STAT 6021 Final Project Report

Predicting House Prices in King County, USA

Jiangxue Han (jh6rg)

Shaoran Li (sl4bz)

Jingyi Luo (jl6zh)

Mengyao Zhang (mz6jv)

Wenxi Zhao (wz8nx)

December 6th, 2018

1 Executive Summary

This report serves the purpose of conducting analysis on features which drive housing price in King County of Washington State in United States and searching for the best statistical model to interpret the relationship among features and housing price, as well as making predictions of housing price in future. The dataset used in this study is obtained from web-based online data repository Kaggle. Both discrete and continuous variables will be addressed and massaged for better model development. Statistical techniques such as stepwise-type procedures, various forms of transformations and influential point analysis will be exploited for model development and model adequacy validation. Multiple graphs and statistical output for methods selection are attached in the Appendix section to support the analysis of this report.

2 Introduction

Housing price has been a trendy topic for machine learning and regression techniques for recent years. People might not necessarily associate an ideal home with features such as the square footage of basement or the latitude of the house; nevertheless, we will explore these features in this study and inspect their relevance with housing price closely.

2.1 Data Description

The dataset used in this study is obtained from the Kaggle data repository and it contains house sale prices for King County of Washington State in United States, including Seattle. This dataset collects housing prices from King County between May 2014 and May 2015, and the data entries are observational. There are 20 features including both quantitative and qualitative variables. The target response variable for prediction is housing price. Refer to (Appendix-Table.1) for a detailed description of these variables.

2.2 Objectives

The objective of the study is to indicate the most influential parameters that drive the housing price in King County. In this exercise, creative feature engineering and advanced regression techniques will be explored, namely stepwise-type procedures, various forms of transformation and influential point analysis. Patterns found in residual plots and influential points will be evaluated and addressed to gear towards a more adequate model. The model is evaluated on the root mean square error between our predictions and the true value. The expectation is that at least more than one feature will be of relevance to the housing price and we will build a statistical model with adequate measures of R-squared and minimum squared errors of residuals to project the housing price in King Country. Patterns that we extract from this dataset could potentially be extrapolated to future studies in other US areas that have similar features with King County, and be used for those housing price predictions.

3 Data Preparation and Initial Exploration

3.1 Data Preprocessing

To clean the data for analysis, we performed several data preprocessing procedures. After checking in R, we found that there is no missing values in the data and thus no imputation was required. We also dropped identical observations in the dataset, reducing the number of observations to 21,608. As shown in (Appendix-Table.1), the raw data has *id*, *date* and *zipcode* as its variables. However, we decided to remove these variables since they are either not relevant in the problem or representing duplicated information. Specifically, *id* is excluded because it is simply a notation for the house. Since we are not looking at the problem from a time series perspective, *date* was also removed. *Zipcode* is duplicated information because the other two variables, *long* and *lat*, reflect the location of a house.

Next, we identified categorical variables to ensure candidate models would interpret them correctly. Among all the variables, we decided to code *yr_renovated*, *waterfront* and *condition* as categorical. Most of value in *yr_renovated* is 0, therefore we decided to create a dummy variable *isRenovated* indicating the presence of renovation. *Waterfront* indicates whether a house has a waterfront view and it is either 0 or 1. We did not create dummy variable for *waterfront* as there are only two levels and the variable is already coded in the right form. *condition* reflects the overall condition of the house and its values range from 1 to 5. These values are discrete and do not have a natural scale of measurement like *sqft_living* does. To ensure our analytic methods will appropriately identify the five levels of condition, we created four dummy variables as shown in (Appendix-Table.2).

3.2 Basic Features and Summary Statistics

After data preprocessing, there are 18 features in total including 15 numeric variables and three categorical variables. We checked the minimum value, maximum value, medium, mean and standard deviation, and found that the variables are measured in different scales (Ex: Price is ranging from $75,000 to $7,700,000, while bedrooms has a range from 0 to 33.). Besides, *price*, *bedrooms*, *bathrooms*, *sqft_living*, *sqft_lot* and *sqft_basement* have outliers, the analysis on the outliers and influential points is discussed in more detail in following sections.

From the correlation analysis, *sqft_living* have high correlation with *price*, *bathrooms*, *grade*, *sqft_above* and *sqft_living*. *sqft_above* has correlation with *grade*. *sqft_above* has correlation with grade and *sqft_living15*. From the distribution plot (Appendix-Figure.3), the response variable *price* is right-skewed. According to pairs plot (Appendix-Figure.4, Figure.5 and Figure.6), and scatter plot(Appendix-Figure.7, Figure.8 and Figure.9), we found that price might have a relationship with *bedrooms*, *bathrooms*, *sqft_living*, *sqft_lot*, *condition* and *grade*.

4 Data Analysis and Statistical Inference

4.1 Methodology

In achieving the objective of uncovering the relationship between parameters and housing price, we utilized several statistical methods in an iterative way to build a satisfactory model. Since multicollinearity could lead to exaggerated estimations and variances for coefficients, we applied variance inflation factor (VIF) analysis and Eigensystem analysis to identify variables that may be involved in multicollinearity. Then the model would be respecified with these variables eliminated for re-evaluation. To ensure our model is not affected by only a few of the observations, we employed different diagnosis methods to locate the influential points for further treatment. We assessed model adequacy by examining the residual plots. Specifically, we focused on the normal probability plot and the plot of residuals against fitted value as these plots would reveal departures from the major assumptions for linear regression. Transformation techniques then help us address the identified model inadequacies. In the spirit of parsimony, we also implemented variable selection techniques such as forward selection to eliminate redundant regressors, if any.

4.2 Regression Techniques

4.2.1 VIF and Eigensystem Analysis

As mentioned in the Methodology section above, we implemented an iterative model building process. We first built the full linear regression model using all variables after data cleaning. The summary of the model (Appendix-Figure.10) indicates that the estimated coefficient for *sqft_basement* is NA, suggesting possible multicollinearity problem. After checking dependency of the terms in the model, we found that *sqft_basement* is a linear combination of *sqft_living* and *sqft_above*, suggesting removal of *sqft_basement* in model. To investigate additional issues with multicollinearity, we examined the VIF for each term in a new linear model without *sqft_basement*. It seems *sqft_living*, *sqft_above* and all condition dummy variables are involved in multicollinearity. After removing *sqft_above* and one of the condition levels (*condition.3*), the multicollinearity in the model is effectively reduced. This improvement is supported by the variance decomposition proportions analysis in Appendix-Figure.13 as none of the condition index is above the widely used cutoff of 30.

4.2.2 Influential Points Diagnosis

Fo influential diagnosis, we tried four different methods: Cook's distance, DFFITS, DFBETAS, and COVRATIO. See each normal probability and residual-by-fitted-value plots in Appendix-Figure.15, Figure.16 and Figure.17.

- **Cook's Distance**: Since we usually consider points for which $Di > 1$ to be influential, the cutoff value is too large, so there are no influential points for our dataset.
- **DFFITS**: It shows that there are 1,117 influential points, and after removing those influential points, the normality probability and residual-by-fitted-value plots are much better, since we can see that the number

of outliers reduced a lot. Also the Adjusted R-squared for original model is 0.6949, and the Adjusted R-squared for DFFITS model is 0.7026, which improves a lot.

- **DFBETAS**: By using DFBETAS we just have three influential points. Because we have 17 columns, and all of those columns are very important to consider, we need to sum up the the number of columns whose value is larger than cutoff value and choose the sum rows which have the sum equals to 17. Only by doing so we can be sure that those points are influential points. But after removing these three points the normal probability and residual-by-fitted-value plots are almost the same since there are 21,605 rows, three outliers will not make any impact on the whole dataset. Also the Adjusted R-squared for original model is 0.6947, and the Adjusted R-squared for DFBETAS model is 0.6956, which are almost the same.
- **COVRATIO**: By using COVRATIO we have 1,324 influential points. After removing those points the normal probability and residual-by-fitted-value plots are better, we can see that the number of outliers reduced a lot. Also the Adjusted R-squared for original model is 0.6947, and the Adjusted R-squared for COVRATIO model is 0.7008 which also improves a lot.

After comparing those four methods, we conclude that DFFITS method gives us best result and we will remove those 1,117 influential points in order to get a better model with higher Adjusted R-squared.

4.2.3 Residual Plots and Transformation Attempts

In this study we compared residual plots for several different models along the data exploration process. The first residual plot came from the first model without any alterations to the features. From Appendix-Figure.19 we can see that the normal probability plot of residuals has a curvy shape with considerably heavy tails. This suggests that there are flaws in assumptions for linear regression. Also, in the R-student residual against fitted values plot, we can see a clear funnel shape. The assumption of constant variance is a basic requirement of regression analysis. From the second residual plot, this assumption is violated due to response variable following a certain distribution, in which variance is functionally related to the mean. Both of these plots imply that transformation is required. We had employed several transformation methods: log transformation, Box-Cox transformation, Box-Tidwell transformation. We also implemented polynomial regression.

- **Log Transformation**: Under log transformation of response variable housing price, both normal probability plot and residual plot of R-student against fitted values are improving. As shown in Figure.1, the normal probability plot barely has any skewness in both upper and lower tails and approaches to the straight line. The fitted value residual plot is scattered along the x-axis randomly. This aligns with our assumption for constant variance for linear regression, which means log transformation should be used for response variable. Also, log transformation gives an improved R-squared from 0.6949 to 0.7674.

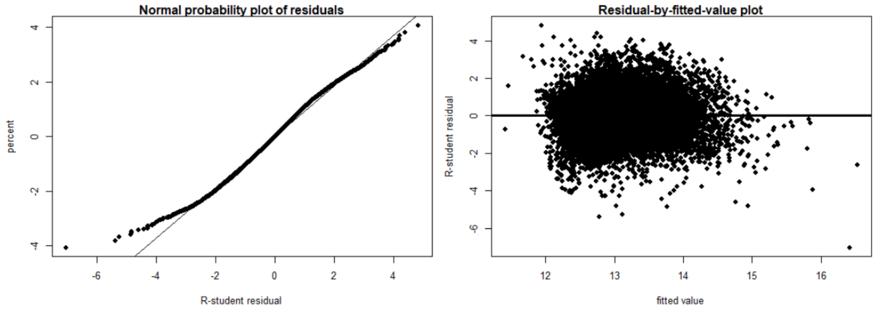


Figure 1: Residual Plot after Log Transformation

- **Box-Cox Transformation**: We applied Box-Cox transformation on the original full model. However, the selection plot (Appendix-Figure.20 and sum of squared residuals table (Appendix-Figure.21) implied that transformation of power parameter lambda as 0 gives smallest sum of squared residual. We conducted the transformation with lambda as 0, and got the R-squared of 0.7674. This is not an improvement compared to R-squared from log transformation, hence we do not apply Box-Cox transformation.

- **Box-Tidwell Transformation**: Since in previous section we already indicated that log transformation should be applied to response variable, now we explore transformation on regressor variables. The four continuous variables are *sqft_living*, *sqft_lot*, *sqft_living15*, *sqft_lot15*, and power transformation is applied to these variables using Box-Tidwell transformation. According to the suggestion of Box-Tidwell transformation, we could apply power transformation of 1/4 to *sqft_lot* and *sqft_lot15*. We applied power transformation of 1/4 to *sqft_lot* and *sqft_lot15* and built one model with both the original regressors and the transformed regressors, as well as another model with the transformed regressors only. The Adjusted R-squared (0.7704) of former model is better than that of latter (0.7673), therefore, we decided to check the model with transformed regressors as the additive terms. The residual plots are shown in Figure.2:

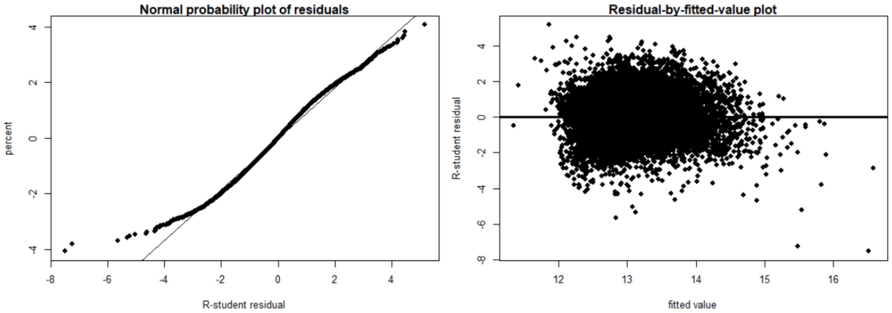


Figure 2: Residual Plot after Box-Tidwell Transformation

R-student residual scatters around the x axis in a clustered shape and qq plot has heavy tail. These residual plots are not ideal but they improved from previous plots.

- **Polynomial Regression**: We applied polynomial transformation to *sqft_living*, *bedrooms* by using their squared values as well as the interaction term. Both the quadratic terms are significant, while the interaction term is not significant. We still keep the interaction term since it should be kept in high-order polynomial regression. Also, we applied polynomial transformation to *sqft_lot*, *bathrooms* by using their squared values and the interaction term. From the summary statistics in Appendix-Figure.27 and Figure.28, the Adjusted R-squared 0.77 does not have much improvement from before. Hence we leave the terms for feature selection.

4.2.4 Forward Selection, Backward Elimination, Stepwise Regression

Stepwise-type procedures, which mainly include forward selection, backward elimination, and stepwise regression, were used to fulfill feature selection. The resulted models from the three procedures were compared.

Forward selection supposes no regressor in the model and takes variables into account one at a time. The variable that has the largest simple correlation with the response has the priority to enter the model. In our project, we set the entry threshold to be 0.25. If the variable whose F statistic exceeded 0.25 and also was the largest among all candidate variables, it got the entry first. In forward selection, the first variable entered the model was *bedrooms*. Then, *bathrooms*, *sqft_living* and other variables entered the model sequentially. The interaction term *living_bedroom* and the quadratic term *bathrooms2* were excluded, which indicating all other variables contributed to a house price based on forward selection.

Backward elimination begins with a model with all variables. The variable that has the least simple correlation with the response is first excluded. For backward elimination, we set the exit threshold to be 0.10. If the smallest F statistic among all variables was less than this threshold value, the variable first left the model. The result from running backward elimination also excluded two variables: *living_bedrooms* and *bathrooms2* as forward selection, and keep all other variables that have significant contribution to predicting house price.

Stepwise regression is a combination of forward selection and backward elimination. It is a procedure to reassess the variables which previously entered the model. We set the entry threshold to be 0.25 and exit threshold to be 0.10. In a similar way, the variable whose F statistic was the largest and exceeded 0.25 entered first, and the variables whose F statistic was the smallest and below 0.10 exited the model. The final result of stepwise regression has an agreement with the other two procedures, keeping all variables except for the two quadratic terms *living_bedrooms* and *bathrooms2*.

Based on the consistency among the three procedures, the final model is:

$log.price\ bedrooms + bathrooms + sqft\_living + sqft\_lot + floors + waterfront + view + grade + yr\_built + lat + long + sqft\_living15 + sqft\_lot15 + isRenovated + condition.4 + condition.5 + sqft\_living2 + bedrooms2$

4.3 Conclusion

The final model is:

$$\begin{aligned}\ln(price) = {} & -5.48 \times 10 - 2.95 \times 10^{-2} \times bedrooms + 7 \times 10^{-2} \times bathrooms + 2.42 \times 10^{-4} \\ & \times sqft\_living + 4.63 \times 10^{-7} \times sqft\_lot + 6.08 \times 10^{-2} \times floors + 3.83 \times 10^{-1} \\ & \times waterfront + 5.9 \times 10^{-2} \times view + 1.58 \times 10^{-1} \times grade - 3.31 \times 10^{-3} \\ & \times yr\_built + 1.36 \times lat - 6.38 \times 10^{-2} \times long + 9.31 \times 10^{-5} \times sqft\_living15 \\ & - 2.58 \times 10^{-7} \times sqft\_lot15 + 7.1 \times 10^{-2} \times isRenovated + 6.69 \times 10^{-2} \\ & \times condition.4 + 1.29 \times 10^{-1} \times condition.5 - 1.58 \times 10^{-8} \times sqft\_living^2 + 1.06 \\ & \times 10^{-3} \times bedrooms^2\end{aligned}$$

The final model has an Adjusted R-squared as 0.77. The main techniques that we used to derive this model from the original full model are: dummy variable treatment, influential points analysis, log transformation, polynomial regression and stepwise-type procedures. This result is achieved with model adequacy checking using residual plots so that there are no violations of variance assumptions. The majority of the regressors remained, which means most fields in this dataset are crucial drivers for housing price in King County of Washington. Based on the summary statistics of the final model, we can see that latitude, year built, and square foot of living are the three most important features that people are paying for their home in King county.

5 Appendix

Table 1: Variable description

Variable Name	Description
id	A notation for a house
date	Date house was sold
price	House sale price (the prediction target)
bedrooms	Number of Bedrooms/House
bathrooms	Number of bathrooms/bedrooms
sqft_living	Square footage of the home
sqft_lot	Square footage of the lot
floors	Total floors (levels) in house
waterfront	Whether house has a view to a waterfront
view	Number of times the house has been viewed
condition	Overall condition of house
grade	Overall grade given to the housing unit, based on King County grading system
sqft_above	Square footage of house apart from basement
sqft_basement	Square footage of the basement
yr_built	Built year
yr_renovated	Year when house was renovated
zipcode	zip
lat	Latitude coordinate
long	Longitude coordinate
sqft_living15	Living room area in 2015 (implies– some renovations)
sqft_lot15	Lot size area in 2015 (implies– some renovations)

Table 2: Dummy Variables for Condition

The following table shows the levels of the dummy variables for condition. Since there are five levels for condition, four dummy variables are required to incorporate these levels.

condition.2	condition.3	condition.4	condition.5	condition (original value)
0	0	0	0	1
1	0	0	0	2
0	1	0	0	3
0	0	1	0	4
0	0	0	1	5

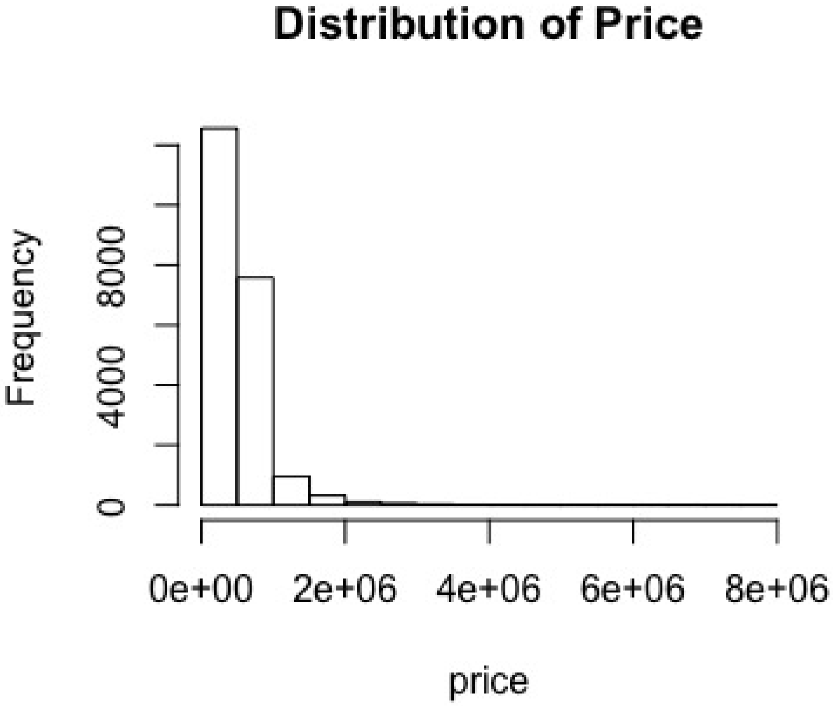


Figure 3: Distribution of *price*

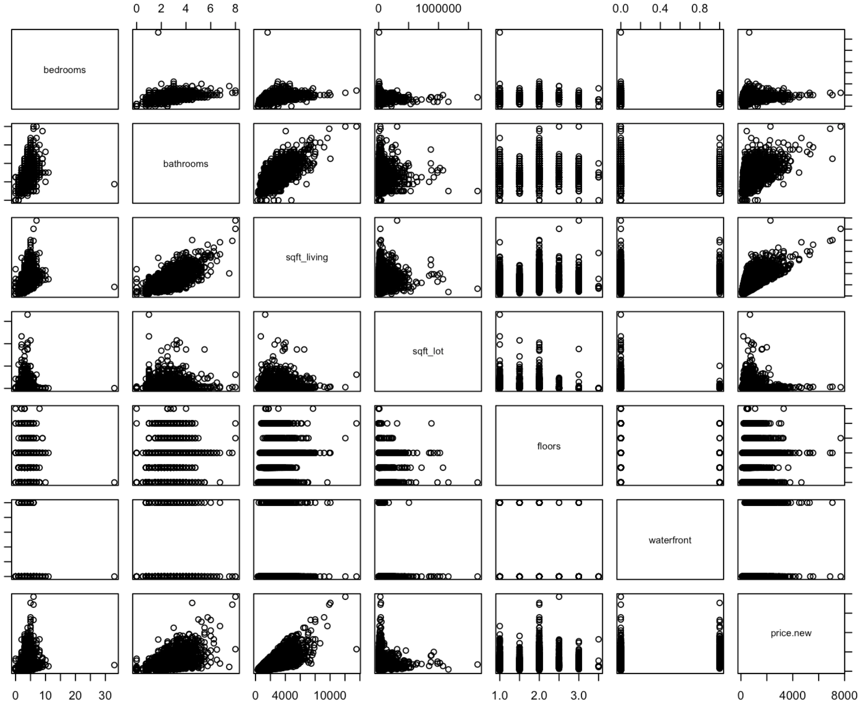


Figure 4: Pairs Plot

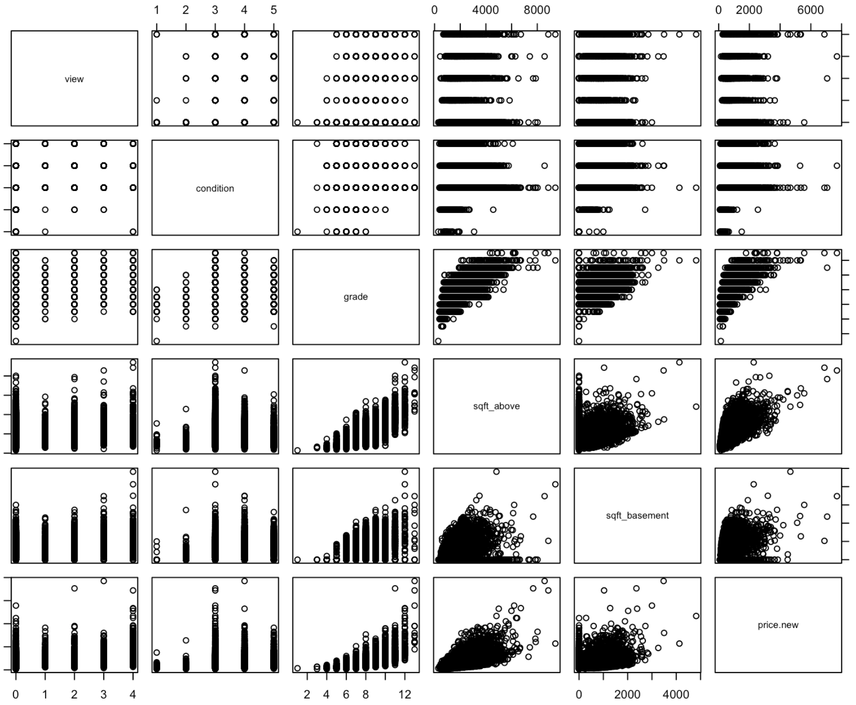


Figure 5: Pairs Plot

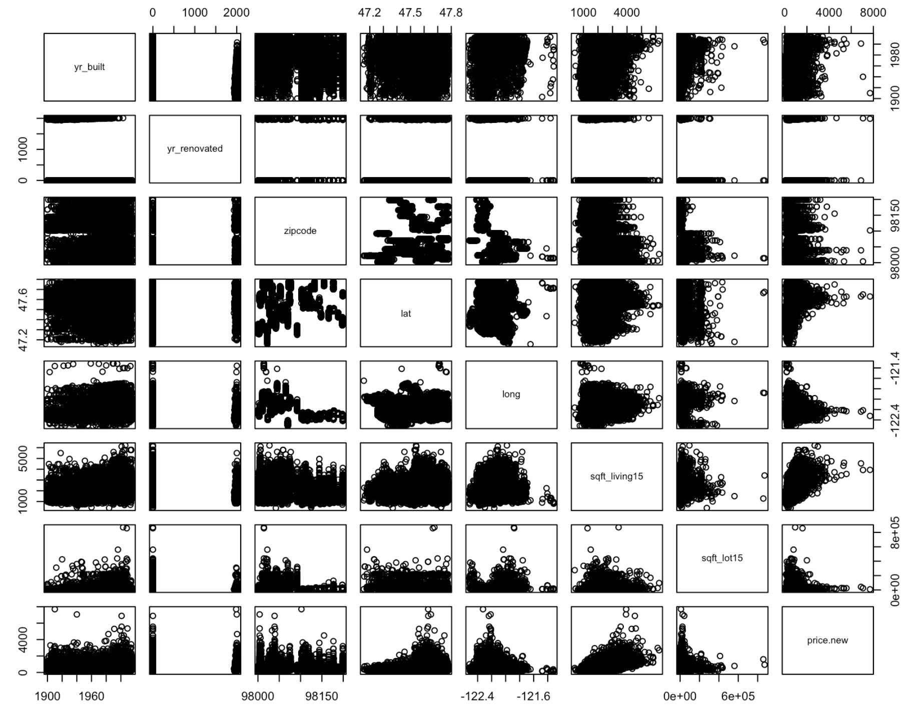


Figure 6: Pairs Plot

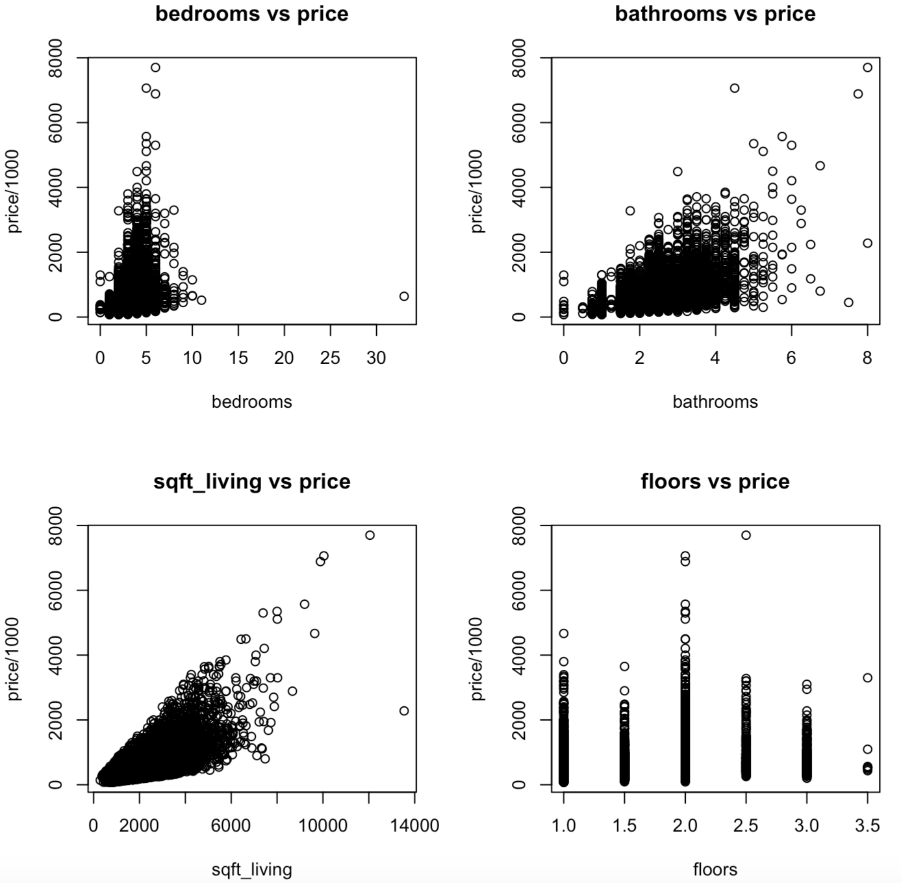


Figure 7: Scatter Plot

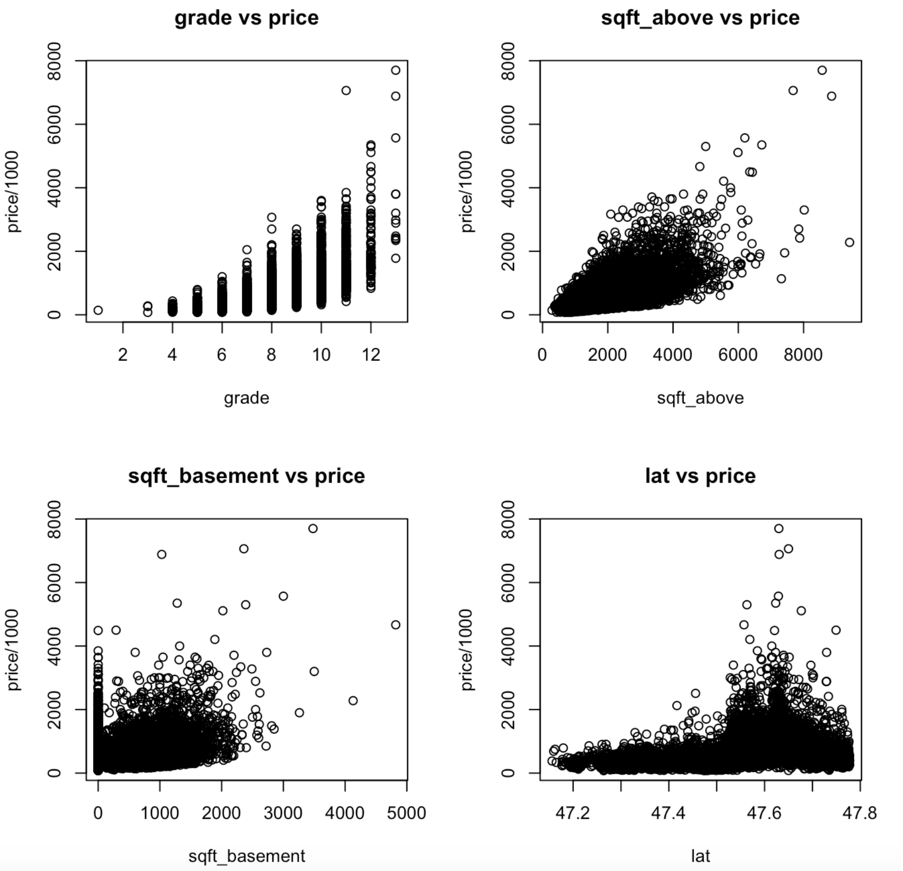


Figure 8: Scatter Plot

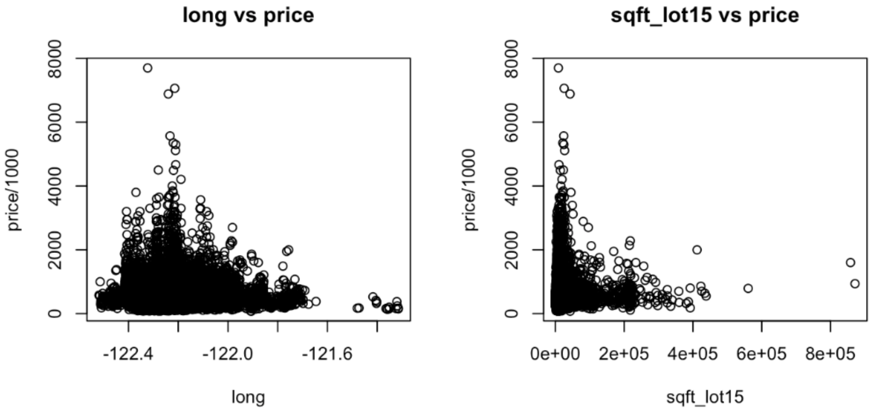
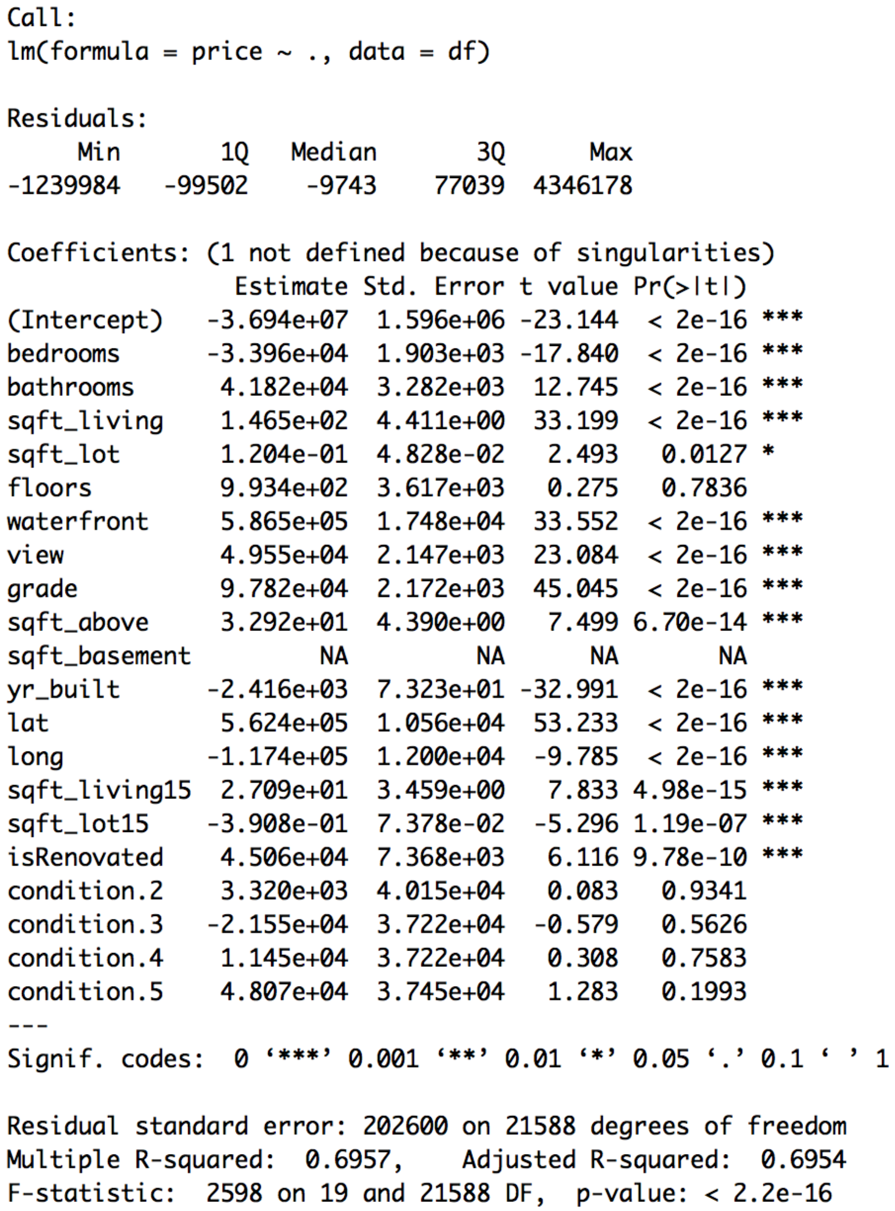
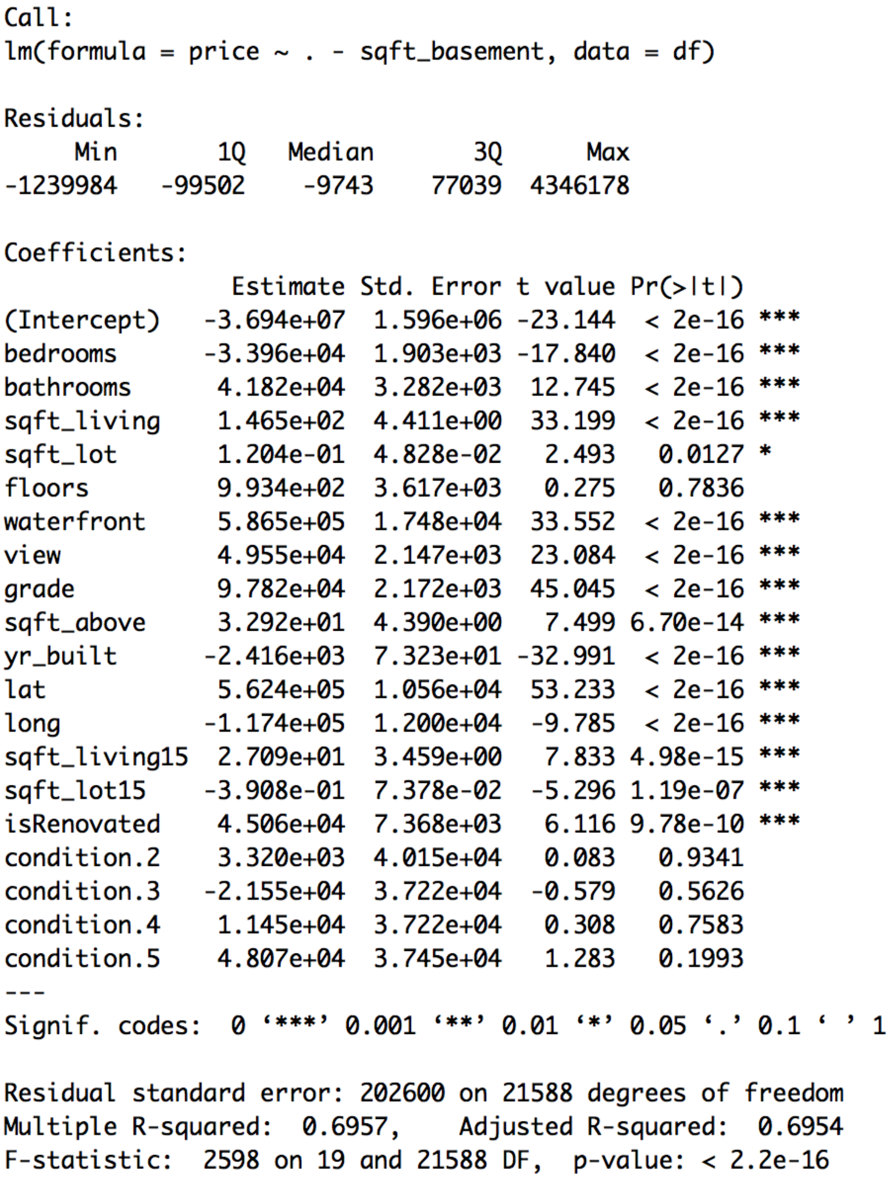


Figure 9: Scatter Plot



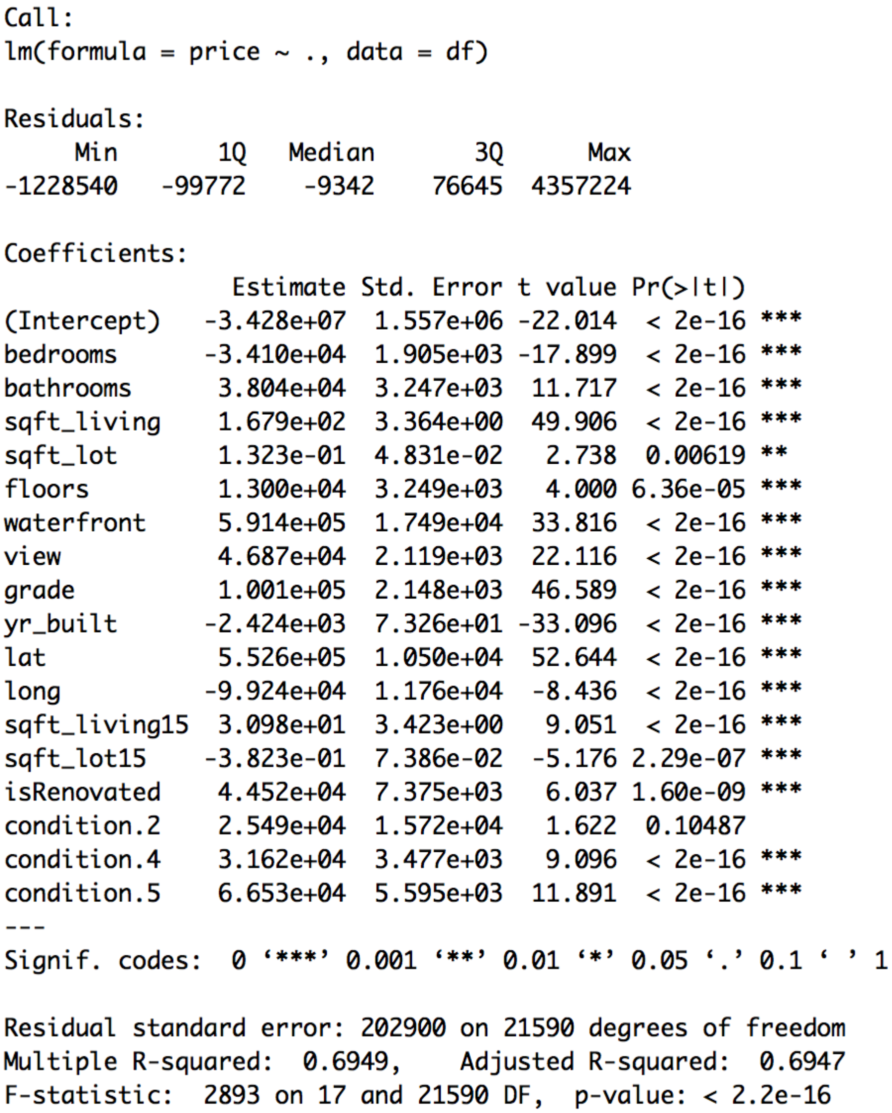
```
Call:
lm(formula = price ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1239984   -99502    -9743    77039  4346178 

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -3.694e+07  1.596e+06 -23.144  < 2e-16 ***
bedrooms      -3.396e+04  1.903e+03 -17.840  < 2e-16 ***
bathrooms      4.182e+04  3.282e+03  12.745  < 2e-16 ***
sqft_living    1.465e+02  4.411e+00  33.199  < 2e-16 ***
sqft_lot       1.204e-01  4.828e-02   2.493   0.0127 *  
floors         9.934e+02  3.617e+03   0.275   0.7836    
waterfront     5.865e+05  1.748e+04  33.552  < 2e-16 ***
view           4.955e+04  2.147e+03  23.084  < 2e-16 ***
grade          9.782e+04  2.172e+03  45.045  < 2e-16 ***
sqft_above     3.292e+01  4.390e+00   7.499 6.70e-14 ***
sqft_basement         NA         NA      NA       NA    
yr_built      -2.416e+03  7.323e+01 -32.991  < 2e-16 ***
lat            5.624e+05  1.056e+04  53.233  < 2e-16 ***
long          -1.174e+05  1.200e+04  -9.785  < 2e-16 ***
sqft_living15  2.709e+01  3.459e+00   7.833 4.98e-15 ***
sqft_lot15    -3.908e-01  7.378e-02  -5.296 1.19e-07 ***
isRenovated    4.506e+04  7.368e+03   6.116 9.78e-10 ***
condition.2    3.320e+03  4.015e+04   0.083   0.9341    
condition.3   -2.155e+04  3.722e+04  -0.579   0.5626    
condition.4    1.145e+04  3.722e+04   0.308   0.7583    
condition.5    4.807e+04  3.745e+04   1.283   0.1993    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 202600 on 21588 degrees of freedom
Multiple R-squared:  0.6957,	Adjusted R-squared:  0.6954 
F-statistic:  2598 on 19 and 21588 DF,  p-value: < 2.2e-16
```

Figure 10: Summary 1: Full model



```
Call:
lm(formula = price ~ . - sqft_basement, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1239984   -99502    -9743    77039  4346178 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -3.694e+07  1.596e+06 -23.144  < 2e-16 ***
bedrooms      -3.396e+04  1.903e+03 -17.840  < 2e-16 ***
bathrooms      4.182e+04  3.282e+03  12.745  < 2e-16 ***
sqft_living    1.465e+02  4.411e+00  33.199  < 2e-16 ***
sqft_lot       1.204e-01  4.828e-02   2.493   0.0127 *  
floors         9.934e+02  3.617e+03   0.275   0.7836    
waterfront     5.865e+05  1.748e+04  33.552  < 2e-16 ***
view           4.955e+04  2.147e+03  23.084  < 2e-16 ***
grade          9.782e+04  2.172e+03  45.045  < 2e-16 ***
sqft_above     3.292e+01  4.390e+00   7.499 6.70e-14 ***
yr_built      -2.416e+03  7.323e+01 -32.991  < 2e-16 ***
lat            5.624e+05  1.056e+04  53.233  < 2e-16 ***
long          -1.174e+05  1.200e+04  -9.785  < 2e-16 ***
sqft_living15  2.709e+01  3.459e+00   7.833 4.98e-15 ***
sqft_lot15    -3.908e-01  7.378e-02  -5.296 1.19e-07 ***
isRenovated    4.506e+04  7.368e+03   6.116 9.78e-10 ***
condition.2    3.320e+03  4.015e+04   0.083   0.9341    
condition.3   -2.155e+04  3.722e+04  -0.579   0.5626    
condition.4    1.145e+04  3.722e+04   0.308   0.7583    
condition.5    4.807e+04  3.745e+04   1.283   0.1993    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 202600 on 21588 degrees of freedom
Multiple R-squared:  0.6957,	Adjusted R-squared:  0.6954 
F-statistic:  2598 on 19 and 21588 DF,  p-value: < 2.2e-16
```

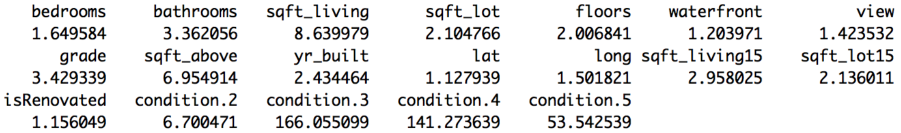
Figure 11: Model without sqft_basement



```
Call:
lm(formula = price ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1228540   -99772    -9342    76645  4357224 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -3.428e+07  1.557e+06 -22.014  < 2e-16 ***
bedrooms      -3.410e+04  1.905e+03 -17.899  < 2e-16 ***
bathrooms      3.804e+04  3.247e+03  11.717  < 2e-16 ***
sqft_living    1.679e+02  3.364e+00  49.906  < 2e-16 ***
sqft_lot       1.323e-01  4.831e-02   2.738  0.00619 ** 
floors         1.300e+04  3.249e+03   4.000 6.36e-05 ***
waterfront     5.914e+05  1.749e+04  33.816  < 2e-16 ***
view           4.687e+04  2.119e+03  22.116  < 2e-16 ***
grade          1.001e+05  2.148e+03  46.589  < 2e-16 ***
yr_built      -2.424e+03  7.326e+01 -33.096  < 2e-16 ***
lat            5.526e+05  1.050e+04  52.644  < 2e-16 ***
long          -9.924e+04  1.176e+04  -8.436  < 2e-16 ***
sqft_living15  3.098e+01  3.423e+00   9.051  < 2e-16 ***
sqft_lot15    -3.823e-01  7.386e-02  -5.176 2.29e-07 ***
isRenovated    4.452e+04  7.375e+03   6.037 1.60e-09 ***
condition.2    2.549e+04  1.572e+04   1.622  0.10487    
condition.4    3.162e+04  3.477e+03   9.096  < 2e-16 ***
condition.5    6.653e+04  5.595e+03  11.891  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 202900 on 21590 degrees of freedom
Multiple R-squared:  0.6949,    Adjusted R-squared:  0.6947 
F-statistic:  2893 on 17 and 21590 DF,  p-value: < 2.2e-16
```

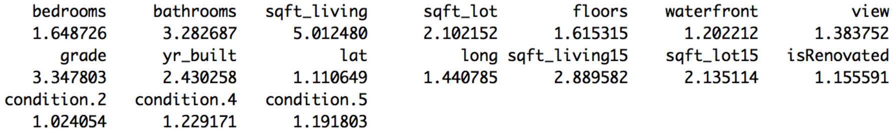
Figure 12: Model without sqrt_basement, sqft_above and condition.3

VIF output for linear model without *sqft_basement*



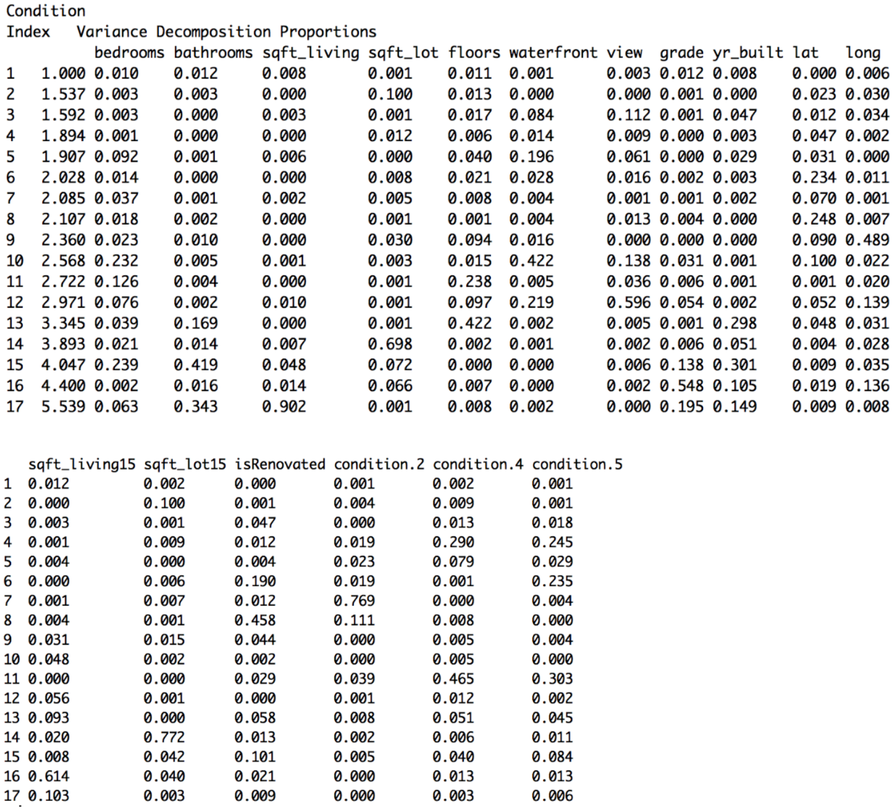
```
   bedrooms   bathrooms sqft_living    sqft_lot      floors  waterfront        view 
   1.649584    3.362056    8.639979    2.104766    2.006841    1.203971    1.423532 
      grade  sqft_above    yr_built         lat        long sqft_living15  sqft_lot15 
   3.429339    6.954914    2.434464    1.127939    1.501821    2.958025    2.136011 
isRenovated condition.2 condition.3 condition.4 condition.5 
   1.156049    6.700471  166.055099  141.273639   53.542539 
```

VIF output for linear model without *sqft_basement*, *sqrt_above* and *condition.3*



```
   bedrooms   bathrooms sqft_living    sqft_lot      floors  waterfront        view 
   1.648726    3.282687    5.012480    2.102152    1.615315    1.202212    1.383752 
      grade    yr_built         lat        long sqft_living15  sqft_lot15 isRenovated 
   3.347803    2.430258    1.110649    1.440785    2.889582    2.135114    1.155591 
condition.2 condition.4 condition.5 
   1.024054    1.229171    1.191803 
```

Figure 13: VIF



```
Condition
Index   Variance Decomposition Proportions
          bedrooms bathrooms sqft_living sqft_lot floors waterfront view  grade yr_built lat   long
1   1.000 0.010    0.012     0.008       0.001    0.011  0.001      0.003 0.012 0.008    0.000 0.006
2   1.537 0.003    0.003     0.000       0.100    0.013  0.000      0.000 0.001 0.000    0.023 0.030
3   1.592 0.003    0.000     0.003       0.001    0.017  0.084      0.112 0.001 0.047    0.012 0.034
4   1.894 0.001    0.000     0.000       0.012    0.006  0.014      0.009 0.000 0.003    0.047 0.002
5   1.907 0.092    0.001     0.006       0.000    0.040  0.196      0.061 0.000 0.029    0.031 0.000
6   2.028 0.014    0.000     0.000       0.008    0.021  0.028      0.016 0.002 0.003    0.234 0.011
7   2.085 0.037    0.001     0.002       0.005    0.008  0.004      0.001 0.001 0.002    0.070 0.001
8   2.107 0.018    0.002     0.000       0.001    0.001  0.004      0.013 0.004 0.000    0.248 0.007
9   2.360 0.023    0.010     0.000       0.030    0.094  0.016      0.000 0.000 0.000    0.090 0.489
10  2.568 0.232    0.005     0.001       0.003    0.015  0.422      0.138 0.031 0.001    0.100 0.022
11  2.722 0.126    0.004     0.000       0.001    0.238  0.005      0.036 0.006 0.001    0.001 0.020
12  2.971 0.076    0.002     0.010       0.001    0.097  0.219      0.596 0.054 0.002    0.052 0.139
13  3.345 0.039    0.169     0.000       0.001    0.422  0.002      0.005 0.001 0.298    0.048 0.031
14  3.893 0.021    0.014     0.007       0.698    0.002  0.001      0.002 0.006 0.051    0.004 0.028
15  4.047 0.239    0.419     0.048       0.072    0.000  0.000      0.006 0.138 0.301    0.009 0.035
16  4.400 0.002    0.016     0.014       0.066    0.007  0.000      0.002 0.548 0.105    0.019 0.136
17  5.539 0.063    0.343     0.902       0.001    0.008  0.002      0.000 0.195 0.149    0.009 0.008


   sqft_living15 sqft_lot15 isRenovated condition.2 condition.4 condition.5
1  0.012         0.002      0.000       0.001       0.002       0.001
2  0.000         0.100      0.001       0.004       0.009       0.001
3  0.003         0.001      0.047       0.000       0.013       0.018
4  0.001         0.009      0.012       0.019       0.290       0.245
5  0.004         0.000      0.004       0.023       0.079       0.029
6  0.000         0.006      0.190       0.019       0.001       0.235
7  0.001         0.007      0.012       0.769       0.000       0.004
8  0.004         0.001      0.458       0.111       0.008       0.000
9  0.031         0.015      0.044       0.000       0.005       0.004
10 0.048         0.002      0.002       0.000       0.005       0.000
11 0.000         0.000      0.029       0.039       0.465       0.303
12 0.056         0.001      0.000       0.001       0.012       0.002
13 0.093         0.000      0.058       0.008       0.051       0.045
14 0.020         0.772      0.013       0.002       0.006       0.011
15 0.008         0.042      0.101       0.005       0.040       0.084
16 0.614         0.040      0.021       0.000       0.013       0.013
17 0.103         0.003      0.009       0.000       0.003       0.006
```

Figure 14: Condition Index and Variance Decomposition Proportions

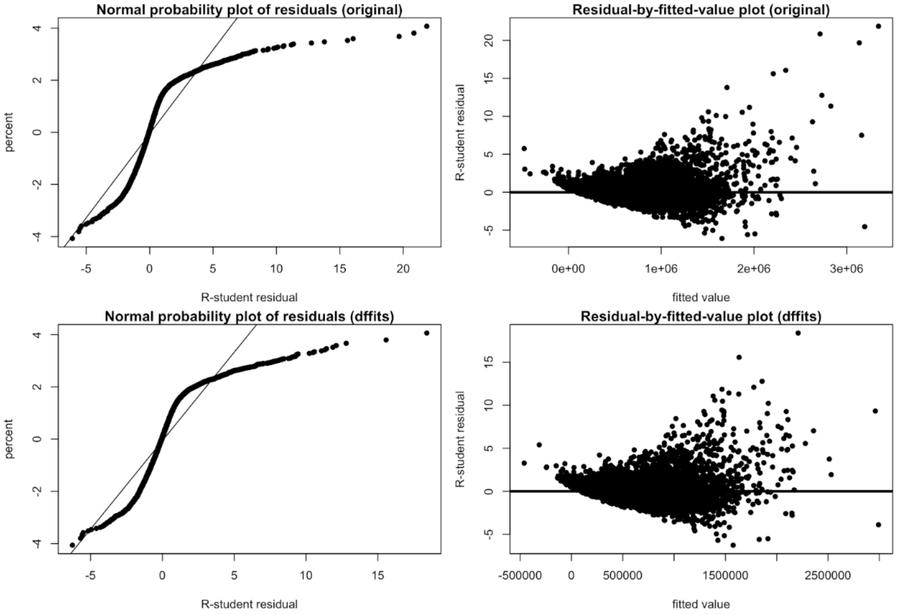


Figure 15: Residual plot (DFFITS)

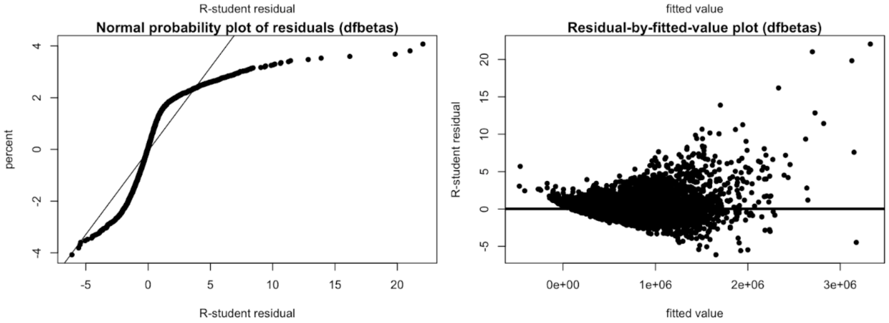


Figure 16: Residual plot (DFBETAS)

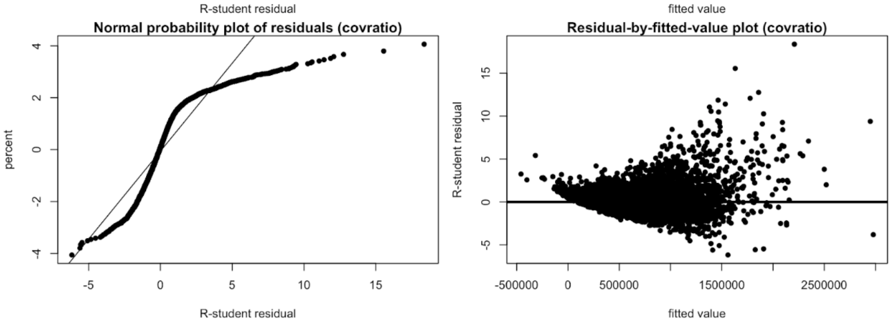


Figure 17: Residual plot (COVRATIO)

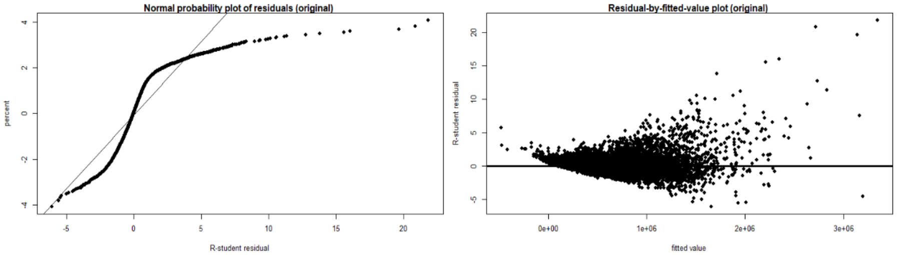


Figure 18: Residual plot (Full Model)

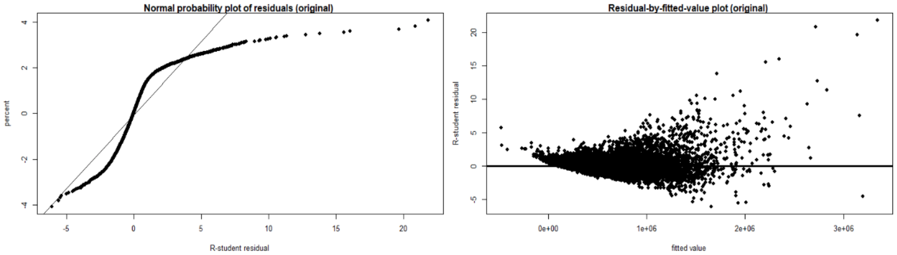


Figure 19: Residual plot (Full Model)

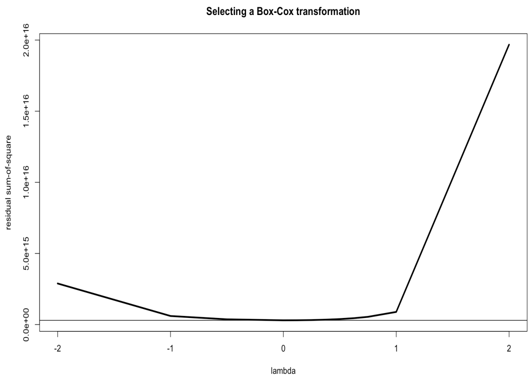
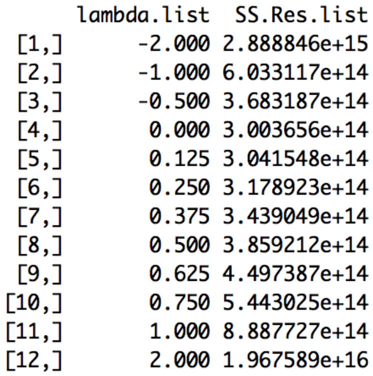
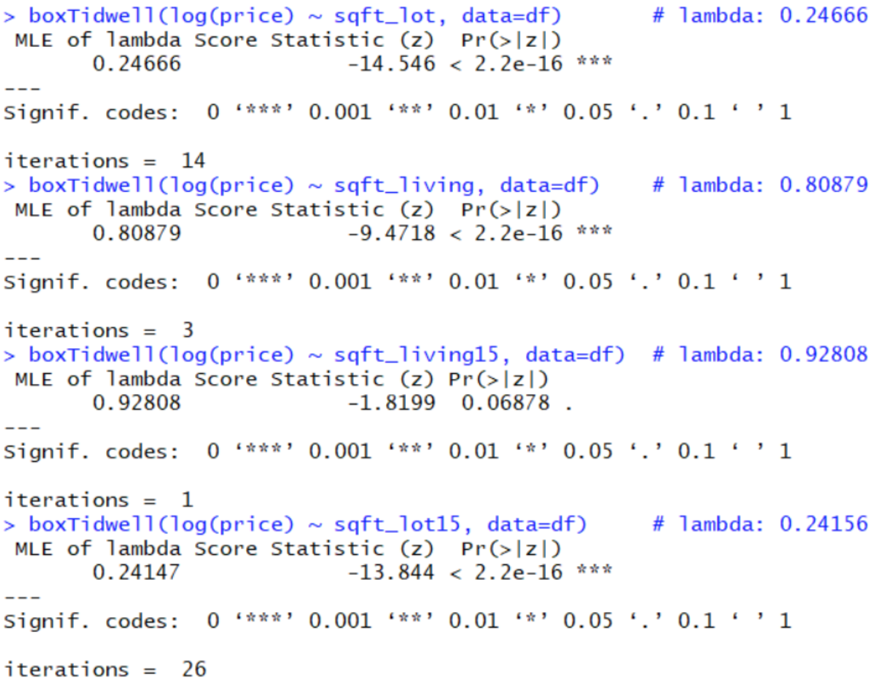


Figure 20: Selection of Box-Cox Transformation



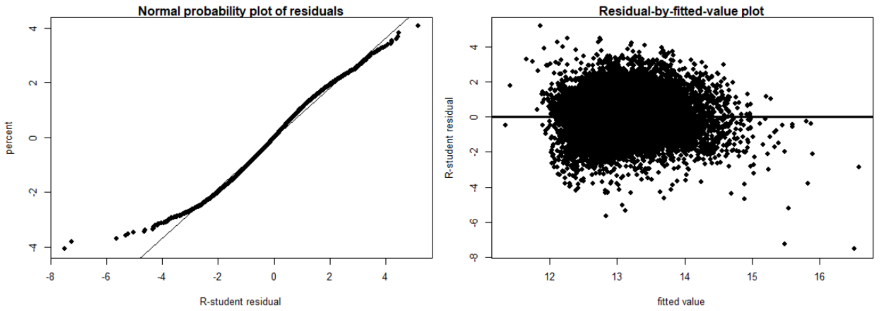
```
      lambda.list  SS.Res.list
 [1,]      -2.000 2.888846e+15
 [2,]      -1.000 6.033117e+14
 [3,]      -0.500 3.683187e+14
 [4,]       0.000 3.003656e+14
 [5,]       0.125 3.041548e+14
 [6,]       0.250 3.178923e+14
 [7,]       0.375 3.439049e+14
 [8,]       0.500 3.859212e+14
 [9,]       0.625 4.497387e+14
[10,]       0.750 5.443025e+14
[11,]       1.000 8.887727e+14
[12,]       2.000 1.967589e+16
```

Figure 21: Lambda vs Residual sum squared



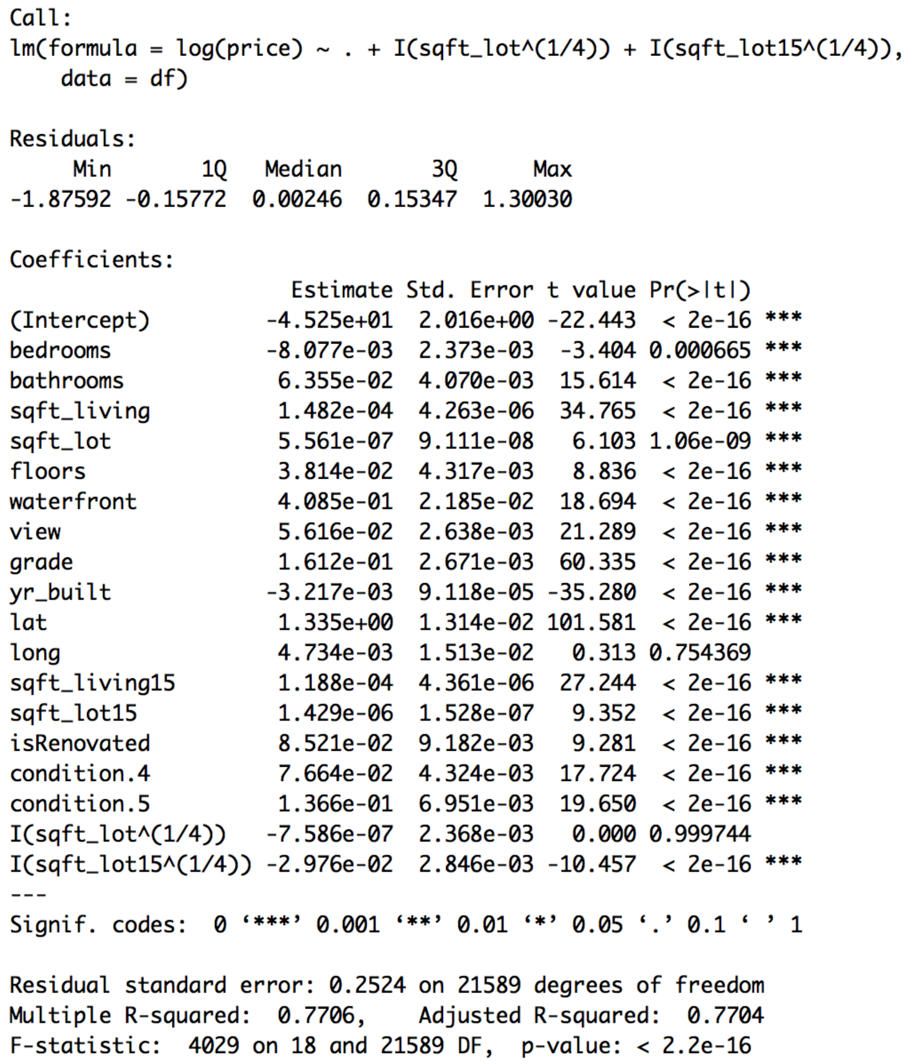
```
> boxTidwell(log(price) ~ sqft_lot, data=df)       # lambda: 0.24666
 MLE of lambda Score Statistic (z)  Pr(>|z|)    
       0.24666              -14.546 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

iterations =  14
> boxTidwell(log(price) ~ sqft_living, data=df)    # lambda: 0.80879
 MLE of lambda Score Statistic (z)  Pr(>|z|)    
       0.80879              -9.4718 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

iterations =  3
> boxTidwell(log(price) ~ sqft_living15, data=df)  # lambda: 0.92808
 MLE of lambda Score Statistic (z) Pr(>|z|)  
       0.92808              -1.8199  0.06878 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

iterations =  1
> boxTidwell(log(price) ~ sqft_lot15, data=df)     # lambda: 0.24156
 MLE of lambda Score Statistic (z)  Pr(>|z|)    
       0.24147              -13.844 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

iterations =  26
```

Figure 22: Box-Tidwell Tables)



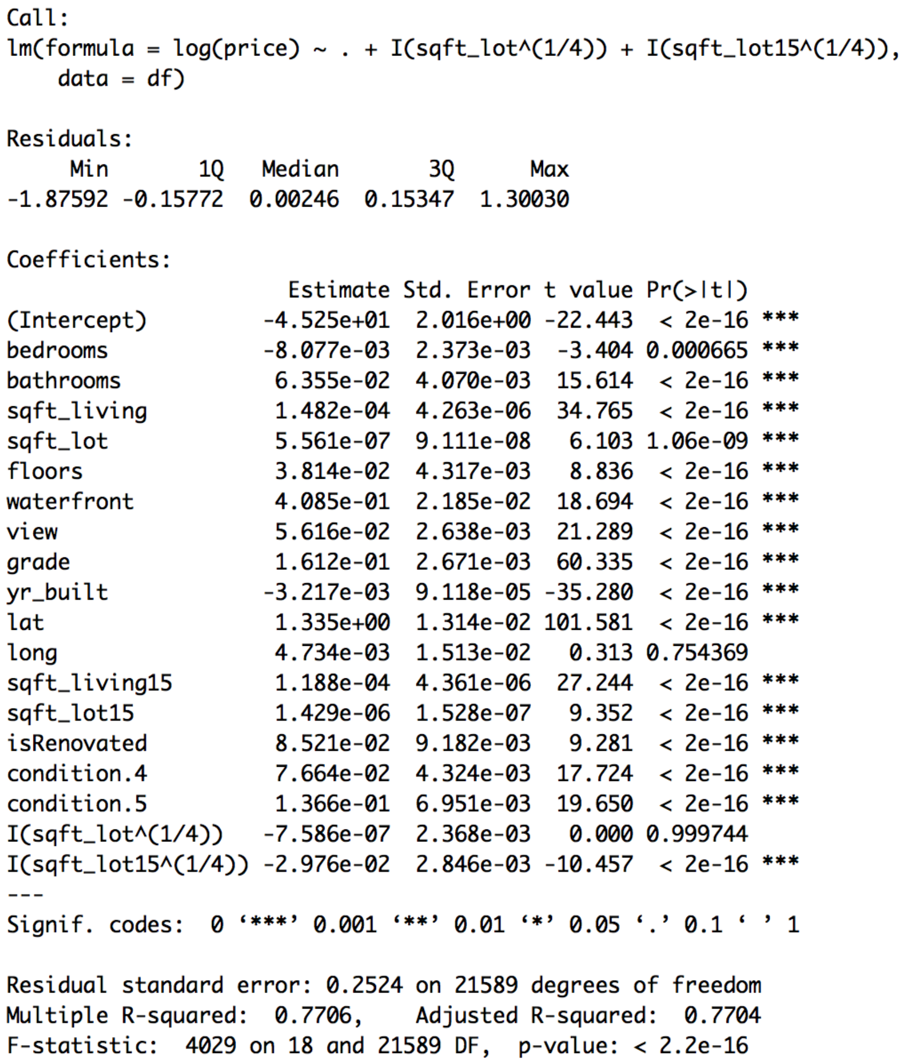
]

Figure 23: Residual plot of Polynomial Regression)



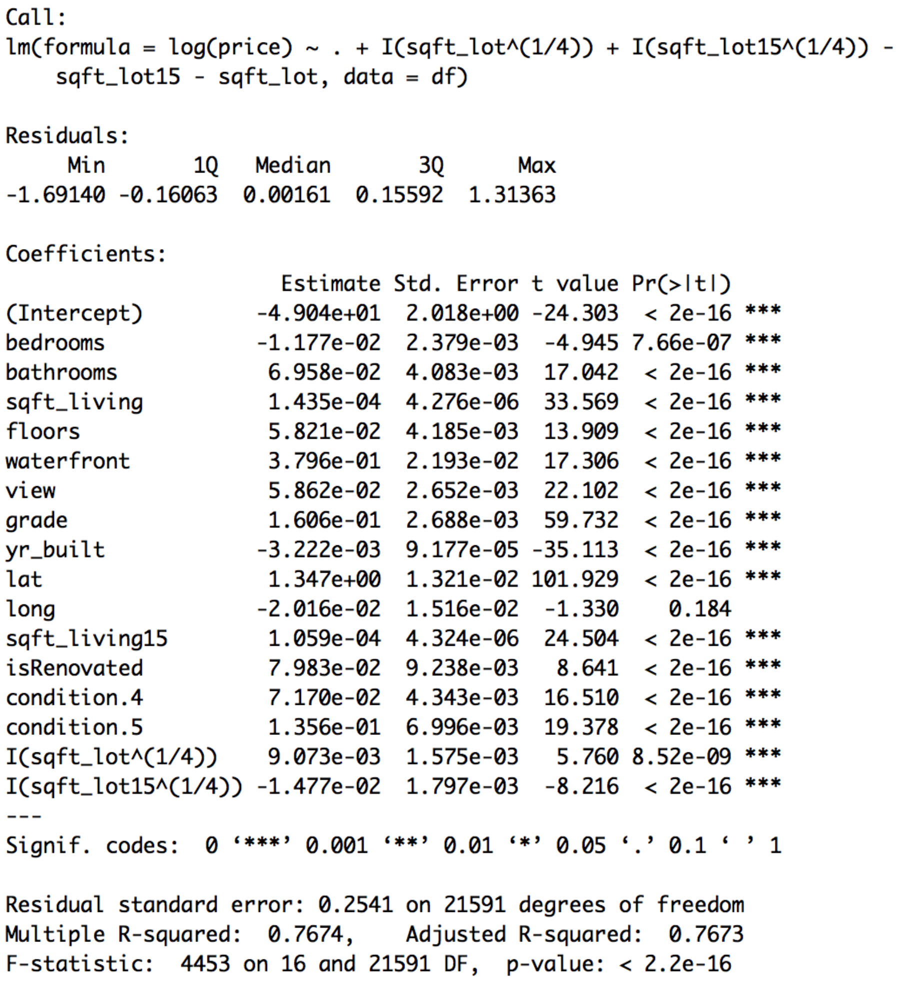
```
Call:
lm(formula = log(price) ~ . + I(sqft_lot^(1/4)) + I(sqft_lot15^(1/4)),
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.87592 -0.15772  0.00246  0.15347  1.30030 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -4.525e+01  2.016e+00 -22.443  < 2e-16 ***
bedrooms            -8.077e-03  2.373e-03  -3.404 0.000665 ***
bathrooms            6.355e-02  4.070e-03  15.614  < 2e-16 ***
sqft_living          1.482e-04  4.263e-06  34.765  < 2e-16 ***
sqft_lot             5.561e-07  9.111e-08   6.103 1.06e-09 ***
floors               3.814e-02  4.317e-03   8.836  < 2e-16 ***
waterfront           4.085e-01  2.185e-02  18.694  < 2e-16 ***
view                 5.616e-02  2.638e-03  21.289  < 2e-16 ***
grade                1.612e-01  2.671e-03  60.335  < 2e-16 ***
yr_built            -3.217e-03  9.118e-05 -35.280  < 2e-16 ***
lat                  1.335e+00  1.314e-02 101.581  < 2e-16 ***
long                 4.734e-03  1.513e-02   0.313 0.754369    
sqft_living15        1.188e-04  4.361e-06  27.244  < 2e-16 ***
sqft_lot15           1.429e-06  1.528e-07   9.352  < 2e-16 ***
isRenovated          8.521e-02  9.182e-03   9.281  < 2e-16 ***
condition.4          7.664e-02  4.324e-03  17.724  < 2e-16 ***
condition.5          1.366e-01  6.951e-03  19.650  < 2e-16 ***
I(sqft_lot^(1/4))   -7.586e-07  2.368e-03   0.000 0.999744    
I(sqft_lot15^(1/4)) -2.976e-02  2.846e-03 -10.457  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2524 on 21589 degrees of freedom
Multiple R-squared:  0.7706,	Adjusted R-squared:  0.7704 
F-statistic:  4029 on 18 and 21589 DF,  p-value: < 2.2e-16
```

Figure 24: Summary of Model after Box-Cox transformation(Additive)



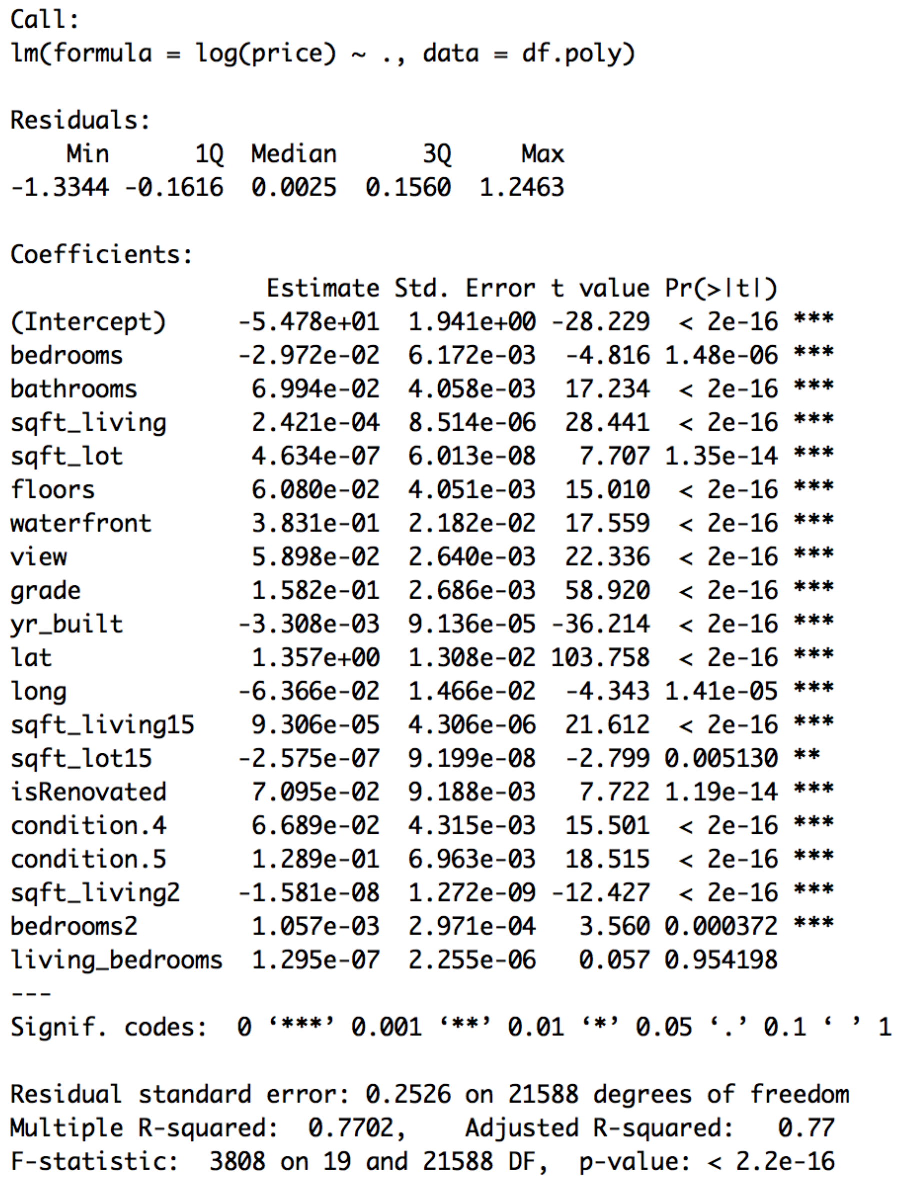
```
Call:
lm(formula = log(price) ~ . + I(sqft_lot^(1/4)) + I(sqft_lot15^(1/4)),
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.87592 -0.15772  0.00246  0.15347  1.30030 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -4.525e+01  2.016e+00 -22.443  < 2e-16 ***
bedrooms            -8.077e-03  2.373e-03  -3.404 0.000665 ***
bathrooms            6.355e-02  4.070e-03  15.614  < 2e-16 ***
sqft_living          1.482e-04  4.263e-06  34.765  < 2e-16 ***
sqft_lot             5.561e-07  9.111e-08   6.103 1.06e-09 ***
floors               3.814e-02  4.317e-03   8.836  < 2e-16 ***
waterfront           4.085e-01  2.185e-02  18.694  < 2e-16 ***
view                 5.616e-02  2.638e-03  21.289  < 2e-16 ***
grade                1.612e-01  2.671e-03  60.335  < 2e-16 ***
yr_built            -3.217e-03  9.118e-05 -35.280  < 2e-16 ***
lat                  1.335e+00  1.314e-02 101.581  < 2e-16 ***
long                 4.734e-03  1.513e-02   0.313 0.754369    
sqft_living15        1.188e-04  4.361e-06  27.244  < 2e-16 ***
sqft_lot15           1.429e-06  1.528e-07   9.352  < 2e-16 ***
isRenovated          8.521e-02  9.182e-03   9.281  < 2e-16 ***
condition.4          7.664e-02  4.324e-03  17.724  < 2e-16 ***
condition.5          1.366e-01  6.951e-03  19.650  < 2e-16 ***
I(sqft_lot^(1/4))   -7.586e-07  2.368e-03   0.000 0.999744    
I(sqft_lot15^(1/4)) -2.976e-02  2.846e-03 -10.457  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2524 on 21589 degrees of freedom
Multiple R-squared:  0.7706,	Adjusted R-squared:  0.7704 
F-statistic:  4029 on 18 and 21589 DF,  p-value: < 2.2e-16
```

Figure 25: Summary of Model after Box-Cox transformation(Additive)



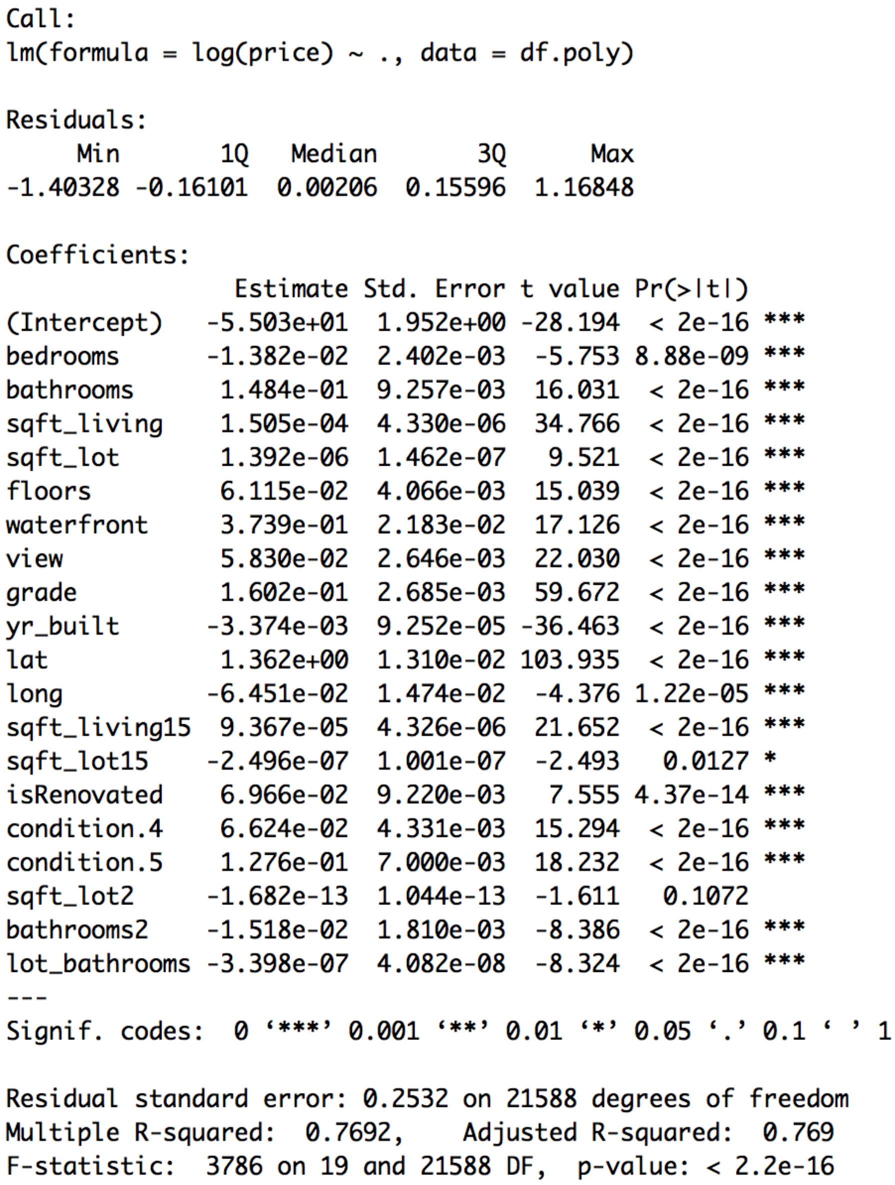
```
Call:
lm(formula = log(price) ~ . + I(sqft_lot^(1/4)) + I(sqft_lot15^(1/4)) -
    sqft_lot15 - sqft_lot, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.69140 -0.16063  0.00161  0.15592  1.31363

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -4.904e+01  2.018e+00 -24.303  < 2e-16 ***
bedrooms            -1.177e-02  2.379e-03  -4.945 7.66e-07 ***
bathrooms            6.958e-02  4.083e-03  17.042  < 2e-16 ***
sqft_living          1.435e-04  4.276e-06  33.569  < 2e-16 ***
floors               5.821e-02  4.185e-03  13.909  < 2e-16 ***
waterfront           3.796e-01  2.193e-02  17.306  < 2e-16 ***
view                 5.862e-02  2.652e-03  22.102  < 2e-16 ***
grade                1.606e-01  2.688e-03  59.732  < 2e-16 ***
yr_built            -3.222e-03  9.177e-05 -35.113  < 2e-16 ***
lat                  1.347e+00  1.321e-02 101.929  < 2e-16 ***
long                -2.016e-02  1.516e-02  -1.330    0.184
sqft_living15        1.059e-04  4.324e-06  24.504  < 2e-16 ***
isRenovated          7.983e-02  9.238e-03   8.641  < 2e-16 ***
condition.4          7.170e-02  4.343e-03  16.510  < 2e-16 ***
condition.5          1.356e-01  6.996e-03  19.378  < 2e-16 ***
I(sqft_lot^(1/4))    9.073e-03  1.575e-03   5.760 8.52e-09 ***
I(sqft_lot15^(1/4)) -1.477e-02  1.797e-03  -8.216  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2541 on 21591 degrees of freedom
Multiple R-squared:  0.7674,	Adjusted R-squared:  0.7673
F-statistic:  4453 on 16 and 21591 DF,  p-value: < 2.2e-16
```

Figure 26: Summary of Model after Box-Cox transformation



```
Call:
lm(formula = log(price) ~ ., data = df.poly)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3344 -0.1616  0.0025  0.1560  1.2463 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -5.478e+01  1.941e+00 -28.229  < 2e-16 ***
bedrooms        -2.972e-02  6.172e-03  -4.816 1.48e-06 ***
bathrooms        6.994e-02  4.058e-03  17.234  < 2e-16 ***
sqft_living      2.421e-04  8.514e-06  28.441  < 2e-16 ***
sqft_lot         4.634e-07  6.013e-08   7.707 1.35e-14 ***
floors           6.080e-02  4.051e-03  15.010  < 2e-16 ***
waterfront       3.831e-01  2.182e-02  17.559  < 2e-16 ***
view             5.898e-02  2.640e-03  22.336  < 2e-16 ***
grade            1.582e-01  2.686e-03  58.920  < 2e-16 ***
yr_built        -3.308e-03  9.136e-05 -36.214  < 2e-16 ***
lat              1.357e+00  1.308e-02 103.758  < 2e-16 ***
long            -6.366e-02  1.466e-02  -4.343 1.41e-05 ***
sqft_living15    9.306e-05  4.306e-06  21.612  < 2e-16 ***
sqft_lot15      -2.575e-07  9.199e-08  -2.799 0.005130 ** 
isRenovated      7.095e-02  9.188e-03   7.722 1.19e-14 ***
condition.4      6.689e-02  4.315e-03  15.501  < 2e-16 ***
condition.5      1.289e-01  6.963e-03  18.515  < 2e-16 ***
sqft_living2    -1.581e-08  1.272e-09 -12.427  < 2e-16 ***
bedrooms2        1.057e-03  2.971e-04   3.560 0.000372 ***
living_bedrooms  1.295e-07  2.255e-06   0.057 0.954198    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2526 on 21588 degrees of freedom
Multiple R-squared:  0.7702,	Adjusted R-squared:   0.77 
F-statistic:  3808 on 19 and 21588 DF,  p-value: < 2.2e-16
```

Figure 27: Summary of Polynomial Regression (*sqft_living* & *bedrooms*))



```
Call:
lm(formula = log(price) ~ ., data = df.poly)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.40328 -0.16101  0.00206  0.15596  1.16848 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -5.503e+01  1.952e+00 -28.194  < 2e-16 ***
bedrooms      -1.382e-02  2.402e-03  -5.753 8.88e-09 ***
bathrooms      1.484e-01  9.257e-03  16.031  < 2e-16 ***
sqft_living    1.505e-04  4.330e-06  34.766  < 2e-16 ***
sqft_lot       1.392e-06  1.462e-07   9.521  < 2e-16 ***
floors         6.115e-02  4.066e-03  15.039  < 2e-16 ***
waterfront     3.739e-01  2.183e-02  17.126  < 2e-16 ***
view           5.830e-02  2.646e-03  22.030  < 2e-16 ***
grade          1.602e-01  2.685e-03  59.672  < 2e-16 ***
yr_built      -3.374e-03  9.252e-05 -36.463  < 2e-16 ***
lat            1.362e+00  1.310e-02 103.935  < 2e-16 ***
long          -6.451e-02  1.474e-02  -4.376 1.22e-05 ***
sqft_living15  9.367e-05  4.326e-06  21.652  < 2e-16 ***
sqft_lot15    -2.496e-07  1.001e-07  -2.493   0.0127 *  
isRenovated    6.966e-02  9.220e-03   7.555 4.37e-14 ***
condition.4    6.624e-02  4.331e-03  15.294  < 2e-16 ***
condition.5    1.276e-01  7.000e-03  18.232  < 2e-16 ***
sqft_lot2     -1.682e-13  1.044e-13  -1.611   0.1072    
bathrooms2    -1.518e-02  1.810e-03  -8.386  < 2e-16 ***
lot_bathrooms -3.398e-07  4.082e-08  -8.324  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2532 on 21588 degrees of freedom
Multiple R-squared:  0.7692,	Adjusted R-squared:  0.769 
F-statistic:  3786 on 19 and 21588 DF,  p-value: < 2.2e-16
```

Figure 28: Summary of Polynomial Regression (*sqft_lot* & *bathrooms*))

6 Reference

Introduction to Linear Regression Analysis (Fifth Edition) Wiley, Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012).