**STAT 6021**

**Project 2: Housing Prices Analysis**

**Project Report**

**Group 2 Members:**

James Caldwell

Sarah Christen

Nathan Koh

Margaret Lindsay

**Section 1: Summary of Findings**

For this project, the team analyzed a dataset containing information about house sale prices for King County, Washington. The dataset included 21,613 observations of homes sold between May 2014 and May 2015. The team explored issues with the data, created visualizations with the data, fit a linear regression to predict house sale prices, and fit a logistic regression model to predict if a home is of good quality.

Before creating the linear and logistic models, the dataset needed to be analyzed and cleaned to fix data quality issues. From this data analysis, 17 of the observations were removed and not used in the models. There are 21 variables included in the dataset to describe qualities of each home sold. After a thorough review of each variable, 6 variables that were not useful and/or relevant to the analysis were removed and 2 variables were created.

After exploring the dataset’s variables, the team investigated fitting a multiple linear regression (MLR) model on the data to predict a house’s price from its features. Using half the data as a training dataset, the team created a model using a house’s living square footage, zipcode/region, waterfront view, grade, and year built. Using the test dataset for model evaluation, the team found that the model can predict a house’s price within 16% over half of the time. The MLR model tended to be the least accurate for lower value homes. Generally, the team found that increasing the square footage, grade, and waterfront view increases the value of a home. Older homes tend to be less valuable, and certain regions of Seattle, such as Eastern Seattle, tend to have more expensive homes.

This report also explains how a model was developed to assess whether a home is of good quality. A home is classified as good quality homes if its condition value is greater than 3 and its grade value is greater than 7. Since whether a home is good quality is a binary variable, meaning there are only two possible values (i.e., yes or no), a logistic regression model can be used. The logistic regression model summarizes whether a home is good quality using probability and odds.

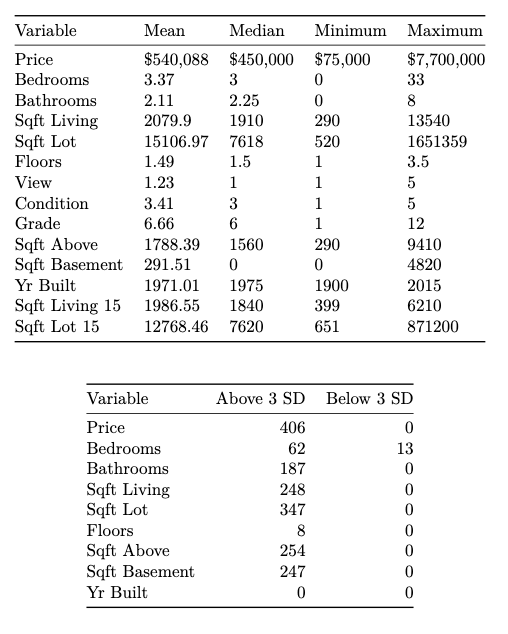
The model developed predicted the odds that a home was good quality using the following variables: price, bedrooms, bathrooms, floors, view, year built, rear built, year renovated, square foot of interior living space of the nearest 15 neighbors, and region. The model also indicates how much influence these variables have over the probability of a home being good quality. The results of the logistic regression model would benefit homeowners who are looking to sell their home and estimate whether their home would be considered good quality. The analysis could also be a resource for realtors or developers looking to sell good quality homes or build homes that would likely be considered good quality.

**Section 2: Description of Data and Variables**

The data set used is from kaggle and contains house sale prices for King County, Washington for homes sold between May 2014 and May 2015. Below are definitions for each of the variables in the data set and summary statistics of the relevant numerical variables.

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| id | Unique ID for each home sold |
| date | Date of the home sale |
| price | Price of each home sold |
| bedrooms | Number of bedrooms |
| bathrooms | Number of bathrooms, where 0.5. accounts for a room with a toilet but no shower |
| sqft\_living | Square footage of the interior living space |
| sqft\_lot | Square footage of the land space |
| sqft\_above | Square footage of the interior housing space that is above ground level |
| sqft\_basement | Square footage of the interior housing space that is below group level |
| floors | Number of floors |
| Categorical | |
| waterfront | A dummy variable for whether the home was overlooking the waterfront or not |
| view | An index from 0 to 4 of how good the view of the property was |
| condition | An index from 1 to 5 on the condition of the property |
| grade | An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high-quality level of construction and design |
| Build Details + Geography | |
| yr\_built | The year the house was initially built |
| yr\_renovated | The year of the house’s last renovation |
| zipcode | What zipcode area the house is in |
| lat | Latitude |
| long | Longitude |
| sqft\_living15 | The square footage of interior housing living space for the nearest 15 neighbors |
| sqft\_lot15 | The square footage of the land lots of the nearest 15 neighbors |

**Summary Statistics**



To better represent location and geography as a variable, we created a "region" variable to the original data set. We sourced this data from the King County, Washington Geographic Information System (GIS). We cross referenced regions used by GIS for health reporting in 20201 to the GIS zip code data2.

We also created a binary variable to differentiate good quality homes. We defined a home to be of good quality if its condition value is greater than 3 and its grade value is greater than 7.

**Section 3: Data Observations and Issues**

During our initial data analysis, we found three types of data entry errors:

1. Homes with 0 bedrooms
2. Homes with 0 bathrooms
3. Home with improbable number of bedrooms

There were 17 homes in the original data set that fit into one of these three error types.

Based on the summary of variables we noticed that the minimum value for bedrooms and bathrooms is 0. This does not make sense; a home should have at least 1 bedroom and 1 bathroom. There are 13 homes with 0 bedrooms and 10 homes with 0 bathrooms. Among these 23 homes, there are 7 that have both 0 bathrooms and 0 bedrooms. To address these errors, we have removed the 16 homes from the data set that have either 0 bedrooms or 0 bathrooms.

Another look at the summary of variables we noticed that the maximum value for bedrooms was 33. This is much higher than the median 3 and mean 3.37. We isolated the data for this home to see if the price, square feet, bathrooms, and floors match what we would expect for a 33-bedroom home. This home has 1.75 bathrooms, 1,620 square feet of living space, 1 floor, and it was sold for $640,000. Thirty-three bedrooms is not feasible for a home of that size with the corresponding attributes. We decided to also remove this data entry from the data set.

After addressing data entry errors, we took a deeper look at the variables in the data set and their usefulness in a linear regression and logistic regression model for house prices. We determined we could remove the date, sqft\_above, and sqft\_basement columns. The date variable is not an attribute of a home and not relevant to predicting house prices. There were multiple variables to describe the size in square feet of each home (living, lot, above, and basement). The square feet of interior living space is a summation of the square footage of the interior housing space that is above ground (sqft\_above) and below ground level (sqft\_basement). Due to the dependency these variables have, we decided to only use the sqft\_living variable to encapsulate the full living space.

The variables that characterize the location of the homes in the original data set are zipcode, lat, and long. These variables are not useful in a linear or logistic regression model. We believe location will be an important variable to use as a predictor of home prices. To incorporate an effective location variable, we use the region variable we created from King County, Washington GIS data.

Finally, the waterfront variable in the data set is read into R as an integer. We set this column as a factor to use it as an indicator variable.

**Section 4: Visualizations**

The histogram below shows the distribution of selling prices for homes in the data set. The histogram clearly demonstrates the selling price of most homes in the data set was under $1 million. However, there are homes represented that sold for over $7 million.

A graph showing a blue line

Description automatically generated

Bivariate visualizations can assist in determining how the price of a homes is related to other variables, such as variables related to size (i.e., number of bedrooms or bathrooms, and interior square footage) or location (i.e., views and whether the home is waterfront property).

As seen in the scatterplot below, the cleaned train data set contains homes ranging from 0 to 10 bedrooms. The plot shows how the selling price varies for houses with the same number of bedrooms. For example, most 6-bedroom homes sold for less than $2 million. However, there are 2 observations that sold for over $3 million.

A graph of a number of bedrooms

Description automatically generated

In addition to the number of bedrooms, the number of bathrooms is an important factor many homeowners consider when purchasing a home. The scatterplot of selling price against number of bathrooms shows the distribution of prices for homes with the same number of bathrooms.

A graph showing a number of bathrooms

Description automatically generated

The overall living space also influences a home’s selling price. As expected, there appears to be a positive relationship between the home’s interior square footage and the selling price. However, as seen by the observation with greater than 12,500 square feet of interior living space, other factors can influence lead to a house selling for a price comparable with homes with half of the square footage.

A graph showing a number of black dots

Description automatically generated

Other variables in the data set, such as a home’s view or whether it is waterfront property, relate to the home’s location rather than its size. As seen in the boxplots below, homes with the highest view ranking (4) have a higher median price than homes with the worst view ranking (0). Moreover, homes with a waterfront view have a higher median price than those without waterfront views.

A graph showing a number of points

Description automatically generated with medium confidence A graph showing a number of houses

Description automatically generated with medium confidence

The spread of prices even within the same view or waterfront category indicates other variables influence a home’s selling price. The variable grade can be broken into categories to describe a home’s level of construction and design. The categories are defined based on the discussion of the Kaggle data set: low (1-3), medium low (4-6), average (7), medium high (8-10), and high (11-13). The boxplot below shows the median selling price increases for increasing grade scores.

A graph showing the average and high prices

Description automatically generated with medium confidence

Another variable to consider in relation to selling price is the age of a home. The density plot below shows the distribution of homes by year built and whether the home was over or under the median selling price. For homes built roughly between 1937 and 1987, more homes sold under the median price than above. However, for homes built between 1900 and 1937 as well as 1987 to 2015, more homes sold over the median price than under.

A graph showing the difference between a house construction and a construction site

Description automatically generated

Multivariate visualizations can also aid in understanding how the variables in the data set relate to the selling price. Taking the scatterplot discussed previously of price against number of bedrooms and adding the grade provides further insight into the data. Most of the observations at higher price points are lighter blue, which indicates they are a higher grade. Conversely, observations at lower price points are darker blue.

A graph of blue dots

Description automatically generated

Comparing price against number of bedrooms and condition shows that for the most part, the higher ranked conditions are priced higher than lower ranked condition homes with the same number of bedrooms.

A graph of a number of bedrooms

Description automatically generated

**Section 5.1: Linear Model Building for Predicting Price**

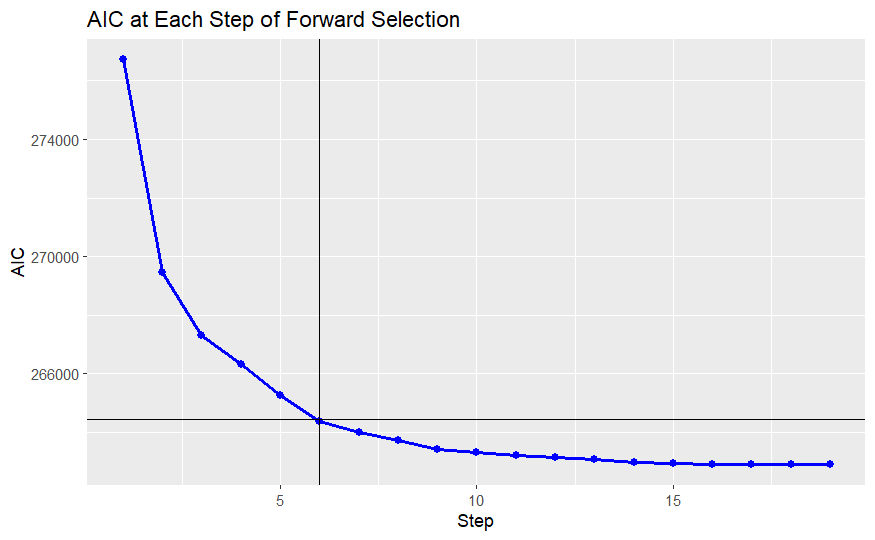
This section outlines how the team used linear regression to build a model that explains how prices of homes in King County, Washington are based on the other variables. Before building a model, the team cleaned the data and added a new zip code variable as described in section 2. Next, the team split the data into a random, 50-50 split of training and testing data.

To find an optimal multiple-linear regression (MLR) model for the dataset, the team used an automated search procedure: Forward Selection. This process begins with a model with no predictors, and then it adds in predictor variables one at a time until a desirable stopping point is reached. The team felt that this process would hopefully lead to a simple model that would not overfit the training data. The forward selection used here bases each iteration on Akaike information criterion (AIC), which is a measure of fit with penalty for additional parameters. Other evaluation metrics could have been added here, but the team felt that using AIC alone should suffice for choosing an accurate model.

The automated search procedure was based on 22 predictor variables for price. The procedure iterated 20 times before stopping on its ideal model choice. However, the model that was automatically selected was quite complex:

price ~ sqft\_living + region + grade + yr\_built + waterfront + zipcode + lat + view + bedrooms + long + bathrooms + condition + sqft\_above + sqft\_living15 + yr\_renovated + id + sqft\_lot15 + sqft\_lot

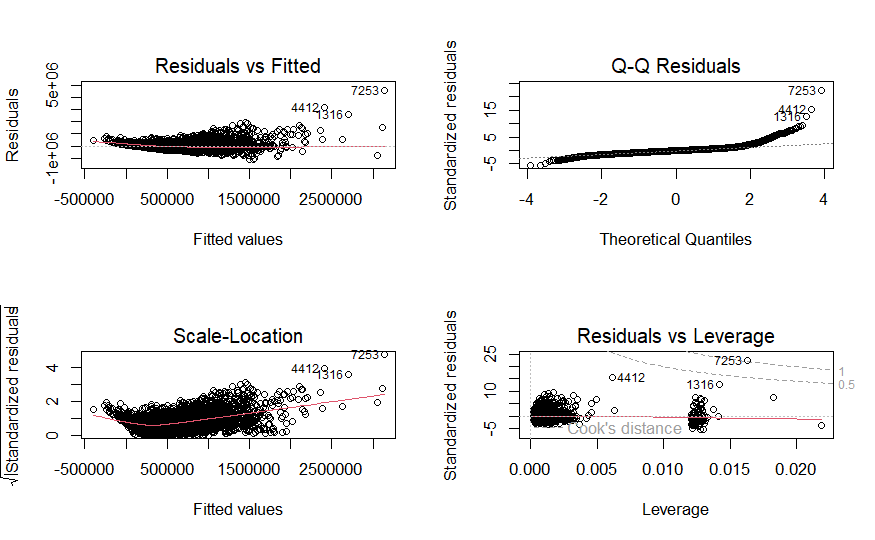
The team noticed there was a diminishing return in reduction of AIC after a few iterations (as shown below). It is likely that adding more variables leads to overfitting, and the reduction in AIC is fairly insignificant. The team decided to select the model after just 6 iterations, which should balance accuracy and improved interpretability without overfitting. Also, choosing a simpler model avoids potentially correlated variables, such as region, zip code, and latitude + longitude.



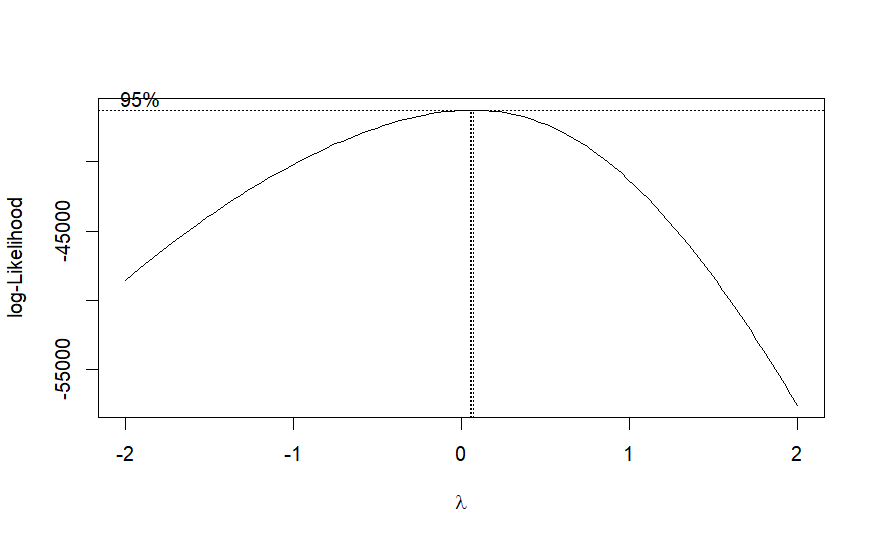
The model after step 6 was chosen (highlighted in the figure above). The variables selected for modeling are:

price ~ sqft\_living + region + grade + yr\_built + waterfront

The diagnostic plots for this model are shown below.



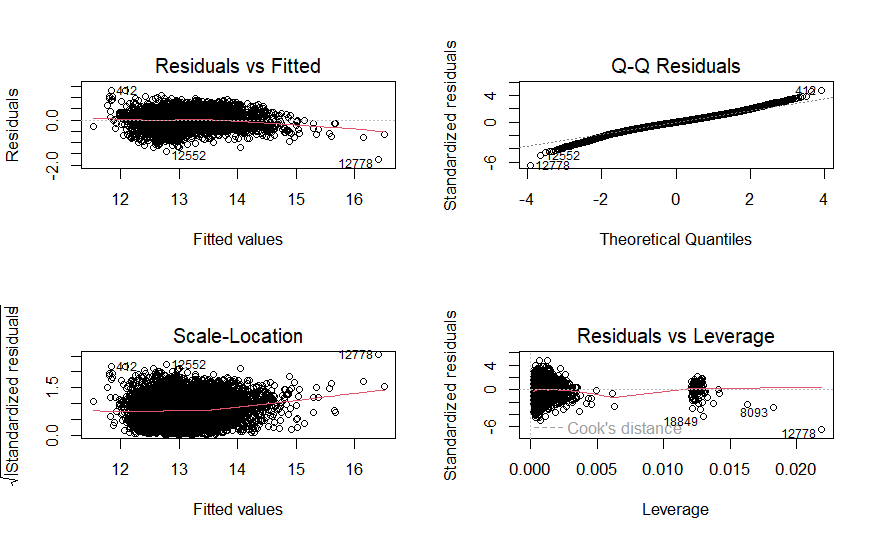
From the residuals plot, the errors seem to mostly be centered around zero, but there is a clear trend with a non-constant variance. Assumption 2 of linear regression does not appear to be met. To remedy this, the team explored transforming the response variable. The box-cox plot is shown below.



The 95% confidence interval for this model is very small, but since it is close centered around 0, the team decided to take the natural log of price. This allows for easier interpretation later as well.

ystar = log(price)

Ystar is introduced as a new variable, the transformation of price. Fitting an MLR with ystar as the predicted variable produces the following diagnostic plots:



The residuals have a mean of 0 with constant variance, so the regression assumptions appear to be met.

Before settling on the chosen model, the team investigated high leverage datapoints and outliers. From the above figure, there are no datapoints with a cook’s distance greater than 1, which is a good sign. The team found that many datapoints were flagged for various reasons: 1) 6389 entries were flagged as high leverage using externally studentized residuals. 2) 89 homes were flagged as outliers with externally studentized residuals greater than 3. 3) 1268 houses were flagged as influential observations using the DFITS method. After a thorough investigation of data entry errors in section 3, the team does not feel confident removing all these homes from the dataset as they are most likely legitimate datapoints. Thus, the team decided to keep all datapoints for two reasons. There are a lot of datapoints in this dataset, so the most egregious outliers/high leverage points should still not carry that much weight on a model, and most likely the datapoints already removed in section three removed the worst datapoints in the set.

Thus, the model chosen for predicting a house’s price from the given dataset is:

Log(price) = sqft\_living + region + waterfront + grade + yr\_built

With coefficients, the model is:

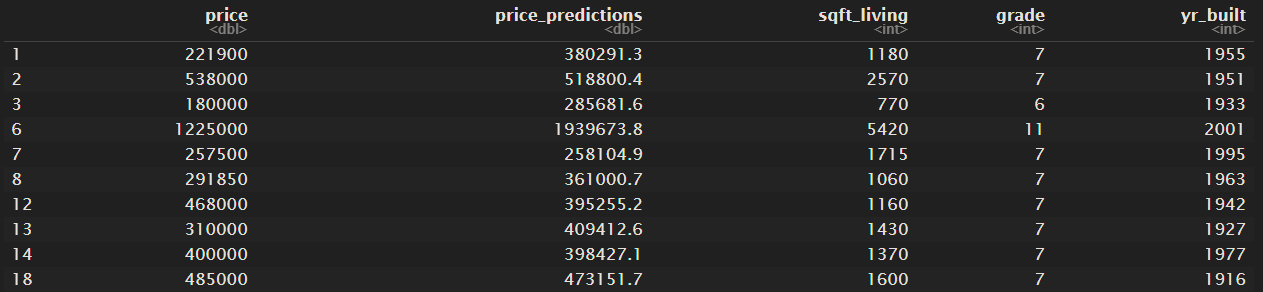
Log(price)=17.7+ 0.0002× sqft\_living−0.2×regionNORTH−0.1 ×regionSEATTLE− 0.4× regionSOUTH +0.5×waterfront+ 0.2×grade−0.003×yr\_built

Interpretation of the model:

Generally, the model seems reasonable. The intercept term is close to 0, which conceptually seems logical: a house of “nothing” is worth nothing. Many of the terms are positively correlated: it is logical that increasing the square footage, grade, and waterfront increase the value of a home. Alternatively, an older home is worth less (negative correlation). Some regions, presumably more affluent neighborhoods, such as the East are associated with an increase in home value. Whereas less desirable places to live, such as the southern region, correlate with less expensive homes.

**Section 5.2: Linear Model Evaluation**

Next, the team made predictions on the test data using the model created in section 5.1. The first several rows of the price vs price predictions are shown in a table below.

 To add points of comparison, two other models were also fit. One model was fit with all the predictors recommended by the Foward Selection automated search procedure. This model, or “model full” was:

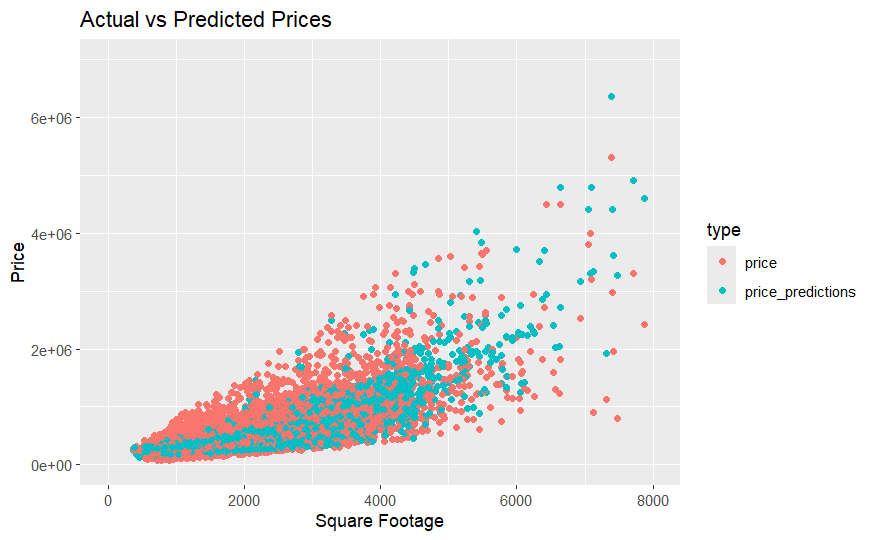
ystar ~ sqft\_living + region + waterfront + grade + yr\_built + view + bedrooms +

bathrooms + condition + sqft\_lot15 + sqft\_living15 + id + floors + sqft\_lot

A third model was chosen that was a purposefully “bad” model, so that a baseline for performance could be set. This model was:

ystar ~ floors + sqft\_lot

To visualize our ideal model’s predictions, the team plotted price and predicted price vs the most influential predictor variable, living space square footage. Clearly, the model is predicting reasonably well. The model appears to be more conservative often, predicting housing prices that are slightly lower than the actual house price.



The RMSE’s for the three models were:

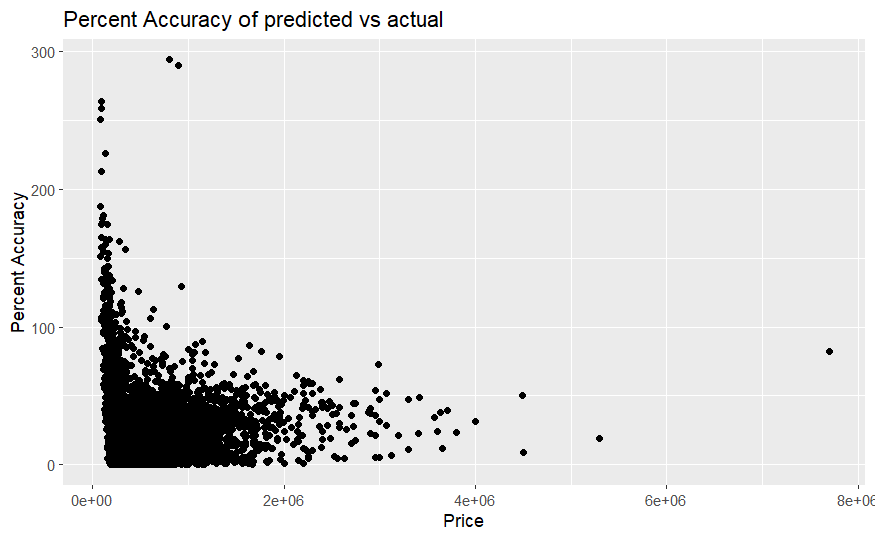
Ideal model: 205,889.2  
 Bad model: 355,894.3  
 Full model: 196,551.9

Surprisingly, the full model seems to be predicting slightly better than our simplified, ideal model. However, our team still justifies using the simpler model since it avoids correlated variables and increases interpretability.

A final visualization that the team used was to calculate the percent accuracy of our chosen model for each datapoint.

% Accuracy = ((predicted – actual)/actual )\*100%

The figure below shows this accuracy vs price:



And the summary below shows the percent accuracy of all the datapoints:

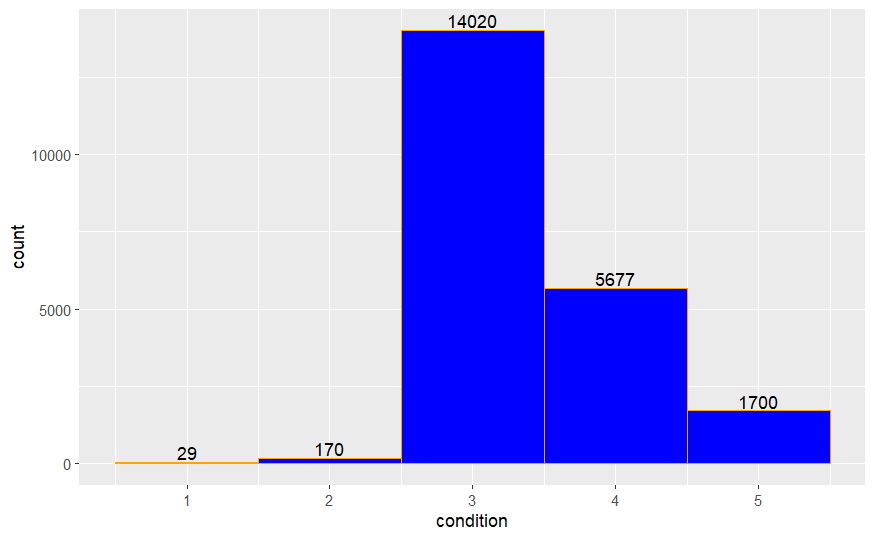
Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.00245 7.23068 15.75443 20.70642 27.86723 294.31886

Notably, our model seems to have trouble predicting lower value houses. However, the model can predict half of the homes with an accuracy below 16%, which the team feels is excellent performance given the complexity of the given problem.

In conclusion, the team was able to create an MLR that can predict housing prices with relative accuracy. The team made several improvements to the model, and the model meets all the regression assumptions. The team evaluated our model against other models using RMSE, percent accuracy, and visual performance plots. The team is happy with the model’s performance but recognizes that more complex methods, different modeling techniques, and further data engineering (especially more outlier/influence removal) could improve upon the model’s accuracy.

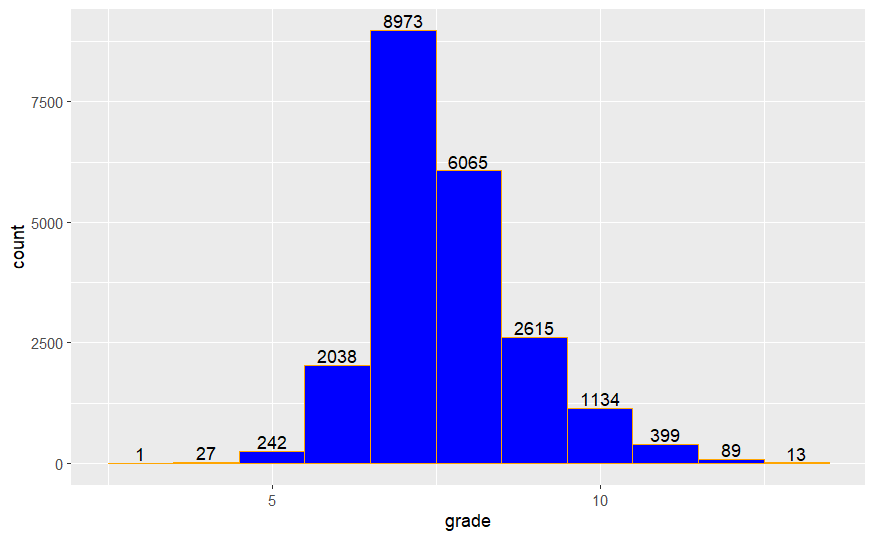
**Section 6: Defining a Good Quality Home**

The purpose of this section is to provide visualizations that explore what characteristics are associated with good quality homes. For the purposes of this project, any homes that have a condition value greater than 3 and grade greater than 7 fulfill these criteria. Additionally, we removed the variables *Date, Sqft\_above, Sqft\_basement, Zipcode, Lat, and Long* due to agreeing that these variables were mainly informative. In lieu of this, we created a new variable called *Region* that summarized the zip codes of the dataset into the following regions: East, North, Seattle, South, with each of them representing a certain part of King County, Washington. Using this information, we started by creating bar charts that visualized the data spread for homes that qualify as good quality homes.

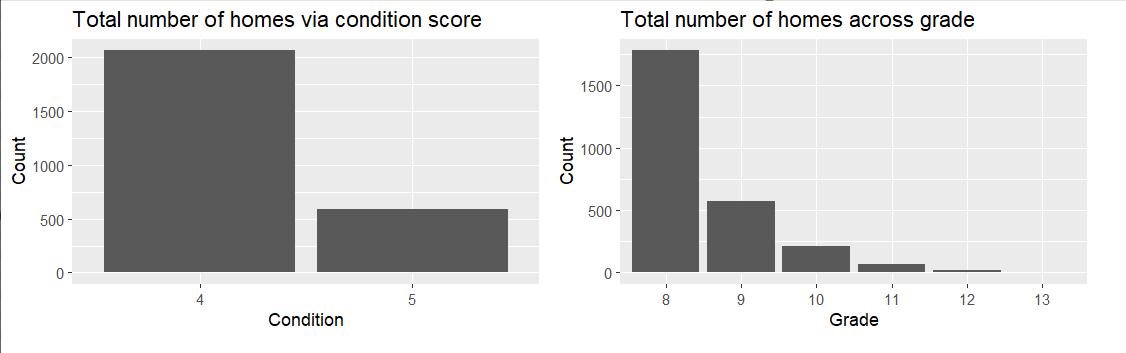
Figure 1

Observing the dataset overall, we see that not many homes classified as condition 1 or 2, but a majority were classified between conditions 3 and 5. Since our focus is on analyzing homes that were classified between condition 4 and 5, we removed the homes that did not qualify from our dataset.

Similarly, when visualizing the comparison of the total number of homes that fulfill the mentioned grade threshold, we see the following in Figure 2:

Figure 2

Like in Figure 1, we see that there are not as many homes that classify as grades 1 to 6. However, the majority of the data lies between grades 7 and 13. Due to the main focus of this section, we removed grades 1 through 7 from our dataset, and performed our analysis on homes graded 8 to 13 and conditions of 4 and 5. As such, the following bar charts represent the distribution of the dataset that we will be using for the remainder of this section. Visually, there are more homes with a condition of 4 than 5, and will help us determine which variables to analyze to determine their significance when comparing them to the condition and grade of a home. When looking at the total number of homes by grade, a majority of homes have a grade of 8 with the total number of homes decreasing for each grade level. Overall, there are 2,652 homes which we will use to determine what characteristics (variables) define these homes as good quality.

Figure 3

To begin our analysis, we started by comparing the condition of homes to their prices as boxplots. Note that due to the extreme outliers, it made the boxplots difficult to see. Two visuals were created – one with and one without – to provide context regarding the observations made.

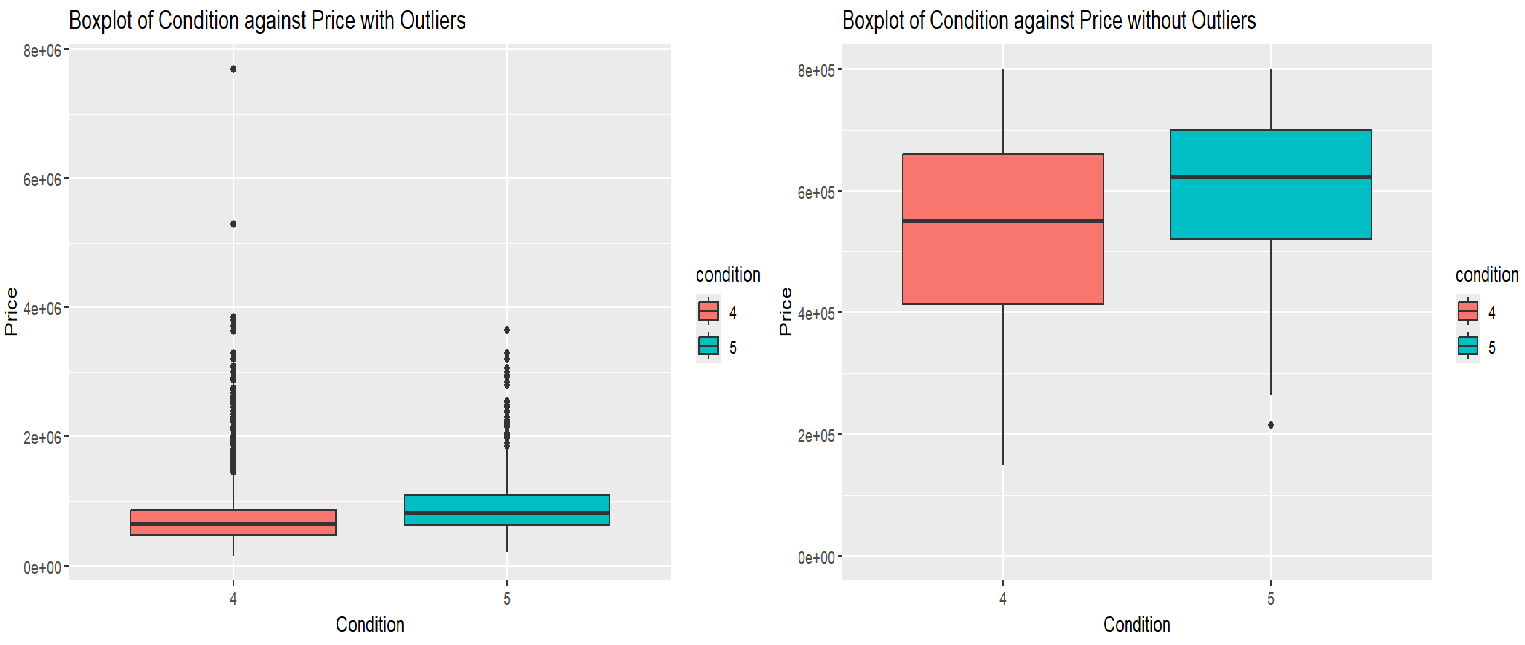


Figure 4

From the boxplots, we can observe the following. When comparing the median price of homes based on both conditions, there is an increase when comparing homes with condition 4 and condition 5. Based on our initial observations, this was to be expected. Secondly, we can also see that the boxplot for condition 4 is wider than that of condition 5. This indicates that there were more data points for homes categorized as condition 4 than condition 5.

Using this information as pretext, we created a boxplot comparing the prices of homes based on their grade score. This boxplot is shown below.

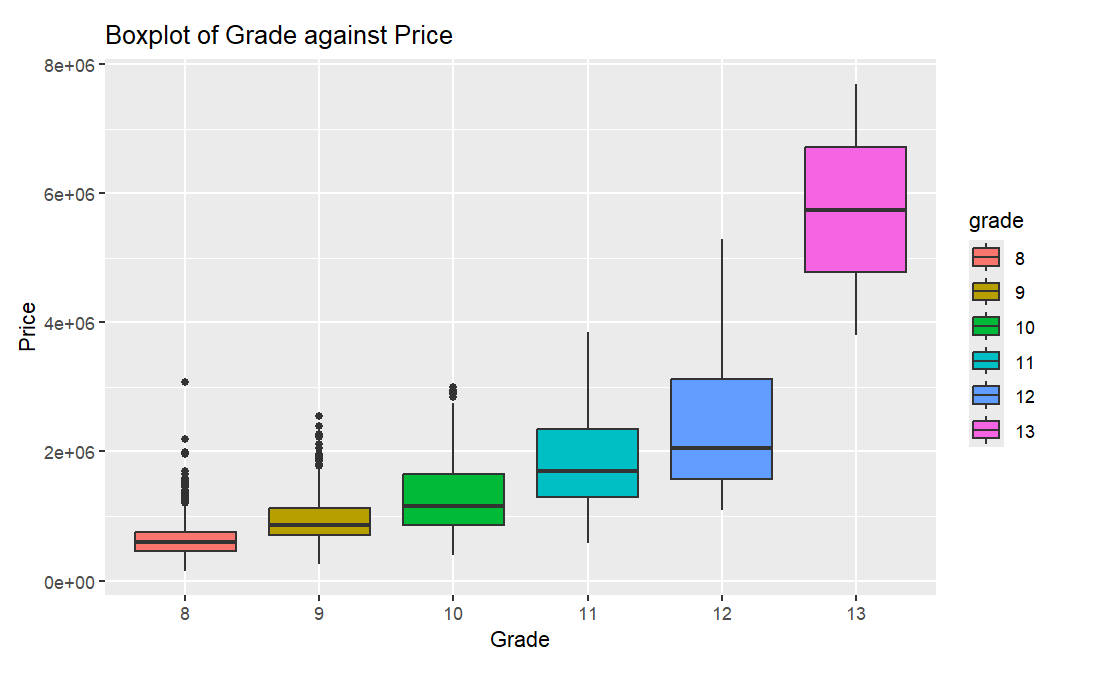
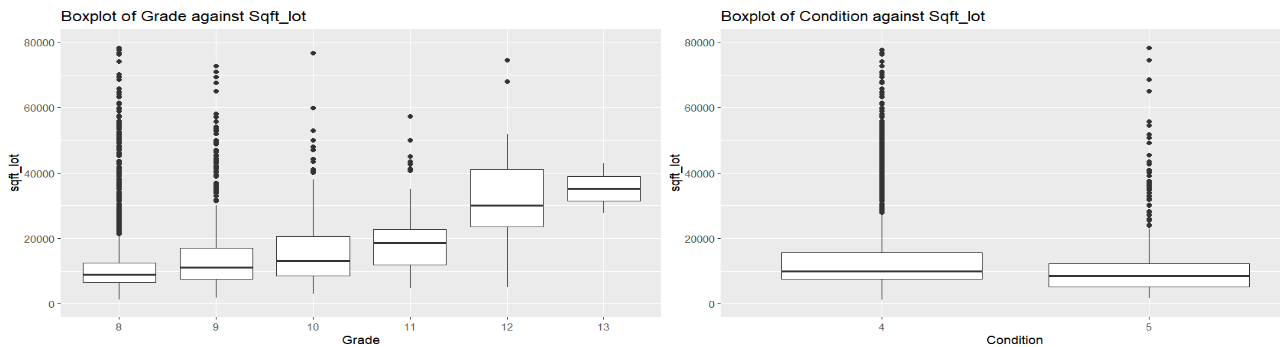


Figure 5

Like in Figure 4, we see an increasing trend in the median price based on the grade of the home, with grade 13 defined as a home with a “high-quality level of construction and design.” Additionally, the interquartile ranges for each boxplot become wider which indicates a wide data spread. From our initial observations, this is what we also expected since each grade is a level below the next. From our initial observations, we can begin comparing other predictors to grade and condition to determine which variables are the most significant.

From these two observations, we decided to compare *sqft\_lot* to *grade* due to agreeing that the amount of land a property has would increase the valuation of a home in its price, therefore possibly improving its grade and condition. Shown below is Figure 6 and 7, which visualizes these variables. We found the following observations with these boxplots:

* There is a positive trend with the median sqft\_lot increasing as grade improves
* Out of the 2,652 homes that qualify as good quality homes, most of the data is skewed towards grades 8 to 10. There are only 85 homes that are classified between grades 11 and 13
* The frequency of outliers decreases across each grade level also indicating the lack of observations in the higher grades

Figure 6 and Figure 7

We also included a two-way table that provides the data distribution for the total number of observed homes throughout this section which is shown in Figure 8 (n = 2652 homes). As mentioned above, a majority of the homes lie between grades 8 to 10.



Figure 8

So, we concluded that sqft\_lot was significant when analyzing grade, but not condition. Lastly, we created a combined side-by-side boxplot from prior observations, to determine if any new observations could be made. Figure 9 shows the median line for homes that are condition 5 and grade 11 are outliers when compared to the entire boxplot of condition 5 homes. We concluded that the lack of data between grades 11 to 13 led to this result.

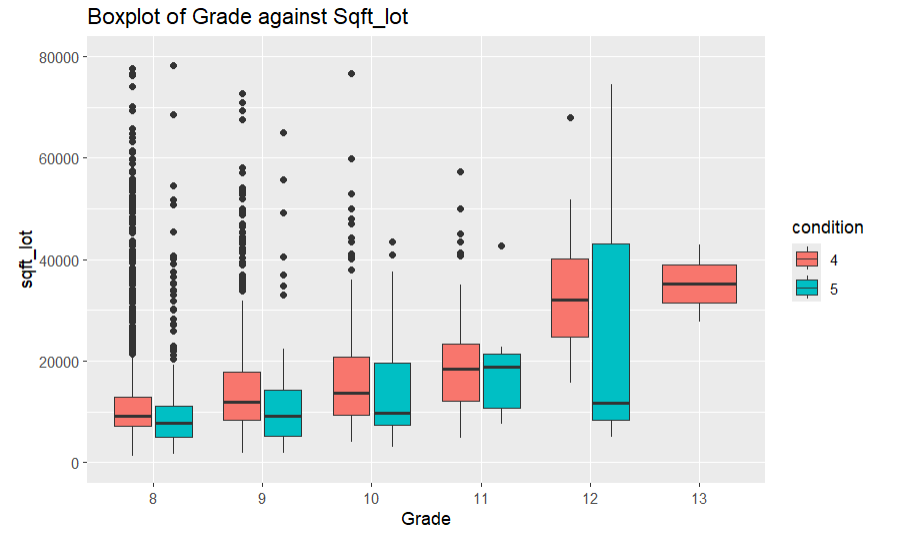
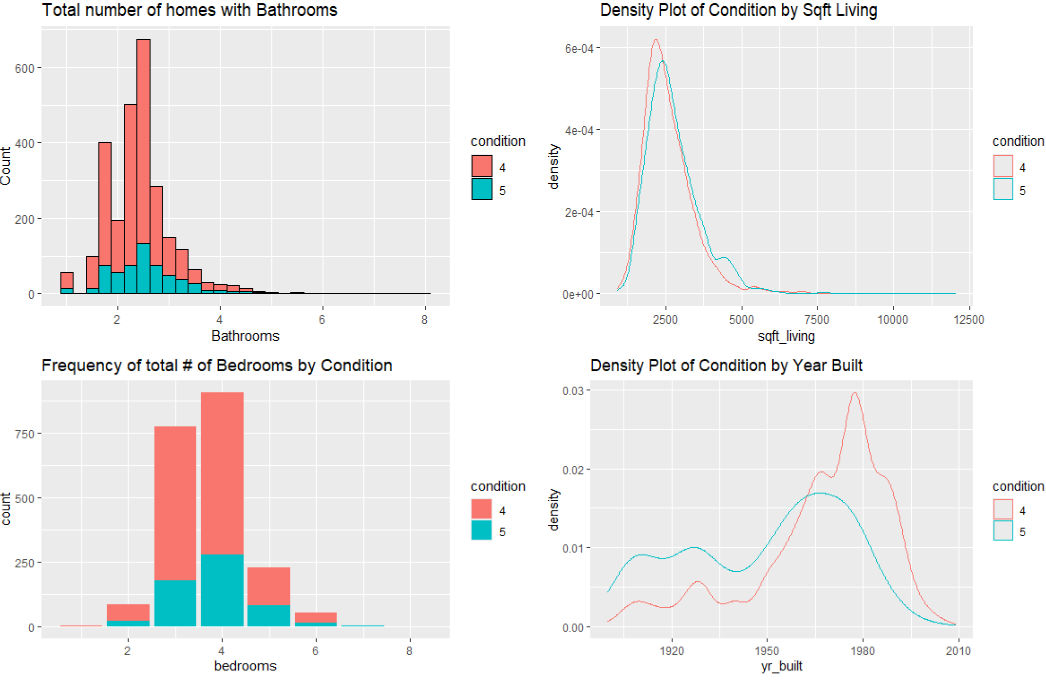


Figure 9

Due to a lack of variables that were characteristics of determining the condition of a home, we created four different visuals based on what would most likely affect it. In this case, we created visuals based on the number of *bathrooms, bedrooms, sqft\_living* (square footage of the interior living space), and *yr\_built* because these were features every home had. We attempted to confirm these assumptions by creating the following visual plots.



From left to right, Figure 10, 11, 12, 13

Observing the histogram in Fig. 10, comparing the number of bathrooms based on the home’s condition, we see that most homes that are condition 4 have 2.5 bathrooms, with second most bathrooms being 2.25 bathrooms. For homes that are condition 5, a majority have 2.5 bathrooms, with the second most bathrooms being shared among homes with 1.75, 2.25, and 2.75 bathrooms. Despite these observations, they do not indicate that bathrooms are a feature to determining the condition of an average home because of how the data is distributed (i.e. homes are considered condition 5 with 1 bathroom and 3+ bathrooms).

In Figure 11, we created a density plot that compares the sqft\_living based on condition. Visually, we see that the density plot looks almost identical to each other. We can interpret this to state that square footage of the interior living space is not a characteristic that affects the condition score of an average home and can proceed to the final two observations.

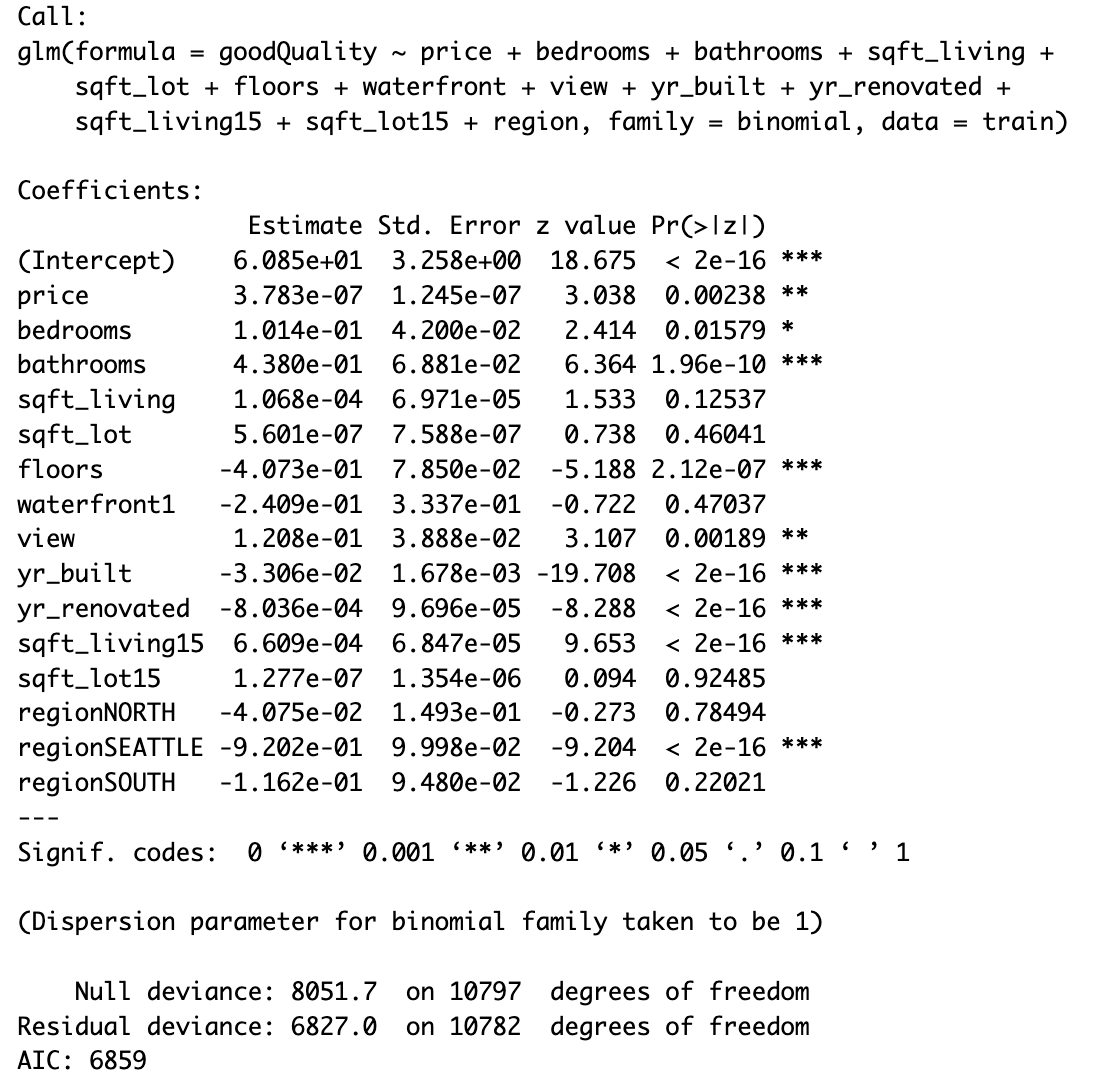
Figure 12 visualizes the frequency of bedrooms across homes that are either condition 4 or 5. We see that a majority of homes, despite their condition score, have 4 bedrooms while 3 comes second most. Like our observation in Figure 9, despite bedrooms being a feature in homes, we cannot state that they affect the condition of a home which invalidates our initial assumption.

Lastly, Figure 13 shows a density plot between the year a home was built and organizes the curves by condition. From initial observations, there is a higher proportion of homes that are condition 5 from the 1900s to 1960s. This then changes to a higher proportion of homes that are condition 4 from the 1960s to 2010. This change in pattern could be due to either using cheaper alternatives during construction or the demand for housing.

In conclusion, our observations led us to believe that *price* and *sqft\_lot* affect the grade of a home. For *condition*, the most significant variables were *price* and *yr\_built*. But *sqft\_lot, sqft\_living, bathrooms,* and *bedrooms,* were insignificant.

**Section 7: Logistic Model for Predicting Good Quality Homes**

To build a model to predict whether a home is considered good quality, we first began by with a logistic regression model using all the variables in the cleaned data set, as described in section 3, except the ID, grade, and condition. Grade and condition were removed as predictor variables because the response variable is based on the values for grade and condition. The 13 predictors first considered were: price, bedrooms, bathrooms, sqft\_living, sqft\_lot, floors, waterfront, view, yr\_built, yr\_renovated, sqft\_living15, sqft\_lot15, and region. The estimated coefficients are shown in the output below:



The output above also includes the results of the Wald test, which assesses whether a single term can be dropped. We then considered a model with only the significant predictor variables based on the results of the Wald test: price, bedrooms, bathrooms, floors, view, yr\_built, yr\_renovated, sqft\_living15, and region.

A screenshot of a computer

Description automatically generated

We then performed a likelihood ratio test to determine if the data support dropping the predictors sqft\_living, sqft\_lot, waterfront, and sqft\_lot15. The null hypothesis for the likelihood ratio test states that all coefficients of the dropped predictors are equal to 0, and the alternative hypothesis states that at least one of the dropped predictors has a coefficient not equal to 0.

The test statistic is the difference in the deviance of the reduced and full models, which we calculated to be 4.764. The associated p-value, found using the test statistic and the chi-squared distribution with four degrees of freedom, is 0.312. Since the p-value is not less than alpha = 0.05, we fail to reject the null hypothesis. Therefore, the data support the reduced model over the 13 predictor model. The estimated logistic regression equation is:

60.75 – 4.42e-07(price) + 0.1196(bedrooms) – 0.4882(bathrooms)

–0.4016(floors) + 0.1138(view) – 3.305e-02(yr\_built)

– 8.123e-04(yr\_renovated) + 7.117e-04(sqft\_living15) – 2.359e-02(I1)

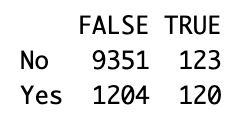
– .9227(I2) – 9.586e-02(I3)

I1 is 1 when the region is North region and 0 otherwise; I2 is 1 when Seattle is the region and 0 otherwise; I3 is 1 when South is the region and 0 otherwise.

The relatively large coefficients for I2 suggest the region variable we created is a helpful in determining if a home is considered good quality. The coefficient for I2 indicates that the log odds of a home being good quality is 0.9227 lower for homes in Seattle than in the East region when controlling for the other variables.

The coefficient of 4.420e-07 for price indicates that the estimated odds that a home is good quality is multiplied by ~1.0000 for each dollar increase in price when controlling for other variables. Thus, the selling price of a home does not appear to have impact on the odds that a home is good quality, so good quality homes are represented in the dataset at a range of prices. The coefficients for yr\_renovated and sqft\_living15 are also small and therefore indicate they have a smaller impact on the odds than some of the other predictor variables, such as region or bathrooms.

Before assessing how well the model does at classifying test data, we first checked if the goodQuality variable is balanced. We found the cleaned dataset had only 2,652 homes considered good quality, and 18,944 homes not considered good quality. The data set therefore appears unbalanced, as it has many fewer good quality homes than not good quality homes. The confusion matrix at threshold 0.5 confirms the imbalance. While accuracy is high (0.877) and error rate is low (0.123), the false negative rate is 0.909.



We decided to lower the threshold to try to lower the false negative rate. The confusion matrix below uses a threshold of 0.15.

A close-up of a number

Description automatically generated

The 0.2 threshold decreases the false negative rate, as shown in the following calculations:

* False negative rate (FNR) =
* False positive rate (FPR) =
* True negative rate (TNR) =
* True positive rate (TPR) =
* Error rate =
* Accuracy =

We used an ROC curve to further assess our logistic regression model. As seen in the plot below, the ROC curve is above the red diagonal line representing a model that classifies at random, which indicates the model performs better than random guessing. The curve is fairly far from the diagonal and close to the top left of the plot, which indicates it does relatively well in classifying observations correctly. The AUC for the model is 0.7966245, which is close to 1 and therefore confirms the model classifies observations well.

A graph with a red line

Description automatically generated

Overall, despite the imbalanced number of good quality homes and not good quality homes in the data set, the model we developed performs better than random guessing at determining if a home is good quality or not.

**Reference**

1 https://gis-kingcounty.opendata.arcgis.com/datasets/kingcounty::health-reporting-areas-2020-health-reporting-2020-area/explore

2 https://gis-kingcounty.opendata.arcgis.com/datasets/kingcounty::zipcodes-for-king-county-and-surrounding-area-shorelines-zipcode-shore-area/explore?location=47.447291%2C-121.477600%2C7.49