Section 1: Findings Summary

Researchers at the University of Virginia have conducted a comprehensive analysis of housing data to provide valuable insights into predicting home prices and grading. By applying multiple linear regression and logistic regression models, they have identified key predictors that have the highest impact on home prices and grades.

In the multiple linear regression analysis, researchers discovered that certain factors play a significant role in determining home prices. The grade of home construction and design, interior square footage, above ground level interior square footage, land square footage of the 15 nearest neighbors, the view of the home, home condition, number of floors, and number of bedrooms were found to be the most influential predictors. While characteristics like the number of bathrooms could relate to home pricing, the listed factors above yielded the highest accuracy in predicting home prices. Notably, the sizing of homes, indicated by square footage of the interior and land, as well as the number of floors and bedrooms, were crucial factors affecting home prices. Additionally, the quality of the home, represented by the grade of construction and design, and the home condition, significantly impacted the cost.

Similarly, the logistic regression analysis focused on determining whether a home has a higher or lower grade based on various housing characteristics. Ten predictors were considered, including price of the home, number of bedrooms, number of bathrooms, square footage of the interior living space, square footage of the land space, number of floors, condition, square footage of interior space above ground level, square footage of interior living space of the nearest fifteen neighbors, and square footage of the land lots of the nearest fifteen neighbors. The findings revealed that bathrooms played a particularly crucial role in predicting the grade of a home. Most predictors had a positive influence on the grade, except for the home condition, square footage of the lot, and square footage of the nearest fifteen homes, which negatively impacted the grade. Overall, all predictors, except for bedrooms, were found to be influential in predicting grade classification.

These findings provide valuable insight for homebuyers by understanding the significant predictors of home prices and grade, potential buyers can make more informed decisions. These finding empower them to consider the most influential factors when evaluating home values.

Section 2: Data Summary

For the purposes of this project, we are using a Kaggle dataset that contains house sale prices for King County, which includes Seattle. The dataset includes homes sold between May 2014 and May 2015. There are 21613 homes (rows) and 21 variables (columns) in this dataset. For our linear regression (question 1) we used price as the response variable and 7 predictors: sqft_living, sqft_lot, sqft_above, sqft_living15, and sqft_lot15, which are all numeric. For our logistic regression in question 2, we created a new binary variable called grade_binary, which classifies a home as either "low" (6) or "high" (7) grade.

Data Dictionary

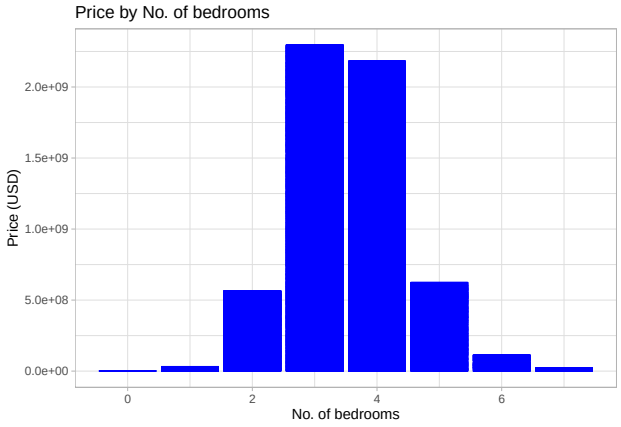
- **price:** price of each home sold (in dollars)
- **price_log:** the natural log of the price variable
- **bedrooms:** number of bedrooms
- **bathrooms:** number of bathrooms
- **sqft_living:** Square footage of the apartments interior living space
- **sqft_lot:** Square footage of the land space
- **floors:** Number of floors
- **view:** An index from 0 to 4 of how good the view of the property was
- **condition:** An index from 1 to 5 on the condition of the apartment
- **grade:** An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.
- **grade_binary:** classifies a home as either "low" (6) or "high" (7)
- **sqft_above:** The square footage of the interior housing space that is above ground level
- **sqft_living15:** The square footage of interior housing living space for the nearest 15 neighbors
- **sqft_lot15:** The square footage of the land lots of the nearest 15 neighbors

Section 3: Questions of Interest

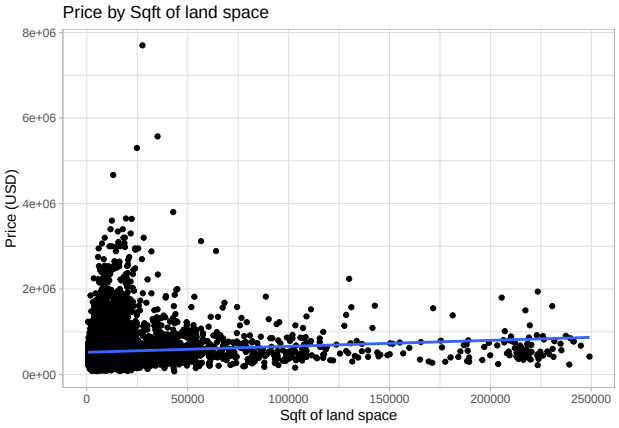
- **Linear Regression:** What characteristics of a home affect the pricing of a home? Can we use these characteristics to predict the price of a home?
  - With this question, we want to explore how various attributes of a home, such as the number of bedrooms, number of floors, or the condition of the place, influence the price of a home. We are motivated to answer this question because price is one of (if not) the most important factors when it comes to purchasing a home, so it is important to explore how the cost may change based on the different characteristics of a home. Therefore, when people are trying to buy a home, they may want to identify what qualities they want in a home (size, condition, location, etc.) so that they can choose a reasonable price range within their budget. Furthermore, homesellers putting their home on the market can determine their pricing not only based on houses around them, but also on different qualities that their house possesses that could affect the home cost as well.
- **Logistic Regression:** Can we classify whether or not a home has a higher or lower grade based on different housing characteristics in the dataset?
  - The motivation behind this question is to determine if we can categorize a home into a higher or lower grade based on its characteristics. We want to explore if attributes such as size or location impact of a home has a higher grade for design and construction. For homeowners or buyers that aren't aware or are not educated in the world of real estate, the categorized grade of the home can be a starting indicator for which home grade to explore. Homebuilders could also see what attributes need to be added to a home to receive a higher grade for their home design and construction. For these reasons, people may be interested in what home characteristics could relate to a higher or lower construction and design grade.

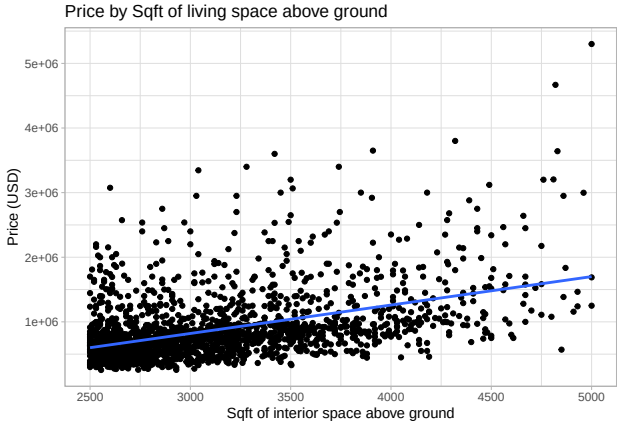
Section 4: Multiple Linear Regression Data Visualizations

Scatterplots of Predictors Against Price



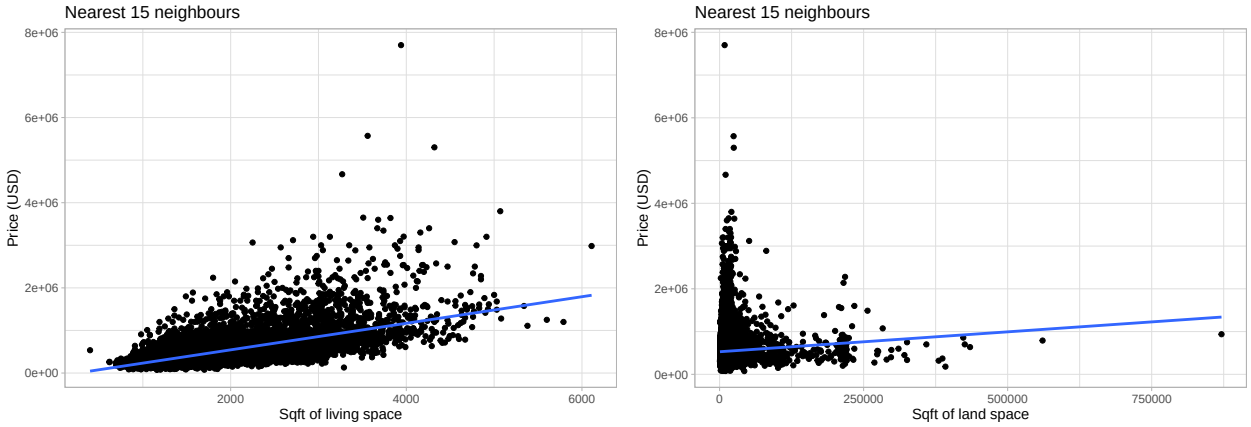
There is a great hike in prices for 3 and 4 bedroom houses, but the price decreases after 4 bedrooms. Logically, the price of houses with more than 5 bedrooms cannot be inexpensive. Either there are fewer observations for houses with 5 bedrooms or more, which has skewed the data, or those houses are situated in a region with relatively low house values. Another way to put it is that 3 and 4 are more common.



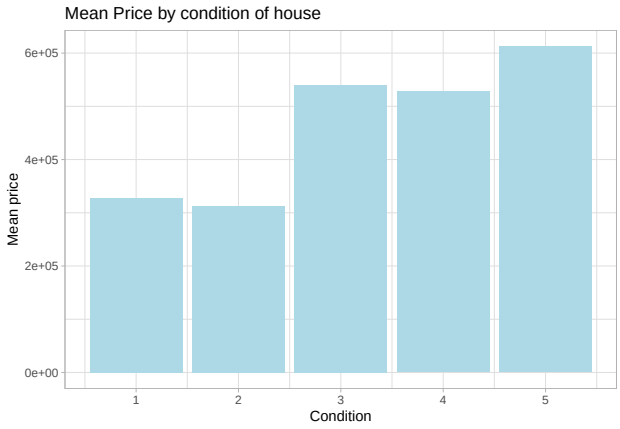


From the above scatterplots we can see a positive linear relationship between prices and sqft of living space, sqft of living spaces above the ground level. Increase in these sqft areas result in increase of prices.

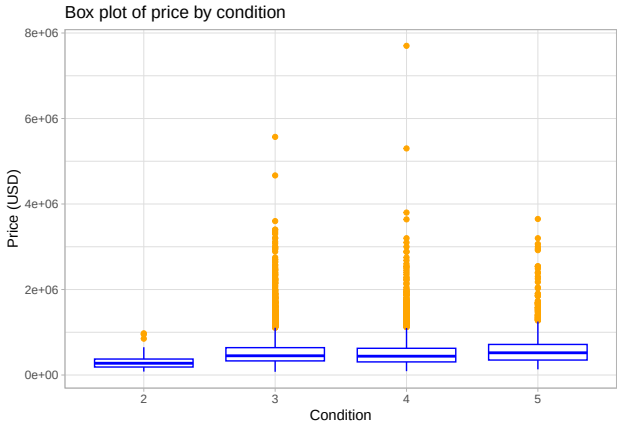
However, if you see the scatterplot (top right) of prices against sqft of land space, the relationship between them is linear but it is very subtle. This may be because most observations are crowded below 50,000 square feet. Also, there can be a possibility that prices are inexpensive in some areas of Washington. As a result, this scatterplot may be biased and fail to demonstrate the linear link between prices and square feet of land.



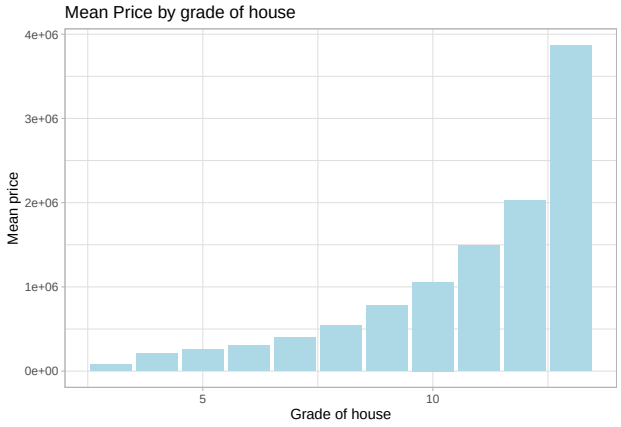
In addition, we compared the price of houses for the 15 closest neighbors based on living space and land space. From above scatter plots, it is clear that there is a direct correlation between price and square feet. Prices rise with increased square footage area.



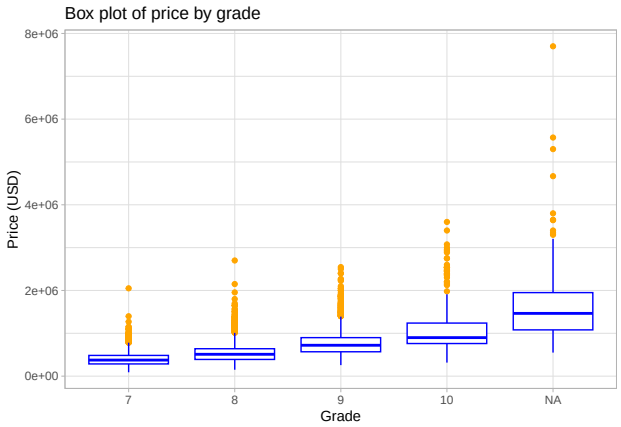
We created a bar chart to check the relationship between house prices and condition of the house. Although there is a price increase from condition 2 to condition 3, the condition of the house does not significantly affect price.



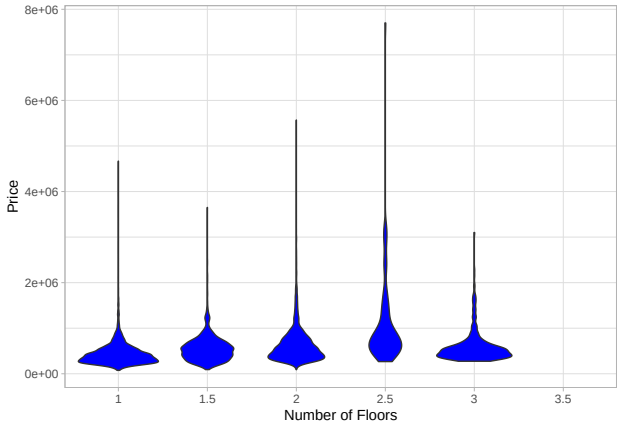
The box plot of price against condition demonstrates the same result, that there is no significant increase in price by improving condition. Additionally, the distribution of data points is somewhat uniform across all conditions. Condition might not be the key differentiator for setting prices.



The bar chart of mean prices versus grade of house provides an impressive visual clue. As the grade rises, the prices exhibit an increasing trend. Based on this plot, we can say that a house's grade may have a big impact on prices. But further in depth statistical analysis is needed to provide a conclusive answer.

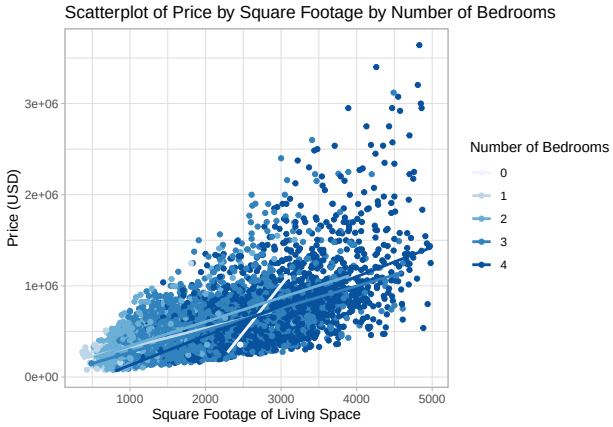


To check the distribution of data points by grade, we created a box plot. Prices rise as the grade rises. There are numerous observations that are not graded but are priced higher than grade 10. Those must be really high-grade residences. Thus, grades could have a significant effect on house prices.

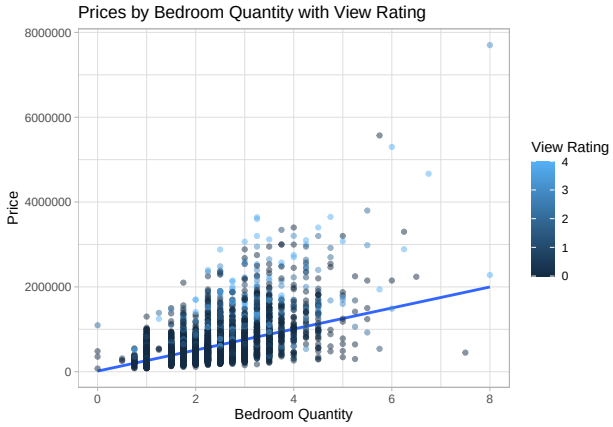


The violin graph explores the relationship between the density of a number of floors with a price. From the

, visualization we can conclude that the majority of homes, no matter how many floors, have prices below 1,000,000 USD. homes with 2-2.5 floors have the greatest range in price. This visualization provides us with an insight that homes that we will be exploring regression model have similar values with some exceptions of homes greater than 2,000,000 USD.

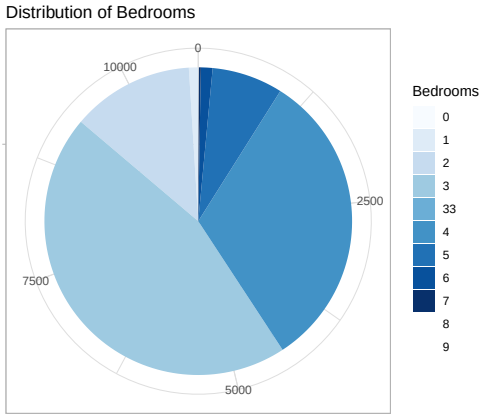


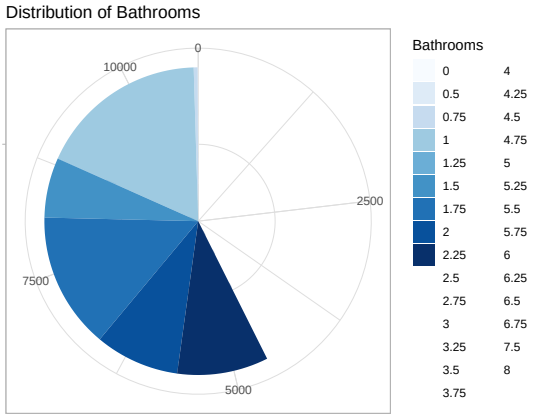
The scatter plot compares the price to sq ft of living space differentiated by the number of bedrooms. The graph shows that as sq ft increases, there is a steady price increase, indicating a linear relationship. When we differentiate based on the number of bedrooms, it can be noted that homes with two or fewer bedrooms have smaller sq ft and have lower prices as compared to homes with three or more bedrooms with homes valued at $1000000 and above. The regression line indicates that homes with bedrooms between 1-4 have similar positive slopes.



In addition to displaying the distribution of view ratings, the scatter plot shown below demonstrates the connection between pricing and the distribution of bedrooms. As previously said, we see a price increase as the number of bedrooms increases, which suggests a linear relationship. Within this range, prices rise specifically from $1,700,000 to $2,100,000. However, whenever there are more than 4 bedrooms, rates either stay the same or even go down.

We discovered minimal statistically significant link between the price variable and the view rating. View ratings range from 0 to 4 for homes in various pricing ranges and bedroom counts. This finding indicates view ratings with prices of 0 and 1 are on a lower price scale. In contrast, although we see lower ratings at lower prices, the graph does not show a consistent trend across price or bedroom ranges.



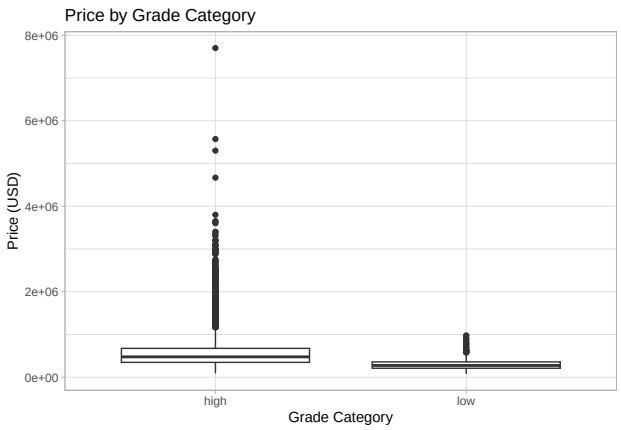


To evaluate the distribution of our outputs with regard to their proportions, it is crucial to acquire insight into the distribution of the number of data points before diving into the analysis of regression models. A cluster of data points around this feature suggests that 3 bedrooms are the most common number of bedrooms in homes in this sample. Similar to bedrooms, most houses typically feature between 1-2 bathrooms.

The presence of a cluster of data points around 3 bedrooms indicates that homeowners frequently and widely choose this number of bedrooms. It can be a sign of a typical or preferred arrangement of residential properties in the dataset. This data enables us to comprehend the current patterns in the distribution of bedrooms in households.

Additionally, the presence of a sizable percentage of houses with 1-2 bathrooms indicates that this range is also a popular choice among homeowners. This knowledge reveals the normal bathroom distribution within the dataset and aids in determining the typical bathroom layout for residential structures.

We may learn more about the dominant patterns and preferences in the dataset by looking at the distribution of data points for bedrooms and bathrooms. These conclusions provide a helpful framework for additional investigation, such as using regression modeling to examine the correlation between the number of bedrooms or bathrooms and other factors, such as the cost.

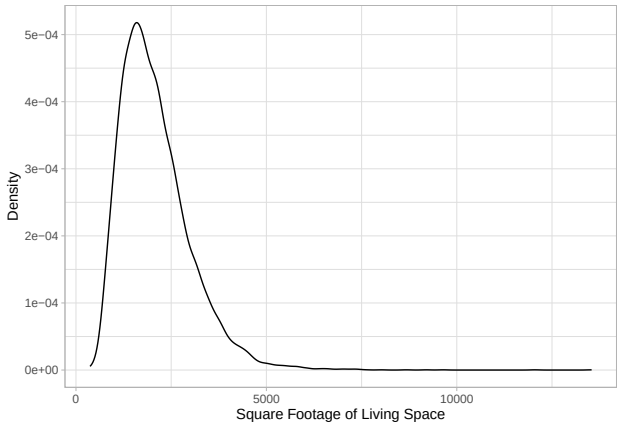


According to the boxplot study, there is a tendency for “high” grade binary categories to have slightly higher pricing than “low” grade binary categories. It’s crucial to remember that there isn’t a big pricing difference between the two groups. The small price rise seen in the “high” grade binary category may be due to the existence of outliers in the data.

It is also essential to keep in mind that the data description does not include any more details regarding the precise standards or traits that define a building as having high-quality construction and design. As a result, the dataset may contain inconsistent and varying definitions of what high-quality building and design are.

As a result, even though there appears to be a minimal correlation between the grade binary category and

pricing, it is crucial to interpret these results with care. We can only make generalizations regarding the relationship between grade and pricing due to the existence of outliers and the absence of specific information about the standards for high-quality construction and design. More research and context-specific data would be required to offer more accurate and trustworthy conclusions.



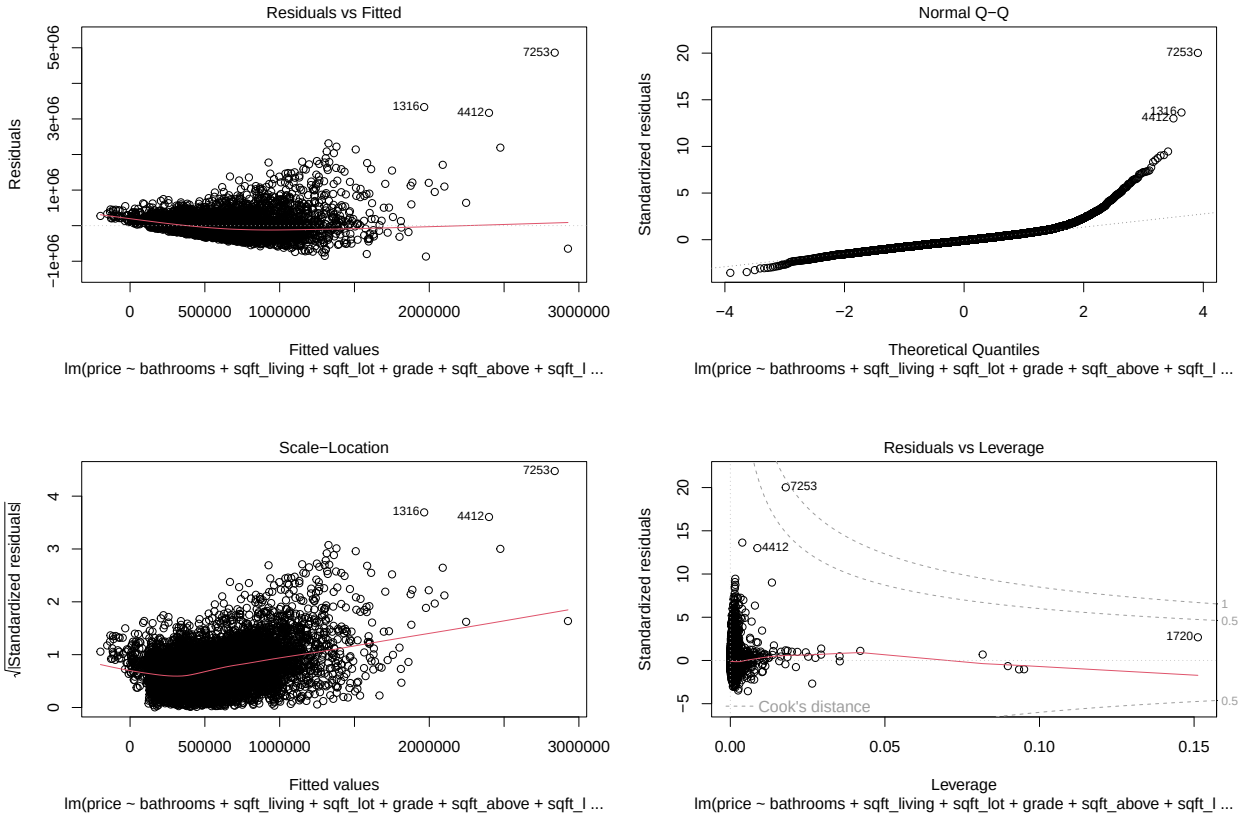
Section 5: Multiple Linear Regression

What characteristics of a home affect the pricing of a home? Can we use these characteristics to predict the price of a home?

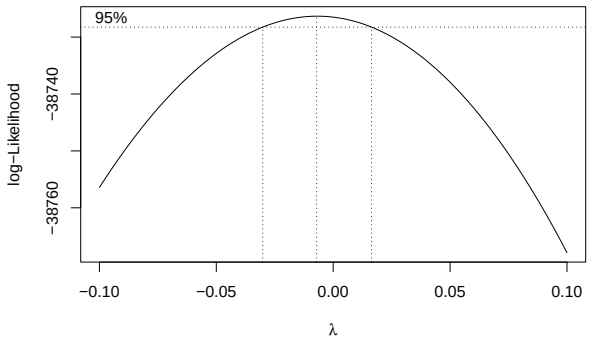
Model Selection

When looking at the plots, we notice a positive relationship between price and interior square footage for the home and its 15 nearest neighbors, above ground level interior square footage, grade of the housing design and construction, bathrooms, bedrooms, and maybe the land square footage for the home and its 15 nearest neighbors. The linear regression line seems flat for predictors floors, view, and condition.

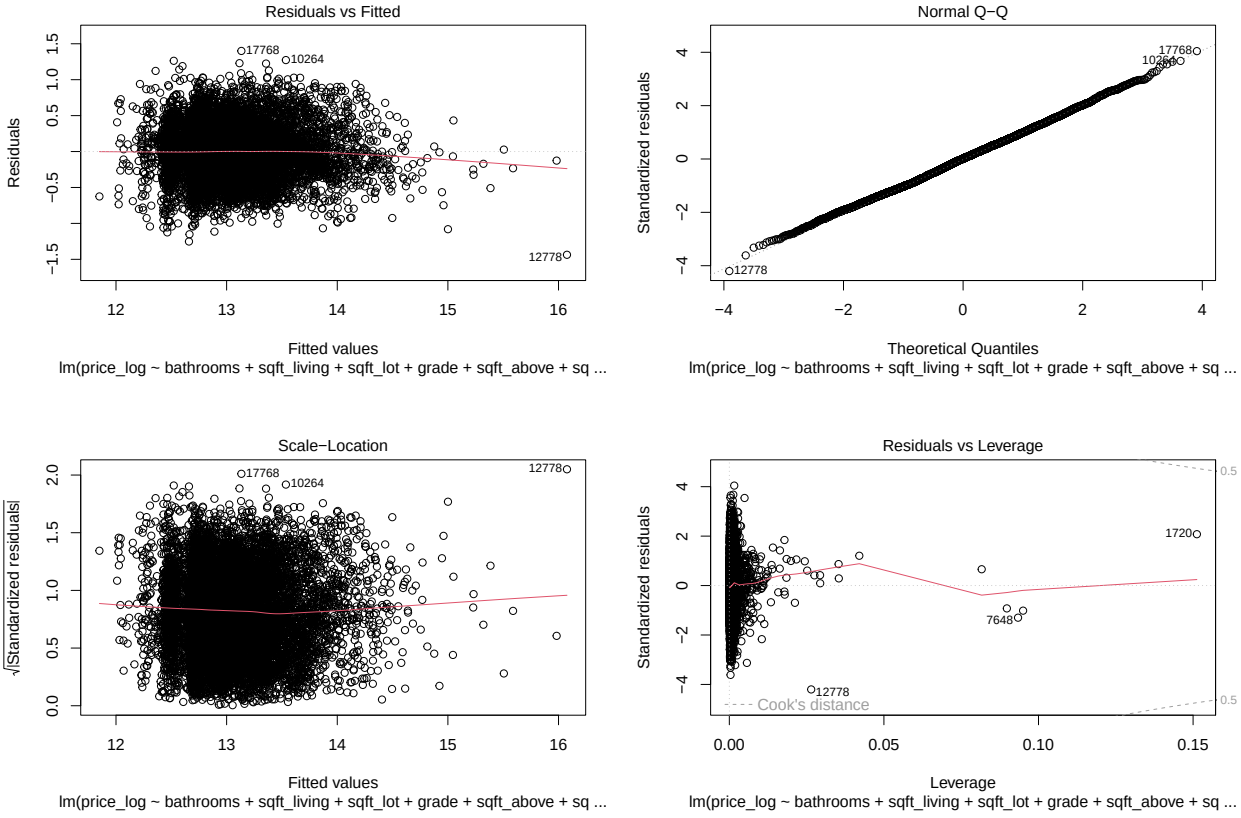
We decide to model our data with price against number of bedrooms and bathrooms, interior square footage, land space square footage, above ground level interior square footage, interior square footage for the nearest 15 neighbors, and the grade of design and construction. We chose these predictors because out of all the scatterplots of price against all potential predictors in the original dataset, these six predictors seemed to display somewhat of a trend.



The residual plots display an increase in variance (a fanning out of the points), as well as a potential non-linear curving pattern in the data. Therefore, the constant variance and the 0 mean assumptions are not met. We plan to transform the y-variable to stabilize the variance.



The Boxcox plot displays that 0 is included in the interval for lambda, so we choose to log transform our y-variable price.



When looking at the residual plot, the points seem to be evenly scattered around 0, meeting the 0 mean assumption, and the variance seems to be constant, meeting the constant variance assumption.

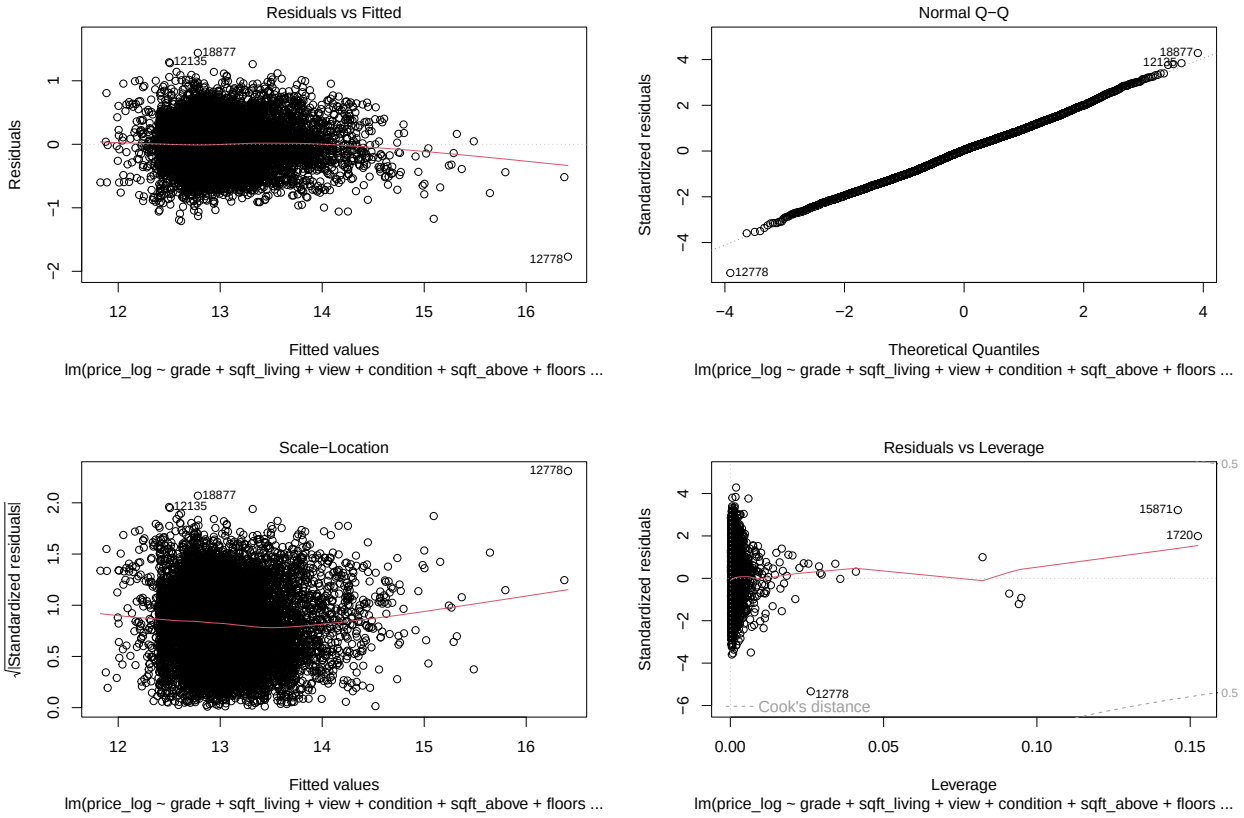
T-Test

We see that the bathrooms predictor has an insignificant t-test, so we run a hypothesis test to drop the predictor from the model.

H0: B1 = 0 Ha: B1 does not equal 0 Test-Statistic: -1.873 p-value: 0.0611 Because the p-value of 0.06 is

smaller than the 0.05 alpha level, we can reject the null hypothesis and drop the bathrooms predictor from the model.

Modeling with Search Procedures



After running a stepwise search procedure, it output the a model with the log of price against grade, sqft_living, view, condition, sqft_above, floors, bedrooms, sqft_lot15, sqft_lot, and bathrooms.

The residual plots of this model also seem to meet the 0 mean assumption, with its points evenly scattered around 0, and the constant variance assumption, with no increase or decrease in variance.

Partial Linear F-Test

With the model found with the search procedure, we see that the sqft_lot and bathrooms have an insignificant t-test, meaning both can be dropped INDIVIDUALLY in the presence of the other predictors. Therefore, we run a partial F-test to see if these two predictors could be dropped together

H0: Bsqft_lot = Bbathrooms = 0 Ha: at least one of the coefficients in H0 is not 0 test-statistic: 2.9468 p-value: 0.05255 Because our p-value of 0.05255 is greater than the 0.05 alpha level, we fail to reject the null hypothesis and can conclude we should drop the sqft_lot and bathrooms predictors and go with the reduced model.

Comparing Potential Models

Model Criteria

When using the regsubsets() function to fit all possible regression models based on the dataframe and response variable, we see that the model chosen by the stepwise selection procedure gives the highest adjusted R-squared, lowest Mallow's CP, and lowest BIC values, which gives us a good balance of model fit and model complexity compared to the other models.

Comparing our adjusted R-squared value of our model chosen with selection procedures and the model selection criteria (0.59095) with the adjusted R-squared value of our model chosen through exploring our visualizations (0.0.5672), we see that the adjusted R-squared for the model chosen with selection procedure is higher.

Comparing Predictive Ability of Test Data

- Test MSE for model selected with visualizations: 0.1183
- Test MSE for model selected with selection procedure: 0.1124
- RMSE for model selected with visualizations: 0.3439
- RMSE for model selected with selection procedure: 0.3353

We see that the MSE and RMSE are both lower for the model selected using the stepwise selection procedure (with price_log against predictors grade, sqft_living, view, condition, sqft_above, floors, bedrooms, and sqft_lot15). Therefore, it has a better 'prediction accuracy' than our model with just sqft_living, sqft_lot, grade, sqft_above, sqft_living15, and sqft_lot15 predictors.

We decide to go with the model with the higher prediction accuracy:

Regression Equation: - logpricehat = 10.81 + 0.1942grade + 2.620e-04sqft_living + 8.310e-02view +9.849e-02condition -8.928e-05sqft_above + 6.008e-02floors - 2.190e-02bedrooms - 4.611e-07sqft_lot15.

For a 1% increase in each of the predictors INDIVIDUALLY, the predicted response of log price increases by approximately it's slope/100, while holding all other predictors constant.

Detecting High Leverage Observations and Outliers

Looking at the high leverage point values, it seems like a lot of the values are barely above the 0.0013 cutoff, as shown through the first couple high leverage point values. Comparing the mean values of the variables of the high leverage points to the training dataset, we see there is not a huge difference, but the high leverage points seem to have a slightly higher mean price, larger bedroom count, and larger square footage of interior space. Though there are 687 points flagged as high leverage, they are mostly only slightly higher than the cutoff, and we notice a general higher price and larger home size for high leverage points.

Detecting Influential Observations

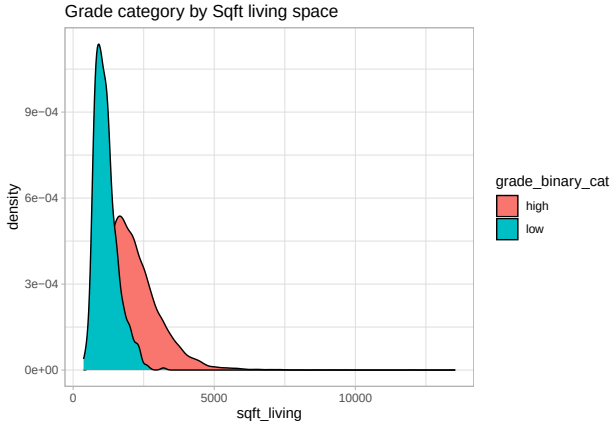
We see that the DFFITS criteria of 2*sqrt(p/n) is equal to 0.0509, and looking at a couple of the DFFITS magnitudes, some of the points are barely higher than the criteria. When comparing the values of the data points with DFFITS larger than the criteria, we do not see any apparent, large differences in values of the variables in the training data versus just the flagged DFFITS points.

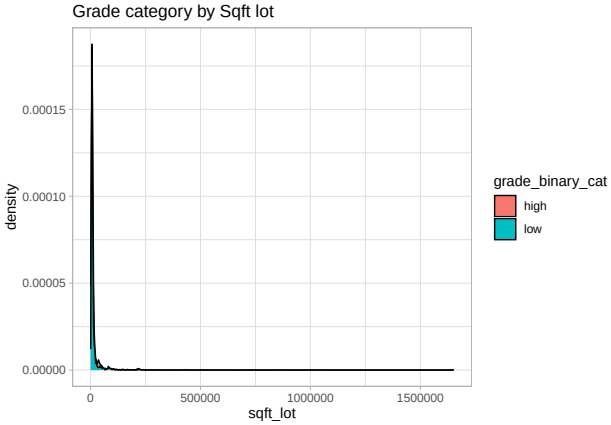
Taking a closer look at our flagged observations, we do not notice any unusual circumstances, in which something seems fundamentally wrong, and we did not calculate any Cook's distance which is larger than 1, so we decided not to remove any observations.

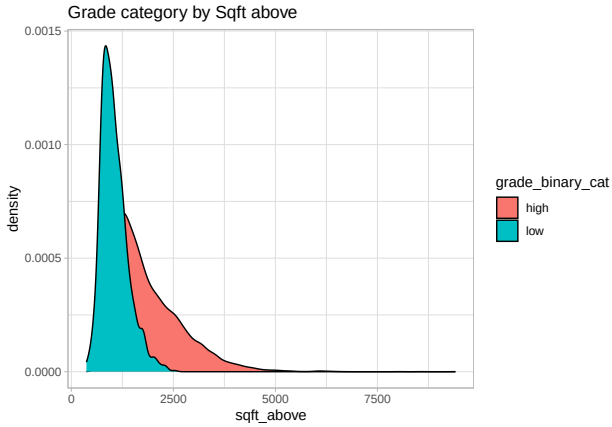
Conclusion

With our final model, we can answer our first question: What characteristics of a home affect the pricing of a home? We were exploring two models– one found through a stepwise selection procedure and the other found through data visualization exploration. We ultimately decided to go with the selection procedure model with sqft_living, sqft_lot, grade, sqft_above, sqft_living15, and sqft_lot15 predictors, as it had lower MSE/RMSE values, with a higher predictive ability. Therefore, we can answer the question by stating that the grade of the home construction and design, interior square footage, and above ground level interior square footage, land square footage of the 15 nearest neighbors, the view of the home, home condition, number of floors, and number of bedrooms as predictors have the highest predictive ability of the price of the homes.

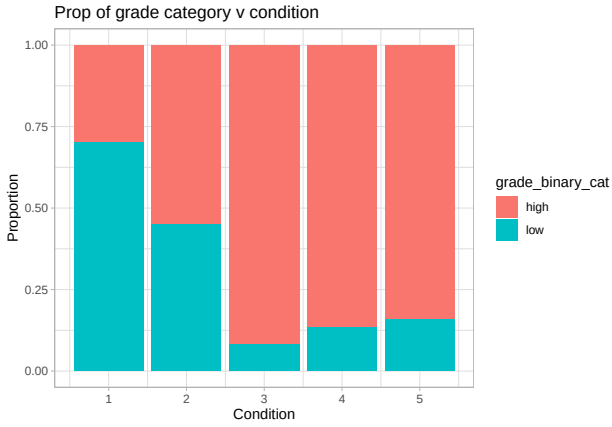
Section 6: Logistic Regression Data Visualizations

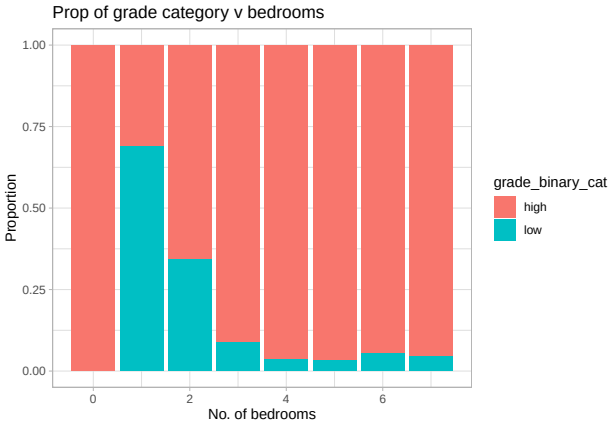


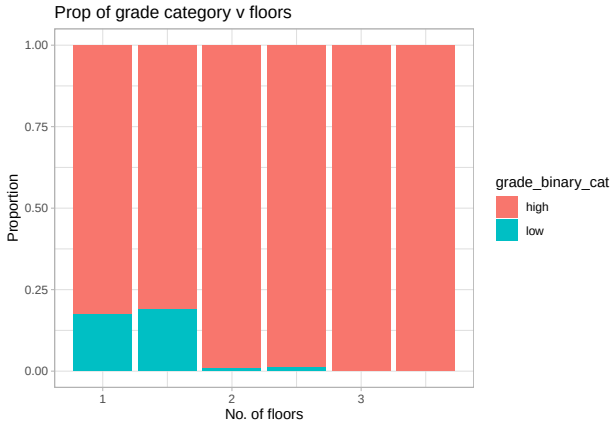




Low grade houses have a higher proportion of sq ft of living space than high grade houses. For sqft_above as well, low grade houses are more in proportion than high grades. However, we see no significant proportion for high and low in sqft_lot.







For conditions 1 and 2, we can see (top left) the significant proportion of low grade houses. But as the condition gets better i.e. 3, 4, 5, high grade houses have higher proportion. For bedrooms, 1 and 2 bedroom houses show some low graded houses, but otherwise most of the houses are high graded. In case of bathrooms, only few houses are labelled low grade, and most of the houses are high grade houses. Overall, high grade houses have higher proportion over low grade houses.

Section 7: Logistic Regression

Question 2: Logistic Regression: Can we classify whether or not a home has a higher or lower grade based on different housing characteristics in the dataset?

Full Model Based on Visualizations

```
##   sqft_living      sqft_lot    sqft_above sqft_living15    sqft_lot15
##      2.908739      1.987847      2.115022      1.299889      2.021879
##      bedrooms     bathrooms         price        floors     condition
##      1.507038      1.569605      1.141804      1.206175      1.053053
```

There is no evidence of multicollinearity because all of the vifs are less than 3.

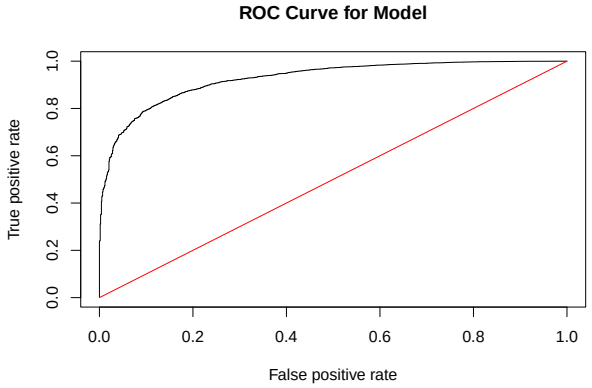
A Wald test suggests that we can remove bedrooms as a predictor because it is the only insignificant p-value (0.862271) based on the 0.05 threshold in the full model.

H0: Bbedrooms = 0 Ha: Bbedrooms does not equal 0 test-statistic: 0.173 p-value: 0.862271 Because the p-value is higher than the 0.05 alpha level, we fail to reject the null hypothesis and can drop the bedrooms predictor from the model.

After removing bedrooms, all of the predictors in the reduced model are individually significant.

The estimated regression for model 1 (without bedrooms) is:

log(pi_hat/1-pi_hat) = -5.327 + 7.797e-04(sqft_living) - 3.804e-06(sqft_lot) + 1.272e-03(sqft_above) + 8.048e-04(sqft_living15) - 7.066e-06(sqft_lot15) + 1.212(bathrooms) + 4.643e-06(price) + 4.283e-01(floors) - 2.148e-01(condition)



Since the ROC curve is above the diagonal line, the log regression performs better than random guessing.

The AUC of 0.9259531 means the log regression performs better than random guessing.

Model Criteria for Model Selection

We use the regsubsets() function from the leaps package to run all possible regressions, and we find the regression equations with the highest adjusted R-squared, lowest Mallow's CP, and lowest BIC.

Highest adjusted R2 Regression Equation:

log(pi_hat/1-pi_hat) = 4.062816e-01 + 3.586136e-05(sqft_living) - 1.951235e-07(sqft_lot) - 4.660069e-05(sqft_above) + 5.570982e-05(sqft_living15) + 2.697893e-02(bedrooms) + 1.024884e-01(bathrooms) - 2.069822e-08(price) + 6.105036e-02(floors)

Lowest Mallow's Cp Regression Equation:

log(pi_hat/1-pi_hat) = 4.062816e-01 + 3.586136e-05(sqft_living) - 1.951235e-07(sqft_lot) - 4.660069e-05(sqft_above) + 5.570982e-05(sqft_living15) + 2.697893e-02(bedrooms) + 1.024884e-01(bathrooms) - 2.069822e-08(price) + 6.105036e-02(floors)

Lowest BIC Regression Equation:

log(pi_hat/1-pi_hat) = 4.062816e-01 + 2.937433e-05(sqft_living) - 1.880538e-07(sqft_lot) - 4.549863e-05(sqft_above) + 5.424850e-05(sqft_living15) + 2.815571e-02(bedrooms) + 1.024452e-01(bathrooms) + 6.065376e-02 (floors)

It turns out we have 2 candidate models. They all have sqft_living, sqft_lot, sqft_above, sqft_living15, bedrooms, bathrooms, and floors. The model with the best adjusted R2 and Cp has one additional predictor: price.

Selection Procedures for Model Selection

Forward selection:

The estimated regression equation based on the forward selection is:

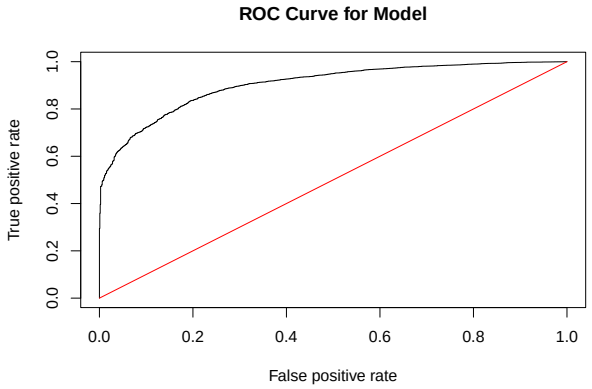
log(pi_hat/1-pi_hat) = 0.4063 + 0.1025(bathrooms) + 5.571e-05(sqft_living15) + 2.698e-02(bedrooms) + 6.105e-02(floors) - 4.660e-05(sqft_above) + 3.586e-05(sqft_living) - 1.951e-07(sqft_lot) - 2.070e-08(price)

Backward selection:

The backwards selection produced the same predictors as the forward selection; both removed condition from the model.

The estimated regression equation based on the backward selection is:

log(pi_hat/1-pi_hat) = 0.4063 + 3.586e-05(sqft_living15) - 1.951e-07(sqft_lot) - 4.660e-05(sqft_above) + 5.571e-05(sqft_living15) + 2.698e-02(bedrooms) 1.025e-01(bathrooms) - 2.070e-08(price) + 6.105e-02(floors)



Since the ROC curve is above the diagonal line, the log regression performs better than random guessing.

The AUC of 0.9037643 means the log regression performs better than random guessing. However, this AUC value is less than model1, which follows the Wald test where bathrooms is removed and condition is maintained. Since the AUC value is higher, we will go with the reduced model.

Comparing the models (the model chosen by selection procedures and the model with all predictors except bedroom included) we see that the initial full model has a higher AUC, meaning a higher predictive ability. Therefore, we choose to go with the first model based off all the predictors.

Regression Equation:

- log(pi_hat/1-pi_hat) = -5.327 + 7.797e-04(sqft_living) - 3.804e-06(sqft_lot) + 1.272e-03(sqft_above) + 8.048e-04(sqft_living15) - 7.066e-06(sqft_lot15) + 1.212(bathrooms) + 4.643e-06(price) + 4.283e-01(floors) - 2.148e-01(condition)