Analyzing and Predicting the Housing Market of King County, WA

By: Erik Boldt, GTID# -264

Eboldt3@gatech.edu

ISYE 7406: Data Mining & Statistical Learning

May 17th, 2022

Abstract:

In competitive markets, buyers can gain the upper hand by understanding which characteristics of a property have the highest effect on sale price. Buyers also need an estimate on property values to place a competitive bid without overpaying. Machine learning algorithms can be used to solve both issues, sometimes simultaneously. In this report, housing market data from King County, WA between May 2014 and May 2015 was analyzed to determine which of the 20 variables impacted sale price the most. Linear regression and Random Forests models were constructed and tuned to maximize adjusted R2 values while minimizing complexity. The full dataset was then partitioned into training and testing sets, predictive models constructed and optimized through cross-validation, and then the performances were compared against the testing set. From the analysis, location data had the highest impact on price with oceanfront properties fetching high premiums. Most predictive models tended to underestimate sale price, which may be the result of the wide range of property types included in the dataset, from oceanfront properties in Seattle in small, rural plots more inland.

Introduction/Data Sources:

In recent years America's housing market has exploded to new highs as buyers are competing with dozens of others for limited supply. To stay competitive, prospective buyers need to understand which factors are driving the higher premiums in prices to make informed decisions and prepare competitive offers. Some aspects of a property may not be a top priority for some, and determining which features affect prices most will allow buyers to forgo more expensive, unnecessary features to keep their costs low.

On the same foot, compiling a competitive offer has never been more difficult, with many properties going under contract hours after listing for thousands over listing price. Basing an offer on most recent comparable home sales has typically been the best option, but in rapidly changing markets these may not capture the entire picture or miss hidden value in unique aspects of the property. Developing machine learning models to accurately predict housing prices would give potential buyers a better idea as to what to expect a home to sell for.

Determining high-value features of a property poses a unique challenge as no one feature is the top priority of all buyers. A family with children may value a great school district and a larger lot than the view from the property or being in/close to a city. Someone who is looking for a vacation home may desire a great location and view, while a larger building or lot could be viewed as a negative due to the increased upkeep.

Machine learning models can help parse these trends and calculate the effects of selected features and whether they are statistically significant. Linear models are one of the most used models due to their versatility and interpretability. Although most data are not strictly linear, various manipulation techniques can be used to form linear trends from non-linear data while remaining easy to interpret. Ensemble methods, such as Random Forests, also aid in feature interpretation as they assign an importance to each feature used to build the model, with higher rated features having a greater impact on the dependent variable.

Using these insights, multiple regression models can be built and compared to give estimates of sales price to guide buyers in constructing their offers. Linear regression and Random Forest regression can provide estimates as well as determining feature importance. However, other methods may provide more accurate predictions by focusing on predictive performance over explanatory power. Stepwise, LASSO, and Ridge regression can be used to simplify linear models and optimize predictive performance. Principal component regression can also reduce model complexity while minimizing predictive error. Additionally, generalized additive models incorporating splines may provide better predictive power by smoothing out high variance predictors.

In this report, housing data from King County, WA, which consists of Seattle and the surrounding area, from May 2014 to May 2015 (Kaggle: https://www.kaggle.com/shivachandel/kc-house-data) was analyzed to determine which property features buyers valued the most and measured their impact on sale price.

Additionally, predictive models were built and compared to estimate sale price of potential newly listed properties. The dataset consists of 21,613 datapoints with 21 variables collected, which are outlined below:

Id	Unique identification number for each listing
Date	Date the property was sold
Price	Sale price, in USD
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms
Sqft_living	Square footage of the building on the property
Sqft_lot	Square footage of the lot
Floors	Number of floors in the building
Waterfront	Whether the property is a waterfront property or not (0, no, 1, yes)
View	A rating from 0 – 4 on how nice the view is from the property
Condition	Condition of the property, graded from 1 – 5
Grade	Quality of the construction and design of the building, from 1 – 13
Sqft_above	Square footage of the building above ground
Sqft_basement	Square footage of the basement
Yr_built	Year the building was built
Yr_renovated	Year the property was last renovated
Zipcode	Zip code of the property's address
Lat	Latitude of the property's location
Long	Longitude of the property's location
Sqft_living15	Average square footage of the nearest 15 properties
Sqft_lot15	Average square footage of the nearest 15 lots

Table 1: Parameters from the King County housing market dataset

An example of the first 5 rows of data is presented below:

id	date	price	bedrooms	bathrooms	sqft_living
7129300520	20141013T000000	221900	3	1	1180
6414100192	20141209T000000	538000	1	4	2570
5631500400	20150225T000000	180000	5	6	770
2487200875	20141209T000000	604000	2	4	1960
1954400510	20150218T000000	510000	4	9	1680
7237550310	20140512T000000	1225000	5	2	5420

sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above
1180	5650	1	0	0	3	7	1180
2570	7242	2	0	0	3	7	2170
770	10000	1	0	0	3	6	770
1960	5000	1	0	0	5	7	1050
1680	8080	1	0	0	3	8	1680
5420	101930	1	0	0	3	11	3890

sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
0	1955	0	98178	47.5112	-122.257	1340	5650
400	1951	1991	98125	47.7210	-122.319	1690	7639
0	1933	0	98028	47.7379	-122.233	2720	8062
910	1965	0	98136	47.5208	-122.393	1360	5000
0	1987	0	98074	47.6168	-122.045	1800	7503
1530	2001	0	98053	47.6561	-122.005	4760	101930

Table 2: First 5 rows from the King County housing market dataset

Methodology:

All data analysis, model building, and testing were performed in the R statistical software.

The data was investigated for missing data and the presence of any unusual inputs. The date variable was input as characters and could not be analyzed by the models. The day, month, and year were extracted and converted to datetime, which is a compatible format with the models. No data was missing but there were some peculiar data points; one property was

listed as having 33 bedrooms while the building was only 1,620 square feet. It was believed that this was a typo when data was collected and the proper number was 3, but due to the large number of datapoints, this property was removed all together. It was also noted that several listings had 0 bedrooms (13 properties) or 0 bathrooms (10 properties). These were also removed from the dataset. The waterfront and zipcode variables were converted to factors as well. There was also perfect collinearity between sqft_above, sqft_basement, and sqft_living (since above + basement = living). To remove this correlation, the price was divided by the sqft_living to give the price_sqft variable, which is the price per square foot of living space. This metric is much more useful than sale price as it tends to remain more consistent between various property types.

First, linear regression was used to determine which factors impacted price the most. For the model to be a good fit, 4 separate assumptions must be met: the linearity (that the relationships between x and y are linear), constant variance (variance of the residuals for all values of x are the same), independence (observations are not correlated), and normality assumptions (that the residuals are normally distributed). These assumptions were tested on the base model and any deviance was addressed through transformations. Box-Cox testing was used to determine whether any transformations to the dependent variable would result in more normally distributed residuals. Cook's distance was also utilized to detect the presence of and remove outliers in the data.

A Random Forest model was also built using the cleaned data to determine which parameters were deemed most important for determining sale price. This ensemble method utilizes bagging, which takes samples with replacement of the training data to build decision trees. This process reduces the variance without increasing bias; each tree is trained on a different dataset (reducing bias) while averaging the results of many trees reduces variance. This model calculates the importance of each predictor based on the change in error of the out-of-bag samples when the values of the predictor in question are randomly shuffled.

Next, several predictive models were optimized and tested to compare their predictive performance on the housing market data. These models include linear regression, stepwise regression, LASSO regression, Ridge regression, principal component regression, random forest regression, and a generalized additive model. A pseudo-ensemble method was also performed, which took the average prediction of the 7 models. The data was first split into training and testing sets, with a 9:1 ratio respectively. This training set was used in the cross-validation of each model to optimize the parameters. Once the optimal models were found, they were used to predict the sale price of the testing set and the mean squared error calculated. Finally, these errors were compared to determine which model is optimal for predicting price.

The linear regression model optimized from the first half of the report was used as a predictive model. A log transformation of the dependent variable was performed because of the Box-Cox test and colinear predictors were removed to ensure the model wasn't overfit. This

model appears to fit all 4 of the linear model assumptions, so it was determined to be a good fit and should have fairly accurate predictive power.

Stepwise regression was used on the full model, carrying over the log transformation from the first model due to Box-Cox testing. This model adds and removes variables in the model and calculates the Akaike information criteria, which is an estimate of the prediction error of the model. The optimal model is then selected when the AIC is minimized.

LASSO regression is a form of linear regression that utilizes L1 regularization, which is equal to the absolute value of the coefficients. This shrinkage parameter can reduce some coefficients to zero, removing them from the model altogether. A tuning parameter is also used to modulate the strength of the penalty term. The optimal lambda value is found through cross-validation. This method is suitable for models with high levels of multicollinearity and tends to generate sparse models.

Ridge regression is similar to LASSO regression, except that it utilizes L2 regularization, which is equal to the square of the coefficients. Unlike LASSO regression, the L2 penalty cannot reduce coefficients to zero. A tuning parameter is also used in Ridge regression to modulate the strength of the penalty term. Since this method does not reduce coefficients to zero, it does not result in sparse models. However, by reducing coefficients to very small numbers it is still capable of handling models with multicollinearity.

Principal component regression uses the same concepts as principal component analysis, using dimensionality reduction to prevent overfitting and eliminating multicollinearity. First, the data is broken down into its principal components and then linear regression is ran using the principal components as predictors. The optimal number of principal components is determined by cross-validation. Principal component regression reduces the complexity of the model while maintaining the variation from the full dataset.

Random forest regression, as explained above, uses numerous decision trees, and averages the results. Cross-validation is used to determine the optimal number of trees to use and the number of predictors to use for each tree. Since Random Forest models scale with the quantity of data used, to reduce computational strain during cross-validation, the training data was sampled into smaller sets prior to cross-validation. Monte Carlo simulation was used while selecting the subsample to reduce bias.

Generalized additive models are similar to linear models in their capability to predict and explain the data. However, GAMs can model non-linear data as well as linear data. Instead of predicting coefficients for each variable, GAMs predict splines, which are flexible functions that can fit non-linear trends. Initial exploratory analysis showed non-linear trends between price and several predicting variables. As a result, splines were fitted to these variables and combined with the coefficients to produce the final model.

Finally, a pseudo-ensemble method was analyzed by taking the average predicted value from each of the 7 models and averaging them. Ensemble methods work by averaging the results of numerous models in order to reduce bias and variance. This is typically seen in Random Forests, where numerous decision trees are averaged to get the final model.

Analysis/Results:

After initial data cleaning and adjustments, a correlation plot was generated to visually interpret any trends in the data and check for obvious signs of multicollinearity (figure 1).

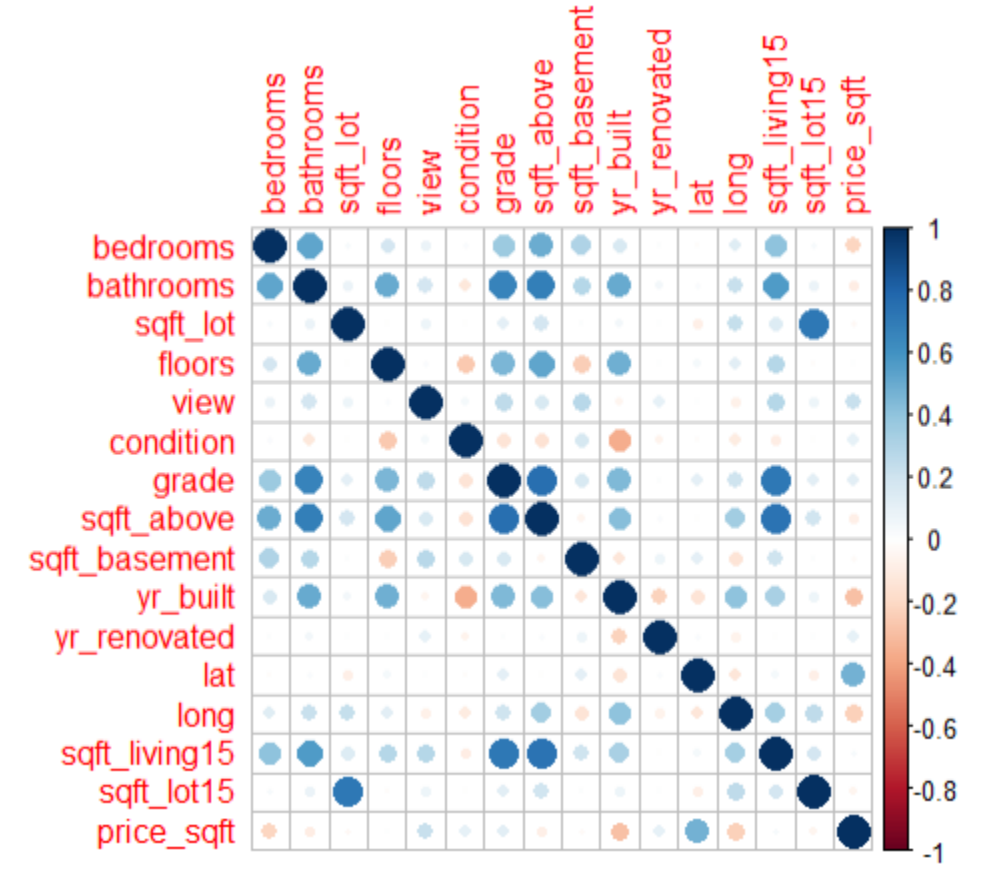


Figure 1: Correlation plot of the data

There were no obvious trends between price_sqft and the predictor variables. However, there appears to be multicollinearity between grade, sqft_above, and sqft_living15. This was further investigated later using variance inflation factors.

Graphing each of the predictors against the price_sqft showed several variables did not have a linear relationship with the dependent variable. After fitting the full model, it was noted that all 4 assumptions were not met so the model was not considered a good fit. To better fit the normality assumption, Box-Cox transformation was performed on the price_sqft with a log

transformation being optimal. Fitting the log-transformed model increased the explained variance by about 5%.

To remove the presence of any outliers or overleveraged points, the Cook's distance for each point was calculated and datapoints with a distance greater than 4/nrow(data) (standard cutoff point) were removed (1222 rows of data). A new log-transformed model was fitted to the reduced dataset, increasing the explained variance by another 7% to 83.39%.

Variance inflation factors were then calculated to determine if multicollinearity was present between any of the predictors. The results showed that the zipcode carriable was highly correlated with the latitude and longitude variables. This is expected since both are descriptors of location. Initially a new model without the zipcode variable was built but the explained variance dropped by over 20%. Another model was built removing latitude and longitude instead, which resulted in 83.35% of the variance explained. The summary of this model indicated that sqft_lot15 was not statistically significant, with a p-value of 0.6963. The final model removed this variable and resulted in 83.35% of the variance being explained. Additionally, these modifications resulted in the 4 assumptions holding up, as shown by figure 2.

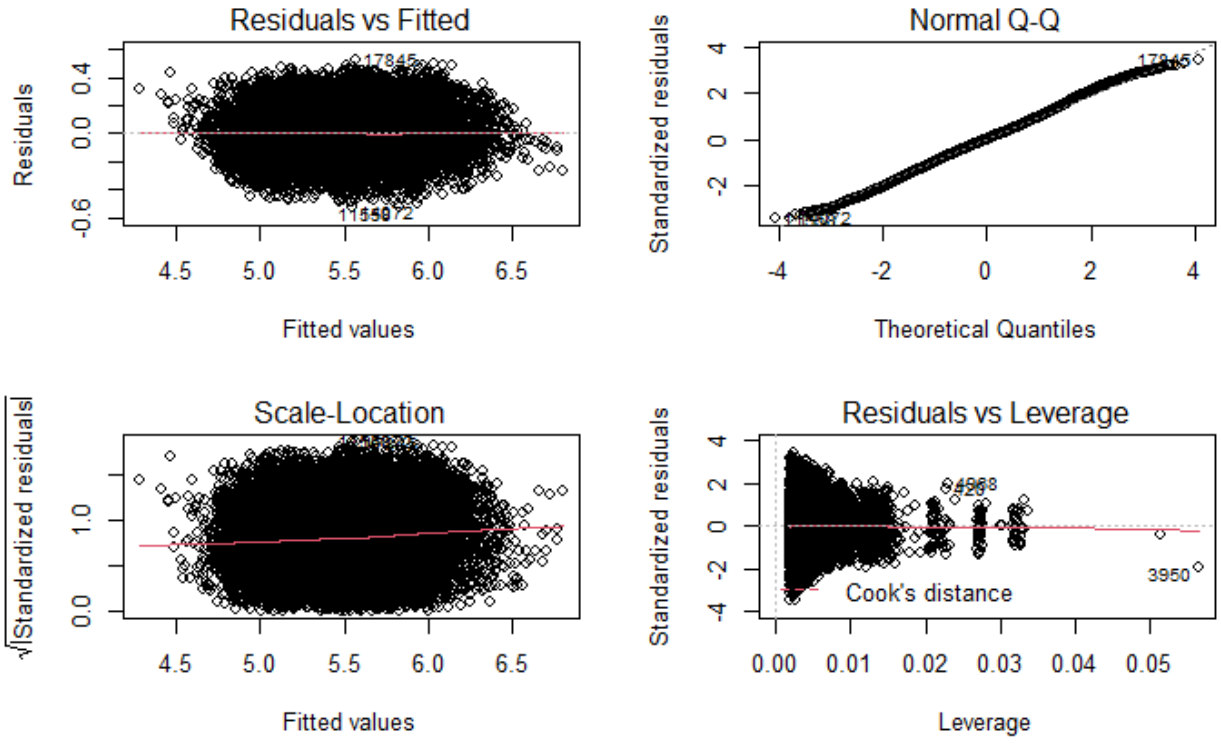


Figure 2: Diagnostic plots of the final linear model

There appeared to be some highly leveraged points through the Cook's distance plot, but when the outliers were removed once more, the explained variance did not change much (83.46%). As a result, the second removal of outliers was not kept. Graphing the residuals versus the

predictors confirmed that the constant variance assumption was upheld. Based on the well-fitted model and adjusting for predictor magnitudes, location-specific parameters appear to have the largest effect on price. Some zip codes appear to be highly desirable with large, positive coefficients. These areas are near or within Seattle while also residing close to the waterfront. Surprisingly, the number of bedrooms, sqft_above, and sqft_basement all had negative coefficients, indicating that they have a negative impact on price. This seems to indicate that size of a property is not one of the main drivers of price.

A random forest model was also fitted to the full dataset using the default parameters (ntree = 500, mtry = ncol/3 = 5) and variable importance was calculated. This model explained 85.78% of the variance in the data, which was higher than the linear model. However, variable importance indicated that the location-specific variables were the main drivers of price compared to the others. Latitude and longitude had the highest importance (0.12 each) while view had the next highest (0.10). Interestingly the date had the lowest importance out of the predictors, indicating that seasonality did not play an important role in housing prices.

Predictive performance of various models was next analyzed using a training set and testing set split. To maintain consistency, all models were evaluated on the log-transformed response variable. The optimized linear model from the explanatory data analysis was used, recording a mean squared error (MSE) of 2061.5. Stepwise regression was then used, which resulted in a similar model, except including lat and long. However, the additional predictors did not increase predictive performance and the MSE was 2061.

LASSO and Ridge regression were used to reduce the presence of multicollinearity that was seen through the correlation plot and variance inflation factors. Optimal lambda values were found through cross-validation and utilized to build the optimal models. Surprisingly no variables were completely removed from the model, though some were shrunken to very small coefficients. Although not identical, both LASSO and Ridge regression models were very similar in their coefficient values. LASSO regression did outperform Ridge regression by a small amount (MSEs of 4931 and 5068 respectively), but neither were able to outperform the optimized linear model where multicollinearity was removed manually.

Principal component regression was applied next and the optimal number of principal components to be used was determined through cross-validation. Error was minimized using all 86 components, indicating that dimensionality reduction was not achievable without losing out on predicting power. The resultant full linear model performed similarly to the optimized linear model, with an MSE of 2061.5 as well. It is worth noting that the optimized linear model is still preferred as it reduces model complexity while maintaining full predictive power.

Random forest regression was then analyzed. Due to the large quantity of data being directly correlated with computational strain while building these models, the training data was subsampled into smaller sets and cross-validation was performed to determine the optimal number of trees and predictors to try. Monte Carlo simulation with 10 repetitions was used

during data subselection to reduce bias. The optimal number of parameters to analyze at each split was 1 while the optimal number of trees was 100. The optimal model explained 80.5% of the variance and resulted in an MSE of 4274. Interestingly, no one variable appeared more important than the others with respect to this model, though date was much less important than the others.

Spline smoothing was used while generating a generalized additive model, which combines linear and nonlinear terms. From the original exploratory data analysis, it was noted that several predictors showed non-linear trends with respect to price, including sqft_lot, sqft_above, sqft_basement, yr_built, yr_renovated, sqft_living15, and sqft_lot15. These predictors were fitted with splines using the default parameters. After fitting, it was noted that all predictors were statistically significant except for a couple zip codes. The model explained 85% of the deviance and resulted in a MSE of 1887. This model outperformed all subsequent models and ended up being the best performing predictive model.

Finally, a pseudo-ensemble model was formed by averaging the predicted prices from the previous seven models. In theory averaging the predictions over several models should reduce variance without increasing bias. Averaging the predicted prices resulted in an MSE of 2245.7, slightly worse than the optimized linear model. A summary table of all models and their MSEs is presented below:

Model	**MSE**
Linear Regression	2061.513
Stepwise Regression	2060.988
LASSO Regression	4931.009
Ridge Regression	5068.003
Principal Component Regression	2061.569
Random Forest Regression	4274.357
Generalized Additive Model	1887.063
Pseudo-ensemble Method	2245.721

Table 3: Mean squared errors of predictive models on the testing set

The generalized additive model incorporating splines for non-linear data provided the lowest MSE and was chosen as the optimal model.

Investigating the optimal model further, the difference between the predicted price and the actual price was determined and analyzed. The generalized additive model, on average, underpredicted the actual sale price of a property by about $3 per square foot. The median error on predictions was only $0.68 underestimated. A plot of the prediction errors is provided below:

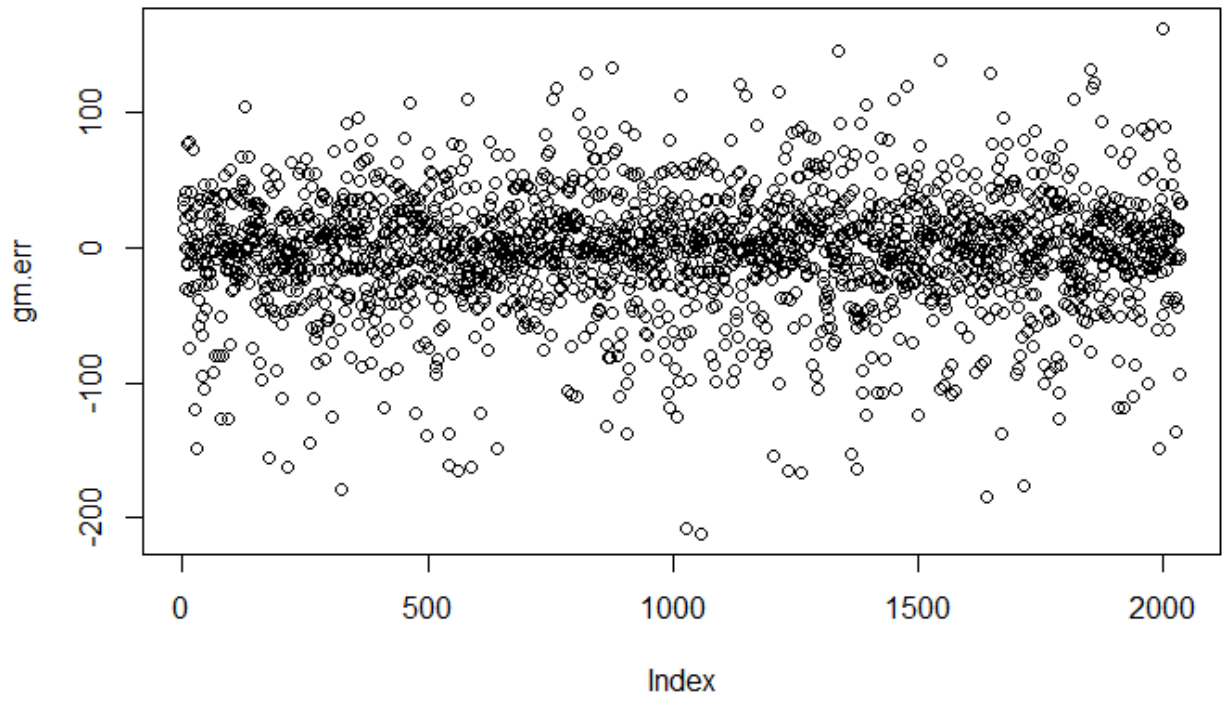


Figure 3: Scatterplot of the prediction errors from the optimal model, GAM

There appears to be several properties that the model was unable to predict accurately and was off by over $150 per square foot. The property that was underpredicted the most does not appear out of the ordinary: a 3-bedroom, 1-bathroom 1140 square foot house with neither a great view nor on the waterfront. However, the zip code of 98115 indicates that this property is in northern Seattle, where properties are currently valued over $1 million. The sale price of $702k ($617.54 / sqft) does not seem out of place for 7 years ago, so it may be that this area is highly desirable, and the location factor outweighs any other parameter. The most overpredicted property also does not appear unusual: a 1-bedroom, 1-bathroom, 590 sqft condo located in central Seattle. It is unclear whether the location parameters were overvalued for this property or if the lower square footage of the building and lot threw off the model. The average lot size and building size of the nearest 15 properties were also quite low, which may have added to the overestimation of the price. From this it appears that the optimized model undervalues highly desirable neighborhoods while overpredicting on smaller properties like condos.

Conclusions:

Descriptive analytic modeling showed that location was one of the most desirable aspects of properties sold in King County, WA between May 2014 and May 2015. Linear and random forest models both showed that zip code or latitude and longitude had larger effects on price than any other variable. The view and whether the properties were on the waterfront

also played outsized roles in determining price. This aligns with the most common saying in real estate: “Location, location, location!”. It also aligns with the fact that most aspects of a property can be changed, except for the location. Most buyers appear to be giving up move-in readiness and lot/building size to live in highly desirable locations. Taking this under consideration, potential homebuyers that are flexible with their location could investigate less popular towns to keep their costs lower. By sacrificing location, homebuyers may be able to afford bigger, more modern homes than if they were to purchase in more pricey neighborhoods. Limitations do apply to this analysis as the models reflect the entirety of King County and trends may differ based on area. As such, smaller models could be built by separating the data by zip code and reperforming the analysis. This would provide better insights into location specific trends and would reduce the number of outliers.

Predictive models were also able to give accurate predictions on sale price for most properties. However, just like the descriptive models above, some of the most and least desirable areas may have added various outliers that skew the data. The optimal generalized additive model gave the best predictive performance but still tended to underpredict sale price. Potential homebuyers who may utilize these models to guide their offer should take this drawback under consideration. Although the predictions on average were close to the sale price, homebuyers will need to add a little more to their offer to reliably have it accepted. These models could be improved the same way as the descriptive models: separating the data by location (zip code) and reoptimizing each model. This would reduce the number of outliers (oceanfront properties versus farmhouses in the country) for each location and most likely provide more accurate predictions on sale price.

Lessons Learned:

One of the biggest lessons learned from this report is that large data is much more difficult to work with than smaller datasets. There could be several subgroups within the dataset that should be isolated and analyzed separately from one another. Unlike the housing dataset, which was nicely separated by zip code, other datasets may not have this luxury and clustering/classification models may need to be used to separate the data by likeness.

Another lesson learned was that a model with high predictive performance may not have the best explanatory power, and vice versa. Linear regression appeared to explain a large portion of the variance but was unable to outperform the generalized additive model. Transformations were able to make the data more linear, but non-linear trends persisted and added too much error. Utilizing non-linear models adds flexibility and can result in better predictive performance without adding too much bias. However, small changes in the data can have large effects on the model’s fit. Care should be taken to clean the data properly prior to fitting models to avoid time-consuming rework.

Appendix:

Summary statistics of the complete dataset, note the 33 bedroom property

```
##        id                 date               price            bedrooms
##  Min.   :1.000e+06   Length:21613       Min.   :  75000   Min.   : 0.000
##  1st Qu.:2.123e+09   Class :character   1st Qu.: 321950   1st Qu.: 3.000
##  Median :3.905e+09   Mode  :character   Median : 450000   Median : 3.000
##  Mean   :4.580e+09                      Mean   : 540088   Mean   : 3.371
##  3rd Qu.:7.309e+09                      3rd Qu.: 645000   3rd Qu.: 4.000
##  Max.   :9.900e+09                      Max.   :7700000   Max.   :33.000
##    bathrooms      sqft_living       sqft_lot           floors
##  Min.   :0.000   Min.   :  290   Min.   :    520   Min.   :1.000
##  1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040   1st Qu.:1.000
##  Median :2.250   Median : 1910   Median :   7618   Median :1.500
##  Mean   :2.115   Mean   : 2080   Mean   :  15107   Mean   :1.494
##  3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688   3rd Qu.:2.000
##  Max.   :8.000   Max.   :13540   Max.   :1651359   Max.   :3.500
##    waterfront            view          condition        grade
##  Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 1.000
##  1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000
##  Median :0.000000   Median :0.0000   Median :3.000   Median : 7.000
##  Mean   :0.007542   Mean   :0.2343   Mean   :3.409   Mean   : 7.657
##  3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000
##  Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :13.000
##    sqft_above   sqft_basement       yr_built     yr_renovated
##  Min.   : 290   Min.   :   0.0   Min.   :1900   Min.   :   0.0
##  1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0
##  Median :1560   Median :   0.0   Median :1975   Median :   0.0
##  Mean   :1788   Mean   : 291.5   Mean   :1971   Mean   :  84.4
##  3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.0
##  Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015.0
##     zipcode           lat             long        sqft_living15
##  Min.   :98001   Min.   :47.16   Min.   :-122.5   Min.   : 399
##  1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490
##  Median :98065   Median :47.57   Median :-122.2   Median :1840
##  Mean   :98078   Mean   :47.56   Mean   :-122.2   Mean   :1987
##  3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360
##  Max.   :98199   Max.   :47.78   Max.   :-121.3   Max.   :6210
```

```
##    sqft_lot15
##  Min.   :   651
##  1st Qu.:  5100
##  Median :  7620
##  Mean   : 12768
##  3rd Qu.: 10083
##  Max.   :871200
```

33-bedroom property does not look realistic as it's too small (1620 sqft). This was removed

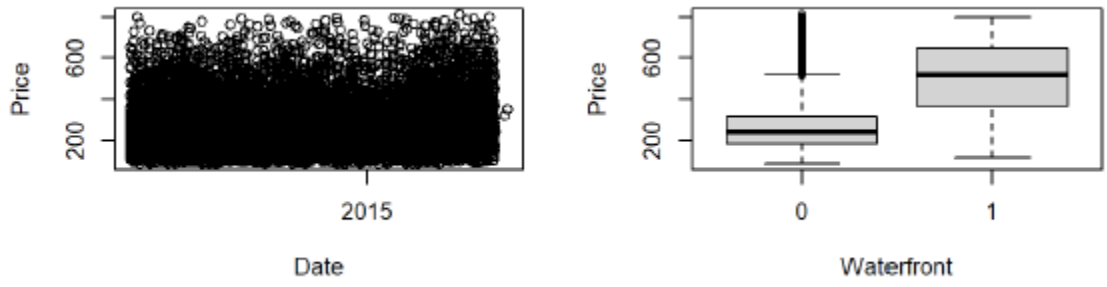
```
##              id            date  price bedrooms bathrooms sqft_living sqft_lot
## 15871 2402100895 20140625T000000 640000       33      1.75        1620     6000
##       floors waterfront view condition grade sqft_above sqft_basement yr_built
## 15871      1          0    0         5     7       1040           580     1947
##       yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 15871            0   98103 47.6878 -122.331          1330       4700
```

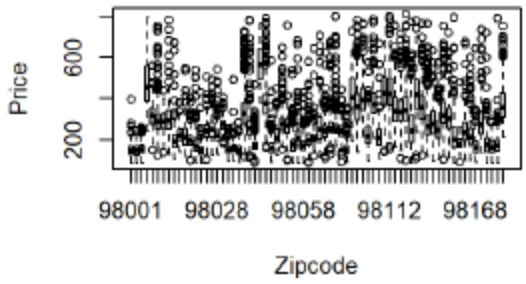
Properties without bedrooms or bathrooms were also removed

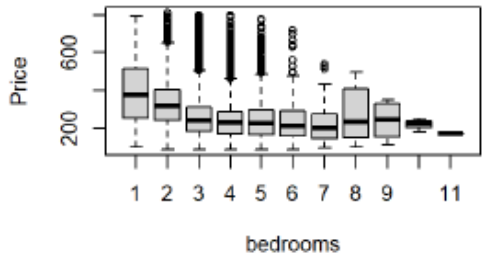
New summary data after cleaning it up

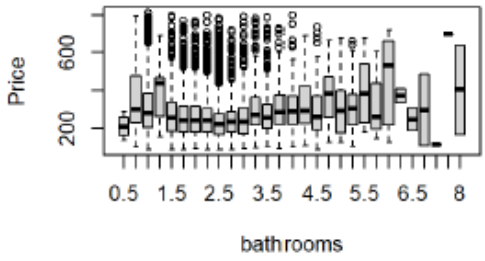
```
##      date              bedrooms        bathrooms        sqft_lot
##  Length:21596       Min.   : 1.000   Min.   :0.500   Min.   :    520
##  Class :character   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.:   5040
##  Mode  :character   Median : 3.000   Median :2.250   Median :   7619
##                     Mean   : 3.372   Mean   :2.116   Mean   :  15100
##                     3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.:  10686
##                     Max.   :11.000   Max.   :8.000   Max.   :1651359
##
##      floors      waterfront      view          condition        grade
##  Min.   :1.000   0:21433    Min.   :0.0000   Min.   :1.00   Min.   : 3.000
##  1st Qu.:1.000   1:  163    1st Qu.:0.0000   1st Qu.:3.00   1st Qu.: 7.000
##  Median :1.500              Median :0.0000   Median :3.00   Median : 7.000
##  Mean   :1.494              Mean   :0.2343   Mean   :3.41   Mean   : 7.658
##  3rd Qu.:2.000              3rd Qu.:0.0000   3rd Qu.:4.00   3rd Qu.: 8.000
##  Max.   :3.500              Max.   :4.0000   Max.   :5.00   Max.   :13.000
##
##    sqft_above   sqft_basement       yr_built     yr_renovated
##  Min.   : 370   Min.   :   0.0   Min.   :1900   Min.   :   0.00
##  1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.00
##  Median :1560   Median :   0.0   Median :1975   Median :   0.00
##  Mean   :1789   Mean   : 291.7   Mean   :1971   Mean   :  84.47
##  3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.00
##  Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015.00
##
##     zipcode           lat             long        sqft_living15
##  98103  :  601   Min.   :47.16   Min.   :-122.5   Min.   : 399
##  98038  :  589   1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490
##  98115  :  583   Median :47.57   Median :-122.2   Median :1840
##  98052  :  574   Mean   :47.56   Mean   :-122.2   Mean   :1987
##  98117  :  553   3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360
##  98042  :  547   Max.   :47.78   Max.   :-121.3   Max.   :6210
##    sqft_lot15       price_sqft
##  Min.   :   651   Min.   : 87.59
##  1st Qu.:  5100   1st Qu.:182.29
##  Median :  7620   Median :244.63
##  Mean   : 12759   Mean   :264.11
##  3rd Qu.: 10083   3rd Qu.:318.27
##  Max.   :871200   Max.   :810.14
```

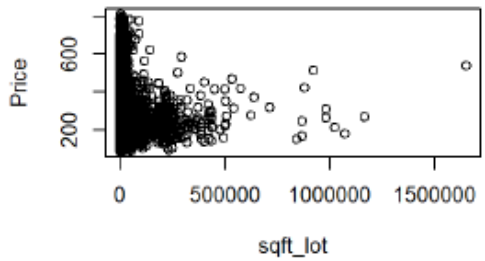
Plotting predictors against price_sqft

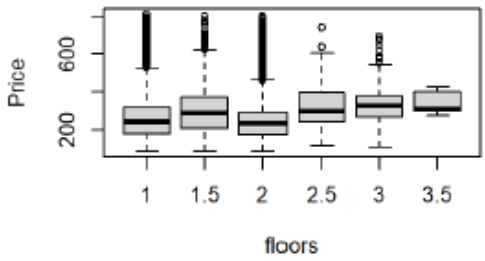


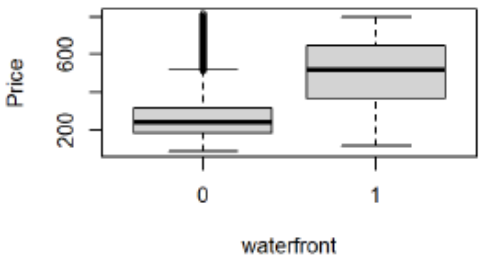

Price
200
600
98001
98028
98058
98112
98168
Zipcode

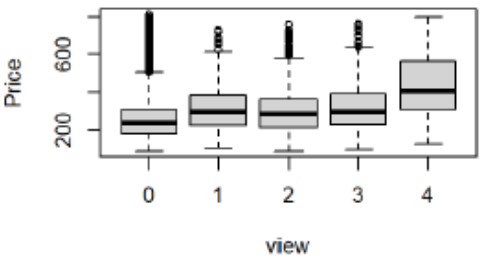

Price
200
600
1
2
3
4
5
6
7
8
9
11
bedrooms

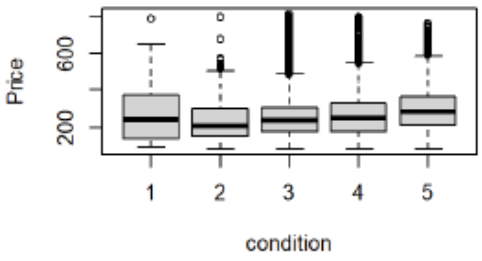

Price
200
600
0.5
1.5
2.5
3.5
4.5
5.5
6.5
8
bathrooms

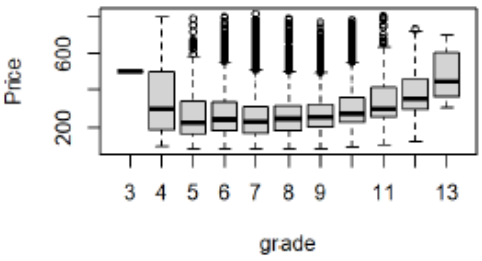

Price
200
600
0
500000
1000000
1500000
sqft_lot

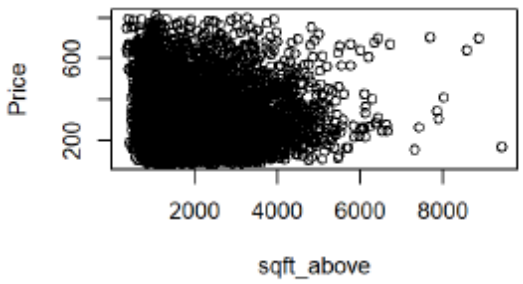

Price
200
600
1
1.5
2
2.5
3
3.5
floors

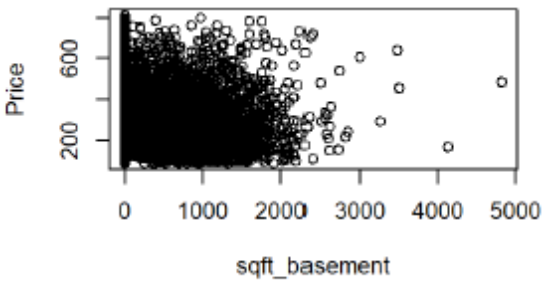

Price
200
600
0
1
waterfront

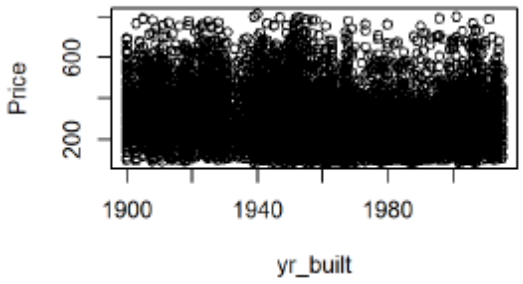

Price
200
600
0
1
2
3
4
view

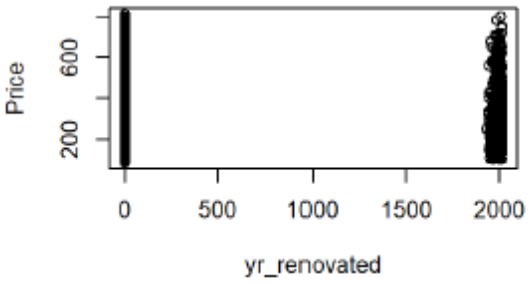

Price
200
600
1
2
3
4
5
condition

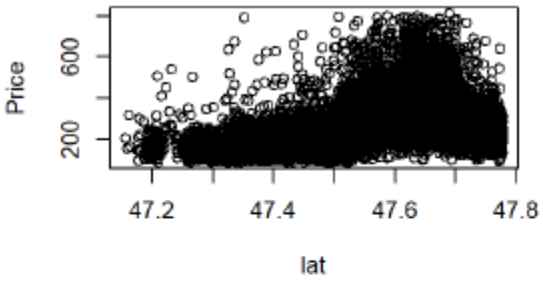

Price
200
600
3
4
5
6
7
8
9
11
13
grade

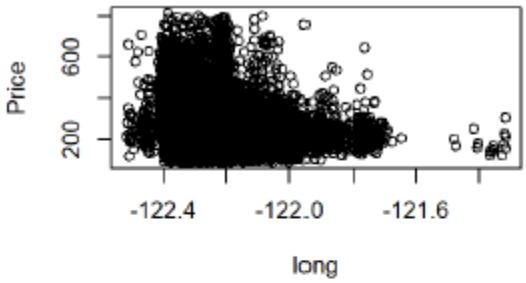

Price
200
600
2000
4000
6000
8000
sqft_above

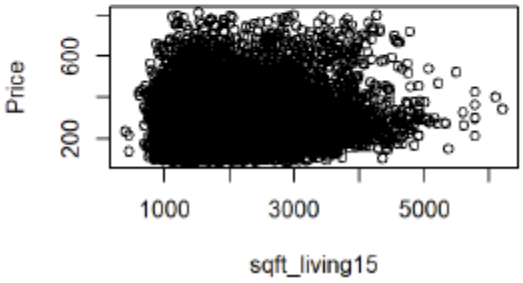

Price
200
600
0
1000
2000
3000
4000
5000
sqft_basement

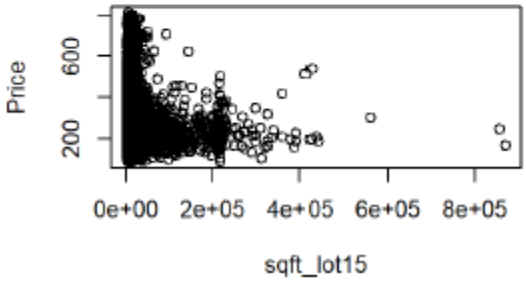

Price
200
600
1900
1940
1980
yr_built


Price
200
600
0
500
1000
1500
2000
yr_renovated

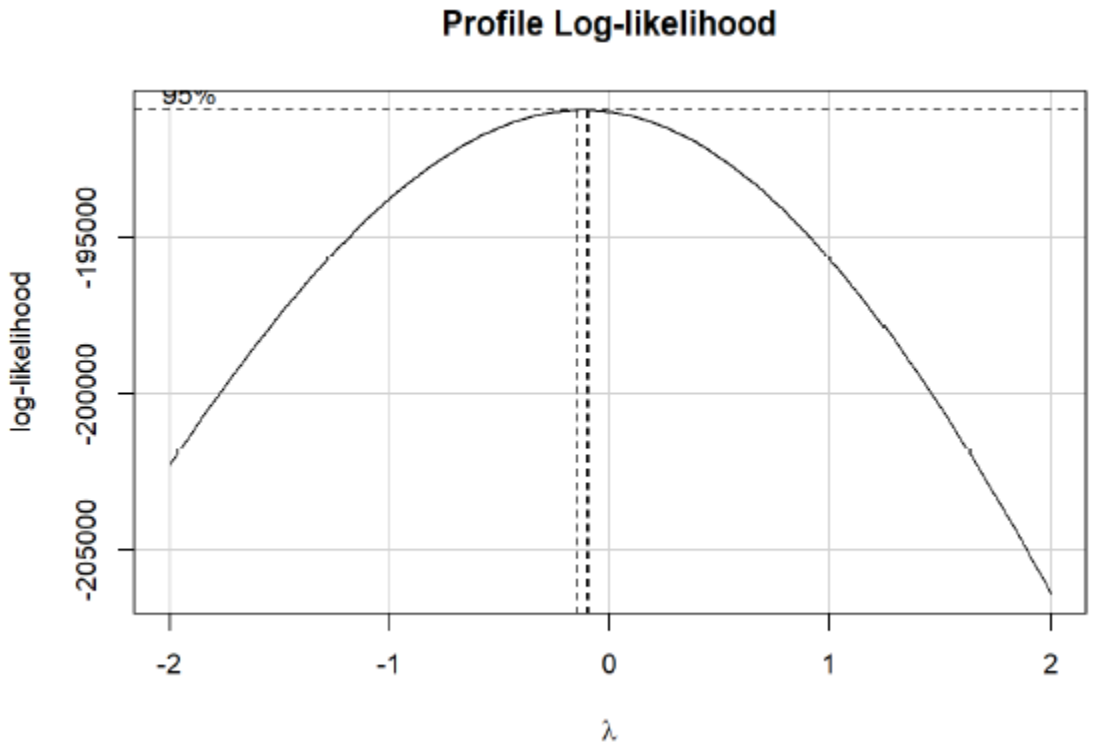

Price
200
600
47.2
47.4
47.6
47.8
lat


Price
200
600
-122.4
-122.0
-121.6
long

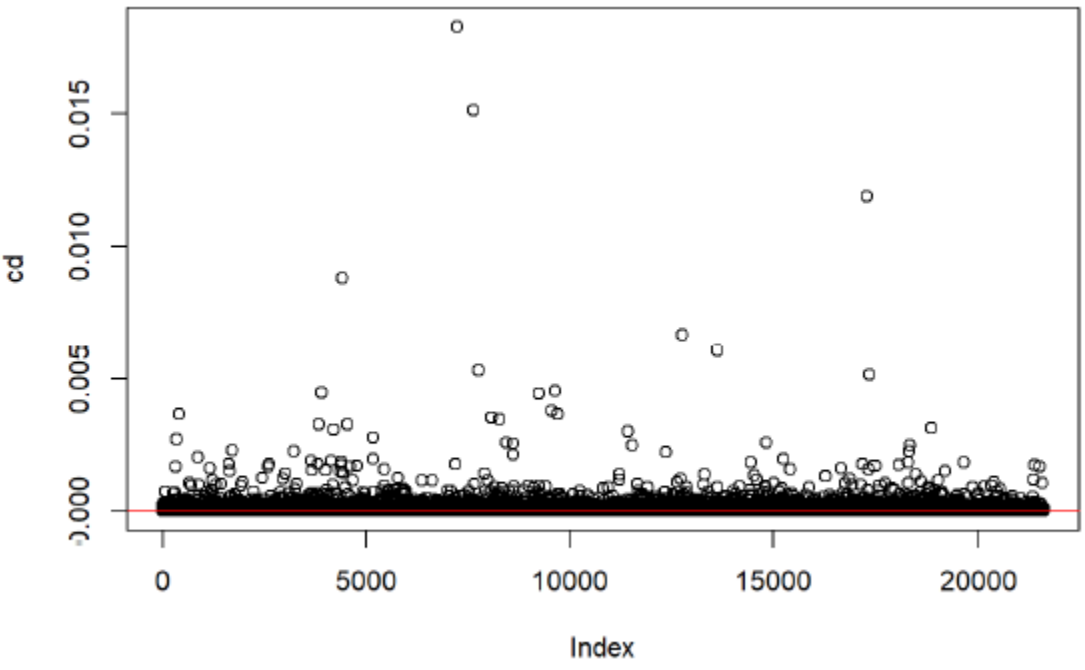

Price
200
600
1000
3000
5000
sqft_living15


Price
200
600
0e+00
2e+05
4e+05
6e+05
8e+05
sqft_lot15

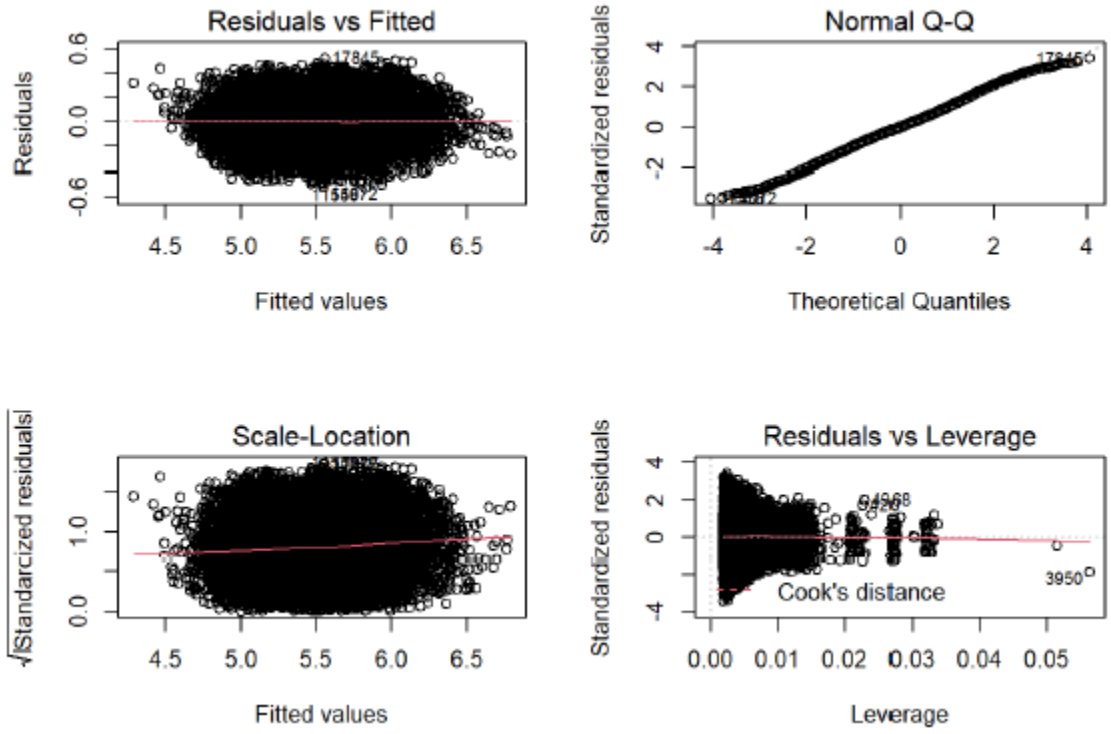
Box-Cox testing of the full linear model



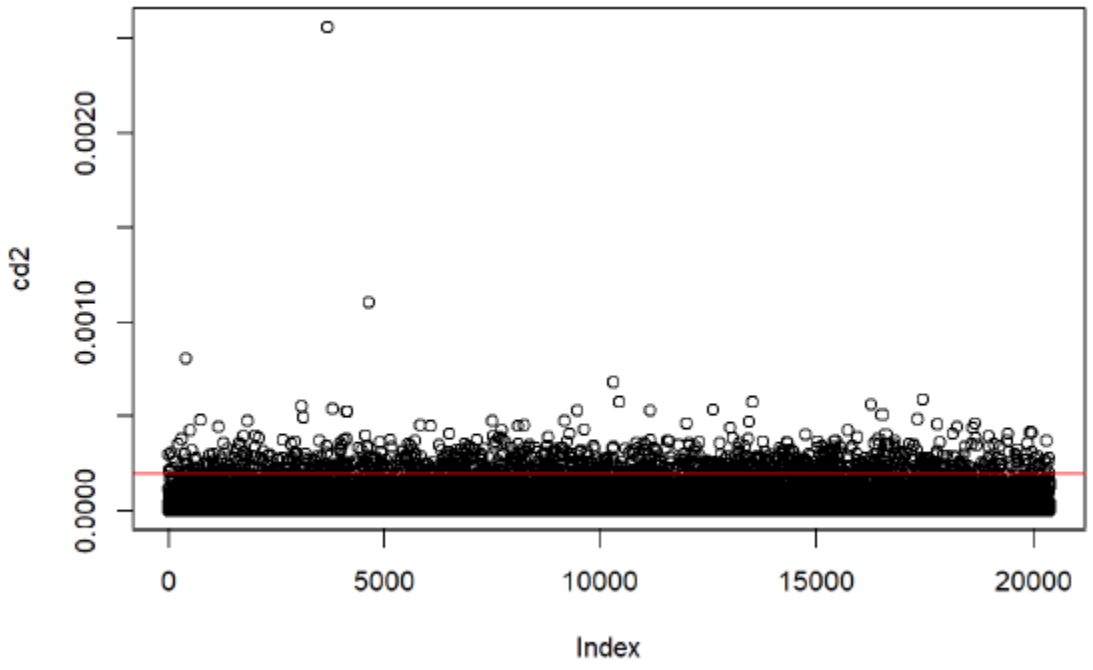
Cook's distance of the log-transformed model, redline = 4/nrow(data) (cutoff point)



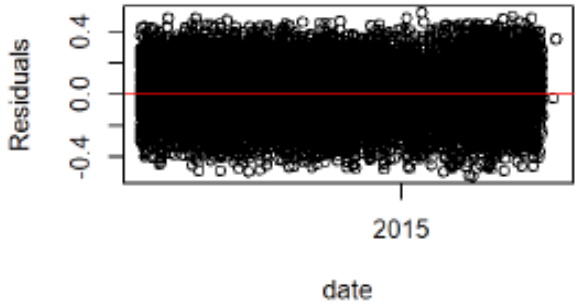
Checking the assumptions of the optimized linear model

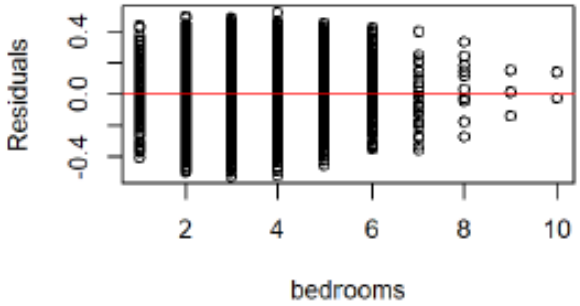


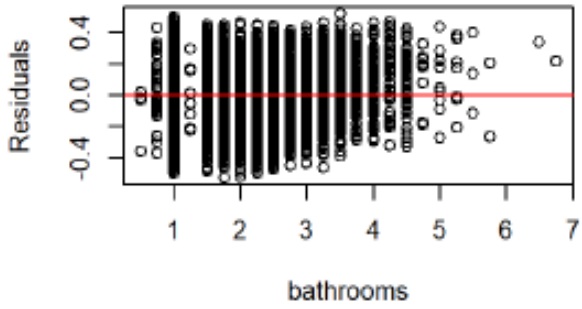
Cook's distance testing of the optimized model, redline = 4/nrow(data) (cutoff point)

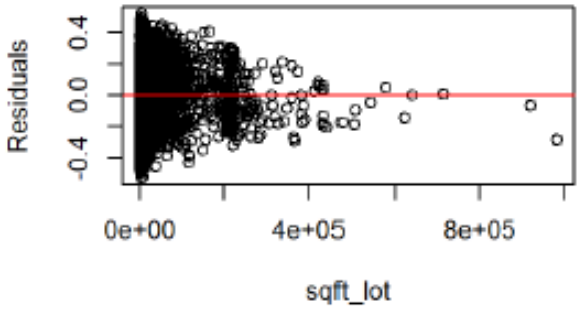


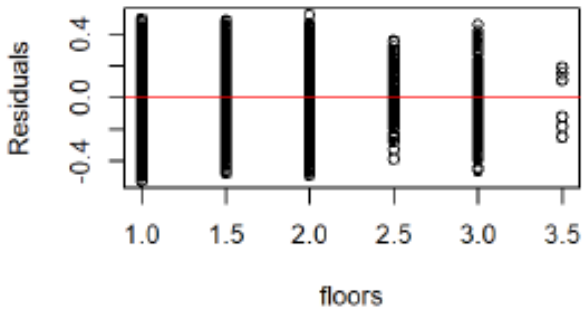
Checking the distribution of the residuals for the final linear model

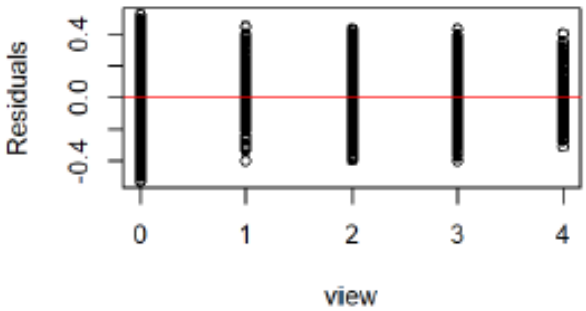


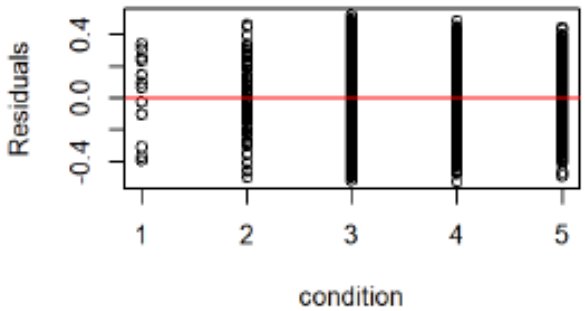


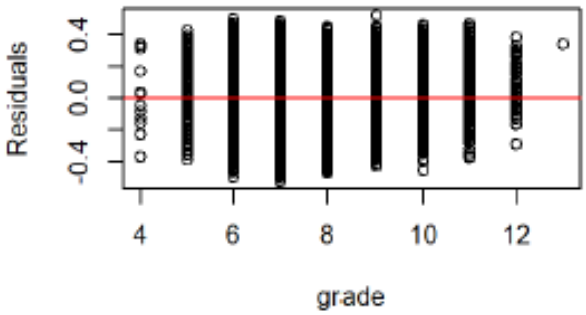


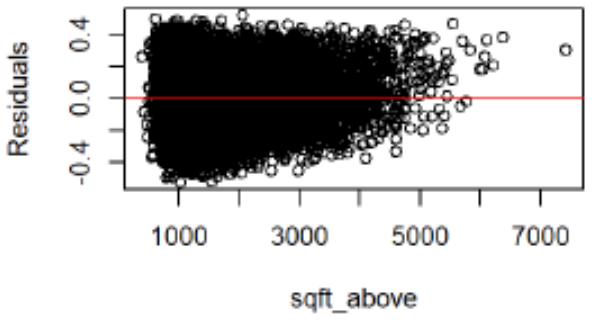


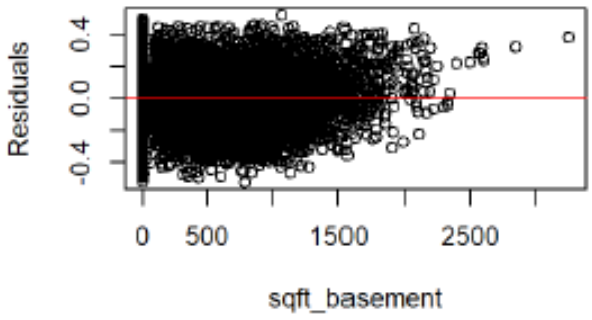


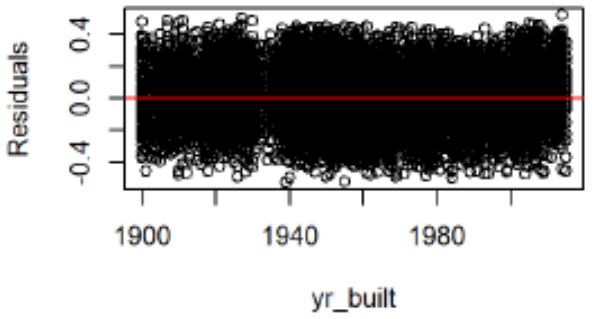


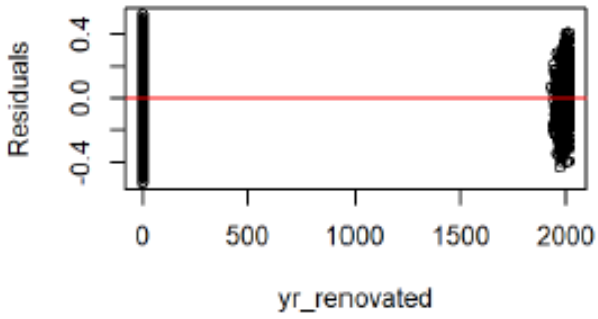












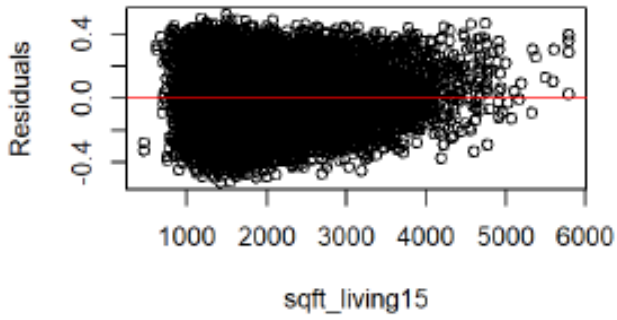
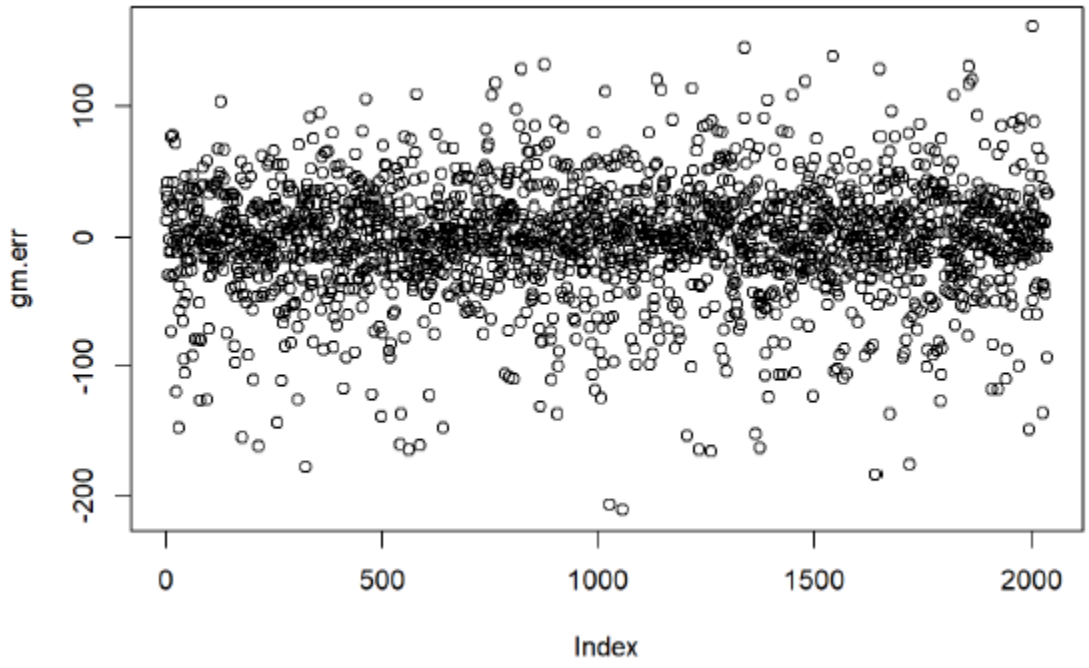


Table of MSEs of the predictive models

```
##           Model      MSE
## 1  Linear Model 2061.513
## 2      Stepwise 2060.988
## 3         LASSO 4931.009
## 4         Ridge 5068.003
## 5           PCR 2061.569
## 6 Random Forest 4274.357
## 7           GAM 1887.063
## 8   Combination 2245.721
```

Statistics and plot on the errors of the best model, the generalized additive model

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -211.3528  -22.2442   -0.6779   -2.9227   21.0632  161.7432
```



The property with the largest error

```
##             date bedrooms bathrooms sqft_lot floors waterfront view condition
## 10695 2015-02-25        3         1     6120    1.5          0    0         3
##       grade sqft_above sqft_basement yr_built yr_renovated zipcode     lat
## 10695     7       1140             0     1926            0   98115 47.6822
##           long sqft_living15 sqft_lot15 price_sqft
## 10695 -122.309          1800       4080   617.5439
```

The property that was the most overpredicted

```
##             date bedrooms bathrooms sqft_lot floors waterfront view condition
## 3988 2014-07-01        1         1      833      1          0    0         4
##      grade sqft_above sqft_basement yr_built yr_renovated zipcode     lat
## 3988     7        590             0     1926            0   98122 47.6082
##          long sqft_living15 sqft_lot15 price_sqft
## 3988 -122.299           780       1617   342.3729
```

Please see the attached pdf document for the entirety of the code used in this analysis.