STAT 6021: Project 2

Image source: https://www.racialequityalliance.org/wp-content/uploads/2016/10/assessors_social-1.jpg

Group #14

Anne Louise Seekford (bng3be), Humaira Halim (hbh4bv), Tatev Kyosababyan (ers5af), Ali Roth (wat6sv)

Section 1: Executive Summary

Using data on houses sold in King County, Washington, we examined two questions of interest that would be applicable on property listing service applications. The first question was whether the square footage inside a home could be predicted using other characteristics of the house. Oftentimes on listing services, square footage is missing from a listing, but is an important decision for a home buyer. Being able to provide an estimated square footage based on other characteristics could increase ease of use for a buyer and could flag potential data entry errors by selling agents. The goal of the second question is to predict whether a house is located in Seattle or elsewhere in King County, given certain qualities of the home. This could be used to determine if you should buy a house in Seattle or outside Seattle based on the characteristics you desire in a property.

To address our first question, we created a model that predicts the amount of living space in square feet of a home based on the number of bedrooms, bathrooms, floors, if there is a view, the condition of the home, the year the home was built, and whether the home was located inside Seattle or elsewhere in King County. We believe that this estimated square footage could be used two ways: to provide an estimate when square footage is unknown, like in a foreclosure or condemned building, or to raise an error if a square footage is entered drastically incorrect by a listing agent for a property. In most states, listing an accurate square footage is required by law, and even a data entry error could result in a lawsuit. Our model was able to use common listing information to estimate the square footage of homes.

The model produced suggests the effects of some characteristics increase the living space of a home, while others decrease the living space when the other characteristics stay the same. The characteristics that increase the living space are more bedrooms and bathrooms, more floors, a view above ‘Fair’ (1), and better conditions. The characteristics that tend to decrease the amount of living space are being built more recently, being in Seattle, and having a view score of ‘Fair’ (1) or no view at all. The exact change that each characteristic creates in the number of predicted square feet is difficult to interpret, but we see the pattern that is suggested.

For our second question, we found that we were able to produce a model that could find the probability that a house with specified characteristics would be available in Seattle, via logistic regression. The home characteristics included are price of the home, whether the home is on a waterfront, the square footage of the lot, the condition of the home, the number of bedrooms and bathrooms, the square footage of the home above ground level, and the view score. Each characteristic changes the likeliness of a house being in Seattle when all the other characteristics remain the same. Higher price, more bedrooms, and views all increase the likeliness of a house being in Seattle, while waterfront properties, bigger lots, better conditions, more bathrooms, and more square footage above ground level all decrease the likeliness of a house being in Seattle. These characteristics are used to produce the “log odds" of a home having a Seattle zipcode. The “log odds” are a relative of probability, which can be calculated to determine how likely a home with those characteristics is to exist in Seattle.

We used this model to determine if a 2 bedroom, 2 bathroom home on a waterfront with a view rating of "Excellent" (4), with a 2,000 square foot lot, in condition "Good" (4) with 1,000 square feet above ground level for $660,000 would be available in Seattle. Our model predicts that the probability of the house being in Seattle is 91.09%. This information would be valuable to a client looking for a home could look in the city or outside of the city. Since the probability of a house with these characteristics being inside Seattle is high, this client would likely have success looking for a home in Seattle.

The data from King County, Washington were useful in producing models that predict the number of square feet in a home and whether a home was in Seattle. Since different regions of the country have such different housing markets, this model may not be applicable to other regions, or even other counties in the state of Washington. However, the same methodology could be applied to create a model for other regions using data on houses sold in that area.

Section 2: Description of Data and Variables

Data Description

The dataset has been acquired through Kaggle. The Data represents the homes sold between May 2014 and May 2015 in King County, including relevant information regarding different aspects of the houses. The dataset has 21613 observations and 22 attributes providing details about each observation. There are no missing values in the dataset. The following characteristics provided were used or considered in this analysis:

- *Id*: For identification purposes, every house has its unique ID.
- d*ate*: The date when the houses were sold
- *price*: The houses vary by price, ranging from $75 thousand to almost $8 million. In some graphs the price will be represented by millions or hundred thousands for easier interpretation.
- *bedrooms, bathrooms*: Numeric amount of designated bedrooms and bathrooms correspondingly. While the number of bedrooms are indicated in the form of integers, the bathrooms can be represented by floating points as well. The value 0.5 stands for a toilet without a shower in the bathroom, and a 0.75 value represents a bathroom that has either a shower or a bath.
- *sqft_living, sqft_lot*: For each house in our dataset, the square footage of the interior living space as well as the footage of the land space are recorded as numeric values.
- *floors*: The number of floors as a numeric value.
- *waterfront*: Dummy variable, where values of 1 mean the apartment is overlooking the waterfront, and 0 means otherwise.
- *view:* Values from 0 to 4 assessing how good the view is from the property.
- *condition*: Values from 1 to 5 assessing the condition of the apartment.
- *grade*: Values assessing the level of construction and design.

1-3: lower quality level,
7: average level,
11-13 high quality level.

- *Sqft_above, sqft_basement*: The square footage of interior housing space correspondingly above and below ground level.
- *yr_built , yr_renovated*: The year the house was built and the year it last got renovated correspondingly.
- *zipcode*: The zipcode where the house belongs to.
- *lat*: Latitude as numeric.
- *long*: Longitude as numeric.
- *sqft_living_15, sqft_lot15*: Square footage of correspondingly the interior housing living space and land lots for the nearest 15 neighbors.

In addition to the existing variables, we added another one for our findings:

- *Seattle*: Dummy variable based off of the provided zipcodes of the houses, where values of 1 represent the house is in Seattle and 0 otherwise.

One objective of this project is to predict the square footage of a house with the help of a model that considers influential variables for the question of interest. As a result, prospective home buyers or participants of the real estate market would be able to approximate the square footage of a house based on other components. Furthermore, it is illegal for home selling agencies to provide incorrect information about the square footage of a home. For this reason, our predictive model can be used to ensure the provided information about a home makes sense, and thus predict potential unlawful actions by the selling home agencies.

The second objective of our predictive models is having an idea of where the house would most likely be located. With the application of Logistic Regression, we want to predict whether the house is located inside or outside of the Seattle area considering other provided characteristics. We believe this model would be useful for prospective home buyers and real estate agencies to find the best fit considering available resources and defining top priority attributes in a home.

Visualizations

Question 1

Living Space Square Footage (*sqft_living*)

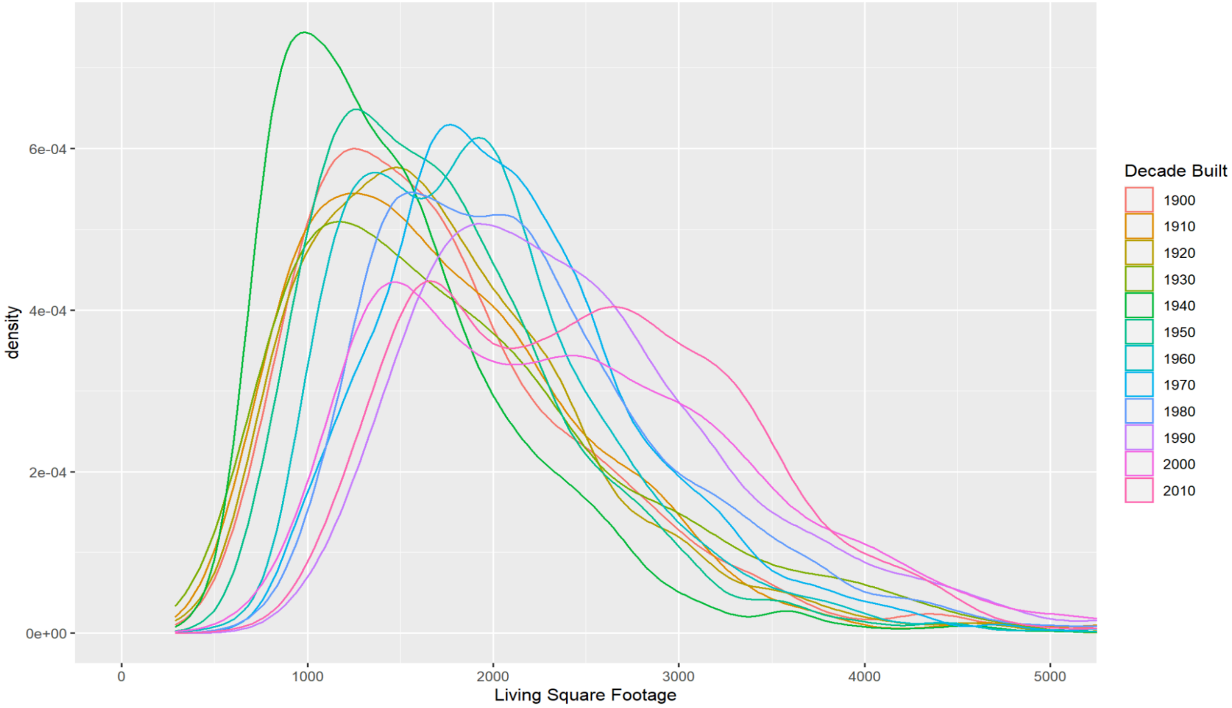


Figure 1: Living square footage and year built (grouped by decade)

In Figure 1, we can see that there seems to be a shift in bigger houses being built in later years (1990s-2010s) and smaller houses built in earlier years (1900s-1950s). We looked at decades instead of individual years because of the vast number of years in the dataset. Instead, grouping by decade made for a plot significantly easier to read.

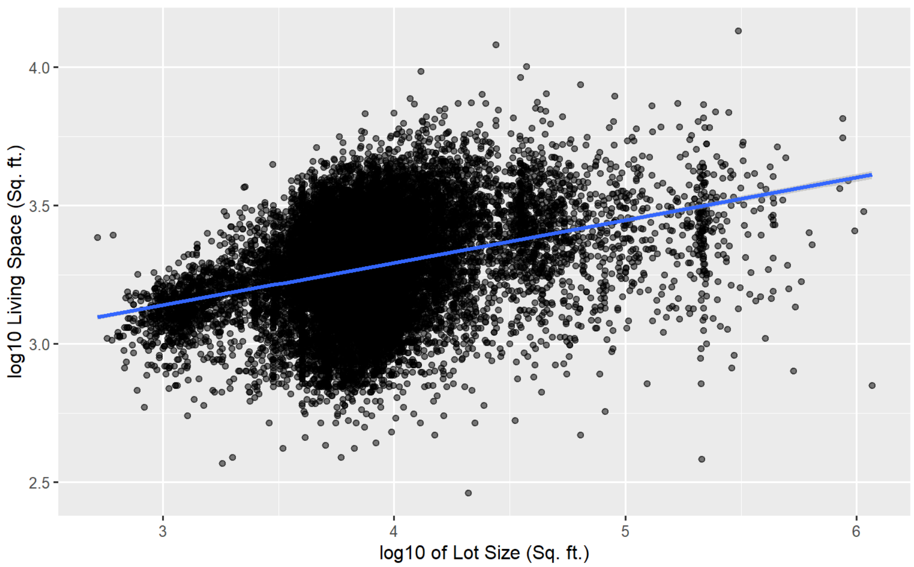


Figure 2: log(sqft_lot) by log(log10(sqft_living)

The data for lot square footage and living square footage was condensed around lower square footage for each variable. In order to better see the relationship, we transformed each with a log based 10 function to better visualize the spread. There appears to be a positive, roughly linear relationship between lot square footage and living square footage.

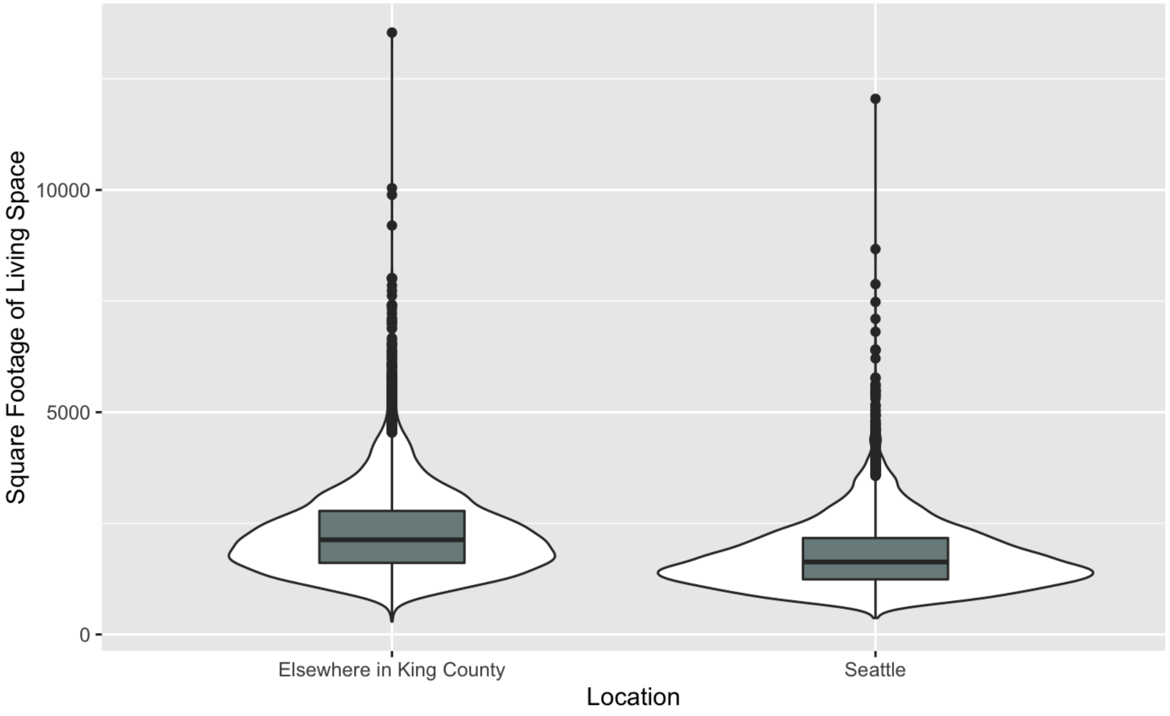


Figure 3: Square Footage of Living Space by Location

From the violin plot in figure 3, we can see that the square footage of the homes in Seattle tends to be smaller than that of elsewhere in King County. Furthermore, the square footage in King County varies a little more than the houses in the Seattle area.

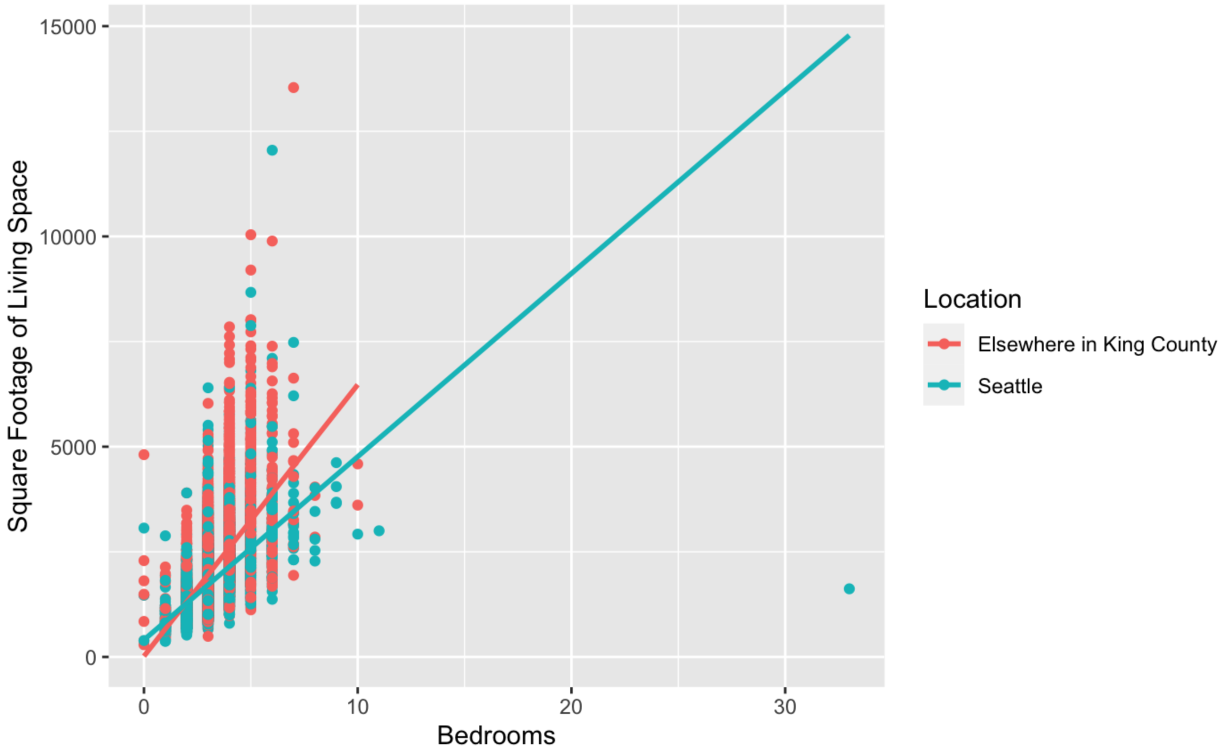


Figure 4: Number of Bedrooms by Square Footage of Living Space (grouped by Location)

The scatter plot demonstrates that there is an obvious difference in the slopes of the two categories of location we are considering.

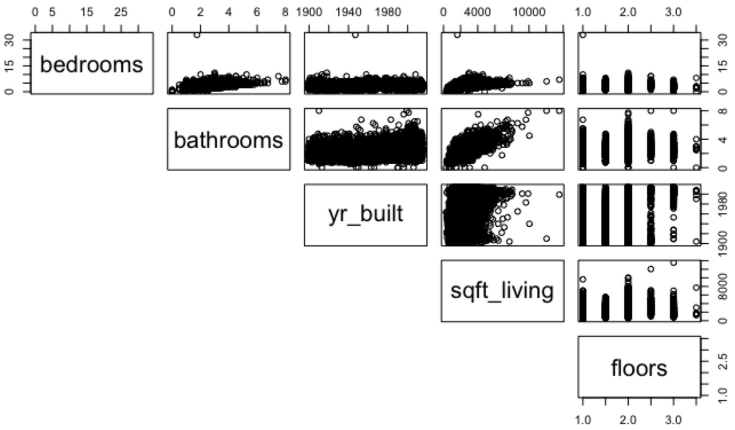
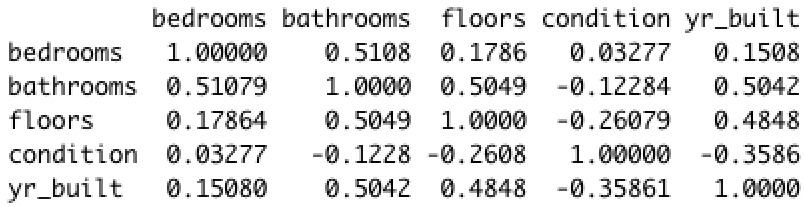


Figure 5: Correlation plots for MLR.



```
          bedrooms bathrooms  floors condition yr_built
bedrooms   1.00000    0.5108  0.1786   0.03277   0.1508
bathrooms  0.51079    1.0000  0.5049  -0.12284   0.5042
floors     0.17864    0.5049  1.0000  -0.26079   0.4848
condition  0.03277   -0.1228 -0.2608   1.00000  -0.3586
yr_built   0.15080    0.5042  0.4848  -0.35861   1.0000
```

Figure 6: Correlation matrix for predictors in MLR.

From the correlation plot in Figure 5 we can see there is some positive correlation between the bedrooms and bathrooms, along with each of these having a positive correlation with the squared footage of living space. This agrees with the numeric values of the correlation matrix in Figure 6.

Question 2

Number of Bathrooms (*bathrooms*)

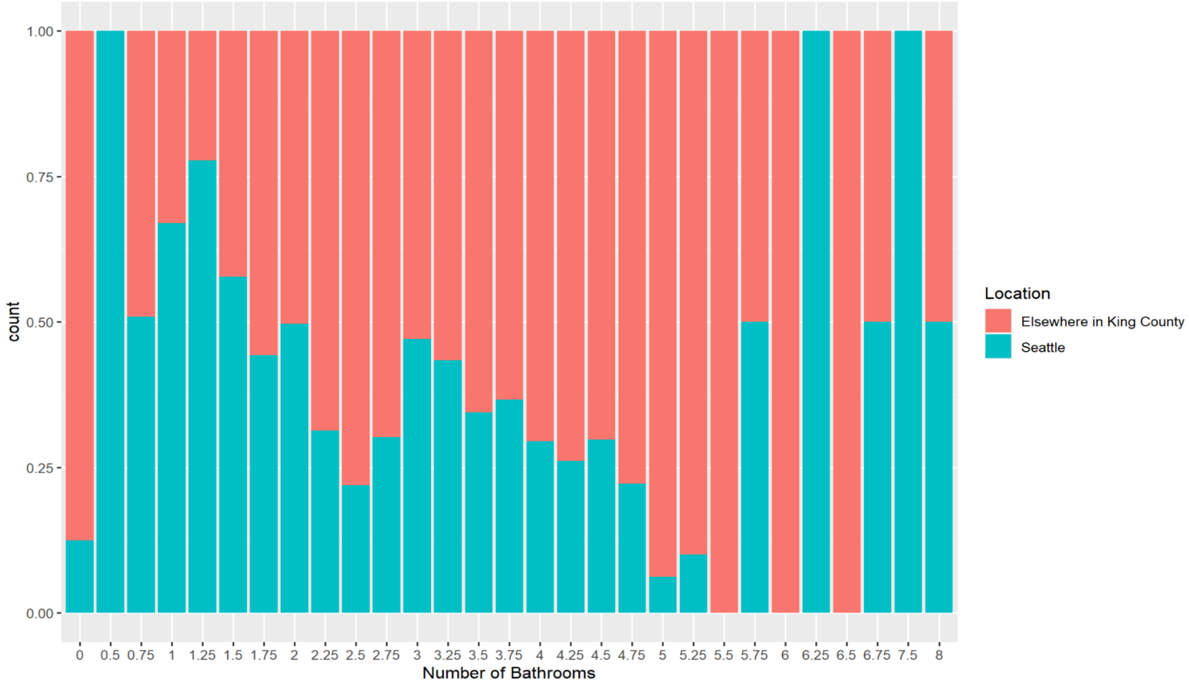


Figure 7: Proportion of bathrooms by location

When you ignore the extreme lower and upper values (less than 1 and more than 5 bathrooms), there appears to be a linear relationship between bathrooms and whether a house is in Seattle. More houses in Seattle have less bathrooms, more houses outside of Seattle have more bathrooms.

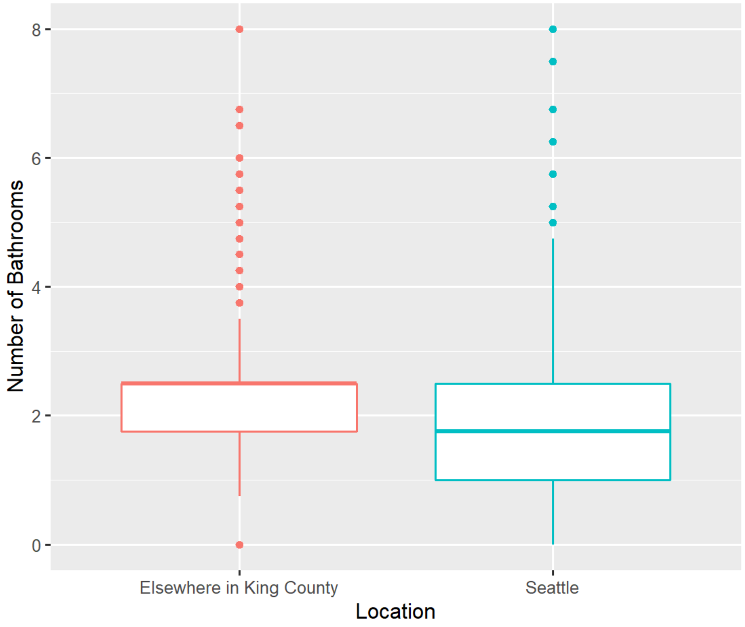


Figure 8: Number of bathrooms by location

The boxplots, shown in Figure 8, agree with the pattern seen on the bar chart - more houses with less bathrooms are in Seattle than elsewhere in King county.

Number of Bedrooms (*bedrooms*)

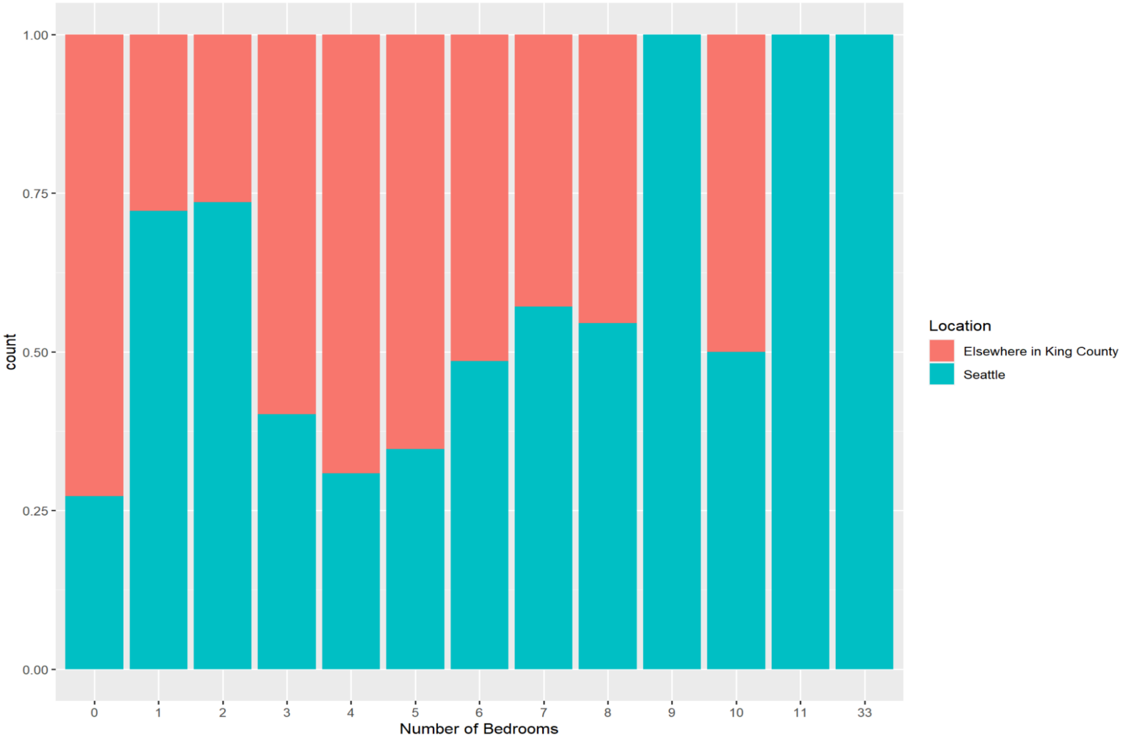


Figure 9: Proportion of bedrooms by location

Looking solely at Figure 9, we can see that there is not an obvious relationship between the number of bedrooms and if a house is in Seattle.

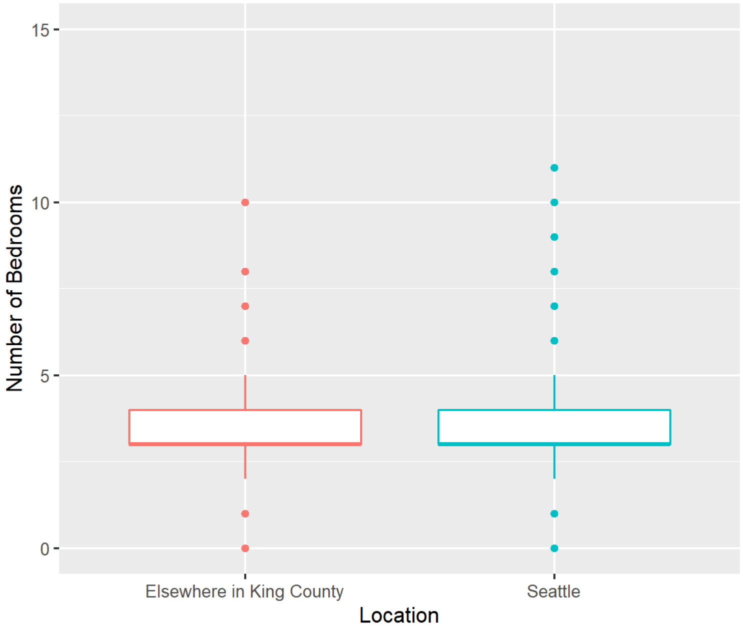


Figure 10: Number of bedrooms by location

It should be noted that the y axis was limited to 15, excluding an outlying data point with 33 bedrooms in Seattle. This gives us a better visualization of the majority of the data. The boxplot shows that the number of bedrooms in Seattle and outside of Seattle is relatively similar, and may not be a good predictor alone.

Good View Index (*view*)

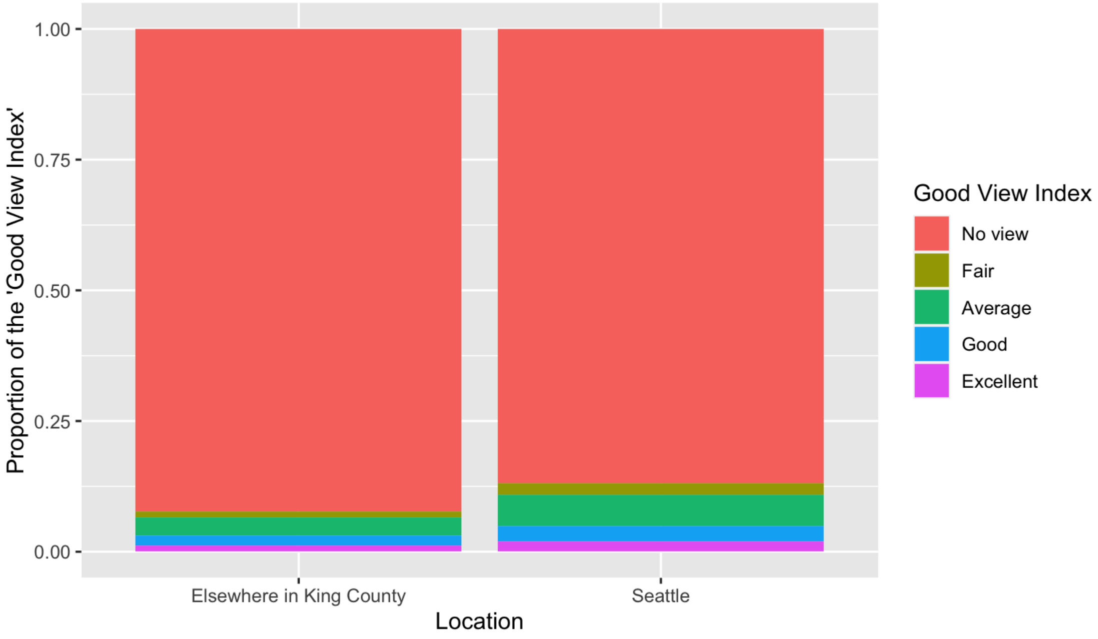


Figure 11: Proportion of the ‘Good View Index’ by location

The Good View Index is a scale ranging from 0 to 4 of how good the view of the property was: 0 = No view, 1 = Fair 2 = Average, 3 = Good, 4 = Excellent. Seattle has a larger proportion of the better views than elsewhere in King County, as expected. We expected this as Seattle has city views, the bay, and Lake Washington - the coastline of the Puget sound.

Land Square Footage (*sqft_lot*)

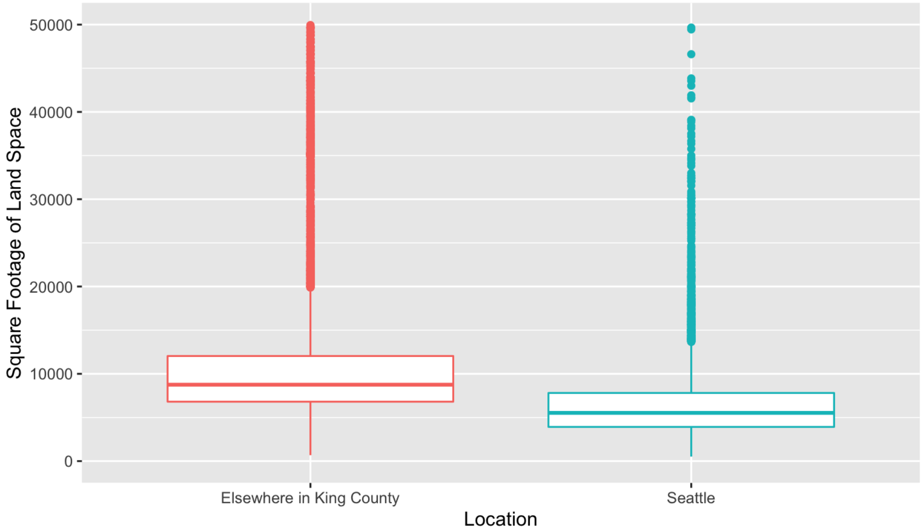


Figure 12: Boxplot of the Land Space (sq. ft.) by location

Figure 12 shows box plots for the distribution of square footage of land space in Seattle and elsewhere in King County. After setting the y-limit to a smaller value to get a better understanding of the statistical summary, the median square footage of land in Seattle is less than elsewhere in King County, as expected.

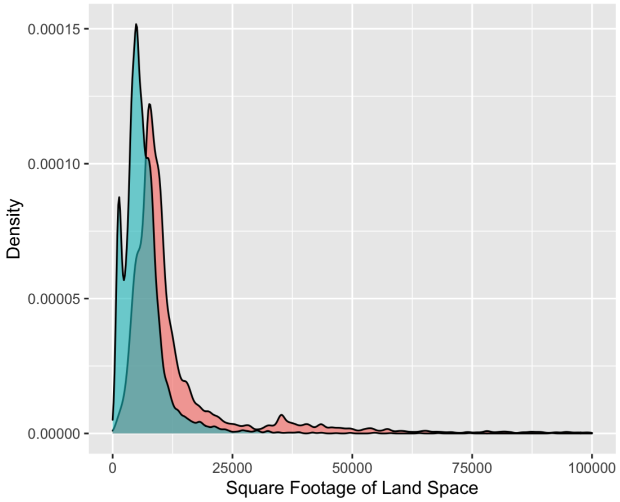


Figure 13.1: Density plot of least land space (sq. ft.) by location

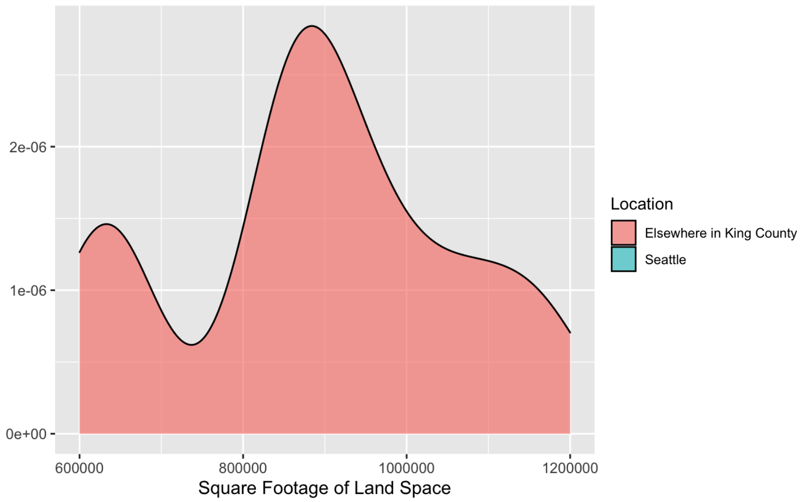


Figure 13.2: Density plot of most land space (sq. ft.) by location

It should be noted here the unit differences on the y axis'. The plots are adjusted to show the density of the smallest lots and largest lots, and their corresponding locations.

Waterfront Homes (*waterfront*)

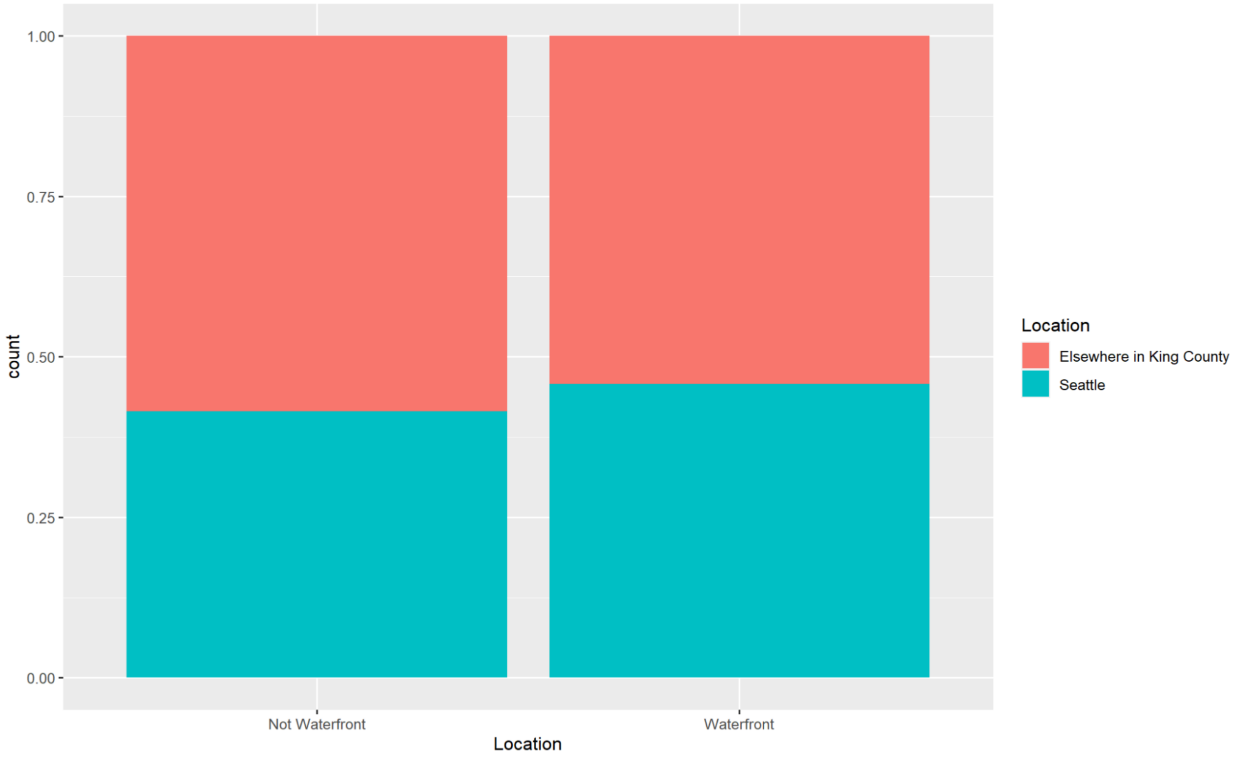


Figure 14: Proportion of waterfront homes by location

The proportion of waterfront homes in Seattle and elsewhere in King County is roughly equal. This is an indication that *waterfront* may not be a good predictor alone.

Condition of Home (*condition*)

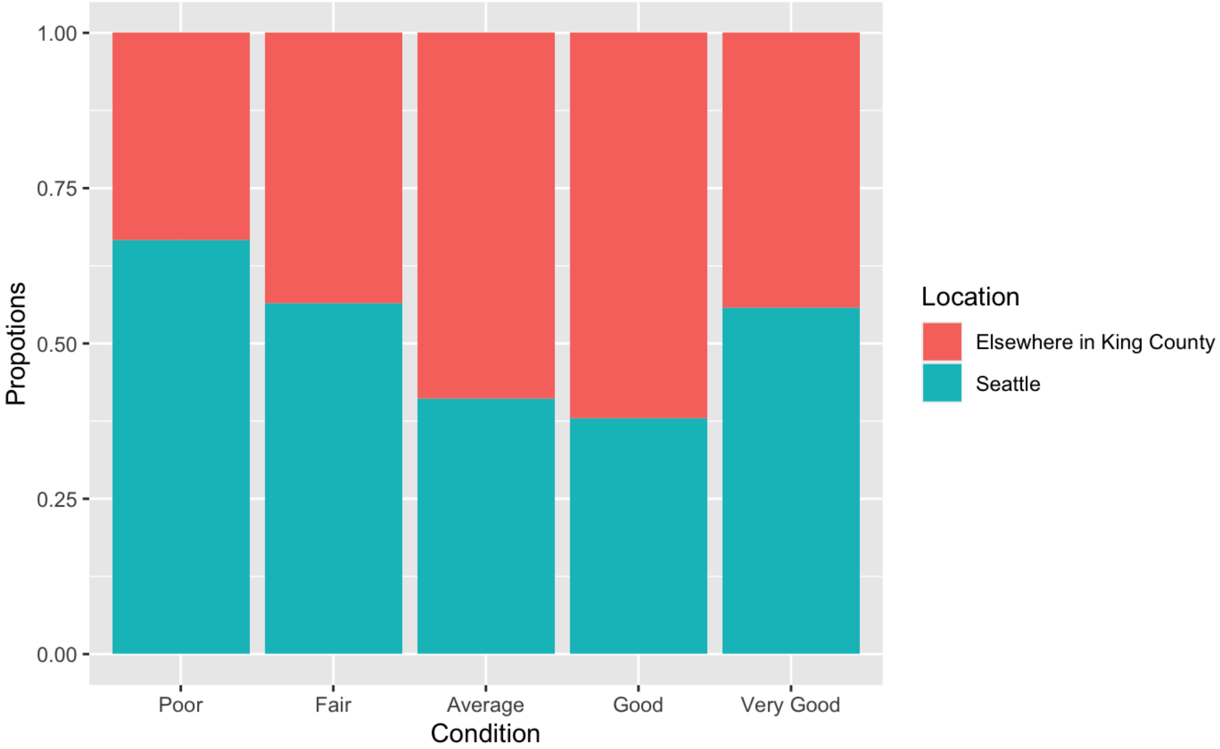


Figure 15: Proportion of home condition by location

There is some difference demonstrated in the visualization based on the condition category and the location. For example, there are more houses of "Poor" condition in Seattle than elsewhere in King County.

Price (*price*)

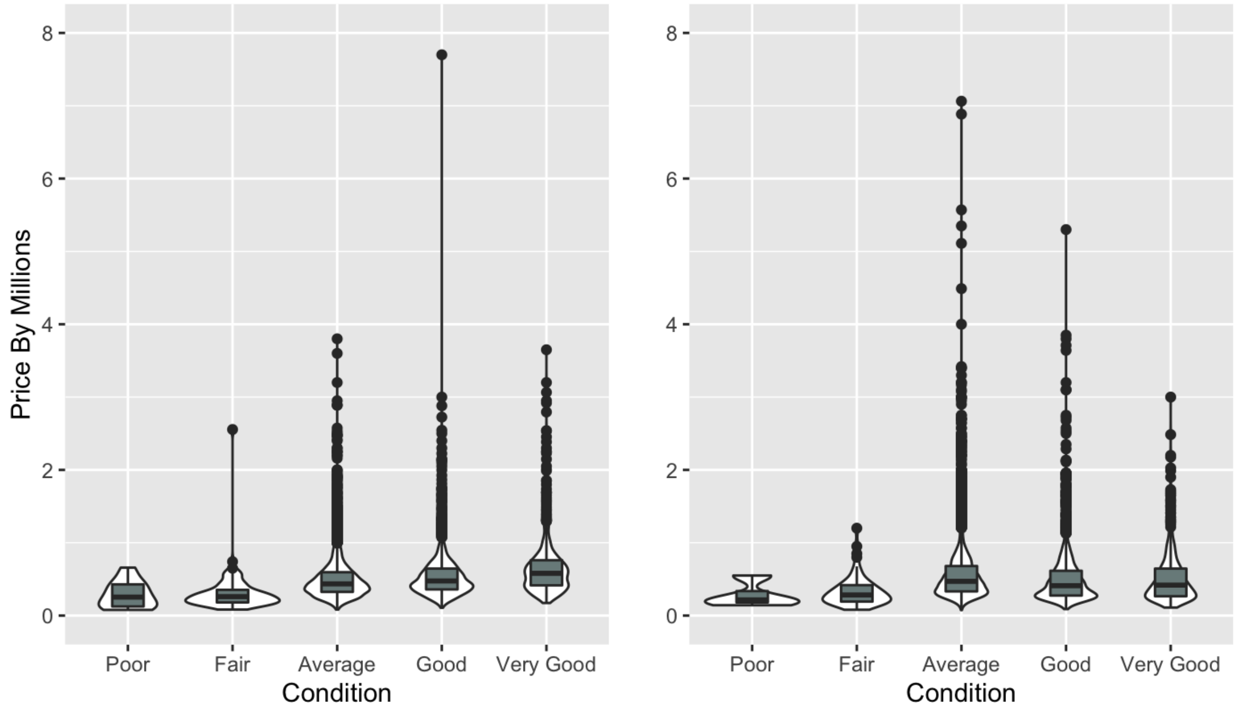


Figure 16: Price (in millions of dollars) by home condition (left: Seattle, right: Elsewhere in King County)

We can see in figure 16 that the distributions of price for each category of *condition* follows the same trend in Seattle as elsewhere in King County. This indicates that there are most likely no interactions between these predictors.

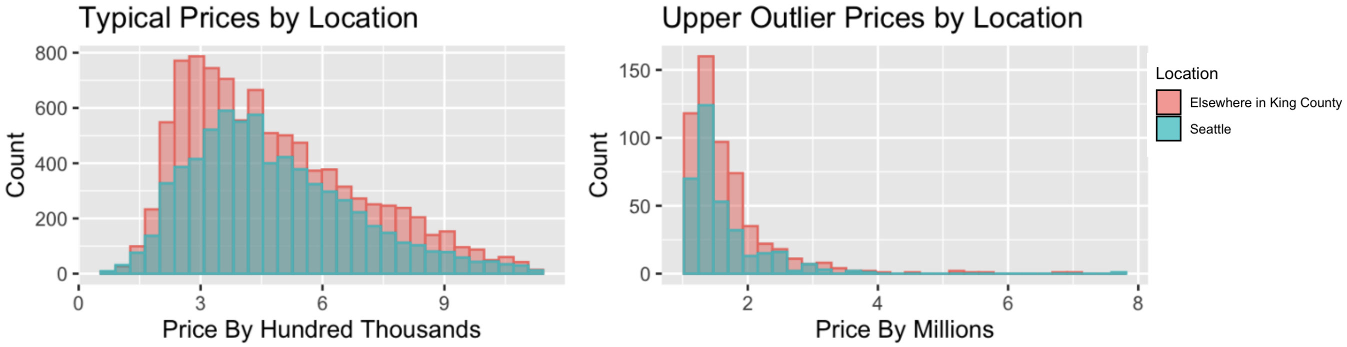


Figure 17: Price Distribution by Location.

There are some visible differences by location. We can see that there are more expensive houses outside of Seattle by sheer quantity. The typical prices by location follows a slightly right skewed normal distribution. The majority of homes, in both Seattle and elsewhere in King County are in the rough \$200,000-\$500,000 range. There are, however, more homes in King County than Seattle for that price range.

Section 3: Linear Regression

Initial Model

The first question we wanted to pursue through linear regression, was to create a model which predicts the square footage of an apartment's interior living space (*sqft_living*) using other qualities of the house itself. This question is useful to prospective home buyers as many times the size of the interior is unlisted on websites, despite it being a large factor of interest towards a buyer's decision. This would also be useful for detecting data entry errors for real estate agents, who are legally obligated to correctly report square footage of properties they list.

We decided to populate our model with characteristics of a house that are most likely to be found on a listing of a home, such as the number of *bedrooms*, *bathrooms*, and *floors*. Additionally, other factors we also included given their likelihood to be found on a website are the *view* from the property (rated 0-4 increasing by how good the view was), whether or not the home was on a *waterfront* (binary 0-1), the *condition* of the house (rated 1-5 increasing by the quality of the place), the year it was built (*yr_built*), and its *zipcode* (which we transformed to a binary variable called *Seattle* of 0-1 by whether or not the house was in Seattle as opposed greater King County).

After choosing characteristics that were relevant to a home listing, we produced several visualizations to examine the relationship between living space and each characteristic. This gave us an idea of what should be looked for when fitting a model and considering removing predictors from the model. From our visualizations, we found that there were relationships between the *sqft_living* and *sqft_lot*, *yr_built*, and *Seattle.* The relationship between number of bedrooms and square footage of living space are slightly different when the house is in Seattle and elsewhere in King county, so we will keep this in mind as we move forward with our model.

After our explorations with *sqft_living* and our predictor variables, we fit our model with all our chosen predictors as well as running a forward selection. In both practices, we found *waterfront* to be insignificant at predicting *sqft_living* in the presence of the other predictors. This decision was supplemented through our next task of performing pairwise comparisons for our categorical predictors.

We conducted pairwise comparison tests using Bonferroni Procedure to examine the interactions between the classes of our categorical predictors. We were concerned about the *waterfront* category as we expect there may be a relationship between *waterfront* and *view*, and *waterfront* produced an insignificant t test value in our model. We determined that the *waterfront* predictor could be dropped because the confidence interval produced from the pairwise comparison was -89.9987 to 175.7931, which includes 0. Going forward, we removed the *waterfront* predictor and refit the model.

Finally, we end up with our first official model:

$$\hat{y} = 6310.07 + 703.19x_1 + 234.77x_2 + 360.33x_3 + 363.76x_4 + 518.92x_5 + 805.50x_6 - 247.49x_7 - 3.31x_8 + 96.70x_9 - 20.37x_{10}$$

where,

x_1 = *bathrooms*

x_2 : *bedrooms*

x_3: 1 for a *view* index of 1, 0 otherwise

x_4: 1 for a *view* index of 2, 0 otherwise

x_5: 1 for a *view* index of 3, 0 otherwise

x_6 = 1 for a *view* index of 4, 0 otherwise

x_7 = 1 for indicating a *Seattle* home, 0 otherwise

x_8: *yr_built*

x_9 = *floors*

x_{10} = *condition*

Model Improvements

In the interest of testing the usefulness of our model, we must check the assumptions for linear regression. First, we start off with the scatterplot of our fitted y values vs. our residuals in order to see if we need to do any transformations to our variables.

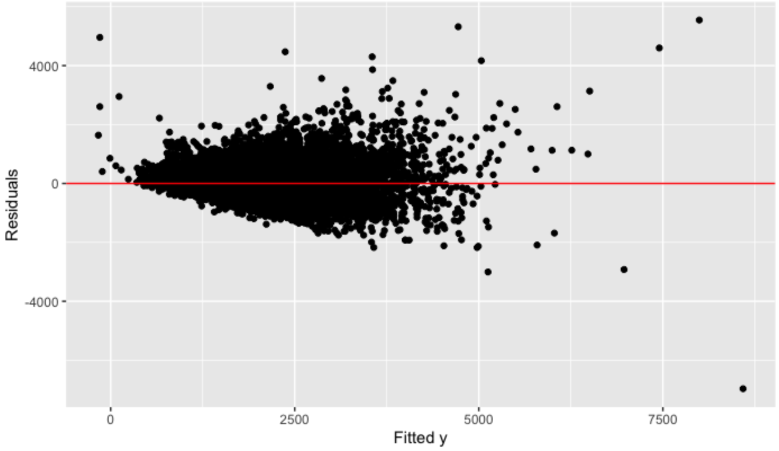


Figure 18: Residual Plot

The scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values). The first assumption states that the residuals should be evenly scattered on both sides of the red horizontal line as we move from left to right. We see this in the residual plot, so assumption 1 is met.

The second assumption states that the vertical spread of the residuals should be constant as we move from left to right. The spread seems to have a fanning out pattern which makes it inconsistent; therefore, assumption 2 is violated and we might elect to transform our model starting with our y variable, *sqft_living* and then moving to the x variables if needed.

In order to understand how to transform our y variable, we generated a boxcox plot:

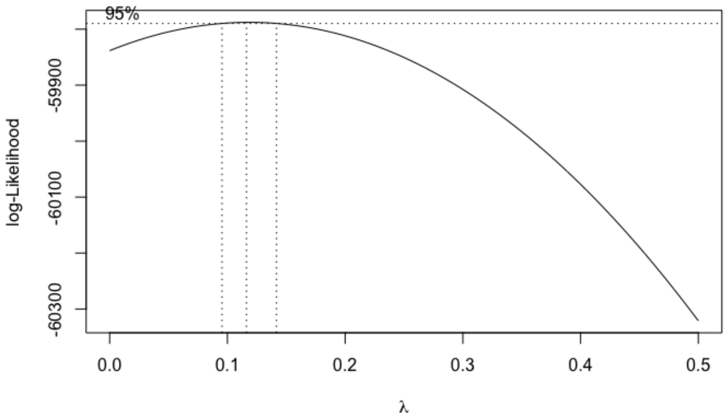


Figure 19: Box Cox Plot

The transformation takes the form y* = y^(λ) , with the value of λ to be chosen. Given λ=1 is not within our confidence interval (illustrated by the dotted lines) it is more evident that y must be transformed. We decided to go with λ=0.1 since it is clearly within that range. After this transformation, we have and improved model of:

$$y* = 2.2859 + 0.0284x_1 + 0.0652x_2 + 0.0108x_3 - 0.0021x_4 + 0.0322x_5 + 0.0441x_6 + 0.0584x_7 + 0.0009x_8 - 0.0002x_9 - 0.0253x_{10}$$

where,

$y * : y^{0.1}$

x_1 : *bedrooms*

x_2 : *bathrooms*

x_3: *floors*

x_4: 1 for a *view* index of 1, 0 otherwise

x_5: 1 for a *view* index of 2, 0 otherwise

x_6: 1 for a *view* index of 3, 0 otherwise

x_7 : 1 for a *view* index of 4, 0 otherwise

x_8 : *condition*

x_9 : *yr_built*

x_{10} : 1 for indicating a *Seattle* home, 0 otherwise

Now that we have refitted our model with the ystar variable, we can recheck our model assumptions again. We will start with the scatterplot for assumption 1 and 2:

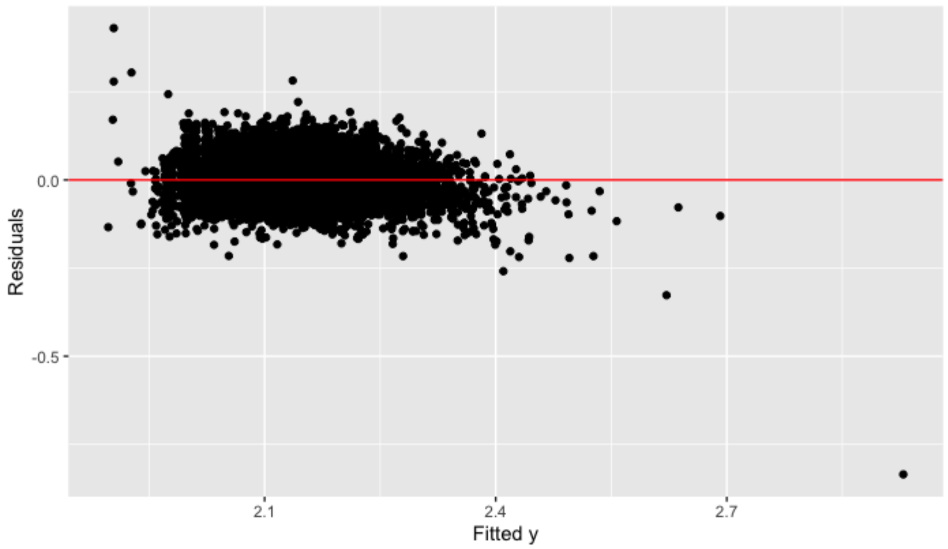


Figure 20: Residual Plot, with ystar

The scatter plot shows the distribution of ystar's residuals (errors) vs fitted values (predicted values). The first assumption states that the residuals should be evenly scattered on both sides of the red horizontal line as we move from left to right. We see this in the residual plot, so assumption 1 is met.

The second assumption states that the vertical spread of the residuals should be constant as we move from left to right. In contrast to our previous regression, this vertical spread is much more consistent, aside from the glaring outlier on the bottom right. We will be looking into that point specifically in the next section on outliers.

Next, we will be looking at the ACF plot:

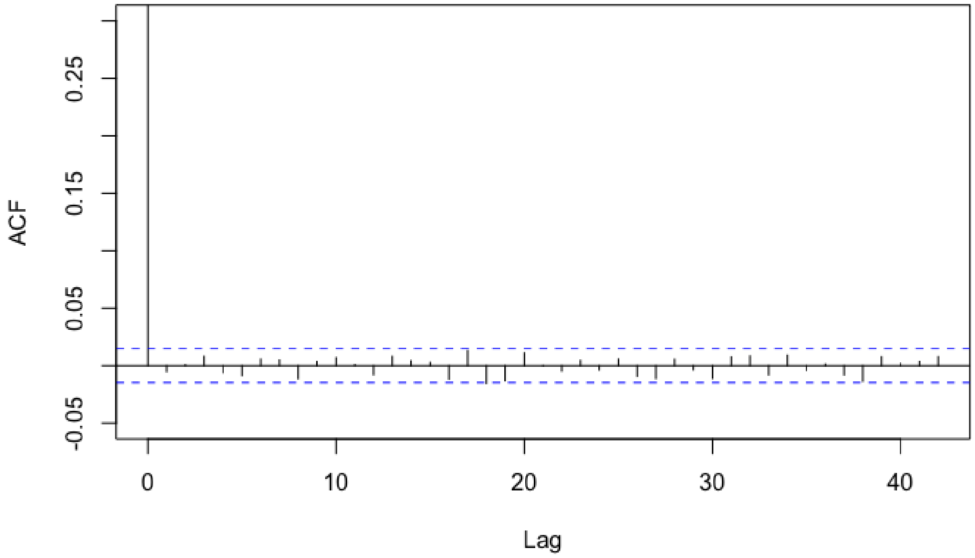


Figure 21: ACF Plot

The ACF at lag 0 is always 1, by definition (the correlation of the vector of residuals with itself). Insignificant ACFs at lag 1 and greater indicate that the residuals are uncorrelated, so we have no evidence that assumption 3 is not met. Since the lags don't go past the dotted line, this informs us that there is no significant correlation in residuals/significant ACF's and more importantly, that the observations within our dataset are independent.

To check our last assumption, we will look at the qqplot of the residuals:

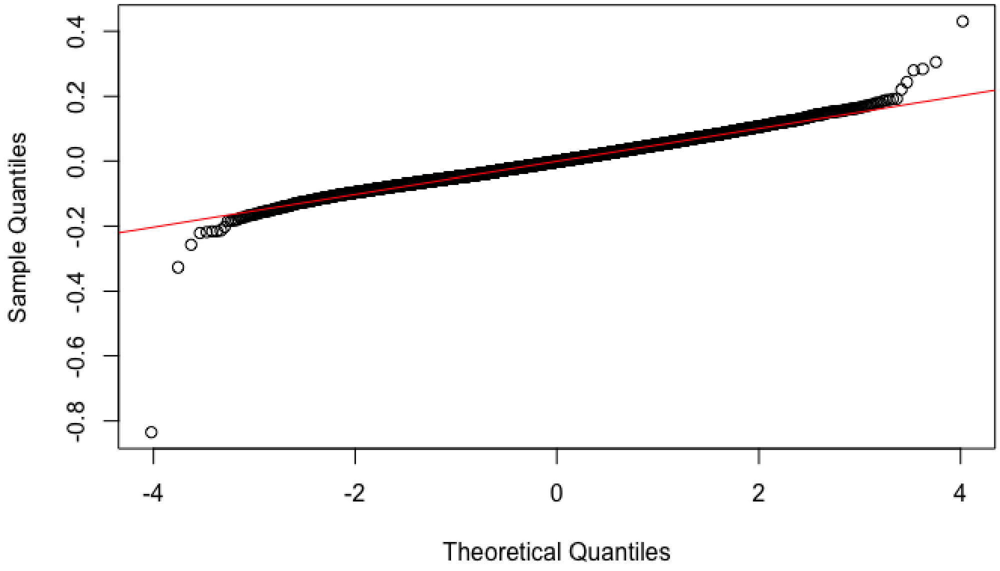


Figure 22: Normal Q-Q Plot

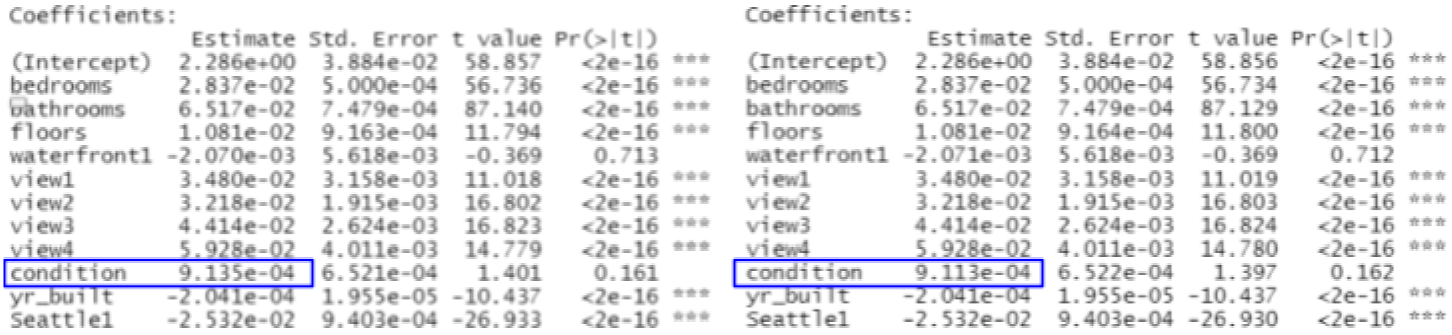
The distribution with a tail will have both the ends of the Q-Q plot to deviate from the straight line and its center follows a straight line. The tail means that compared to the normal distribution there is much more data located at the extremes of the distribution and less data in the center of the distribution. The normality assumption is the least important assumption since regression models are fairly robust to deviations from this assumption, so despite the tails we will leave this assumption checked.

Outliers, High Leverage Points, & Influential Observations

We identified the outlier in the residual plot as a home with 33 bedrooms and approximately 1,600 square feet. We think that this may be a data entry error (3 instead of 33), since the next highest number of bedrooms is 11 with 3,000 square feet, but since we are not sure we have decided to leave it in the data set and see if it is a high leverage point or influential.

We found that there were many observations that were found to be high leverage points. We decided to use Cook's distance to determine if the high leverage points were influential in our data because our question is general and we want our model to work broadly. Since Cook's Distance measures the influence of each observation on all of the fitted values, it will identify observations that influence the overall model broadly. DFFITS is more precise and will identify all observations in which the point is influential on its own predicted values. This is not practical in the case of our model to estimate square footage for a wide range of homes.

Only one point was found to be influential using Cook's Distance with a cutoff of $F_{0.5,p,n-p}$. We removed this point and refitted the model without it, which produced a model with almost identical values as shown below in figure 23. After examining the point of interest, we decided to keep it in the training data set as it did not change the model and there was nothing notable enough to justify removing it.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.286e+00  3.884e-02  58.857   <2e-16 ***
bedrooms     2.837e-02  5.000e-04  56.736   <2e-16 ***
bathrooms    6.517e-02  7.479e-04  87.140   <2e-16 ***
floors       1.081e-02  9.163e-04  11.794   <2e-16 ***
waterfront1 -2.070e-03  5.618e-03  -0.369    0.713    
view1        3.480e-02  3.158e-03  11.018   <2e-16 ***
view2        3.218e-02  1.915e-03  16.802   <2e-16 ***
view3        4.414e-02  2.624e-03  16.823   <2e-16 ***
view4        5.928e-02  4.011e-03  14.779   <2e-16 ***
condition    9.135e-04  6.521e-04   1.401    0.161    
yr_built    -2.041e-04  1.955e-05 -10.437   <2e-16 ***
Seattle1    -2.532e-02  9.403e-04 -26.933   <2e-16 ***
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.286e+00  3.884e-02  58.856   <2e-16 ***
bedrooms     2.837e-02  5.000e-04  56.734   <2e-16 ***
bathrooms    6.517e-02  7.479e-04  87.129   <2e-16 ***
floors       1.081e-02  9.164e-04  11.800   <2e-16 ***
waterfront1 -2.071e-03  5.618e-03  -0.369    0.712    
view1        3.480e-02  3.158e-03  11.019   <2e-16 ***
view2        3.218e-02  1.915e-03  16.803   <2e-16 ***
view3        4.414e-02  2.624e-03  16.824   <2e-16 ***
view4        5.928e-02  4.011e-03  14.780   <2e-16 ***
condition    9.113e-04  6.522e-04   1.397    0.162    
yr_built    -2.041e-04  1.955e-05 -10.437   <2e-16 ***
Seattle1    -2.532e-02  9.403e-04 -26.930   <2e-16 ***
```

Figure 23: Model including influential point (left) and after removing influential point (right). Note the changed coefficient for condition.

Model Predictive Ability

In order to assess how well our model does in predicting *sqft_living*, we have plotted the expected vs. observed values in a scatterplot:

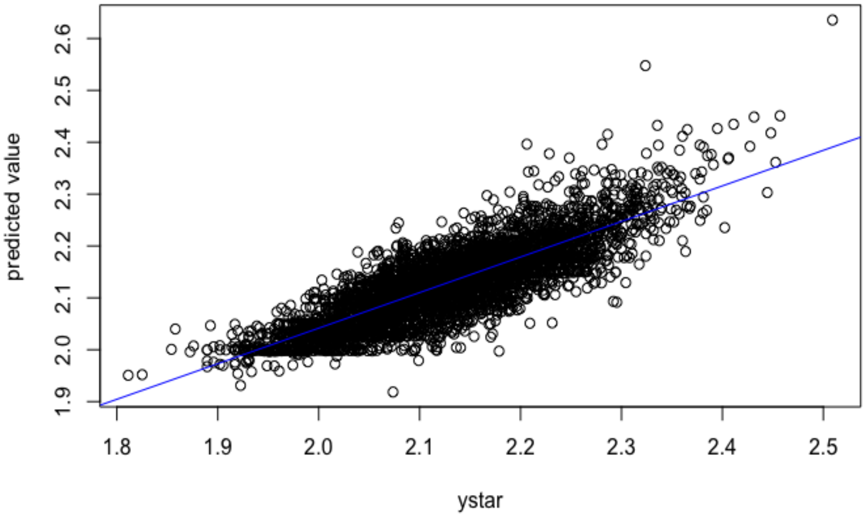


Figure 24: Scatterplot, ystar vs predicted values

The points seem to follow the linear trend in an even manner, which makes this graph quite promising regarding the success of our predictive model. The r-squared of this relationship is 0.693. Therefore 69.3% of the variance in our observed vs predicted values is explained by the regression model, which is a pretty good value.

Model Conclusions

Given that we transformed our y variable, it's more difficult to interpret the estimated coefficients besides positive and negative effects on *sqft_living*. The predictor variables which have a positive effect on *sqft_living* (other predictors remaining constant) are *bedrooms*, *bathrooms*, *floors*, *view 2*, *view 3*, *view 4*, and *condition*. The predictor variables which have a negative effect on *sqft_living* (other predictors remaining constant) are *view 1*, *yr_built*, and homes in *Seattle*. Our model proved to be useful in predicting the square foot of living space for a home as seen in its predictive ability.

Section 4: Logistic Regression

Initial Model

Using logistic regression, our goal is to model and predict whether a home, given certain characteristics, is in Seattle or not (aka located elsewhere in King County, Washington). Our initial model consists of eight parameters hand-selected by our group. Given an extensive list of feature variables to choose from, we first considered the data from a conceptual standpoint, followed by exploratory data analysis of the features chosen. Thus, this initial model includes:

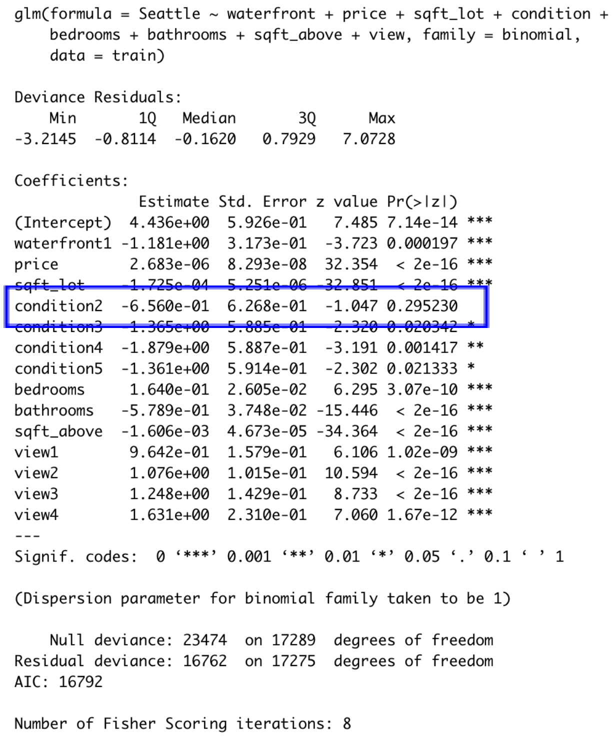
- *waterfront*: whether the apartment was overlooking the waterfront (1) or not (0)
- *price*: price of each home sold ($)
- *sqft_lot*: square footage of the land space
- *condition*: an index from 1 (worst) to 5 (best) on the condition of the apartment
- *bedrooms*: # of bedrooms
- *bathrooms*: # of bathrooms, where .5 accounts for a room with a toilet but no shower
- *view*: an index from 0 (no view) to 4 (excellent view) of how good the view of the property was
- *sqft_above*: the square footage of the interior housing space that is above ground level,

to predict whether the home is located in Seattle (1), or elsewhere in King County (0).

This model was carefully curated, considering general home factors and special characteristics unique to Seattle. For example, the Puget Sound cuts right through Seattle, thus we included *waterfront* in the initial model. Similarly, from a conceptual standpoint we expected the square footage of the home above ground level (*sqft_above*) to be significantly smaller in homes in Seattle than not, due to more apartment complexes (where a "house" is only on one floor in a larger building) likely being in Seattle. Via exploratory data analysis, we found that there was a significant difference in the land space square footage for homes in Seattle compared to those not, thus we included the *sqft_lot* parameter. All model parameters were well thought out through either conceptual understanding of the data, via exploratory data analysis, or both.

Model Improvements

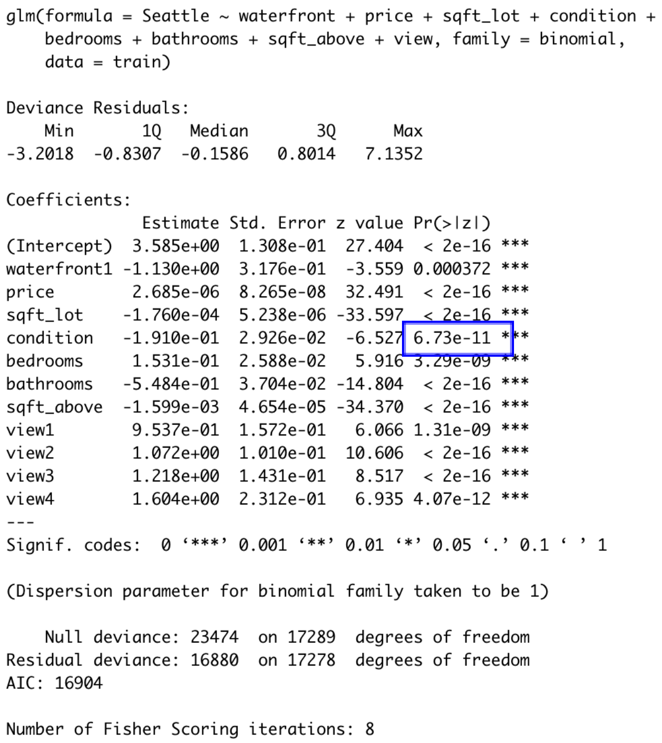
After fitting the initial model, a Wald test was performed on the predictor *condition*.



```
glm(formula = Seattle ~ waterfront + price + sqft_lot + condition +
    bedrooms + bathrooms + sqft_above + view, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2145  -0.8114  -0.1620   0.7929   7.0728

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.436e+00  5.926e-01   7.485 7.14e-14 ***
waterfront1 -1.181e+00  3.173e-01  -3.723 0.000197 ***
price        2.683e-06  8.293e-08  32.354  < 2e-16 ***
sqft_lot    -1.725e-04  5.251e-06 -32.851  < 2e-16 ***
condition2  -6.560e-01  6.268e-01  -1.047 0.295230
condition3  -1.365e+00  5.885e-01  -2.320 0.020342 *
condition4  -1.879e+00  5.887e-01  -3.191 0.001417 **
condition5  -1.361e+00  5.914e-01  -2.302 0.021333 *
bedrooms     1.640e-01  2.605e-02   6.295 3.07e-10 ***
bathrooms   -5.789e-01  3.748e-02 -15.446  < 2e-16 ***
sqft_above  -1.606e-03  4.673e-05 -34.364  < 2e-16 ***
view1        9.642e-01  1.579e-01   6.106 1.02e-09 ***
view2        1.076e+00  1.015e-01  10.594  < 2e-16 ***
view3        1.248e+00  1.429e-01   8.733  < 2e-16 ***
view4        1.631e+00  2.310e-01   7.060 1.67e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23474  on 17289  degrees of freedom
Residual deviance: 16762  on 17275  degrees of freedom
AIC: 16792

Number of Fisher Scoring iterations: 8
```

Figure 25: summary output of our initial model

As we can see, one of the p-values for this predictor is larger than our alpha (α) value of 0.05. This value is telling us that there is not a significant difference between Condition 1 and 2, so we will treat our predictor, *condition*, as a quantitative variable moving forwards.



```
glm(formula = Seattle ~ waterfront + price + sqft_lot + condition +
    bedrooms + bathrooms + sqft_above + view, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2018  -0.8307  -0.1586   0.8014   7.1352

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.585e+00  1.308e-01  27.404  < 2e-16 ***
waterfront1 -1.130e+00  3.176e-01  -3.559 0.000372 ***
price        2.685e-06  8.265e-08  32.491  < 2e-16 ***
sqft_lot    -1.760e-04  5.238e-06 -33.597  < 2e-16 ***
condition   -1.910e-01  2.926e-02  -6.527 6.73e-11 ***
bedrooms     1.531e-01  2.588e-02   5.916 3.29e-09 ***
bathrooms   -5.484e-01  3.704e-02 -14.804  < 2e-16 ***
sqft_above  -1.599e-03  4.654e-05 -34.370  < 2e-16 ***
view1        9.537e-01  1.572e-01   6.066 1.31e-09 ***
view2        1.072e+00  1.010e-01  10.606  < 2e-16 ***
view3        1.218e+00  1.431e-01   8.517  < 2e-16 ***
view4        1.604e+00  2.312e-01   6.935 4.07e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23474  on 17289  degrees of freedom
Residual deviance: 16880  on 17278  degrees of freedom
AIC: 16904

Number of Fisher Scoring iterations: 8
```

Figure 26: summary output of initial model with *condition* as a qualitative parameter

After refitting the initial model with *condition* as a numeric variable, the p-value is small, less than our alpha (α) value of 0.05. Thus, we can reject the null hypothesis (H_0) for the Wald test, and conclude we should keep the *condition* score of the home in the model. Although both values have significant p-values, we next wanted to explore whether the predictors *waterfront* and *view* were correlated.

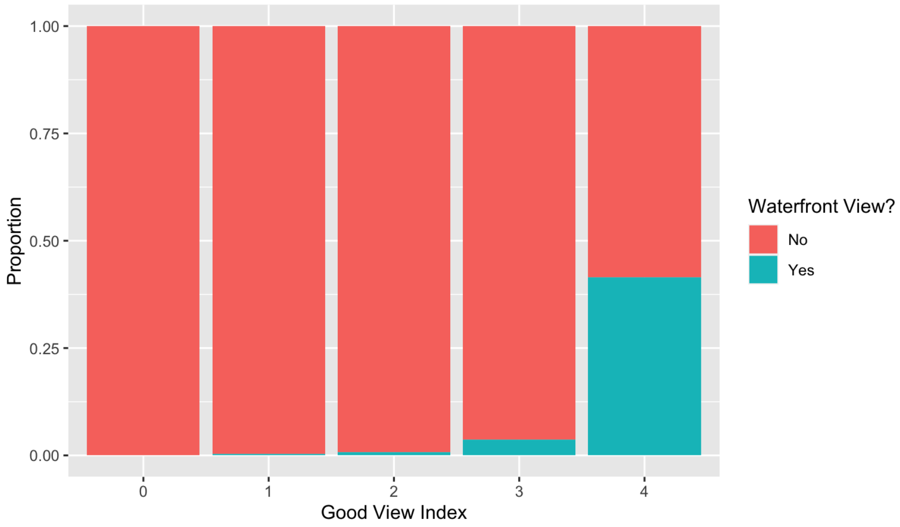


Figure 27: proportion of *waterfront* homes by *view* index

Looking at figure 27, it is highly probable that the Good View Index and Waterfront view are correlated - thus providing similar information (especially in Good View Index 0, 1, and 2) - and could potentially result in a multicollinearity issue in our model. To test this, a Likelihood Ratio Test was performed. For this test, we chose to not include *view* in the reduced model as, again, it may have a relationship with the other categorical predictor, *waterfront*. Our null hypothesis states that the coefficients pertaining to the predictor view are not significant (H_0: $\beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$), whereas the alternative states at least one coefficient is significant (H_a: at least one $\neq 0$). Via the Likelihood Ratio Test, our ΔG^2 statistic, which measures the difference in loglikelihoods of the initial and reduced models, equaled 6593.96. After comparing this value to a χ^2 statistic with one degree of freedom, our p-value is 0. Thus, we reject the null hypothesis, and prefer the initial, eight-predictor model including the *view* parameter, over the reduced predictor model. Based on these analyses, we will carry on with the full model.

Our estimated regression equation for the logistic regression model is:

$$log(\frac{\hat{\pi}}{1-\hat{\pi}}) = 3.585 - 1.130x_1 + 0.000002685x_2 - 0.0001760x_3 - 0.1910x_4 + 0.1531x_5$$
$$- 0.5484x_6 - 0.001599x_7 + 0.9537x_8 + 1.072x_9 + 1.218x_{10} + 1.604x_{11}$$

where,

x_1 : 1 for indicating a *waterfront* view, 0 otherwise

x_2 : *price*

x_3: *sqft_lot*

x_4: *condition*

x_5: *bedrooms*

x_6: *bathrooms*

x_7: *sqft_above*

x_8 : 1 for a *view* index of 1, 0 otherwise

x_9 : 1 for a *view* index of 2, 0 otherwise

x_{10} : 1 for a *view* index of 3, 0 otherwise

x_{11} : 1 for a *view* index of 4, 0 otherwise

Base case: for a *view* index of 0: $x_8 = x_9 = x_{10} = x_{11} = 0$

In terms of the relationship between predictors and log odds of a home being in Seattle, holding all other predictors constant, the estimated log odds of a house being located in Seattle is affected the most by whether or not the home is *waterfront*. Looking at our regression equation and coefficients, the estimated odds of a waterfront home to be located in Seattle is 0.323 times the odds of a home that is not waterfront, while controlling for the other seven predictors.

Model Predictive Ability

The first test to see how well our model does in classifying whether a home is in Seattle or elsewhere in King County is by plotting a Receiver Operating Characteristic (ROC) curve. The ROC curve plots the sensitivity for every value of the threshold. It is important to note that the ROC curve uses our model, but predicts on the test portion of the split dataset.

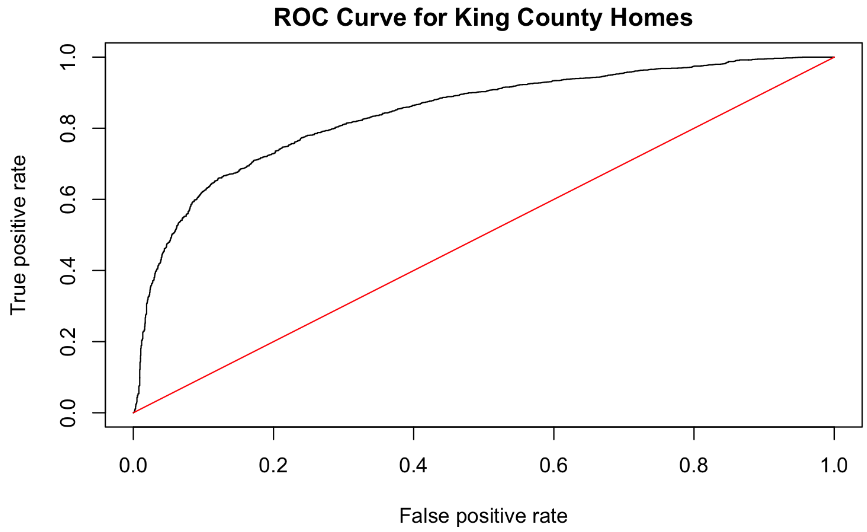
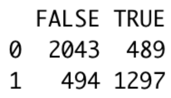


Figure 28: ROC Curve for King County Homes

Our ROC curve falls above the diagonal red line, thus performing better than random guessing. The second measure of classification ability is the Area Under the ROC Curve (AUC). With a value close to 1 the AUC of 0.8420609 means the logistic regression performs much better than random guessing. To further this analysis, a confusion matrix (Figure 29) was created to compare the predicted and actual values of our model.



```
  FALSE TRUE
0  2043  489
1   494 1297
```

Figure 29: Confusion Matrix

To preface, regarding this particular question we are not necessarily too worried about false positives over false negatives or vice versa, thus will begin with a cutoff of 0.5. The accuracy of our model is fairly high, at 0.7726116. Both the false positive rate and false negative rate were fairly low, equaling 0.193128 and 0.2758236, respectively. The sensitivity = 0.7241764 and specificity = 0.806872. Comparing our False Positive Rate (of 0.193128) and the ROC curve, it is apparent that the logistic regression at the threshold cutoff of 0.5 is significantly better than random guessing.

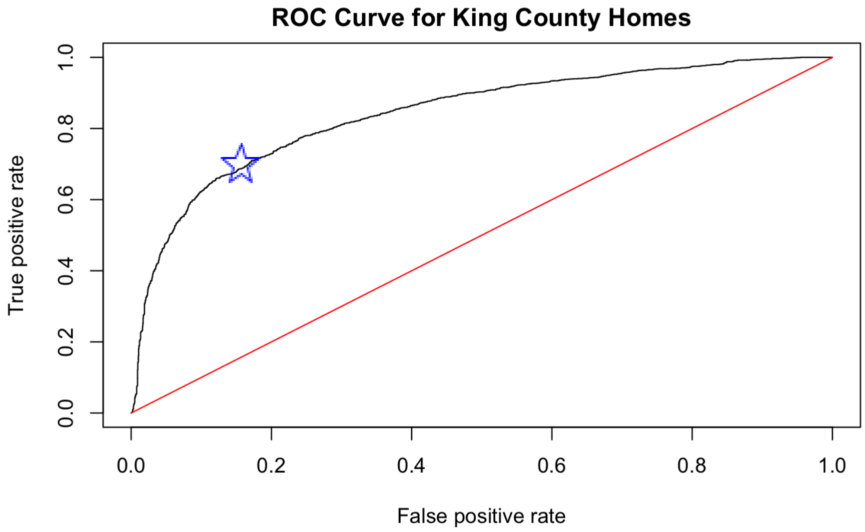


Figure 30: ROC Curve and Validation

Model Interpretability and Application

To further complete this analysis, we would like to determine the odds, log odds, and prediction of a specific home. For the sake of example, say we are a realty company. Our client is looking for her dream home. Ideally, she wants to live in Seattle. Given her preferences, what are the log odds the house she's looking for is in Seattle? We can use our model to answer this question. Her wish list:

- A *waterfront* home
- An ideal *price* of $660,000
- At least a small yard for her dog to play in, so an approximate *sqft_lot* size of 2,000 square feet
- A home in good *condition*
- 2 *bedrooms* (one for her, one for guests)
- 2 full *bathrooms*
- Roughly 1,000 square feet above ground level (*sqft_above*)
- An excellent *view*

After adding this new data and predicting using our model, the log odds of the home she's looking for being in Seattle is 2.324848. The corresponding odds is 10.22512, and the corresponding probability is 0.9109141!