SCHOOL *of* DATA SCIENCE

Group 17

STAT 6021: Project 2

Rachel Holman, Serene Lu, Taryn Trimble

Table of Contents

1 Summary of Findings

Are you a prospective house buyer who wants to be able to get the best value on their house? Or perhaps a contractor who wants to receive the best grade for the houses you construct? Researchers at the University of Virginia have conducted a comprehensive analysis of over a 20 thousand houses sold in King County, Washington from 2014 to 2015 to evaluate several different metrics that can measure and predict the value and grade of these houses. This research has shown various interesting trends that can help you get the very most out of your home!

For those of you price-contentious buyers in King County, it has been shown that the grade of the house typically has the greatest impact on the price of a house. In fact, for every one unit increase in grade, the price of the house increased by approximately 26%. The grade is an indicator of how well the house was constructed and designed. This statistic is a comprehensive index of how well-built the windows and roof were, how high quality the building material and design were, and even considers the location the house is located in – all of which make it a reliable characteristic to predict house prices by. A house with a high grade is definitely a house you want to live in! The number of floors in a house also appeared to have a significant impact on the price of a house with an approximately 11.71% increase for every additional floor. This is interesting considering the number of square feet of the house above ground had an extremely minimal impact on the price. It appears that, at least in King County, how well a house was built is more important in terms of pricing than how large it is.

As a contractor focused on earning an above average grade on construction and design of a house, it is important to note that houses not overlooking waterfronts were about 10 times more likely to have above average grades than houses near water! It also seems that open floor plans are the latest craze as each additional closed-off bedroom in a house decreased the odds of an above average construction and design grade by nearly 40%! However, this open floor plan mindset does not apply to bathrooms so remember that the more bathrooms the better! Finally, keep in mind that renovations did not have any notable impact on whether a house earned an above or below average grade for construction and design, so rather than focusing on revamping an old space, turn your attention toward finding spaces with the best views!

2 Data Description

The data set we worked with contains house sale prices for King County, Washington, which includes Seattle. It includes homes sold between May 2014 and May 2015. This data set is available to download from kaggle.com. More information on the variables in the data set shown below:

Variable	Type	Description
id	Double	Unique ID for each home sold
date	Character	Date of the home sale
price	Double	Price of each home sold
bedrooms	Integer	Number of bedrooms
bathrooms	Double	Number of bathrooms, where .5 accounts for a room with a toilet but no shower
sqft_living	Integer	Square footage of the apartments interior living space
sqft_lot	Integer	Square footage of the land space
floors	Double	Number of floors
waterfront	Integer	A dummy variable for whether the apartment was overlooking the waterfront or not
view	Integer	An index from 0 to 4 of how good the view of the property was
condition	Integer	An index from 1 to 5 on the condition of the apartment
grade	Integer	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design
sqft_above	Integer	The square footage of the interior housing space that is above ground level
sqft_basement	Integer	The square footage of the interior housing space that is below ground level
yr_built	Integer	The year the house was initially built
yr_renovated	Integer	The year of the house's last renovation
zipcode	Integer	What zipcode area the house is in
lat	Double	Latitude
long	Double	Longitude
sqft_living15	Integer	The square footage of interior housing living space for the nearest 15 neighbors
sqft_lot15	Integer	The square footage of the land lots of the nearest 15 neighbors

In addition to the variables proved in the data, we also created our own variables outlined below:

Variable	Type	Description
hightGrade	Integer	A 0/1 dummy variable for whether a home have above average grade or not (grade >7 or not)
highGradeFtr	Character	A yes/no dummy variable for whether a home have above average grade or not (grade >7 or not)
renovated	Integer	A 0/1 dummy variable for whether a home has been renovated or not
waterfront	Character	A yes/no dummy variable for whether the apartment was overlooking the waterfront or not
houseAge	Integer	2023 - yr_built
final_date	Date	Date of each home sale as a date value

We chose to treat the variables view and condition as numeric values rather than converting them into categorical predictors. In doing so, we assumed that there is an equal change between each level of view and between each ranking for condition. In other words, we interpreted a one unit change in each of these variables as constant for the ease of modeling.

We also chose to remove the id, date, lat, long, sqft_above, sqft_living15, sqft_lot15, yr_renovated, and yr_built variables from our data when modeling to avoid blatant multicollinarity, hard to interpret variables, and variables that did not apply to our questions of interest.

Additionally, we split the full data into a training data set with 70% of the information, and a test set with the remaining 30% to evaluate predictive power.

3 Questions of Interest

We placed ourselves in the shoes of people that we thought would most likely be able to use the data and apply it to their lives. We found that buyers, architects, constructors, real estate agents, etc. would find this data useful, but we ultimately decided that we could best help people who were looking for houses to buy and contractors building the houses in the King County area. We believed that regression would be able to help us better understand the associations between variables in our data set and even predict qualities about these houses based on other variables; our questions of interest are as follows.

Linear Regression: Which predictors make up the most accurate model to predict the housing price?

Thinking as someone who is interested in buying a house in the King County area, we would probably be most curious about the price of the houses and how the price is affected by other variables. Thus, we assigned the price of the house to be the response variable. Of course, we had our theories about which predictors would be the most influential on price, but creating a statistical model would allow us to truly see the importance of each of these variables and how they compared to one another. This model would also be able to help us predict the price of a house based on specific conditions. Thus, we believed multiple linear regression would best be able to achieve our goals.

Logistic Regression: Can we predict with relatively high accuracy the odds of a house being rated as having above average construction and design (grade)?

Thinking as a contractor who constructs houses in the King County area and is interested in figuring out how to improve the construction and design of the houses they build to achieve a higher grade, a logistic regression would be most useful. We created a new binary response variable to describe whether an observation had an above or below average grade. We were interested in figuring out which variables had the greatest impact on the grade in order for contractors to be more

aware of them and adjust their construction or design strategies accordingly. A logistic model allows us to figure out the association between the predictors and grade variables, and be able to predict whether a house would receive an above or below average grade based on specific conditions.

4 Visualizations for Linear Regression

To begin exploratory data analysis, we were initially curious to see how strongly correlated each of the predictor variables were with our response variable, the price of the houses.

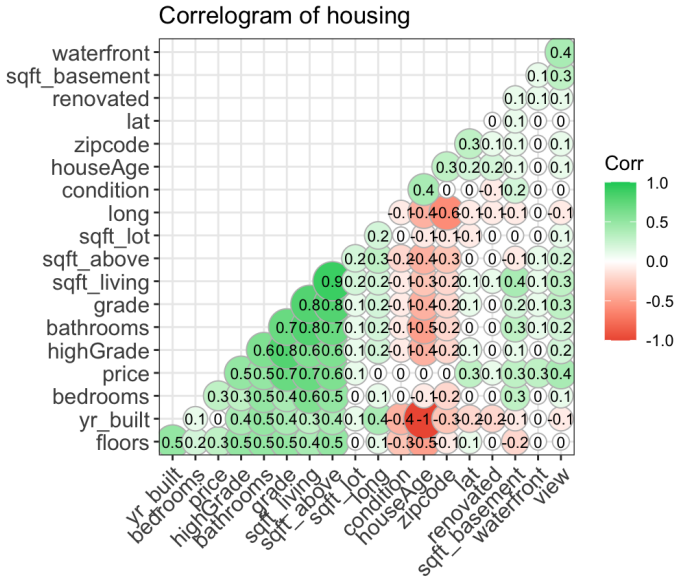

Figure 1: Correlogram of Training Data

From this correlogram, we found the predictors that had the strongest relationships with price were the grade and square foot living variables which had a correlation coefficient of 0.7, followed by the square foot above variable with a coefficient of 0.6, and the high grade and bathroom variables with coefficients of 0.5. This visualization gave us a basis of how to continue exploring these variables.

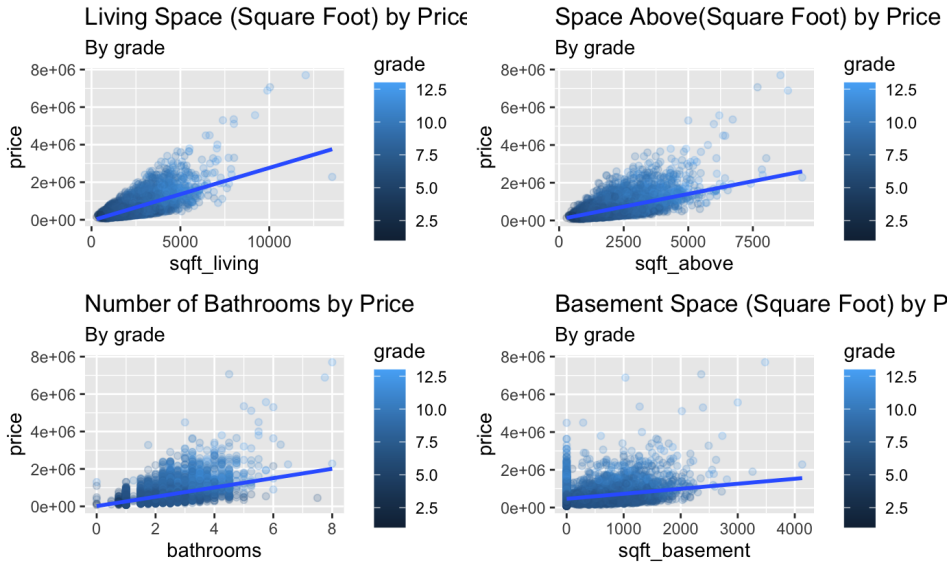

Figure 2: Scatterplots of Highly Correlated Data

Using the identified variables from the correlogram, we created a series of scatterplots to view the relationship between these variables and price by the grade variable. These scatterplots showed a relatively strong linear relationship between living space v. price and space above v. price. These predictors seem to have very similar relationships to price because of the fact that the value of the space above variable contributes to the value of the living space. There were also more moderate linear relationships between the number of bathrooms v. price and the basement space v. price. This indicates that these variables might have some sort of impact on how houses are priced. This agreed with our correlation matrix. We also noticed that the grade generally increased as the predictor variables increased (gradient from darker blue to lighter blue moving from left to right and from bottom to top). This indicates that the grade might also have some sort of effect on the pricing of the houses.

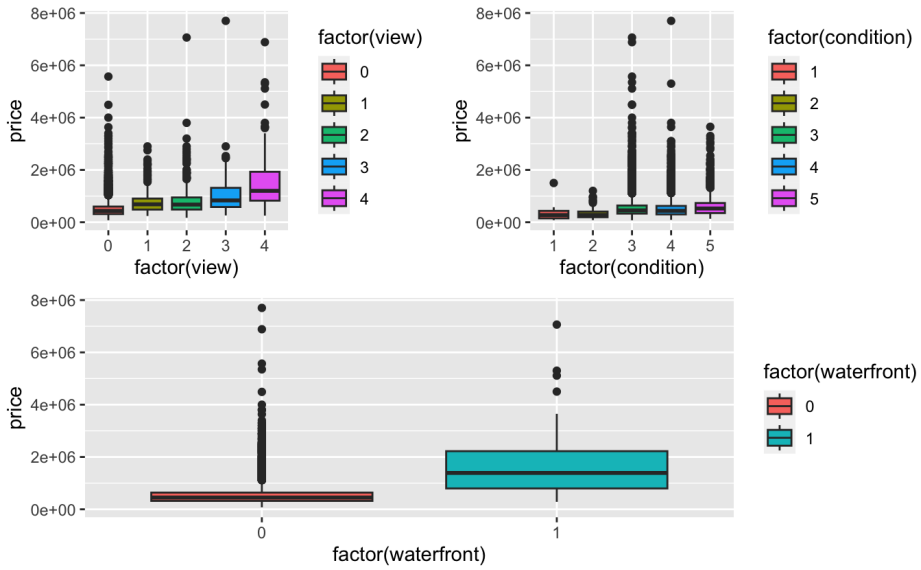

Figure 3: Boxplots of Categorical Variables

We also wanted to see the relationship between the categorical predictors and price. We decided to create boxplots to display these relationships. The first boxplot shows that the houses with the best view tend to have a significantly higher prices. We saw that there was quite a clear linear relationship between these variables. The second graph shows a similar relationship between price and the condition of the house. It appears that those houses in the best condition have the highest prices although the houses at the top of this range do not have as significant a difference in price. There are also many outliers in these plots which might suggest that condition is not the most reliable variable to predict housing price on. The bottom plot shows that there is quite a significant difference between houses located on a waterfront versus elsewhere. Those built on the waterfront typically have a higher price which makes sense, since waterfront views are typically the most highly valued in the housing market.

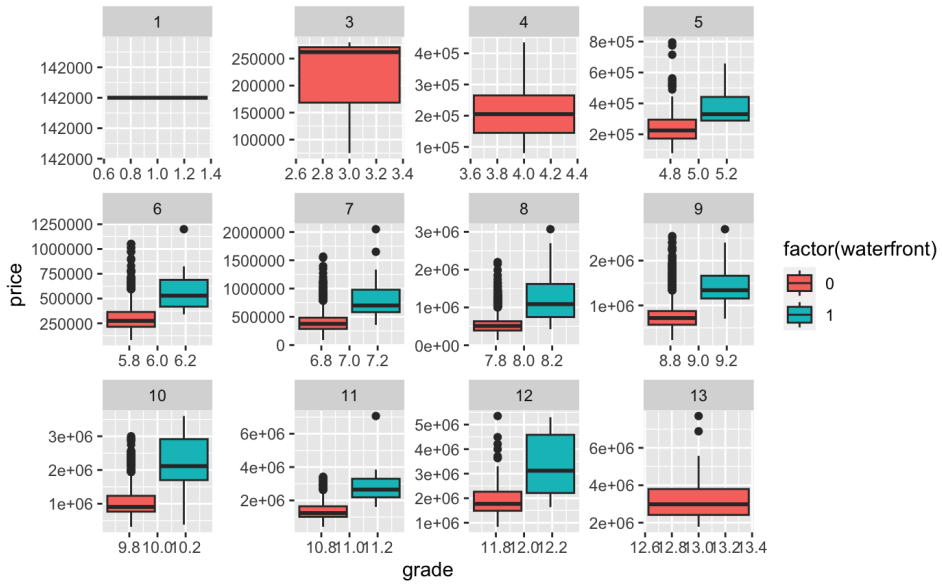

Figure 4: Boxplots of Waterfront Variable by Grade

Taking a closer look at this waterfront variable, we found that separating this variable into separate boxplots based on grade would help us narrow in on which houses in particular had the highest prices. From this plot, we found that houses with a grade of 12 on the waterfront had the highest prices of all other houses, followed by houses with a grade of 13 that were not on the waterfront and houses with a grade of 11 with a waterfront view. These are characteristics of houses that might be of interest when buyers are looking.

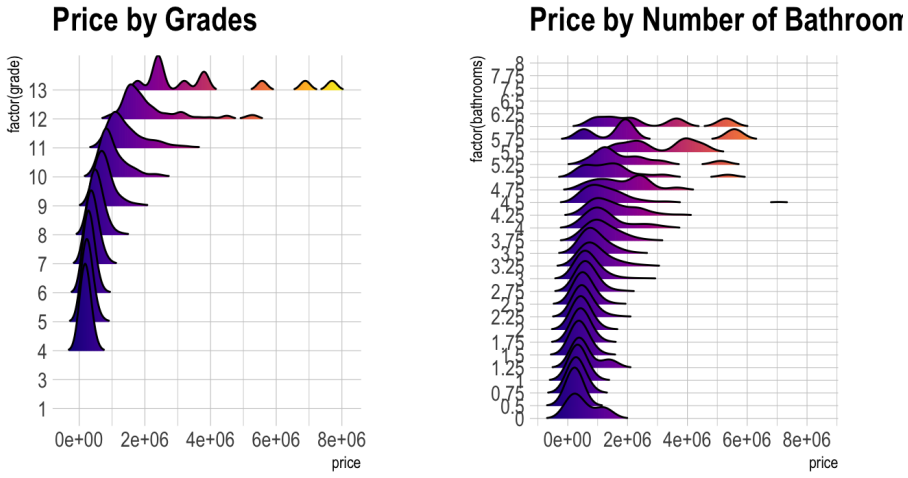

Figure 5: Ridgeline Plots of Grade and Bathrooms

Lastly, these ridgeline plots highlighted the individual relationships between price v. grade and

price v. number of bathrooms. The graph on the left shows that there is a larger range of prices for houses with higher grades. These grades refer to the construction and design of the houses which explains the lack of points in the lower grade range. Buyers do not want houses with an extremely low build grade, and buyers that have a more extensive budget are more willing to spend more extreme amounts of money to purchase houses that are assigned higher grades. The graph on the right also shows a similar story, with the number of bathrooms increasing dramatically at higher price ranges. It seems that most houses have about the same price for an increasing number of bathrooms until about 4.5 bathrooms and beyond.

5 Linear Regression

5.1 Initial Linear Model

The question we wanted to answer by linear regression was: Which predictors make up the most efficient model to predict the housing price? Of the 21 predictors we were originally given, bedrooms, bathrooms, sqft_lot, floors, condition, grade, sqft_above, sqft_basement, renovated, and houseAge were used initially for modeling. These variables were decided contextually to be significant predictors that would structure the first linear regression model. There were other variables such as id, date (sold), waterfront, view, and zip code that were decided to be less contextually significant for the model. We decided to start with the most basic predictors that were more likely to have a significance. Other variables that were calculated by existing variables were renovated and houseAge, which we found were more interpretable and quantifyable than yr_renovated and yr_built. Once we established our initial predictor variables, we were interested to see how accurate this model was at determining the price of a house.

```
Call:
lm(formula = price ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1118845  -115880   -10953    89542  4386299

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.098e+06  2.174e+04 -50.499  < 2e-16 ***
bedrooms      -4.567e+04  2.481e+03 -18.406  < 2e-16 ***
bathrooms      4.710e+04  4.358e+03  10.807  < 2e-16 ***
sqft_lot      -2.182e-01  4.696e-02  -4.646 3.41e-06 ***
floors         2.838e+04  4.690e+03   6.050 1.48e-09 ***
condition      2.083e+04  3.137e+03   6.642 3.20e-11 ***
grade          1.290e+05  2.689e+03  47.958  < 2e-16 ***
sqft_above     1.848e+02  4.349e+00  42.503  < 2e-16 ***
sqft_basement  2.007e+02  5.657e+00  35.477  < 2e-16 ***
renovated      4.185e+04  9.682e+03   4.323 1.55e-05 ***
houseAge       3.929e+03  8.877e+01  44.260  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 227500 on 15118 degrees of freedom
Multiple R-squared:  0.6183,    Adjusted R-squared:  0.6181
F-statistic:  2449 on 10 and 15118 DF,  p-value: < 2.2e-16
```

Figure 6: Initial linear regression model for predicting house prices.

Above is a summary of the first linear regression model, this allowed us to decide which predictors

needed to be removed and how significant they were in predicting house price. All of the p-values showed statistical significance but the largest p-values were for sqft_lot, renovated, and floors. Although the values were larger, they were still considered significant. Thus, there was not enough evidence to conclude that removing one or all of the variables was necessary. Other values that was noted were the $R^2 = 0.6183$ and the Adjusted $R^2 = 0.6181$, this shows that when accounting for multicollinearity, about 61.81% of variability in price was explained by our model. This showed that our model was moderately effective without any further change.

5.2 Model Improvements

Our next step was to plot the residual plots to see if the response variable and/or predictors needed to be transformed. This is important because it helped us recognize any outliers or highly leveraged data points that may have been skewing our model.

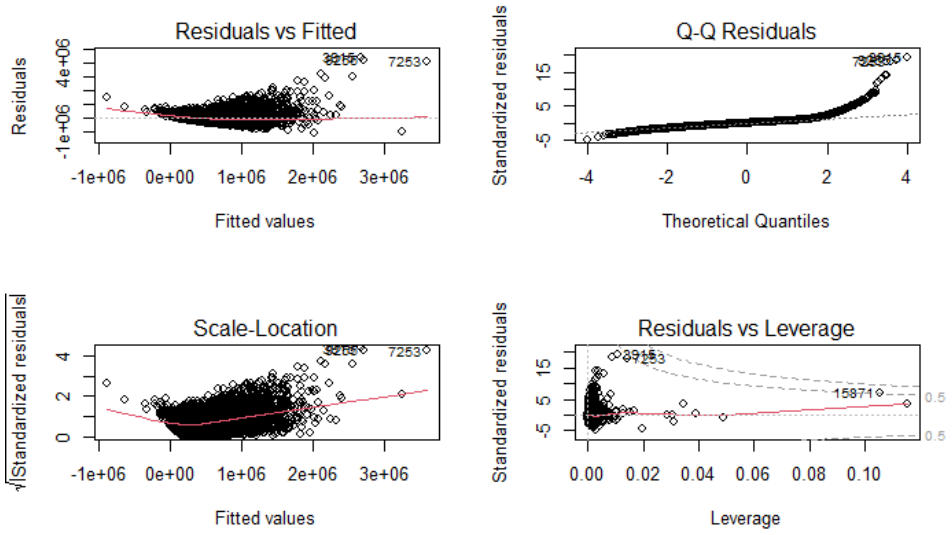

Figure 7: Initial residual plot

The initial residual plots showed that assumption #2 was not met due to failure to have constant variance throughout the points. However, it did appear that assumption #1 was somewhat met due to the points having a linear trend. Next, it was necessary to plot the Box Cox plot to find the optimum lambda value to properly transform the response variable. If the Box Cox plot showed a lambda variable with zero in the interval (or close to zero) then we would use the log transformation.

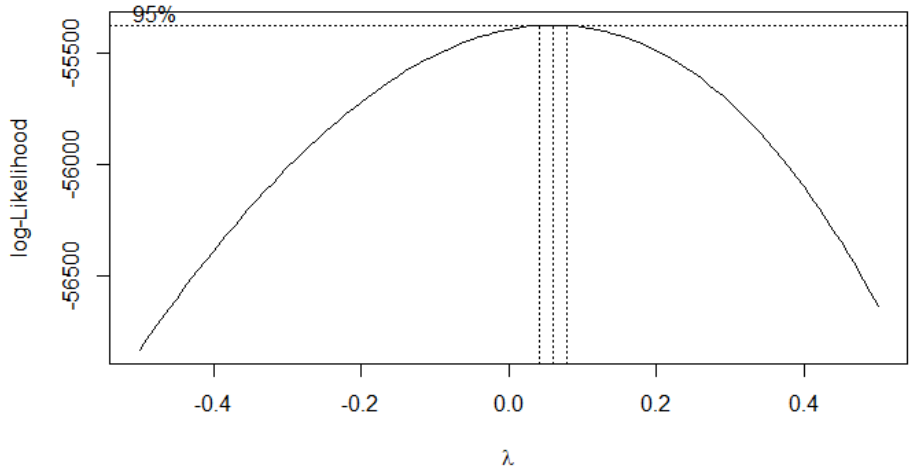

Figure 8: Box Cox plot to find the ideal lambda value

Since the lambda value was close to zero, we decided to use log transformation for the response variable. Once we applied the appropriate methods, we re-plotted the residual plots and decided how to improve the model next.

5.2.1 Log Transformation

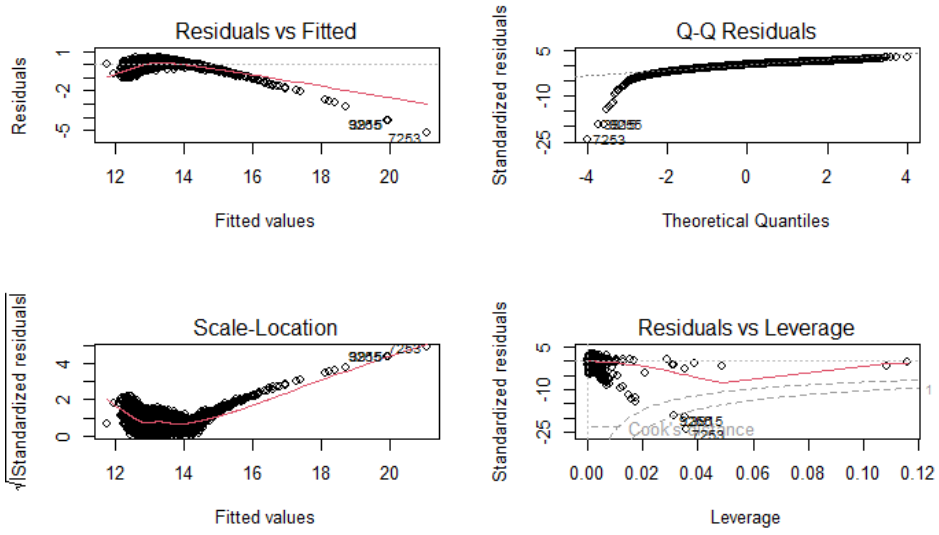

Figure 9: Residual plot after log transformation of response variable

Now assumption #2 was met, but assumption #1 was no longer met. We concluded that after the log transformation of the predictor variable we would focus on removing any outliers and insignificant predictor variables. The Residuals v. Leverage plot supported this idea, leading us to hypothesize that the removal of outlying points would improve our residual plot.

Because we could not run all the partial residual plots at once to assess the relationship between predictors against the response, we opted to split our model to evaluate the six predictors with the lowest p-values first.. This allowed us to more closely consider which of these variables needed to be removed. A horizontal trend line is considered to be insignificant within a partial residual plot.

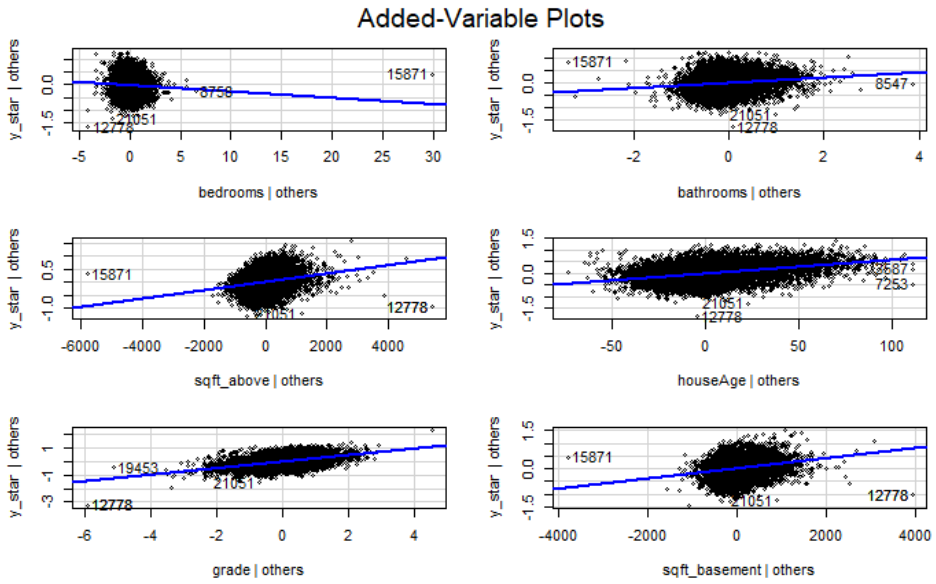

Figure 10: Partial residual plots of predictors we assume would be more significant. There were 6 variables we expected to have more influence on predicting the housing price.

Looking at the partial regression model above showed us that all of the predictors with the lowest p-values have a somewhat linear relationships to the log(price) of houses. The number of bedrooms had a surprisingly negative relationship with price, which we intended to look into further. However, the overall appearance of the plots agreed with our hypothesis that these 6 variables should be kept in the model. The p-values of these predictors were also significant in the first linear regression model summary output which contributed to our conclusion.

We then moved on to the variables we assumed were less significant to the log(price) and looked at the partial regression plots for these predictors based on the p-values.

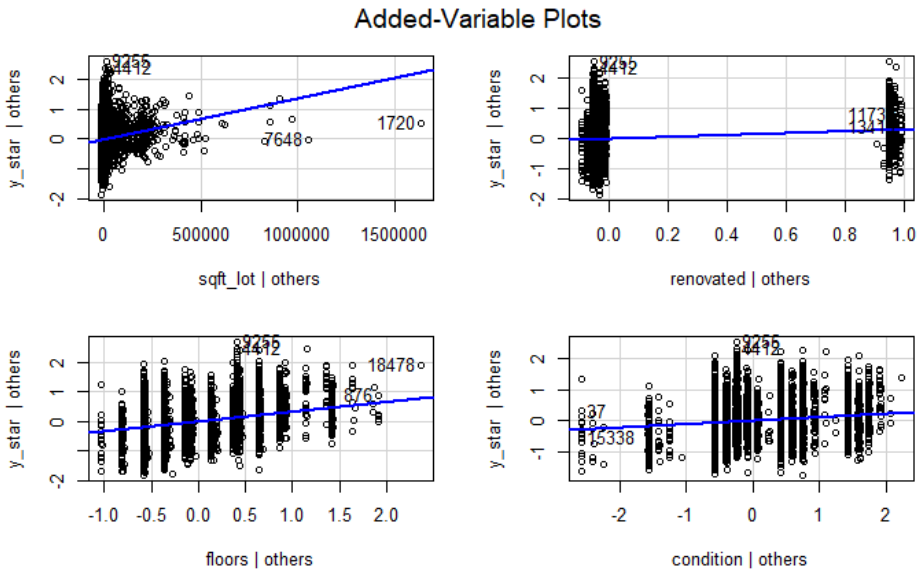

Figure 11: Partial residual plots we assumed would be less significant. There were 4 variables we expected to have less influence compared to the other 6 predictors.

The partial residual plots for sqft_lot and floors appeared to have significant linear relationships to the log(price). This is because the slope of the line was non-zero, meaning the variables were linearly associated. On the other hand, the slope for renovated and condition were less severe and showed there was a weaker linear correlation between log(price) and the two predictors. In our initial linear model we found that these two had the largest p-values; this method supported our decision to remove them and continue with an 8 variable linear regression model.

5.2.2 Removing Insignificant Predictors

After removing the renovated and condition variables we wanted to run a VIF test to quickly identify if there was a high degree of multicollinearity within the model.

bedrooms	bathrooms	floors	condition	grade	sqft_above	sqft_basement	houseAge
1.664832	3.283488	1.854635	1.184401	2.928643	3.736872	1.814899	1.763094

Figure 12: VIF output for 8 predictor model

A VIF value of 10 or larger indicates very significant degrees of multicollinearity, but since all of the VIF values in this model are well below that criteria, we are assured that there is no concerning multicollinearity. We shifted our focus from the predictors to outlying points by looking at the leverage graph in our diagnostic plots below.

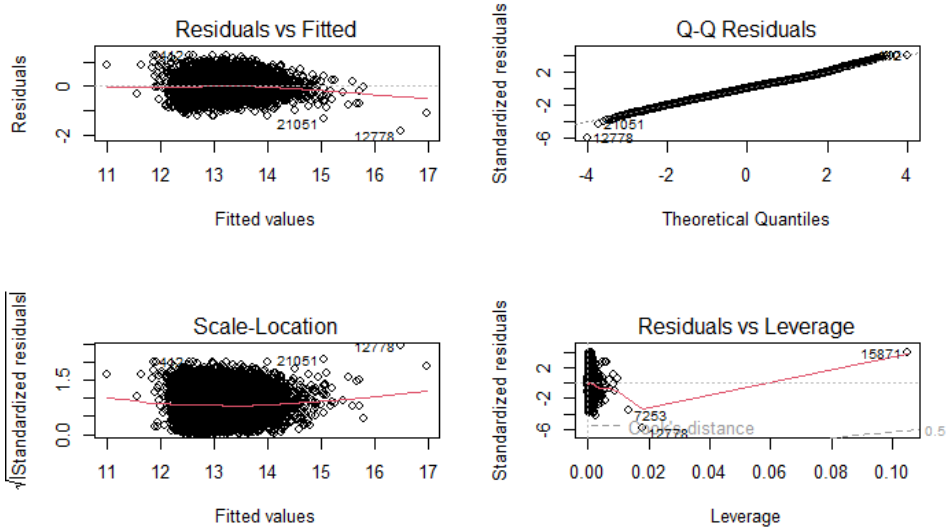

Figure 13: Residual plot with 8 predictor variables

After removing two more variables, our residual plot looked quite different than it had in the earlier section. We had satisfied both assumptions #1 and #2.

Looking at the bottom right hand graph above indicated that there was a high leverage outlier within our data. This instance was identified as the 15871st observation in our training data set which mapped to a house that had 33 bedrooms recorded. We could see that it was an entry error because the 33 bedroom house only had 1.75 bathrooms and a relatively small number of square feet of living space. Rather than correcting this value, we decided to remove the data point then re-plot the diagnostics plots to confirm there were no other influential outliers.

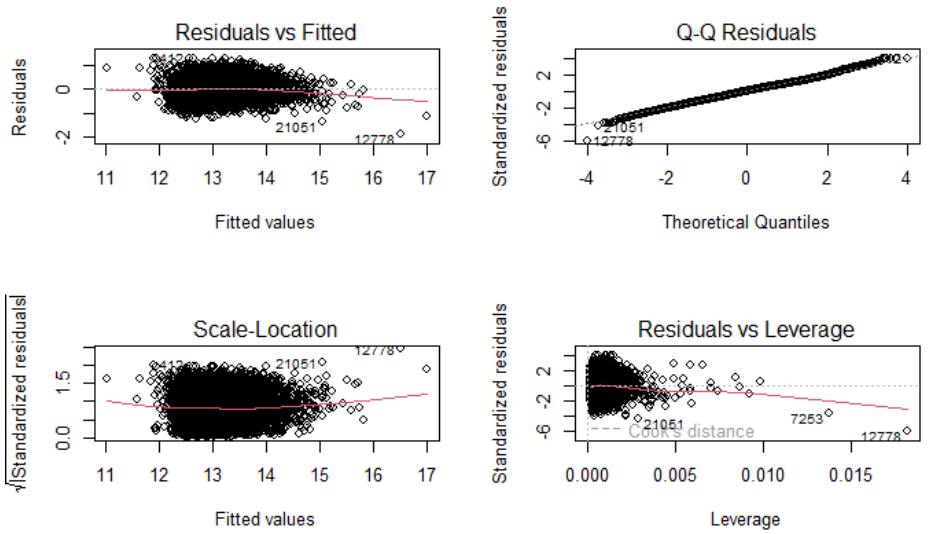

Figure 14: Final linear residual model

Our Residuals vs Leverage model looks significantly better but we wanted to investigate one more data point that was flagged in all four of our plots. We decided to use Cook's Distance on observation 12778 to see if this point was highly influential.

5.2.3 Cook's Distance

Cook's Distance is the estimation of influence of one data point on the response variable price. It takes into account the leverage and the residual of the instance. It is also known to be the summary of how much a regression model changes when that instance is removed.

```
cooks.distance(result)[which.max(cooks.distance(result))]
```

Figure 15: R code to find the max Cook's Distance, which we knew instance 12778 would have.

Cook's Distance of observation 12778: 0.0735863. Since the Cook's distance value is less than 1, we concluded that it is not significantly influential and does not need to be removed from the data.

5.3 Final Model

In order to determine if our linear regression model was useful at predicting the log(price), we set up a hypothesis test as outlined below.

H_o : $\beta_{bedrooms} = \beta_{bathrooms} = \beta_{floors} = \beta_{condition} = \beta_{grade} = \beta_{sqft_above} = \beta_{sqft_basement} = \beta_{houseAge} = 0$
H_a : At least one coefficient in is nonzero, meaning the full model is useful for estimating the $log(price)$ of a house.

Since the p-value is essentially zero ($2.2 \cdot 10^{-16}$), we reject the null hypothesis and conclude that the full model is useful for estimating the $log(price)$ of the house. The F-stat was also significantly larger than the critical value: $3326 > 1.027116$.

5.4 Assessing Predictive Capabilities of the Model

To assess the predictive capabilities of the model we calculated the mean squared error (MSE) using the testing data as well as the equation $MSE = \sum(y_i - \hat{y})^2$. MSE is used to check how close estimates are to the actual value. In other words, the lower the score the closer the predicted value is to the actual. We found that our MSE score is 0.4500993 which means that our model's predictions are relatively close to the actual values.

5.5 Model Interpretations

```
Call:
lm(formula = y_star ~ bedrooms + bathrooms + floors + condition +
    grade + sqft_above + sqft_basement + houseAge, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.87713 -0.21299  0.01378  0.21040  1.26743

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.025e+01  3.031e-02 338.102   <2e-16 ***
bedrooms      -3.041e-02  3.630e-03  -8.377   <2e-16 ***
bathrooms      8.211e-02  6.025e-03  13.629   <2e-16 ***
floors         1.059e-01  6.481e-03  16.337   <2e-16 ***
condition      3.898e-02  4.305e-03   9.053   <2e-16 ***
grade          2.341e-01  3.740e-03  62.595   <2e-16 ***
sqft_above     1.544e-04  5.966e-06  25.887   <2e-16 ***
sqft_basement  2.395e-04  7.885e-06  30.373   <2e-16 ***
houseAge       5.761e-03  1.166e-04  49.401   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3164 on 15119 degrees of freedom
Multiple R-squared:  0.6376,    Adjusted R-squared:  0.6375
F-statistic:  3326 on 8 and 15119 DF,  p-value: < 2.2e-16
```

Figure 16: Final linear regression model

Final regression equation:
$log(price) = 10.25 - 0.03041(bedrooms) + 0.08211(bathrooms) + 0.1059(floors) + 0.03898(condition) + 0.2341(grade) + 0.0001544(sqft\_above) + 0.0002395(sqft\_basement) + 0.005761(houseAge)$

- $\beta_{bedrooms} = -0.03041$. The estimated price of a house is multiplied by exp(-0.03041) = 0.97004 for each additional bedroom in a home, when controlling for the other predictors. This means that the price will decrease if there are more bedrooms.
- $\beta_{bathrooms} = 0.08211$. The estimated price of a house is multiplied by exp(0.08211) = 1.08557 for each additional bathroom in a home, when controlling for the other predictors. This means that the price will increase if there are more bathrooms.
- $\beta_{floors} = 0.1059$. The estimated price of a house is multiplied by exp(0.1059) = 1.1171 for each additional floor in a home, when controlling for the other predictors. This means that the price will increase if there are more floors.
- $\beta_{condition} = 0.03898$. The estimated price of a house is multiplied by exp(0.03898) = 1.03975 for each level of increasing condition, when controlling for other predictors. This means that the better condition the house is in, the higher the house will be priced.
- $\beta_{grade} = 0.2341$. The estimated price of a house is multiplied by exp(0.2341) = 1.26377 for each level of increasing grade, when controlling for other predictors. This means that the better the house is built, the higher it will be priced.

- $\beta_{sqft\_above} = 0.0001544$. The estimated price of a house is multiplied by exp(0.0001544) = 1.00015 for each additional square foot above ground, when controlling for the other predictors. This means that the price will increase if there is more square footage above ground.

- $\beta_{sqft\_basement} = 0.0002395$. The estimated price of a house is multiplied by exp(0.0002395) = 1.0002 for each additional square foot of the basement, when controlling for the other predictors. This means that the price will increase if there is more square footage in the basement.

- $\beta_{houseAge} = 0.005761$. The estimated price of a house is multiplied by exp(0.005761) = 1.00578 for each additional year of the home's age, when controlling for the other predictors. This means that the price of the home will increase slightly if the house is older.

5.6 Relevant Conclusions Addressing Our First Question of Interest

Based on the interpretation of the model we can draw multiple conclusions regarding the predictors for price. From the final regression model, it is clear that the grade level of the house has the most impact on how expensive the house is. For every unit increase of the grade there is a 26.4% increase in the house price. The next most impactful predictor is the number of floors the house contains. For every additional floor there is an 11.71% increase in the house price.
The predictors that have less of an impact are the sqft_above and sqft_basement of the house, meaning that regardless of how large a house is, how it was built is more important when taking all the variables into account. How the house was built can be defined by grade, floors, and houseAge. HouseAge could represent how the house was built because there are certain styles and layouts that are more popular during certain years, which could effect how well they "hold up" in the future.

6 Visualizations for Logistic Regression

In logistic regression, we were looking at several predictor variables as well as the binary response variable of above or below average grade. Focusing on the bar graphs below, we see that there are several variables that seem to have some sort of relationship with the grade of the house.

Figure 17: How Each Categorical Variable Affects Grade

Starting from the top left, we noticed that surprisingly the renovation status, whether or not a house has been renovated, did not seem to impact whether a house had an above average grade. Waterfront, on the other hand, seemed to have a strong impact as we noticed there were nearly half the number of above average grade houses by a waterfront as opposed to not near a waterfront.

Not surprisingly, the view status, number of floors, and number of bathrooms all seemed to have positive relationships with a house being above average for grade. It seemed that the better the view and the larger the number of floors and bathrooms in a house, the more likely the house was to have above average construction and design. Condition seemed to have a nonlinear impact.

We noticed a large outlier when looking at the number of bedrooms, there is a house with 33 bedrooms! Upon exploration we determined that this is likely a data entry error that could skew our modelling results, so we opted to remove it before modeling.

Figure 18: How Each Numerical Variable Affects Grade

When looking at the density plots shown above, we saw large differences between houses with above and below average grade for zip codes, house age, square footage of living space, and price! This indicates that these variables may be very impactful in our model.

Next, we created ridgeline density plots above to display a visual representation of where the houses with the highest grades were most likely to be located.

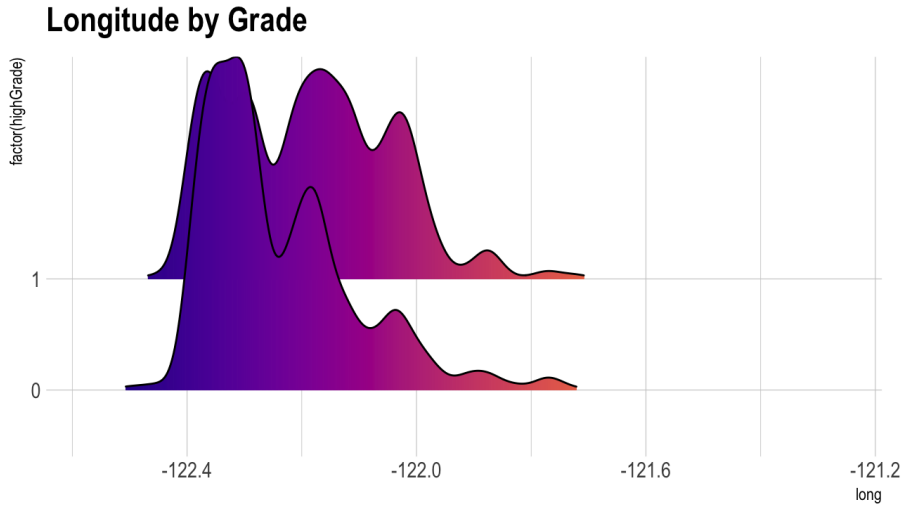

Figure 19: Ridgeline Density Plot for Longitude

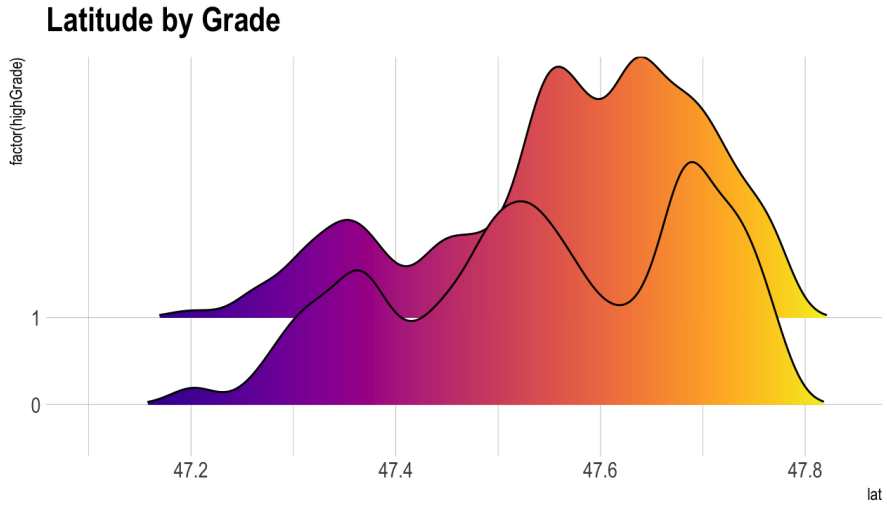

Figure 20: Ridgeline Density Plot for Latitude

In the longitude plot, the area in the upper density plot that is greater than that of the lower density plot is approximately in the range of (-122.2,-122.1). This is the approximate range of the longitudes where the houses of the highest grades are located. Similarly, on the latitude plot, the area in the upper density plot that is greater than that of the lower density plot is approximately in the range of (-47.55,-47.65). This is the approximate range of the latitudes where the houses of the highest grades are located.

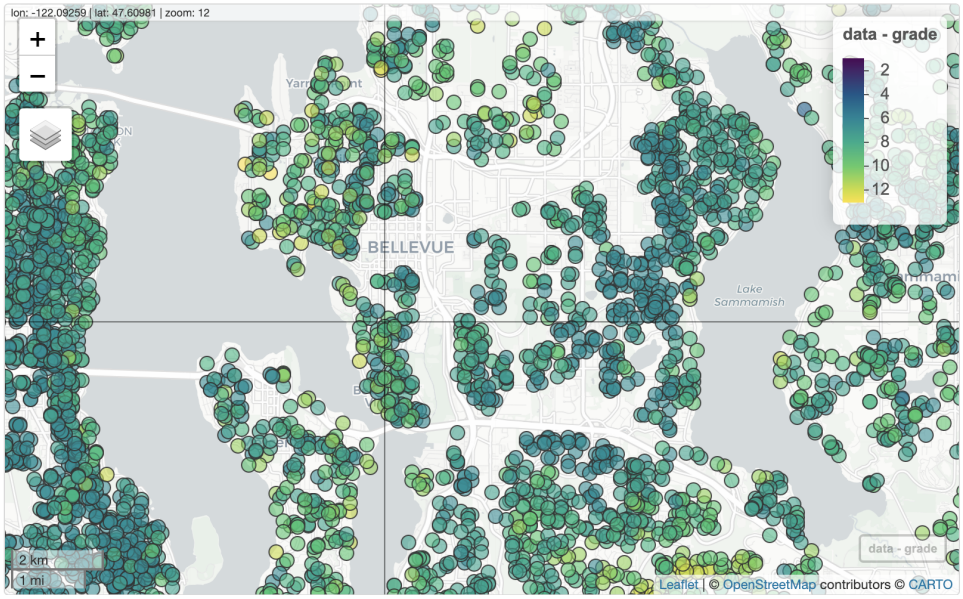

Figure 21: Map of Bellevue

Figure 21 is a view of a map that shows the approximate area that we denoted to be the location

of the houses with the highest grades. These houses appear to be located in Bellevue, which upon further research, turns out to be an up-and-coming technology hub. This well-known city is located in an area with moderately large bodies of water surrounding it – this is extremely surprising that the houses here had the higher grades despite there being more waterfront houses. Perhaps the great view outweighed the waterfront status! This would be an interesting place to take a closer look into to discover what other factors make real estate in this area so much more expensive than other areas in King County.

7 Logistic Regression

7.1 Modeling

One of the primary goals of our analysis was to create a logistic regression model that could predict with relatively high accuracy the odds of a house being rated as having above average construction and design (grade) using variables in the data set. We initially considered the price, number of bedrooms, number of bathrooms, number of floors, square footage of the living space, square footage of the land space, square footage of the basement, view rating, waterfront status, condition, zip code, house age, and renovation status in the model. This full model was then evaluated and improved on using various improvement methods outlined below.

7.2 Model Improvement Methods

7.2.1 Regsubets

Firstly, we utilized the regsubsets function in the leaps package to fit all possible regression models based on the supplied data frame and specified response variable, then calculate the values of R2, adjusted R2, SSres, Mallows Cp, and BIC of each model. After extracting information regarding adjusted R2, Mallow's Cp, and BIC, we found that there was one model that returned as the best model based on all three of these criteria. That model predicts the odds of a house being rated as having above average construction and design (grade) based on the following 8 predictors: price, number of bedrooms, number of bathrooms, number of floors, square footage of the living space, view rating, waterfront status, and house age.

```
##
## Call:
## glm(formula = highGrade ~ price + bedrooms + bathrooms + sqft_living +
##     floors + waterfront + view + houseAge, family = "binomial",
##     data = train)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.313e+00  1.599e-01 -26.981  < 2e-16 ***
## price        6.173e-06  1.811e-07  34.079  < 2e-16 ***
## bedrooms    -4.133e-01  3.752e-02 -11.016  < 2e-16 ***
## bathrooms    2.292e-01  5.956e-02   3.848 0.000119 ***
## sqft_living  1.526e-03  6.489e-05  23.511  < 2e-16 ***
## floors       6.036e-01  5.170e-02  11.675  < 2e-16 ***
## waterfront1 -2.273e+00  4.809e-01  -4.726 2.29e-06 ***
## view         2.630e-01  4.645e-02   5.662 1.50e-08 ***
## houseAge    -3.619e-02  1.255e-03 -28.843  < 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: 20946  on 15127  degrees of freedom
## Residual deviance: 10452  on 15119  degrees of freedom
## AIC: 10470
##
## Number of Fisher Scoring iterations: 6
```

Figure 22: Summary of the regsubsets model

When looking at the summary output, we performed the Wald's test to determine if there were any additional variables that could be removed, but all the predictors appeared significant and therefore were not removed in the presence of the other variables.

```
   price    bedrooms   bathrooms sqft_living      floors waterfront1        view
2.500468    1.695369    3.201961    3.990681    1.518701    1.232696    1.366830
houseAge
1.717170
```

Figure 23: VIFs for the regsubsets model predictors

We then calculated the variance inflation factors (VIFs) to determine how associated each predictor is with the other predictor variables, we see that all the VIF values are extremely small. Being that each VIF value is smaller than 5, we were relatively confident that there was not strong evidence of multicollinearity between these variables. Before moving forward with this model, we wanted to perform other model selection procedures to confirm this was the best model possible.

7.2.2 Forward, Backward, and Both Stepwise Selection

Next we performed three stepwise selection methods: forward selection which progressively adds variables to the intercept-only model until it achieves the best AIC value possible, backward selection which removes variables one at a time from the full model until it achieves the best AIC value, and finally the stepwise selection which both adds and removes variables until it achieves the best AIC value for the model. Surprisingly, all three of these selection methods resulted in the same final model, but this was a different model than was chosen by the regsubsets function. Based on the stepwise methods, the ideal model predicts above-average grade based on these 11 variables: price, number of bedrooms, number of bathrooms, number of floors, square footage of the living space,

square footage of the land space, square footage of the basement, view rating, waterfront status, condition, and house age. In other words, the full model after removing zip code and renovation status is chosen.

```
##
## Call:
## glm(formula = highGrade ~ sqft_living + floors + price + houseAge +
##     bedrooms + sqft_basement + bathrooms + view + waterfront +
##     sqft_lot + condition, family = "binomial", data = train)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.016e+00  2.034e-01 -19.745  < 2e-16 ***
## sqft_living    1.750e-03  7.144e-05  24.503  < 2e-16 ***
## floors         3.619e-01  5.825e-02   6.214 5.17e-10 ***
## price          6.254e-06  1.831e-07  34.158  < 2e-16 ***
## houseAge      -3.436e-02  1.315e-03 -26.125  < 2e-16 ***
## bedrooms      -4.259e-01  3.790e-02 -11.238  < 2e-16 ***
## sqft_basement -6.348e-04  8.074e-05  -7.862 3.80e-15 ***
## bathrooms      2.916e-01  6.049e-02   4.821 1.43e-06 ***
## view           2.871e-01  4.651e-02   6.172 6.74e-10 ***
## waterfront1   -2.262e+00  4.771e-01  -4.742 2.12e-06 ***
## sqft_lot      -2.857e-06  8.015e-07  -3.564 0.000365 ***
## condition     -1.041e-01  4.319e-02  -2.409 0.015987 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20946  on 15127  degrees of freedom
## Residual deviance: 10374  on 15116  degrees of freedom
## AIC: 10398
##
## Number of Fisher Scoring iterations: 6
```

Figure 24: Summary of the stepwise model

When looking at the summary output, we performed the Wald's test to determine if there were any insignificant variables that could be removed, but all the predictors appeared significant, so we did not remove any.

```
      price      bedrooms     bathrooms   sqft_living      sqft_lot        floors
   2.524588      1.715498      3.307762      4.536039      1.067337      1.870609
waterfront1          view     condition sqft_basement      houseAge
   1.233055      1.397637      1.187696      1.738806      1.888094
```

Figure 25: VIFs for the stepwise model predictors

In the model chosen by the stepwise procedures, the VIF values were all relatively small (less than 5) so there was no alarming evidence of multicollinearity to be aware of.

This model was larger than the previous one we found, so we reflected on the question we were truly interested in answering to determine whether the additional predictor variables were necessary. The larger model chosen by the stepwise selection methods includes three different square footage measurements (living space, land space, and basement) while the smaller model chosen earlier only uses the one square footage measurement for the living space. While having all three measurements may have improved the accuracy of the model, we determined that these variables would likely be very correlated with one another. In order to avoid unnecessary multicollinearity, we opted to remove

the square foot measurements for the basement and land space. The only other predictor that was up for debate was the condition variable. Because condition was the only variable differentiating the two models, we decided to avoid overfitting by dropping condition and proceeding with the model chosen by the regsubsets function.

7.3 Final Model

To ensure that the regsubsets model was useful for predicting the odds of a house being rated as having above average construction and design (grade), we preformed a likelihood ratio test as outlined below.

$H_0 : \beta_{price} = \beta_{bedrooms} = \beta_{bathrooms} = \beta_{sqft\_living} = \beta_{floors} = \beta_{waterfront} = \beta_{view} = \beta_{houseAge} = 0$ (The full model is not useful for estimating the grade status of a house)
H_A : At least one coefficient in H_0 is nonzero (The full model is useful for estimating the grade status of a house)

Test Statistic, $\triangle G^2$, = (null deviance - full model deviance) = 10494.06
Critical value = 14.06714
P-value = 0

Because the test statistic was larger than the critical value, and the p-value was smaller than $\alpha = 0.05$, we rejected H_0 in favor of H_A. In other words, we had enough evidence to conclude that this logistic regression model with eight predictors was useful in estimating the odds of a house being rated as having above average construction and design compared to the intercept-only model.
We then were confident in proceeding with the model found using the regsubsets function. The estimated logistic regression equation for our final model is as follows:

$highGrade = -4.313 + 0.000006173(price) + -0.4133(bedrooms) + 0.2292(bathrooms) + 0.001526(sqft\_living) + 0.6036(floors) + -2.273(waterfront1) + 0.2630(view) + -0.03619(houseAge)$

7.4 Assessing Predictive Ability

To assess the predictive ability of our final logistic regression model we used the model to estimate the predicted probabilities of the test data, then used a threshold of 0.5 to classify the test data. To summarize the number of correct and incorrect classifications, based on our test data, we produced a confusion matrix as shown below.

	FALSE	TRUE
0	2914	501
1	517	2552

Figure 26: Confusion matrix

From this matrix, we are able to compute valuable metrics such as error rate, accuracy, and more.

Error rate: $\frac{501+517}{(2914+501+517+2552)} = \frac{1018}{6484} = 0.1570019$
False Positive Rate: $\frac{501}{(2914+501)} = \frac{501}{3415} = 0.1467057$
False Negative Rate: $\frac{517}{(517+2552)} = \frac{517}{3069} = 0.1684588$

True Positive Rate: $\frac{2914}{(2914+501)} = \frac{2914}{3415} = 0.8532943$
True Negative Rate: $\frac{2552}{(517+2552)} = \frac{2552}{3069} = 0.8315412$
Accuracy: $\frac{2914+2552}{(2914+501+517+2552)} = \frac{5487}{6484} = 0.8462369$
Precision: $\frac{2552}{(501+2552)} = \frac{2552}{3053} = 0.8358991$

The metrics computed above show that we have a very small error rate of 15.70%, and a relatively high classification accuracy of 84.62%. We see that our precision (83.59%) is also quite large. This shows that not only is our threshold level just right, but we can also feel confident that this model has a relatively high and accurate predictive ability.

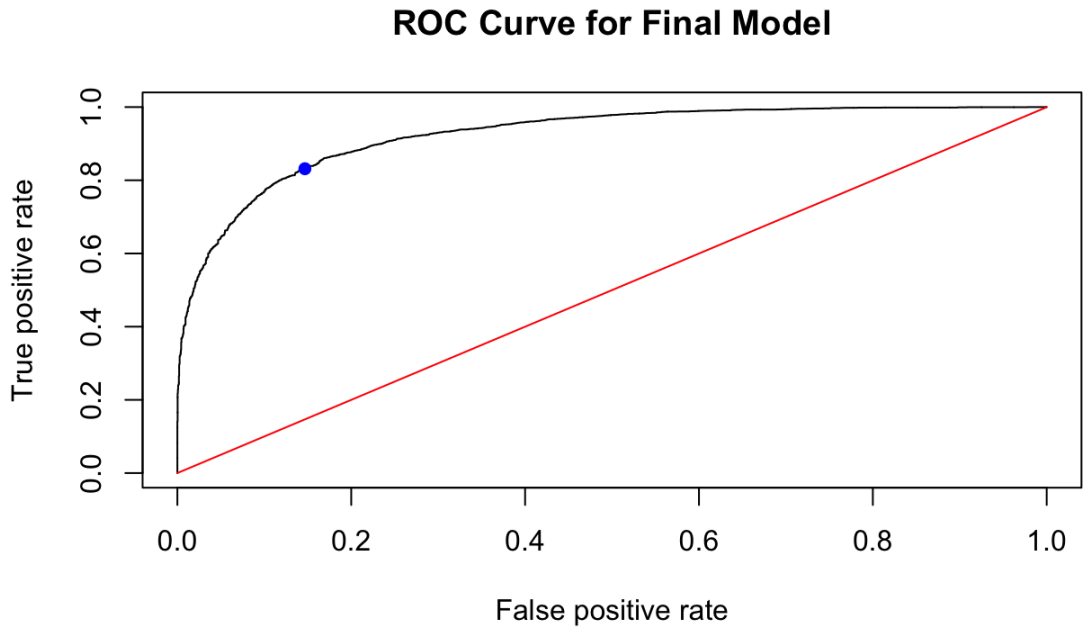

Figure 27: ROC plot

The receiver operating characteristic (ROC) curve shown in black plots the sensitivity (true positive rate TNR) on the y-axis and 1 − specificity (false positive rate FPR) on the x-axis. This shows our TPR and FPR at every threshold level. Because this curve was far above the diagonal red line which represented a model classifying at random, we were assured that our logistic regression model does much better than random guessing.

The blue dot on the figure marked the true positive rate and false positive rate at our chosen threshold 0.5. Because this point was in the top left-hand corner of the curve, this threshold was ideal for our predictive classification.

Finally, we computed the area under the curve (AUC) value and found it equal to 0.9238871. An AUC value of 0.5 is the same as random guessing, but the closer the value is to 1, the closer the model is to correctly classifying all observations. Because our AUC was relatively close to 1, we noted that our logistic regression was much better at classifying houses with above average construction and design (grade) than a random guessing model.

7.5 Model Interpretations

Using our estimated logistic regression equation, we interpreted the impact each predictor variable had on the odds of a house having above average construction and design (grade) as follows:

- $\beta_{price} = 0.000006173$. The estimated odds of a house being ranked with above average condition and design is multiplied by exp(0.000006173) = 1.000006 for each additional dollar of a house is sold for, when controlling for the other predictors. In other words, the odds increase a small bit for each dollar.

- $\beta_{bedrooms} = -0.4133$. The estimated odds of a house being ranked with above average condition and design is multiplied by exp(-0.4133) = 0.6614638 for each additional bedroom in a house, when controlling for the other predictors. In other words, the odds decrease a bit for each bedroom.

- $\beta_{bathrooms} = 0.2292$. The estimated odds of a house being ranked with above average condition and design is multiplied by exp(0.2292) = 1.257594 for each additional bathroom in a house, when controlling for the other predictors. In other words, the odds increase a bit for each bathroom.

- $\beta_{sqft\_living} = 0.001526$. The estimated odds of a house being ranked with above average condition and design is multiplied by exp(0.001526) = 1.001527 for each additional square foot of interior living space, when controlling for the other predictors. In other words, the odds increase a small bit for each square foot of living space.

- $\beta_{floors} = 0.6036$. The estimated odds of a house being ranked with above average condition and design is multiplied by exp(0.6036) = 1.82869 for each additional floor in a house, when controlling for the other predictors. In other words, the odds increase by nearly double for each additional floor.

- $\beta_{waterfront} = -2.273$. The estimated odds of a house being ranked with above average condition and design for a house overlooking a waterfront is exp(-2.273) = 0.1030027 times the odds for a house not overlooking a waterfront, when controlling for the other predictors. In other words, the odds decrease by nearly nine-tenths when being my a waterfront.

- $\beta_{view} = 0.2630$. The estimated odds of a house being ranked with above average condition and design is multiplied by exp(0.2630) = 1.300827 for each additional rating of a house's view, when controlling for the other predictors. In other words, the odds increase a bit for better views.

- $\beta_{houseAge} = -0.03619$. The estimated odds of a house being ranked with above average condition and design is multiplied by exp(-0.03619) = 0.964457 for each additional year since a house was built, when controlling for the other predictors. In other words, the odds decrease a bit for older houses.

The intercept term cannot be interpreted because it would not make sense in this context to classify a house that sold for $0 and had 0 bedrooms, bathrooms, floors, etc.

7.6 Relevant Conclusions Addressing Our Second Question of Interest

Based on our model interpretations and visualizations, we drew various conclusions. Firstly, it was very clear that whether a house is overlooking a waterfront or not had the largest impact on whether it would be rated as having above average condition and design. Houses that overlook the water were nearly 10 times less likely to have above average condition and design than houses not

near water. It was also interesting to note that when holding all other variables constant, additional bedrooms reduce the odds of above average construction and design, but additional bathrooms increase these odds. This may be due in part to modern architecture favoring an open floor plan, or fewer enclosed rooms, as being a better design choice.

The number of floors and higher quality of view directly and relatively strongly improve the odds that a house had above average construction and design (grade). Although one-unit increases in price, square footage of the living space, and the house age had smaller impacts in this model, all of these values worked together to produce an 84.62% accurate classification model for houses with above average construction and design (grade).