F

Authors: Chiu Chu-Chuan – Farnaz Ghashami – Hang Le

STAT 628 – Applied Regression Analysis

3/19/20

Professor: Christopher Gaffney

Table of Contents

I. Introduction

The housing market is a key element of any nation’s economy. Housing expenditures (i.e. construction and renovation) increase in the gross domestic product through the two services and manufacturing sectors. They also simulate the demand for relevant industries such as household durables. The oscillation of house prices affects the value of asset portfolio for most households for whom a house is the largest single asset. Moreover, price movements influence the profitability of financial institutions and the soundness of the financial system. Recent studies further justify the necessity of housing price analysis with a conclusion that housing sector plays a significant role in acting as a leading indicator of the real sector of the economy and assets prices help forecast both inflation and output (Forni, Hallin, Lippi, and Reichlin, 2003; Stock and Watson, 2003; Das, Gupta, and Kabundi, 2009a).

A housing market can be understood as any market for properties which are negotiated either directly from their owners to buyers, or through the services of real estate brokers. People and companies are drawn to this market, which presents many profit opportunities that come from housing demands worldwide. These demands are influenced by several factors, such as demography, economy, and politics (Afonso, Melo, Dihanster, Sousa, and Berton, 2019).

By the decade of 1990, the increasing popularity of the Internet made it suitable for hosting advertisements which previously were published in newspapers and magazines. As of today, there is a huge number of property advertisements on the Web and exploiting knowledge from them is a topic which is observed in the recent ML literature (Wu and Brynjolfsson 2015, Sirignano et al. 2016). Due to the large amounts of data from these advertisements, deep learning approaches seem to perform feature extraction for housing prices prediction with good performance (Poursaeed et al. 2018).

For the context of the regression project using R, the intention was to apply the regression knowledge of R to such context and data. The initial goal for this project is to build a model that would predict the house prices in Washington State. As a result of such a goal, we would finally aim to understand what factors impact house prices in Washington State. Furthermore, the dataset, which will be elaborated on in the following section is a relatively large one found using Kaggle.

The remainder of this paper is organized as follows. In Section 2, a discussion of variables (in particular, the relation between the dependent variable and the independent variables) is presented. Section 3 details the variable selection, statistical tests, and so on. The results (validation and interpretation) are described in Section 4. Finally, in Section 5 we present the conclusion, being the limitations and possibilities of future work.

II. Data

A dataset was found through the Kaggle website (https://www.kaggle.com/shree1992/housedata/data) containing 4600 house sale prices with 18 variables in Washington state during the year of 2014. The variables of data later used for inclusion as independent variables are bedrooms (the number of bedrooms), bathrooms (the number of full bathrooms), sqft_living (the number of square feet of the living room space), sqft_lot (the number of square feet of the underlying lot), floors (the number of floors), waterfront (binary - 1 implies inclusion of waterfront and 0 implies otherwise), view, condition (discrete positive values below five, implying quality with a direct relationship with numbers), sqft_above (the number of square feet of above), sqft_basement (the number of square feet of basement - the addition of variables “sqft_above” and “sqft_basement” result in the variable “sqft_living”), yr_built (the year in which the house was built), yr_renovated (the year in which the house was renovated), street (address of the house), city, statezip (state followed by the zip code), and country (USA for all cases). The dependent variable for which we are trying to create estimations is simply the “price.”

Data Processing & Cleaning

Not all variables in the dataset were relevant enough to be used for regression estimation. The variables omitted are country (due to all items having the same country), street (due to less relevance with respect to city, and possible correlation with city if used simultaneously), statezip (due to less relevance with respect to city, and possible correlation with city if used simultaneously), view (due to ambiguity of the variable by nature), and sqft_living (due to existence of correlation with two other variables - the addition of variables “sqft_above” and “sqft_basement” result in the variable “sqft_living”).

Additionally, the outlier data for price was deleted from the dataset in order to increase the precision of the regression equation. Prices with zero were deleted from the dataset. For those cases, it was assumed that the houses were received as heritage, gift, or simply the data was nonexistent. By checking other basic features of the house such as number of bedrooms and bathrooms, we also found out that there are 2 houses without any bedrooms and bathrooms. We also delete these data point due to errors in the data entry process. After removing outliers and all data points that are not correct, our final data have 4547 rows.

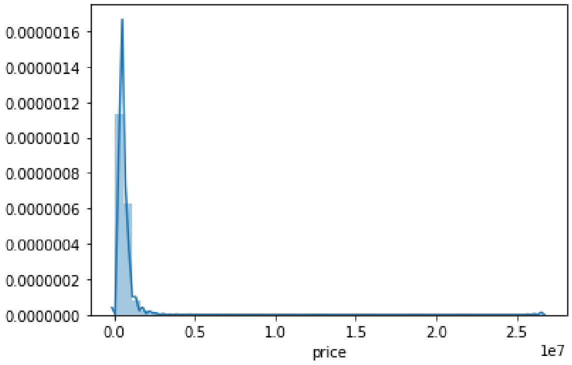
Regarding manipulations within the independent variables, the variable yr_built (showing the year in which the building was initially constructed) was not directly used as given. Instead, we used 2014 (the year in which the database was constructed) subtracted the numbers given in the variable yr_built to get the age of the building so as to use that instead. Furthermore, due to the buildings renovated not being all the buildings, we decided to create a variable showing solely if the building was renovated or not - that is, we turned yr_renovated into a binary, assigning 0 to any non-renovated building and 1 for any building that already had a renovation year. As a result, it should be noted that yr_built and yr_renovated were not used as initially given.

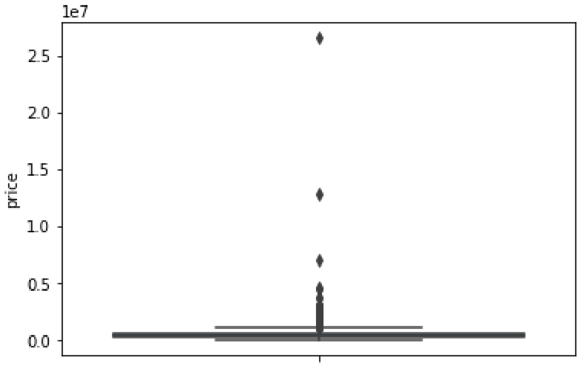
All variables later taken into account for our regression equation (after the cleanings elaborated on in the previous paragraph were implicated) were numerical. As mentioned, yr_renovated was turned into binary, and the variable waterfront was binary as well. For the city, also, a number of dummy variables equivalent to the number of cities were created. From those dummies, any city that was used had a value of 1 for that certain dummy and 0 for the rest. For instance, for any house built in Seattle, the dummy variable of Seattle was set equal to 1 and the rest of the cities' dummy variables (Shoreline, Kent, Redmont, and so on) were set equal to 0.

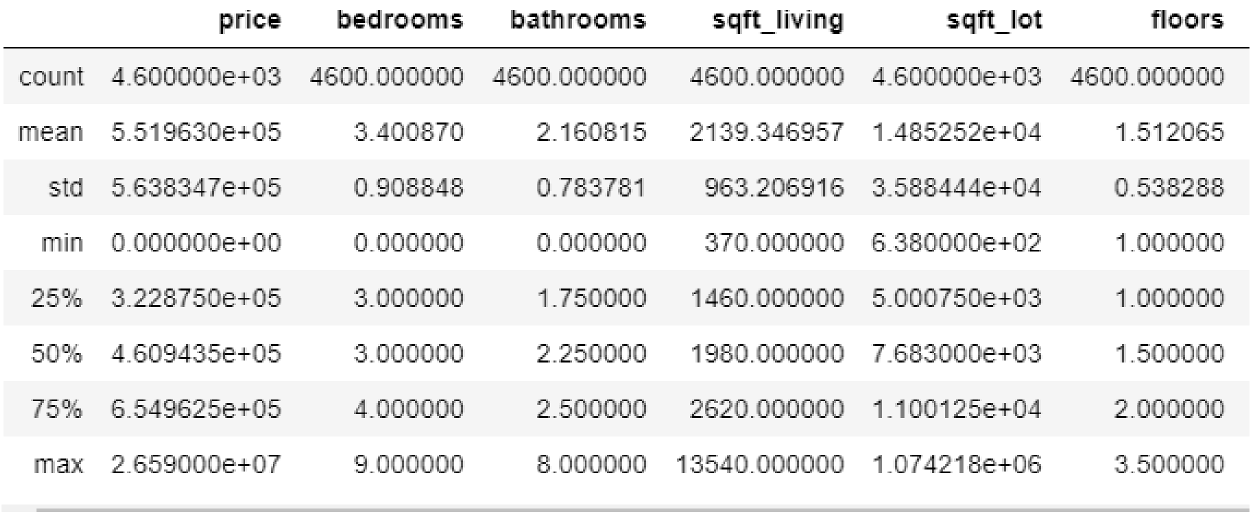
As the result of our process, we delete 8 columns in the original data: date, country, street, statezip, view, sqft_living, year_built and year_renovated. We also create 2 new variables which are Age of House and categorical variable Renovate to indicate whether the house has undergone renovation process until the time they are sold. At the end of the data cleaning process, our final data has 4547 rows and 12 variables. We created the binary variables of the names of the cities for further analyses and ended up 55 columns.

III. Analysis

Price distribution: We first look at the distribution of our dependent variable: house's price. By looking at the histogram of price, we can see that the distribution is skewed right with long right tail. That tells us there are some houses with extremely high prices in the data set. We next examine outliers by looking at the boxplot of price and some basic statistics.







	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors
count	4.600000e+03	4600.000000	4600.000000	4600.000000	4.600000e+03	4600.000000
mean	5.519630e+05	3.400870	2.160815	2139.346957	1.485252e+04	1.512065
std	5.638347e+05	0.908848	0.783781	963.206916	3.588444e+04	0.538288
min	0.000000e+00	0.000000	0.000000	370.000000	6.380000e+02	1.000000
25%	3.228750e+05	3.000000	1.750000	1460.000000	5.000750e+03	1.000000
50%	4.609435e+05	3.000000	2.250000	1980.000000	7.683000e+03	1.500000
75%	6.549625e+05	4.000000	2.500000	2620.000000	1.100125e+04	2.000000
max	2.659000e+07	9.000000	8.000000	13540.000000	1.074218e+06	3.500000

As we can see from the box plot of price, there are three houses with extremely high price: 26.5, 12.9 and 7.0 USD millions. We examined houses with prices of 26.5 and 12.9 USD millions respectively and found out that these houses have 3 bedrooms with 2 and 2.5 bathrooms. The living square feet are 1180 and 1580 which are below mean square feet of living space in our data set. These two houses were built in 1992 and 1956 which are also considered older than the majority of houses in our data. Locations are Kent city and Seattle where the mean price of houses are $439,492 and $579,837 respectively. These examinations confirm our thoughts that these extremely high prices are due to mistakes in data entry and we choose to delete these records in our further analysis. The minimum price is 0 and there are 49 houses having price as 0 in our data set. We consider removing these houses as discussed in the previous part. The mean value of price in the data is $551,963 USD with standard deviation of $563,834 so we conclude that our data is widely spread from the mean value. After removing outliers and houses with price = 0, new mean value for price is $549,186 with standard deviation of $368,056.

Correlation among variables:

In order to investigate the association among all the variables, we ran a correlation test and have results as figure 1. From the heatmap, we can learn what variables are highly related to each other and if they should be further examined to include or exclude in the model.

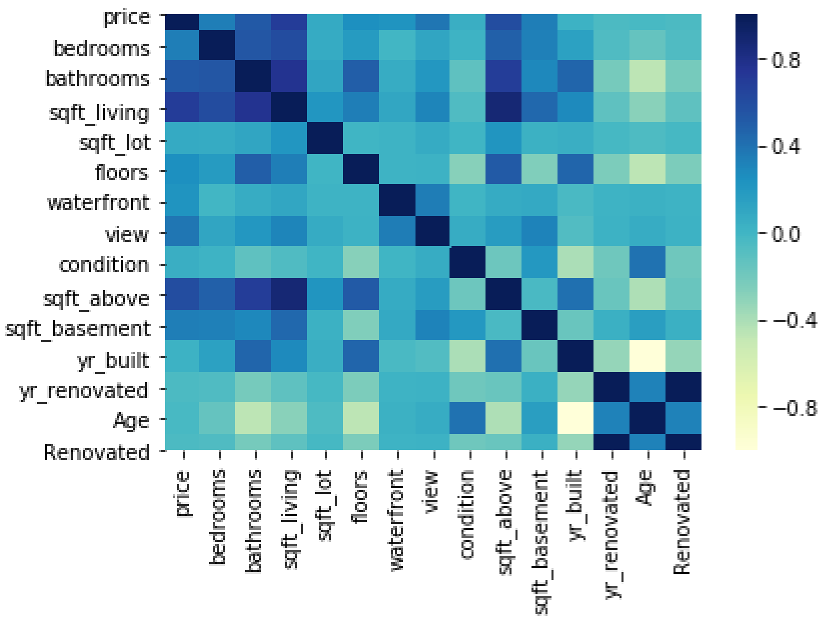
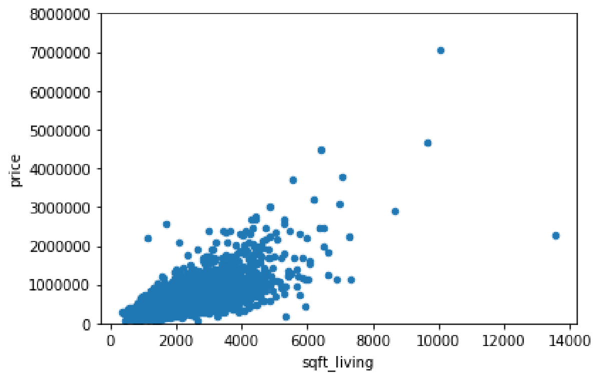


Figure 1 Correlations among Variables

Dependent variable “price” is positively correlated with “sqft_living” (0.70, highest), “sqft_above” (0.60, second), and “bathrooms” (0.53, third). The data suggest the more space inside the house, the higher the price. From the three Figures 2 and 3, we can tell the mean prices increase as the bedroom or living square feet increase.



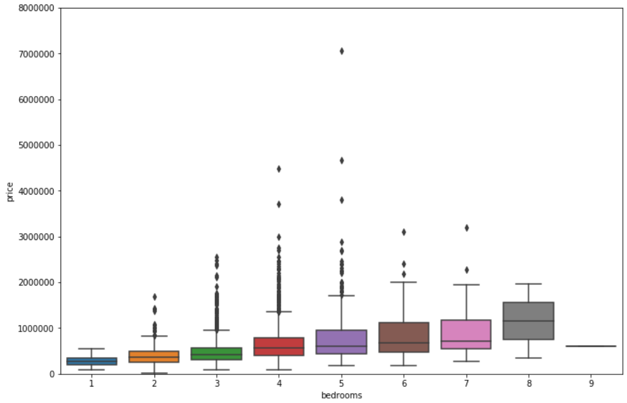


Figure 2 Price v.s. sqft_living **Figure 3 Price v.s. Number of Bedrooms**

However, the more floors do not reflect higher prices because we found the mean price of 2.5 floors could be higher than the mean of three floors house. See Figure 4.

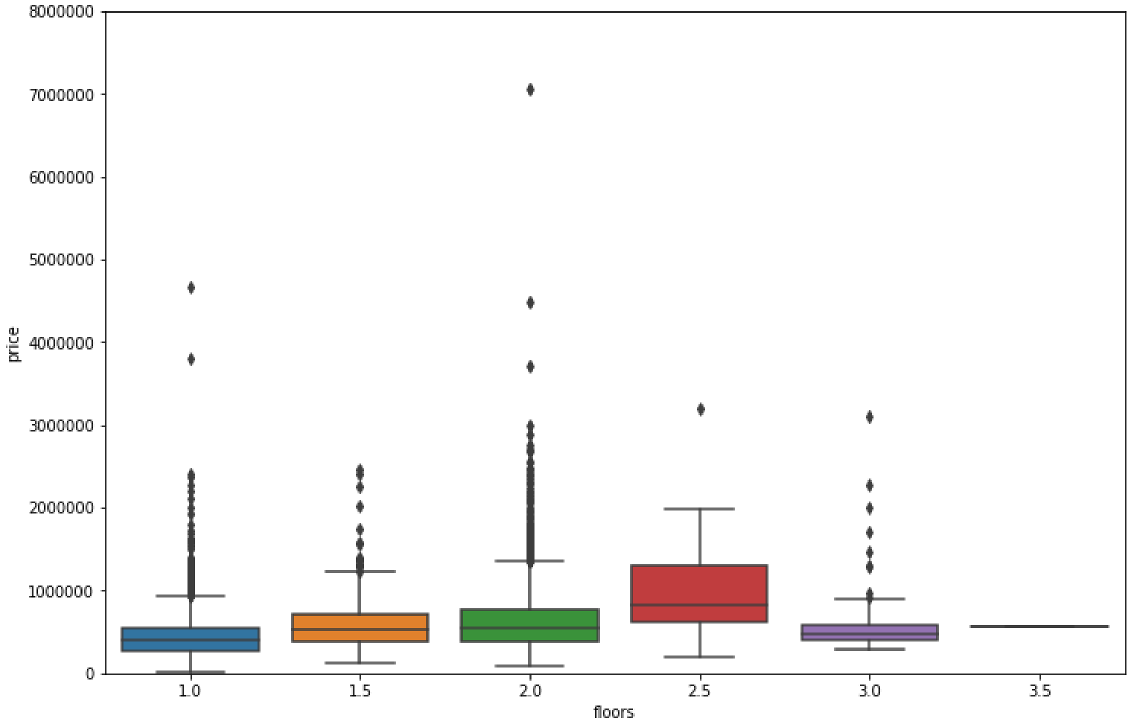


Figure 4 Price v.s. Floors

We also explored the bathroom since the bathroom also has a high positive relation with the price. Based on Figure 5, we could see a trend that as the number of bedrooms increases, the mean price increases, but not all of the cases. For example, the mean price of the 5 bathrooms is smaller than 4.5 bathrooms. Next, we look at condition to see its relationship with prices. Condition is a categorical variable that presents the quality of the houses. Conditions 1 and 2 have mean prices relatively lower; condition 5 have the highest mean of the prices. We also have the variable called "waterfront" as a binary variable to inform if the house has a waterfront view or not. From Figure 7, we can see the mean price of those that have waterfront is much higher than those that do not have waterfront.

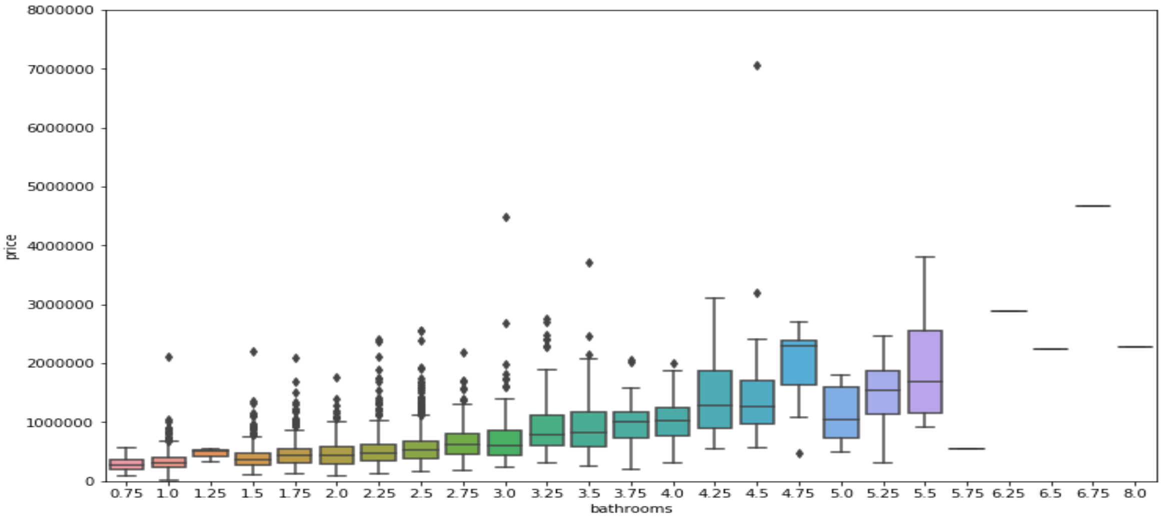
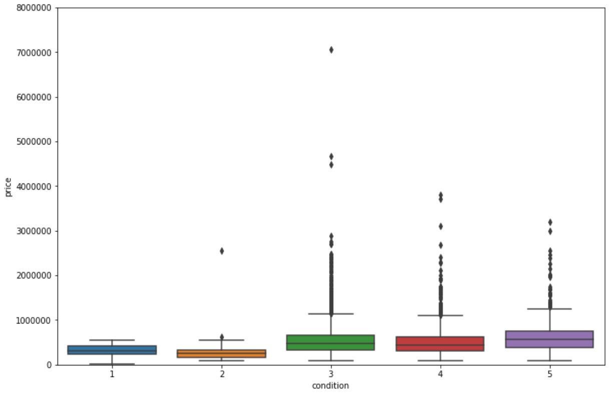


Figure 5 Price v.s. Number of Bathrooms



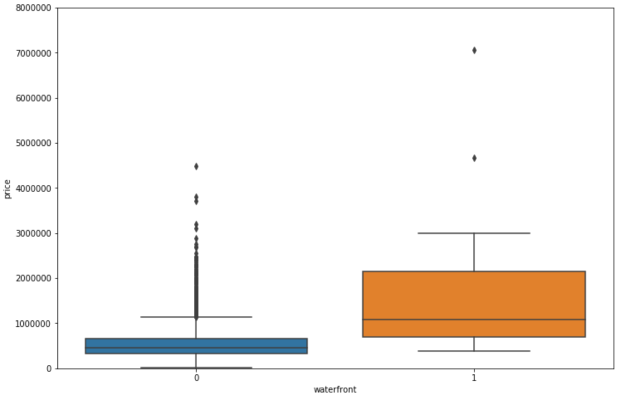


Figure 6 Price v.s. Conditions **Figure 7 Price v.s. Waterfront**

The independent variables that have negatively correlated with “price” are “Age” (-0.03) and “ Renovated” (- 0.04). For variable “Age” is easier to understand. The older the house, the lower the price. From Figure 8, we could see the price distributions among different ages and the relation with price. Figure 9 tells us that the majority of houses in our data set is from 0 to 20 year olds.

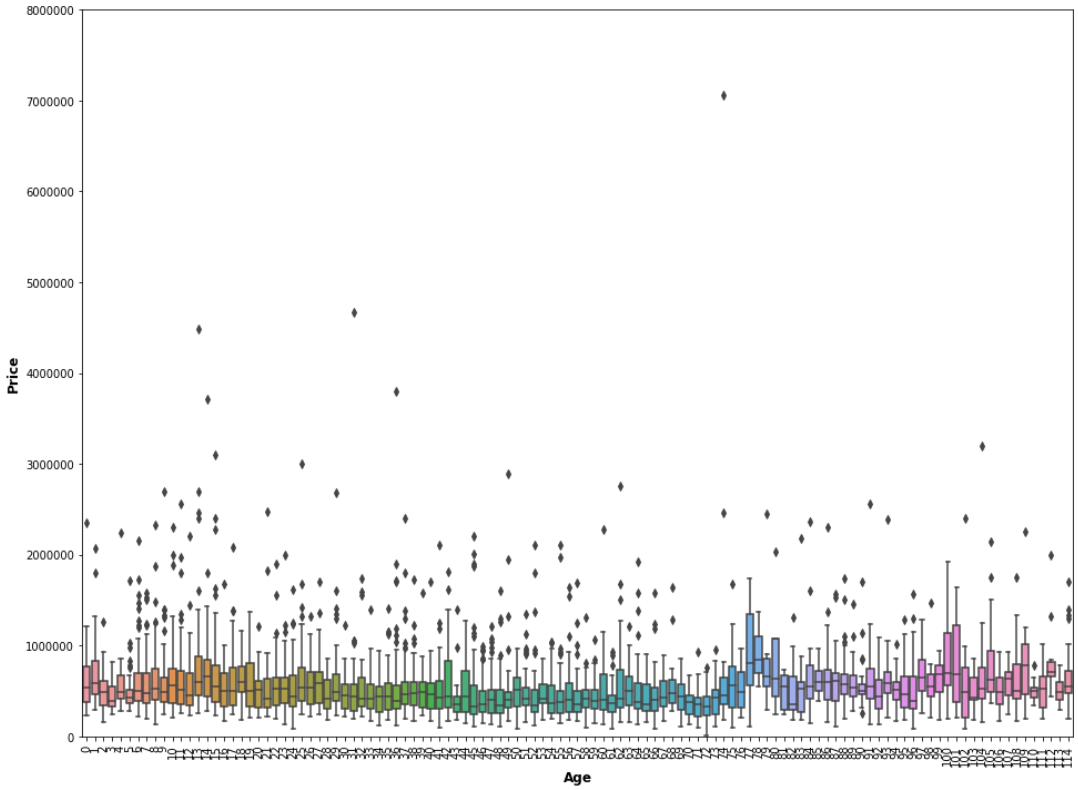


Figure 8 Price v.s. Age

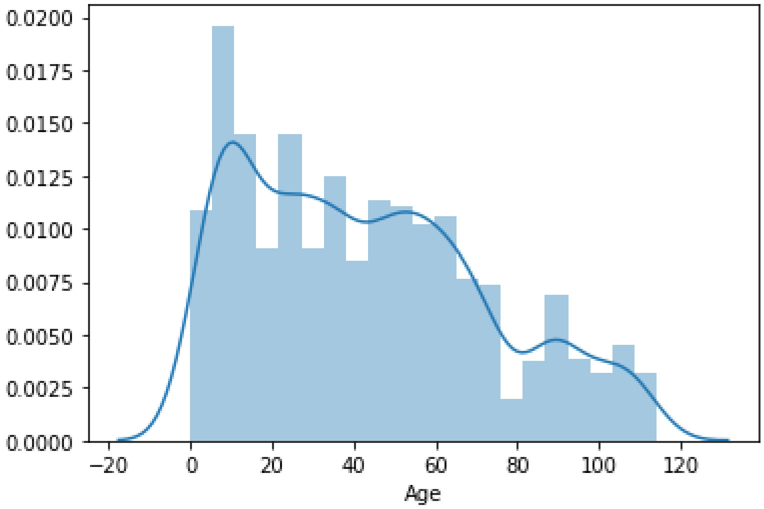


Figure 9 House's Age Distribution

For the variable "Renovated", the negative association implies that even though the house has been renovated at some point, the house renovation could already signal the poor condition of the house that could lead to decrease the price of the house. Figure 10 helps explain the above statement to reveal that the mean age of the renovated house is higher than the mean age of the non-renovated house.

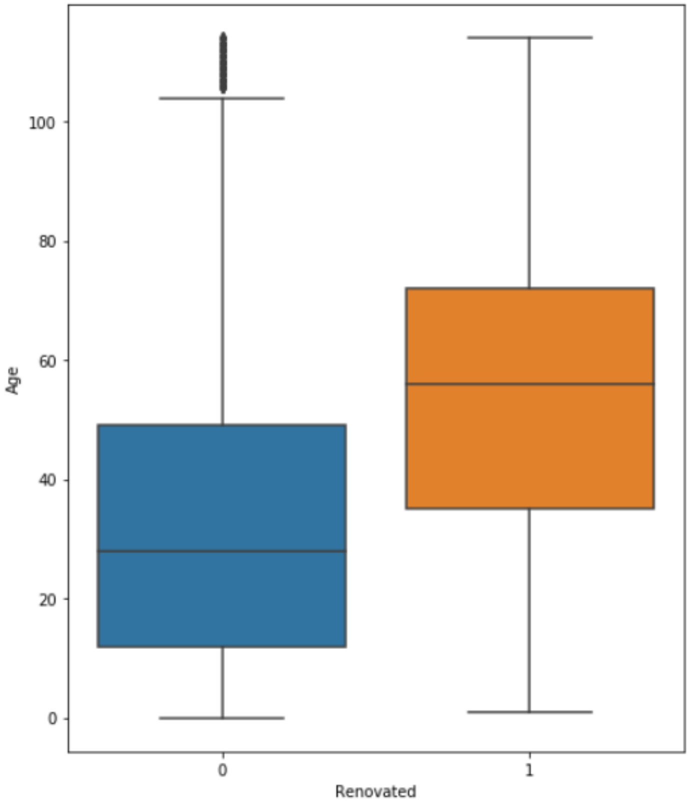


Figure 10 Mean of Age by. Renovated variable

In the house market, location matters. To understand the distribution of our data, we created a frequency chart to illustrate the counts of each city. We have 44 different cities and 77 different zip codes. Figure 11 illustrates the first 15 cities in our dataset. There are 1,573 (43%) of prices from Seattle. We tried to convert our address data to county by US Census website but could not find an efficient tool to deal with our large data. Due to the limited time, we chose to remain the city and zip code as factors.

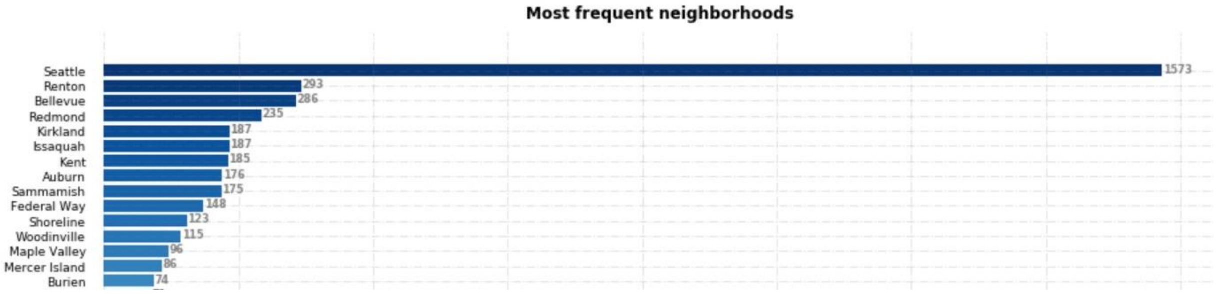


Figure 11 City Counts in Dataset

Most of the houses have a mean price under one million. From the boxplot Figure 12, we can see the means of prices in some cities are over two million such as Medina with a mean of 2,046,559. Also, some cities have larger variance in price than other cities.

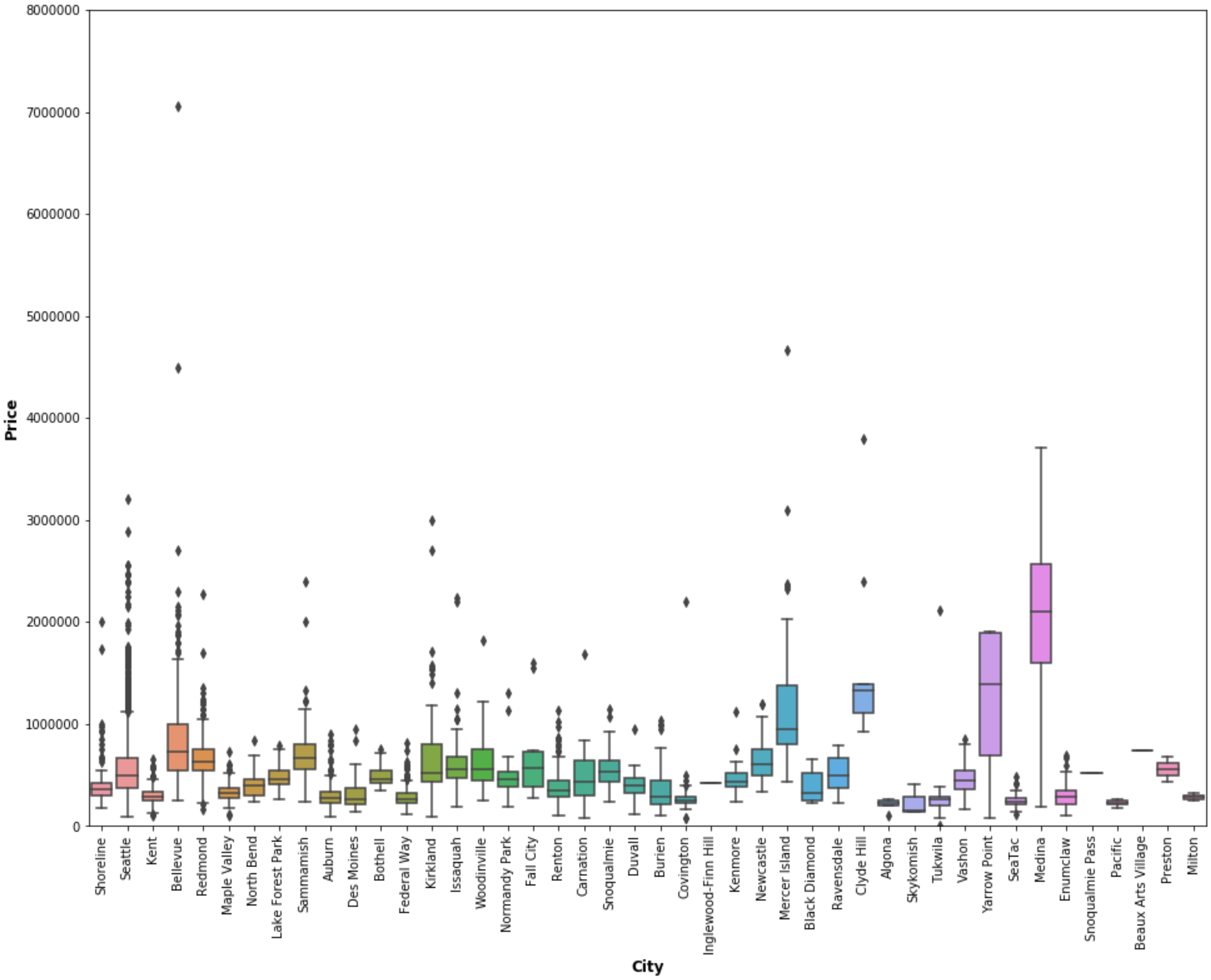


Figure 12 Mean Prices by City

Model building

To avoid overfitting data to our prediction models, we first divide our data set into training and testing sets. Our training data will account for 75% of the total data set and the rest will be used to test our models later. As a result, we have 3410 rows in training and 1137 rows in our testing data. We first fit the first-order model with only 4 variables: Bedrooms, Bathrooms, Condition and Waterfront, which we think it's significant to predict house price. The results of this analysis are shown in Table 1 in the Appendix. Below is a summary of the results.

- All variables included in the model have positive relationships with dependent variable price. An increase in one bedroom will increase the mean price of a house's price $27,092. If the house has a waterfront view, its mean price will increase $899,841.
- The first-order model with only 4 variables can explain 33% of variance in our data set.

As the next step we try to include all variables in our model. The results of this regression are shown in the Appendix in Table 2. The findings are summarized below.

- As we expected in our data exploratory process, significant variables are number of bedrooms, number of bathrooms, Age, square feet above and basement, condition of the house, waterfront view and square feet of lots. Variable Renovated is significant at 10 % level. Locations which are represented by the city of the house seem to be significant in some cases. These cases include Medina, Mercer Island and Cycle Hill city. These cities are top 4 cities with the highest mean price in our data set. As we discovered in the previous part, Medina city has the highest mean value for price ($2.01 millions).
- Our R square is 0.6953 which means 69.52% variance in the data set was explained by using all predictor variables.

Then, we diagnose fit of the model by looking at plots between standardized residuals and fitted values and other predictor variables. The full results of this analysis are shown in Figure 13 through Figure 17 in the Appendix. The findings are summarized below.

- The residual doesn't have constant variance. The errors seem to increase when square feet above increase.
- Data still has many outliers.
- Normal QQ plot suggests the distribution of error has long tails.
- This model suffers from singularity when we include all city dummy variables.

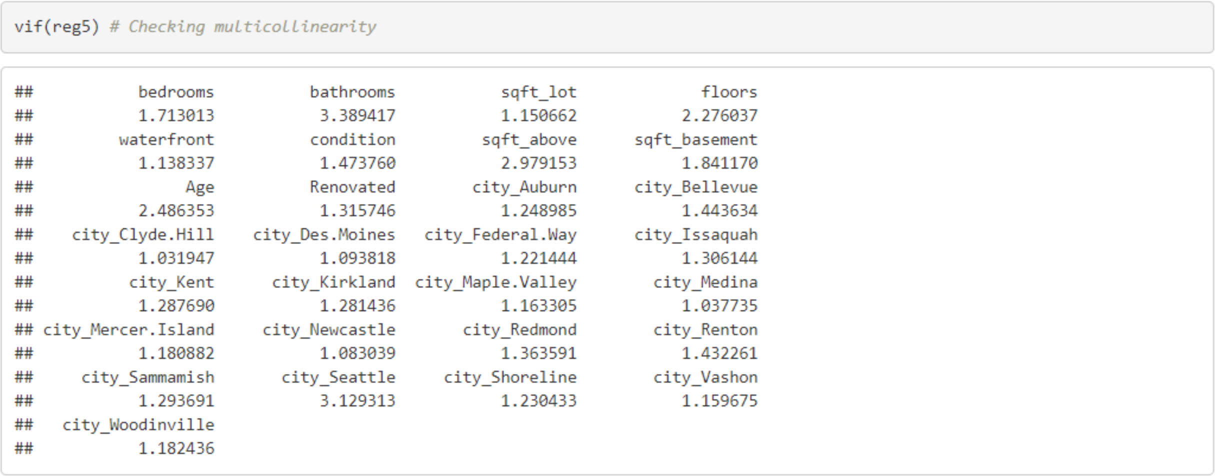
From model diagnostics, we can see that by including all city dummy variables, our model is suffering from singularity. Therefore, we try to build Model 3 without the variable city. The results of this analysis are shown in Table 3 in the Appendix. As a result, our R-square decreases from 0.69 to 0.57 and RMSE from the test data set increases 16%, from 198,388 to 230,168 as we exclude the variable city from our model. However, the result is not so surprising. From our industry research, we also noticed that the location of the house is critical to determine house price. For example, Medina city is the neighborhood for billionaires like Jeff Bezos and Bill Gate. That's the reason why houses in this city are extremely more expensive compared to other neighborhoods.

As our goal is building a model as simple as possible but still can capture as much as information from our data, we formulate a new variable called **Residential** indicating whether the house is located in the neighborhood which has a mean house price over $1 million. These includes: Mercer Island, Medina city, Yarrow Point city and Clyde Hill. Our Model 4 will include all other variables and new variable Residential but without variable city. The results of this analysis are shown in Table 4 in the Appendix. As we expected, model 4 performs better than Model 3 which doesn't have any information about the house's location. R-squared slightly increases from .57 to .60 and RMSE on test data also decrease 4%. However, compared to our model with all cities (model 2), model 4 performance is not as good as this full model. RMSE from Model 4 is 11% higher than RMSE from model 2 when testing on validation data. Next, we consider selecting variables using AIC criteria. The results of this regression are shown in the Appendix in Table 5. We start with the model without any variables and then use both backward and forward direction to choose the subset of variables that gives us the lowest AIC. By performing variable selection, we have a new model (Model 5) with only 29 variables but still keep similar R-square. This reduced model also performs well on our test data: RMSE only increases 0.27% compared to our full model.

Our next step is checking multicollinearity among variables in the reduced model. Our initial analysis shows us some potential multicollinearity between:

1. Variable bedrooms and sqft_above: since the more bedroom, the larger the house.
2. Variable bathrooms and floors: Houses with more floors usually have more bathrooms.

We use Variance Inflation Factor (VIF) criterion to identify variables with potential multicollinearity problems. We use the threshold of 10 to decide whether there are any serious multicollinearity issues. Although bathrooms and sqft_above have high VIFs compared to other models, these VIFs are under our threshold 10. Therefore, although there are some variables that might correlate with others in our data set, there is no serious multicollinearity issue when including all these variables in our final model.



```
vif(reg5) # Checking multicollinearity
```

```
##           bedrooms          bathrooms           sqft_lot             floors
##           1.713013           3.389417           1.150662           2.276037
##         waterfront          condition         sqft_above      sqft_basement
##           1.138337           1.473760           2.979153           1.841170
##                Age          Renovated        city_Auburn      city_Bellevue
##           2.486353           1.315746           1.248985           1.443634
##    city_Clyde.Hill    city_Des.Moines   city_Federal.Way      city_Issaquah
##           1.031947           1.093818           1.221444           1.306144
##          city_Kent      city_Kirkland  city_Maple.Valley        city_Medina
##           1.287690           1.281436           1.163305           1.037735
## city_Mercer.Island     city_Newcastle       city_Redmond        city_Renton
##           1.180882           1.083039           1.363591           1.432261
##     city_Sammamish       city_Seattle     city_Shoreline        city_Vashon
##           1.293691           3.129313           1.230433           1.159675
##   city_Woodinville
##           1.182436
```

From our model diagnostics, we try to transform our price variable to fix the problem of non-constant variance in our error term. We tried Box-Cox transformation to find our lambda value for the transformation but the value of lambda is close to zero (The results of this analysis are shown in Figure 18 in the Appendix).

However, in the previous step, we found that price has a right-skewed distribution, so we transformed our price variable by using natural log. When we transformed price by using log, more variance now can be explained by using all predictor variables. Then, we diagnose fit of the model by looking at plots between standardized residuals and fitted values and other predictor variables. The full results of this analysis are shown in Figure 19 through Figure 21 in the Appendix. From the residual plot of model 2, the error terms still don’t have constant variance. However, by using log of price, we can reduce outliers in our data set. Model 6 with the transformed log price variable also performs better compared to our previous model. 72.67% variance in our data can be explained using the log price model (Model 6) and RSME on our test data also decreases approximately 5.6%. We then perform model selection using AIC criterion and stepwise procedure to choose the subset of our models with price transformation. Interestingly, the reduced model (Model 8) with only 40 variables has better predictions on our validation data even though it has lower R-squared compared to the full model. This suggests that the reduced model is a better option to go with because although the full model has better performance on training data, it might suffer from overfitting problems and can’t predict as well as the reduced model. The full results of this analysis are shown in Table 6 through Table 10 in the Appendix.

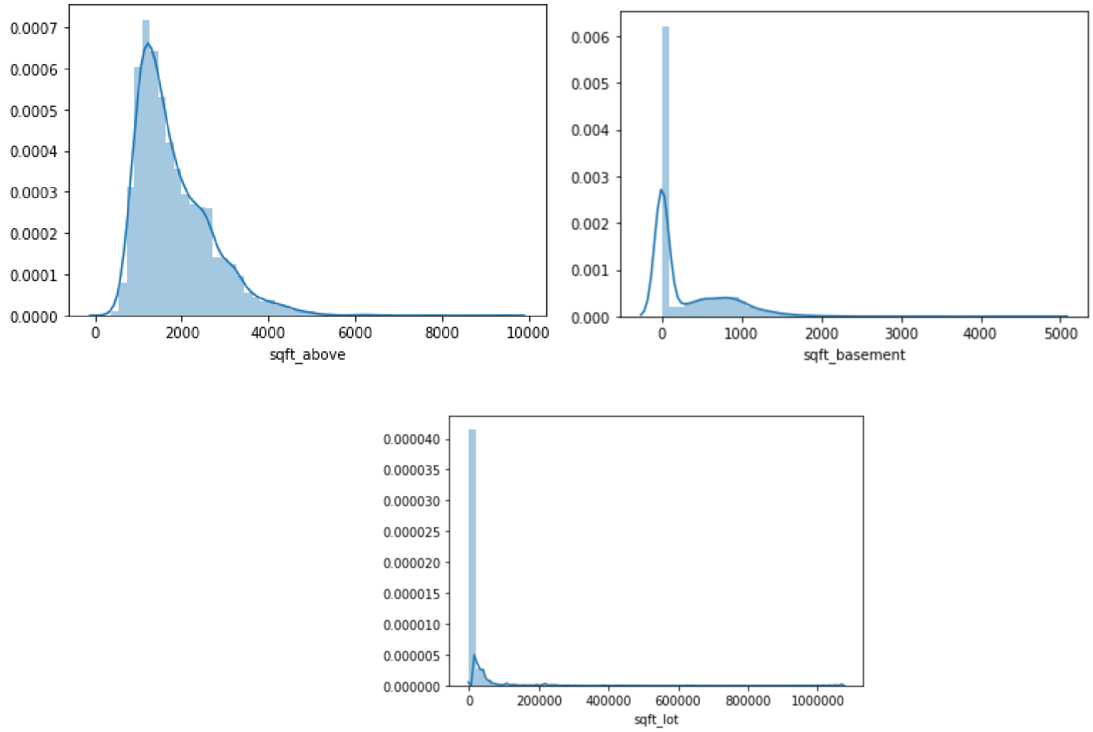


Figure 13 Distribution of Sqft_above, sqft_Basement and Sqft_lot

Next, in our previous data exploratory process, we also see that square feet above, square feet for basement, square feet for lots also suffer from right-skewed distribution. The transformation on these variables might help to bring constant variance for error terms. As our reduced model from the previous step doesn't include square feet for lots, we only performed log transformation for square feet above (we chose to perform only transformation on sqft_above because there was error arising with log transformation for both sqft_above and sqft_basement). By transforming square feet for the above area, our model could perform better than the model without log transformation on this variable. RMSE on test data also decreased .85% compared to our previous model, indicating the better performance on predicting new data.

Final Model

Model Comparison:

Variables	R-Square (Train Data)	RMSE (Test data)
Model 1: Bedrooms, Bathrooms, Condition and Waterfront	**0.3313**	**281,267.2**
Model 2: Price ~ all variables (54 variables)	**0.6952**	**198,388.9**
Model 3: Price ~ all variables but city	**0.5741**	**230,168.9**
Model 4: Price ~ all variables but city + new variable residential	**0.6004**	**220,536.6**
Model 5: Model selection using AIC from model 2 (29 variables)	**0.6944**	**198,936.6**
Model 6: Log(Price) ~ all variables	**0.7267**	**187,145.8**
Model 7: Log(Price) ~ all variables but city + new variable residential	**0.5420**	**222,406.6**
Model 8: Model selection using AIC from model 6 (40 variables)	**0.7262**	**186,484.1**
Model 9: Log(Price) ~ other variables + log(sqft_lot) + log(sqft_above)	**0.5541**	**216,771.4**
Model 10: Log(Price) ~ All variables in model 8 with transformation log (sqft_above)	**0.7425**	**184,885.8**

IV. Results

Model Interpretation

Our final model is the Model 10 including 40 variables in which natural log transformation on our Price and sqft_above were performed. The results of this regression are shown in Table 10 in Appendix.

Our final model can be interpreted as follow:

- In our final model, all of the variables are significant except city_Tukwila. Variable Renovated, city_Inglrwood.Finn.Hill and city_Algona are significant at 10 % level.

The data set that we had doesn’t have much information about renovation. Only information regarding the year that the house was renovated was available. If we have more information about renovation such as which parts of the house or the value of total renovation, we can use this variable more efficiently.

- All variables included in the model have positive relationships with dependent variable price except the number of bedrooms, city_Algona and city_Tukwila. As discussed in the previous part, the variable bedroom has positive impacts on predicting house’s prices. However, when including both bedroom and square feet for living areas in the model, the negative impact started due to interaction between two variables.
- The comparison among the coefficient estimates of the different cities can suggest the impact of these different cities on housing prices. For example, if the house located in city_Bellevue (7.619e-01) will have an increase as twice as the house located in city_Shoreline (3.896e-01) based on their coefficient estimates.
- The impact of all predictor variables in the final model can be interpreted by their regression coefficient on the percentage increase/decrease with price. For example, 1% increase in square feet of living area above will lead to an 0.73% increase in the mean of the house price. Similarly, if the house has a waterfront view, there will be an 0.53% increase to the mean of predicted house’s price. If the house has one more bathroom, there will be an increase of 0.03% in the mean price of the house. By estimating the relationship among predictor variables through percentage increase in the dependent variable price, it is also easier to interpret than using dollar values.
- R square is 0.7425 which means 74.25% variance in the data set was explained by using all variables in model 10 with transformation log (sqft_above).

V. Conclusion

The preceding analysis has shown that the most important factors based on the regression coefficient are locations, sqft_above, and waterfront. In the first portion of our analysis, we were able to show that the logarithmic dependent variable gave us more accurate results, so we transformed our price variable by using natural log. We were also able to identify that sqft_basement and age have significant positive relations with the housing prices but not weighted as much as other variables like sqft_above or locations.

Limitations

The final model is built under some limitations and might affect the application of the model and its interpretation. Our variables selection and model building rely on our industry research, data exploration, and some statistical tests such as VIF. We did not get to explore other methods such as Lasso and Ridge Regression due to the time constraint. In addition, our knowledge of real estate is based on our research and understanding. There might be some external variables worth including when building the models such as median household income for the city variable. Or, some combined variables like counties or metropolitan areas might be good potential predictor variables for predicting housing prices. We also understand incorporating local realtors' inputs on variable selections could provide qualitative evaluations to our model, but we did not get to do it. Last but not least, our predictive model is built without considering any factors of the economic environment such as inflation and employment rate, which might relate to housing prices as well.

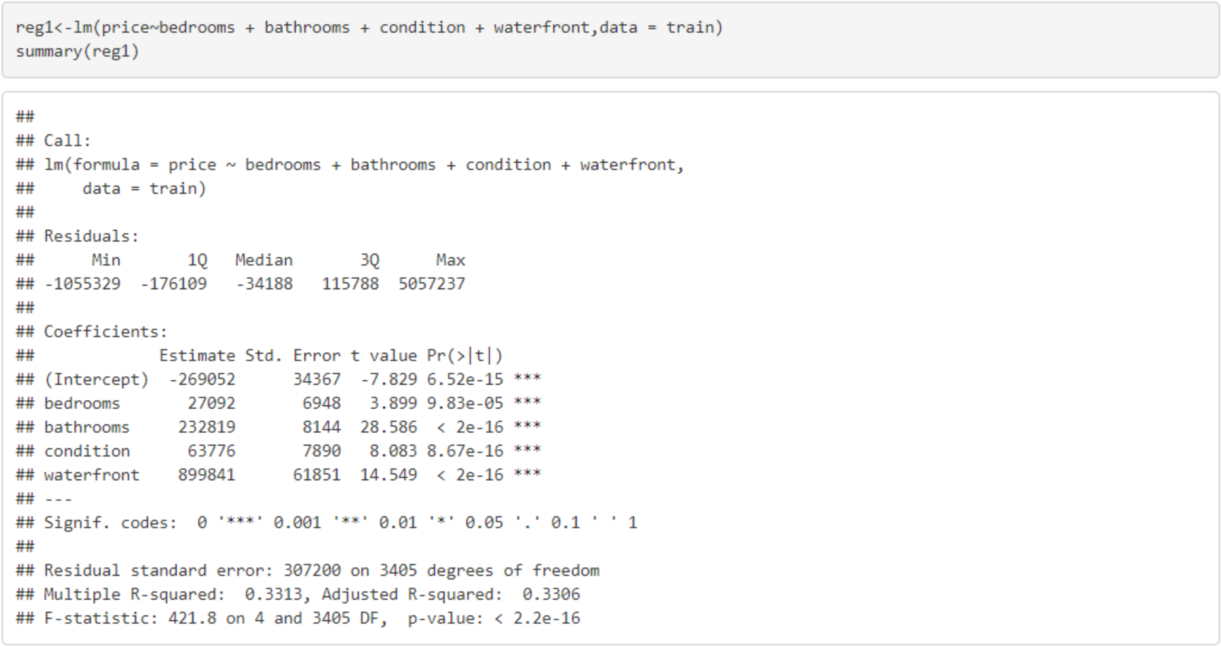
Future Study

There are several possible extensions of our work for researchers to pursue in the future:

First, there are several additional years of data available, so the analysis can be extended to cover a longer time period. Second, researchers can consider the effect of recession on the house price in Washington by creating three different scenarios (Mild, Moderate, and Severe). Also, researchers can investigate a similar question using other variables such as Lot shape, Utilities or Sale condition to predict house price in Washington.

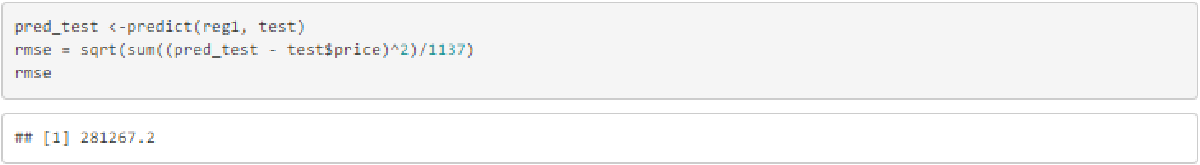
Appendix

Table 1: Regression results - Model with 4 variables: Bedrooms, Bathrooms, Condition and Waterfront



```
reg1<-lm(price~bedrooms + bathrooms + condition + waterfront,data = train)
summary(reg1)
```

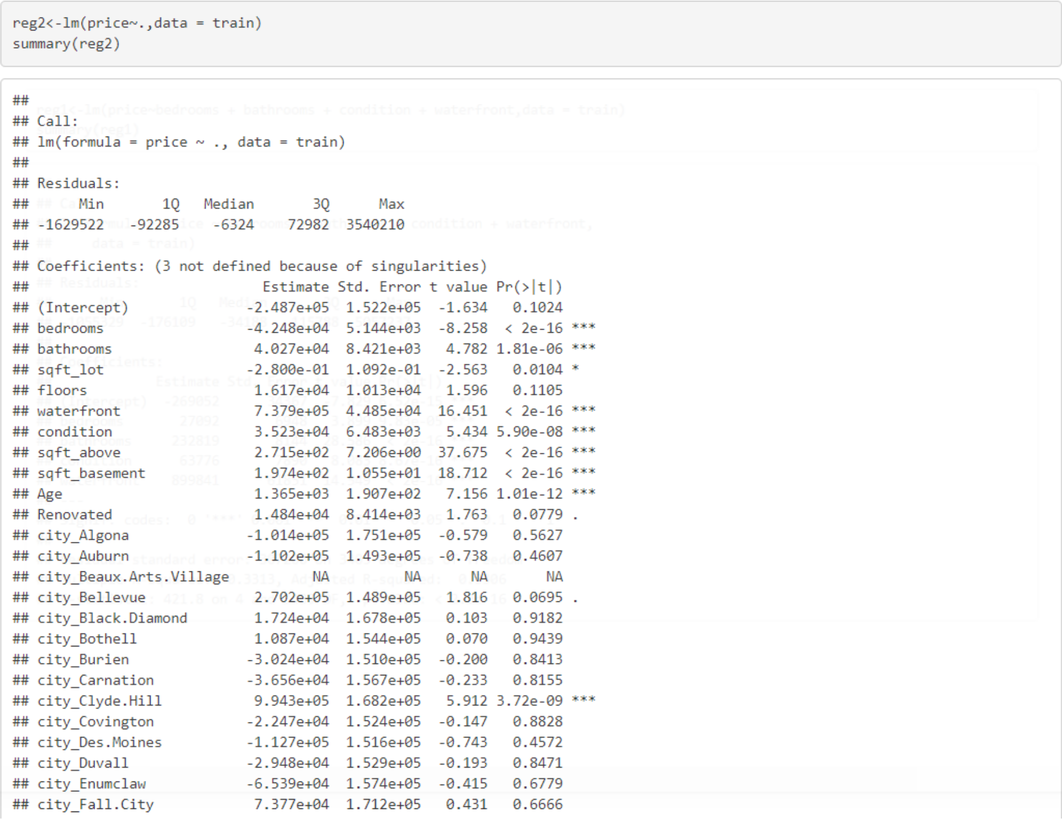
```
##
## Call:
## lm(formula = price ~ bedrooms + bathrooms + condition + waterfront,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1055329  -176109   -34188   115788  5057237
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -269052      34367  -7.829 6.52e-15 ***
## bedrooms       27092       6948   3.899 9.83e-05 ***
## bathrooms     232819       8144  28.586  < 2e-16 ***
## condition      63776       7890   8.083 8.67e-16 ***
## waterfront    899841      61851  14.549  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 307200 on 3405 degrees of freedom
## Multiple R-squared:  0.3313, Adjusted R-squared:  0.3306
## F-statistic: 421.8 on 4 and 3405 DF,  p-value: < 2.2e-16
```



```
pred_test <-predict(reg1, test)
rmse = sqrt(sum((pred_test - test$price)^2)/1137)
rmse
```

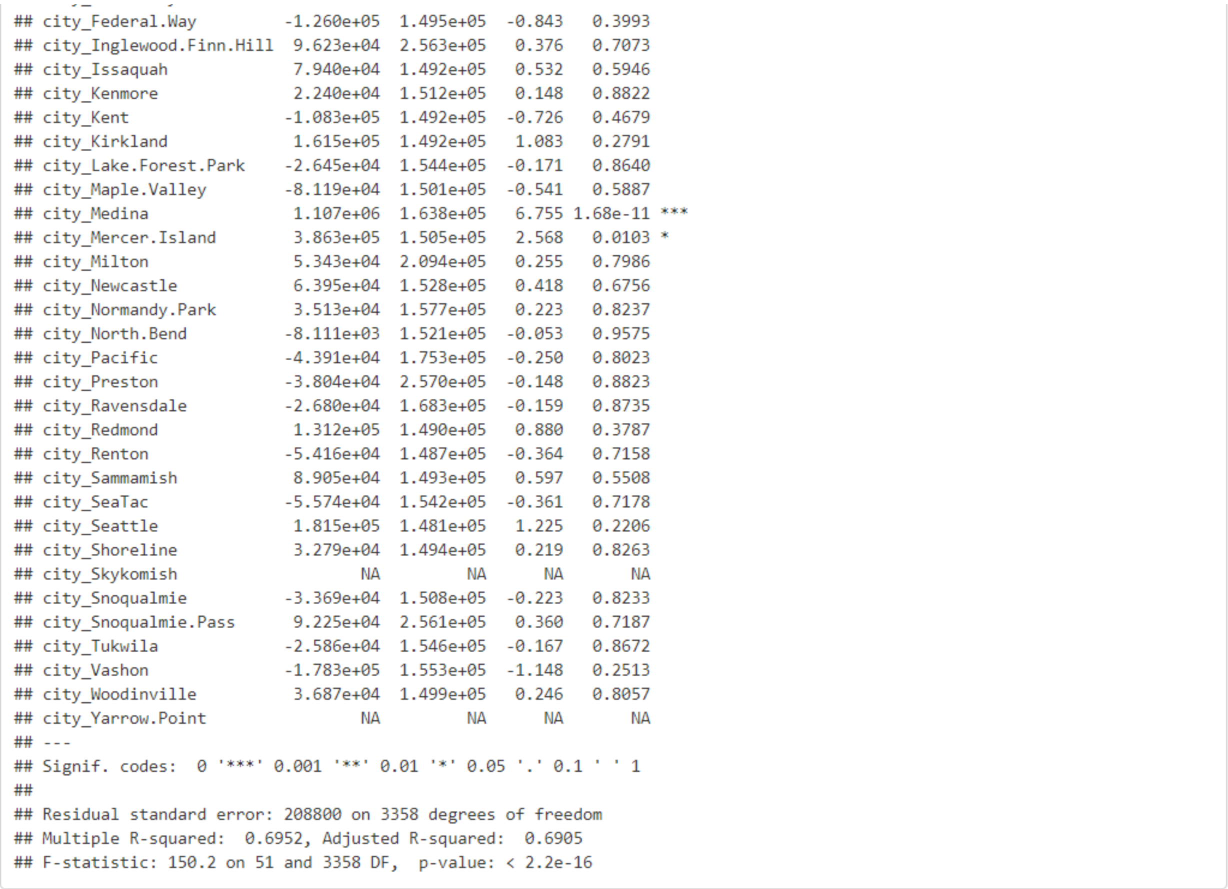
```
## [1] 281267.2
```

Table 2: Regression results - Model with all variables

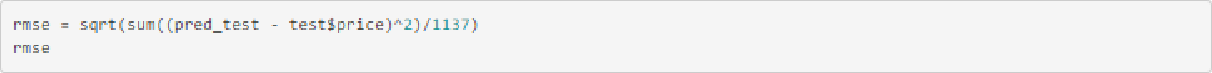


```
reg2<-lm(price~.,data = train)
summary(reg2)
```

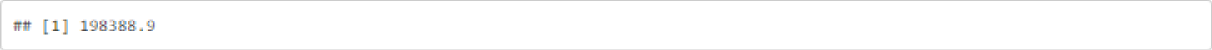
```
##
## Call:
## lm(formula = price ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1629522   -92285    -6324    72982  3540210
##
## Coefficients: (3 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -2.487e+05  1.522e+05  -1.634   0.1024
## bedrooms               -4.248e+04  5.144e+03  -8.258  < 2e-16 ***
## bathrooms               4.027e+04  8.421e+03   4.782 1.81e-06 ***
## sqft_lot               -2.800e-01  1.092e-01  -2.563   0.0104 *
## floors                  1.617e+04  1.013e+04   1.596   0.1105
## waterfront              7.379e+05  4.485e+04  16.451  < 2e-16 ***
## condition               3.523e+04  6.483e+03   5.434 5.90e-08 ***
## sqft_above              2.715e+02  7.206e+00  37.675  < 2e-16 ***
## sqft_basement           1.974e+02  1.055e+01  18.712  < 2e-16 ***
## Age                     1.365e+03  1.907e+02   7.156 1.01e-12 ***
## Renovated               1.484e+04  8.414e+03   1.763   0.0779 .
## city_Algona            -1.014e+05  1.751e+05  -0.579   0.5627
## city_Auburn            -1.102e+05  1.493e+05  -0.738   0.4607
## city_Beaux.Arts.Village        NA         NA      NA       NA
## city_Bellevue           2.702e+05  1.489e+05   1.816   0.0695 .
## city_Black.Diamond      1.724e+04  1.678e+05   0.103   0.9182
## city_Bothell            1.087e+04  1.544e+05   0.070   0.9439
## city_Burien            -3.024e+04  1.510e+05  -0.200   0.8413
## city_Carnation         -3.656e+04  1.567e+05  -0.233   0.8155
## city_Clyde.Hill         9.943e+05  1.682e+05   5.912 3.72e-09 ***
## city_Covington         -2.247e+04  1.524e+05  -0.147   0.8828
## city_Des.Moines        -1.127e+05  1.516e+05  -0.743   0.4572
## city_Duvall            -2.948e+04  1.529e+05  -0.193   0.8471
## city_Enumclaw          -6.539e+04  1.574e+05  -0.415   0.6779
## city_Fall.City          7.377e+04  1.712e+05   0.431   0.6666
```



```
## city_Federal.Way         -1.260e+05  1.495e+05  -0.843   0.3993
## city_Inglewood.Finn.Hill  9.623e+04  2.563e+05   0.376   0.7073
## city_Issaquah             7.940e+04  1.492e+05   0.532   0.5946
## city_Kenmore              2.240e+04  1.512e+05   0.148   0.8822
## city_Kent                -1.083e+05  1.492e+05  -0.726   0.4679
## city_Kirkland             1.615e+05  1.492e+05   1.083   0.2791
## city_Lake.Forest.Park    -2.645e+04  1.544e+05  -0.171   0.8640
## city_Maple.Valley        -8.119e+04  1.501e+05  -0.541   0.5887
## city_Medina               1.107e+06  1.638e+05   6.755 1.68e-11 ***
## city_Mercer.Island        3.863e+05  1.505e+05   2.568   0.0103 *
## city_Milton               5.343e+04  2.094e+05   0.255   0.7986
## city_Newcastle            6.395e+04  1.528e+05   0.418   0.6756
## city_Normandy.Park        3.513e+04  1.577e+05   0.223   0.8237
## city_North.Bend          -8.111e+03  1.521e+05  -0.053   0.9575
## city_Pacific             -4.391e+04  1.753e+05  -0.250   0.8023
## city_Preston             -3.804e+04  2.570e+05  -0.148   0.8823
## city_Ravensdale          -2.680e+04  1.683e+05  -0.159   0.8735
## city_Redmond              1.312e+05  1.490e+05   0.880   0.3787
## city_Renton              -5.416e+04  1.487e+05  -0.364   0.7158
## city_Sammamish            8.905e+04  1.493e+05   0.597   0.5508
## city_SeaTac              -5.574e+04  1.542e+05  -0.361   0.7178
## city_Seattle              1.815e+05  1.481e+05   1.225   0.2206
## city_Shoreline            3.279e+04  1.494e+05   0.219   0.8263
## city_Skykomish                   NA         NA      NA       NA
## city_Snoqualmie          -3.369e+04  1.508e+05  -0.223   0.8233
## city_Snoqualmie.Pass      9.225e+04  2.561e+05   0.360   0.7187
## city_Tukwila             -2.586e+04  1.546e+05  -0.167   0.8672
## city_Vashon              -1.783e+05  1.553e+05  -1.148   0.2513
## city_Woodinville          3.687e+04  1.499e+05   0.246   0.8057
## city_Yarrow.Point                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 208800 on 3358 degrees of freedom
## Multiple R-squared:  0.6952, Adjusted R-squared:  0.6905
## F-statistic: 150.2 on 51 and 3358 DF,  p-value: < 2.2e-16
```



```
rmse = sqrt(sum((pred_test - test$price)^2)/1137)
rmse
```



```
## [1] 198388.9
```

Figure 13 -17:

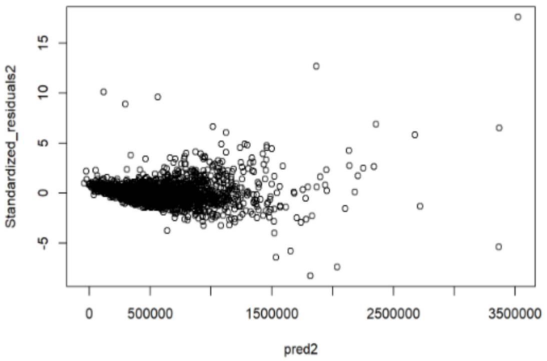


Figure 13

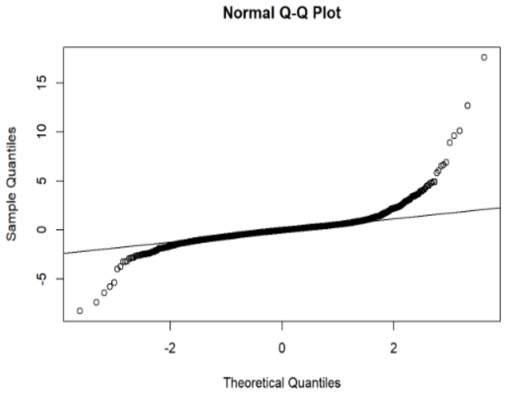


Figure 14

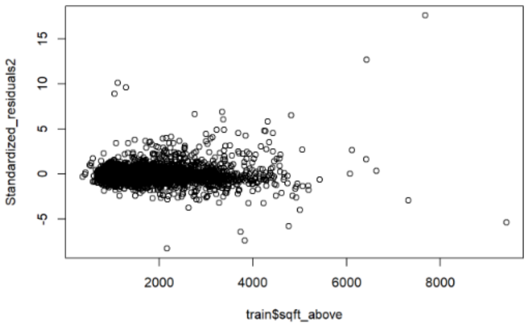


Figure 15

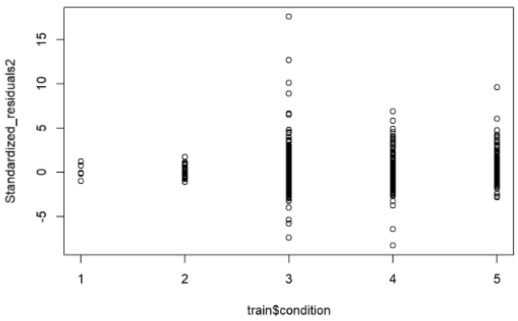


Figure 16

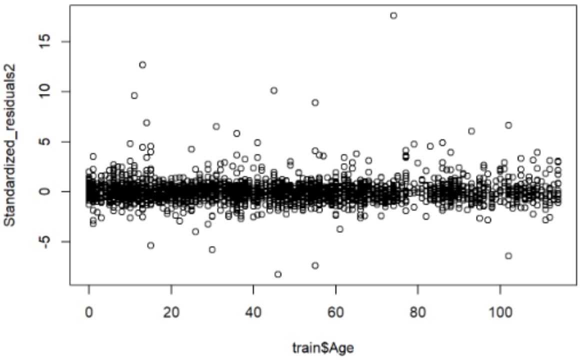
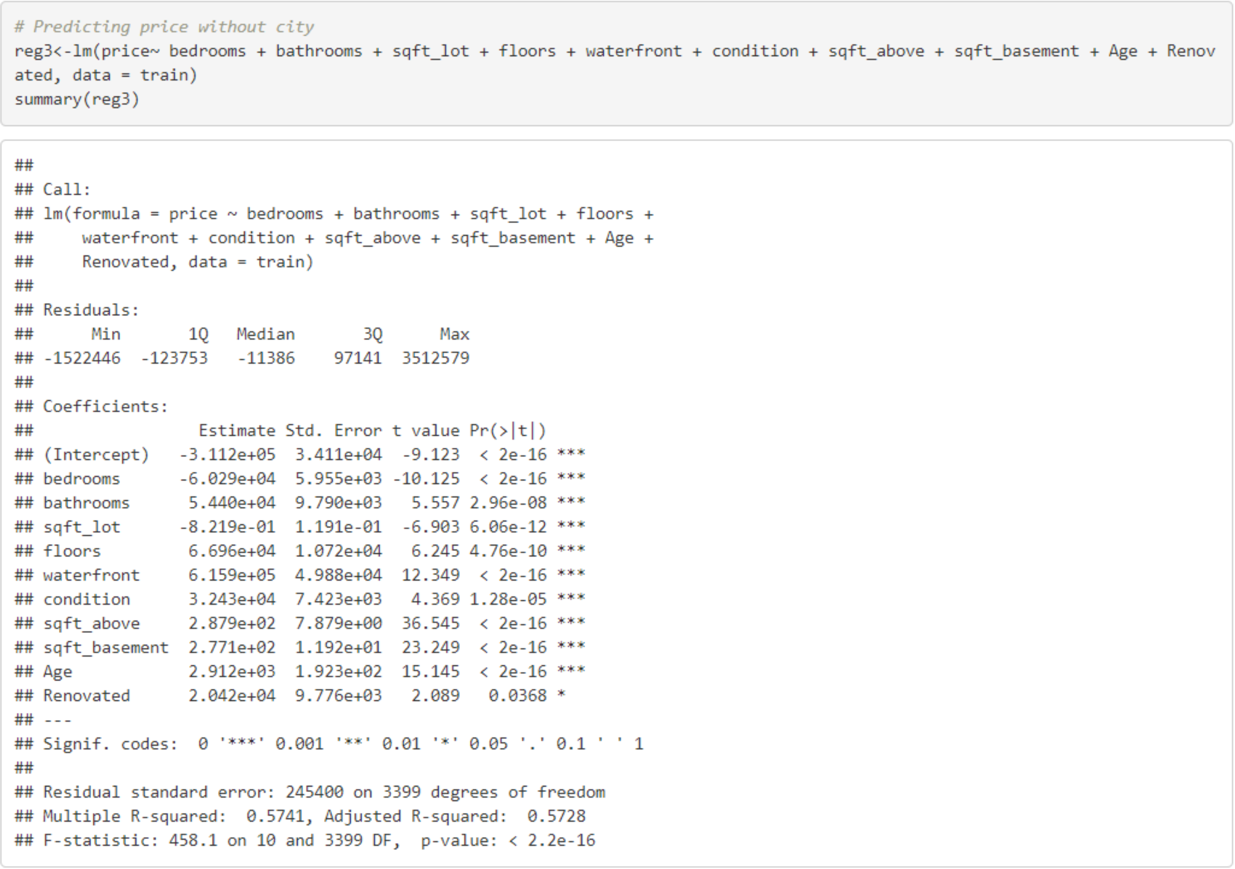


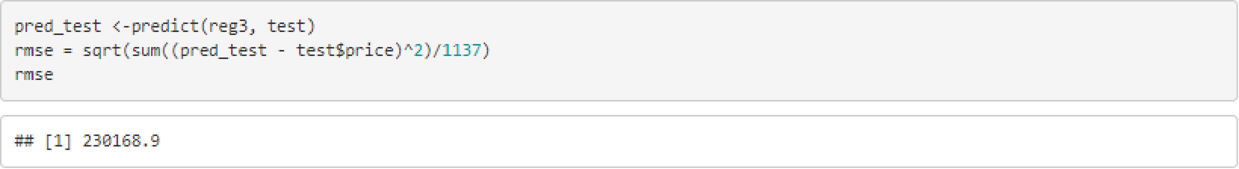
Figure 17

Table 3: Regression results - Without city



```
# Predicting price without city
reg3<-lm(price~ bedrooms + bathrooms + sqft_lot + floors + waterfront + condition + sqft_above + sqft_basement + Age + Renov
ated, data = train)
summary(reg3)
```

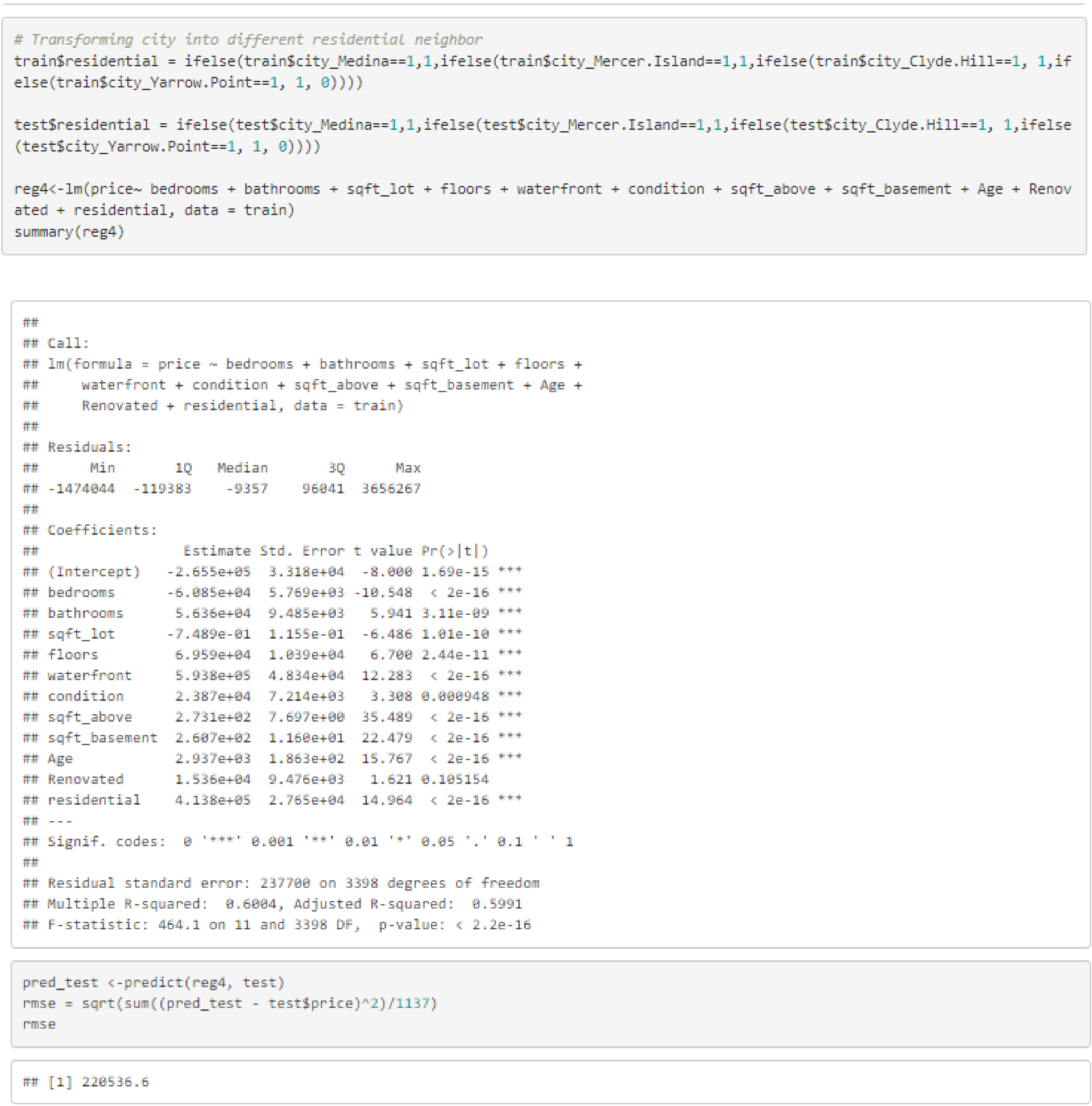
```
##
## Call:
## lm(formula = price ~ bedrooms + bathrooms + sqft_lot + floors +
##     waterfront + condition + sqft_above + sqft_basement + Age +
##     Renovated, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1522446  -123753   -11386    97141  3512579
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.112e+05  3.411e+04  -9.123  < 2e-16 ***
## bedrooms      -6.029e+04  5.955e+03 -10.125  < 2e-16 ***
## bathrooms      5.440e+04  9.790e+03   5.557 2.96e-08 ***
## sqft_lot      -8.219e-01  1.191e-01  -6.903 6.06e-12 ***
## floors         6.696e+04  1.072e+04   6.245 4.76e-10 ***
## waterfront     6.159e+05  4.988e+04  12.349  < 2e-16 ***
## condition      3.243e+04  7.423e+03   4.369 1.28e-05 ***
## sqft_above     2.879e+02  7.879e+00  36.545  < 2e-16 ***
## sqft_basement  2.771e+02  1.192e+01  23.249  < 2e-16 ***
## Age            2.912e+03  1.923e+02  15.145  < 2e-16 ***
## Renovated      2.042e+04  9.776e+03   2.089   0.0368 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 245400 on 3399 degrees of freedom
## Multiple R-squared:  0.5741, Adjusted R-squared:  0.5728
## F-statistic: 458.1 on 10 and 3399 DF,  p-value: < 2.2e-16
```



```
pred_test <-predict(reg3, test)
rmse = sqrt(sum((pred_test - test$price)^2)/1137)
rmse
```

```
## [1] 230168.9
```

Table 4: Regression results - all variables but city + new variable residential



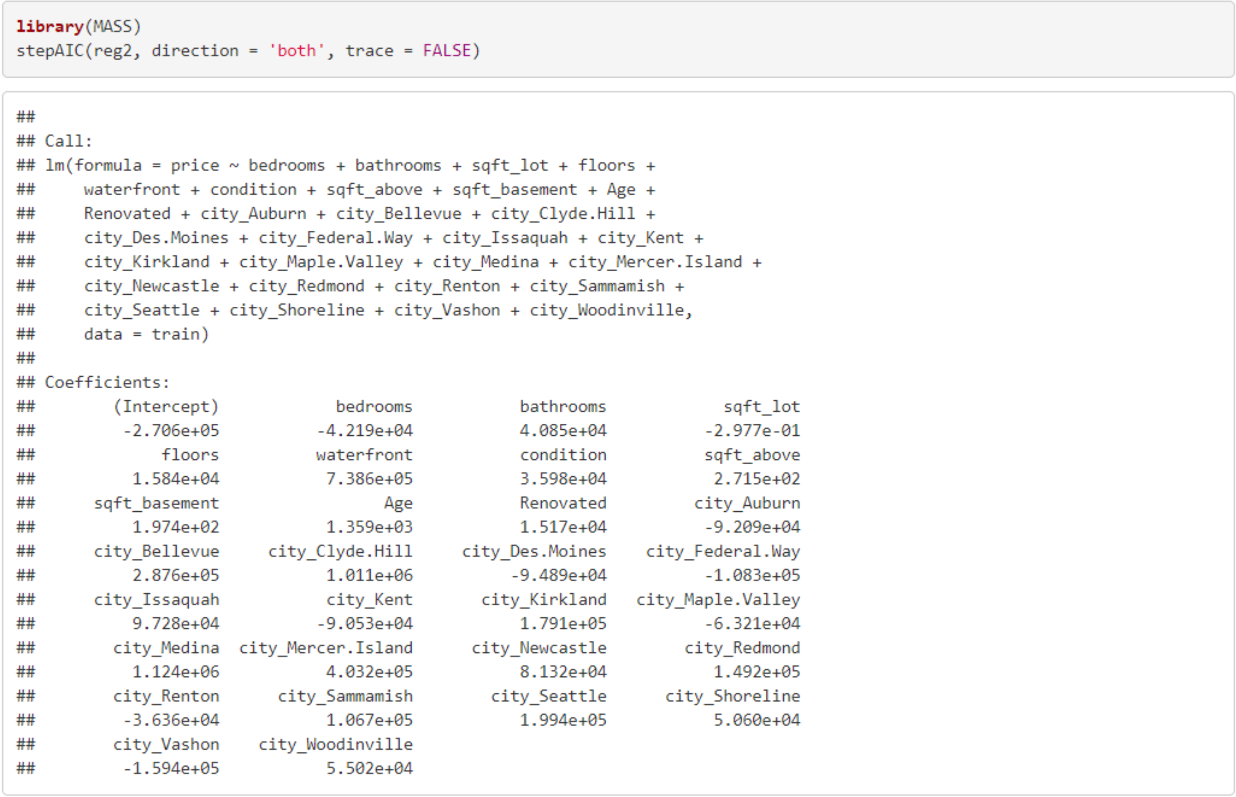
```
# Transforming city into different residential neighbor
train$residential = ifelse(train$city_Medina==1,1,ifelse(train$city_Mercer.Island==1,1,ifelse(train$city_Clyde.Hill==1, 1,if
else(train$city_Yarrow.Point==1, 1, 0))))

test$residential = ifelse(test$city_Medina==1,1,ifelse(test$city_Mercer.Island==1,1,ifelse(test$city_Clyde.Hill==1, 1,ifelse
(test$city_Yarrow.Point==1, 1, 0))))

reg4<-lm(price~ bedrooms + bathrooms + sqft_lot + floors + waterfront + condition + sqft_above + sqft_basement + Age + Renov
ated + residential, data = train)
summary(reg4)
```

```
## 
## Call:
## lm(formula = price ~ bedrooms + bathrooms + sqft_lot + floors + 
##     waterfront + condition + sqft_above + sqft_basement + Age + 
##     Renovated + residential, data = train)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1474044  -119383    -9357    96041  3656267 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   -2.655e+05  3.318e+04  -8.000 1.69e-15 ***
## bedrooms      -6.085e+04  5.769e+03 -10.548  < 2e-16 ***
## bathrooms      5.636e+04  9.485e+03   5.941 3.11e-09 ***
## sqft_lot      -7.489e-01  1.155e-01  -6.486 1.01e-10 ***
## floors         6.959e+04  1.039e+04   6.700 2.44e-11 ***
## waterfront     5.938e+05  4.834e+04  12.283  < 2e-16 ***
## condition      2.387e+04  7.214e+03   3.308 0.000948 ***
## sqft_above     2.731e+02  7.697e+00  35.489  < 2e-16 ***
## sqft_basement  2.607e+02  1.160e+01  22.479  < 2e-16 ***
## Age            2.937e+03  1.863e+02  15.767  < 2e-16 ***
## Renovated      1.536e+04  9.476e+03   1.621 0.105154    
## residential    4.138e+05  2.765e+04  14.964  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 237700 on 3398 degrees of freedom
## Multiple R-squared:  0.6004, Adjusted R-squared:  0.5991 
## F-statistic: 464.1 on 11 and 3398 DF,  p-value: < 2.2e-16
```

```
pred_test <-predict(reg4, test)
rmse = sqrt(sum((pred_test - test$price)^2)/1137)
rmse
```

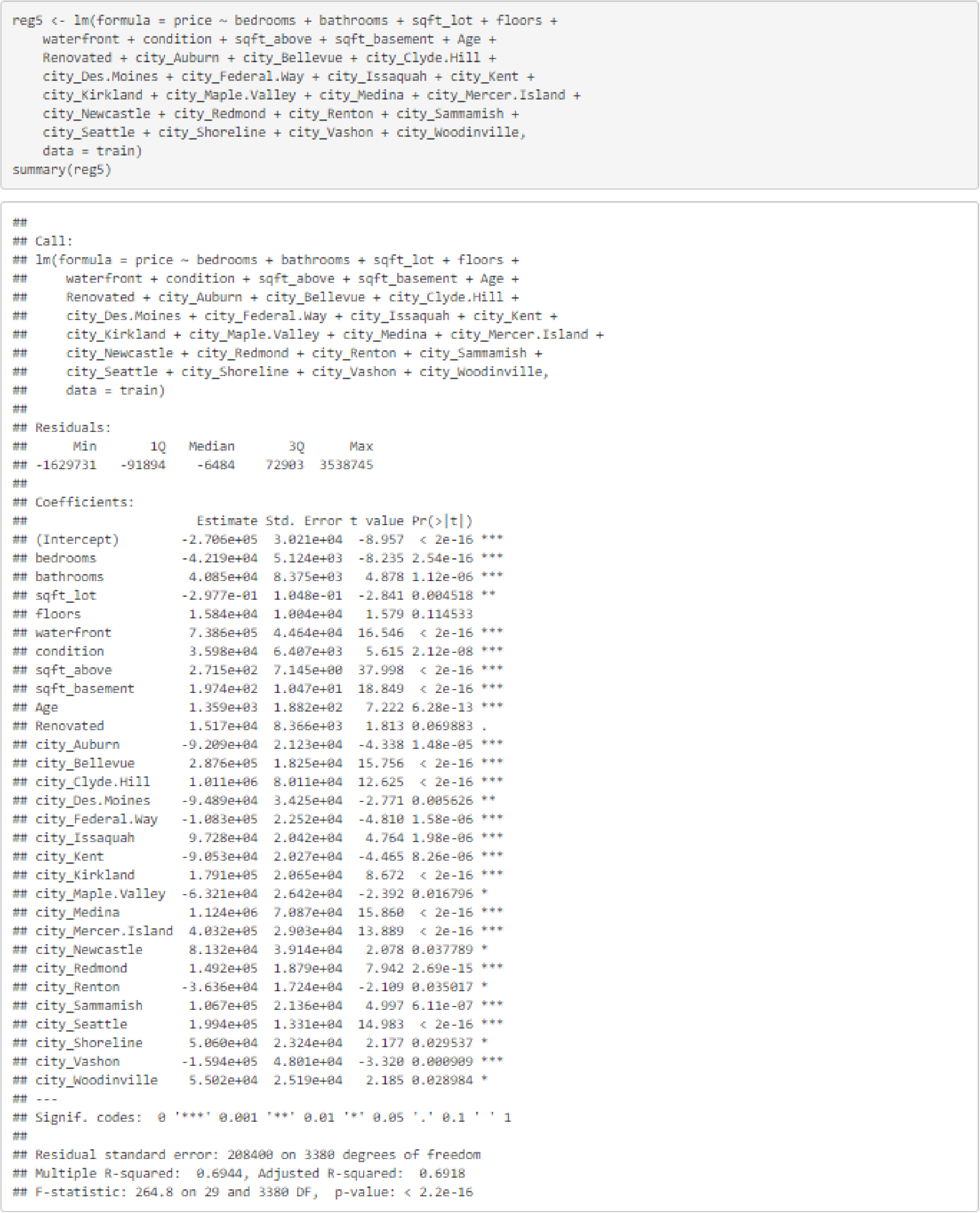
```
## [1] 220536.6
```

Table 5: Regression results - Selecting variables using AIC criteria



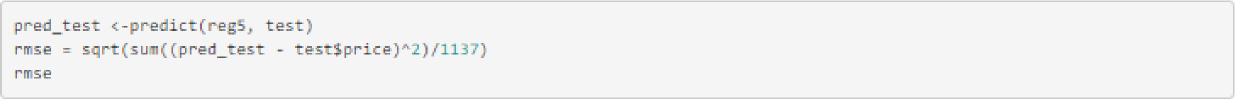
```r
library(MASS)
stepAIC(reg2, direction = 'both', trace = FALSE)
```

```
##
## Call:
## lm(formula = price ~ bedrooms + bathrooms + sqft_lot + floors +
##     waterfront + condition + sqft_above + sqft_basement + Age +
##     Renovated + city_Auburn + city_Bellevue + city_Clyde.Hill +
##     city_Des.Moines + city_Federal.Way + city_Issaquah + city_Kent +
##     city_Kirkland + city_Maple.Valley + city_Medina + city_Mercer.Island +
##     city_Newcastle + city_Redmond + city_Renton + city_Sammamish +
##     city_Seattle + city_Shoreline + city_Vashon + city_Woodinville,
##     data = train)
##
## Coefficients:
##        (Intercept)            bedrooms           bathrooms            sqft_lot
##         -2.706e+05          -4.219e+04           4.085e+04          -2.977e-01
##             floors          waterfront           condition          sqft_above
##          1.584e+04           7.386e+05           3.598e+04           2.715e+02
##      sqft_basement                 Age           Renovated         city_Auburn
##          1.974e+02           1.359e+03           1.517e+04          -9.209e+04
##      city_Bellevue     city_Clyde.Hill     city_Des.Moines    city_Federal.Way
##          2.876e+05           1.011e+06          -9.489e+04          -1.083e+05
##      city_Issaquah           city_Kent       city_Kirkland   city_Maple.Valley
##          9.728e+04          -9.053e+04           1.791e+05          -6.321e+04
##        city_Medina  city_Mercer.Island      city_Newcastle        city_Redmond
##          1.124e+06           4.032e+05           8.132e+04           1.492e+05
##        city_Renton      city_Sammamish        city_Seattle      city_Shoreline
##         -3.636e+04           1.067e+05           1.994e+05           5.060e+04
##        city_Vashon    city_Woodinville
##         -1.594e+05           5.502e+04
```

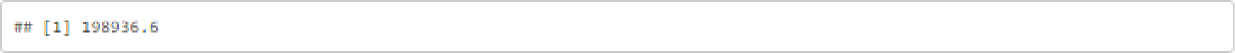


```
reg5 <- lm(formula = price ~ bedrooms + bathrooms + sqft_lot + floors +
    waterfront + condition + sqft_above + sqft_basement + Age +
    Renovated + city_Auburn + city_Bellevue + city_Clyde.Hill +
    city_Des.Moines + city_Federal.Way + city_Issaquah + city_Kent +
    city_Kirkland + city_Maple.Valley + city_Medina + city_Mercer.Island +
    city_Newcastle + city_Redmond + city_Renton + city_Sammamish +
    city_Seattle + city_Shoreline + city_Vashon + city_Woodinville,
    data = train)
summary(reg5)
```

```
##
## Call:
## lm(formula = price ~ bedrooms + bathrooms + sqft_lot + floors +
##     waterfront + condition + sqft_above + sqft_basement + Age +
##     Renovated + city_Auburn + city_Bellevue + city_Clyde.Hill +
##     city_Des.Moines + city_Federal.Way + city_Issaquah + city_Kent +
##     city_Kirkland + city_Maple.Valley + city_Medina + city_Mercer.Island +
##     city_Newcastle + city_Redmond + city_Renton + city_Sammamish +
##     city_Seattle + city_Shoreline + city_Vashon + city_Woodinville,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1629731   -91894    -6484    72903  3538745
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.706e+05  3.021e+04  -8.957  < 2e-16 ***
## bedrooms           -4.219e+04  5.124e+03  -8.235 2.54e-16 ***
## bathrooms           4.085e+04  8.375e+03   4.878 1.12e-06 ***
## sqft_lot           -2.977e-01  1.048e-01  -2.841 0.004518 **
## floors              1.584e+04  1.004e+04   1.579 0.114533
## waterfront          7.386e+05  4.464e+04  16.546  < 2e-16 ***
## condition           3.598e+04  6.407e+03   5.615 2.12e-08 ***
## sqft_above          2.715e+02  7.145e+00  37.998  < 2e-16 ***
## sqft_basement       1.974e+02  1.047e+01  18.849  < 2e-16 ***
## Age                 1.359e+03  1.882e+02   7.222 6.28e-13 ***
## Renovated           1.517e+04  8.366e+03   1.813 0.069883 .
## city_Auburn        -9.209e+04  2.123e+04  -4.338 1.48e-05 ***
## city_Bellevue       2.876e+05  1.825e+04  15.756  < 2e-16 ***
## city_Clyde.Hill     1.011e+06  8.011e+04  12.625  < 2e-16 ***
## city_Des.Moines    -9.489e+04  3.425e+04  -2.771 0.005626 **
## city_Federal.Way   -1.083e+05  2.252e+04  -4.810 1.58e-06 ***
## city_Issaquah       9.728e+04  2.042e+04   4.764 1.98e-06 ***
## city_Kent          -9.053e+04  2.027e+04  -4.465 8.26e-06 ***
## city_Kirkland       1.791e+05  2.065e+04   8.672  < 2e-16 ***
## city_Maple.Valley  -6.321e+04  2.642e+04  -2.392 0.016796 *
## city_Medina         1.124e+06  7.087e+04  15.860  < 2e-16 ***
## city_Mercer.Island  4.032e+05  2.903e+04  13.889  < 2e-16 ***
## city_Newcastle      8.132e+04  3.914e+04   2.078 0.037789 *
## city_Redmond        1.492e+05  1.879e+04   7.942 2.69e-15 ***
## city_Renton        -3.636e+04  1.724e+04  -2.109 0.035017 *
## city_Sammamish      1.067e+05  2.136e+04   4.997 6.11e-07 ***
## city_Seattle        1.994e+05  1.331e+04  14.983  < 2e-16 ***
## city_Shoreline      5.060e+04  2.324e+04   2.177 0.029537 *
## city_Vashon        -1.594e+05  4.801e+04  -3.320 0.000909 ***
## city_Woodinville    5.502e+04  2.519e+04   2.185 0.028984 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 208400 on 3380 degrees of freedom
## Multiple R-squared:  0.6944, Adjusted R-squared:  0.6918
## F-statistic: 264.8 on 29 and 3380 DF,  p-value: < 2.2e-16
```

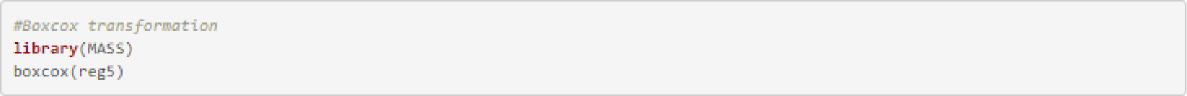


```
pred_test <-predict(reg5, test)
rmse = sqrt(sum((pred_test - test$price)^2)/1137)
rmse
```



```
## [1] 198936.6
```

Figure 18 : Boxcox transformation



```
#Boxcox transformation
library(MASS)
boxcox(reg5)
```

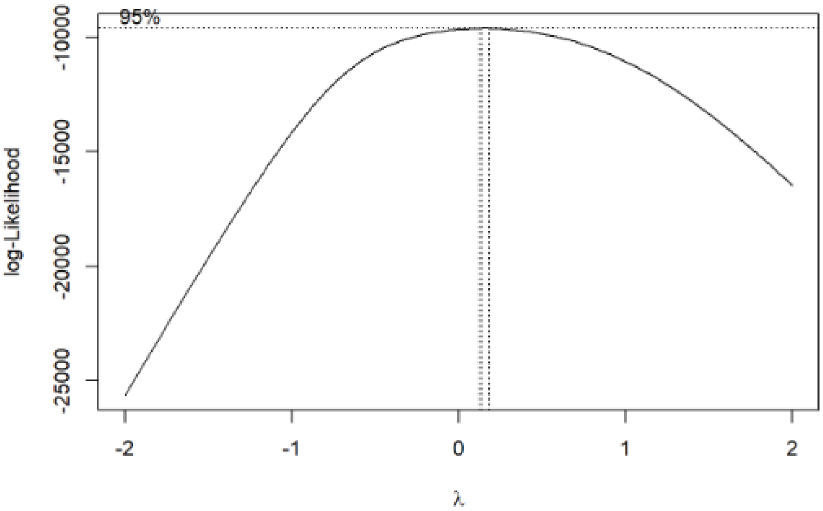
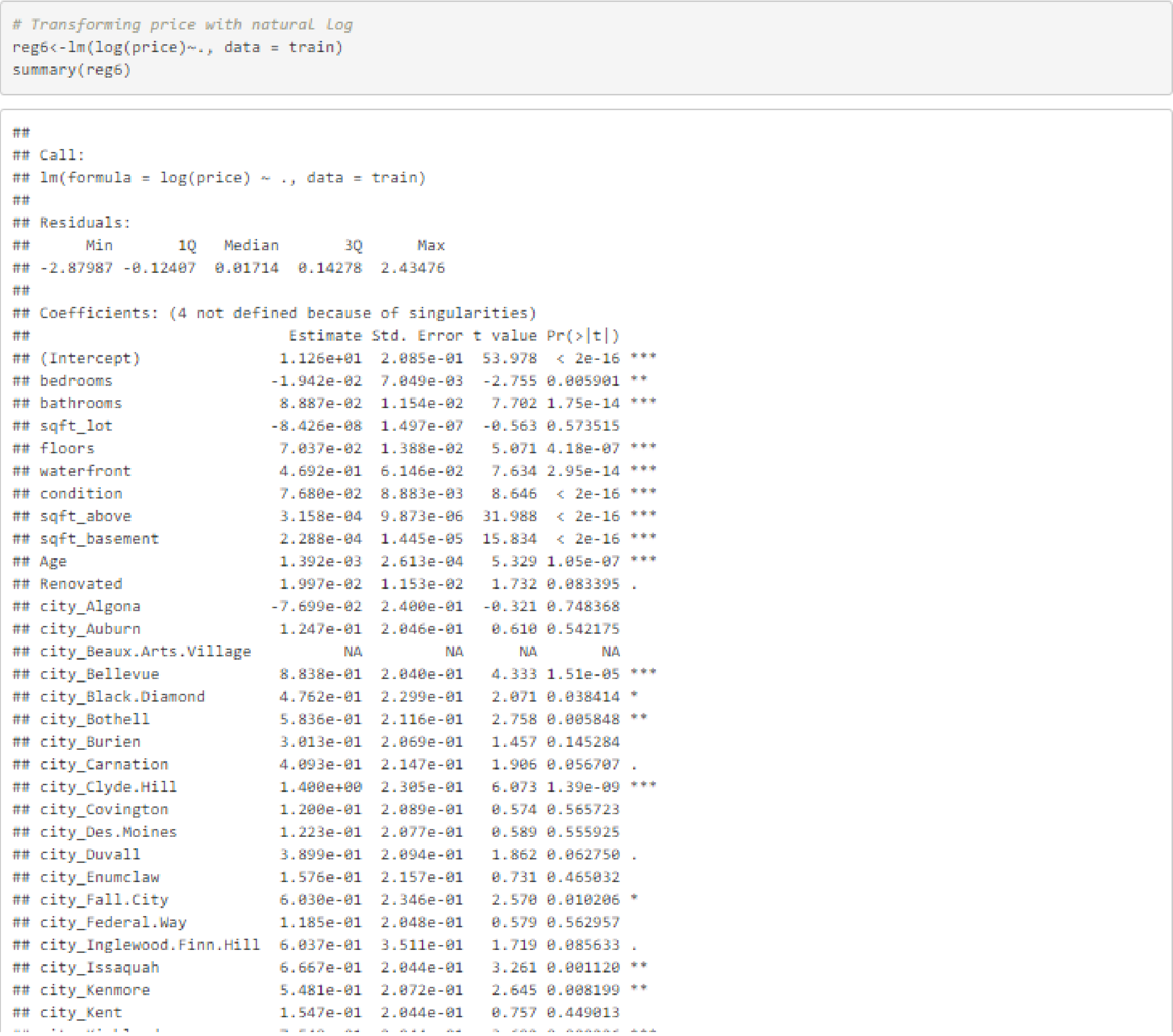
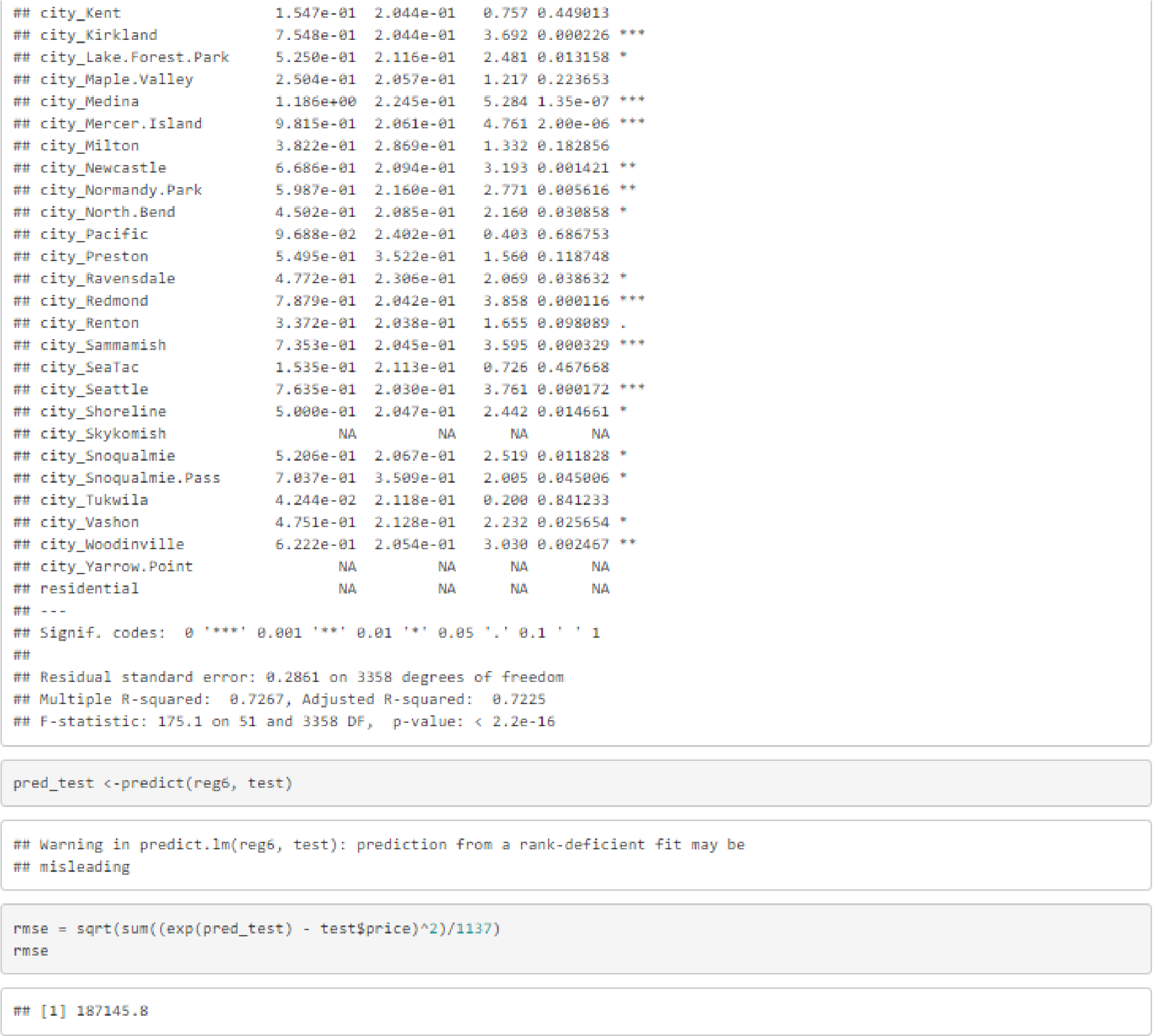


Table 6: Regression results -Logarithmic dependent variable



```r
# Transforming price with natural log
reg6<-lm(log(price)~., data = train)
summary(reg6)
```

```
##
## Call:
## lm(formula = log(price) ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.87987 -0.12407  0.01714  0.14278  2.43476
##
## Coefficients: (4 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.126e+01  2.085e-01  53.978  < 2e-16 ***
## bedrooms                 -1.942e-02  7.049e-03  -2.755 0.005901 **
## bathrooms                 8.887e-02  1.154e-02   7.702 1.75e-14 ***
## sqft_lot                 -8.426e-08  1.497e-07  -0.563 0.573515
## floors                    7.037e-02  1.388e-02   5.071 4.18e-07 ***
## waterfront                4.692e-01  6.146e-02   7.634 2.95e-14 ***
## condition                 7.680e-02  8.883e-03   8.646  < 2e-16 ***
## sqft_above                3.158e-04  9.873e-06  31.988  < 2e-16 ***
## sqft_basement             2.288e-04  1.445e-05  15.834  < 2e-16 ***
## Age                       1.392e-03  2.613e-04   5.329 1.05e-07 ***
## Renovated                 1.997e-02  1.153e-02   1.732 0.083395 .
## city_Algona              -7.699e-02  2.400e-01  -0.321 0.748368
## city_Auburn               1.247e-01  2.046e-01   0.610 0.542175
## city_Beaux.Arts.Village          NA         NA      NA       NA
## city_Bellevue             8.838e-01  2.040e-01   4.333 1.51e-05 ***
## city_Black.Diamond        4.762e-01  2.299e-01   2.071 0.038414 *
## city_Bothell              5.836e-01  2.116e-01   2.758 0.005848 **
## city_Burien               3.013e-01  2.069e-01   1.457 0.145284
## city_Carnation            4.093e-01  2.147e-01   1.906 0.056707 .
## city_Clyde.Hill           1.400e+00  2.305e-01   6.073 1.39e-09 ***
## city_Covington            1.200e-01  2.089e-01   0.574 0.565723
## city_Des.Moines           1.223e-01  2.077e-01   0.589 0.555925
## city_Duvall               3.899e-01  2.094e-01   1.862 0.062750 .
## city_Enumclaw             1.576e-01  2.157e-01   0.731 0.465032
## city_Fall.City            6.030e-01  2.346e-01   2.570 0.010206 *
## city_Federal.Way          1.185e-01  2.048e-01   0.579 0.562957
## city_Inglewood.Finn.Hill  6.037e-01  3.511e-01   1.719 0.085633 .
## city_Issaquah             6.667e-01  2.044e-01   3.261 0.001120 **
## city_Kenmore              5.481e-01  2.072e-01   2.645 0.008199 **
## city_Kent                 1.547e-01  2.044e-01   0.757 0.449013
```



```
## city_Kent                 1.547e-01  2.044e-01   0.757 0.449013    
## city_Kirkland             7.548e-01  2.044e-01   3.692 0.000226 ***
## city_Lake.Forest.Park     5.250e-01  2.116e-01   2.481 0.013158 *  
## city_Maple.Valley         2.504e-01  2.057e-01   1.217 0.223653    
## city_Medina               1.186e+00  2.245e-01   5.284 1.35e-07 ***
## city_Mercer.Island        9.815e-01  2.061e-01   4.761 2.00e-06 ***
## city_Milton               3.822e-01  2.869e-01   1.332 0.182856    
## city_Newcastle            6.686e-01  2.094e-01   3.193 0.001421 ** 
## city_Normandy.Park        5.987e-01  2.160e-01   2.771 0.005616 ** 
## city_North.Bend           4.502e-01  2.085e-01   2.160 0.030858 *  
## city_Pacific              9.688e-02  2.402e-01   0.403 0.686753    
## city_Preston              5.495e-01  3.522e-01   1.560 0.118748    
## city_Ravensdale           4.772e-01  2.306e-01   2.069 0.038632 *  
## city_Redmond              7.879e-01  2.042e-01   3.858 0.000116 ***
## city_Renton               3.372e-01  2.038e-01   1.655 0.098089 .  
## city_Sammamish            7.353e-01  2.045e-01   3.595 0.000329 ***
## city_SeaTac               1.535e-01  2.113e-01   0.726 0.467668    
## city_Seattle              7.635e-01  2.030e-01   3.761 0.000172 ***
## city_Shoreline            5.000e-01  2.047e-01   2.442 0.014661 *  
## city_Skykomish                   NA         NA      NA       NA    
## city_Snoqualmie           5.206e-01  2.067e-01   2.519 0.011828 *  
## city_Snoqualmie.Pass      7.037e-01  3.509e-01   2.005 0.045006 *  
## city_Tukwila              4.244e-02  2.118e-01   0.200 0.841233    
## city_Vashon               4.751e-01  2.128e-01   2.232 0.025654 *  
## city_Woodinville          6.222e-01  2.054e-01   3.030 0.002467 ** 
## city_Yarrow.Point                NA         NA      NA       NA    
## residential                      NA         NA      NA       NA    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2861 on 3358 degrees of freedom
## Multiple R-squared:  0.7267, Adjusted R-squared:  0.7225 
## F-statistic: 175.1 on 51 and 3358 DF,  p-value: < 2.2e-16
```

```
pred_test <-predict(reg6, test)
```

```
## Warning in predict.lm(reg6, test): prediction from a rank-deficient fit may be
## misleading
```

```
rmse = sqrt(sum((exp(pred_test) - test$price)^2)/1137)
rmse
```

```
## [1] 187145.8
```

Figure 18 - 22 :

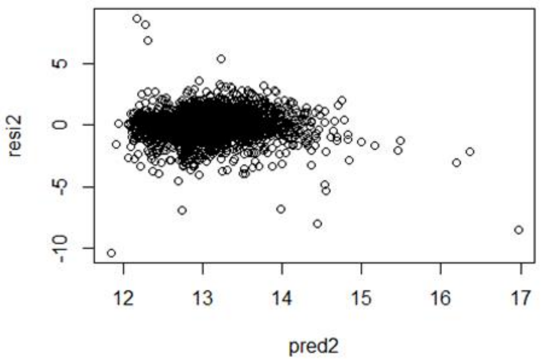


Figure 18

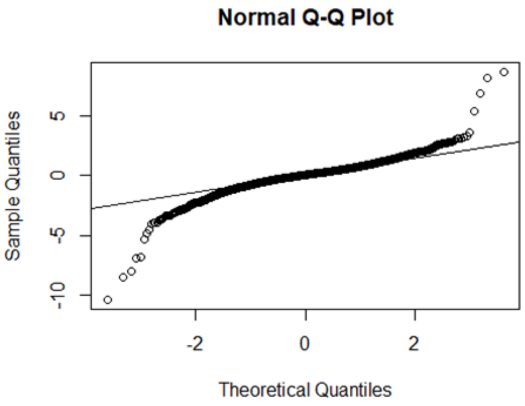


Figure 19

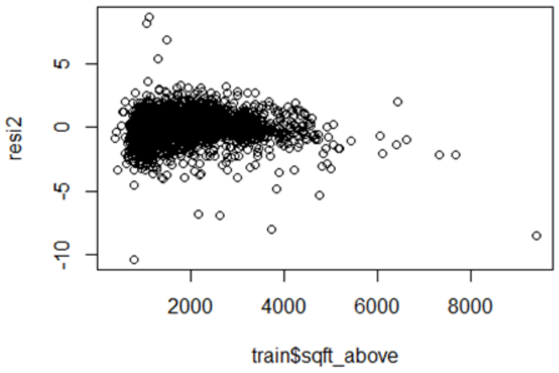


Figure 20

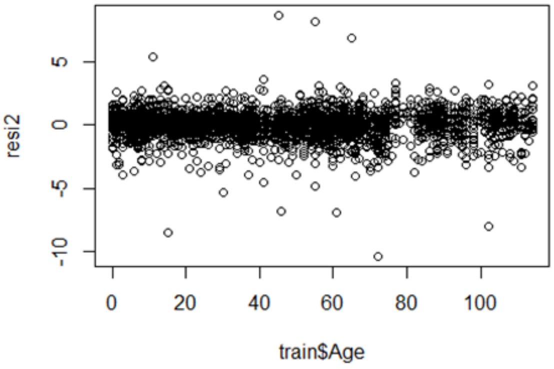
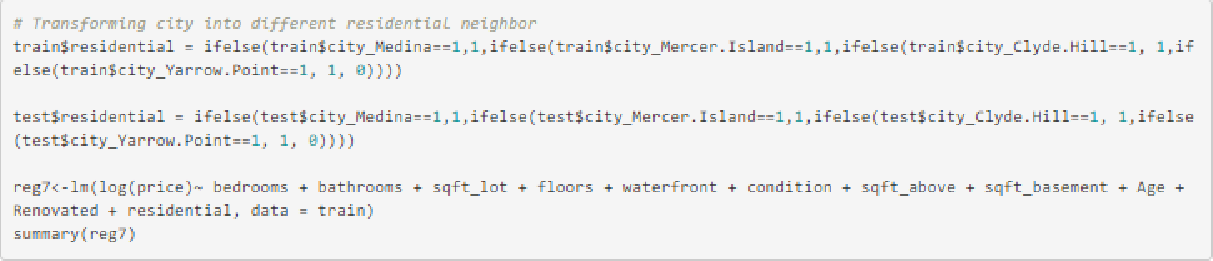
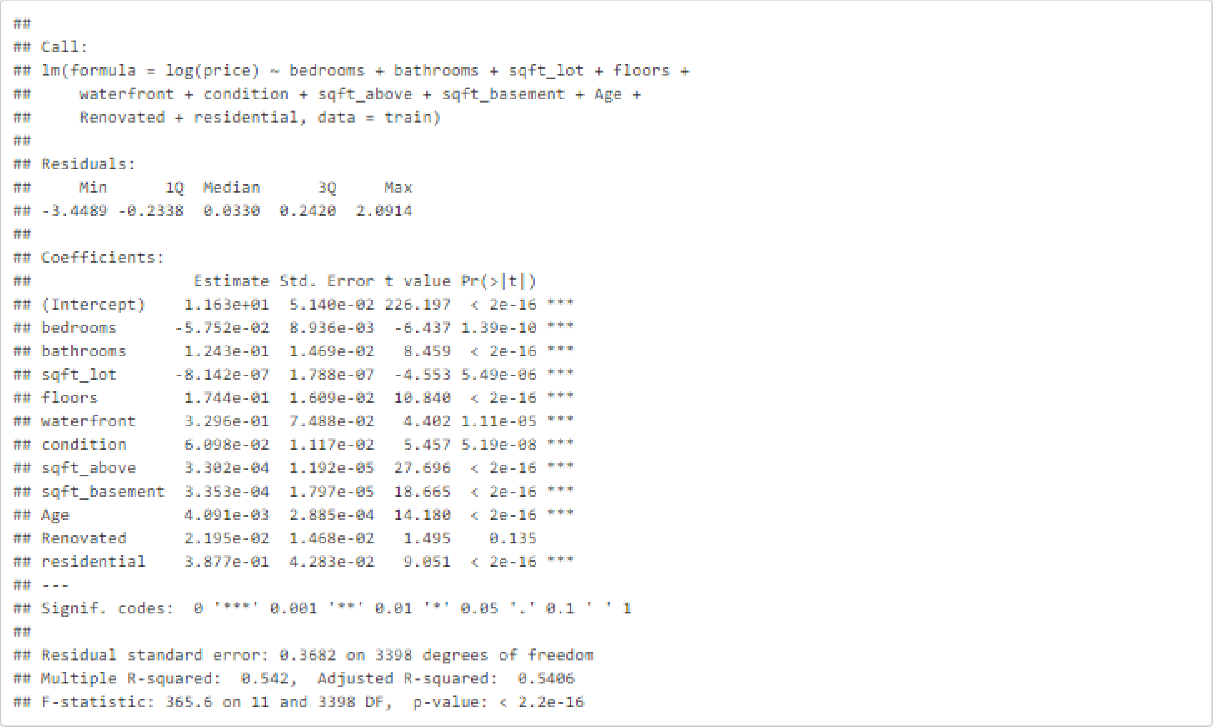


Figure 21

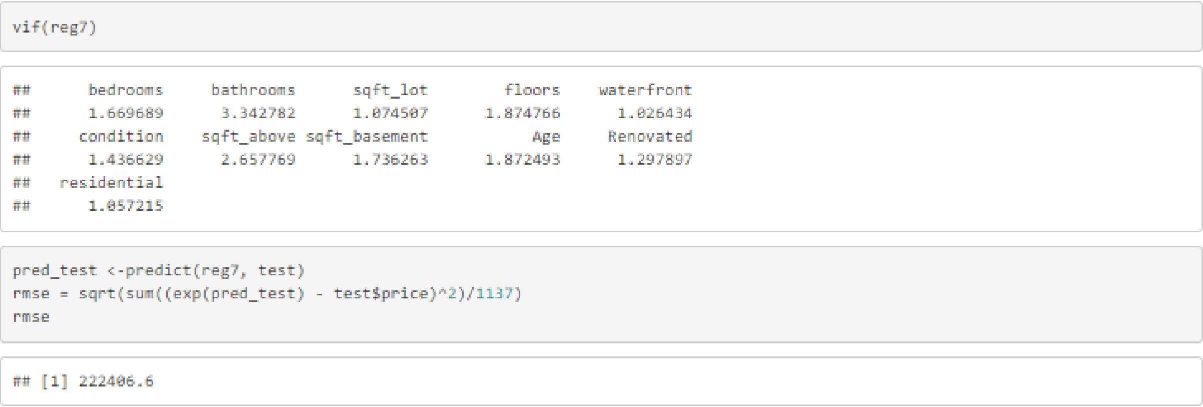
Table 7: Regression results - All variables but city + new variable residential



```
# Transforming city into different residential neighbor
train$residential = ifelse(train$city_Medina==1,1,ifelse(train$city_Mercer.Island==1,1,ifelse(train$city_Clyde.Hill==1, 1,if
else(train$city_Yarrow.Point==1, 1, 0))))

test$residential = ifelse(test$city_Medina==1,1,ifelse(test$city_Mercer.Island==1,1,ifelse(test$city_Clyde.Hill==1, 1,ifelse
(test$city_Yarrow.Point==1, 1, 0))))

reg7<-lm(log(price)~ bedrooms + bathrooms + sqft_lot + floors + waterfront + condition + sqft_above + sqft_basement + Age +
Renovated + residential, data = train)
summary(reg7)
```



```
##
## Call:
## lm(formula = log(price) ~ bedrooms + bathrooms + sqft_lot + floors +
##     waterfront + condition + sqft_above + sqft_basement + Age +
##     Renovated + residential, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4489 -0.2338  0.0330  0.2420  2.0914
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.163e+01  5.140e-02 226.197  < 2e-16 ***
## bedrooms      -5.752e-02  8.936e-03  -6.437 1.39e-10 ***
## bathrooms      1.243e-01  1.469e-02   8.459  < 2e-16 ***
## sqft_lot      -8.142e-07  1.788e-07  -4.553 5.49e-06 ***
## floors         1.744e-01  1.609e-02  10.840  < 2e-16 ***
## waterfront     3.296e-01  7.488e-02   4.402 1.11e-05 ***
## condition      6.098e-02  1.117e-02   5.457 5.19e-08 ***
## sqft_above     3.302e-04  1.192e-05  27.696  < 2e-16 ***
## sqft_basement  3.353e-04  1.797e-05  18.665  < 2e-16 ***
## Age            4.091e-03  2.885e-04  14.180  < 2e-16 ***
## Renovated      2.195e-02  1.468e-02   1.495    0.135
## residential    3.877e-01  4.283e-02   9.051  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3682 on 3398 degrees of freedom
## Multiple R-squared:  0.542,  Adjusted R-squared:  0.5406
## F-statistic: 365.6 on 11 and 3398 DF,  p-value: < 2.2e-16
```



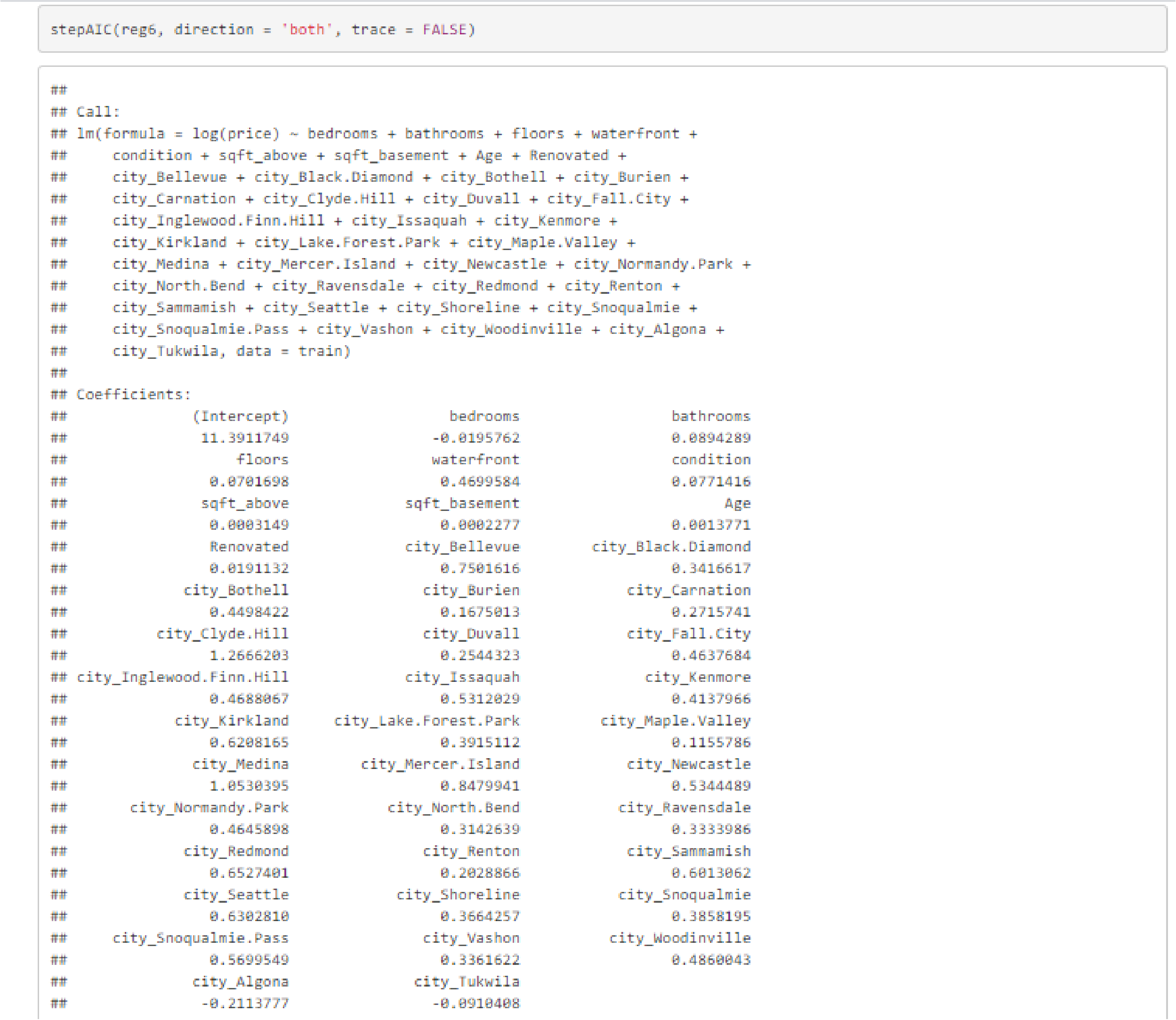
```r
vif(reg7)
```

```
##      bedrooms     bathrooms      sqft_lot        floors    waterfront 
##      1.669689      3.342782      1.074507      1.874766      1.026434 
##     condition    sqft_above sqft_basement           Age     Renovated 
##      1.436629      2.657769      1.736263      1.872493      1.297897 
##   residential 
##      1.057215
```

```r
pred_test <-predict(reg7, test)
rmse = sqrt(sum((exp(pred_test) - test$price)^2)/1137)
rmse
```

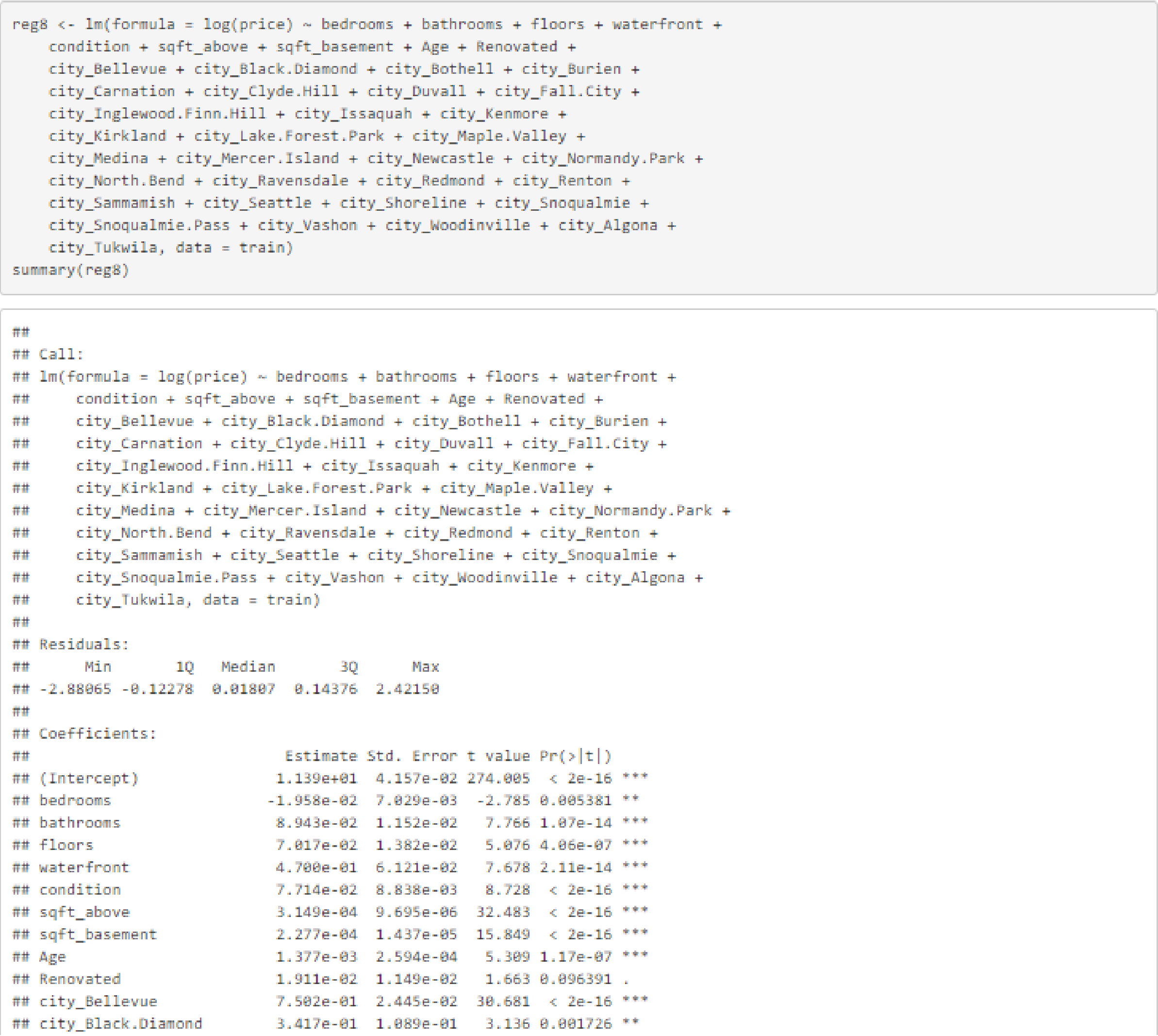
```
## [1] 222406.6
```

Table 8: Regression results - Model selection using AIC from model 6 (40 variables)



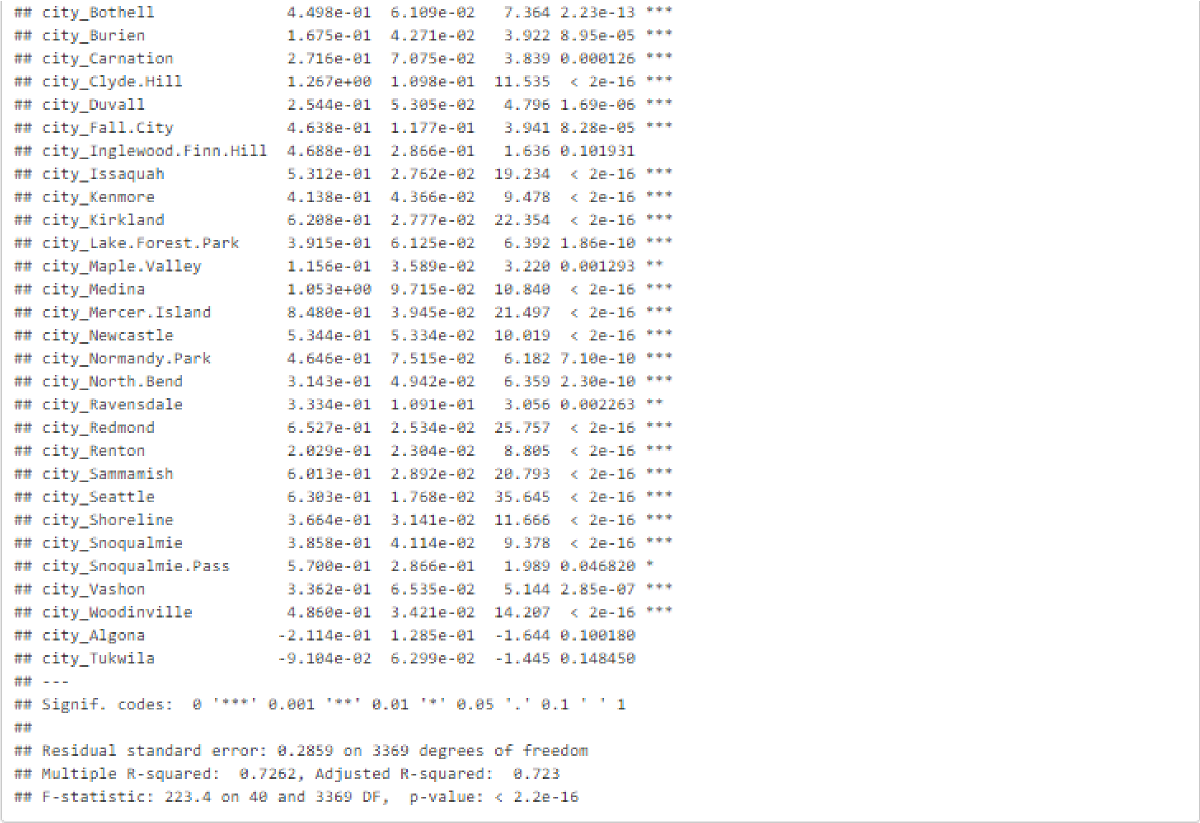
```
stepAIC(reg6, direction = 'both', trace = FALSE)
```

```
##
## Call:
## lm(formula = log(price) ~ bedrooms + bathrooms + floors + waterfront +
##     condition + sqft_above + sqft_basement + Age + Renovated +
##     city_Bellevue + city_Black.Diamond + city_Bothell + city_Burien +
##     city_Carnation + city_Clyde.Hill + city_Duvall + city_Fall.City +
##     city_Inglewood.Finn.Hill + city_Issaquah + city_Kenmore +
##     city_Kirkland + city_Lake.Forest.Park + city_Maple.Valley +
##     city_Medina + city_Mercer.Island + city_Newcastle + city_Normandy.Park +
##     city_North.Bend + city_Ravensdale + city_Redmond + city_Renton +
##     city_Sammamish + city_Seattle + city_Shoreline + city_Snoqualmie +
##     city_Snoqualmie.Pass + city_Vashon + city_Woodinville + city_Algona +
##     city_Tukwila, data = train)
##
## Coefficients:
##              (Intercept)                 bedrooms                bathrooms
##               11.3911749               -0.0195762                0.0894289
##                   floors               waterfront                condition
##                0.0701698                0.4699584                0.0771416
##               sqft_above            sqft_basement                      Age
##                0.0003149                0.0002277                0.0013771
##                Renovated            city_Bellevue       city_Black.Diamond
##                0.0191132                0.7501616                0.3416617
##             city_Bothell              city_Burien           city_Carnation
##                0.4498422                0.1675013                0.2715741
##          city_Clyde.Hill              city_Duvall           city_Fall.City
##                1.2666203                0.2544323                0.4637684
## city_Inglewood.Finn.Hill            city_Issaquah             city_Kenmore
##                0.4688067                0.5312029                0.4137966
##            city_Kirkland    city_Lake.Forest.Park        city_Maple.Valley
##                0.6208165                0.3915112                0.1155786
##              city_Medina       city_Mercer.Island           city_Newcastle
##                1.0530395                0.8479941                0.5344489
##       city_Normandy.Park          city_North.Bend          city_Ravensdale
##                0.4645898                0.3142639                0.3333986
##             city_Redmond              city_Renton           city_Sammamish
##                0.6527401                0.2028866                0.6013062
##             city_Seattle           city_Shoreline          city_Snoqualmie
##                0.6302810                0.3664257                0.3858195
##     city_Snoqualmie.Pass              city_Vashon         city_Woodinville
##                0.5699549                0.3361622                0.4860043
##              city_Algona             city_Tukwila
##               -0.2113777               -0.0910408
```

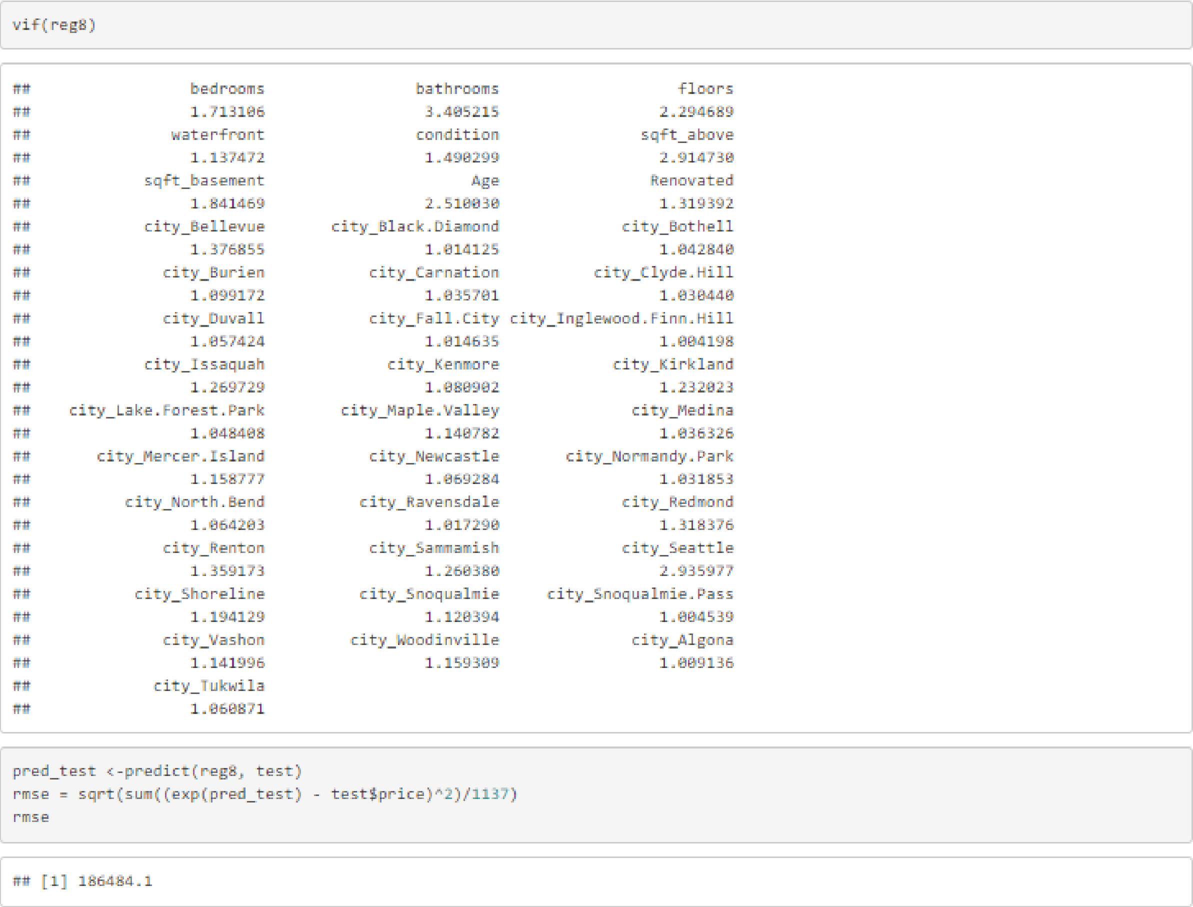


```
reg8 <- lm(formula = log(price) ~ bedrooms + bathrooms + floors + waterfront +
    condition + sqft_above + sqft_basement + Age + Renovated +
    city_Bellevue + city_Black.Diamond + city_Bothell + city_Burien +
    city_Carnation + city_Clyde.Hill + city_Duvall + city_Fall.City +
    city_Inglewood.Finn.Hill + city_Issaquah + city_Kenmore +
    city_Kirkland + city_Lake.Forest.Park + city_Maple.Valley +
    city_Medina + city_Mercer.Island + city_Newcastle + city_Normandy.Park +
    city_North.Bend + city_Ravensdale + city_Redmond + city_Renton +
    city_Sammamish + city_Seattle + city_Shoreline + city_Snoqualmie +
    city_Snoqualmie.Pass + city_Vashon + city_Woodinville + city_Algona +
    city_Tukwila, data = train)
summary(reg8)
```

```
## 
## Call:
## lm(formula = log(price) ~ bedrooms + bathrooms + floors + waterfront + 
##     condition + sqft_above + sqft_basement + Age + Renovated + 
##     city_Bellevue + city_Black.Diamond + city_Bothell + city_Burien + 
##     city_Carnation + city_Clyde.Hill + city_Duvall + city_Fall.City + 
##     city_Inglewood.Finn.Hill + city_Issaquah + city_Kenmore + 
##     city_Kirkland + city_Lake.Forest.Park + city_Maple.Valley + 
##     city_Medina + city_Mercer.Island + city_Newcastle + city_Normandy.Park + 
##     city_North.Bend + city_Ravensdale + city_Redmond + city_Renton + 
##     city_Sammamish + city_Seattle + city_Shoreline + city_Snoqualmie + 
##     city_Snoqualmie.Pass + city_Vashon + city_Woodinville + city_Algona + 
##     city_Tukwila, data = train)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.88065 -0.12278  0.01807  0.14376  2.42150 
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               1.139e+01  4.157e-02 274.005  < 2e-16 ***
## bedrooms                 -1.958e-02  7.029e-03  -2.785 0.005381 ** 
## bathrooms                 8.943e-02  1.152e-02   7.766 1.07e-14 ***
## floors                    7.017e-02  1.382e-02   5.076 4.06e-07 ***
## waterfront                4.700e-01  6.121e-02   7.678 2.11e-14 ***
## condition                 7.714e-02  8.838e-03   8.728  < 2e-16 ***
## sqft_above                3.149e-04  9.695e-06  32.483  < 2e-16 ***
## sqft_basement             2.277e-04  1.437e-05  15.849  < 2e-16 ***
## Age                       1.377e-03  2.594e-04   5.309 1.17e-07 ***
## Renovated                 1.911e-02  1.149e-02   1.663 0.096391 .  
## city_Bellevue             7.502e-01  2.445e-02  30.681  < 2e-16 ***
## city_Black.Diamond        3.417e-01  1.089e-01   3.136 0.001726 ** 
```



```
## city_Bothell             4.498e-01  6.109e-02   7.364 2.23e-13 ***
## city_Burien              1.675e-01  4.271e-02   3.922 8.95e-05 ***
## city_Carnation           2.716e-01  7.075e-02   3.839 0.000126 ***
## city_Clyde.Hill          1.267e+00  1.098e-01  11.535  < 2e-16 ***
## city_Duvall              2.544e-01  5.305e-02   4.796 1.69e-06 ***
## city_Fall.City           4.638e-01  1.177e-01   3.941 8.28e-05 ***
## city_Inglewood.Finn.Hill  4.688e-01  2.866e-01   1.636 0.101931
## city_Issaquah            5.312e-01  2.762e-02  19.234  < 2e-16 ***
## city_Kenmore             4.138e-01  4.366e-02   9.478  < 2e-16 ***
## city_Kirkland            6.208e-01  2.777e-02  22.354  < 2e-16 ***
## city_Lake.Forest.Park     3.915e-01  6.125e-02   6.392 1.86e-10 ***
## city_Maple.Valley        1.156e-01  3.589e-02   3.220 0.001293 **
## city_Medina              1.053e+00  9.715e-02  10.840  < 2e-16 ***
## city_Mercer.Island       8.480e-01  3.945e-02  21.497  < 2e-16 ***
## city_Newcastle           5.344e-01  5.334e-02  10.019  < 2e-16 ***
## city_Normandy.Park       4.646e-01  7.515e-02   6.182 7.10e-10 ***
## city_North.Bend          3.143e-01  4.942e-02   6.359 2.30e-10 ***
## city_Ravensdale          3.334e-01  1.091e-01   3.056 0.002263 **
## city_Redmond             6.527e-01  2.534e-02  25.757  < 2e-16 ***
## city_Renton              2.029e-01  2.304e-02   8.805  < 2e-16 ***
## city_Sammamish           6.013e-01  2.892e-02  20.793  < 2e-16 ***
## city_Seattle             6.303e-01  1.768e-02  35.645  < 2e-16 ***
## city_Shoreline           3.664e-01  3.141e-02  11.666  < 2e-16 ***
## city_Snoqualmie          3.858e-01  4.114e-02   9.378  < 2e-16 ***
## city_Snoqualmie.Pass      5.700e-01  2.866e-01   1.989 0.046820 *
## city_Vashon              3.362e-01  6.535e-02   5.144 2.85e-07 ***
## city_Woodinville         4.860e-01  3.421e-02  14.207  < 2e-16 ***
## city_Algona             -2.114e-01  1.285e-01  -1.644 0.100180
## city_Tukwila            -9.104e-02  6.299e-02  -1.445 0.148450
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2859 on 3369 degrees of freedom
## Multiple R-squared:  0.7262, Adjusted R-squared:  0.723
## F-statistic: 223.4 on 40 and 3369 DF,  p-value: < 2.2e-16
```



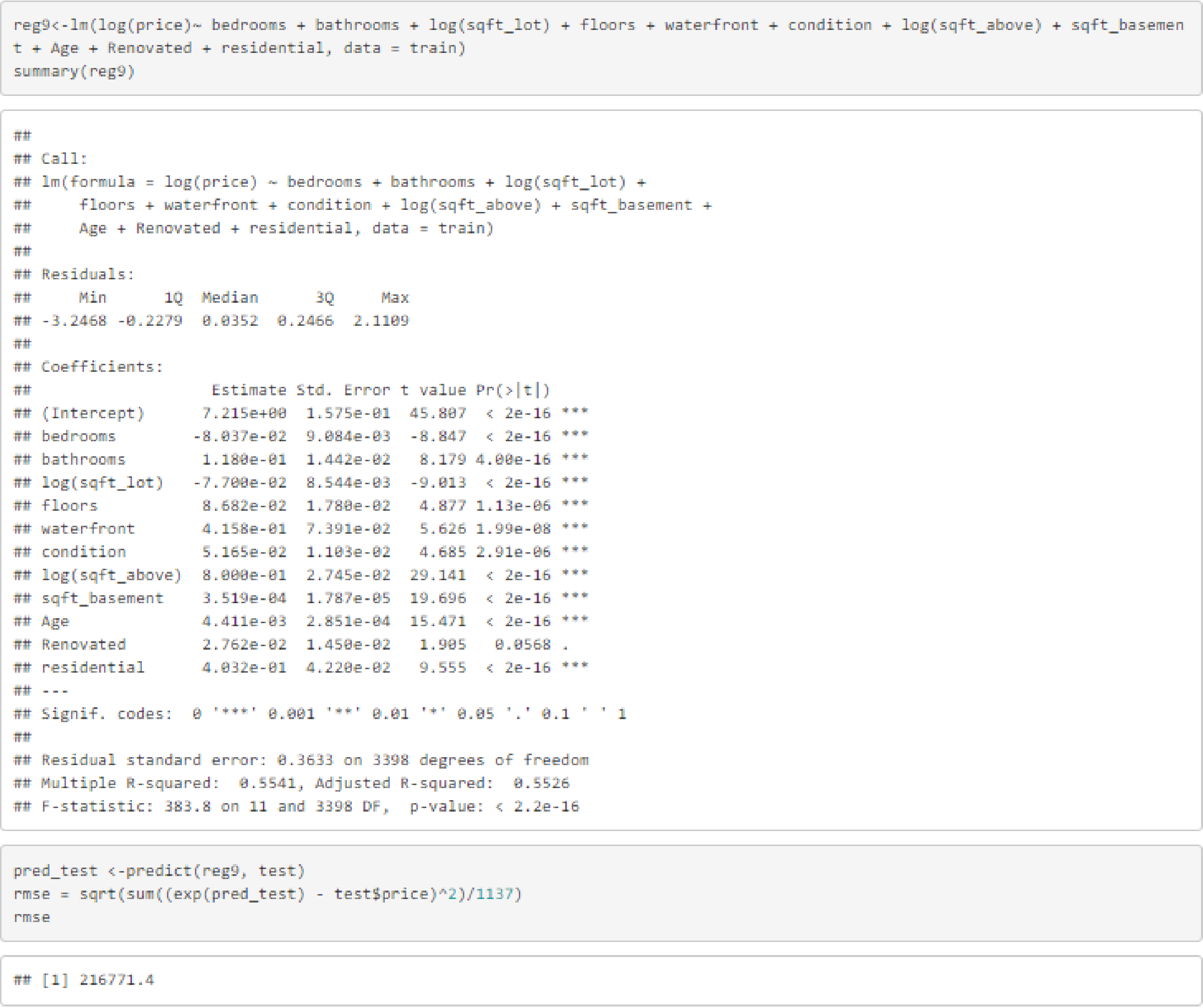
```
vif(reg8)
```

```
##                 bedrooms                bathrooms                   floors
##                 1.713106                 3.405215                 2.294689
##               waterfront                condition               sqft_above
##                 1.137472                 1.490299                 2.914730
##            sqft_basement                      Age                Renovated
##                 1.841469                 2.510030                 1.319392
##            city_Bellevue       city_Black.Diamond             city_Bothell
##                 1.376855                 1.014125                 1.042840
##              city_Burien           city_Carnation          city_Clyde.Hill
##                 1.099172                 1.035701                 1.030440
##              city_Duvall           city_Fall.City city_Inglewood.Finn.Hill
##                 1.057424                 1.014635                 1.004198
##            city_Issaquah             city_Kenmore            city_Kirkland
##                 1.269729                 1.080902                 1.232023
##    city_Lake.Forest.Park        city_Maple.Valley              city_Medina
##                 1.048408                 1.140782                 1.036326
##       city_Mercer.Island           city_Newcastle       city_Normandy.Park
##                 1.158777                 1.069284                 1.031853
##          city_North.Bend          city_Ravensdale             city_Redmond
##                 1.064203                 1.017290                 1.318376
##              city_Renton           city_Sammamish             city_Seattle
##                 1.359173                 1.260380                 2.935977
##           city_Shoreline          city_Snoqualmie     city_Snoqualmie.Pass
##                 1.194129                 1.120394                 1.004539
##              city_Vashon         city_Woodinville              city_Algona
##                 1.141996                 1.159309                 1.009136
##             city_Tukwila
##                 1.060871
```

```
pred_test <-predict(reg8, test)
rmse = sqrt(sum((exp(pred_test) - test$price)^2)/1137)
rmse
```

```
## [1] 186484.1
```

Table 9: Regression results - other variables + log(sqft_lot) + log(sqft_above)



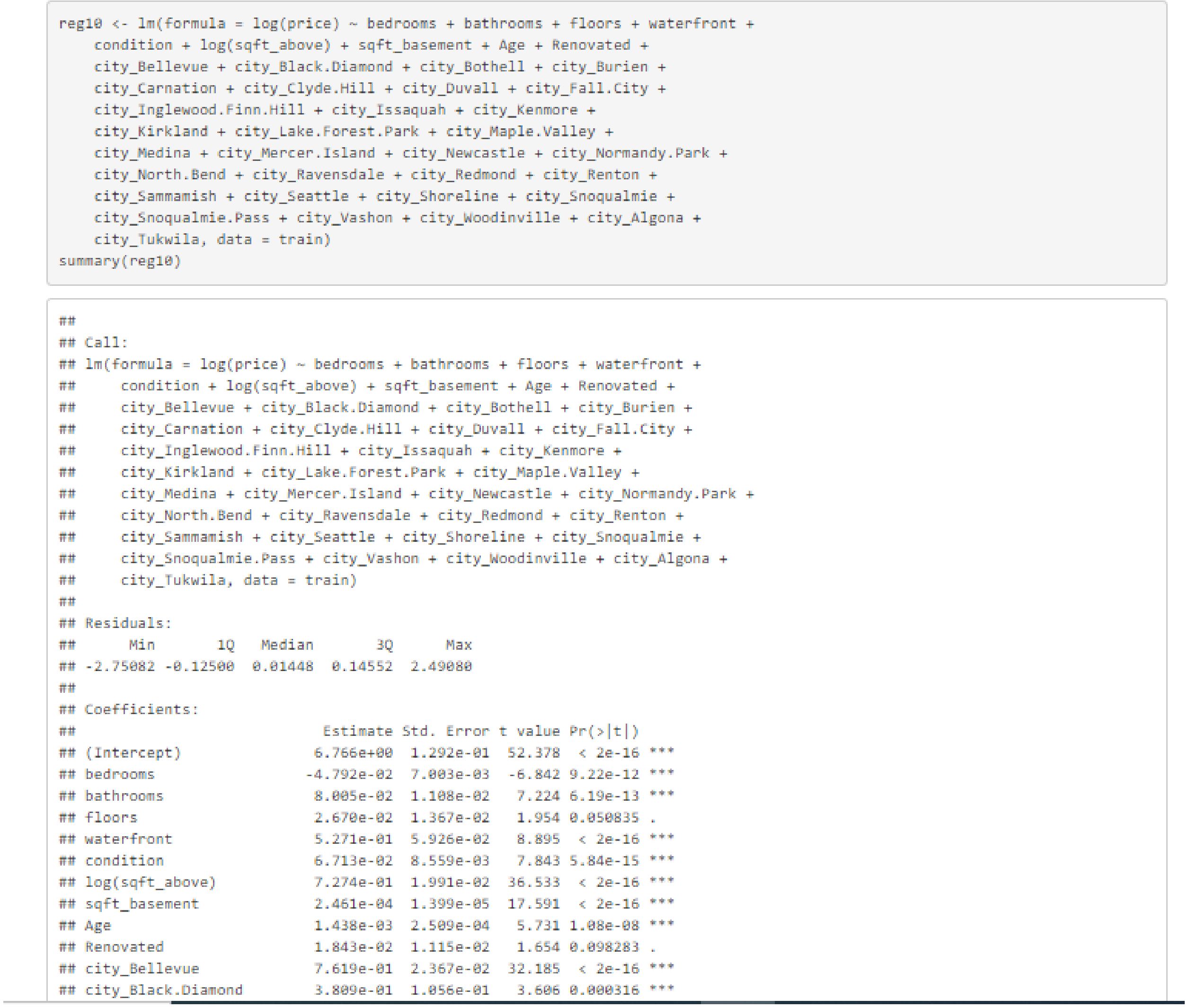
```
reg9<-lm(log(price)~ bedrooms + bathrooms + log(sqft_lot) + floors + waterfront + condition + log(sqft_above) + sqft_basemen
t + Age + Renovated + residential, data = train)
summary(reg9)
```

```
##
## Call:
## lm(formula = log(price) ~ bedrooms + bathrooms + log(sqft_lot) +
##     floors + waterfront + condition + log(sqft_above) + sqft_basement +
##     Age + Renovated + residential, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2468 -0.2279  0.0352  0.2466  2.1109
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.215e+00  1.575e-01  45.807  < 2e-16 ***
## bedrooms        -8.037e-02  9.084e-03  -8.847  < 2e-16 ***
## bathrooms        1.180e-01  1.442e-02   8.179 4.00e-16 ***
## log(sqft_lot)   -7.700e-02  8.544e-03  -9.013  < 2e-16 ***
## floors           8.682e-02  1.780e-02   4.877 1.13e-06 ***
## waterfront       4.158e-01  7.391e-02   5.626 1.99e-08 ***
## condition        5.165e-02  1.103e-02   4.685 2.91e-06 ***
## log(sqft_above)  8.000e-01  2.745e-02  29.141  < 2e-16 ***
## sqft_basement    3.519e-04  1.787e-05  19.696  < 2e-16 ***
## Age              4.411e-03  2.851e-04  15.471  < 2e-16 ***
## Renovated        2.762e-02  1.450e-02   1.905   0.0568 .
## residential      4.032e-01  4.220e-02   9.555  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3633 on 3398 degrees of freedom
## Multiple R-squared:  0.5541, Adjusted R-squared:  0.5526
## F-statistic: 383.8 on 11 and 3398 DF,  p-value: < 2.2e-16
```

```
pred_test <-predict(reg9, test)
rmse = sqrt(sum((exp(pred_test) - test$price)^2)/1137)
rmse
```

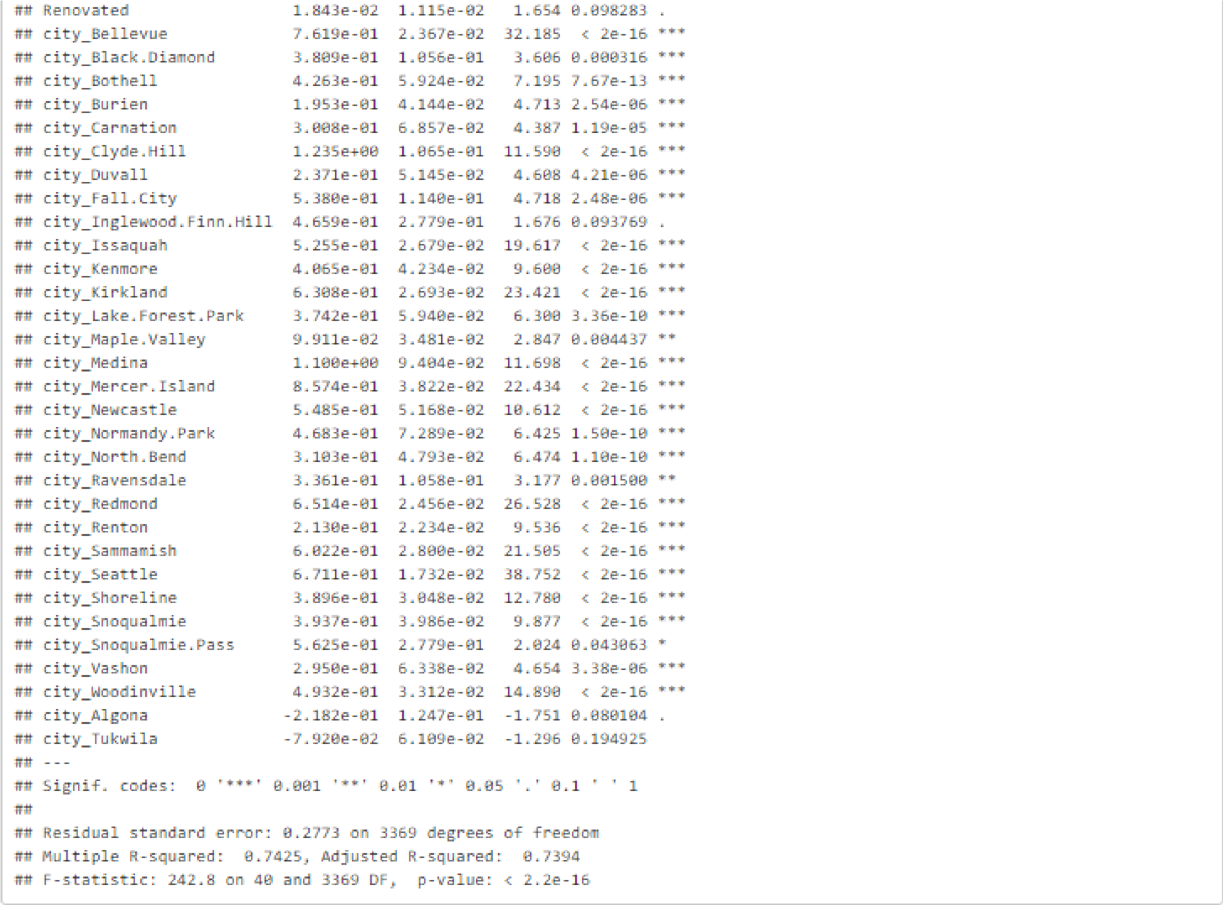
```
## [1] 216771.4
```

Table 10: Regression results - All variables in model 8 with transformation log (sqft_above)

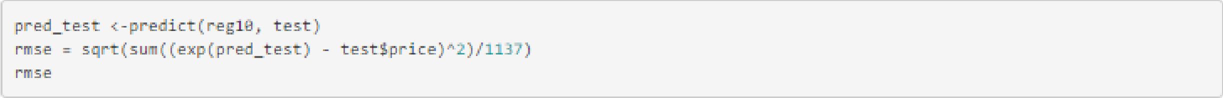


```
reg10 <- lm(formula = log(price) ~ bedrooms + bathrooms + floors + waterfront +
    condition + log(sqft_above) + sqft_basement + Age + Renovated +
    city_Bellevue + city_Black.Diamond + city_Bothell + city_Burien +
    city_Carnation + city_Clyde.Hill + city_Duvall + city_Fall.City +
    city_Inglewood.Finn.Hill + city_Issaquah + city_Kenmore +
    city_Kirkland + city_Lake.Forest.Park + city_Maple.Valley +
    city_Medina + city_Mercer.Island + city_Newcastle + city_Normandy.Park +
    city_North.Bend + city_Ravensdale + city_Redmond + city_Renton +
    city_Sammamish + city_Seattle + city_Shoreline + city_Snoqualmie +
    city_Snoqualmie.Pass + city_Vashon + city_Woodinville + city_Algona +
    city_Tukwila, data = train)
summary(reg10)
```

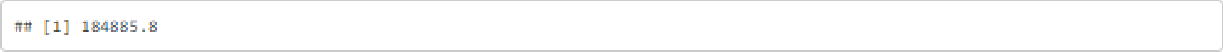
```
##
## Call:
## lm(formula = log(price) ~ bedrooms + bathrooms + floors + waterfront +
##     condition + log(sqft_above) + sqft_basement + Age + Renovated +
##     city_Bellevue + city_Black.Diamond + city_Bothell + city_Burien +
##     city_Carnation + city_Clyde.Hill + city_Duvall + city_Fall.City +
##     city_Inglewood.Finn.Hill + city_Issaquah + city_Kenmore +
##     city_Kirkland + city_Lake.Forest.Park + city_Maple.Valley +
##     city_Medina + city_Mercer.Island + city_Newcastle + city_Normandy.Park +
##     city_North.Bend + city_Ravensdale + city_Redmond + city_Renton +
##     city_Sammamish + city_Seattle + city_Shoreline + city_Snoqualmie +
##     city_Snoqualmie.Pass + city_Vashon + city_Woodinville + city_Algona +
##     city_Tukwila, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.75082 -0.12500  0.01448  0.14552  2.49080
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.766e+00  1.292e-01  52.378  < 2e-16 ***
## bedrooms                 -4.792e-02  7.003e-03  -6.842 9.22e-12 ***
## bathrooms                 8.005e-02  1.108e-02   7.224 6.19e-13 ***
## floors                    2.670e-02  1.367e-02   1.954 0.050835 .
## waterfront                5.271e-01  5.926e-02   8.895  < 2e-16 ***
## condition                 6.713e-02  8.559e-03   7.843 5.84e-15 ***
## log(sqft_above)           7.274e-01  1.991e-02  36.533  < 2e-16 ***
## sqft_basement             2.461e-04  1.399e-05  17.591  < 2e-16 ***
## Age                       1.438e-03  2.509e-04   5.731 1.08e-08 ***
## Renovated                 1.843e-02  1.115e-02   1.654 0.098283 .
## city_Bellevue             7.619e-01  2.367e-02  32.185  < 2e-16 ***
## city_Black.Diamond        3.809e-01  1.056e-01   3.606 0.000316 ***
```



```
## Renovated                1.843e-02  1.115e-02   1.654 0.098283 .  
## city_Bellevue            7.619e-01  2.367e-02  32.185  < 2e-16 ***
## city_Black.Diamond       3.809e-01  1.056e-01   3.606 0.000316 ***
## city_Bothell             4.263e-01  5.924e-02   7.195 7.67e-13 ***
## city_Burien              1.953e-01  4.144e-02   4.713 2.54e-06 ***
## city_Carnation           3.008e-01  6.857e-02   4.387 1.19e-05 ***
## city_Clyde.Hill          1.235e+00  1.065e-01  11.590  < 2e-16 ***
## city_Duvall              2.371e-01  5.145e-02   4.608 4.21e-06 ***
## city_Fall.City           5.380e-01  1.140e-01   4.718 2.48e-06 ***
## city_Inglewood.Finn.Hill 4.659e-01  2.779e-01   1.676 0.093769 .  
## city_Issaquah            5.255e-01  2.679e-02  19.617  < 2e-16 ***
## city_Kenmore             4.065e-01  4.234e-02   9.600  < 2e-16 ***
## city_Kirkland            6.308e-01  2.693e-02  23.421  < 2e-16 ***
## city_Lake.Forest.Park    3.742e-01  5.940e-02   6.300 3.36e-10 ***
## city_Maple.Valley        9.911e-02  3.481e-02   2.847 0.004437 ** 
## city_Medina              1.100e+00  9.404e-02  11.698  < 2e-16 ***
## city_Mercer.Island       8.574e-01  3.822e-02  22.434  < 2e-16 ***
## city_Newcastle           5.485e-01  5.168e-02  10.612  < 2e-16 ***
## city_Normandy.Park       4.683e-01  7.289e-02   6.425 1.50e-10 ***
## city_North.Bend          3.103e-01  4.793e-02   6.474 1.10e-10 ***
## city_Ravensdale          3.361e-01  1.058e-01   3.177 0.001500 ** 
## city_Redmond             6.514e-01  2.456e-02  26.528  < 2e-16 ***
## city_Renton              2.130e-01  2.234e-02   9.536  < 2e-16 ***
## city_Sammamish           6.022e-01  2.800e-02  21.505  < 2e-16 ***
## city_Seattle             6.711e-01  1.732e-02  38.752  < 2e-16 ***
## city_Shoreline           3.896e-01  3.048e-02  12.780  < 2e-16 ***
## city_Snoqualmie          3.937e-01  3.986e-02   9.877  < 2e-16 ***
## city_Snoqualmie.Pass     5.625e-01  2.779e-01   2.024 0.043063 *  
## city_Vashon              2.950e-01  6.338e-02   4.654 3.38e-06 ***
## city_Woodinville         4.932e-01  3.312e-02  14.890  < 2e-16 ***
## city_Algona             -2.182e-01  1.247e-01  -1.751 0.080104 .  
## city_Tukwila            -7.920e-02  6.109e-02  -1.296 0.194925    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2773 on 3369 degrees of freedom
## Multiple R-squared:  0.7425, Adjusted R-squared:  0.7394 
## F-statistic: 242.8 on 40 and 3369 DF,  p-value: < 2.2e-16
```



```
pred_test <-predict(reg10, test)
rmse = sqrt(sum((exp(pred_test) - test$price)^2)/1137)
rmse
```



```
## [1] 184885.8
```

References

Afonso, Bruno & Melo, Luckeciano & Dihanster, Willian & Sousa, Samuel & Berton, L. (2019). Housing Prices Prediction with a Deep Learning and Random Forest Ensemble.

Gupta, R., A. Kabundi, and S.M. Miller. “Forecasting the US Real House Price Index: Structural and Non-Structural Models with and without Fundamentals.” Working paper No. 200927, Dept. of Econ., University of Pretonia, 2009a.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin. “Do financial variables help forecasting inflation and real activity in the euro area?” Journal of Monetary Economics 6(2003): 1243-1255.

Li, Yarui & Leatham, David J., 2010. "Forecasting Housing Prices: Dynamic Factor Model versus LBVAR Model," 2011 Annual Meeting, July 24-26, 2011, Pittsburgh, Pennsylvania 103667, Agricultural and Applied Economics Association.

Poursaeed, O., Matera, T., and Belongie, S. (2018). Vision-based real estate price estimation. Machine Vision and Applications, 29(4):667–676.

Sirignano, J., Sadhwani, A., and Giesecke, K. (2016). Deep learning for mortgage risk. SSRN Electronic Journal, pages 1–75.

Stock, J.H., and M.W. Watson. “Forecasting Output and Inflation: The Role of Asset Prices.” Journal of Economic Literature 3(2003):788-829.

Wu, L. and Brynjolfsson, E. (2015). The future of prediction: How google searches foreshadow housing prices and sales. In Economic analysis of the digital economy, pages 89–118. University of Chicago Press.