

# King County Housing Market Analysis

Alex Gromadzki(arg2eu), Karan Kant(kk4ze), Rakesh Ravi(rk9cx),  
Aman Shrivastava(as3ek) and Jingnan Yang(jy4fch)

December 6, 2018

## **Abstract**

The team built a model based on a housing dataset consisting of over 21,000 house sales from September 2014 to January 2015 in order to predict housing prices. After some data cleaning and transformation, we fit a linear model using ordinary least squares estimation. The features most indicative of price variability were the year built and lot size, while house ID and living size were relatively insignificant. Ultimately, the results from this project follow what we would have expected, given previous research in the housing domain.

# 1 Introduction

Ever since the 2008 financial crisis, analyzing housing data has been a particularly fruitful field of study. Exploring real estate pricing as a whole is interesting as the subject matter generates many exploratory models that often exhibit similar characteristics. Over the course of this project, we expect to evaluate and confirm a relationship between common property indicators and housing prices and to develop a suitable model for the given data utilizing statistical techniques applied throughout Linear Models for Data Science. The data was collected and released under public domain anonymously, which we then downloaded from Kaggle. Despite the anonymity involved with the release, it seems likely that the data originated from some sort of official documentation, given the grade variable, which is specifically based on a King County Scale, according to the column description. The data points range from 2014 to 2015 as part of what appears to have been an observational study. The CSV file consists of various indicators that are, heuristically, considered strong indicators of property prices. We wish to use the data to corroborate our hypothesis that a relationship exists between common housing properties including location, property size and quality, and prices.

## 2 Dataset

### 2.1 Dataset Description and Provenance

The dataset contains over 21000 rows of data on sale prices of houses in King County, United States. For the purpose of this project, the response variable will be the sale price of the houses. In the dataset, each row indicates a house and is characterized by information on real estate specific metrics such as number of bedrooms, number of bathrooms, square footage of the house etc., locational features such as latitude, longitude and zip codes and temporal features such as date, month and the year of sale. There are 21595 houses located in over 70 different zip codes within the state of Washington which were sold during the period between January 2014 and December 2015. A majority of the variables in the dataset are continuous like square footage of the lot, basement and the house and the rest are the dataset are either nominal or ordinal in nature.

### 2.2 Exploratory Data Analysis and Data Cleaning

As the dataset contained more than 10 regressors, it was cumbersome to examine binary relationships between them. Correlation analysis is a method of statistically evaluating the strength of relationship between different variables in a dataset. This analysis yields a measure called the correlation coefficient which ranges from -1 to 1 where 1 represents strong positive correlation and -1 represents strong negative correlation. Generally, it is expected that the regressors are not correlated with each other but are correlated with the response variable. The sale price of the house is highly correlated to the area/ size of the house (Figure 1), grade conferred by King County authorities, and the number of bathrooms in the house. In addition, there were linear relationships between area/size of the house, the number of bathrooms, and the grade of the house. Variables with an exceptionally low coefficient of

correlation with the response variable (nearly zero) were removed from analysis.

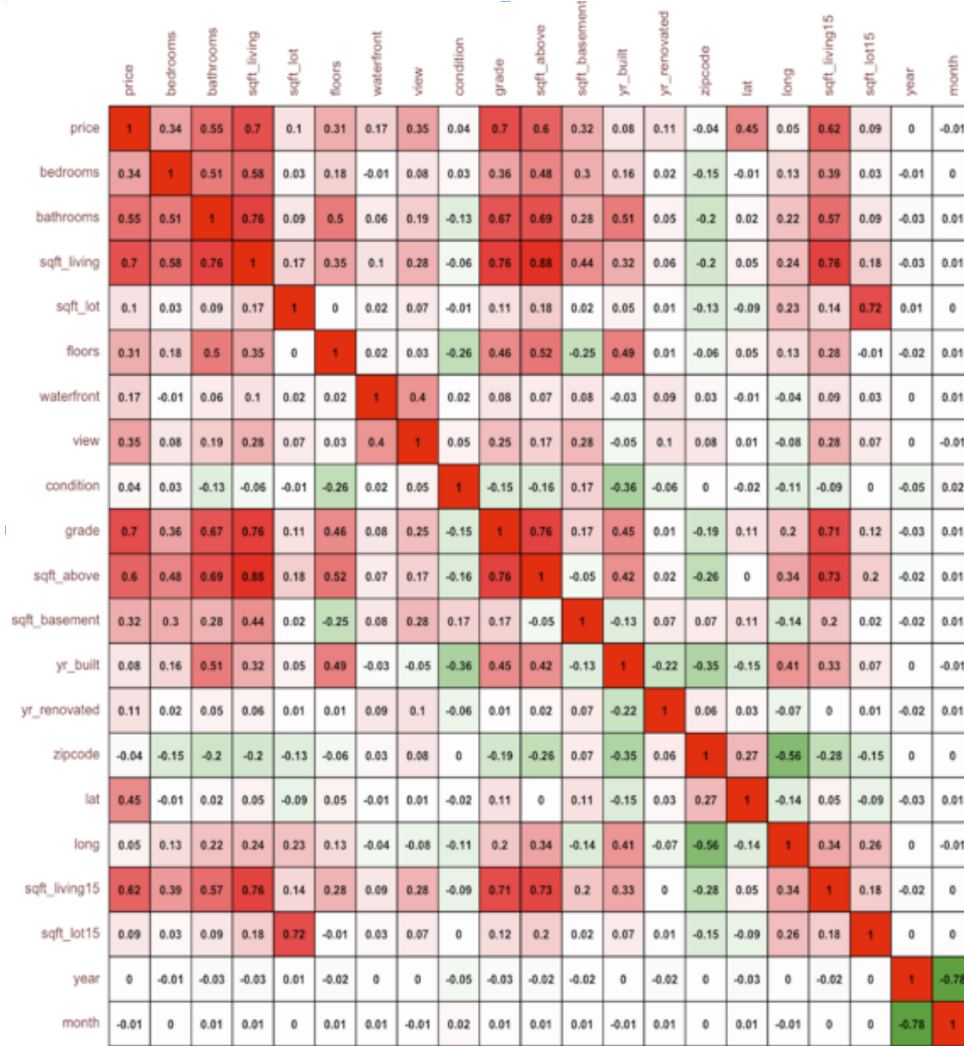


Figure 1: Correlation plot for all variables

Upon observation of the distribution of the response variable, it is clear that the raw house prices exhibit a long tailed distribution (left - **Figure 2**), indicating that a transformation might be necessary before attempting to model the data. Using the Box-Cox procedure (center - **Figure 2**), we decided that the natural log was the most appropriate transformation for the response variable. Applying log transformations is common practice when working with economic data, so by approximating the suggested result from the Box-Cox Procedure, we obtain transformation that agrees with our heuristic. After applying the log transformation, the distribution of the response variable is much smoother and follows a bell shaped curve, as shown on the right hand side of **Figure 2**. The transformed response variable allows us to use Ordinary Least Squares without violating the basic assumptions of Linear Regression.

The real estate prices, generally, depend on the location and the size of the property and

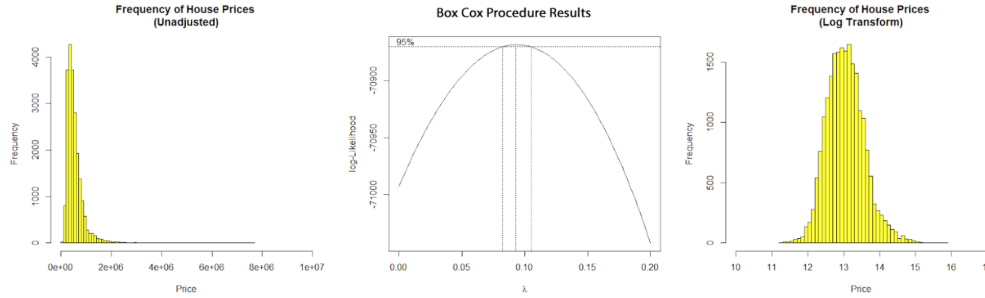


Figure 2: Left to right: Skewed distribution of response variable, Results from Box-Cox Procedure, Distribution of log transformed response variable

this is consistent in the dataset where the sale price has a strong correlation with the area (in square feet) of the house both in the year they were built and when they were measured in 2015.

We observed that the size of some of the houses increased from the time they were built as shown in **Figure 3**. The total area of the house in square feet is a sum of two different variables `sqft_above` (covers the area above the ground) and `sqft_basement` (covers the area beneath the ground). From the scatter plots (Bottom left and right **Figure 3**), it is evident that `sqft_basement` and `sqft_above` follows a trend similar to `sqft_living` and it would be redundant to include both within a single model. Accordingly, we decided to remove `sqft_above` from the final variable list. Upon deep diving into the data on the basement of houses, we found that over 13,000 houses did not have a basement; this proportion represents 50% of the total houses in the dataset. As a result, we converted the area of basement variable to a dummy variable which is equal to one when a property has a basement and zero when it does not. Since nearly 20,000 houses (90% overall) have not been renovated from the time they have been built, we also decided to convert the `yr_renovated` variable into a dummy variable which takes the value of one when a property had renovations and zero when it did not.

There are 15 categorical variables in the dataset where most of them are ordinal and some of them are nominal in nature. Generally speaking, categorical variables that contain a large number of levels increases the dimensionality of the dataset. Understanding the relationship between the response variables and different levels of individual categorical variable will provide more insight into classifying them into ordinal and nominal variables. A box plot provides the best representation of distribution of the response variables within different levels of these categorical variables.

Two houses were listed as having over ten bedrooms; however, closer examination of these two observations revealed that these houses only had two bathrooms and the size of the house did not exceed that of a traditional three or four bedroom house. We believe the abnormality within these rows occur as cases of transcription or data entry error and these two observations were removed from the dataset. The left side of **Figure 4** indicates that there is a

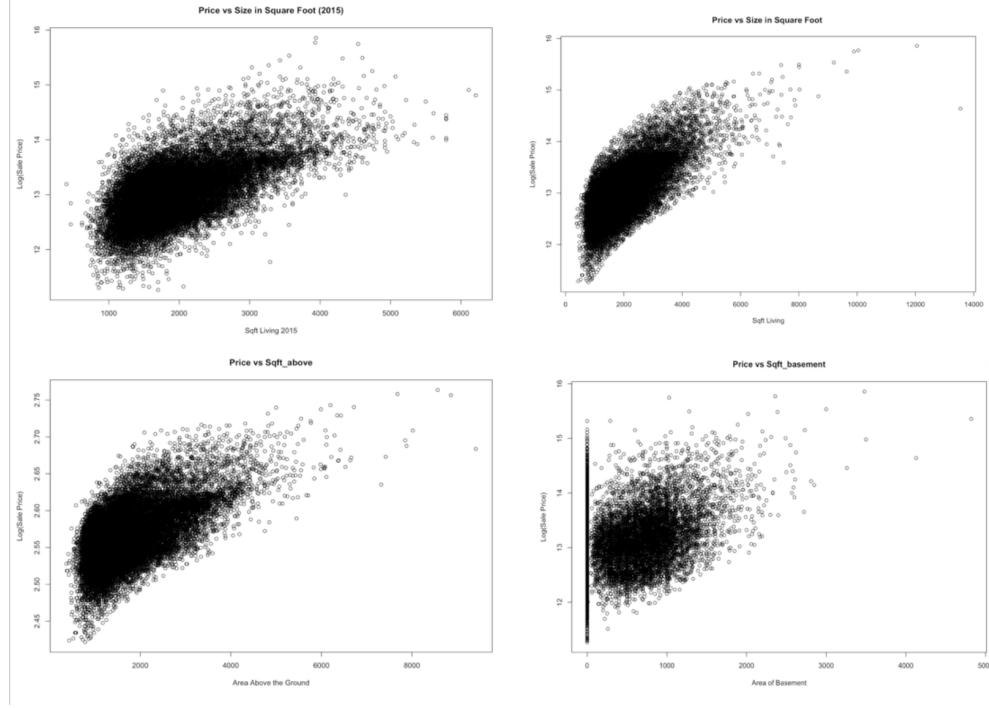


Figure 3: Top Left: Area in square feet versus price, Top Right: Area in square feet as measured in 2015 versus price, Bottom Left: Area in square feet above the ground versus price, Bottom Right: Area of basement in square feet versus price.

rough linear relationship between the number of bedrooms and median sale prices of houses, after removing the houses with 11 and 33 bedroom houses. For the purpose of the project, the number of bedrooms were considered as an ordinal categorical variable. Given that the number of bathrooms and Grade of the house increased with increase in price (Right side of Figure 4), we converted them into ordinal categorical variables. We found floor, view, waterfront and condition of the house to be nominal categorical variables, because they did not exhibit any trends with the sale price.

## 3 Methodology

### 3.1 Variable selection

Various methods were employed for variable selection. The geographical position (latitude and longitude) and ID were heuristically removed, as we determined that the variables may confuse the model rather than improve accuracy. The `sqft_living` variable was also removed, because it represented the sum of two other variables in the dataset, and was thus perfectly correlated. The date variable was also split into month and year predictors.

Multiple tests were performed to confirm variable selection for continuous data. Each test utilizes a different method of evaluating variable importance by entering and exiting the variable and recording changes to model error parameters. Utilizing backward selection,

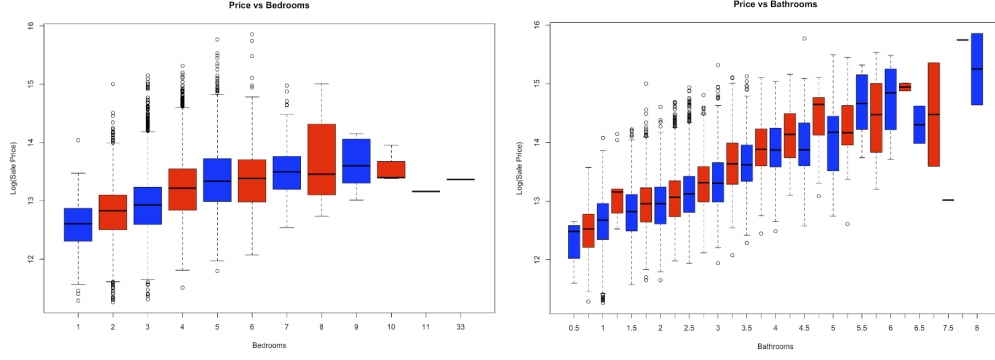


Figure 4: House price trends versus number of bedrooms (left), House price trends versus number of bathrooms (right)

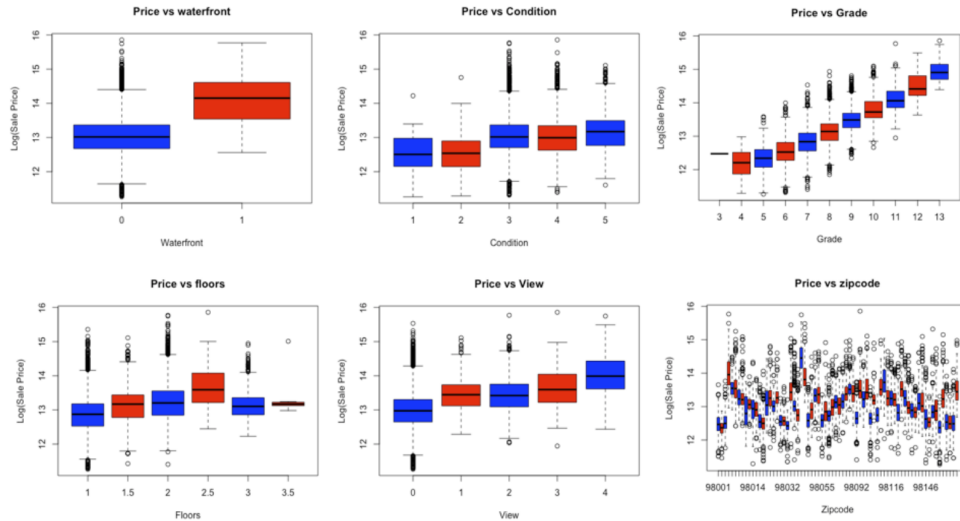


Figure 5: Box Plots for Waterfront, Condition, Grade, Floors, View and Zip Code

forward selection and all possible regressors for four continuous variables confirmed that all continuous variables were of benefit to the model. We also converted all ordinal variable into factor level variables, allowing for `RStudio` to read the predictors as levels rather than as continuous variables. Examples of ordinal variables in the data include bedroom, grade and floors. Our dataset after completing this step consisted of 17 predictor variables and one response variable (price). All variables and their corresponding type are displayed in Table 1 of the Appendix.

### 3.2 Multicollinearity

Detecting multicollinearity is essential for reducing variance and redundant regressors in the model. Allowing correlated variables in the model may result in model skewness and serious variance inflation. In addition, multicollinearity may jeopardize future prediction accuracy if left unresolved. The first step for identifying and mitigating multicollinearity

requires variable scaling. All continuous variables were unit-length and centered, after which multiple detection processes were performed. We first calculated Variance Inflation Factors and ensured our values were within reasonable limits. All variables showcased values less than 10, indicating that serious multicollinearity did not exist in the model. Calculating variance decomposition also resulted in values below the threshold, as did conditional indices for all but `sqft_lot`. All values can be found within **Tables 2, 3 and 4** of the Appendix. We left the dataset as is, because serious multicollinearity did not seem to exist within the tested variables.

### 3.3 Hypothesis testing and residual analysis

We performed tests of significance of regression and partial regression on our models to ensure regressor importance and overall model adequacy. An F test was used to measure the significance of the model and partial F tests were used for evaluating individual regressor significance. We further performed residual analysis on our model and fitted data at various intervals of the modeling process. A model using the raw, untouched data was tested for residual patterns and inadequacies through normal probability and fitted residual plots. The same test was performed on a separate model created on transformed and cleaned data and compared with our original results, thus allowing us to confirm the effectiveness of our data manipulation techniques.

### 3.4 Influence diagnostics and treatment

It is important to identify influential observations in the dataset as leverage points that are possibly influential may result in a skewed model. As a result, we ensured that the model was as general as possible by removing influential points in the dataset using Cooks distance. The method evaluates every observation, and an observation is considered influential if the model significantly changes due to removal of the observation. We specified a threshold and removed any observation that landed beyond this mark. 752 observations were considered influential and removed from the dataset.. We created a final model after all influential points were removed from the dataset using this procedure, given that the data was considered clean after applying all aforementioned strategies in the methodology section.

## 4 Results

We created three models during analysis, all at different stages of the process and used as benchmarks and as corroboration of our data manipulation methods. The adjusted R Squared values of the three models are displayed in **Table 5** of the Appendix.

The first model was created as a testing unit for our ground truth dataset. We employed a simple linear regression model using all 20 variables, without any response transformations or removals. The adjusted R Squared value equaled 0.7023, translating to 70.23% of variability in housing prices being explained by the model. A significance of regression test for the model was also performed with a corresponding F value of 5222.791. Thus, we inferred that

the base model does an average job of representing the data. Despite a higher than expected R squared value, we were confident that removing variables and performing transformations would significantly improve the model.

Performing residual analysis on the model is also a critical step and helps display skewness in the data. We calculated studentized and PRESS residuals for the model and fit the resulting data into a normal probability and fitted residual plots. **Figure 5** displays the two residual plots, and it is immediately evident that the residuals exhibit patterns, outliers and are heavily skewed. The normal probability plots show a very heavy tailed distribution, while the fitted residual plots show a funnel pattern and outliers in the data. We can thus conclude that the model is not adequate enough to represent our dataset.

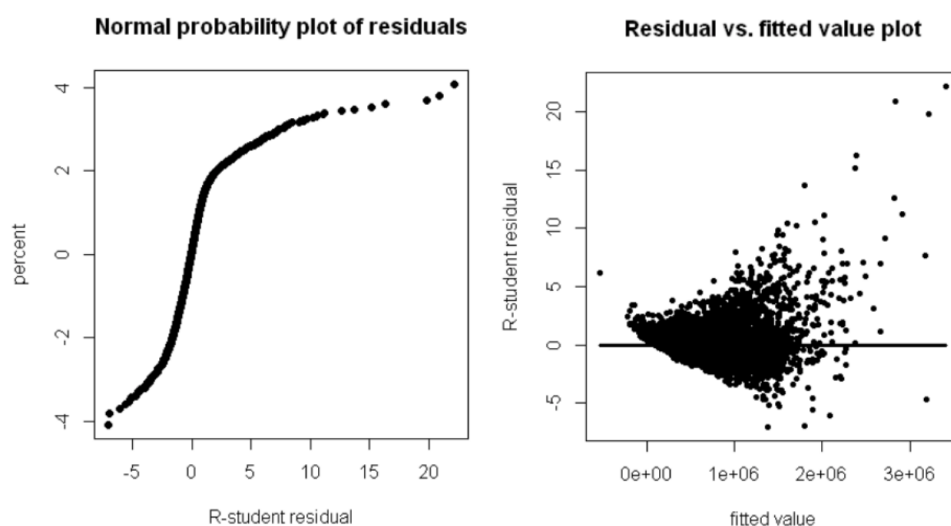


Figure 6: Normal Probability Plot (Left) and Fitted Residual Plot (Right)

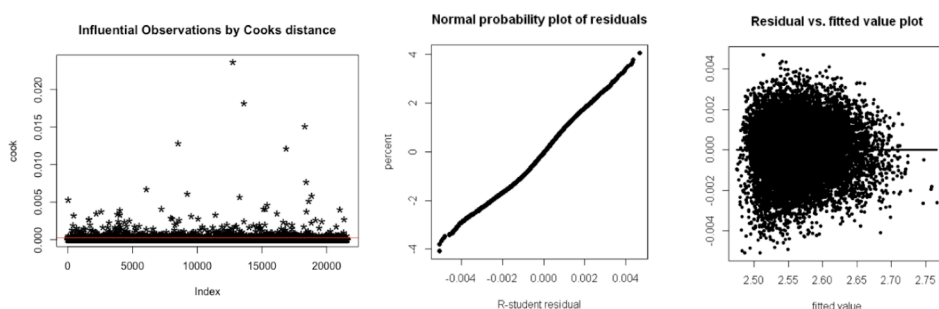


Figure 7: Normal Probability Plot (Left) and Fitted Residual Plot (Right)

The second step in our process involved removing variables and log transforming the response with techniques outlined in previous sections of this report. A second model was



created after working with the data, and it was immediately evident to us that our data transformation techniques resulted in a significant boost to our model. The model exhibited an Adjusted R Squared value of 0.8862 and F value of 1475.468, confirming our improvements to the dataset and significance of regression. Partial F tests performed on the model regressors further confirmed significance for all variables in our model.

Detecting and removing influential observations was the final step in our process. We created our final model after using the Cooks distance method to handle influential observations. Our final model showcased an Adjusted R Squared value of 0.9093 and F value of 2079.985. We also performed residual analysis by outputting probability and fitted residual plots (Figure 6), both of which are much improved from the plots created in the first model. The normal probability plot shows a reasonably normal distribution, and the fitted residual plot displays a horizontal band that houses most of the observations. Both plots indicate that the assumption of normality in errors is justified.

## 5 Conclusion

Ultimately, we were successful in our goal of creating a linear model to capture the relationship between housing prices and common predictors. The study began by exploring our dataset in order to determine how to properly clean and transform our variables so that we were most accurately modelling what we were intending. We arrived at a log-transformation of the response through the Box-Cox procedure and modified several regressors by either making them into categorical variables or extracting features from them. After testing for multicollinearity, we ran an updated model, which led us to believe there were still influential points. We removed these by relying on the Cooks Distance and ran one final model, for which we explained about 91% of the variability in housing prices with 17 regressors.

The results from this study are not foundational, rather supplement an already extensive field of research into the housing market. If we were to continue working in this field, we might consider other choices of models beyond simple linear regression, such as Random Forests or Extreme Gradient Boosting. Additionally, our dataset was relatively small and bounded to a single geographic region. It would be useful in future work to compare neighboring counties if the data were available. We are also aware that we may have been lacking some fundamental domain knowledge, as none of the team members are from King County, thus may have run the risk of being unfamiliar with underlying trends that a domain expert or on this case, a resident would have known. Accordingly, we might better model an area that we were familiar with as we could design more intricate features. Regardless, we were able to successfully use common property characteristics to predict housing prices. Assuming similar data is available in other areas, we imagine that our model could produce satisfactory, albeit potentially extrapolated, results when applied to other similar housing markets.

# Appendices

Table 1: Variable Descriptions

Variable	Type	
price	Response (continuous)	In the final model? (Y/N)
date	Ordinal	Y
id	Ordinal	N
bedrooms	Ordinal	N
bathrooms	Ordinal	Y
sqft_living	Continuous	Y
sqft_lot	Continuous	N
floors	Nominal	Y
waterfront	Nominal	Y
view	Nominal	Y
condition	Nominal	Y
grade	Ordinal	Y
sqft_above	Continuous	Y
sqft_basement	Continuous	Y
yr_built	Ordinal	Y
yr_renovated	Nominal	Y
zipcode	Ordinal	N
lat	Ordinal	N
long	Ordinal	N
sqft_living15	Continuous	Y
sqft_lot15	Continuous	Y
year	Ordinal	Y
month	Ordinal	Y

Table 2: Variance Decomposition

Variance Decomposition	sqft_above	sqft_basement	sqt_living15	sqft_lot15	sqft_lot
sqft_living	0.0072	0.06	0.072	0.025	0.437
sqft_basement	0.0016	0.2073	0.00343	0.6932	0.53
sqft_living15	0.0119	0.5095	0.1206	0.2797	0.034
sqft_lot15	0.9146	0.223	0.804	0.002	0.001

Table 3: Conditional Indices

Conditional Indices	1.000000	1.427707	1.6141	3.1666	25.22
---------------------	----------	----------	--------	--------	-------

Table 4: Variance Inflation Factors

Conditional Indices	1.000000	1.427707	1.6141	3.1666	25.22
---------------------	----------	----------	--------	--------	-------

Table 3: Model Results

Model	Adjusted R Squared Value	F Value
Baseline	0.7023	5222.791
Transformations + Variable Removal	0.8862	1475.468
Transformations + Variable Removal + Influential Points Removal	0.9093	2079.985

## References

- [1] *KC Housing Data*, available at <https://www.kaggle.com/swathiachath/kc-housesales-data>.
- [2] <https://cran.r-project.org/web/packages/olsrr/olsrr.pdf>.
- [3] [https://cran.r-project.org/web/packages/olsrr/vignettes/influence\\_measures.html](https://cran.r-project.org/web/packages/olsrr/vignettes/influence_measures.html).

## Contributions:





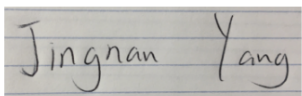
Student	Contribution	Signature
Alex Gromadzki	Data Transformation	
Karan Kant	Variable Selection and Multicollinearity	
Rakesh Ravi	Exploratory Data Analysis and Data Cleaning	
Aman Shrivastava	Model consolidation and document preparation	
Jingnan Yang	Residual Analysis	

Figure 8: Load distribution