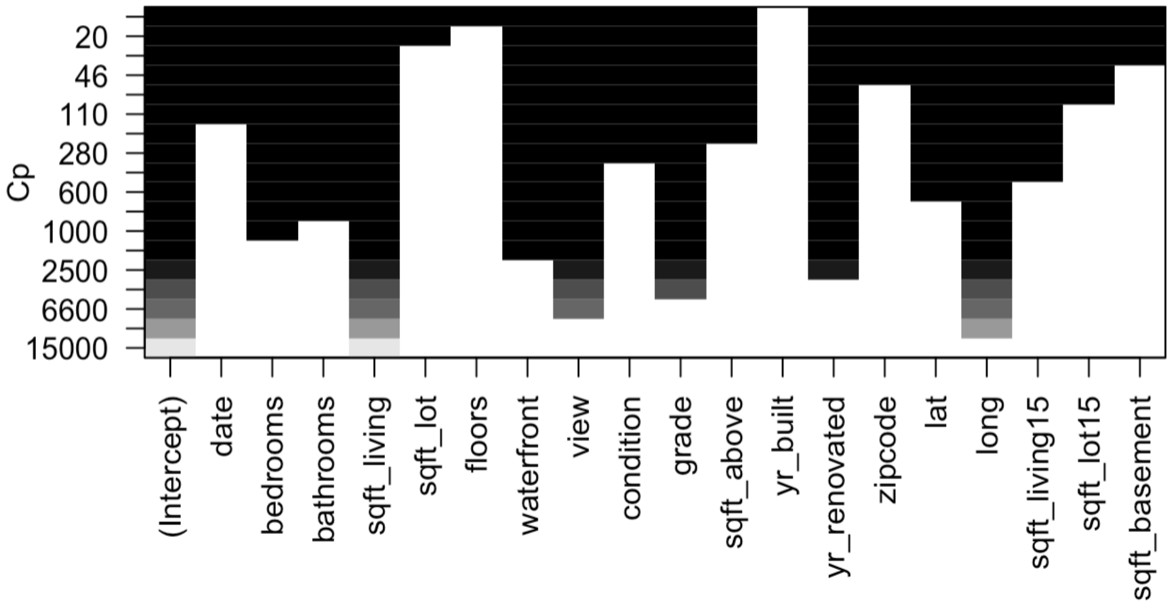
Distribution Plot of Given Columns in King County Dataset



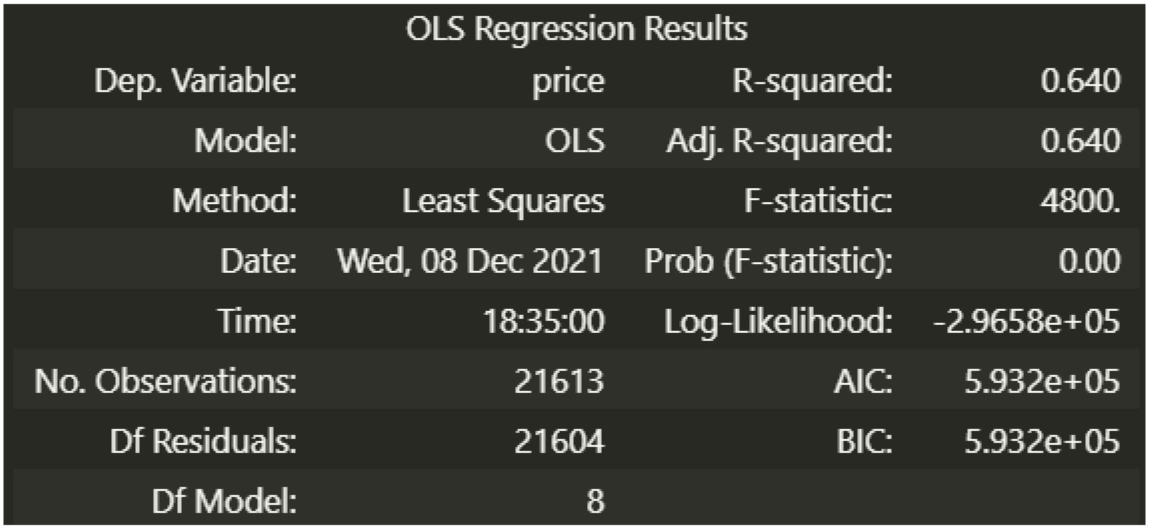
Distribution of House Prices Based on Latitude and Longitude



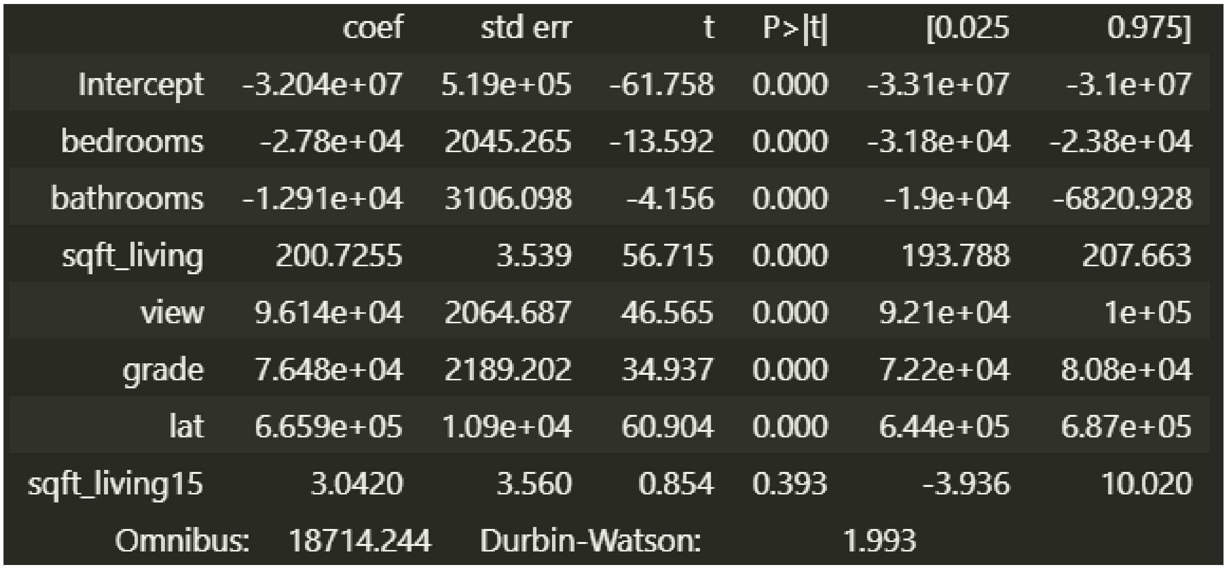
Model 1 - Mallow’s Cp Graph



Model 1 - Baseline Model Results

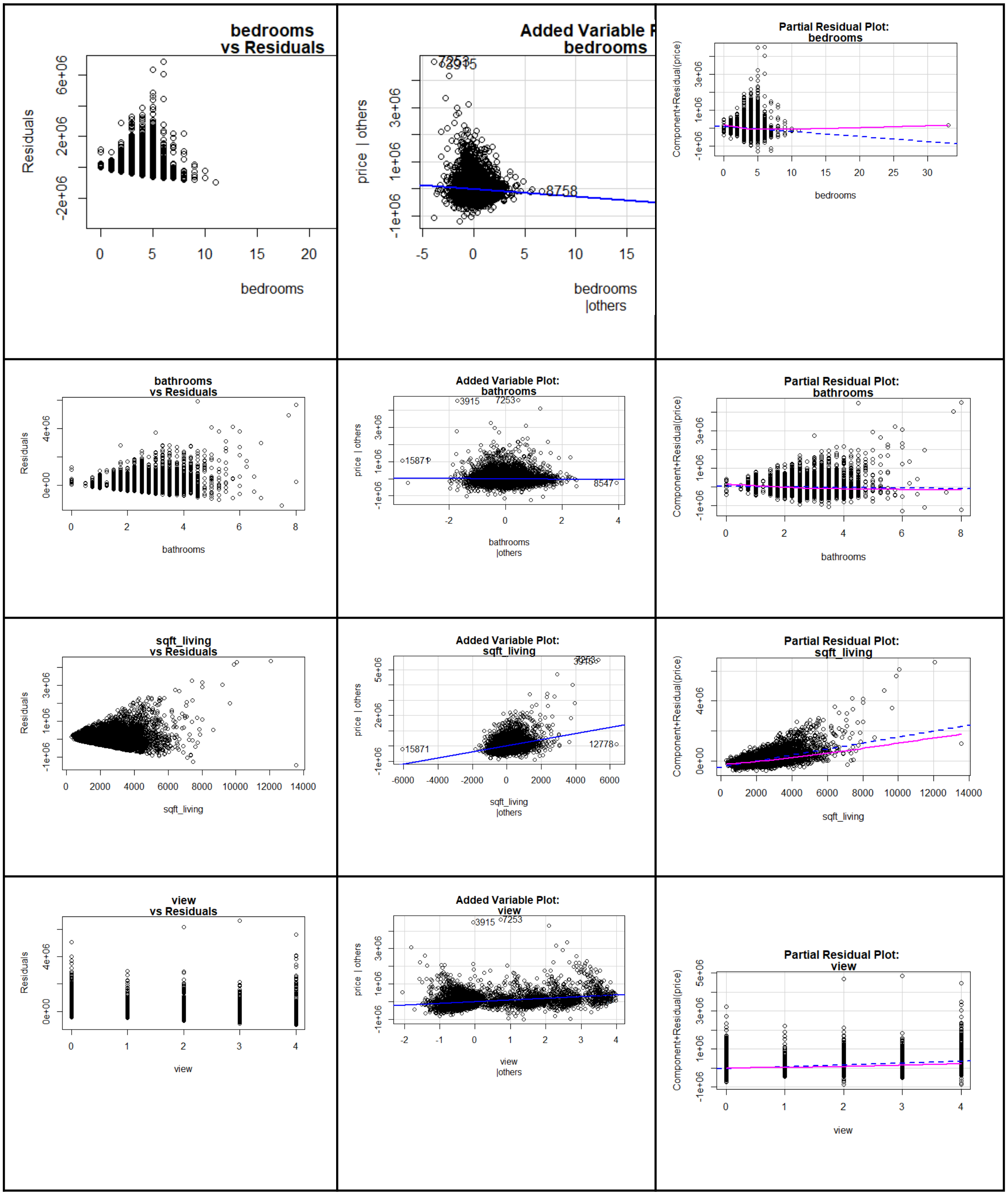


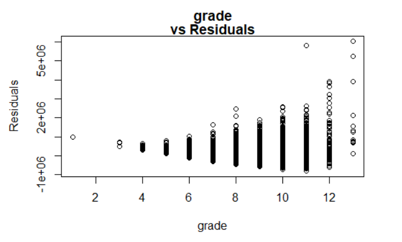
```
                         OLS Regression Results
   Dep. Variable:              price           R-squared:        0.640
           Model:                OLS      Adj. R-squared:        0.640
          Method:      Least Squares         F-statistic:        4800.
            Date:   Wed, 08 Dec 2021  Prob (F-statistic):         0.00
            Time:           18:35:00      Log-Likelihood:  -2.9658e+05
No. Observations:              21613                 AIC:    5.932e+05
    Df Residuals:              21604                 BIC:    5.932e+05
        Df Model:                  8
```

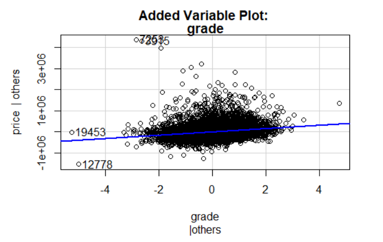


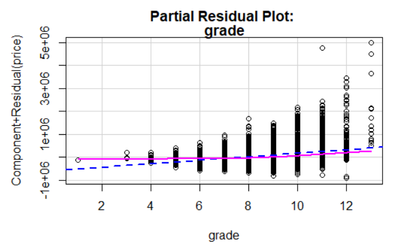
```
                      coef     std err         t    P>|t|      [0.025      0.975]
    Intercept   -3.204e+07    5.19e+05   -61.758    0.000   -3.31e+07     -3.1e+07
     bedrooms    -2.78e+04    2045.265   -13.592    0.000   -3.18e+04    -2.38e+04
    bathrooms   -1.291e+04    3106.098    -4.156    0.000    -1.9e+04    -6820.928
  sqft_living     200.7255       3.539    56.715    0.000     193.788      207.663
         view    9.614e+04    2064.687    46.565    0.000    9.21e+04        1e+05
        grade    7.648e+04    2189.202    34.937    0.000    7.22e+04     8.08e+04
          lat    6.659e+05    1.09e+04    60.904    0.000    6.44e+05     6.87e+05
sqft_living15       3.0420       3.560     0.854    0.393      -3.936       10.020
        Omnibus:   18714.244     Durbin-Watson:          1.993
```

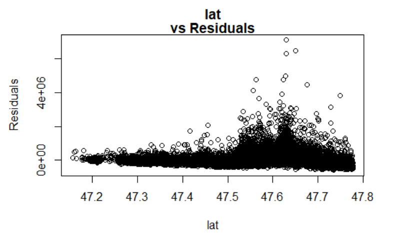
Analysis of Baseline Model Predictors Model 1

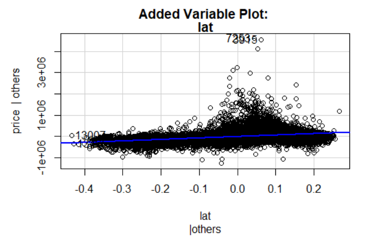


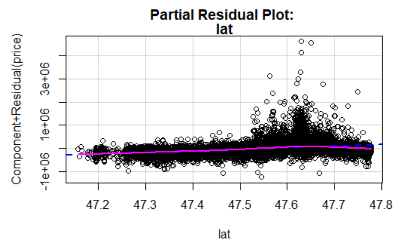

grade
vs Residuals

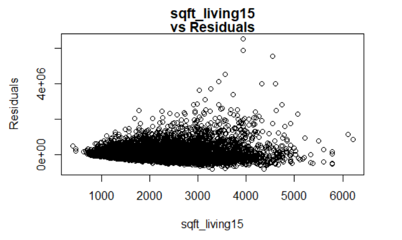

Added Variable Plot:
grade

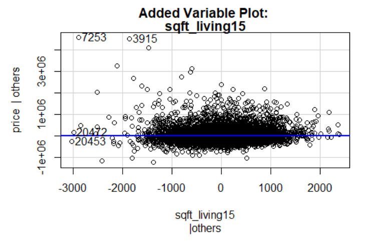

Partial Residual Plot:
grade

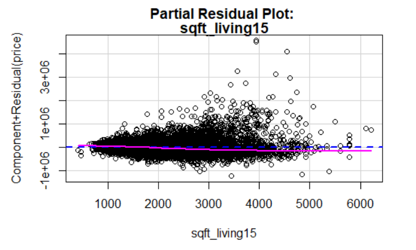

lat
vs Residuals


Added Variable Plot:
lat

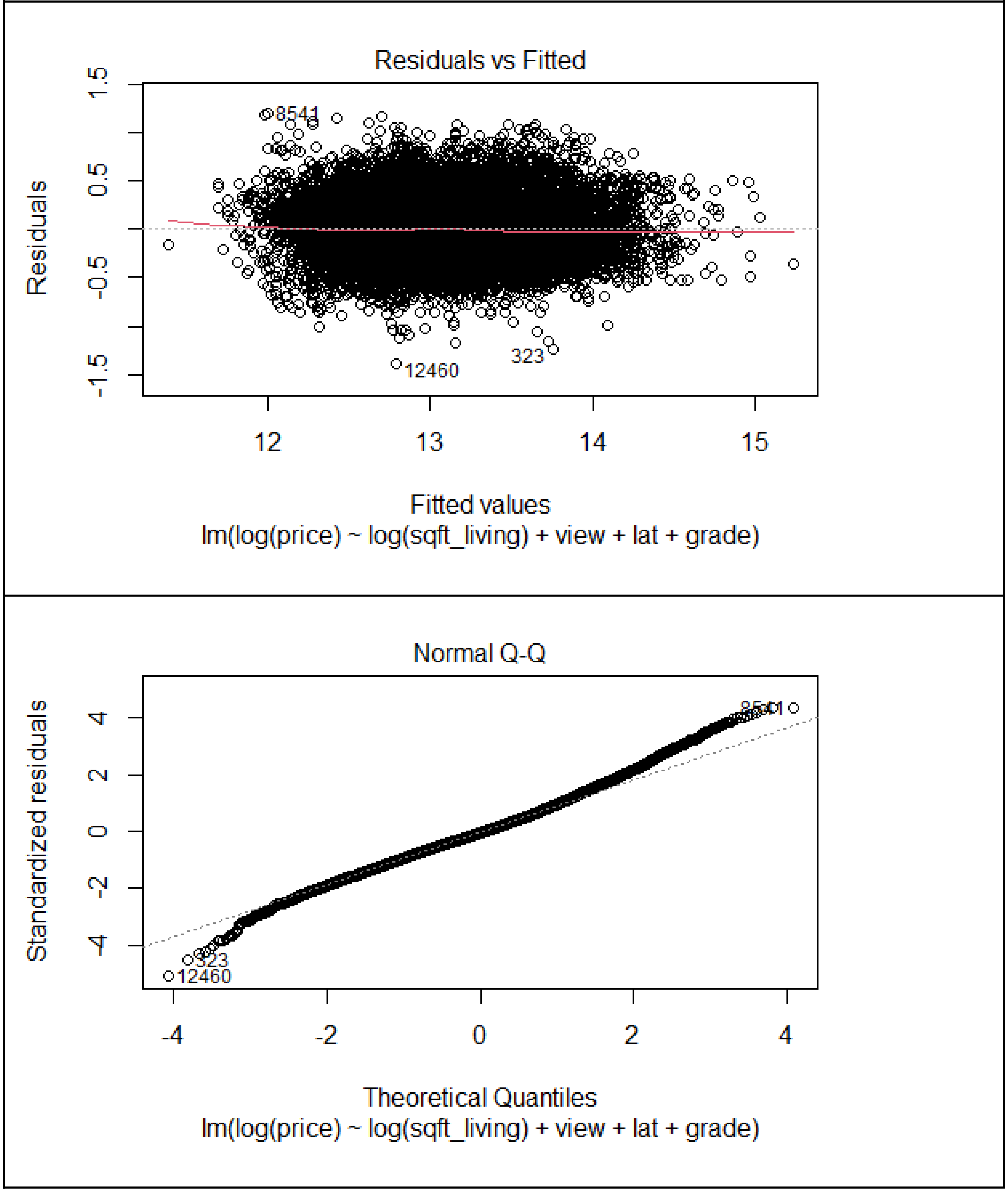

Partial Residual Plot:
lat


sqft_living15
vs Residuals

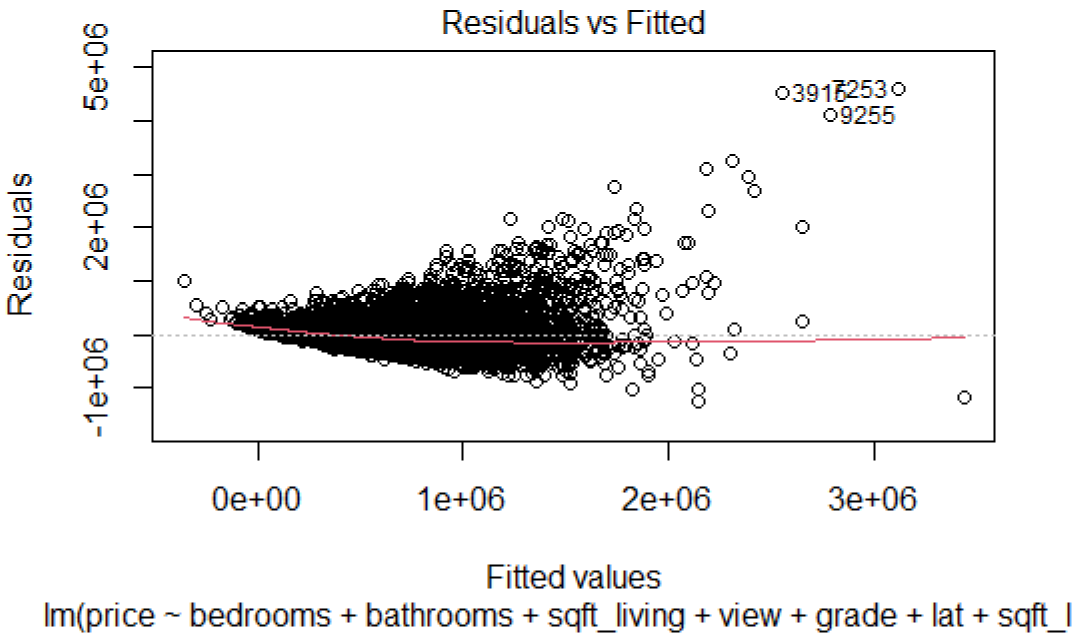

Added Variable Plot:
sqft_living15

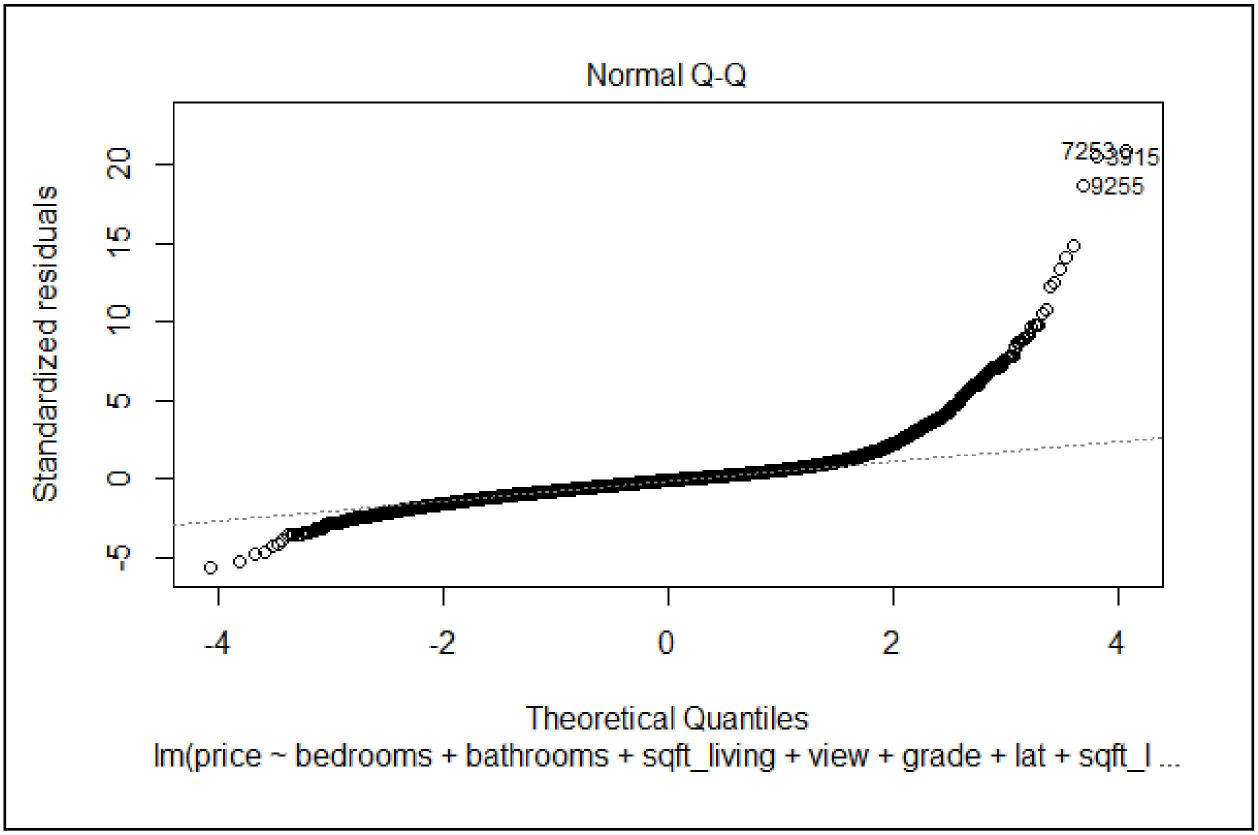

Partial Residual Plot:
sqft_living15

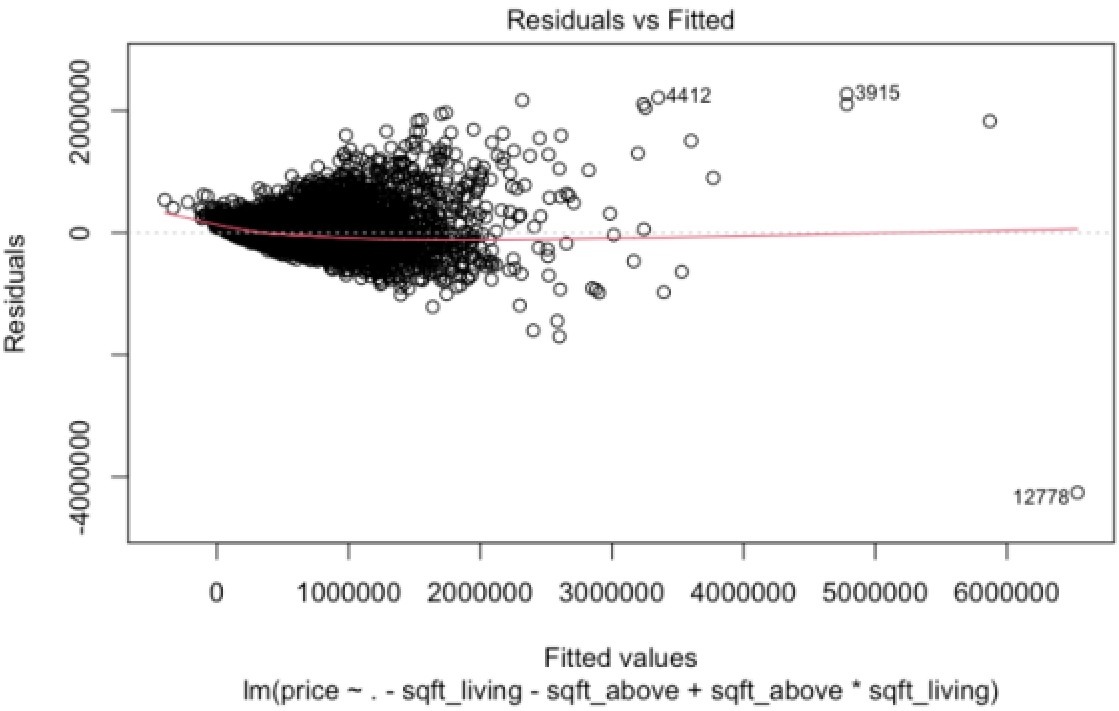
Residuals vs Fitted and Q-Q Plot for Transformed Model 1

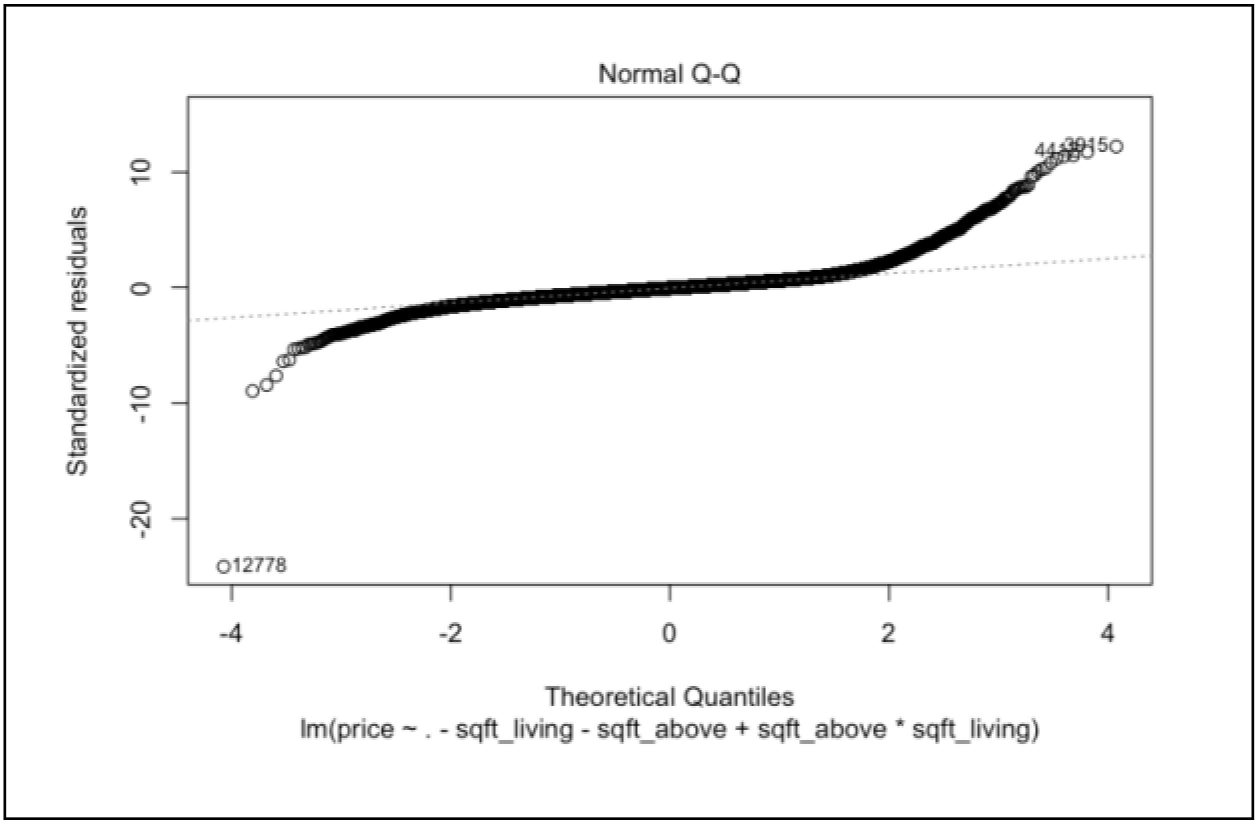


Residuals vs Fitted and Q-Q Plot for Baseline Model 1

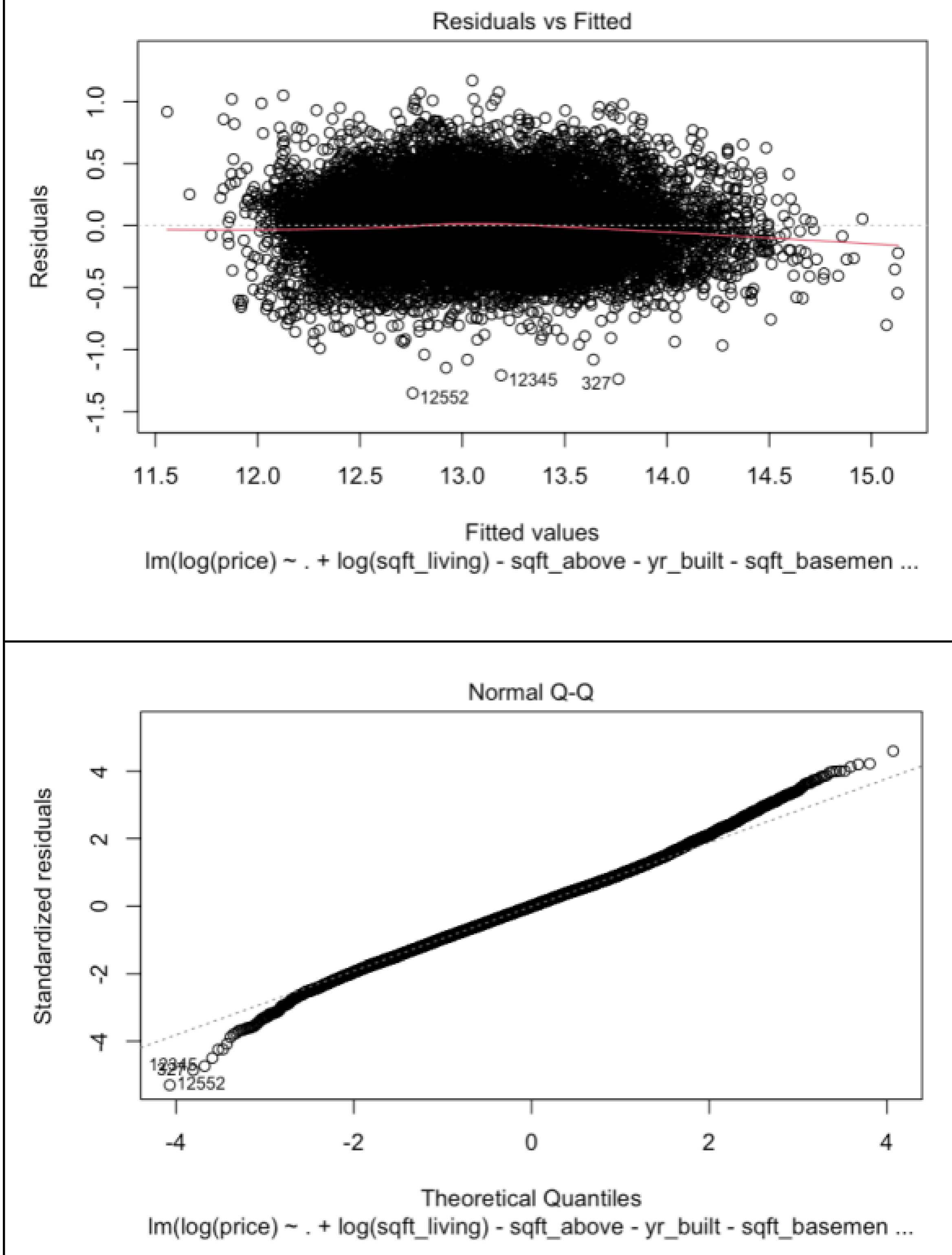



Normal Q-Q
Standardized residuals
-5
0
5
10
15
20
-4
-2
0
2
4
Theoretical Quantiles
lm(price ~ bedrooms + bathrooms + sqft_living + view + grade + lat + sqft_l ...
7253
3915
9255

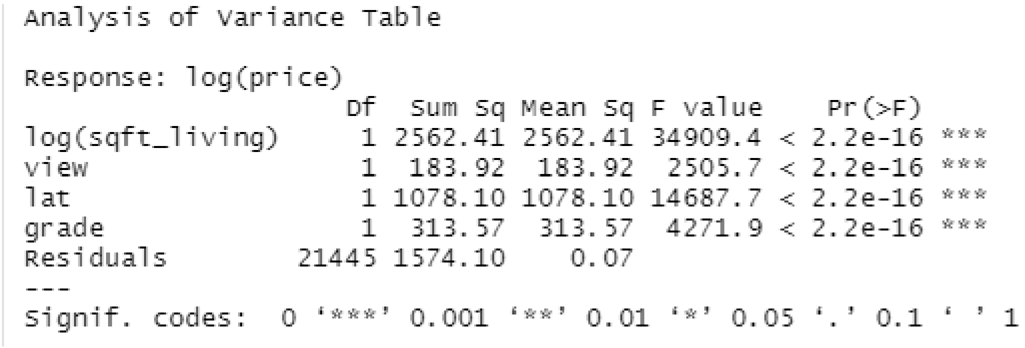
Residuals vs. Fitted Values and Q-Q plot for Baseline Model 2



Normal Q-Q
Standardized residuals
10
0
-10
-20
3915
44
12778
-4
-2
0
2
4
Theoretical Quantiles
lm(price ~ . - sqft_living - sqft_above + sqft_above * sqft_living)

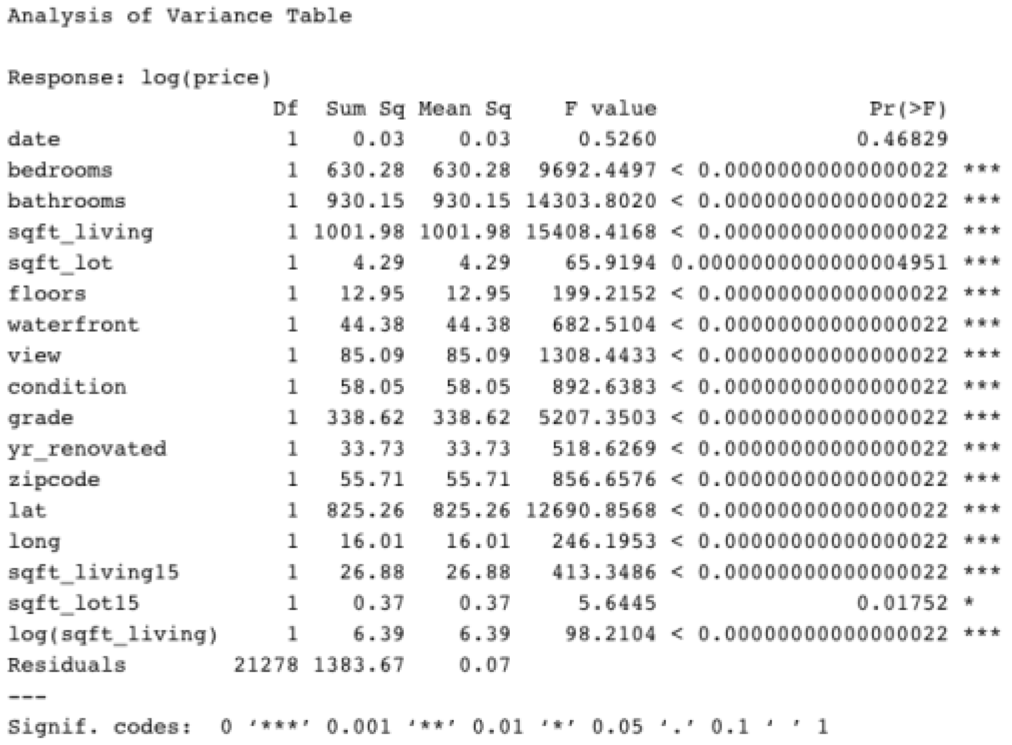
Residuals vs Fitted and Q-Q Plot for Transformed Model 2



ANOVA Tables for Model 1 and Model 2

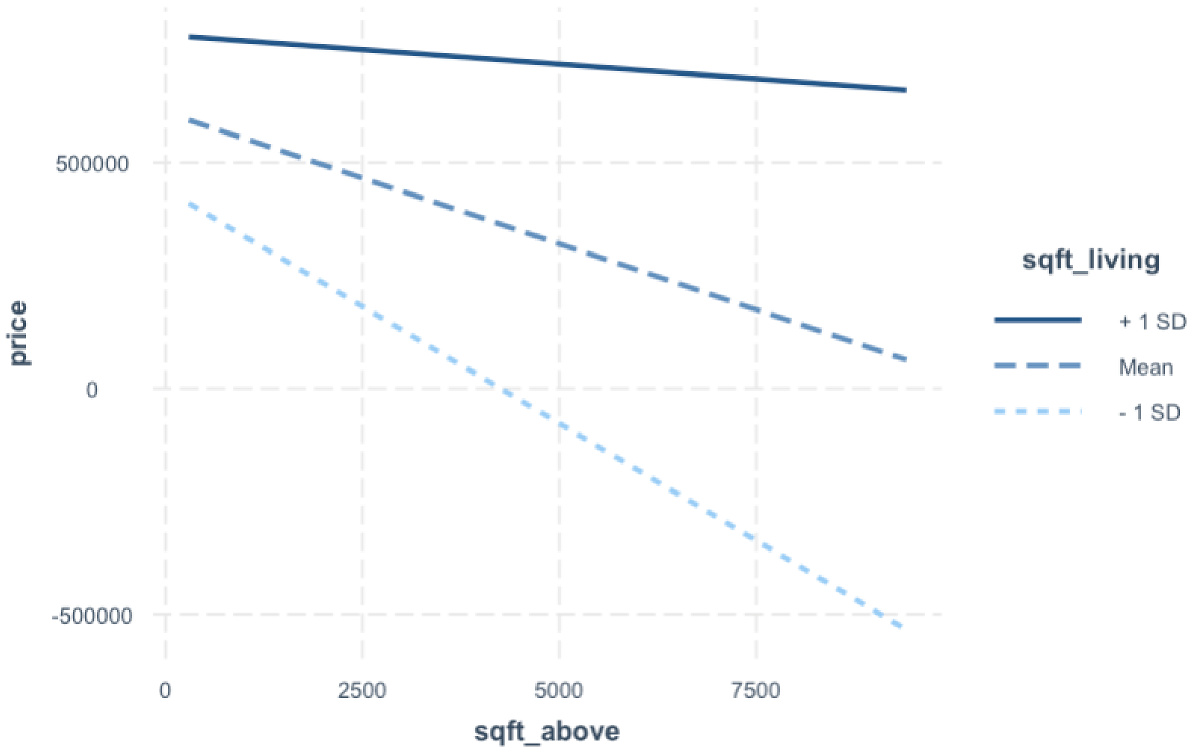


```
Analysis of Variance Table

Response: log(price)
                    Df  Sum Sq Mean Sq F value    Pr(>F)    
log(sqft_living)     1 2562.41 2562.41 34909.4 < 2.2e-16 ***
view                 1  183.92  183.92  2505.7 < 2.2e-16 ***
lat                  1 1078.10 1078.10 14687.7 < 2.2e-16 ***
grade                1  313.57  313.57  4271.9 < 2.2e-16 ***
Residuals        21445 1574.10    0.07                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

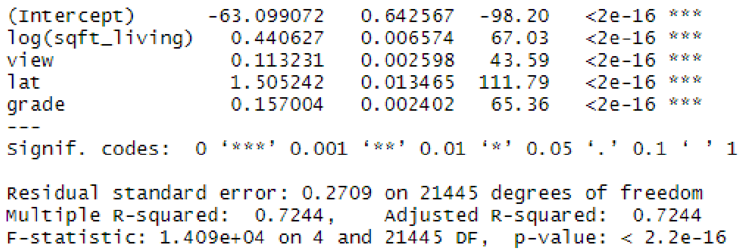


```
Analysis of Variance Table

Response: log(price)
                    Df  Sum Sq Mean Sq    F value               Pr(>F)    
date                 1    0.03    0.03     0.5260              0.46829    
bedrooms             1  630.28  630.28  9692.4497 < 0.00000000000000022 ***
bathrooms            1  930.15  930.15 14303.8020 < 0.00000000000000022 ***
sqft_living          1 1001.98 1001.98 15408.4168 < 0.00000000000000022 ***
sqft_lot             1    4.29    4.29    65.9194 0.0000000000000004951 ***
floors               1   12.95   12.95   199.2152 < 0.00000000000000022 ***
waterfront           1   44.38   44.38   682.5104 < 0.00000000000000022 ***
view                 1   85.09   85.09  1308.4433 < 0.00000000000000022 ***
condition            1   58.05   58.05   892.6383 < 0.00000000000000022 ***
grade                1  338.62  338.62  5207.3503 < 0.00000000000000022 ***
yr_renovated         1   33.73   33.73   518.6269 < 0.00000000000000022 ***
zipcode              1   55.71   55.71   856.6576 < 0.00000000000000022 ***
lat                  1  825.26  825.26 12690.8568 < 0.00000000000000022 ***
long                 1   16.01   16.01   246.1953 < 0.00000000000000022 ***
sqft_living15        1   26.88   26.88   413.3486 < 0.00000000000000022 ***
sqft_lot15           1    0.37    0.37     5.6445              0.01752 *  
log(sqft_living)     1    6.39    6.39    98.2104 < 0.00000000000000022 ***
Residuals        21278 1383.67    0.07                                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interactive Plot between sqft_living and sqft_above

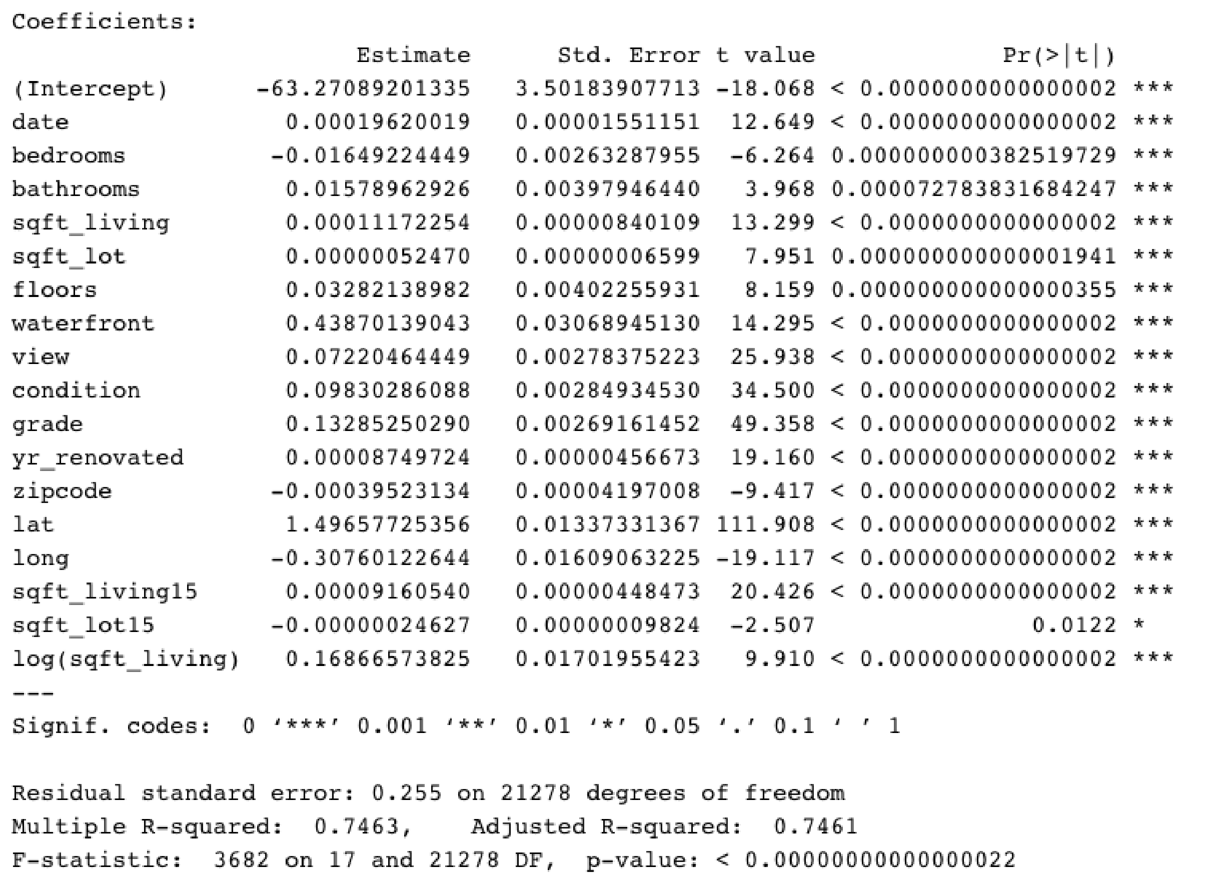


Summary of Model 1



```
(Intercept)      -63.099072   0.642567  -98.20   <2e-16 ***
log(sqft_living)   0.440627   0.006574   67.03   <2e-16 ***
view               0.113231   0.002598   43.59   <2e-16 ***
lat                1.505242   0.013465  111.79   <2e-16 ***
grade              0.157004   0.002402   65.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2709 on 21445 degrees of freedom
Multiple R-squared:  0.7244,	Adjusted R-squared:  0.7244 
F-statistic: 1.409e+04 on 4 and 21445 DF,  p-value: < 2.2e-16
```

Summary of Model 2



```
Coefficients:
                        Estimate      Std. Error t value             Pr(>|t|)    
(Intercept)      -63.27089201335   3.50183907713 -18.068 < 0.0000000000000002 ***
date               0.00019620019   0.00001551151  12.649 < 0.0000000000000002 ***
bedrooms          -0.01649224449   0.00263287955  -6.264 0.000000000382519729 ***
bathrooms          0.01578962926   0.00397946440   3.968 0.000072783831684247 ***
sqft_living        0.00011172254   0.00000840109  13.299 < 0.0000000000000002 ***
sqft_lot           0.00000052470   0.00000006599   7.951 0.000000000000001941 ***
floors             0.03282138982   0.00402255931   8.159 0.000000000000000355 ***
waterfront         0.43870139043   0.03068945130  14.295 < 0.0000000000000002 ***
view               0.07220464449   0.00278375223  25.938 < 0.0000000000000002 ***
condition          0.09830286088   0.00284934530  34.500 < 0.0000000000000002 ***
grade              0.13285250290   0.00269161452  49.358 < 0.0000000000000002 ***
yr_renovated       0.00008749724   0.00000456673  19.160 < 0.0000000000000002 ***
zipcode           -0.00039523134   0.00004197008  -9.417 < 0.0000000000000002 ***
lat                1.49657725356   0.01337331367 111.908 < 0.0000000000000002 ***
long              -0.30760122644   0.01609063225 -19.117 < 0.0000000000000002 ***
sqft_living15      0.00009160540   0.00000448473  20.426 < 0.0000000000000002 ***
sqft_lot15        -0.00000024627   0.00000009824  -2.507               0.0122 *  
log(sqft_living)   0.16866573825   0.01701955423   9.910 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.255 on 21278 degrees of freedom
Multiple R-squared:  0.7463,    Adjusted R-squared:  0.7461 
F-statistic:  3682 on 17 and 21278 DF,  p-value: < 0.00000000000000022
```

APPENDIX - LAITH R CODE

ATTACHED IN DROP BOX.

APPENDIX - HERTEG R CODE

ATTACHED IN DROP BOX.