Prediction of House Prices Using Multiple Linear Regression

Aniket Dalvi

Ce Ji

Prasad Marathe

Qimei Wang

Harrisburg University of Science and Technology

February 2018

Contents

Introduction

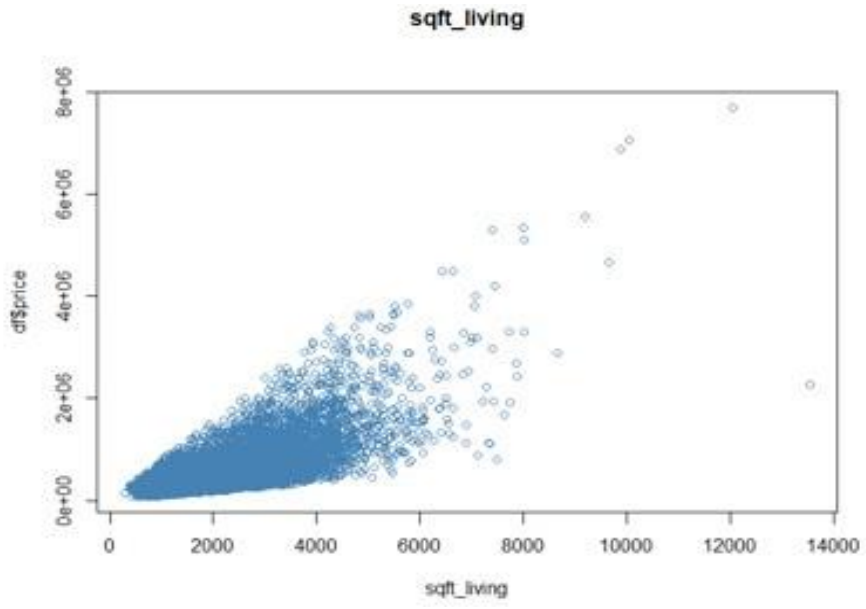
Real estate industry is a complex system. House price is the most important element, and house prices and the living standard of the residents are closely related. There are a lot of factors that affect the price of residential house. The Seattle's median price has grown over 10% from last year. More and more homes are edging above the $1 million mark in Seattle area. Our project utilizes data of real estate in King County, Washington, which includes Seattle. We believe the study in sale price of real estate in King County will be representative. This is the dataset released under COO public domain on the Kaggle website. This dataset contains 21613 observations and 19 variables such as sale price, number of bedrooms, number of bathrooms, floors, sale date etc. It includes homes sold between May 2014 and May 2015. Our goal is to predict house prices in King County by identifying significant factors using multiple linear regression techniques.

Exploratory Data Analysis (EDA)

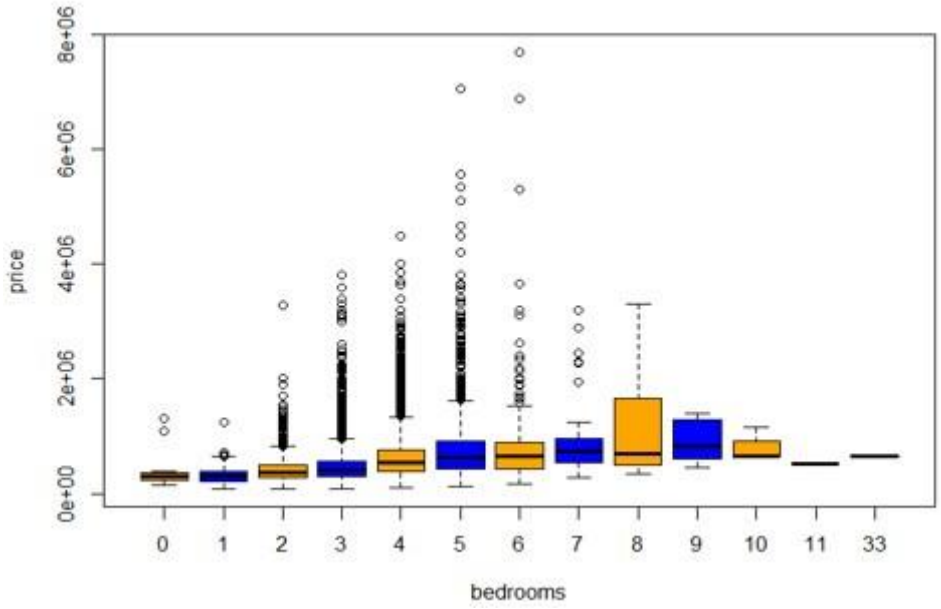
The data set contains 21613 observations and 19 variables. The response variable is house price. We cleaned the data before feeding them into our analytics. We checked for any missing data and outlier. There is no missing data, but one house has 33 bedrooms and 1.75 bathroom. Mean square feet living of 3 bedrooms house is 1805, while the square feet living of 33 bedrooms house is 1620. So, we updated outlier in bedroom with value 33 to 3. Then we found out correlation among all features and house price. The correlation coefficients are shown as below:

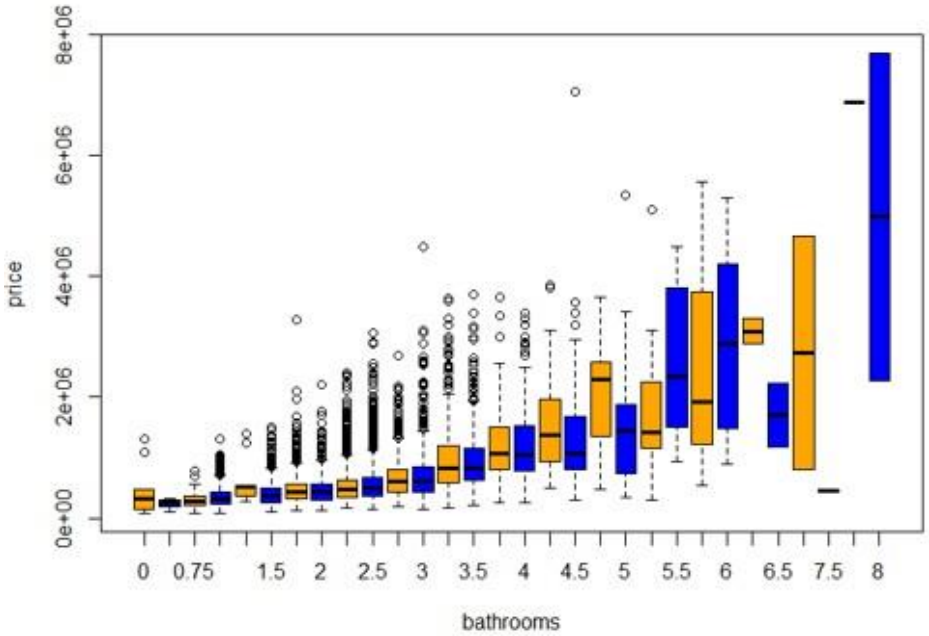
	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
price	1	0.31	0.53	0.7	0.09	0.26	0.27	0.4	0.04	0.67	0.61	0.32	0.05	0.13	-0.05	0.31	0.02	0.59	0.08
bedrooms	0.31	1	0.52	0.58	0.03	0.18	-0.01	0.08	0.03	0.36	0.48	0.3	0.15	0.02	-0.15	-0.01	0.13	0.39	0.03
bathrooms	0.53	0.52	1	0.75	0.09	0.5	0.06	0.19	-0.12	0.66	0.69	0.28	0.51	0.05	-0.2	0.02	0.22	0.57	0.09
sqft_living	0.7	0.58	0.75	1	0.17	0.35	0.1	0.28	-0.06	0.76	0.88	0.44	0.32	0.06	-0.2	0.05	0.24	0.76	0.18
sqft_lot	0.09	0.03	0.09	0.17	1	-0.01	0.02	0.07	-0.01	0.11	0.18	0.02	0.05	0.01	-0.13	-0.09	0.23	0.14	0.72
floors	0.26	0.18	0.5	0.35	-0.01	1	0.02	0.03	-0.26	0.46	0.52	-0.25	0.49	0.01	-0.06	0.05	0.13	0.28	-0.01
waterfront	0.27	-0.01	0.06	0.1	0.02	0.02	1	0.4	0.02	0.08	0.07	0.08	-0.03	0.09	0.03	-0.01	-0.04	0.09	0.03
view	0.4	0.08	0.19	0.28	0.07	0.03	0.4	1	0.05	0.25	0.17	0.28	-0.05	0.1	0.08	0.01	-0.08	0.28	0.07
condition	0.04	0.03	-0.12	-0.06	-0.01	-0.26	0.02	0.05	1	-0.14	-0.16	0.17	-0.36	-0.06	0	-0.01	-0.11	-0.09	0
grade	0.67	0.36	0.66	0.76	0.11	0.46	0.08	0.25	-0.14	1	0.76	0.17	0.45	0.01	-0.18	0.11	0.2	0.71	0.12
sqft_above	0.61	0.48	0.69	0.88	0.18	0.52	0.07	0.17	-0.16	0.76	1	-0.05	0.42	0.02	-0.26	0	0.34	0.73	0.19
sqft_basement	0.32	0.3	0.28	0.44	0.02	-0.25	0.08	0.28	0.17	0.17	-0.05	1	-0.13	0.07	0.07	0.11	-0.14	0.2	0.02
yr_built	0.05	0.15	0.51	0.32	0.05	0.49	-0.03	-0.05	-0.36	0.45	0.42	-0.13	1	-0.22	-0.35	-0.15	0.41	0.33	0.07
yr_renovated	0.13	0.02	0.05	0.06	0.01	0.01	0.09	0.1	-0.06	0.01	0.02	0.07	-0.22	1	0.06	0.03	-0.07	0	0.01
zipcode	-0.05	-0.15	-0.2	-0.2	-0.13	-0.06	0.03	0.08	0	-0.18	-0.26	0.07	-0.35	0.06	1	0.27	-0.56	-0.28	-0.15
lat	0.31	-0.01	0.02	0.05	-0.09	0.05	-0.01	0.01	-0.01	0.11	0	0.11	-0.15	0.03	0.27	1	-0.14	0.05	-0.09
long	0.02	0.13	0.22	0.24	0.23	0.13	-0.04	-0.08	-0.11	0.2	0.34	-0.14	0.41	-0.07	-0.56	-0.14	1	0.33	0.25
sqft_living15	0.59	0.39	0.57	0.76	0.14	0.28	0.09	0.28	-0.09	0.71	0.73	0.2	0.33	0	-0.28	0.05	0.33	1	0.18
sqft_lot15	0.08	0.03	0.09	0.18	0.72	-0.01	0.03	0.07	0	0.12	0.19	0.02	0.07	0.01	-0.15	-0.09	0.25	0.18	1

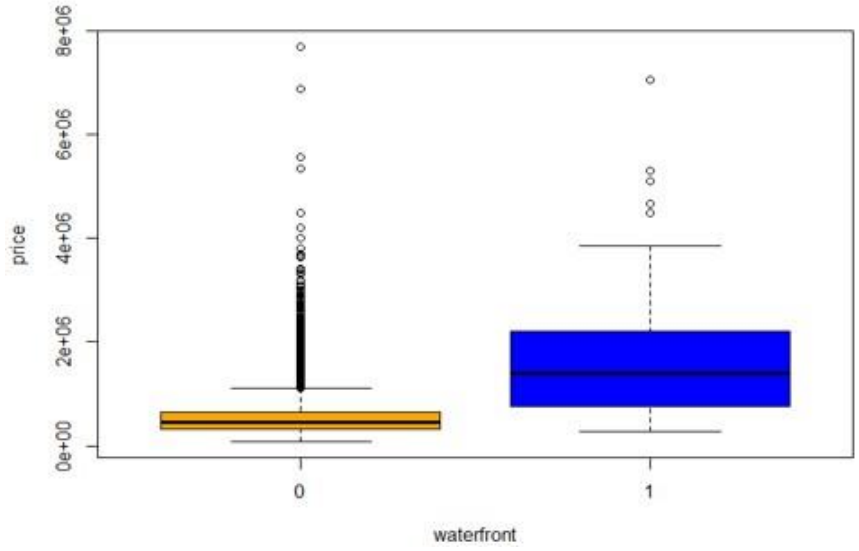
From figure above, sqft_living15 is highly correlated with sqft_living, and sqft_lot15 is highly correlated with sqft_lot, so sqft_living15 and sqft_lot15 were dropped from the model. Sqft_living is the sum of sqft_above and sqft_basement. Clearly there is multicollinearity at play among independent variables, so we only used sqft_living in our analytics. We dropped 'date house was sold' and 'ID' as we don't perform any analysis on time-period. Variables like lat (latitude) and long (longitude) are not very intuitive from customer perspective, so latitude and longitude were dropped. From the figure below, we can see a high correlation between price and sqft_living. The Correlation Coefficient, r equals to 0.7.

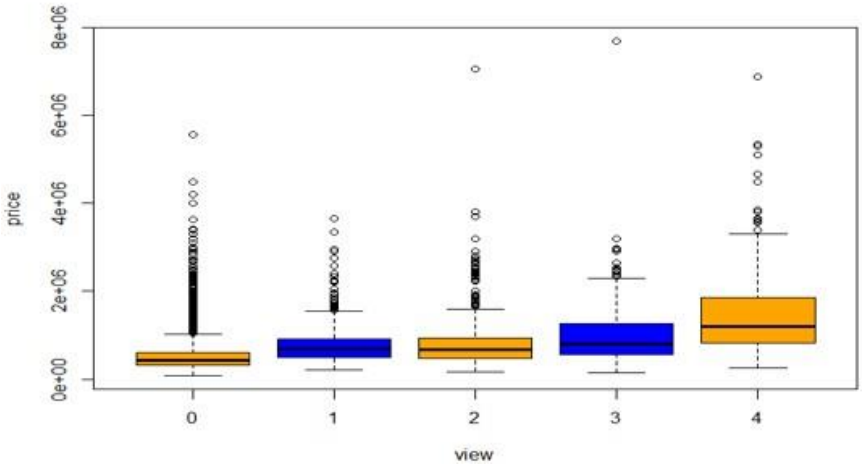


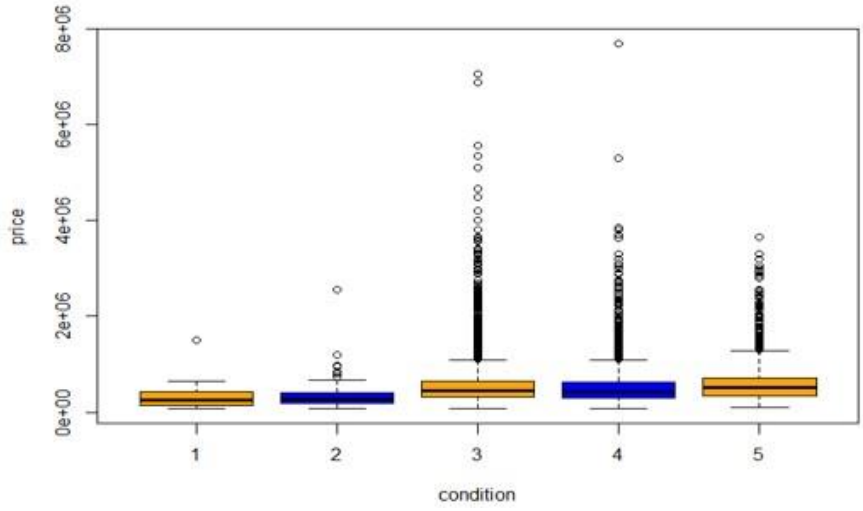
The box plots for categorical variables versus prices are shown as below:

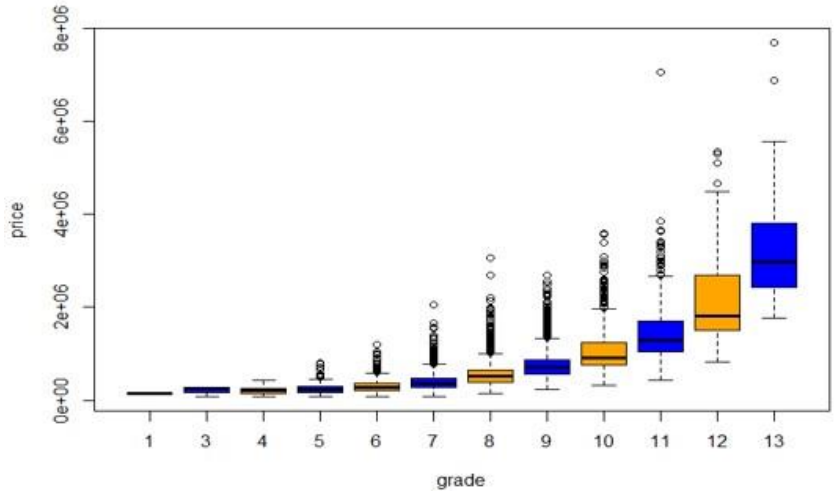


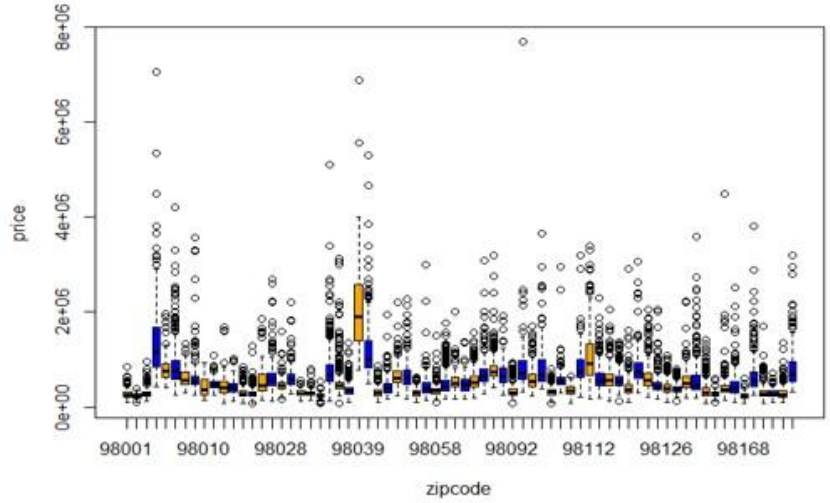


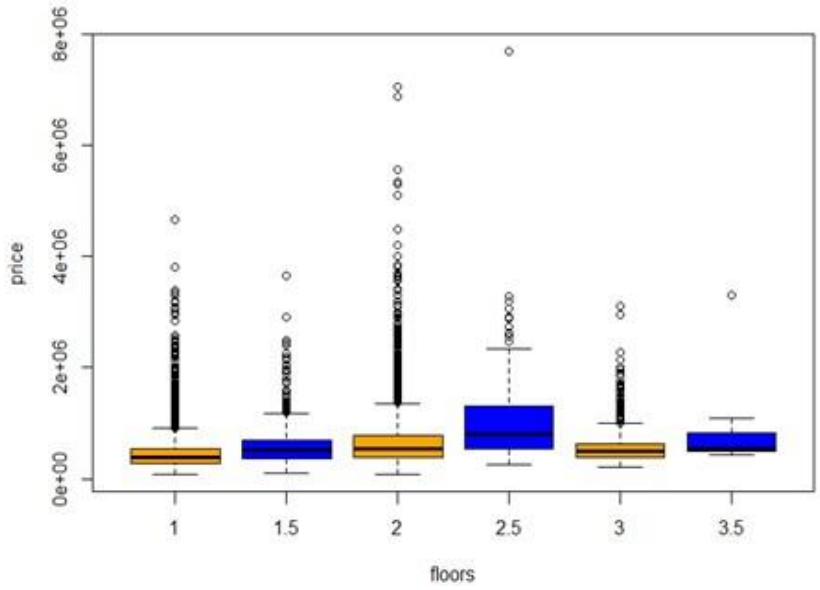

price
0e+00
2e+06
4e+06
6e+06
8e+06
0
1
waterfront


price
0e+00
2e+06
4e+06
6e+06
8e+06
0
1
2
3
4
view


price
0e+00
2e+06
4e+06
6e+06
8e+06
1
2
3
4
5
condition


price
0e+00
2e+06
4e+06
6e+06
8e+06
1
3
4
5
6
7
8
9
10
11
12
13
grade


price
0e+00
2e+06
4e+06
6e+06
8e+06
98001
98010
98028
98039
98058
98092
98112
98126
98168
zipcode


price
0e+00
2e+06
4e+06
6e+06
8e+06
1
1.5
2
2.5
3
3.5
floors

Based on domain knowledge the team had about house prices and correlation between price and independent variables, we selected a subset of regressors for building the model, which included 10 variables. Below is a subset of independent variables we used for building the model.

Variable Name	Data Type	Variable Description
bedrooms	Factor	Number of bedrooms
bathrooms	Factor	Number of bathrooms
sqft_living	Num	Square feet of living space
sqft_lot	Num	Square feet of the lot
floors	Factor	Number of floors
waterfront	Factor	Number of water front view
view	Factor	Number of the views of the house
condition	Factor	Overall condition
grade	Factor	Grade of the house
zipcode	Factor	Zip code

Inference

Based on our domain knowledge we made three hypotheses as below:

1. Effect of square feet of living space: Square feet of living is the important criteria for anyone looking to purchase a house.
2. Effect of bathrooms: Having more bathrooms will increase the price.
3. Effect of bedrooms: The basic building blocks of the house.

We will test our hypothesis using p-value approach while building the model and drop regressors that are not significant using backward elimination.

Modeling (The Best Model)

We divided the original data into training and testing. Training dataset has 17290 observations and represents 80% of the dataset. The testing/validation dataset has 4323 observations and represents 20% of the dataset. After selecting the best multiple linear regression model, we use the testing dataset for predicting house prices in King County.

```
# Now Selecting 80% of data as sample from total 'n' rows of the data
set.seed(1000)
sample <- sample.int(n=nrow(df), size = floor(0.80*nrow(df)), replace = F) # Sample without replacement

# Splitting train and test data
train <- df[sample, ]
test  <- df[-sample, ]
```

The result of model 1 is shown as below:

```
call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot +
    floors + waterfront + view + condition + grade + zipcode,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-2156073   -58415     1583    54279  3587936

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.143e+04  1.493e+05  -0.411 0.680808
```

```
sqft_living    1.467e+02  2.718e+00  53.994  < 2e-16 ***
sqft_lot       2.045e-01  3.212e-02   6.367 1.98e-10 ***
floors1.5      7.167e+03  4.390e+03   1.632 0.102599
floors2       -5.053e+03  3.455e+03  -1.462 0.143629
floors2.5      7.197e+04  1.356e+04   5.307 1.13e-07 ***
floors3       -6.775e+04  7.833e+03  -8.650  < 2e-16 ***
floors3.5      3.813e+04  5.727e+04   0.666 0.505557
waterfront1    6.046e+05  1.692e+04  35.735  < 2e-16 ***
view1          1.016e+05  9.593e+03  10.594  < 2e-16 ***
view2          7.495e+04  5.653e+03  13.259  < 2e-16 ***
view3          1.442e+05  7.873e+03  18.318  < 2e-16 ***
view4          2.717e+05  1.219e+04  22.283  < 2e-16 ***
condition2     5.679e+04  3.280e+04   1.731 0.083436 .
condition3     6.442e+04  3.069e+04   2.099 0.035828 *
condition4     8.642e+04  3.072e+04   2.813 0.004912 **
condition5     1.313e+05  3.091e+04   4.248 2.17e-05 ***
grade3         3.731e+04  1.826e+05   0.204 0.838074
grade4        -5.012e+04  1.652e+05  -0.303 0.761562
grade5        -8.286e+04  1.624e+05  -0.510 0.609932
grade6        -9.049e+04  1.623e+05  -0.557 0.577223
grade7        -8.205e+04  1.623e+05  -0.506 0.613197
grade8        -4.737e+04  1.624e+05  -0.292 0.770483
grade9         3.378e+04  1.624e+05   0.208 0.835261
grade10        1.557e+05  1.625e+05   0.958 0.338043
grade11        3.614e+05  1.628e+05   2.220 0.026435 *
grade12        7.322e+05  1.635e+05   4.478 7.60e-06 ***
grade13        1.632e+06  1.743e+05   9.361  < 2e-16 ***
zipcode98002   3.439e+03  1.458e+04   0.236 0.813465
zipcode98003  -6.526e+03  1.324e+04  -0.493 0.622033
zipcode98004   7.856e+05  1.291e+04  60.848  < 2e-16 ***

Residual standard error: 147900 on 17154 degrees of freedom
Multiple R-squared:  0.8384,    Adjusted R-squared:  0.8371
F-statistic: 659.2 on 135 and 17154 DF,  p-value: < 2.2e-16
```

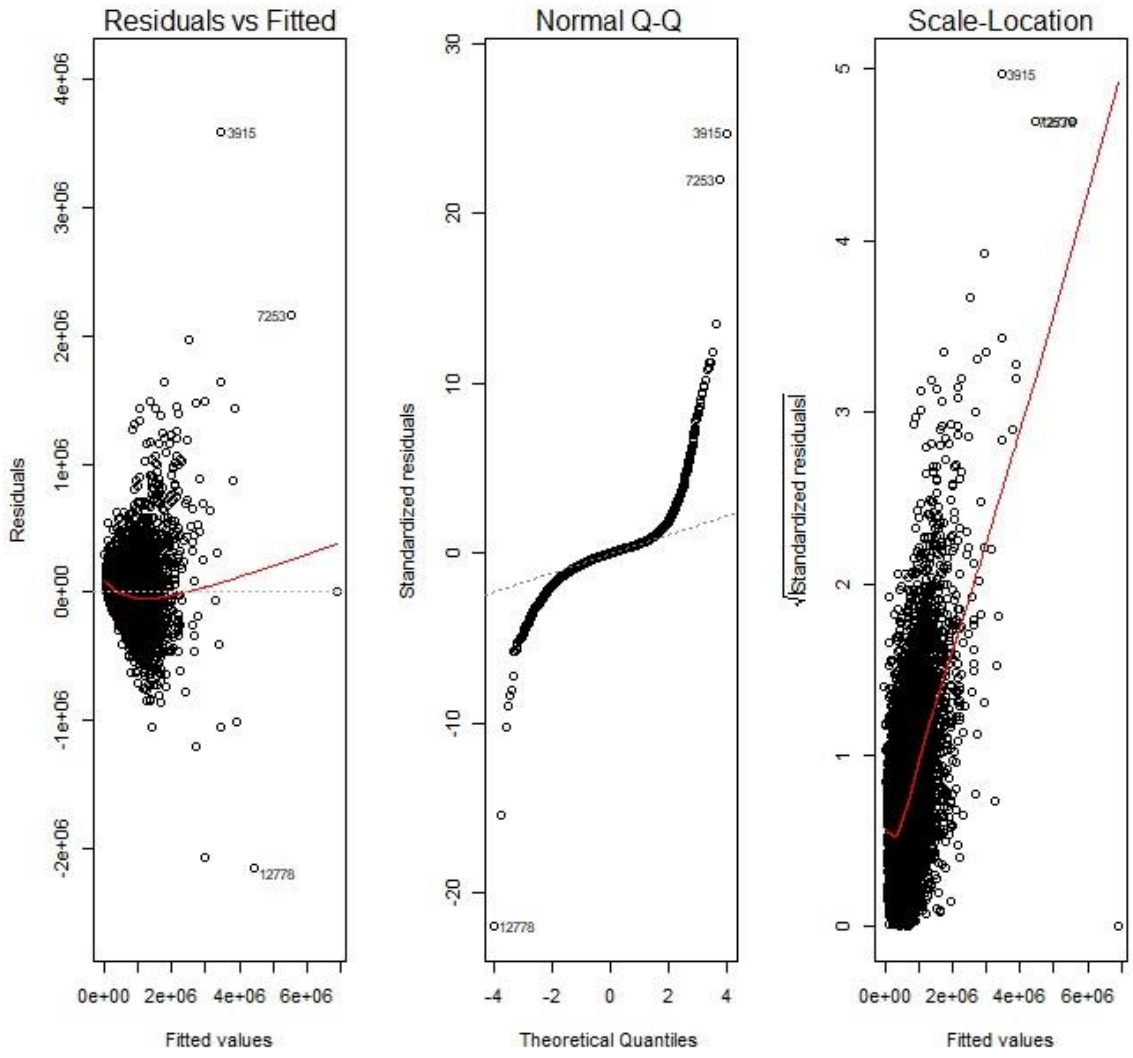
Bedroom is not as significant compared to other variables in the model as p-values are greater than alpha (Fail to Reject H0 for bedrooms ~ Beta co-efficient for bedroom is 0). R-Squared equals to 0.8384. Adjusted R-Squared equals to 0.8371. Residual Error equals to 147900. We used backward elimination method and we eliminated bedrooms in next iteration.

Model Adequacy Check for Model 1:

The assumptions made for linear regression are:

1. Relationship between price (response) and independent variables is linear, at least approximately.
2. The errors are normally distributed.
3. The errors have constant variance.

Normal Probability plot for model 1 violates assumptions 2 while Residual Vs Fitted plot violates Constant error variance assumption, hence model 1 is not adequate.

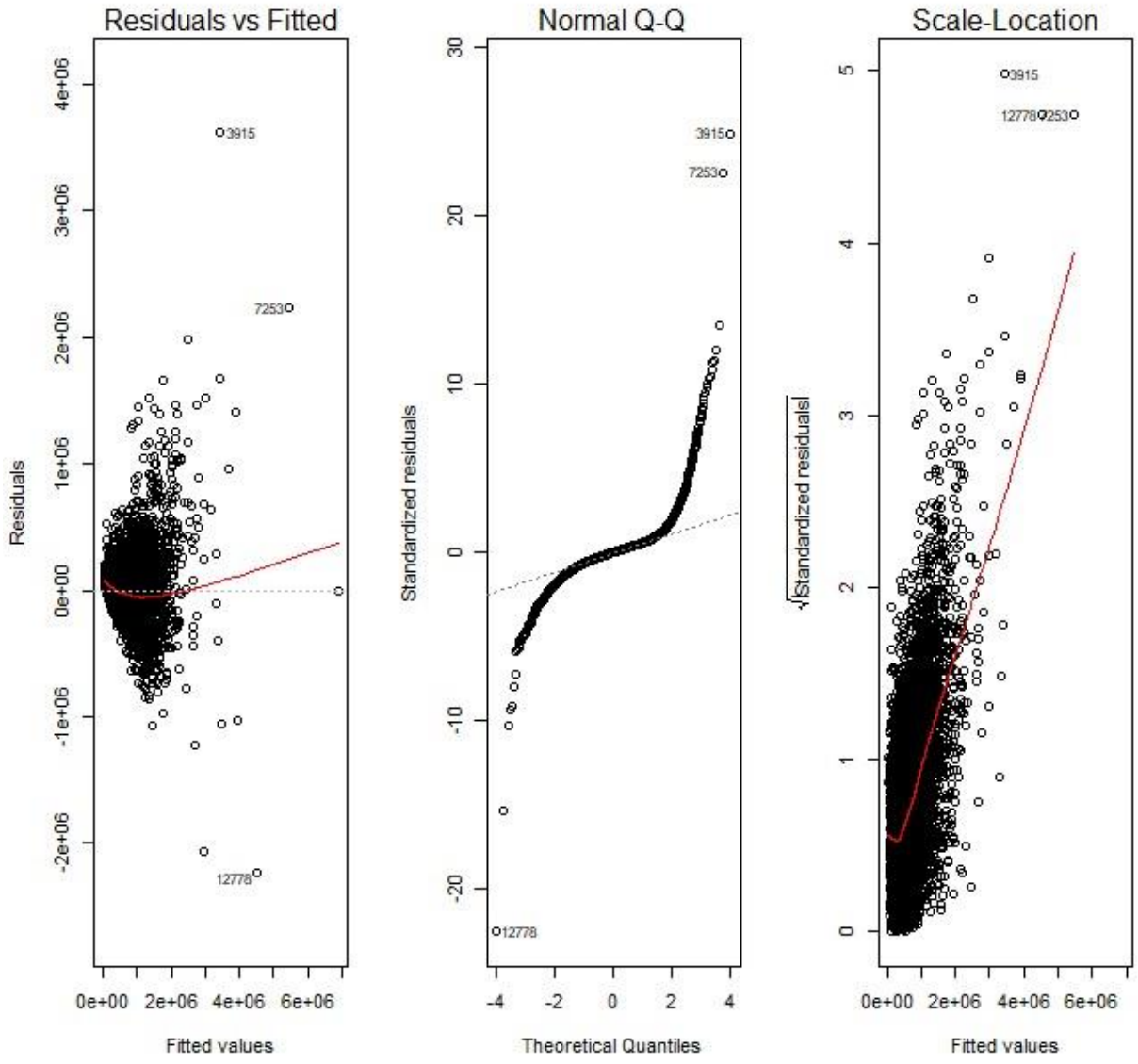


The result of model 2 is shown as below:

```
Call:
lm(formula = price ~ bathrooms + sqft_living + sqft_lot + floors +
    waterfront + view + condition + grade + zipcode, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-2232331   -59093     1847    54372  3618886

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
```

```
sqft_living    1.397e+02  2.488e+00  56.172  < 2e-16 ***
sqft_lot       2.122e-01  3.217e-02   6.595 4.38e-11 ***
floors1.5      6.099e+03  4.354e+03   1.401 0.161312
floors2       -4.948e+03  3.463e+03  -1.429 0.153076
floors2.5      7.125e+04  1.358e+04   5.245 1.58e-07 ***
floors3       -6.466e+04  7.846e+03  -8.241  < 2e-16 ***
floors3.5      5.274e+04  5.700e+04   0.925 0.354862
waterfront1    6.128e+05  1.695e+04  36.158  < 2e-16 ***
view1          1.016e+05  9.624e+03  10.556  < 2e-16 ***
view2          7.500e+04  5.669e+03  13.230  < 2e-16 ***
view3          1.471e+05  7.895e+03  18.637  < 2e-16 ***
view4          2.696e+05  1.223e+04  22.048  < 2e-16 ***
condition2     5.918e+04  3.292e+04   1.797 0.072283 .
condition3     6.736e+04  3.080e+04   2.187 0.028760 *
condition4     8.934e+04  3.083e+04   2.898 0.003762 **
condition5     1.349e+05  3.102e+04   4.348 1.38e-05 ***
grade3         5.454e+04  1.765e+05   0.309 0.757302
grade4        -4.982e+04  1.653e+05  -0.301 0.763109
grade5        -7.572e+04  1.623e+05  -0.466 0.640891
grade6        -7.922e+04  1.623e+05  -0.488 0.625426
grade7        -6.784e+04  1.623e+05  -0.418 0.675888
grade8        -3.156e+04  1.623e+05  -0.194 0.845835
grade9         5.263e+04  1.624e+05   0.324 0.745850
grade10        1.799e+05  1.625e+05   1.107 0.268154
grade11        3.899e+05  1.627e+05   2.396 0.016589 *
grade12        7.731e+05  1.634e+05   4.730 2.26e-06 ***
grade13        1.695e+06  1.743e+05   9.725  < 2e-16 ***
zipcode98002   2.922e+03  1.463e+04   0.200 0.841684
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 148400 on 17164 degrees of freedom
Multiple R-squared:  0.837,     Adjusted R-squared:  0.8358
F-statistic: 705.3 on 125 and 17164 DF,  p-value: < 2.2e-16
```

In model 2, R-Squared equals to 0.837. Adjusted R-Squared equals to 0.8358. Residual error equals to 148400. Model 2 also violates constant error variance and error normality assumptions which is evident from the residual vs fitted and Normal probability plots respectively. For 3rd Iteration, we will use logarithmic transformation to reduce variability (residual error) and to make data conform to normality so that assumptions of multiple linear regression are satisfied.

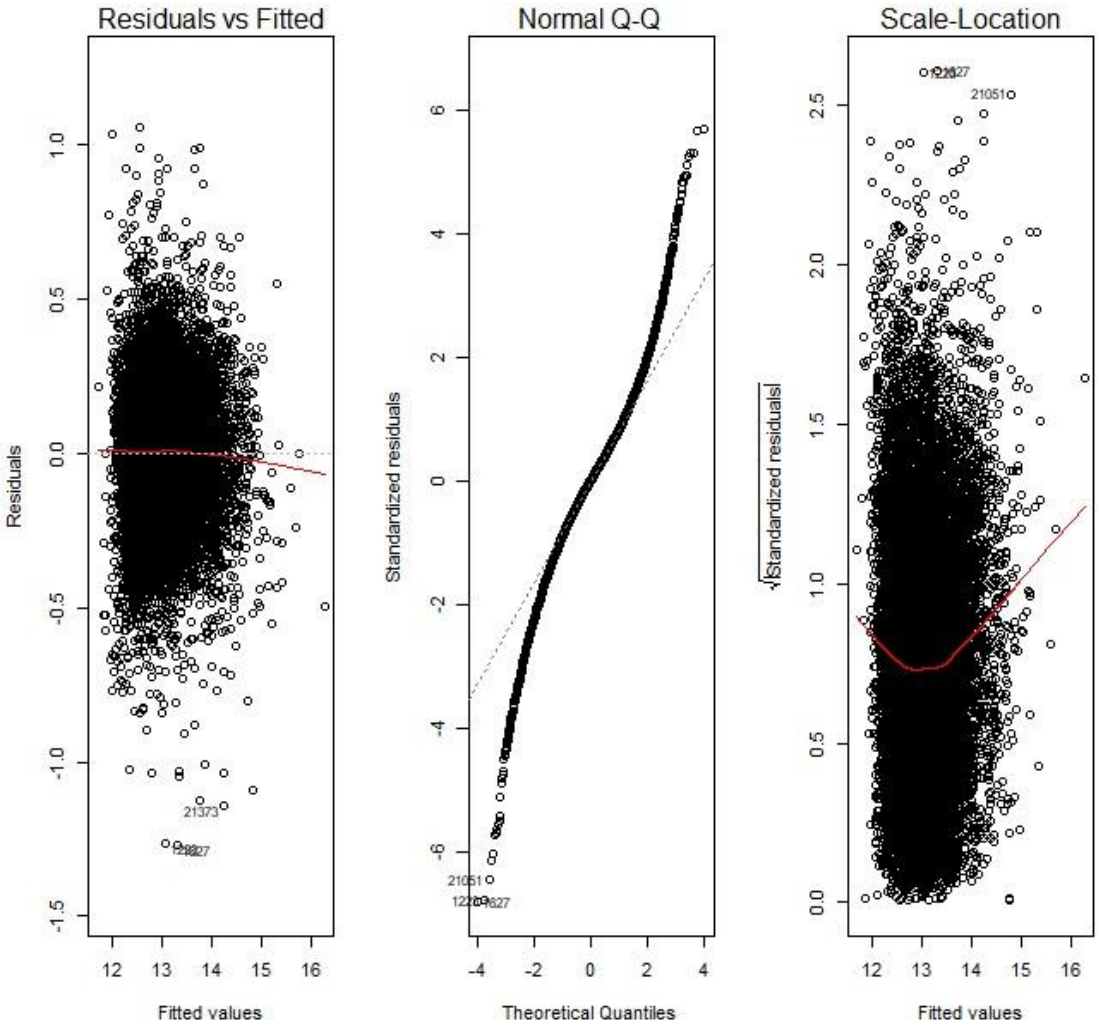


The result of model 3 is shown as below:

```
Call:
lm(formula = log(price) ~ bathrooms + sqft_living + sqft_lot +
    floors + waterfront + view + condition + grade + zipcode,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.26976 -0.09888  0.00614  0.10662  1.05269

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.136e+01  1.888e-01  60.136  < 2e-16 ***
```

```
sqft_living    2.147e-04  3.134e-06  68.507  < 2e-16 ***
sqft_lot       7.495e-07  4.053e-08  18.496  < 2e-16 ***
floors1.5      4.065e-02  5.485e-03   7.410 1.32e-13 ***
floors2       -2.933e-03  4.363e-03  -0.672 0.501441
floors2.5     -1.208e-03  1.711e-02  -0.071 0.943740
floors3       -1.293e-01  9.883e-03 -13.078  < 2e-16 ***
floors3.5     -1.048e-01  7.180e-02  -1.460 0.144362
waterfront1    4.532e-01  2.135e-02  21.230  < 2e-16 ***
view1          1.348e-01  1.212e-02  11.121  < 2e-16 ***
view2          1.205e-01  7.141e-03  16.877  < 2e-16 ***
view3          1.887e-01  9.945e-03  18.977  < 2e-16 ***
view4          3.014e-01  1.541e-02  19.566  < 2e-16 ***
condition2     5.543e-02  4.147e-02   1.336 0.181427
condition3     1.765e-01  3.880e-02   4.548 5.46e-06 ***
condition4     2.173e-01  3.884e-02   5.594 2.25e-08 ***
condition5     2.740e-01  3.907e-02   7.012 2.43e-12 ***
grade3         2.504e-01  2.223e-01   1.126 0.260128
grade4         6.470e-02  2.082e-01   0.311 0.755997
grade5         1.960e-01  2.045e-01   0.959 0.337797
grade6         2.868e-01  2.044e-01   1.403 0.160625
grade7         3.949e-01  2.044e-01   1.932 0.053397 .
grade8         5.027e-01  2.044e-01   2.459 0.013943 *
grade9         6.256e-01  2.045e-01   3.059 0.002226 **
grade10        7.085e-01  2.047e-01   3.462 0.000538 ***
grade11        7.868e-01  2.050e-01   3.839 0.000124 ***
grade12        8.442e-01  2.059e-01   4.100 4.14e-05 ***
grade13        1.099e+00  2.195e-01   5.007 5.59e-07 ***
zipcode98002  -2.460e-02  1.843e-02  -1.335 0.181792
zipcode98003   1.265e-02  1.674e-02   0.756 0.449569
zipcode98004   1.135e+00  1.630e-02  69.591  < 2e-16 ***
zipcode98005   7.418e-01  2.003e-02  37.042  < 2e-16 ***
zipcode98006   6.367e-01  1.473e-02  43.212  < 2e-16 ***
zipcode98007   6.546e-01  2.080e-02  31.475  < 2e-16 ***
zipcode98008   6.470e-01  1.674e-02  38.644  < 2e-16 ***
zipcode98010   2.445e-01  2.404e-02  10.172  < 2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.187 on 17164 degrees of freedom
Multiple R-squared:  0.8746,    Adjusted R-squared:  0.8737
F-statistic: 957.4 on 125 and 17164 DF,  p-value: < 2.2e-16
```

In model 3, R-Squared equals to 0.8746, Adjusted R-Squared equals to 0.8737. Residual error equals to 0.187. The residual vs fitted, and normal probability plots satisfy the constant variance and normality of residuals assumption better than the previous models. So, Model 3 is our best model and we will use this model for predicting house prices in King County.



Prediction

```
# Predicting test data on mod3
pred_log_price <- predict(mod3, test) # Because we used log Transformation for price, this will give predicted values of log of price
pred_price <- exp(pred_log_price) # Now exponentiate to get predicted price
actual_predicted_price_df <- data.frame(obs=test$price, pred=pred_price)
actual_predicted_price_df
```

```
> defaultSummary(actual_predicted_price_df)
        RMSE      Rsquared           MAE
1.434596e+05 8.620248e-01 7.906204e+04
```

We use model 3 for prediction. RMSE equals to 143459.6, and R-Squared equals to 0.86. Below is a screenshot of Observed Vs Predicted Values.

```
> head(actual_predicted_price_df)
      obs     pred
10 323000 302877.2
11 662500 787765.9
19 189000 210988.6
20 230000 221348.7
29 438000 510320.4
31 580500 513472.5
```

Conclusion

- In this study, we followed the multiple linear regression technique to explore the relationship between predictors like sqft_living, bathrooms, grade, etc. and response variable i.e. price of the house.
- The independent variable bathroom and sqft_living are significant regressors. So, this confirms our first two hypothesis.
- The bedrooms did not seem to have much influence on the price of the house. So, we reject the third hypothesis.

References

1. Introduction to Linear Regression Analysis – Douglas Montgomery, Elizabeth Peck, Geoffrey Vining
2. https://www.openintro.org/stat/textbook.php?stat_book=reset
3. Kaggle:

   https://www.kaggle.com/shephalika21/models-for-prediction-linear-regression

   https://www.kaggle.com/sushantsawant/linear-regression

   https://www.kaggle.com/prabhats/linear-regression-on-house-price