An Analysis of House Sales in King County, USA

Group 2

Levi Davis, George Jagtap, Jessica Kimbrell, Jake Weinberg

Table of Contents:

Section 1: Executive Summary

Our project analyzed data on house sales and focused on two main questions: 1) What features of a house and its environment will best predict the house's price?; and, 2) Can we predict the construction and design quality (grade) of a house based on its physical characteristics? To conduct our analysis we used a dataset of 21,613 houses in King County, Washington. The dataset contains the selling price of the home as well as other information about each house like square footage, number of bedrooms, and year built.

Recently, the housing market has experienced sizable fluctuations, which can make it difficult for home buyers to understand house prices. What determines the price of a house? Factors like size, quality, and location come to mind first, but in reality dozens of factors have varying degrees of impact on house price. We wanted to investigate if a small subset of factors can accurately predict house prices. This will give buyers insight into whether a house price is reasonable.

To uncover which factors would best predict the price of the house, we created several different iterations of a model using both automated and deductive methodologies. In the end, we found that the most important variables in predicting a house were the square footage of the living space, whether the house is a waterfront property, the view from the house, the grade of its construction quality, and the year the house was built. We found it interesting that three of these five characteristics did not actually represent physical characteristics of the house. Based on the view and waterfront status being influential, our analysis supports the notion that location is a key factor in determining the price of a house.

House construction and design quality is one of the most important factors when buying a home. Over the years, houses with exceptional construction and design quality could save owners from stress, not to mention thousands of dollars in repairs. However, assessing the quality as a potential buyer can be tricky, as a seller will show off the best parts while hiding the flaws. Being able to predict which houses have above average construction and design quality could save buyers money, time, and peace of mind.

Our analysis revealed that we can moderately predict if a house's construction and design quality is above average or not. To make the model we used lot square footage, living space square footage, above ground square footage, number of bedrooms, bathrooms, and floors as predictors . Using just these 6 house descriptors our model was able to categorize houses as above average or not with an accuracy of 81%.

Section 2: Description of the Data and the Variables

The dataset we used contained information gathered from King County, Washington, about house selling price, location and physical properties. Table 2.1 lists the variables contained in the dataset and a short description of each. Latitude and longitude were excluded from our analysis due to their unique data type. We also created two new variables to aid in our analysis, latest_construct and grade.cat. Both variables are explained in detail in the table below.

Variables	type	values	Description
price	Quantitative	(75000, 7700000)	Price of each home sold.
bedrooms	Quantitative	(0, 33)	Number of bedrooms.
bathrooms	Quantitative	(0, 8)	Number of bathrooms.
sqft_living	Quantitative	(290, 13540) sq ft	Square footage of the apartment's interior living space.
sqft_lot	Quantitative	(520, 1651359) sq ft	Square footage of the land space.
floors	Quantitative	(1, 3.5)	Number of floors
waterfront	Binary	0,1	A dummy variable for whether the apartment was overlooking the waterfront or not.
view	Categorical	Index from 0 to 4	An index from 0 to 4 of how good the view of the property was.
condition	Categorical	Index 1 to 5	An index from 1 to 5 on the condition of the apartment
grade	Categorical	Index from 1 to 13	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.
sqft_above	Quantitative	(290, 9410) sq ft	The square footage of the interior housing space that is above ground level.
sqft_basement	Quantitative	(0, 4820) sq ft	The square footage of the interior housing space that is below ground level.
yr_built	Quantitative	(1900, 2015)	The year the house was initially built.
yr_renovated	Quantitative	(0, 2015)	The year of the house's last renovation. (Never renovated coded as 0).
zipcode	Categorical	(98001, 98199)	What zip code area the house is in

sqft_living15	Quantitative	(399, 6210)	The square footage of interior housing living space for the nearest 15 neighbors.
sqft_lot15	Quantitative	(651, 871200)	The square footage of the land lots of the nearest 15 neighbors
latest_construct	Quantitative	(1900, 2015)	Renovation year for houses that have been renovated, year built for houses that haven't.
grade.cat	Binary	0,1	A transformed variable created from grade. 0 for average or below, grade levels 1-7, and 1 for above average, levels 8-13.

Table 2.1

The first of two questions we are seeking to answer with this analysis is what features of a house and its environment will best predict its price? The housing market has been booming and is in the news daily, making this question relevant. Prices are rising and houses are sellings within days instead of months. These factors lead us to question what attributes of a house contribute to the price and how much influence each attribute has. This will allow homeowners to more accurately evaluate their home's worth and make improvements to the more valuable attributes.

Our second question investigates if house construction and design quality (grade) can be predicted by physical house descriptors such as square footage or number of bedrooms. Are homes with a greater square footage more likely to have an above average grade? One might argue bigger houses use more supplies, take longer to build, and cost more, which could incentivise owners to build something sturdy and worth the investment. However, another could argue that these factors could cause the builders to focus on turning a quick profit instead of quality. The construction quality of a house is often the important factor homebuyers consider when looking for a house. Poor design can lead to tens of thousands of dollars in repairs that must be completed to ensure the safety of the occupants. Being able to accurately predict construction quality would give homeowners a greater peace of mind when purchasing their largest asset.

Before starting the model building process to investigate our questions, we did some exploratory data analysis to get an idea of the relationships between predictors and response variables. We started by investigating the price variable because it has interesting real-world applications. Figure 2.1 shows the histogram of price data. It's not surprising to see that there is a severe right skew to the histogram because most home prices fall within a certain range but there are a few outliers. The graph also shows that the vast majority of houses have prices under $1,000,000. The figure gives us a very good understanding of the range of prices in the dataset.

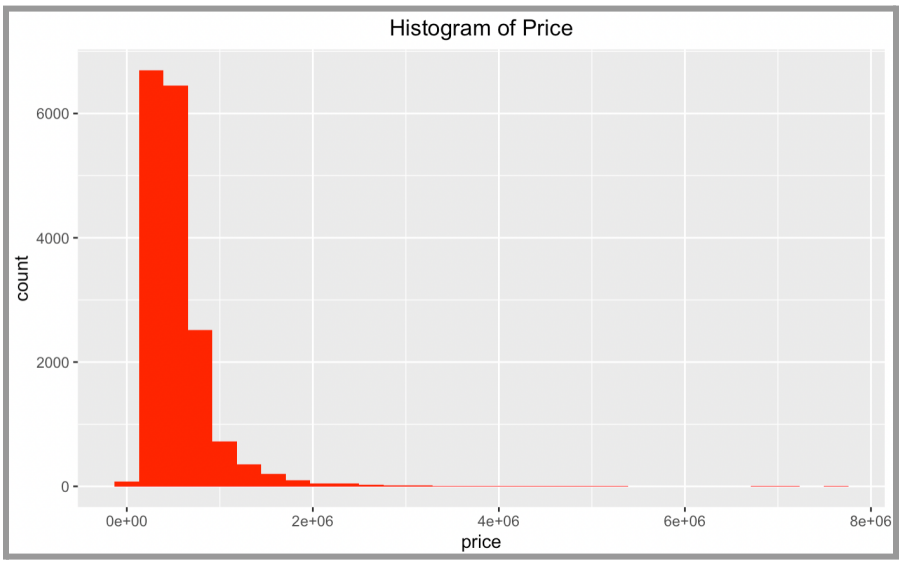

Figure 2.1

Next, we wanted to explore the relationship between year built and square footage of living space. Our hypothesis was that living space would on average increase over time, and that is indeed what is shown in Figure 2.2. The trend is positive and seems linear, although very gradual. This is intuitive, as people have moved into the suburbs where there is more space, and people have had higher incomes over time. This helped us understand more about our potential variables before we dove into the heavy analysis.

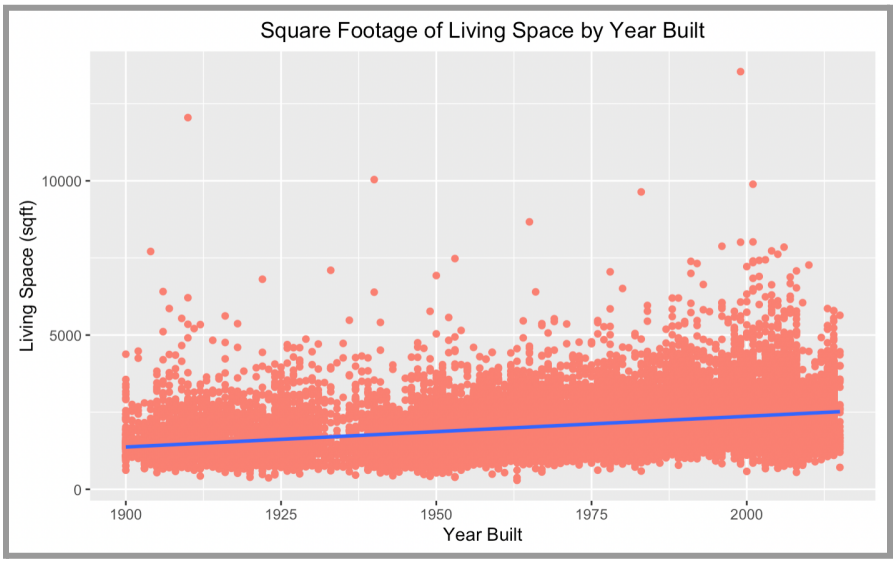

Figure 2.2

Next we explored how certain physical characteristic predictors relate to our grade.cat variable. Figure 2.3 displays a scatter plot of lot square footage against living space square footage by grade. The graph shows that there is an approximate horizontal line that splits the two grades at 2,500 sqft of living space. If the house has more than 2,500 sqft of living space it looks very likely that the house was constructed with an above average design. Square footage of the lot does not seem to have much of an influence on the grade category. This figure will allow us to intuitively check our final model and ensure it matches up with our findings from this graph.

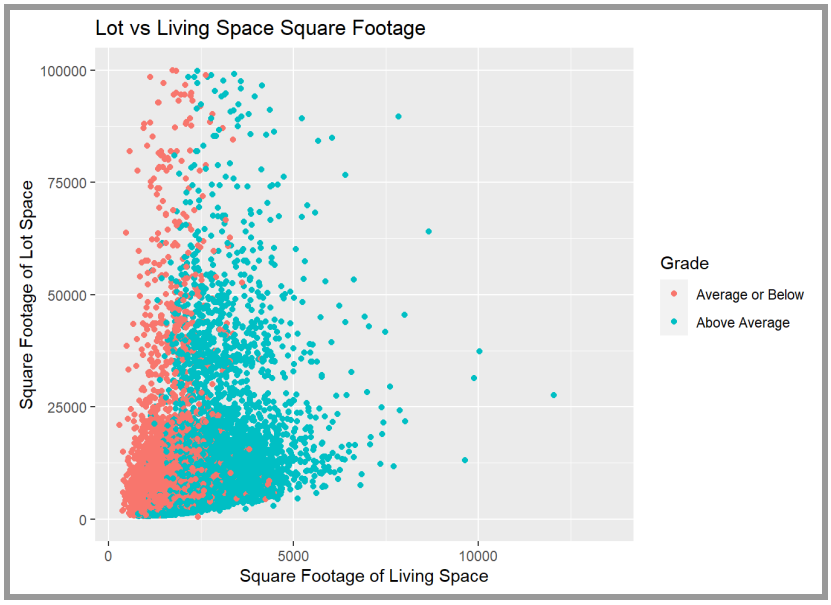

Figure 2.3

The second relationship for predicting grade we explored was the number of floors in the house by grade as shown in Figure 2.4. The graph reveals that if the house has 2 or more floors it is much more likely that the house has an above average grade. It is interesting that after 2 floors the proportion of average or below vs above average is about the same and does not continue increasing as the number of floors increase. Floor will have a positive coefficient in our model because it increases the odds of a house having above average design.

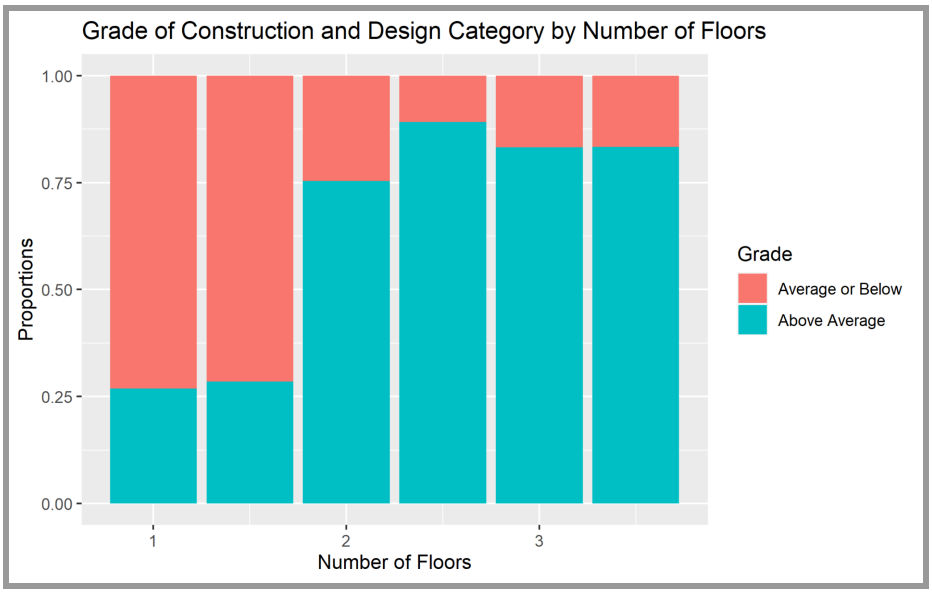

Figure 2.4

The last relationship we decided to investigate was the number of bathrooms and its association with grade. Figure 2.5 displays a density plot of bathrooms and grade which shows the most frequent number of bathrooms for each grade. One bathroom is the most prevalent among houses with an average or below grade and 2.5 bathrooms is very common among houses with an above average grade. Due to these findings it seems likely that as the number of bathrooms increases there is a greater probability the grade will be above average.

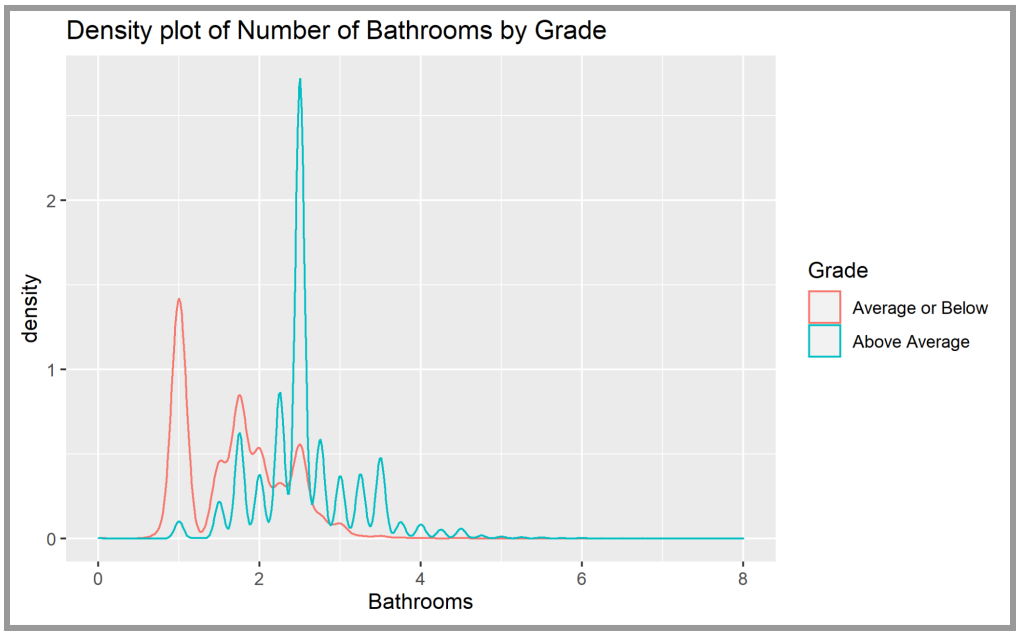

Figure 2.5

Section 3: Linear Regression

The question that we used multi-linear regression to answer was: what features of a house and its environment will best predict the house's price? To figure out which set of predictors will best answer this question, we took two modeling approaches: first, constructing a regression from intuition and second by creating a regression using automatic algorithms. Then, we will compare the models to determine which had the best fit of the data.

To set up our analysis, we split our data into training and testing by an 80/20 split. We decided to exclude the variables yr_renovated, longitude, latitude, and zip code from our data frames, as we didn't think they would offer anything helpful to our analysis.

We performed exploratory data analysis on the variables in the data set to determine which variables we wanted to include in our first pass model. To examine the quantitative predictors, our first step was to plot each of them against price. Figure 3.1.1 shows the scatter plots. From this figure, we see that many of the predictors appear to have a relationship with price; at first glance, some strong predictors appear to be bathrooms, sqft_living, sqft_living_15, sqft_above, and grade. A lot of these make intuitive sense to include (as well as some of the other predictors), but to avoid potential overfitting, we also wanted to examine the interactivity between the variables.

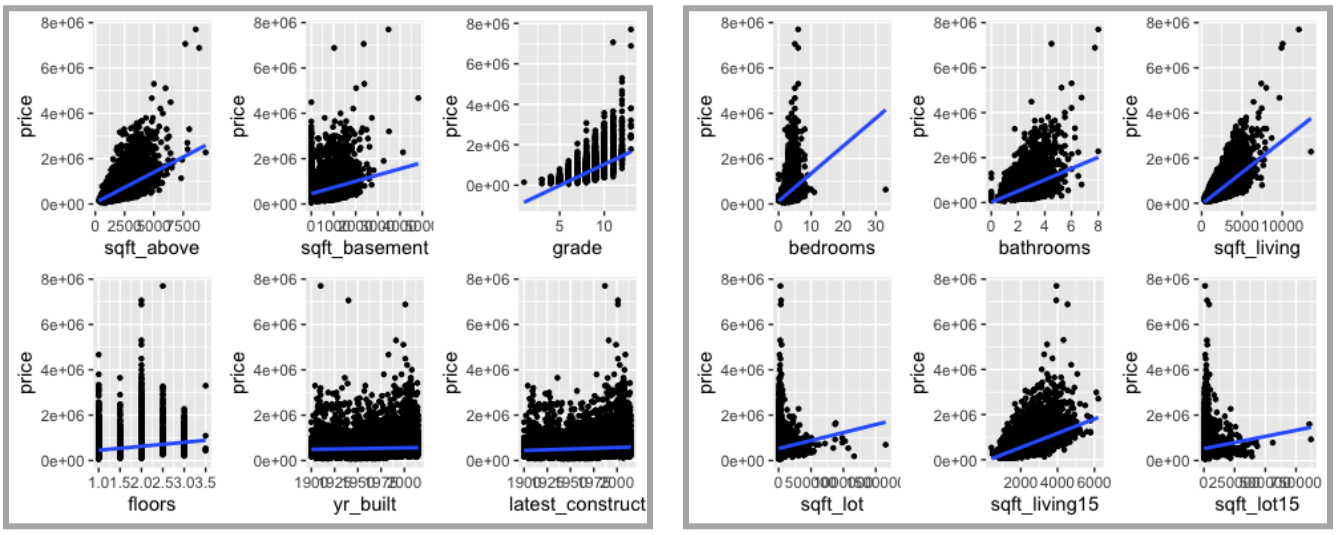

Figure 3.1.1

Figure 3.1.2 shows both the correlation coefficient between the quantitative predictors and a scatter plot for each combination of predictors. Before diving into the numbers, we intuitively knew that many of the variables were proxies for the size of the house (bedrooms, bathrooms, floors, all of the sqft variables), so we expected those predictors to be highly correlated. Additionally, we thought that a house's living and lot sizes would be correlated with the neighborhood living and lot sizes. As predicted, we saw that many, although not all, of the size-related variables had high correlation with each other. Going forward, we will be selective in choosing size-related variables so as not to overfit the model. We also saw some relatively unexpected correlations, such as grade with sqft_living and sqft_above.

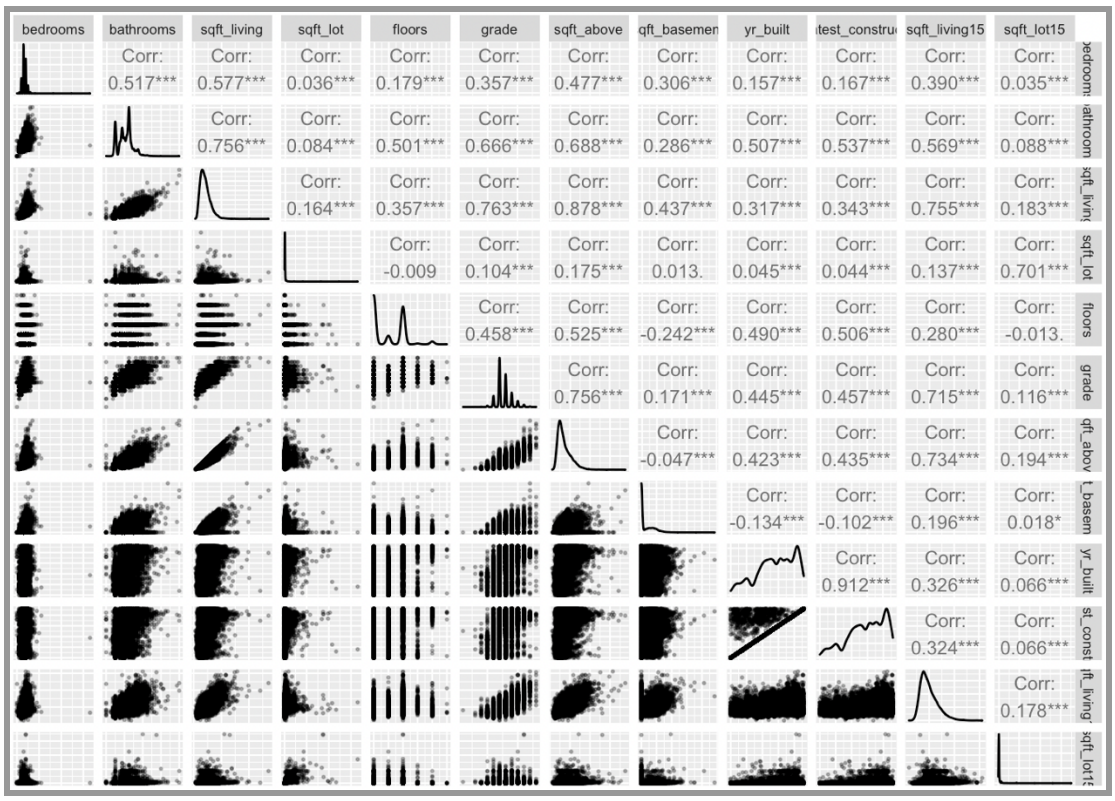

Figure 3.1.2

For the categorical variables, we created boxplots to compare the mean and variance of price within the different classes for each variable. Figure 3.1.3 shows the boxplot for the waterfront variable. We expected a strong effect from waterfront, as we hypothesized that location-related variables would be very important in determining price. There seems to be a relatively clear difference between mean price for waterfront and non-waterfront properties, so we will want to include waterfront in our model.

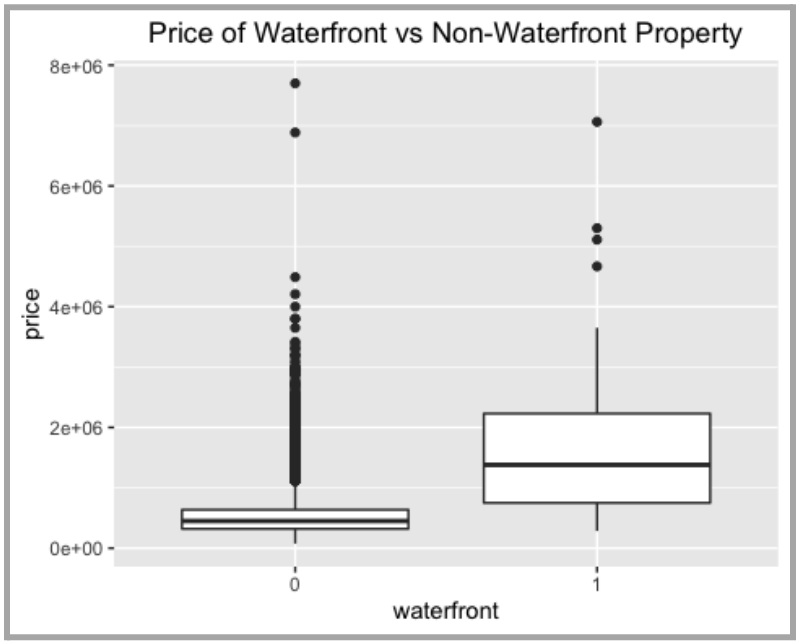

Figure 3.1.3

Figure 3.1.4 shows the property price plotted against the quality of the view at the property. We also thought the view would have an impact on price, because it also is a location-adjacent variable. From the box plot we can see that there is a difference between the five classes, although it doesn't appear to be as clear as waterfront. At the very least, we can see that category 4 seems to have a higher mean than the rest of the categories, demonstrating that view may also be useful in our model.

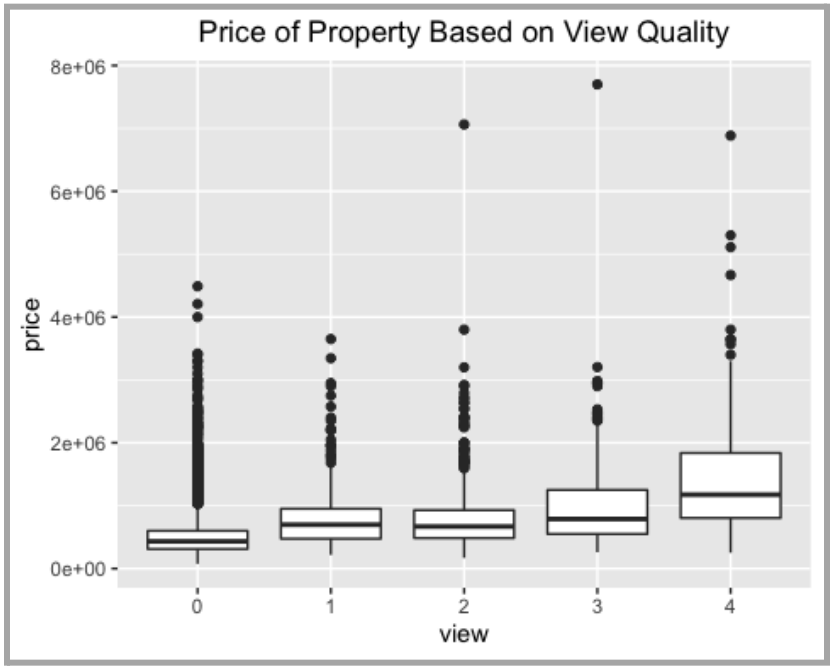

Figure 3.1.4

Figure 3.1.5 plots the property price against the condition variable. There does not appear to be too much of a difference between the buckets of the condition variable. We will leave the condition predictor out of our intuitive model.

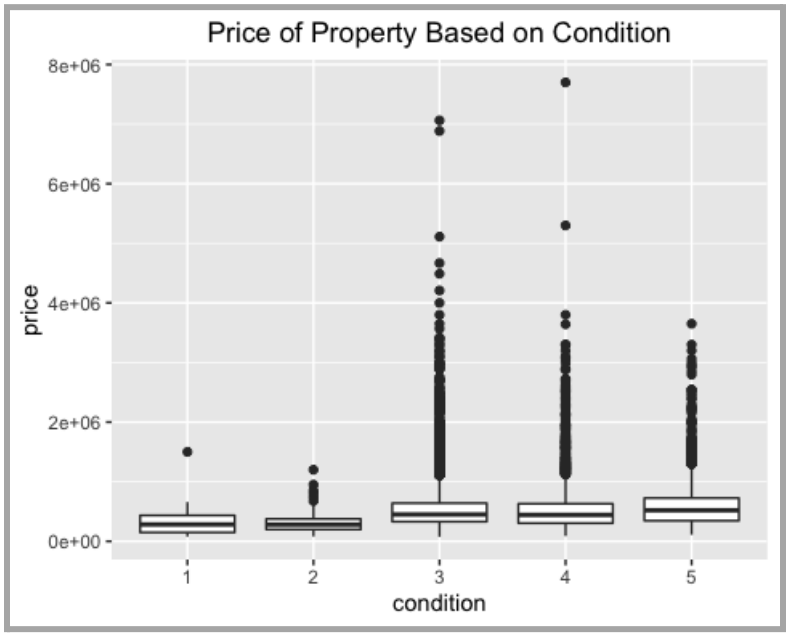

Figure 3.1.5

Figure 3.1.6 is a bar chart that plots the average price of a property in a zip code for each zip code in the data. We can see that there are clear differences in price by zip code, which makes intuitive sense. However, zipcode is a categorical variable with many classes, with no intuitive way to create fewer buckets. Therefore, it will be very complicated to include in the model. Additionally, zipcode likely also has high correlation with other variables in the model. Based on our knowledge, the differences in price for zipcode would likely be partially explained with view, waterfront, and the neighborhood variables. With these reasons in mind, we decided to drop the zipcode from our analysis.

Figure 3.1.6

In our exploratory data analysis, we noticed that the distributions for two of our predictor variables (sqft_living and sqft_living15) were right skewed. Applying a log transformation on these predictors changed their distribution to be approximately normal. The results of this EDA are shown in Figure 3.1.7. From this, we hypothesize that our transformed predictors will help fit a more accurate regression and reduce the high number of influential points that we anticipate.

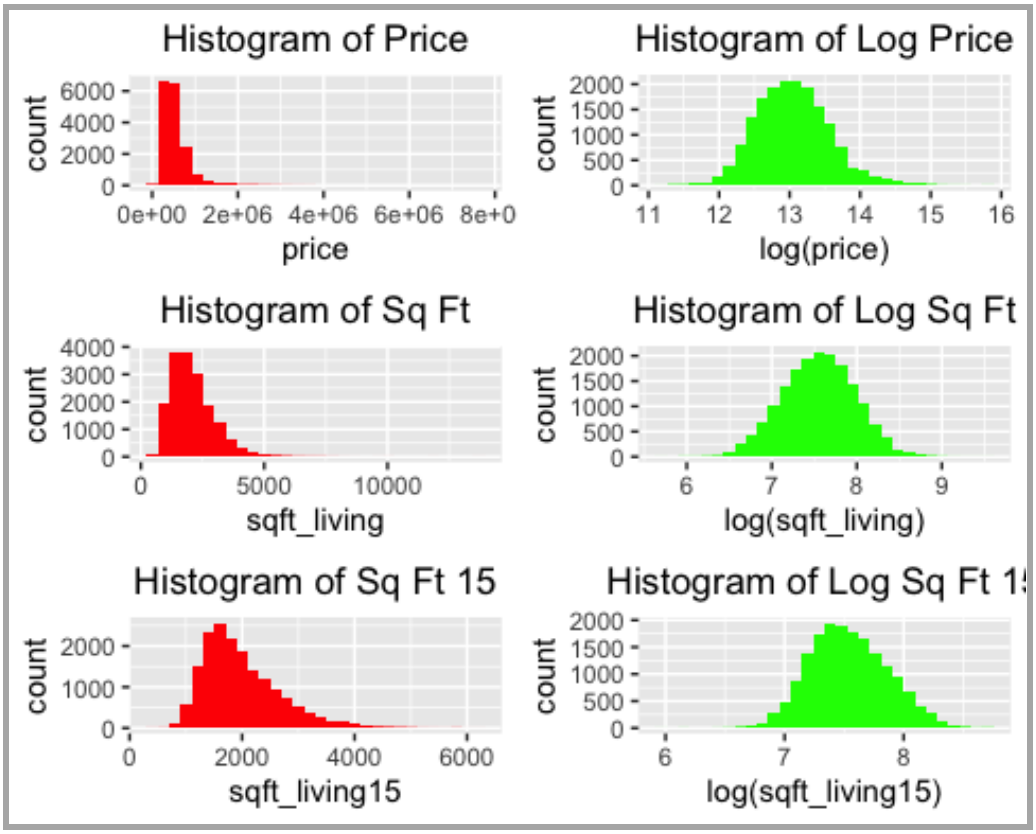

Figure 3.1.7

We fit our first pass at our MLR model, including the variables of log.sqft_living, grade, view, and waterfront. These were the variables that had the most obvious associations with price from our knowledge and Exploratory Data Analysis.

We ran an F-test on this initial model, with the null hypothesis that all coefficients for predictors would be equal to 0. The alternative hypothesis is that at least one of these coefficients would not be equal to 0. Our F-statistic, shown in Table 3.1.1, that we generated was 3554, with a p-value close to 0. This means that we can reject our null hypothesis and that we have a useful model. With our baseline established, we proceeded to test other variables to see if they would improve our model.

```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2057913      39053  -52.70   <2e-16 ***
## log.sqft_living   202002       6688   30.20   <2e-16 ***
## grade             136994       2425   56.49   <2e-16 ***
## view1             204746      15441   13.26   <2e-16 ***
## view2             132171       9264   14.27   <2e-16 ***
## view3             212049      12787   16.58   <2e-16 ***
## view4             412989      19750   20.91   <2e-16 ***
## waterfront1       549537      27573   19.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 249000 on 17282 degrees of freedom
## Multiple R-squared:  0.5415, Adjusted R-squared:  0.5413
## F-statistic:  2916 on 7 and 17282 DF,  p-value: < 2.2e-16
```

Table 3.1.1: Regression Output 1

First, shown in Table 3.12, we tried adding bathrooms, because it appeared to correlate with price. We left it out of our initial model because logically it represents the same concept as sqft_living (size of the house), and they had a fairly high correlation coefficient (0.756). The result was that bathrooms did appear as significant, but the adjusted R-squared of our model only increased by 0.0006. Therefore, following the concept that simpler models are more effective and that conceptually the bathroom variable was redundant, we decided to leave it out of our model.

```
##
## Call:
## lm(formula = price ~ log.sqft_living + grade + view + waterfront +
##     bathrooms, data = brain.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1429034  -135987   -26918    98665  5807810
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1920626      48125 -39.909  < 2e-16 ***
## log.sqft_living   181292       7919  22.893  < 2e-16 ***
## grade             134222       2489  53.920  < 2e-16 ***
## view1             205502      15432  13.317  < 2e-16 ***
## view2             133681       9263  14.432  < 2e-16 ***
## view3             212425      12779  16.624  < 2e-16 ***
## view4             414038      19738  20.977  < 2e-16 ***
## waterfront1       549764      27555  19.951  < 2e-16 ***
## bathrooms          18993       3896   4.876 1.09e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 248900 on 17281 degrees of freedom
## Multiple R-squared:  0.5421, Adjusted R-squared:  0.5419
## F-statistic:  2557 on 8 and 17281 DF,  p-value: < 2.2e-16
```

Table 3.1.2: Regression Output 2

Next we tried adding log.sqft_living15, because that variable also seemed correlated with price during EDA. When we added it, though, we found that it was significant but only improved adjusted R-squared by 0.0004, so we did not include it in the model. This is shown in Table 3.1.3.

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2208518      53019 -41.655  < 2e-16 ***
## log.sqft_living     186930       7588  24.634  < 2e-16 ***
## grade               133867       2536  52.791  < 2e-16 ***
## view1               201420      15454  13.033  < 2e-16 ***
## view2               129596       9279  13.966  < 2e-16 ***
## view3               208937      12802  16.321  < 2e-16 ***
## view4               409694      19756  20.738  < 2e-16 ***
## waterfront1         550624      27561  19.978  < 2e-16 ***
## log.sqft_living15    38282       9120   4.198 2.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 248900 on 17281 degrees of freedom
## Multiple R-squared:  0.542,  Adjusted R-squared:  0.5417
## F-statistic:  2556 on 8 and 17281 DF,  p-value: < 2.2e-16
```

Table 3.1.3: Regression Output 3

Last, we tested yr_built. We tried this because, even though yr_built did not appear to have too strong of a relationship with price during EDA, it logically would make sense to include. We found that yr_built was significant and that it improved our adjusted R-squared by ~0.05, so we included it. We are not concerned with the fact that the coefficient is the opposite sign as expected. MLR coefficients measure the effect of a predictor after accounting for all the other predictors in the model, so given levels of all the other predictors, it is reasonable to assume there is a real estate-related reason for the age effect described by our model. The regression is shown in Table 3.14.

```
##
## Call:
## lm(formula = price ~ log.sqft_living + grade + view + waterfront +
##     yr_built, data = brain.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1396862  -120791   -12980    91187  5485637
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4281211.04  132558.47  32.297   <2e-16 ***
## log.sqft_living  217869.01    6263.53  34.784   <2e-16 ***
## grade            174539.90    2390.42  73.016   <2e-16 ***
## view1            150978.03   14482.61  10.425   <2e-16 ***
## view2             75763.06    8738.01   8.671   <2e-16 ***
## view3            148433.03   12027.60  12.341   <2e-16 ***
## view4            330877.86   18545.46  17.841   <2e-16 ***
## waterfront1      550194.82   25789.23  21.334   <2e-16 ***
## yr_built          -3419.67      68.74 -49.747   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 232900 on 17281 degrees of freedom
## Multiple R-squared:  0.5989, Adjusted R-squared:  0.5987
## F-statistic:  3226 on 8 and 17281 DF,  p-value: < 2.2e-16
```

Table 3.1.4: Regression Output 4

After we tested the other variables that could make sense to include in the model, the last step was to check for interaction terms. Intuitively, we thought that there could be interaction effects between sqft_living and waterfront as well as sqft_living and view. However, we thought that these interaction effects would be similar because the best views were likely waterfront properties. We elected to test the waterfall predictor first, as we thought this showed the relationship more clearly. Figure 3.1.8 shows the clear interaction effect. We found this interaction term to be significant, as shown in Table 3.1.5, so we included it in our model.

```
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.301e+06  1.297e+05  33.174   <2e-16 ***
## log.sqft_living                2.102e+05  6.133e+03  34.270   <2e-16 ***
## waterfront1                   -8.057e+06  3.085e+05 -26.116   <2e-16 ***
## view1                          1.561e+05  1.417e+04  11.017   <2e-16 ***
## view2                          8.139e+04  8.549e+03   9.521   <2e-16 ***
## view3                          1.656e+05  1.178e+04  14.061   <2e-16 ***
## view4                          3.032e+05  1.817e+04  16.688   <2e-16 ***
## grade                          1.717e+05  2.340e+03  73.371   <2e-16 ***
## yr_built                      -3.390e+03  6.724e+01 -50.407   <2e-16 ***
## log.sqft_living:waterfront1    1.088e+06  3.886e+04  27.993   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 227800 on 17280 degrees of freedom
## Multiple R-squared:  0.6163, Adjusted R-squared:  0.6161
## F-statistic:  3084 on 9 and 17280 DF,  p-value: < 2.2e-16
```

Table 3.1.5: Regression Output 5

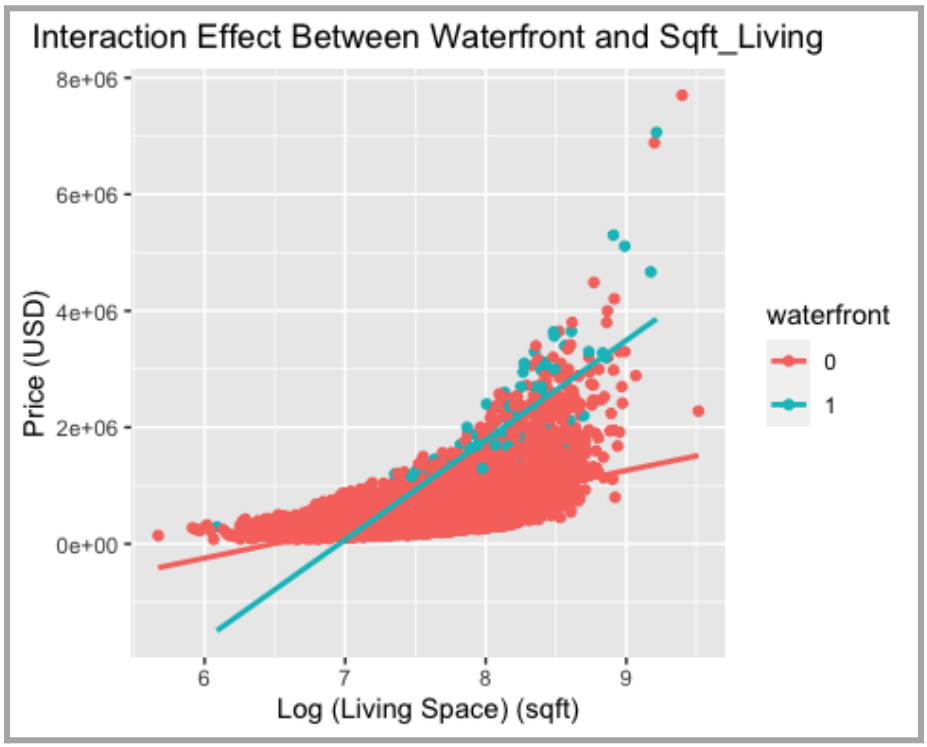

Figure 3.1.8

We created a residual plot (Figure 3.1.9) to test the first two linear regression assumptions. Assumption 1 states that for every level of the predictor, the residuals have a mean variance of zero. Assumption 2 states that for every level of the predictor, the residuals have a constant variance. Both assumptions are violated in the residual plot, assumption 2 more obviously so than assumption 1. Therefore, we will need to transform our y-variable first and reassess.

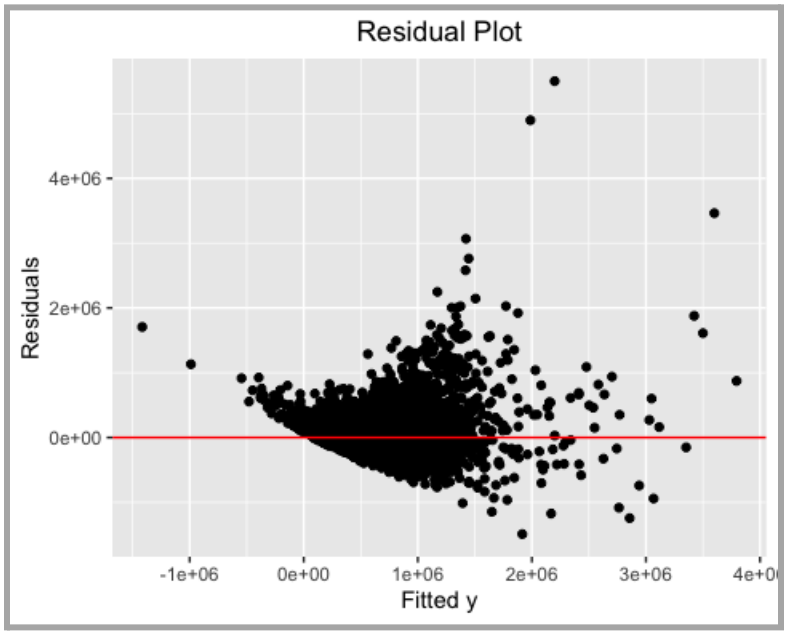

Figure 3.1.9

To determine how to transform our y-variable, we created a Box-Cox plot, shown in Figure 3.1.10. This showed a very small confidence interval for lambda in between 0 and 0.05. For simplicity and due to proximity to this value, we decided to first try transforming the y-variable price by taking log(price).

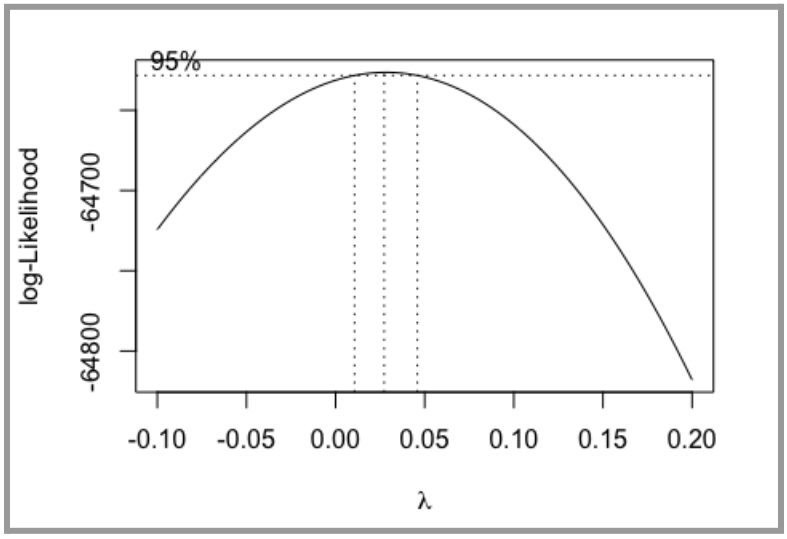

Figure 3.1.10

After transforming the price, we wanted to recheck our regression. In Table 3.1.6, we see that the interaction term is still significant in predicting log(price), but with a p-value of only 0.04. Figure 3.1.11 shows how the difference in the slopes between the two waterfront classes eroded with the y-transformation. Coupled with the extremely high VIF numbers seen in table 3.1.7, we decided to drop the interaction term from the model. Table 3.1.6 shows that we have the exact same adjusted R-squared with and without the interaction term.

```
##
## Call:
## lm(formula = price.star ~ log.sqft_living * waterfront + view +
##     grade + yr_built, data = brain.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28822 -0.21697  0.01633  0.21487  1.29548
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.765e+01  1.819e-01  97.006   <2e-16
## log.sqft_living              4.176e-01  8.604e-03  48.538   <2e-16
## waterfront1                 -5.261e-01  4.328e-01  -1.215   0.2242
## view1                        1.926e-01  1.988e-02   9.689   <2e-16
## view2                        1.097e-01  1.199e-02   9.142   <2e-16
## view3                        1.508e-01  1.653e-02   9.121   <2e-16
## view4                        2.584e-01  2.549e-02  10.136   <2e-16
## grade                        2.445e-01  3.284e-03  74.452   <2e-16
## yr_built                    -4.892e-03  9.435e-05 -51.851   <2e-16
## log.sqft_living:waterfront1  1.117e-01  5.453e-02   2.048   0.0406
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3196 on 17280 degrees of freedom
## Multiple R-squared:  0.6319, Adjusted R-squared:  0.6317
## F-statistic:  3296 on 9 and 17280 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = price.star ~ log.sqft_living + waterfront + view +
##     grade + yr_built, data = brain.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32759 -0.21693  0.01632  0.21496  1.29613
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.765e+01  1.819e-01  96.987   <2e-16 ***
## log.sqft_living 4.184e-01  8.597e-03  48.674   <2e-16 ***
## waterfront1    3.572e-01  3.539e-02  10.093   <2e-16 ***
## view1          1.921e-01  1.988e-02   9.663   <2e-16 ***
## view2          1.091e-01  1.199e-02   9.095   <2e-16 ***
## view3          1.490e-01  1.651e-02   9.026   <2e-16 ***
## view4          2.612e-01  2.545e-02  10.262   <2e-16 ***
## grade          2.448e-01  3.281e-03  74.604   <2e-16 ***
## yr_built      -4.895e-03  9.435e-05 -51.885   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3197 on 17281 degrees of freedom
## Multiple R-squared:  0.6318, Adjusted R-squared:  0.6317
## F-statistic:  3707 on 8 and 17281 DF,  p-value: < 2.2e-16
```

Table 3.1.6: Regression Comparison for Interaction Term

log.sqft_living	waterfront1	view1	view2	view3	view4	grade	yr_built	log.sqft_living:waterfront1
2.273	223.959	1.013	1.042	1.053	1.542	2.522	1.299	224.873

Table 3.1.7: VIF Statistics

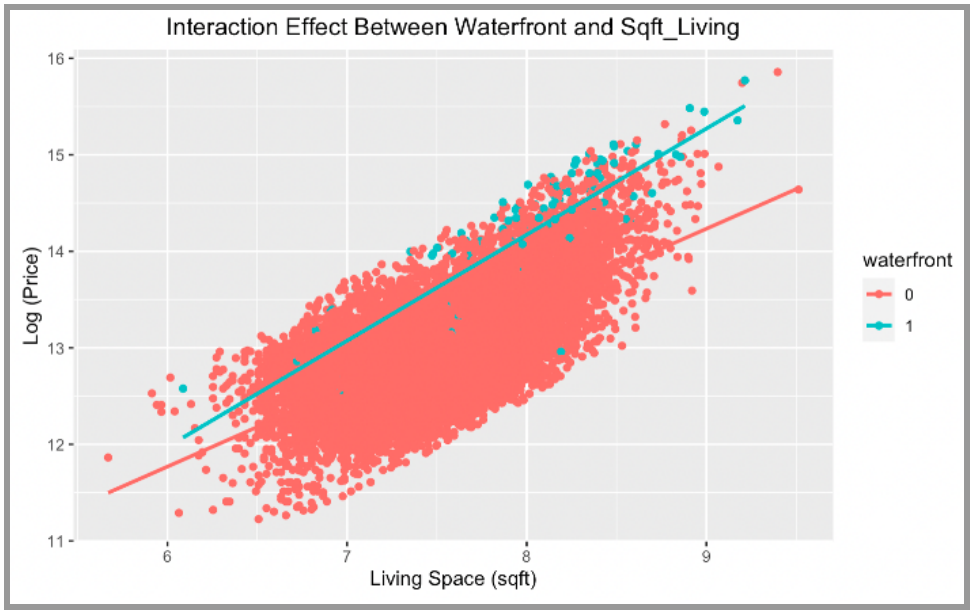

Figure 3.1.11

Now that we transformed price, we are ready to check our regression assumptions again. We started out with Levene's test for equal variance, as we had categorical predictors. Table 3.1.8 shows the result for view and waterfront on the transformed price variable. As shown, we have a significant p-value for both, meaning that the variability of the response variable is different for each class of the categorical variable. This means the tests failed.

```
 Modified robust Brown-Forsythe Levene-type test
 deviations from the median

data:  brain.train$price.star
Test Statistic = 10.569, p-value = 1.496e-08
```

```
 Modified robust Brown-Forsythe Levene-type test
 deviations from the median

data:  brain.train$price.star
Test Statistic = 34.64, p-value = 4.041e-09
```

Table 3.1.8: Levene's Test

Because of the failed Levene's tests, we examined the sample sizes of each class for our categorical variables. The sample sizes were not relatively equal. Then, we examined the variances. For view, the smallest and largest variances of the classes were within 1.5x of each other. For waterfront, the smallest and largest variances of the classes were within 1.85x of each other. While ideally, we would create a regression for each waterfront category, the waterfront variances are not too extremely different, so due to this consideration and project constraints, we cautiously proceeded with one regression.

With our regression finalized, we went to recheck our regression assumptions, starting with creating another residual plot. This time, Assumptions 1 and 2, as described above, appear to be met well. We are comfortable moving on to Assumption 3.

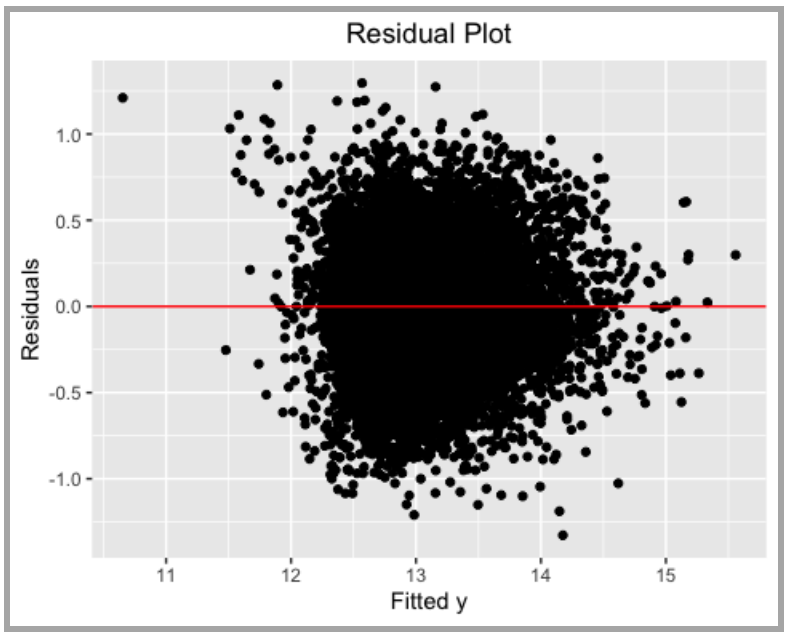

Figure 3.1.12

Assumption 3 of linear regression states that there cannot be any autocorrelation of the residuals. To test this, we plotted the residuals on an ACF plot (Figure 3.1.13). Because all of the lines on the plot were within the confidence bound, we say that there is no autocorrelation of the residuals and that the Assumption 3 is met.

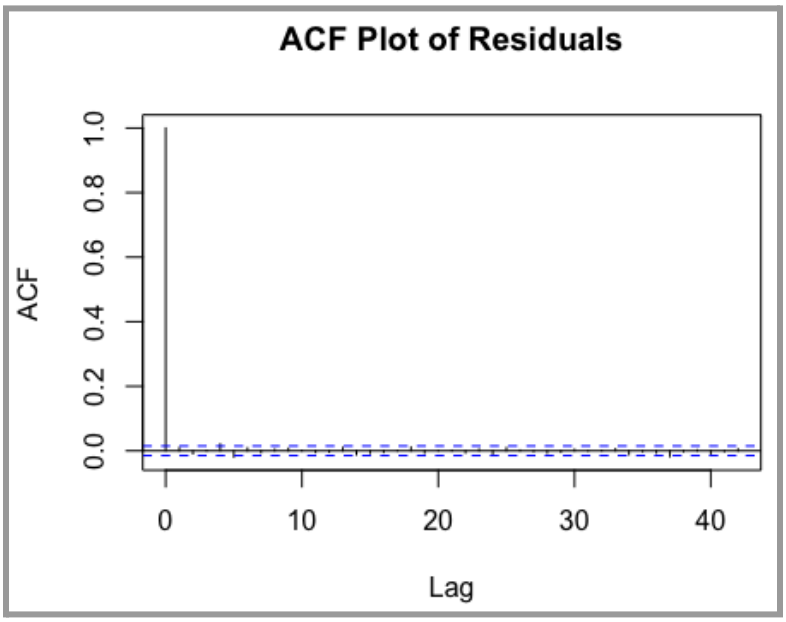

Figure 3.1.13

The final assumption of linear regression states that for each value of the predictor, the error terms follow a normal distribution. Figure 3.1.14 depicts a Normal Q-Q plot that we used to test this

assumption. All of the residuals fall very close to the line representing a normal distribution, so Assumption 4 is also met.

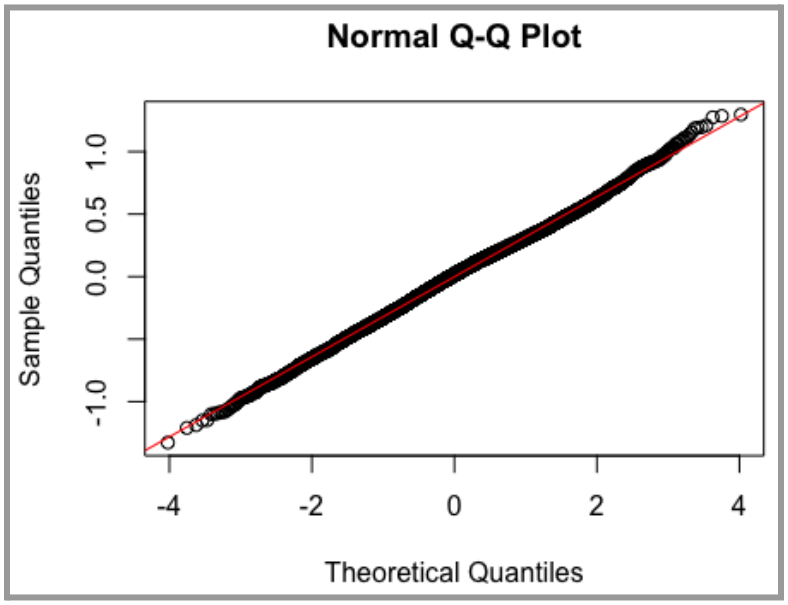

Figure 3.1.14

Before we moved on to testing for influential data points, we wanted to assess our model fit further. To do this, we compared our calculated R-squared prediction value to the R-squared of the model. Our R-squared prediction value was 0.6315 and R-squared was 0.6318. These values are very close to one another, so we are confident that we do not have overfitting in our model.

The first step in checking for influential points is checking for outliers. To do this, we produced a list of externally studentized residuals and compared them to a t(n-1-p) distribution. If any residuals were greater than the critical value, then those points would be considered outliers. However, we did not find any, so we proceeded.

Next, we calculated highly leveraged points, which are particularly influential points in the regression because they are far from the mean of the predictors. We found 1,935 high leverage points for this model, but before we become concerned, we need to check if they are actually influential.

There are two methodologies to test for influential points - Cook's Distance and DFFITS. Because Cook's Distance is the measure that is better for general models with goals of broadly fitting data well and DFFITS is best for assessing models that will be used to predict a specific point with specific parameters, we chose to use Cook's Distance for our test. Using Cook's Distance, none of our data points were deemed influential, so we did not need to address any influential observations.

Finally, to determine the quality of our final intuitive model, we fit our regression onto our test data and calculated the root mean square error of the model. The final intuitive model had an RMSE of 0.3161. We will compare this with the model we develop using the automatic algorithms to determine which is better.

Our second model was built using all possible regressions. After running the automated selection procedure we found that the best model, based on highest R squared, lowest Mallows' CP, and lowest BIC

penalty, had predictors bedrooms, bathrooms, sqft_living, waterfront, view, sqft_above, and sqft_living. t. However, we got a warning stating there are linear dependencies found, and the coefficient for sqft_basement is zero, we can speculate that this variable is highly correlated with sqft_living. When looking at the ANOVA F Test output, we see that sqft_basement does not even appear in the model - therefore it must have an almost perfect degree of multicollinearity with sqft_living. When we drop this variable, the ANOVA F Test output has the same predictors as above.

After doing further research on the variables in this dataset, we hypothesize that if you add sqft_above and sqft_basement, this equals sqft_living, so we want to try dropping the two former variables. From our research and based on the output of the above model, we hypothesize that by dropping sqft_above, we can still predict price with the remaining predictors. We will run a partial F test with our null hypothesis stating that the coefficient for sqft_above equals zero, implying that we can predict price with the remaining variables and we can drop this variable from our model. Since the p-value is below 0.05, we can reject our null hypothesis and conclude that we do need sqft_above in our model. Our resulting model that we will continue with is given in Figure 3.2.1

```
Analysis of Variance Table

Model 1: price ~ bedrooms + bathrooms + sqft_living + waterfront + view +
    sqft_living15
Model 2: price ~ bedrooms + bathrooms + sqft_living + waterfront + view +
    sqft_above + sqft_living15
  Res.Df        RSS Df Sum of Sq      F   Pr(>F)
1  17280 1.0137e+15
2  17279 1.0131e+15  1 5.408e+11 9.2234 0.002393 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.2.1

Now that we have a starting model generated from our automated selection, we want to try improving this model. It is important to keep in mind that in our automated selection process, our starting model had an effect on what final model was generated. If our starting model was different, we could have ended up with a different result that may have had different variables with different predicting accuracy. Furthermore, the automated selection process did not consider any interactions or higher order terms, and we did not check that all of the regression assumptions are met. Thus there is no guarantee that the best model was found yet.

Our next step was to check our regression assumptions for our model. We started by looking at our residual plot. The vertical variation of the data points around the regression equation do not have the same magnitude everywhere. Therefore the errors do not have constant variance. According to the general rules for data transformations, we should first transform the response variable first in an attempt to fix the non-constant variance of the error terms. We will look at a Box Cox plot to determine what the best transformation is. Since 1 is not inside the interval, a transformation is needed for the response variable. The best lambda value to choose in the interval is $\lambda=0$, and we apply a log transformation to our price variable.

The plot improved greatly and the issue of heteroskedasticity is no longer present. However, we cannot say that the data does appear to follow a linear pattern. In our exploratory data analysis, we noticed that the distributions for two of our predictor variables (sqft_living and sqft_living15) were right skewed.

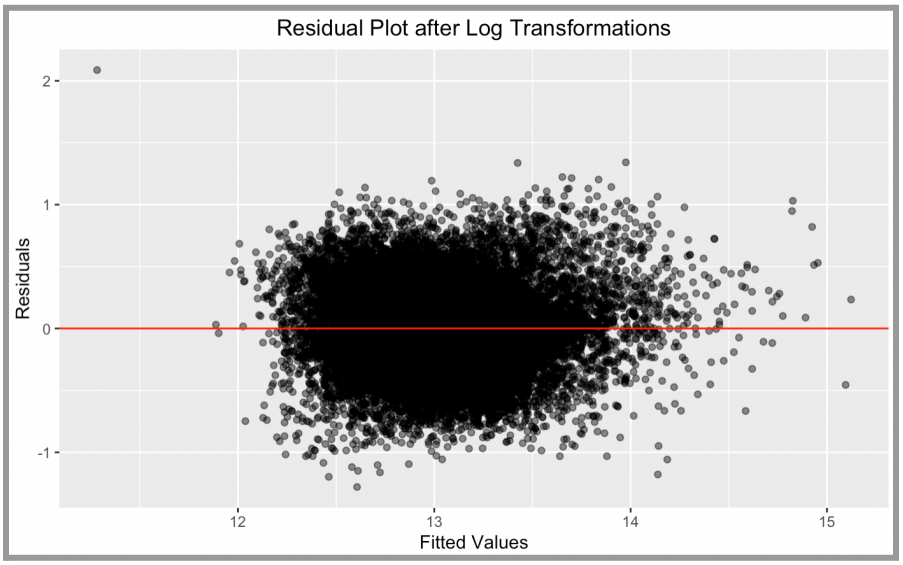

Figure 3.2.2

Applying a log transformation on these predictors changed their distribution to normal. From this we hypothesize that log transforming these two predictors our linearity assumption will be satisfied.

Our residual plot in Figure 3.2.2 continues to show improvement, and we will continue using this model with the new log transformations on price and the square foot living predictors. We still have to continue checking the remaining assumptions. The ACF plot illustrates how strong the correlation is of the current residual with the residual of the previous position, and the dashed blue lines represent critical values. Since all of the lines fall within the range of the dashed lines, this means we can conclude independence, and thus a linear regression is appropriate. The QQPlot allows us to check the normality assumption, and since the line is straight this indicates that the distribution is normal (ignoring the extreme points near the beginning and end of the line). Thus all four assumptions are met for linear regression. We will continue using this model as we proceed.

The next step we wanted to explore was any possible interaction terms. We hypothesized that there might be a possible interaction effect between the view index of the house and the amount of square feet on the price, as well as waterfront and square feet. To explore this, we generated a scatterplot of price versus square foot, with a color for each view index, and the same for waterfront.

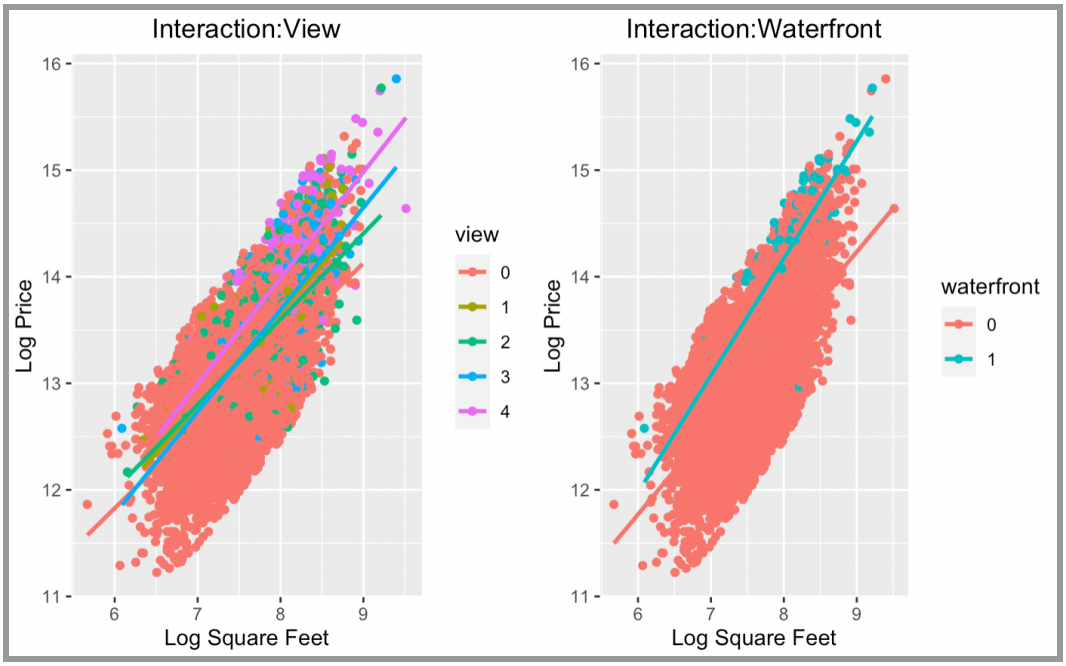

Figure 3.2.3

In Figure 3.2.3 we can see that as view increases, the slope for the log of price versus the log of square foot living seems to increase. The same observation remained true for waterfront in. So there may be an interaction effect. In other words, the type of view and waterfront does have an effect on price when looking at the relationship between square foot and price. When we used a partial F test to determine if we need to add these interaction terms into our model, the results indicated that both interaction terms were significant, so we decided to include the one with view. However, when we took a look at multicollinearity, the variance inflation factors for all of these interaction terms were in the hundreds, so we dropped the interaction completely. There were no other variables that were highly correlated, and in the ANOVA F Test output of the model all predictors were significant, so we decided to continue with the below model in Figure 3.2.4.

```
Analysis of Variance Table

Response: log_price
                   Df  Sum Sq Mean Sq  F value    Pr(>F)
bedrooms            1  565.50  565.50 4247.012 < 2.2e-16 ***
bathrooms           1  907.58  907.58 6816.153 < 2.2e-16 ***
log_sqft_living     1  781.55  781.55 5869.624 < 2.2e-16 ***
view                4  155.79   38.95  292.515 < 2.2e-16 ***
waterfront          1    7.88    7.88   59.218 1.487e-14 ***
log_sqft_living15   1   77.09   77.09  578.933 < 2.2e-16 ***
Residuals       17280 2300.85    0.13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.2.4

Our next step for model diagnostics is to check for outliers, high leverage points, and influential points. Using the Bonferroni procedure, only one outlier was identified so we will explore if it is influential in further steps. We also observed 2070 high leverage observations, and now we need to examine if any of these points are influential. Since our overall question was how well does our model predict price with these 6 predictors, we will use Cook's Distance as the measurement for influential observations. Based on this result, our model does not have any influential observations.

Now we need to test for the equality of variances across all levels of our categorical predictors with Levene's Test. We displayed the distributions for each level of each predictor using boxplots below in Figure 3.2.5. Since the p-values are so low for both categorical predictors - view and waterfront - we will check if the variances fall within a 1.5 scale of the maximum and minimum. Since the minimum variance for view is within a 1.5 scale of the maximum variance for view, we can proceed. The same is true for waterfront, but our scale is slightly bigger at 1.9 so we will proceed with caution.

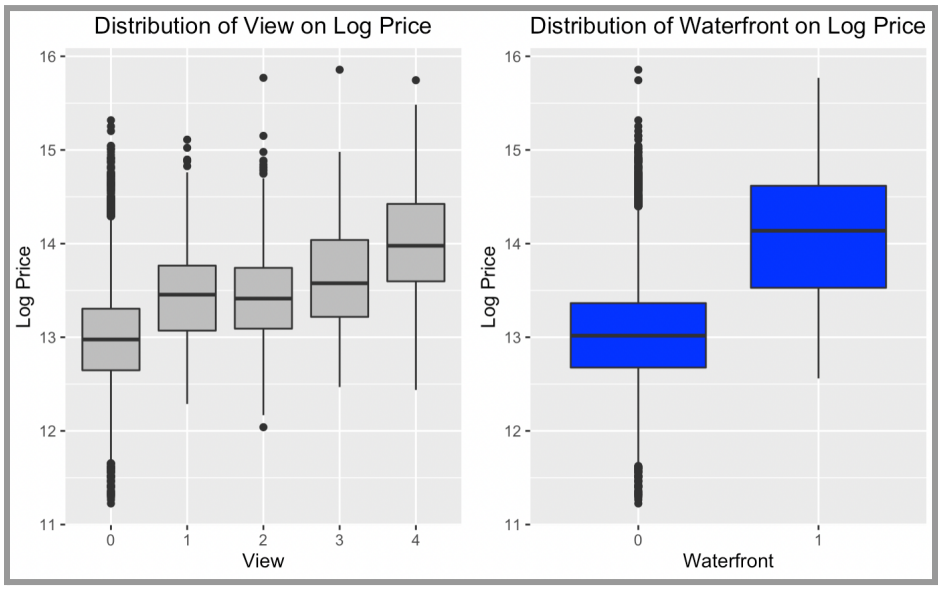

Figure 3.2.5

Our final step is to determine how well our model fits on our test data. The R squared prediction value can be used to evaluate how well our model does on new observations. So in our case, 0.5194858 of the proportion of the variation in price can be explained by our model. Since there is not a huge difference between the R squared prediction of 0.5194858 and the R squared of 0.5202798, we can say this does not indicate over-fitting. We will use the Root Mean Square Error as the metric to determine how well our model predicts price. This model had an RMSE of 0.3651696, which is relatively acceptable.

Comparing the adjusted R squared of the intuitive model with the automated model, the intuitive model had a higher Adjusted R squared at 0.6317, compared to 0.5200. The intuitive model also had a lower RMSE value of 0.3161, compared to 0.3652. We decided to use root mean standard error as the metric for our model because it's a standard way to measure the error of a model in predicting quantitative

data, and can be measured in the same units as our response variable. Based on both of these criteria, we will choose to go with the intuitive model as our final model.

Variables	Estimate	Interpretation:
intercept	16.7500	When all other predictors are zero, the predicted log price of the house is 16.75 dollars.
log.sqft_living	0.4184	For each one unit increase in log square footage, the predicted log price of the house decreases by 0.4184 dollars, controlling for all other variables.
waterfront1	0.3572	For waterfront properties, the predicted log price of the house increases by 0.0049 dollars as compared to non waterfront properties, controlling for all other variables.
view1	0.1921	For houses with a view index of 1, the predicted log price of the house increases by 0.1921 dollars, controlling for all other variables.
view2	0.1091	For houses with a view index of 2, the predicted log price of the house increases by 0.1091 dollars, controlling for all other variables.
view3	0.1490	For houses with a view index of 3, the predicted log price of the house increases by 0.1490 dollars, controlling for all other variables.
view4	0.2162	For houses with a view index of 4, the predicted log price of the house increases by 0.2162 dollars, controlling for all other variables.
grade	0.2448	For each one unit increase in grade index, the predicted log price of the house increases by 0.2448 dollars, controlling for all other variables.
yr_built	-0.0049	For each additional increase in year, the predicted log price of the house decreases by 0.0049 dollars, controlling for all other variables.

Table 3.2.1: Final MLR Model

Ultimately, we found that the most important variables in predicting a house were the square footage of the living space, whether the house is a waterfront property, the view from the house, the grade of its construction quality, and the year the house was built. Table 3.2.1 shows the predictors included in our final regression model and the interpretations for all coefficients. Interestingly, three of the five predictive variables did not actually represent physical characteristics of the house, as may have been predicted going into this analysis. Based on the view quality and waterfront status being influential, our analysis supports the notion that location is a key factor in determining the price of a house. In practice, our model could be used to help find the expected price of a house so a buyer or seller can determine whether a property is under- or over-priced.

Section 4: Logistic Regression

The logistic regression question we explored was: do the physical characteristics of a house (number of floors, size, etc.) influence the quality of craftsmanship during construction? Another way to phrase the above question is can the level of construction and design (grade) be accurately predicted by the physical characteristics of a house. For this study, analysis accurately predicted will be defined by an overall error rate less than or equal to 0.2 (20%). The physical characteristics of a house used as possible predictors for grade include:

- Number of Bedrooms (bedrooms)
- Number of Bathrooms (bathrooms)
- Square Footage of the Home (sqft_living)
- Square Footage of the Lot on which the Home is Built (sqft_lot)
- Number of Floors (floors)
- Square Footage of the Home that is Above Ground (sqft_above)
- Square Footage of the Home that is Below Ground (sqft_basement)

Before analysis can begin we must recognize that the square footage of the home is equal to the amount of square footage above ground plus the amount of square footage below ground. To reconcile this issue of multicollinearity, we will not be considering the square footage of the home that is below ground in our analysis. Secondly, we will be using the new grade variable, grade.cat, which collapses the classes into average or below and above average as our response variable.

To begin our model analysis, we started by randomly splitting our data into a training set and a test set with a 80-20 split respectively. A seed of 101 was used during the splitting process to allow for reproducibility of our results. We will use the training set to build the model and the test set to verify the accuracy of our model. Next, we fit a model with all of the physical home characters to see what the initial findings would yield. Starting with a full model is desired because partial ANOVA F tests can be run to drop parameters, if needed. The output of the full model regression is displayed below in Table 4.1. Upon studying the output it looks as though we have a fairly good model since all predictors are shown to be statistically significant with the other predictors in the model.

```
##
## Call:
## glm(formula = grade.cat ~ bedrooms + bathrooms + sqft_living +
##     sqft_lot + floors + sqft_above, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8830  -0.6033  -0.2410   0.6066   2.6541
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.782e+00  1.128e-01 -51.254  < 2e-16 ***
## bedrooms    -6.250e-01  3.179e-02 -19.658  < 2e-16 ***
## bathrooms    8.625e-01  4.775e-02  18.061  < 2e-16 ***
## sqft_living  1.808e-03  6.830e-05  26.468  < 2e-16 ***
## sqft_lot    -1.681e-06  6.343e-07  -2.650  0.00806 **
## floors       8.715e-01  4.786e-02  18.211  < 2e-16 ***
## sqft_above   6.310e-04  6.630e-05   9.518  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 23929  on 17289  degrees of freedom
## Residual deviance: 14061  on 17283  degrees of freedom
## AIC: 14075
##
## Number of Fisher Scoring iterations: 6
```

Table 4.1: Output of the Full Model Regression

To check if there is possibly a better model we ran a stepwise regression starting with the intercept model to see if the result would also support using the full model. Once the regression was completed the output did confirm and validate that a full model with all predictors is in fact the most desired model.

Next an ANOVA F test was completed to answer the question, is our model useful? A summary of the test can be found in Table 4.2. The resulting p-value is shown to be 0 and therefore we can reject the null hypothesis. The 6-predictor model is chosen over the intercept-only model and it can be said that the 6-predictor model is useful. Our final estimated regression equation is $\log[((\pi / (1 - \pi))] = -5.782 + -0.6250*\text{bedrooms} + 0.8625*\text{bathrooms} + 1.808e\text{-}03*\text{sqft\_living} + -1.681e\text{-}06*\text{sqft\_lot} + 0.8715*\text{floors} + 6.310e\text{-}04*\text{sqft\_above}$.

Null Hypothesis	H0: $\beta1 = \beta2 = \beta3 = \beta4 = \beta5 = \beta6 = 0$
Alternate Hypothesis	Ha: At least one coefficient in H0 $\neq 0$
Test Statistic	9867.681
p-value	0

Table 4.2: Results of the ANOVA F Test

Now that our model has been determined the next step is to check the accuracy of it and see how well it does predicting grade. The first step in assessing the model is creating an ROC curve, using the test set, to see if the model does better, the same or worse than random guessing also known as no info classification. Figure 4.1 shows the resulting ROC curve for our full model. In the graph the red line represents no info classification. Because our curve is above that line, our model does better classifying the observations

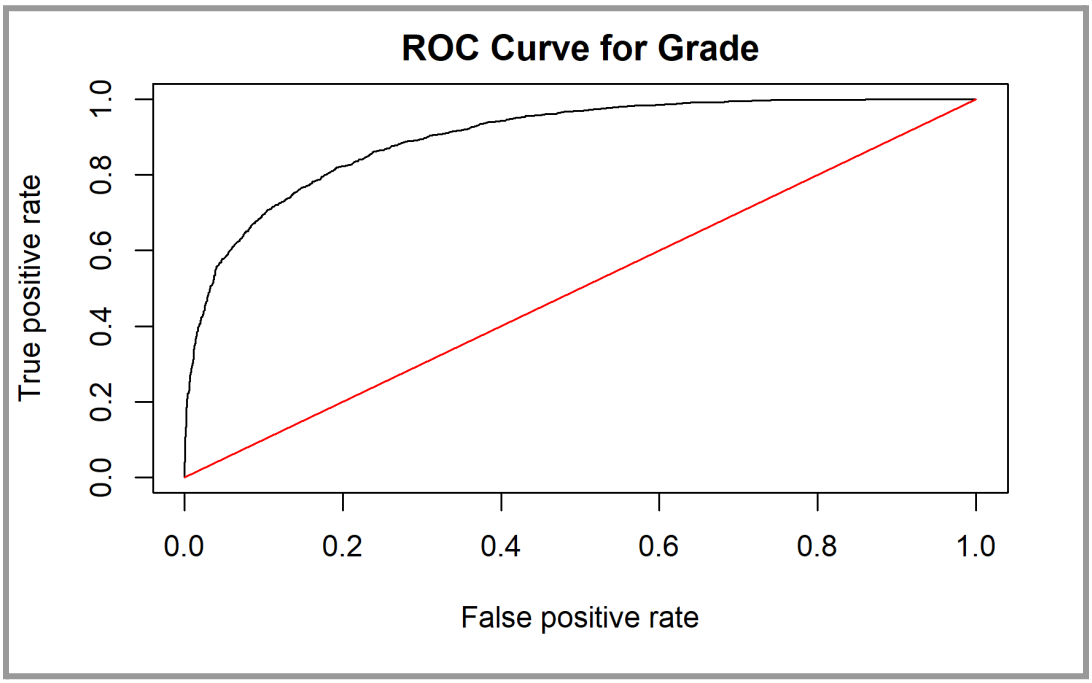

Figure 4.1

Another way to verify that our model is better than no information classification is to look at areas under the ROC curve (AUC). The AUC value associated with random guessing is 0.5, AUC values close to 1 are most desirable because it indicates the model does better than random guessing in classifying observations. Our model produced an AUC value of 0.9005 which verifies our predictions are being classified with a high amount of accuracy. The last way to check the ability of the model is by creating a confusion matrix and calculating the import statistics that are associated with the matrix. The confusion matrix shown in Table 4.3 was created using a threshold value of 0.5. The resulting statistics calculated from the matrix are shown in Table 4.4. Table 4.3 again confirms how our model more accurately classifies the grade then random guessing. The overall error is approximately 19%, the false

positive rate (FPR) is about 17% and the false negative rate is approximately 21%. Each one of these numbers represents error, or incorrectly predicted terms, so the goal is to make the value small. With the max of these error terms being 21%, we conclude our model is making predictions correctly the majority of the time.

Category	FALSE	TRUE
Average or Below	1847	384
Above Average	442	1650

Table 4.3: Confusion Matrix

Statistic	Value
Overall Error Rate	0.191071
Accuracy	0.808929
False Positive Rate	0.1721201
False Negative Rate	0.2112811
Sensitivity	0.7887189
Specificity	0.8278799

Table 4.4: Confusion Matrix Statistics

Reflecting on our results, the two coefficients that stand out are bedrooms and square footage of the lot; both coefficients are negative. Looking at the bedrooms coefficient the estimated log odds of a house having an above average grade decreases by 0.6250 for a one-unit increase in the number of bedrooms, while controlling for all other predictors. As shown by the initial data visualizations the grade seemed more likely to be above average when there were 2 or more bathrooms. One may assume that finding would be the same with bedrooms but our model shows that in fact is not the case. The other interesting coefficient, square footage of the lot, can be interpreted as the estimated log odds of a house having an above average grade decreases by 1.681e-06 for a one-unit increase in the square footage of the lot while controlling for all other predictors. From our initial data visualizations we observed that square footage of the lot did not have much influence on if the grade was average and below or above average. Therefore, it is not surprising the coefficient is very small. The unusual part of the interpretation is that it does decrease the odds even though it is a minimal decrease.

In conclusion, the final model included all predictors considered to be physical characteristics of a house. These predictors were able to predict the level of construction and design accurately as defined by having an error rate less than or equal to 20%.