**Abstract**

House prices, unlike stock prices, appear to be predictable with some degree of accuracy. This paper discusses how Regression Theory can help to predict the sales price of houses. The dataset in this paper includes several house features and the sale prices of houses sold in King County, Washington, the 13th-most populous county in the United States, between May 2014 and May 2015. The proposed model predicts the sales price of the house with the mean square prediction error of 0.062. Also, this paper concludes that freeways are the main reason behind the high price of houses.

## Introduction

The real estate market has experienced a significant change since the severe recession in 2008. The housing prices is growing with the rate of 38% more than inflation. According the Bureau of Labor Statistics, payments for a mortgage or rent account for more than a quarter of the average household income before tax in 2017. Gaining deeper understanding about what factors impacting on the trend of housing prices would help sellers and purchasers have appropriate reactions and make good decisions in housing investment.

This research imports dataset published on Kaggle, an online community of data scientists and machine learners that offers a public data platform and a cloud-based workbench for data science. The dataset for analysis includes several house features and the sale prices of houses sold in King County, Washington, the 13th-most populous county in the United States, between May 2014 and May 2015.

The purpose of our statistical analysis is to predict housing sales prices in King County, WA. We are interested in how home values are affected by the houses' features and launch further investigations to explore possible relationships between them.

## Data and Analytic Approach

*Data Description*

There are 21613 observations, which implicates the number of houses sold in the mentioned period of time, 19 house features, their sales prices and an ID column in the dataset.
House sale price is determined as the response variable with its unit of US dollars. The following is the description of predictor variables implemented in this analysis.

ID: Notation for a house
Date: Date house was sold
Bedrooms: Number of Bedrooms per house
Bathrooms: Number of bathrooms
Sqft_living: Square footage of the house
Sqft_lot: Square footage of the lot
Floors: Total floors (levels) in the house
Waterfront: House which has a view to a waterfront

View: Number of times the house was viewed
Condition: How good the house's condition is, ranking from 1 (poor) to 5 (excellent)
Grade: Overall grade given to the housing unit, based on King County grading system
Sqft_above: Square footage of the house apart from the basement
Sqft_basement: Square footage of the basement
Yr_built: the year in which the house was built
Yr_renovated: the year in which the house was renovated
Lat: Latitude coordinate
Long: Longitude coordinate
Sqft_living15: Square footage of the home in 2015, which might implies some renovations since 2014. This feature might or might not have affected the lot size area
Sqft_lot15: Square footage of the lot in 2015, which might implies some renovations since 2014

Waterfront is selected as the unique categorical variable while the remaining predictors are considered quantitative variables. Waterfront takes on the value 0 if the house has no waterfront and 1 otherwise.

This report excludes the involvement of ID and Date due to their irrelevance to house price. Zip code is also omitted since it could be substituted by longitude and latitude coordinate. We also exclude square footage of the house apart from the basement and square footage of the basement from the dataset because the correlation between the sum of these two variables and square footage of living space equals to 1, suggesting an apparent linear correlation exists among them. The total of predictors now shrinks to 15 variables. The following table shows a summary of our dataset in term of median and range of each variable included in this analysis.

| | Price($) | Bedrooms | Bathrooms | Living Area (sq ft) | Lot size (sq ft) | Floors | Waterfront | View |
|---|---|---|---|---|---|---|---|---|
| **Median** | 450,000 | 3 | 2.25 | 1910 | 7618 | 1.5 | - | 0 |
| **Range** | 75,000-7,700,000 | 0-33 | 0.00-8.00 | 290<br>-13,540 | 520<br>-1,651,359 | 1.0-3.5 | 0<-21450<br>1<-163 | 0-4 |
| | Condition | Grade | Year built | Year renovated | Latitude (degree) | Longitude (degree) | Living Area 2015 (sq ft) | Lot Size 2015 (sq ft) |
| **Median** | 3 | 7 | 1975 | 0 | 47.56 | -122.2 | 1987 | 7620 |
| **Range** | 1-5 | 4-13 | 1900-2015 | 0-2015 | 47.16<br>-47.78 | (-122.5)<br>- (-121.3) | 399<br>-6,210 | 651<br>-871,200 |

Table 1: Summary statistics of housing features

*Sample Determination*

We split the data set in two parts: train and test, using sample function in R. The ratio of the sample size between the train and test sets is chosen to be 70% and 30% respectively. Since our dataset includes 21613 observations, this ratio is assumed to provide sufficient data points for each predictor per set.

*Data Diagnostics*

We initially fit a linear regression model on the response Price that is explained by 15 chosen predictors. A residual analysis is conducted first to check the model assumptions. The scatterplot of Price as a function of Bedrooms, Bathrooms, Living area, Lot size, Floors, Waterfront, View, Condition, Grade, Year built, Year renovated, Latitude, Longitude, Living area 2015, Lot size 2015 shows an obvious triangular pattern in residual variance. Furthermore, there is also evidence of data deviating from the normality assumption in the quantile plot with clear thick tails. Both unsatisfactory residual assumptions suggest the necessity of transformations on either the predictors or response.

*Transformations*

In order to identify an appropriate transformation on the response variable, the Box-Cox method is applied on Price. The result with $\lambda = 0$ indicates that the variance can be stabilized by a logarithmic transformation on Price.

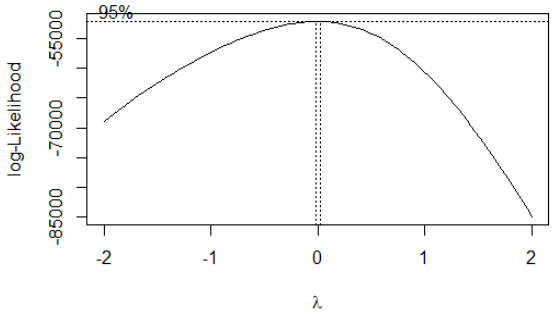


Figure 1: Box-Cox Transformation on Price variable

We re-check the model assumptions by the residual plots of multiple linear regression with log(Price) as the new response. Both scatter plot and quantile plot depict a dramatic change in the data patterns. The residuals are distributed "like confetti in a box" and data points approximately line up along the line in Q-Q plot. The model assumptions are now satisfied.

The transformed response log(Price) is used to check linear relation with the remaining predictors. Box plots of transformed Price against each predictor illustrate approximate linear relationships between log(Price) and Bedrooms, Bathrooms, Floors, View, Condition, Grade, Year built and Longitude individually, with no transformation required on these predictors.
When log(Price) is plotted as a function of Living area, we observe a slight bend log in their scatterplot, whose shape is recognizable in a square root function. A square root transformation on square footage of the house is therefore reasonable and makes the curvature disappeared after we plot log(Price) against $\sqrt{Living\ area}$ . The new plot suggests a clear positive linear relationship between the predictor and response. A similar problem is also found in the graphical analysis of Living area in 2015 and resolved in a same manner.

The relationship between log(Price) and Lot size is also depicted through a curve pattern in scatter plot. Since there seems to exist an asymptote at some certain housing values, we believe an exponential transformation on Lot size will be an appropriate treatment for its non-linear relationship with log(Price). We apply mathematical techniques to find new Lot size, which is defined by $10^6 \times \left(6.5 \times 0.2^{\frac{Lot\ size}{10^4}} + 0.5\right)$. The curvature vanishes after we re-plot log(Price) against transformed Lot size. Log(Price) appears to be linearly dependent on Lot size. We also apply an exponential transformation on Lot size 2015 as it behaves similarly to Lot size. The transformed Lot size 2015, $10^6 \times \left(6.5 \times 0.2^{\frac{Lot\ size\ 2015}{10^4}} + 0.5\right)$ also resolve the problem of the curve pattern in scatter plot.

Performing analysis on the effect of Year renovated alone on home values yields no useful insight as the scatter plot of log(Price) against Year renovated show two separate clusters of housing prices when the house had no renovation before and when it was reconditioned in about 2000. We later notice that Year renovated taking on the value 0 carries similar characteristics to Year built since both variables imply the house has not renovated. Therefore, for any house has value 0 in Year renovated, we replace its Year renovated value by its Year built value to generate more meaningful results in analyzing the effects of home renovation on the sales price.

A curve pattern which bends down near the center of the scatter plot of log(Price) against Latitude urges us to apply a polynomial transformation on Latitude, and this treatment makes the pattern significantly less obvious.

The following is a summary of transformations applied on all variables up to this point.

No transformations:

Bedrooms Bathrooms Floors View Condition Grade Year built Longitude

Transformations:

Price $\mapsto$ log(Price)

Living area $\mapsto \sqrt{Living\ area}$

Living area 2015 $\mapsto \sqrt{Living\ area\ 2015}$

Lot size $\mapsto 10^6 \times \left(6.5 \times 0.2^{\frac{Lot\ size}{10^4}} + 0.5\right)$

Lot size 2015 $\mapsto 10^6 \times \left(6.5 \times 0.2^{\frac{Lot\ size\ 2015}{10^4}} + 0.5\right)$

Year renovated: Year renovated = 0 $\mapsto$ Year built

Year renovated $\neq 0 \mapsto$ No transformation

Latitude $\mapsto$ (Latitude - mean(Latitude))$^2$

We also include several box-plots and scatter plots that illustrate significant changes relationships between the response and predictors individually after transformations (figure 2).

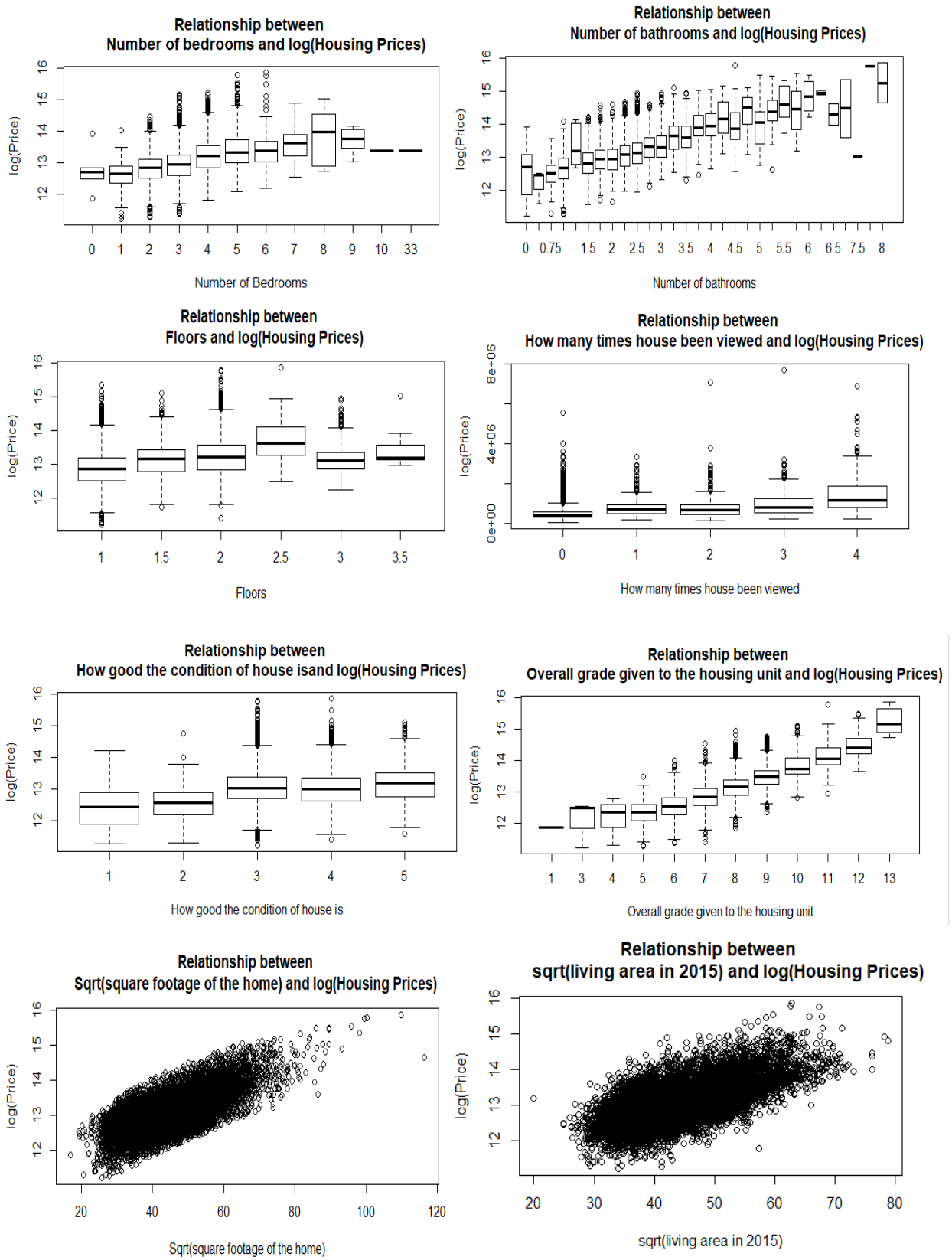


Figure 2: Relationship between each transformed predictor and log(Price)

## Transformed Model Analysis

*Variables Selection and Model building*

We further investigate the important variables which are helpful to predict the price. We apply the exhaustive selection method for variable selection with n_best equal to 3. In total, we get 24 number of models with the different number of predictors in each subset. We choose adjusted $R^2$ instead of $R^2$ to compare different models due to the variation in the number of predictors between models selected. Our criteria is to select the model with low $C_p$, low MSE, and high adjusted $R^2$ . We finally came up with the top three contenders and call it as Model 1, Model 2 and Model 3. Values of MSE, Adjusted R2 and $C_p$ for all three models are summarized below.

| | Top 3 models on the basis of exhaustive method | SSRes | R2 | AdjR2 | MSE | Cp |
|---|---|---|---|---|---|---|
| Model 1 | **log(Price)** ~ $\sqrt{Living\ area}$ + View + Condition + Grade + Year built + (Latitude - mean(Latitude))$^2$ + $\sqrt{Living\ area\ 2015}$ | 955.695 | 0.773 | 0.773 | 0.063 | 894.780 |
| Model 2 | **log(Price)** ~ $\sqrt{Living\ area}$ + View + Condition + Grade + Year built + (Latitude - mean(Latitude))$^2$ | 971.191 | 0.769 | 0.769 | 0.064 | 1152.309 |
| Model 3 | **log(Price)** ~ $\sqrt{Living\ area}$ + View + Grade + Year built + (Latitude - mean(Latitude))$^2$ | 990.385 | 0.765 | 0.765 | 0.066 | 1471.765 |

Table 1: Comparison of top three models from Exhaustive search method

*Model Adequacy Check*

We plotted Q-Q norm and residual plots for the top three models to find the better model. Plots in figure 3 and 4 do not show any deviation from our model assumptions. Errors are normally distributed with constant variance in our top three models.

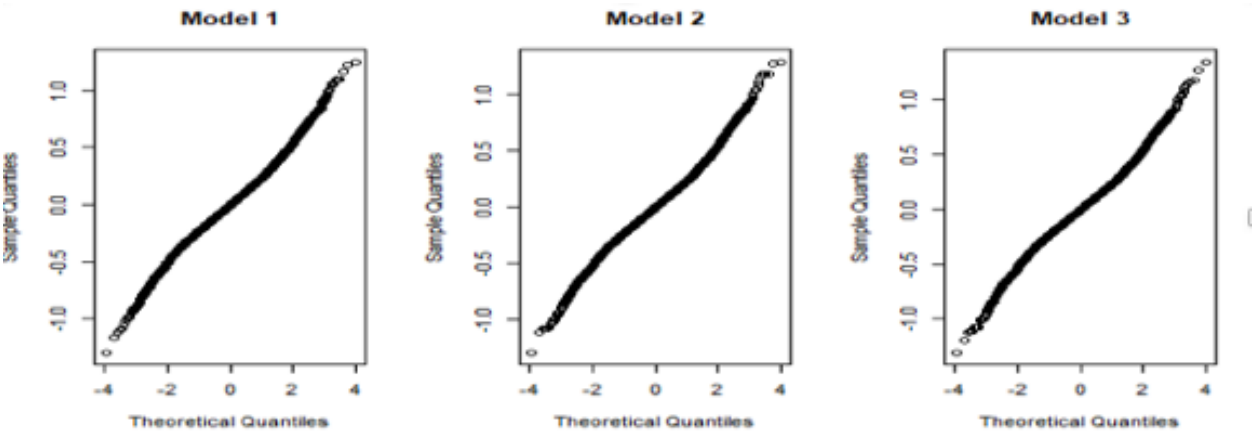


Figure 3: Q-Q plots of top three models

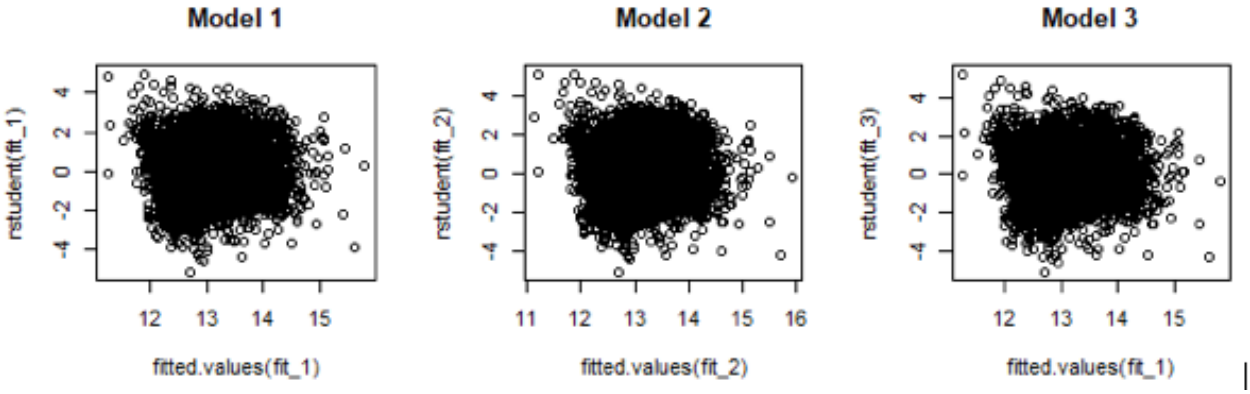


Figure 4: residual plots of top three models

*Multicollinearity analysis*

We look for presence of multicollinearity in our data. We calculated coefficient of correlations and VIF values for each predictor. As being showed in the R-output below, the largest value is 3.12, and all are less than 10. This means that there is no severe multicollinearity problem between predictors. Therefore, there is no correlation between any of the predictors in our top three models.

```
      view      condition          grade       yr_built  sqft_living_new
  1.144334       1.167103       3.008840       1.569739       3.125492
  lat_new.1      lat_new.2 sqft_living15_new
  1.086186       1.022219       2.613465
```

**Final Model Selection and Further Analysis**

*Final Regression Model*

Among top three models, we chose model 1 based on criteria of lowest $C_p$ (894.7791), lowest $MS_{res}$ (0.0632) and highest adjusted $R^2$ (0.7727).

Our final regression model is:

$$\textbf{log(Price)} \sim \sqrt{\textit{Living area}} + \textbf{View} + \textbf{Condition} + \textbf{Grade} + \textbf{Year built} + \textbf{(Latitude-mean(Latitude))}^2 + \sqrt{\textit{Living area 2015}}$$

The following is statistical summary of the model we finalized.

```
Call:
lm(formula = log(price) ~ sqft_living_new + view + condition +
    grade + yr_built + lat_new + sqft_living15_new, data = train_hs)

Residuals:
     Min       1Q   Median       3Q      Max
-1.29419 -0.15945 -0.00416  0.15042  1.24805

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.514e+01  1.685e-01   89.85   <2e-16 ***
sqft_living_new   1.809e-02  3.797e-04   47.65   <2e-16 ***
view              8.034e-02  2.836e-03   28.33   <2e-16 ***
condition         6.016e-02  3.397e-03   17.71   <2e-16 ***
grade             1.691e-01  3.010e-03   56.17   <2e-16 ***
yr_built         -2.398e-03  8.699e-05  -27.57   <2e-16 ***
lat_new1          2.353e+01  2.620e-01   89.79   <2e-16 ***
lat_new2         -8.211e+00  2.542e-01  -32.30   <2e-16 ***
sqft_living15_new 7.068e-03  4.514e-04   15.66   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2514 on 15120 degrees of freedom
Multiple R-squared:  0.7728,    Adjusted R-squared:  0.7727
F-statistic:  6430 on 8 and 15120 DF,  p-value: < 2.2e-16
```

*R output of final model*

From the R output, we notice that the estimations of the model parameters are significant as they all have small p-values. Furthermore, around 77.27% of variability in the transformed housing prices can be explained through our final model based on the adjusted R-squared.

*Detecting Influential Observations*

We calculated studentized residual for every observation in our final model set. We have few observation has standardized residual larger than |3|, however deleting those points did not change the model parameter much. For our model, we have k=8 (number of predictors) and n=15129 (training data set). Thus our cut-off for leverage points is $2 \cdot \frac{(8+1)}{15129} = 0.001189768$. We computed the Hii matrix in R and got hii value for each observation. We got few leverage points in our data. To check for influential points, we calculated Cook's distance and it was less than 1 for all the data points. Therefore, we do not have any influential point in our dataset to effect our parameter estimates.

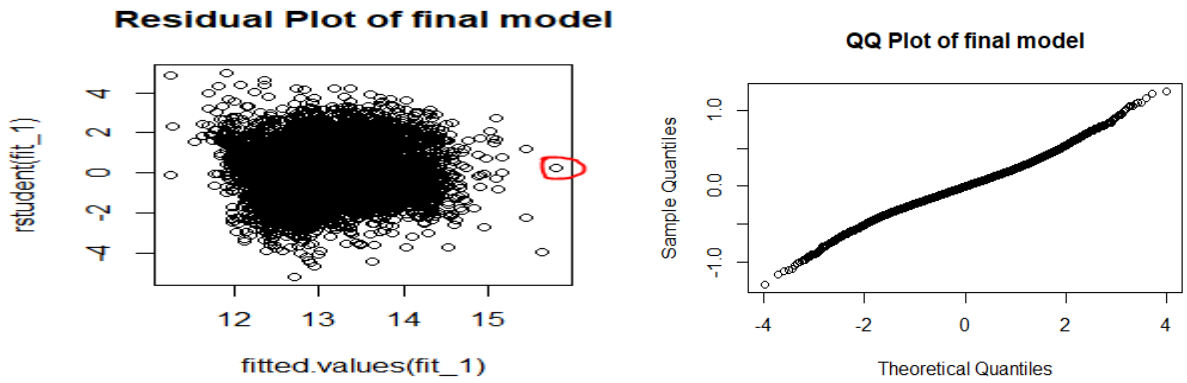


Figure 5: Residual plot and Q-Q plot of the final model

*MSP on validation dataset*

We apply our final regression model on test data to check its performance. Mean squared Prediction error comes out to be 0.06122029, which is close to the MSRes (0.06) of final model. Therefore, our model is fairly successful at making predictions.

*Bootstrap*

We obtained distinct 1000 datasets by repeatedly sampling observations of size 1000 from the test data set. We can see that each estimate is normally distributed. Therefore, all coefficient of predictors are fairly stable.

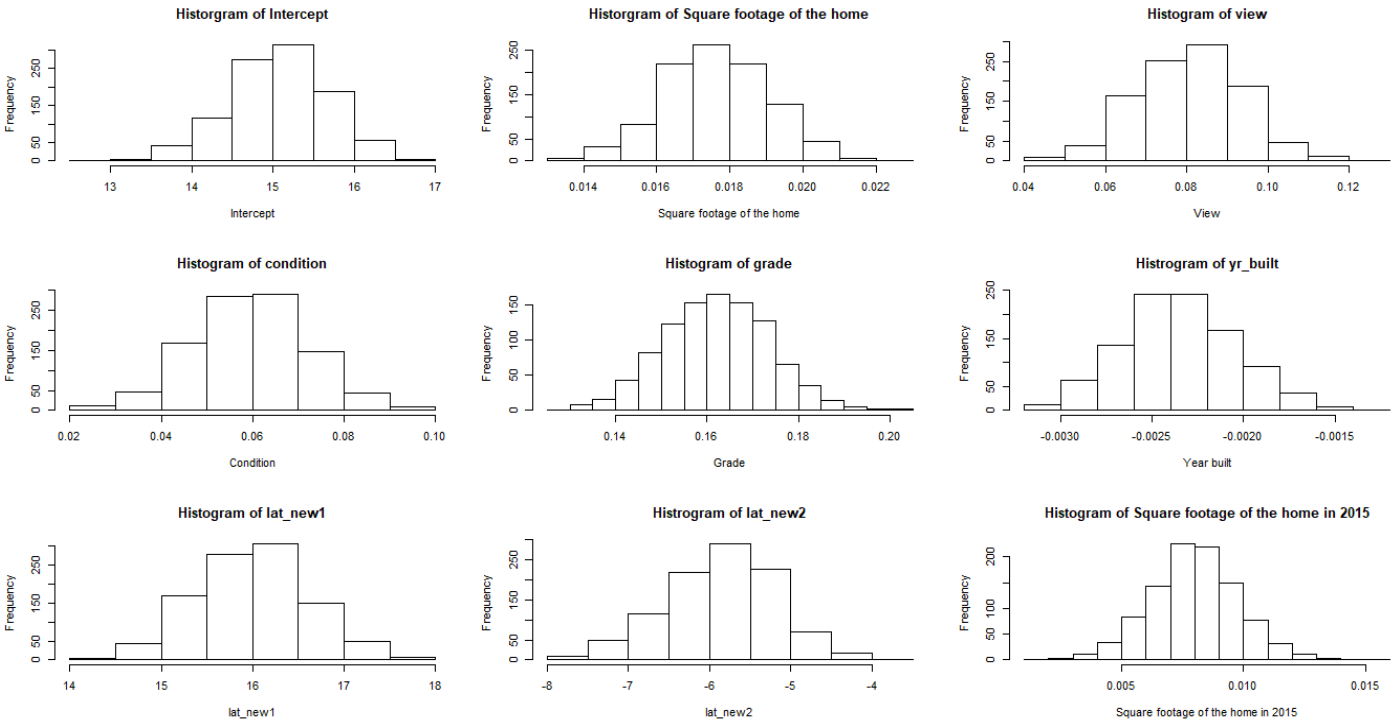


Figure 6: Histogram showing bootstrapping

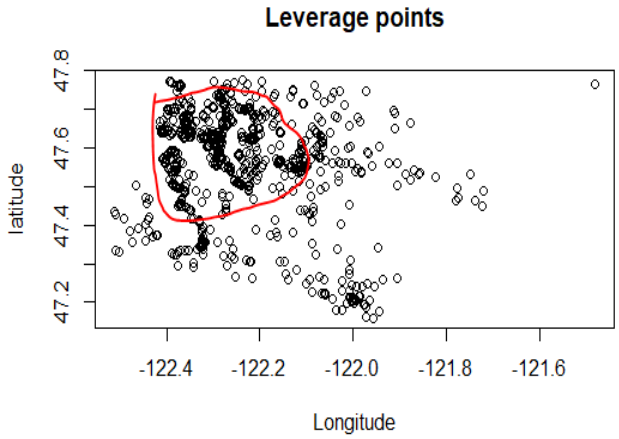


Figure 7: Leverage points

**Analysis of Leverage Points**

In figure 5 and figure 7 we can see leverage points. After analysis we found that most of the leverage points are falling in the latitude between $47.4°$ and $47.7°$, and longitude $-122.4°$ & $-122.1°$. This region has very high or unusual square footage value of living area and the lot.
Consider, red circled data point in figure 4, having latitude $47.6298°$ and longitude $-122.32°$, has highest leverage value. As we can see, it has very high/unusual sq ft value of living area (12,050 sq ft) and lot (27600 sq ft).
Also, predicted value for this data-point is $7,261,304 and actual value is $7,700,000 which means that our final model is able to predict the price very well.

**Conclusion**

Our final model indicates that the variation in King County's home values are indeed significantly affected by several housing characteristics.

Among 19 predictors at the beginning, our model finalized only 7 predictors that influence in predicting the house sales prices. Specifically, they are living area, number of times the house was viewed, condition, grade, year built, latitude coordinate and living area in 2015. Furthermore, how the house is graded seems to be most influential to the home values.

Unlike our prediction that waterfront would significantly affect the housing price, our final model does not include waterfront. The reason might be due to the limited number of houses that have waterfront obtained in our data, which only accounts for .07%

While longitude of the house seems not to make much impact on home values, houses in King County appears to be sold with higher prices when the house's latitude degree increase, especially whose latitude around 47.4 degrees to 47.7 degrees. Further investigations reveal several reasons behind the elevated levels of house sales prices. We create a map through tableau to analyze the variation of home values based on their locations. The home values rise according to the change of light green to dark blue shade for each area in the map. When we take closer look at the geographic map of King County and compare it with our map, we find out there are three interstate highways passing through the zone of latitude around 47.4 degrees to 47.7

degrees (figure 8). Additionally, the area marked in dark blue shade, the area of highest home values, is in the neighborhood of downtown Seattle, the largest city in King County, WA.

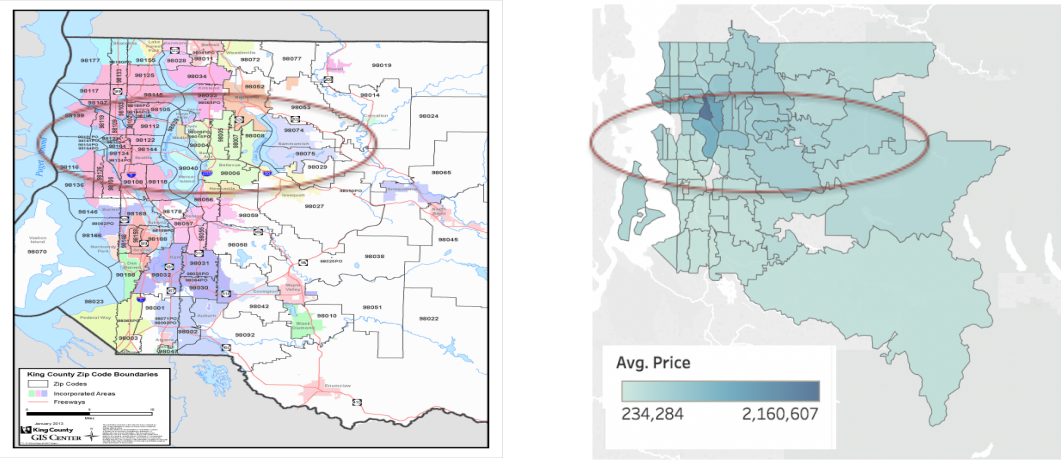


Figure 8: Geographic graph of Kings County

**Suggestion for future study**

-The “Date” of the data we are doing research on are in a short timespan of one year (May 2014-May 2015), therefore, we removed variable “Date”. In future study, if we have a dataset with longer timespan, we might need to take that into account.
-Environmental factors are not provided by dataset. Information like School district, Population density and Income per capita for the area also have influences on house price

## References

"Boundaries." *King County, 24 Apr. 2017,* *www.kingcounty.gov/services/gis/Maps/vmc/Boundaries.aspx*. Accessed 20 Nov. 2018.

Bremer, Martina. (2018). *Regression Theory and Methods*.

"Consumer expenditures 2017." *U.S. Bureau of Labor Statistics,* 13 Sep. 2018, www.bls.gov/news.release/cesan.nr0.htm. Accessed 20 Nov. 2018.

DePillis, Lydia. "How Washington could actually make housing more affordable." *Cable News Network,* 24 *Jul. 2018*, www.cnn.com/2018/07/24/politics/affordable-housing/index.html. Accessed 20 Nov. 2018.

"House Sales in King County, USA." *Kaggle, 2016,* www.kaggle.com/harlfoxem/housesalesprediction. Accessed 20 Nov. 2018.

"Kaggle." *Wikipedia,* 13 November 2018, en.wikipedia.org/wiki/Kaggle. Accessed 20 Nov. 2018.

"King County, Washington." Wikipedia, 20 November 2018, en.wikipedia.org/wiki/King_County,_Washington. Accessed 20 Nov. 2018.

Montgometry, Douglas, et al. *Introduction to Linear Regression Analysis*. 5th ed., John Wiley and Sons, Inc., 2012.