Stat 6021 - Project 2 - Housing Models - Group 9

Alanna Hazlett, Elisa de la Vega Ricardo, Ryan Healy

Section 1 Summary of Findings

Our Analysis was done on house sale prices for King County, Washington between May 2014 and May 2015. During the analysis our group built two models, one for the price of the house sold and the other for whether to renovate the house or not.

Our price model found that features with the largest contribution in predicting the house sale price were if the home was located on the Waterfront or it had a good grade. The smallest contributions of the features that ended up in the model were the `sqft_living` and the `age` of the house. Our `price` model is better suited for prediction than association as some of the regression coefficients such as those in `bedrooms` are the opposite of what we would expect, but the model has a great out of sample predictive (RMSE) value.

The swapping of expected coefficient signs is often a sign of multi-colinearity, which can make the coefficient estimates hard to exactly quantify but doesn't necessarily stop the model from predicting well. In fact we would expect that almost all the variables in the model have some linear dependence on price, as price is a measure of scarcity and demand and a common proxy to how we quantify the world around us including with Real Estate. If we want to really study the associative relationship to price and home sales we would need to conduct a through and careful Experimental Design with that goal in mind. However there is utility in a good prediction, especially in Real Estate.

Our Renovation Modeling, showed us that older houses are more likely to be renovated, suggesting a trend where older structures are updated more frequently to meet modern standards or preferences. Houses with a higher Construction Grade are updated more frequently, likely to maintain or enhance their already high standards and Market Values. Waterfront properties are significantly more likely to see renovations, reflecting the high value placed on maximizing the appeal and functionality of houses in desirable locations and the likely wear and tear of the Pacific Ocean.

We also found that Larger homes have a slightly higher probability of being renovated, which could be due to the greater potential for return on investment or the versatility larger spaces

offer for improvements. Also that Houses in better condition are less likely to be renovated, implying that well-maintained properties require fewer updates and that higher view quality properties are more likely to see renovations, perhaps because of the market demand for premium properties, and homeowners' desire to enhance marketability, and grow long-term value.

A note on the .Rmd, if something is commented out its useful, but we don't need all the output in the report if something wasn't useful it was deleted.

Section 2 Data and Variables

The data set contains house sale prices for King County, Washington, which includes Seattle. It includes homes sold between May 2014 and May 2015. The data set has $n = 21,603$ rows, and split 50/50 into training and test sets, with the training set having $10,801$ rows and the test set having $10,802$ rows.

Data Dictionary

- `id` - Unique ID for each home sold
- `date` - Date of the home sale
- `price` - Price of each home sold (MOLDEL_price TARGET)
- `bedrooms` - Number of bedrooms
- `bathrooms` - Number of bathrooms, where .5 accounts for a room with a toilet but no shower
- `sqft_living` - Square footage of the apartments interior living space
- `sqft_lot` - Square footage of the land space
- `floors` - Number of floors
- `waterfront` - A dummy variable for whether the apartment was overlooking the waterfront or not
- `view` - An index from 0 to 4 of how good the view of the property was
- `condition` - An index from 1 to 5 on the condition of the 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.
- `sqft_above` - The square footage of the interior housing space that is above ground level
- `sqft_basement` - The square footage of the interior housing space that is below ground level
- `yr_built` - The year the house was initially built
- `yr_renovated` - The year of the last renovation for a house (used to build target model 2)
- `zip code` - What zip code area the house is in
- `lat` - Latitude of home

- `long` - Longitude of home
- `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

Data Transformations

- `renovated`: if `year_renovated` > 0 we consider the property to have been renovated, the `year_renovated` is then dropped (MODEL_renovation TARGET)
- `yrs_renovated` : If a house has been renovated, how many years since its was renovated (MODEL_price).
- `house_age` : we look at the different from the year the data house was built to the year this data was collected (MODEL_price).
- `y_star` : the log of the price (MODEL_price).
- `sqft_lot_star`: the log of the square foot of a property lot (MODEL_price).

Data Quality

- Correct observation `15871` where `bedroom = 33.` This observation appears to be an entry error and was incorrectly entered and should have been 3, as a house with 33 bedrooms would need to be larger than 1,620 sqft and have more than 1.75 bathrooms.
- During discovery we realized **that there are observations with 0 bathrooms** and this is impractical and would be uninhabitable or commercial properties. We will remove these 11 observations: row locations: `876 1150 3120 5833 6995 9774 9855 10482 14424 19453`.

Section 3 Questions of Interest

We are interested in exploring two different relationships in our housing data, the price of the house and the whether or not it gets renovated. The first model will be an exploration of `price`, in USD and the second will be exploring the Decision to `Renovate` using data available to use from Kaggle.

In building the first model on price we found that a **logarithmic transformed response** led to a better overall model. The final model has log price, instead of just price as a target.

The data is split into equal training and testing sets in order to validate the Predictive power of our models.

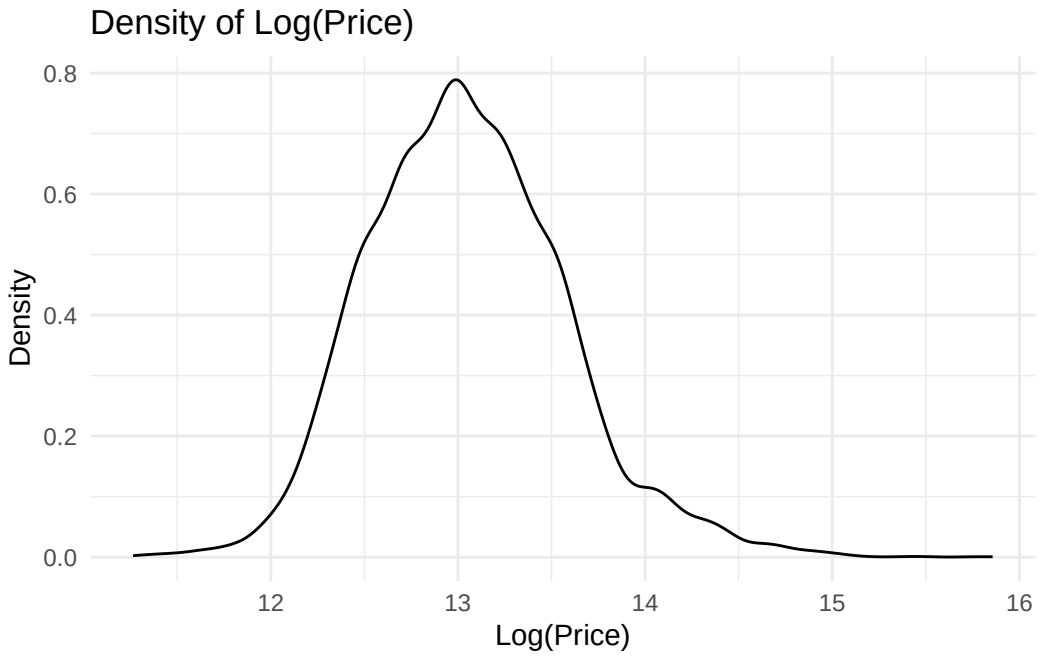
These two relationships, which specifically contain the decisions to buy, sell, refinance, or renovate a house are top of mind for most prospective buyers or current home owners. The price paid for a house dictates where someone can afford to live and in what conditions. The decision to renovate a home is often a huge undertaking requiring months of planning, a fair amount of capital, and often a disruptive demolition and installation process.

Understanding the factors influence the likelihood of a house being renovated, and what goes into price can help homeowners make informed decisions about their properties that could increase their value and appeal on the market.

Our models can used by people actually in the housing industry such as for Realtors (price) or Home Builders (renovation). An understanding of both can help with estimating a fair listing or offer price, as well which neighborhoods to target with Marketing campaigns.

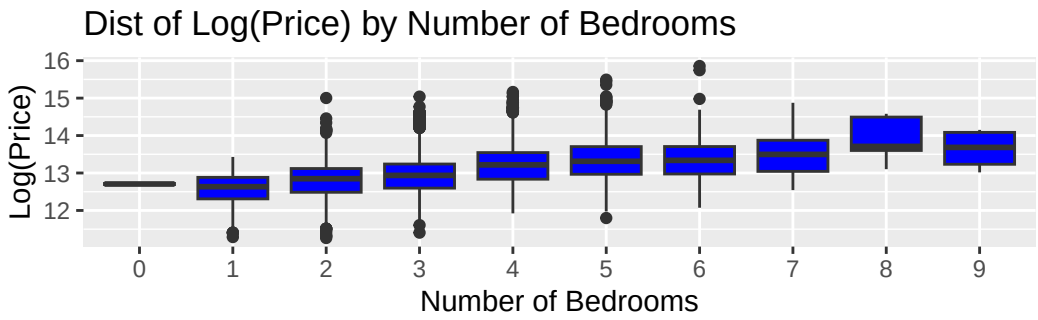
Section 4 Exploring Price

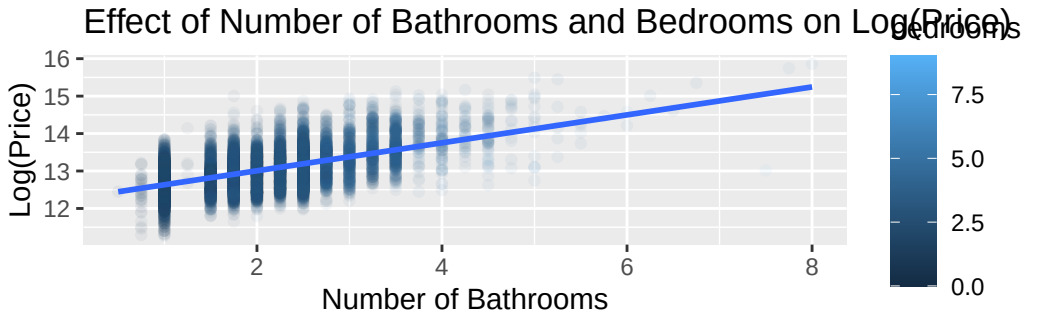
The Distribution of our Target Log(Price)



Most houses have a log(price) of 13, equating to about $440,000. The training data ranges from log(price) = 11, which is about $60,000 to log(price) = 15.5, which is about $5,400,000.

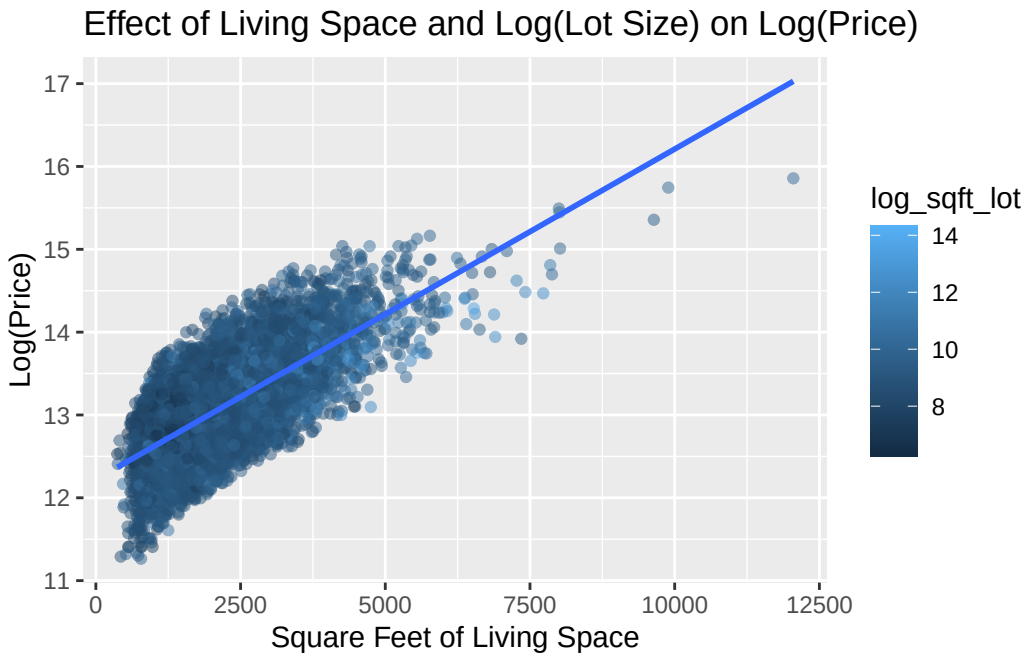
Log(Price) and Bedrooms





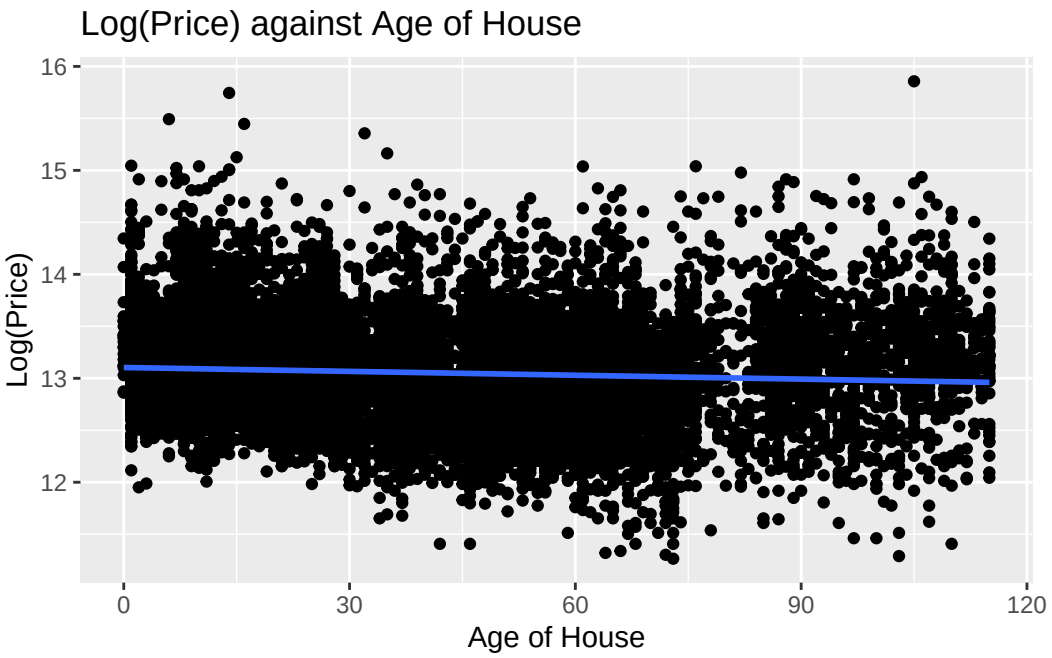
We see as the number of bedrooms increases so does the `log(price)` of a home. We see a large interquartile range as the number of bedrooms increases, indicating a larger variance in price as the number of bedrooms increases.

Log(Price), Living Space, and Log(Lot Size)



Generally we see that both `sqft_living` and `sqft_lot_star` (log) increase as price increases.

House Age and Log(Price)



It appears that the age of the a house's `age` does not play much of a role in log(price), as the `price` remains relatively constant across all house ages.

Section 5 Modeling Price

Question of interest: What factors influence the price of a house?

Initial Modeling

We removed `ID, lat`, `long`. The ID is not necessary for our analysis. lat and long are location information that can be given with the `zipcode`. Remove `sqft_lot15` and `sqft_living15` as they pertain to neighboring houses.

We perform Automated Search Procedures for preliminary model building, using Forward, Backward and Step wise Selection on AiC which show the (best) model as

price ~ `sqft_living` + `grade` + `house_age` + `waterfront` + `view` + `bedrooms` + `bathrooms` + `sqft_lot` + `condition` + `floors`

The Automated Search Procedure using (adjusted R^2, Mallow's Cp, and BIC) gave us a preliminary model as price ~ `sqft_living` + `grade` + `house_age` + `waterfront` + `view` + `bedrooms` + `bathrooms` + `sqft_lot`.

Check Multi collinearity

We check for multicollinearity as `bedrooms` has a negative sign for its coefficient estimate, which is opposite of what we would expect, namely that an increase in the number of bedrooms would mean a higher price. `Sqft_lot` also shows opposite of what we would expect.

The Variable Inflation Factors VIFs are all within acceptable limits, as they are less than 5.

Do we drop condition and floors?

We Conduct an F test to see if we can drop condition and floors from our model:

$$H_0 : \hat{\beta}_{condition} = \hat{\beta}_{floors} = 0$$
$$H_a : at\ least\ one\ \neq 0$$

$$F_0 = \frac{(SS_R(F) - SS_R(R))\ /\ r}{(SS_{res}(F))\ /\ (n-p)} = \frac{(SS_{res}(R) - SS_{res}(F))\ /\ r}{(SS_{res}(F))\ /\ (n-p)}$$
$$F_0 = \frac{(5.0763e+14 - 5.0506e+14)\ /\ 2}{5.0506e+14 /\ (10801-11)} = 27.4525$$

Finding p-value and critical value to compare t statistic to:

```
[1] 0.000000000001398215
```

```
[1] 3.690141
```

Our p-value is smaller than our significance level of 0.05, so we **reject the null hypothesis**. This supports that we should utilize the full model and keep the predictors condition and floors.

Check regression assumptions

Once we fit our regression, we will also test that our four regression assumptions below are met:

(1) **Assumption 1:** Errors have mean 0.

Assumption 1 is the most important assumption as it essentially means that the relationship between the two variables is linear, if this assumption is not met then the predicted values will be biased and will systematically under- or overestimate the true values of the response variable

(2) **Assumption 2:** The errors have constant variance.

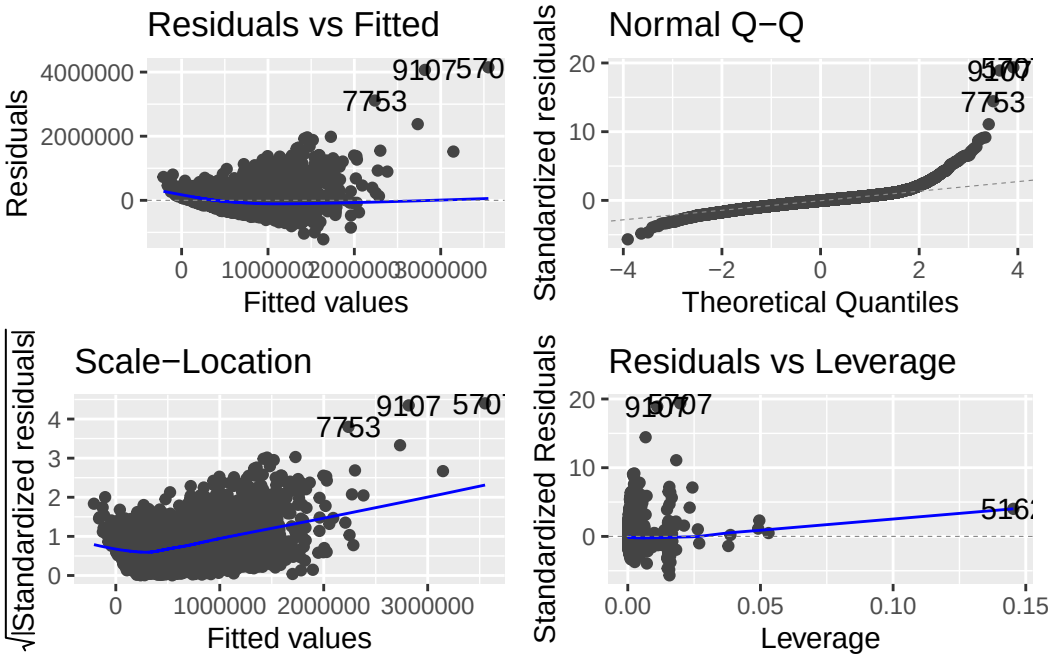
Assumption 2 is an important assumption because it means that hypothesis tests, confidence intervals and prediction intervals are reliable. If assumption 2 is violated then none of those intervals or tests are reliable, however, we will still be able to tell if our relationship is approximately linear.

(3) **Assumption 3:** The errors are independent.

Assumption 3 is important for similar reasons to assumption 2 because violating this assumption results in unreliable confidence intervals, prediction intervals and hypothesis tests.

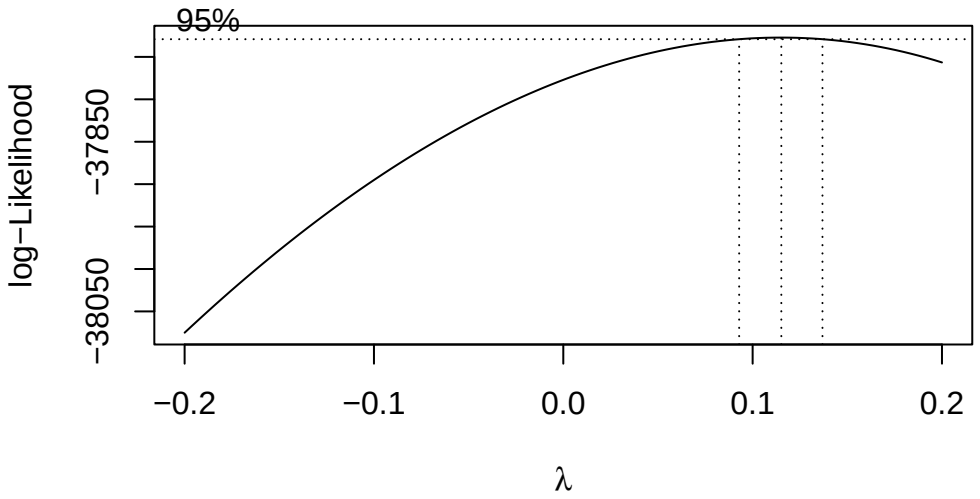
(4) **Assumption 4:** The errors are normally distributed.

The final assumption is the least important assumption as a violation of this assumption does not significantly impact the robustness of the regression model.

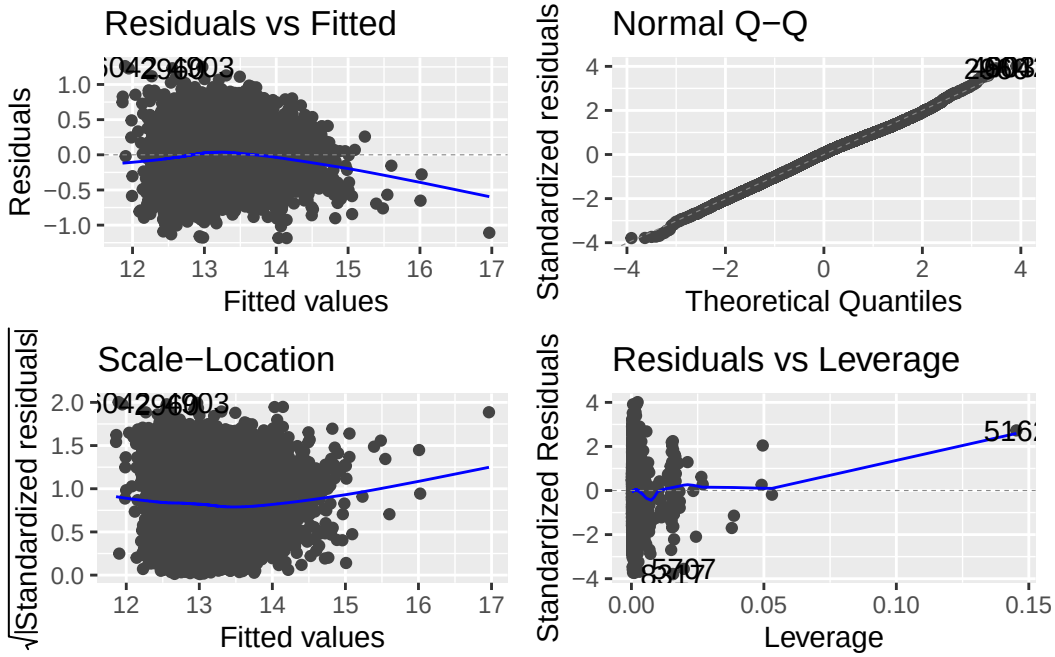


Assumption 1: The errors have a mean of zero, appears met. **Assumption 2**: Errors have constant variance for each value of the predictor, this is not met as we can see it starts small on the left hand side and gets larger as you go right.
Assumption 4: Errors are normally distributed. In Q-Q Residuals we can see that the majority of the studentized residuals follow their theoretical quantities, so we conclude that the assumption is met.

Remedial Measures:
We need to transform our y variable to stabilize the variance. As we can see in the Residuals vs Fitted Values plot the variance increases as fitted values increases, which indicates we will likely transform y with a lambda value < 1. We will use box cox plot to help us determine transformation.



We will transform our y variable with a log transformation, as zero is the closest whole value to our confidence interval of lambda. Perform the necessary transformation to the data. Re-fit the regression with the transformed variable(s) and assess the regression assumptions.



All of the regression assumptions appear to have been met. We double check the partial plot

All quantitative variables have an even amount of observations across the line, indicating that

they are linear. `sqft_lot` has an uneven variance, so we can try transforming it.

Trying log transformation on `sqft_lot` and `sqft_lot_star` appears to have a better variance now after transformation.

We will now evaluate the Model's Residuals.

High Leverage

We have quite a few high leverage points Our threshold value to calculate them is 0.002, but there is generally a gradual increase in all of our observations. There is a jump from 0.00768 to 0.0113. A high leverage point just means a predictors value is extreme, which we would expect from price data.

Detecting outliers

`Studentized residuals`: Studentized residuals greater than 2 are typically flagged as outlying. We have no studentized residuals greater than 2.

`Externally studentized residuals`: Observations with externally studentized residuals with magnitude greater than 3 are typically flagged as being outlying. We have no externally studentized residuals with magnitude greater than 3.

Influential Points

We can also test the influence with DFFITS

There are a lot of observations that are greater than our threshold, however we can see a very steady increase in the DFFITS values. There may be potential argument for `8093` `8451` `18849` `7253` as significantly influential, even more so for `7253`.

```
     bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition
8093        3      1.75        1610    10229      1          0    0         5
8451        3      1.00        1020     6120      1          0    0         4
NA         NA        NA          NA       NA     NA       <NA>   NA        NA
7253        4      3.00        3080     9601      2          0    1         3
     grade sqft_above sqft_basement yr_built yr_renovated sqft_living15
8093     8       1610             0     1961            0          2270
8451     7       1020             0     1967            0          1370
NA      NA         NA            NA       NA           NA            NA
7253    10       3080             0     1990            0          3200
     factor_bedrooms log_sqft_lot    ystar sqft_lot_star  ystar.1
8093               3     9.232982 13.38165      9.232982 13.38165
```

```
8451               3     8.719317 12.24529      8.719317 12.24529
NA              <NA>           NA       NA            NA       NA
7253               4     9.169623 13.23569      9.169623 13.23569
```

Nothing seemes immediately noticeable regarding these observations

MSE Evaluations for models utilized

We can evaluate our Final Linear Regression Model using MSE, which is interpreted approximately as the standard deviation of the prediciton error as well as the RMSE which is the same as the response variable price.

```
[1] "MSE 0.0965306251185823"
```

```
[1] "RMSE 0.310693780302378"
```

The test set Mean Squared Error or`MSE` is *0.0965306251185823* and the test set Root Mean Squared Error or `RMSE` is *0.0965306251185823*

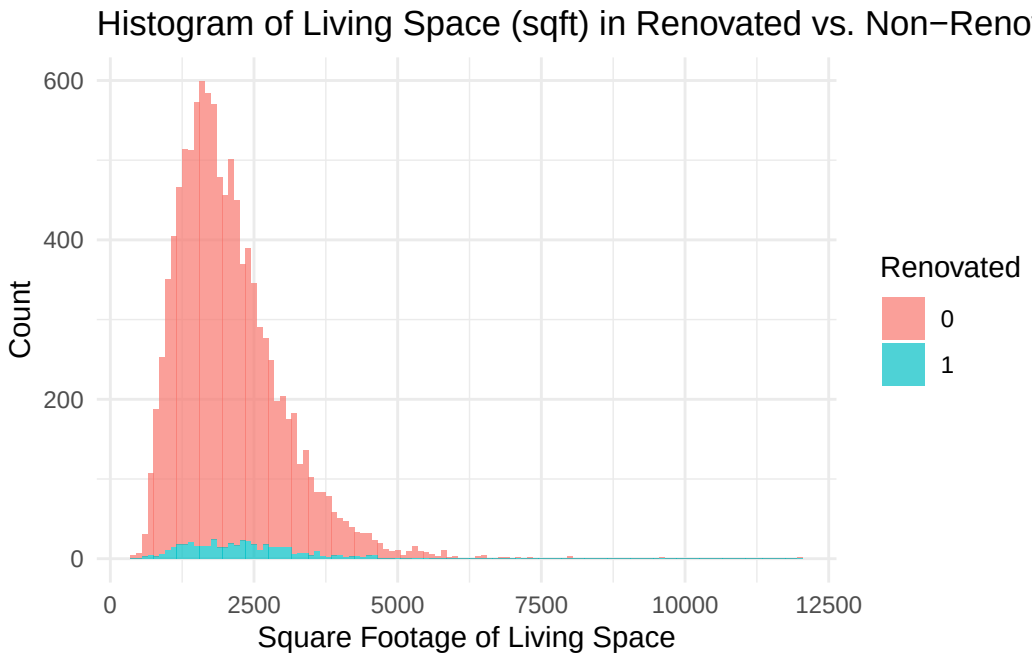
These values do a good job of predicting housing price on our test set.

Section 6 Exploring Renovations

Distribution of Square Footage of Living Space

"Histogram of Living Space Square Footage in Renovated vs. Non-Renovated Houses"

This histogram compares the distribution of square footage in living spaces between houses that have been renovated and those that have not



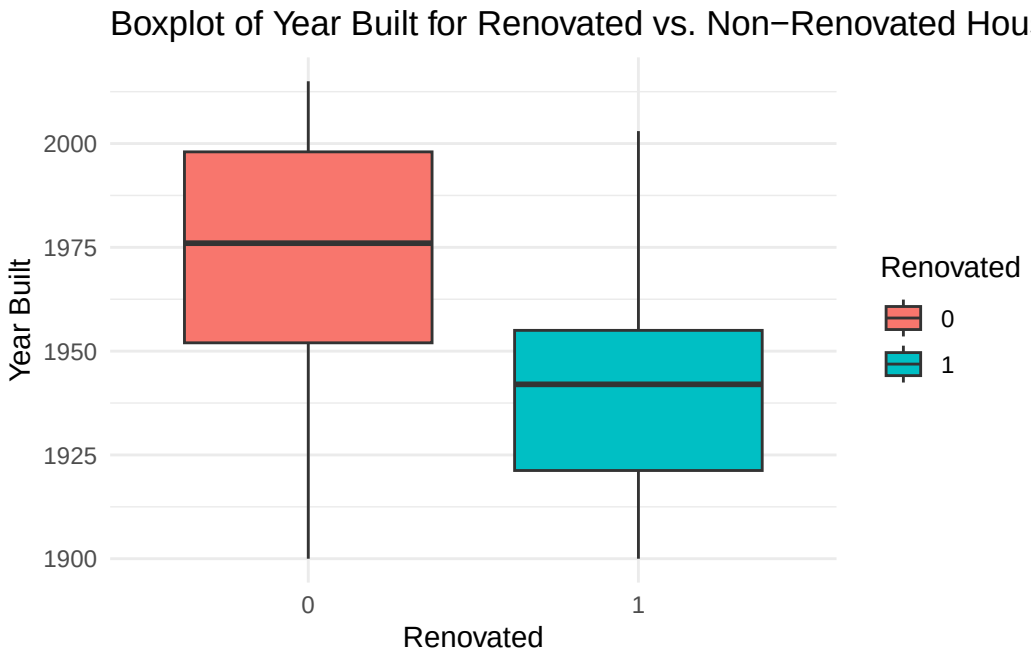
From the graph, we can observe that:

1. The majority of houses, both renovated and non-renovated, have living spaces that range from small to medium square footage, with a peak in frequency at the lower end of the scale.
2. Renovated houses (represented by the color blue) are fewer overall and are distributed across a similar range of square footage as non-renovated houses, indicating that living space size alone does not distinguish renovated from non-renovated homes distinctly.
3. The overlapping distributions suggest that while larger homes may be subject to renovation, the decision to renovate is likely influenced by a combination of factors beyond just the size of the living space.

Year Built vs. Renovation Status

"Box plot of Year Built for Renovated vs. Non-Renovated Houses"

This box plot visualizes the distribution of the years in which houses were built, categorized by whether or not they have been renovated



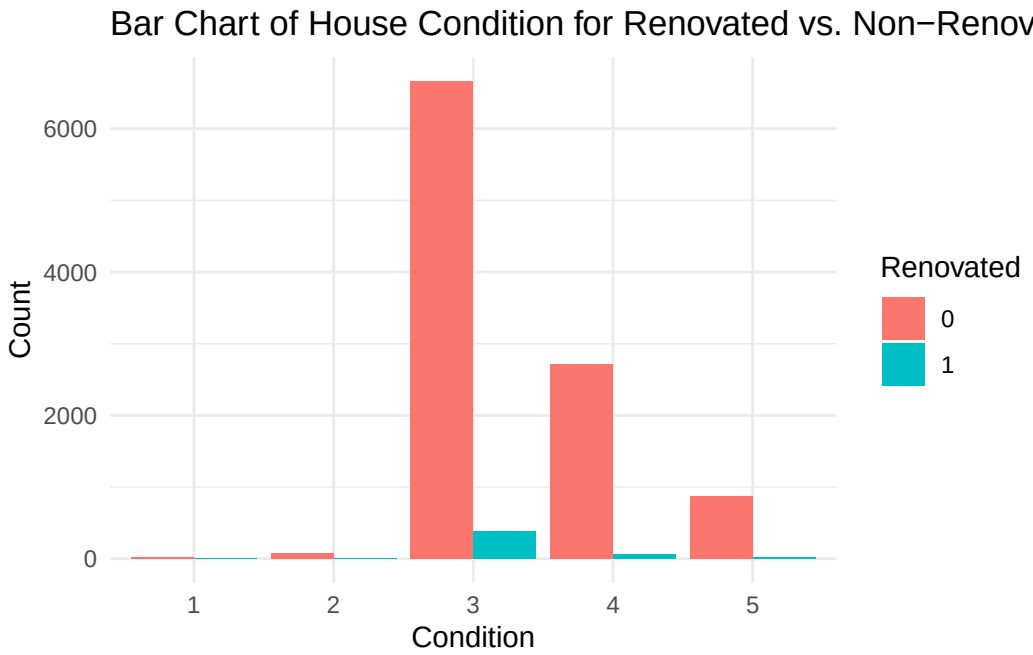
In this box plot, we can observe the distribution of the year built for houses and how it relates to renovation status:

1. Non-renovated houses (coded as "0" and displayed in red) show a wider interquartile range (IQR), suggesting a broad distribution of construction years. The median year built appears to be around the 1970s, indicating that a large proportion of non-renovated houses were built around this time.
2. Renovated houses (coded as "1" and displayed in blue) tend to have a slightly lower median year built, which suggests that renovations are more common in houses that were built earlier. The IQR for renovated houses is narrower, implying less variability in the construction year among renovated houses compared to non-renovated ones.
3. There are outliers present for both renovated and non-renovated houses, indicating some houses built much earlier or later than the typical range for each group. This could reflect historic homes that have been renovated and very new homes that have not required renovation.

Condition of Houses

"Bar Chart of House Condition for Renovated vs. Non-Renovated Houses"

This bar chart shows the counts of houses in each condition level, split by renovation status.



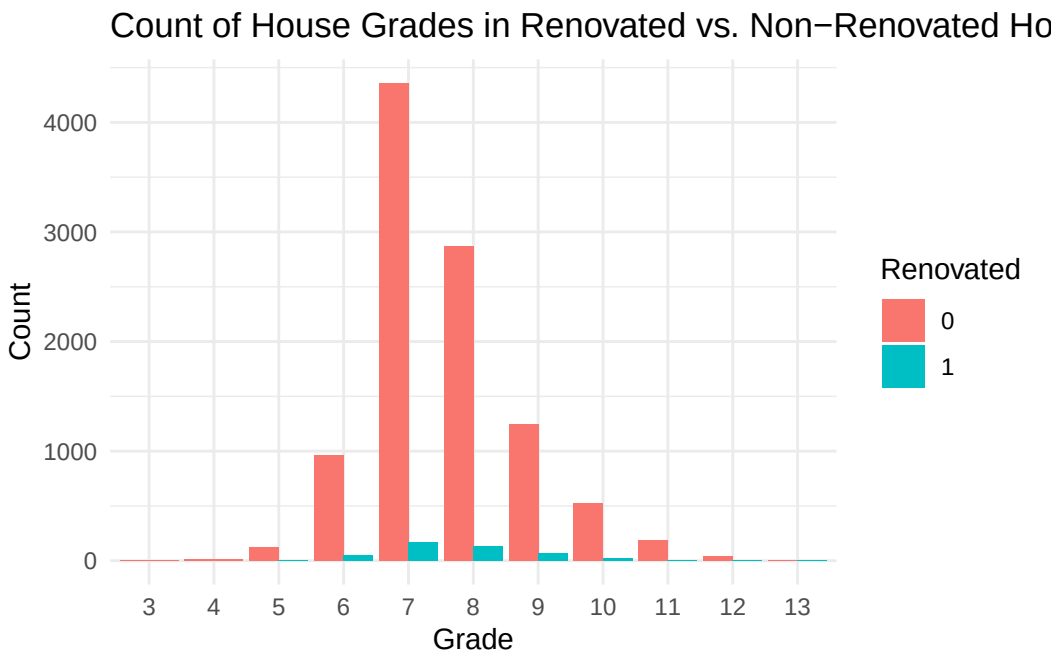
In this bar chart, which depicts house condition ratings for renovated versus non-renovated homes, we can deduce the following:

1. The vast majority of houses, regardless of renovation status, are in condition 3, indicating a moderate level of maintenance is most common in the dataset.
2. Renovated houses (shown in blue) are present in each condition category, but their numbers are significantly smaller in comparison to non-renovated houses (shown in red).
3. There is a notable presence of non-renovated houses in the higher condition ratings (4 and 5), suggesting that houses in better condition are less frequently renovated, possibly because they may not require updates as urgently as those in lower condition ratings.

House Grade and Renovation Status

"Count of House Grades in Renovated vs. Non-Renovated Houses"

This visualization represents the number of houses in each grade category, showing the difference between renovated and non-renovated homes.



This bar chart provides a comparison of the number of houses across different grades, distinguishing between renovated and non-renovated homes:

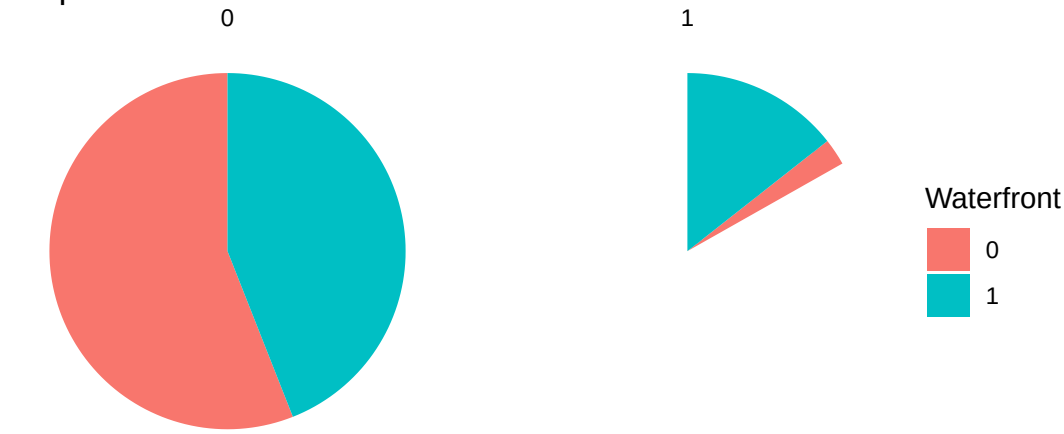
1. **Most Common Grades**: The majority of houses are concentrated in the middle-grade categories, particularly grade 7, with a substantial decline observed on either side of this peak.
2. **Renovated Homes Across Grades**: Renovated houses (blue bars) appear across all grades, but they represent a relatively small proportion of the total houses within each grade category.
3. **Non-Renovated Homes Predominance**: Non-renovated houses (red bars) vastly outnumber renovated ones in every grade category. The predominance is particularly noticeable in the most common grades, suggesting that grade alone does not strongly predict renovation status.

The chart implies that while houses of all grades are being renovated, the decision to renovate is not solely based on the grade of the house, and renovation is less common overall.

Waterfront View and Renovation Status

"Proportion of Renovated Houses with and without Waterfront Views"

This pie chart illustrates the proportion of houses with and without waterfront views that have been renovated.



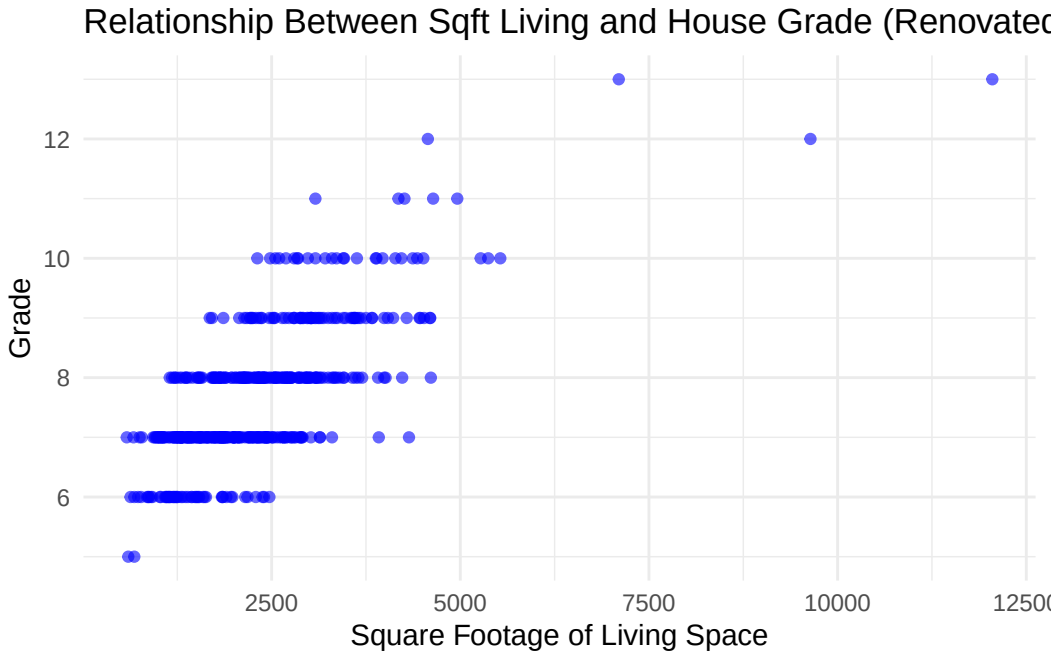
The pie chart is divided into two sections, each representing the proportion of houses with and without waterfront views and their renovation status:

1. **Non-Waterfront Houses (Left Pie Chart)**: The larger portion represents non-waterfront houses (label 0 in red), indicating that most houses in the data set do not have waterfront views. Within this category, the division between renovated and non-renovated homes is depicted, with the vast majority being non-renovated.
2. **Waterfront Houses (Right Pie Chart)**: The smaller pie chart indicates that there are fewer waterfront properties (label 1 in teal) in the data set. Among these, the proportion of renovated to non-renovated houses is visibly greater compared to the non-waterfront properties, suggesting that waterfront houses are more likely to be renovated.
3. **Comparison of Renovation Rates**: When comparing the two charts, it's evident that waterfront homes, despite being fewer in number, have a higher rate of renovation than non-waterfront homes. This aligns with the idea that the unique and desirable feature of a waterfront view increases the likelihood of a property being renovated.

Living Space vs. Grade

"Scatter Plot of Living Space vs. Grade for Renovated Houses"

This scatter plot explores the relationship between the square footage of living space and the house grade, specifically for renovated houses.



In this scatter plot, which depicts the relationship between the square footage of living space and the house grade, specifically for renovated houses, we can deduce the following:

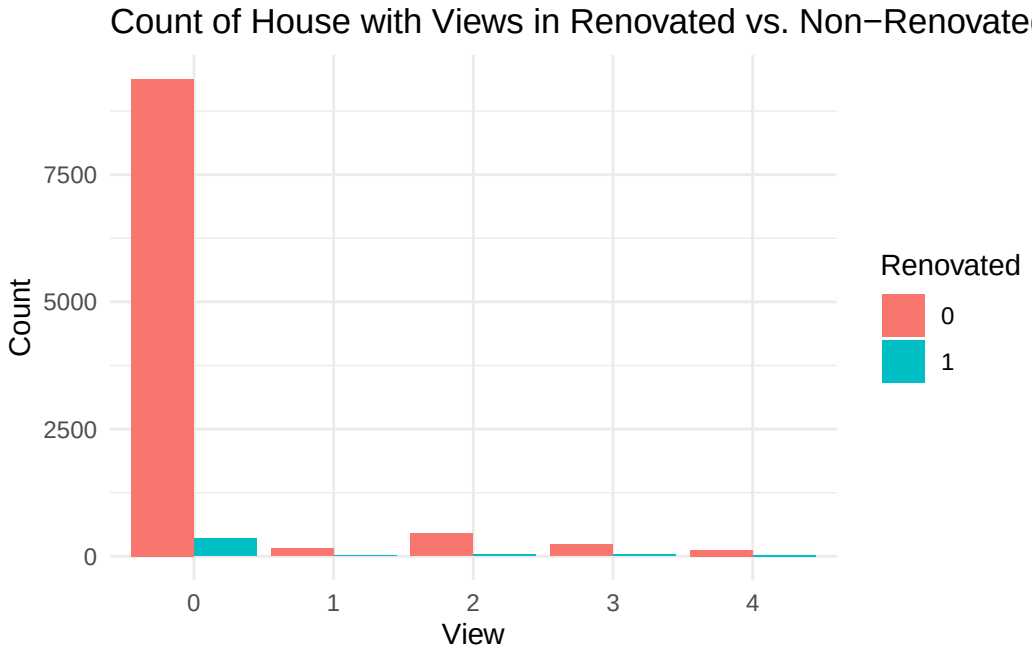
Houses between 800 to 3700 sqft and with grades between 6 to 10 are renovated in more numbers compared to houses bigger than 5000 sqft and with grades lower than 6 and higher than 10. This behavior can be attributed to several factors:

1. ***Market Appeal and Resale Value:*** Houses within the range of 800 to 3700 sqft and grades 6 to 10 may represent a sweet spot regarding market appeal and resale value. Renovations in this range will likely yield a favorable return on investment as they cater to a broad range of potential buyers seeking houses with adequate living space and moderate to high-quality finishes.
2. ***Budget Constraints:*** Homeowners should know that renovations on larger houses or those with lower grades may require a substantial financial investment. This could limit the extent of renovations they can undertake, as some homeowners may not be willing or able to afford such a significant financial commitment.
3. ***Maintenance:*** Homeowners must understand that larger houses and those with lower grades may require more extensive maintenance. This could deter homeowners from undertaking additional renovations. Homeowners should prioritize essential maintenance tasks over discretionary renovations to ensure the long-term integrity and livability of the property.

House View and Renovation Status

"Count of House with Views in Renovated vs. Non-Renovated Houses"

This visualization represents the number of houses in each view category, showing the difference between renovated and non-renovated homes.



Based on the bar plot:

1. The proportions of renovated houses across different view qualities increases as the view quality increases. While it's true that there is a trend of increasing proportions of renovated houses from view qualities 0 to 4, the absolute number of non-renovated houses with view quality 4 is lower compared to the other view qualities. Therefore, while there may be a slight trend suggesting that houses with higher view quality are more likely to be renovated, other factors such as the availability of dwellings with view quality 4, market demand, and homeowner preferences also influence the likelihood of renovation.

Section 7 Modeling Renovations

A logistic regression analysis, where the binary outcome (renovated vs. not renovated) is modeled as a function of the most relevant house characteristics.

```
Call:
glm(formula = yr_renovated_binary ~ bedrooms_filtered + price +
```

```
    yr_built + condition + sqft_living + grade + waterfront +
    view, family = binomial(), data = train)

Coefficients:
                        Estimate     Std. Error z value             Pr(>|z|)
(Intercept)       90.29040698230  3.95169536634  22.849 < 0.0000000000000002
bedrooms_filtered  0.03076286773  0.06360571888   0.484             0.628635
price              0.00000003163  0.00000017393   0.182             0.855698
yr_built          -0.04692160910  0.00207987730 -22.560 < 0.0000000000000002
condition         -1.14507271335  0.08664868801 -13.215 < 0.0000000000000002
sqft_living        0.00049831733  0.00009881547   5.043          0.000000459
grade              0.11947628840  0.07164550827   1.668             0.095394
waterfront         0.74827588649  0.38495516955   1.944             0.051920
view               0.19391560779  0.05622010529   3.449             0.000562

(Intercept)       ***
bedrooms_filtered
price
yr_built          ***
condition         ***
sqft_living       ***
grade             .
waterfront        .
view              ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3816.2  on 10800  degrees of freedom
Residual deviance: 2919.1  on 10792  degrees of freedom
AIC: 2937.1

Number of Fisher Scoring iterations: 7
```

Explaining result for model_8

In this logistic regression model output, we detect several p-values with greater significance than the conventional significance level of 0.05, indicating that these variables are not statistically significant predictors in predicting the likelihood of renovation outcome. Let's examine the non-significant variables and discuss why we may consider removing them from the model. For this purpose we will be implementing the Wald test.

*** Wald test ***

predictor variable 'bedrooms_filtered'

Ho: B1 = 0; Ha: B1 different from 0.

T statistic:

$$Z = \beta 1 - 0/se(\beta 1) = -0.03873 - 0/0.06499 = -0.59594 => Z = -0.596$$

```
[1] 0.5512153
```

p value ~ 0.55122 > 0.05

We fail to reject the null hypothesis. Therefore, we drop the `bedrooms_filtered` predictor from the regression model.

Based on the Wald test results, we may consider dropping the variable `bedrooms_filtered` from the model, as it does not contribute significantly to predicting the likelihood of renovation (yr_renovated_binary). By dropping this variable, we can simplify the model and improve its interpret ability without sacrificing predictive accuracy.

Conducting Logistic Regression

```
Call:
glm(formula = yr_renovated_binary ~ price + yr_built + condition +
    sqft_living + grade + waterfront + view, family = binomial(),
    data = train)

Coefficients:
                  Estimate     Std. Error z value             Pr(>|z|)
(Intercept) 90.46687832050  3.93715281186  22.978 < 0.0000000000000002 ***
price        0.00000002152  0.00000017261   0.125              0.900778
yr_built    -0.04697708584  0.00207790540 -22.608 < 0.0000000000000002 ***
condition   -1.14290954306  0.08646473021 -13.218 < 0.0000000000000002 ***
sqft_living  0.00052078087  0.00008699442   5.986        0.00000000215 ***
grade        0.11781081952  0.07151425439   1.647              0.099481 .
waterfront   0.74416804094  0.38500396060   1.933              0.053250 .
view         0.19191754811  0.05606485542   3.423              0.000619 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 3816.2  on 10800  degrees of freedom
Residual deviance: 2919.3  on 10793  degrees of freedom
AIC: 2935.3

Number of Fisher Scoring iterations: 7
```

Explaining result for model_7

In this new logistic regression model output, we detect another p-value with greater significance than the conventional significance level of 0.05, indicating that the `price` variable is not a statistically significant predictor in predicting the likelihood of renovation outcome. Let's examine the non-significant variable and discuss why we may consider removing it from the model. For this purpose we will be implementing the Wald test.

*** Wald test ***

predictor variable 'price'

- $Ho : \beta_1 = 0$
- $Ha : \beta_1 \neq 0$

T statistic:

$Z = \beta_1 - \beta_0 / se(\beta_1) = -1.342e-08 - 0/1.830e-07 Z = -0.07333$

```
[1] 0.9415435
```

$pvalue = 0.94154 > 0.05$

We fail to reject the null hypothesis. Therefore, **we drop the price predictor** from the regression model.

Based on the Wald test results, we may consider dropping the variable `price` from the model, as it does not contribute significantly to predicting the likelihood of renovation (`yr_renovated_binary`). By dropping this variable, we can improve the model's performance and make it more interpret able.

```
Call:
glm(formula = yr_renovated_binary ~ yr_built + condition + sqft_living +
    view + grade + waterfront, family = binomial(), data = train)

Coefficients:
              Estimate  Std. Error z value             Pr(>|z|)
```

```
(Intercept) 90.62975367  3.71529260  24.394 < 0.0000000000000002 ***
yr_built    -0.04707275  0.00193178 -24.368 < 0.0000000000000002 ***
condition   -1.14269807  0.08645947 -13.217 < 0.0000000000000002 ***
sqft_living  0.00052573  0.00007738   6.794      0.0000000000109 ***
view         0.19271114  0.05568449   3.461             0.000539 ***
grade        0.12114784  0.06629297   1.827             0.067630 .
waterfront   0.75448571  0.37590244   2.007             0.044736 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3816.2  on 10800  degrees of freedom
Residual deviance: 2919.4  on 10794  degrees of freedom
AIC: 2933.4

Number of Fisher Scoring iterations: 7
```

Explaining result for model_6

In this logistic regression model, all variables are statistically significant and contribute to explaining the variation in the outcome variable. It may be appropriate to leave all variables in the model to capture their combined effects on the likelihood of renovation. Removing any variable may lose valuable predictive information and weaken the model's performance.

#7. Model Selection Using AIC Test

The Akaike Information Criterion (AIC) is a widely used statistical measure for model selection in regression analysis. It balances a model's goodness of fit and complexity, aiming to identify the model that best explains the underlying structure of the data while avoiding over fitting. AIC seeks to find the most economical model that adequately describes the data and is useful in comparing models.

```r
aictab(cand.set = models, modnames = model.names)
```

```
Model selection based on AICc:

        K    AICc Delta_AICc AICcWt Cum.Wt       LL
model_6 7 2933.37       0.00   0.66   0.66 -1459.68
model_7 8 2935.36       1.99   0.24   0.90 -1459.67
model_8 9 2937.13       3.76   0.10   1.00 -1459.56
```

Explaining the AIC test result:

From this table we can see that model_6 is the best model.The model is much better than model_7, and model_8, as it carries 65% of the cumulative model weight and has the lowest AIC score. The cumulative model weight means that at least 65% of the variation is explained by the variables from the model.

Comparing and Selecting Models by Likelihood Ratio Test

The likelihood ratio test, also known as ΔG_2, is a statistical method used in logistic regression to assess the overall fit of a model and determine whether adding or removing predictor variables significantly improves the model's performance. In logistic regression, the ΔG_2 test helps researchers determine whether adding predictor variables enhances the model's ability to explain the variation in the outcome variable beyond what would be expected by chance.

Comparing Model_6 to Model_8

Hypotheses:

$$Ho : \beta_1 = \beta_2 = 0;$$

$Ha : \beta_1$ or β_2 different from 0.

```
[1] 0.2492924
```

```
[1] 0.8828092
```

```
[1] 5.991465
```

Explaining the Likelihood Ratio Test result for Model_8 and Model_6:

The test statistic is $\Delta G_2 = 0.2492924$ with a p value 0.8828092. With a 0.05 α significance level, since the p value is greater than the significance level, **we fail to reject the null hypothesis**. This means we should use the reduced model_6 for which we dropped `bedrooms_filtered` and `price` predictor variables, instead of the full model_8.

Generating and Explaining the Confusion Matrix for Model_6

To evaluate the performance of the logistic regression model, a confusion matrix can be used. It will help in understanding how well the model is predicting the renovation status.

```
  preds...0.5 sqft_lot view condition grade sqft_above yr_renovated
1       FALSE     5650    0         3     7       1180            0
2       FALSE     7242    0         3     7       2170         1991
3       FALSE    10000    0         3     6        770            0
4       FALSE   101930    0         3    11       3890            0
5       FALSE     6819    0         3     7       1715            0
```

```
Confusion Matrix and Statistics

          Actual
Predicted     0     1
        0 10323   431
        1    27    21

               Accuracy : 0.9576
                 95% CI : (0.9536, 0.9613)
    No Information Rate : 0.9582
    P-Value [Acc > NIR] : 0.6251

                  Kappa : 0.0766

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.99739
            Specificity : 0.04646
         Pos Pred Value : 0.95992
         Neg Pred Value : 0.43750
             Prevalence : 0.95816
         Detection Rate : 0.95566
   Detection Prevalence : 0.99556
      Balanced Accuracy : 0.52193

       'Positive' Class : 0
```

Confusion Matrix

Here's the breakdown:

- **True Negatives (TN)**: 10,323 - The model correctly predicted that these many houses would not be renovated.
- **False Positives (FP)**: 27 - The model incorrectly predicted these houses would be renovated, but they were not.
- **True Positives (TP)**: 21 - The model correctly predicted that these houses would be renovated.
- **False Negatives (FN)**: 431 - The model incorrectly predicted these houses would not be renovated, but they were.
- **Accuracy**: 0.9596 - Accuracy measures the proportion of correctly predicted cases (both true positives and true negatives) out of the total cases. In this case, it indicates that the model correctly predicted approximately 95.96% of all cases.
- **Sensitivity**: 0.99739 - Sensitivity measures the proportion of actual positive cases (1s) correctly predicted as positive by the model. It represents the model's ability to identify positive cases out of all positive ones correctly. The model is highly sensitive, indicating that it correctly identifies a high proportion (approximately 99.73%) of positive cases.
- **Specificity**: 0.04646 - Specificity measures the proportion of actual negative cases (0s) correctly predicted as unfavorable by the model. It represents the model's ability to identify negative instances out of all actual negative cases correctly. In this case, the specificity is low, suggesting that the model has a high rate of false positives (predicting 1 when the actual is 0).

Confusion Matrix Summary:

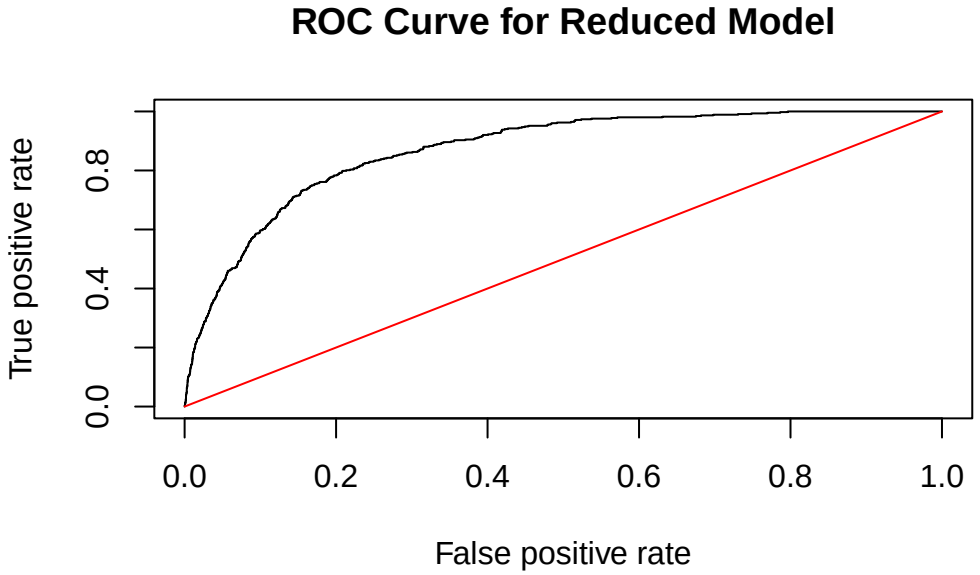
- The model shows high accuracy but exhibits a trade-off between sensitivity and specificity.
- The high sensitivity indicates the model's strong ability to detect positive cases.
- Specificity is low, suggesting a high rate of false positives.

In practical terms, while the model performs well in identifying positive cases, it may need improvement in reducing false positives (FP) to enhance its overall effectiveness, especially in scenarios where a balanced prediction of both positive and negative cases is crucial.

ROC Curve and AUC

A ROC curve helps evaluate how well a model discriminates between the positive and negative classes across different classification thresholds. If the ROC curve lies above the diagonal line (the line of random guessing), it indicates that the model performs better than random guessing. The greater the area under the ROC curve (AUC), the better the model's discrimination ability.The further the ROC curve is from the diagonal line towards the top-left corner, the better the model's performance.

The Area Under the ROC Curve (AUC) is a metric used to evaluate the performance of a binary classification model. Higher AUC values (close to 1) indicate better model performance in terms of discrimination between the positive and negative classes.



A curve that is above the diagonal indicates the model does better than random guessing. Our ROC curve is above the diagonal so it does better than random guessing in terms of discriminating between the positive and negative classes across different classification thresholds.

```
[[1]]
[1] 0.8708405
```

The AUC is around 0.8708405, which is greater than 0.5 and close to 1, which implies that the model performs better than random guessing in terms of discrimination between the positive and negative classes.

Interpretation of Coefficients

```
Call:
glm(formula = yr_renovated_binary ~ yr_built + condition + sqft_living +
    view + grade + waterfront, family = binomial(), data = train)

Coefficients:
               Estimate  Std. Error z value             Pr(>|z|)
(Intercept) 90.62975367  3.71529260  24.394 < 0.0000000000000002 ***
```

```
yr_built    -0.04707275  0.00193178 -24.368 < 0.0000000000000002 ***
condition   -1.14269807  0.08645947 -13.217 < 0.0000000000000002 ***
sqft_living  0.00052573  0.00007738   6.794      0.0000000000109 ***
view         0.19271114  0.05568449   3.461             0.000539 ***
grade        0.12114784  0.06629297   1.827             0.067630 .
waterfront   0.75448571  0.37590244   2.007             0.044736 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3816.2  on 10800  degrees of freedom
Residual deviance: 2919.4  on 10794  degrees of freedom
AIC: 2933.4

Number of Fisher Scoring iterations: 7
```

Logit Regression Model Equation

log odds(renovated/no renovated)= 90.62975367 - -0.04707275(yr_built) - -1.14269807(condition) + 0.00052573(sqft_living) + 0.19271114(view) + 0.12114784grade) + 0.75448571(waterfront)

- **(Intercept)**:
  - **Estimate**: 90.63
  - This is the log-odds of a house being renovated when all other predictor variables are held at zero. The large positive value suggests a high baseline likelihood in the context of the model, although this scenario might not be practically possible given the nature of the predictors (like 'yr_built').
- **yr_built**:
  - **Estimate**: -0.047
  - **Interpretation**: Each additional year in the age of the house decreases the log-odds of the house being renovated by -0.04707275. This indicates that newer houses are less likely to be renovated, perhaps because they require less updating.
- **condition**:
  - **Estimate**: -1.14

  - **Interpretation**: Higher condition ratings decrease the likelihood of renovation. A one-unit increase in the condition rating decreases the log-odds of renovation by -1.14269807, possibly because well-maintained houses need fewer renovations.

- **sqft_living**:
  - **Estimate**: 0.00052573
  - **Interpretation**: Each additional square foot in living area increases the log-odds of the house being renovated by 0.00052573. Larger houses might be renovated more frequently, possibly to update or repurpose space as needs change.

- **view**:
  - **Estimate**: 0.19
  - **Interpretation**: Higher view quality increases the log-odds of the house being renovated by 0.19271114. This indicates that houses with high quality views are more likely to be renovated, perhaps because the intrinsic value associated with scenic views, market demand for premium properties, and homeowners' desire to enhance enjoyment, marketability, and long-term value.

- **grade**:
  - **Estimate**: 0.12
  - **Interpretation**: Higher house grades, which indicate better overall construction and design quality, increase the likelihood of renovation. Each step up in grade increases the log-odds of renovation by 0.12114784.

- **waterfront**:
  - **Estimate**: 0.75
  - **Interpretation**: Houses with waterfronts are significantly more likely to be renovated. The presence of a waterfront increases the log-odds of a renovation by 1.766, which suggests that these desirable properties are often upgraded to maximize value.

Model Fit and Significance

- **p-values**: All predictors have p-values well below 0.05, indicating that they are statistically significant and likely have a meaningful impact on whether a house is renovated.
- **Residual deviance**: Lower than the null deviance, indicating that the model with predictors fits the data better than a model without any predictors.
- **AIC**: The Akaike Information Criterion value is 2933.4, which can be used for model selection purposes; lower AIC values generally indicate a better model.

Show Prediction on New House

In this data analysis project, we have produced a regression model to predict the likelihood of renovating a house based on critical factors such as the house's year of construction, condition, view quality, waterfront access, and the house's grade and square feet of living space. Let us apply this model to a house built in 1975 with a condition rating of 4, a grade of 7, and 1160 square feet of living space. This house, built in 1975, has no significant view and no waterfront access.

The logit regression equation:

log odds(renovated/no renovated)= 90.62975367 - -0.04707275(yr_built) - -1.14269807(condition) + 0.00052573(sqft_living) + 0.19271114(view) + 0.12114784(grade) + 0.75448571(waterfront)

Plugging in the values for our specific house

log odds(renovated/no renovated)= 90.62975367 + -0.04707275(1975) + -1.14269807(4) + 0.00052573(1160) + 0.19271114(0) + 0.12114784(7)) + 0.75448571(0)

log odds(renovated/no renovated)= -5.451838

```
[1] 0.004288415
```

```
[1] 0.004270103
```

Given its characteristics, the resulting log odds of approximately -5.451838suggest a lower probability of renovation for this particular house. Based on the calculated odds of 0.004270103, the model suggests a very low likelihood (or probability of approximately 0.42%) that this specific house, given its characteristics, will undergo renovation. This indicates that, statistically speaking, the odds are heavily against renovation for this particular property. In simpler terms, our model predicts that this house, built in 1975 with a condition rating of 4, no significant view and no waterfront access, among other specified attributes, is less likely to undergo renovation compared to homes with different characteristics. This analysis not only showcases the practical application of statistical modeling in real estate but also provides valuable insights into the factors that influence renovation decisions for residential properties. These insights could potentially guide future renovation strategies and investment decisions.