Group 4 Project 2 Report

Pratham Choksi, Emmanuel Leonce, Vicky Singh, Alec Pixton

Section 1: Summary

Everyone has walked by a really nice house and thought “Wow, that house must cost at least $1 million.” But how often do we stop to think about what actually makes a house worth $1 million? Our team began investigating this by asking what variables are most influential to the price of a house. We started by obtaining data on homes sold in King County, USA. To determine our starting point, our team created a correlation matrix(Figure 1), a tool used to help determine what variables are related. From the matrix, we found that the sqft of the home had the strongest relationship with price, a correlation of 0.7. To determine if the relationship is more than just chance, we looked at a scatterplot of the sqft of living space and the price(Figure 3). We found that as sqft increased price also increased. However, for larger homes, the spread of prices is much higher.

Another variable highly correlated with price is the sqft of the house above ground level which has a correlation of 0.6. We created a scatterplot of sqft above ground against price(Figure 4) which is very similar to Figure 3. The correlation between sqft of the house and sqft of the house above ground is 0.87. During our modeling of variables against the price, we calculated the VIF of each variable. This number is a measurement of how much these variables overlap in their prediction of price. The VIF for sqft_living is 7.314256 and for sqft_above is 5.148665. All these factors together lead us to conclude that sqft_living and sqft_above are so closely related that we cannot include both of them, called collinearity. Moving forward, we chose to only work with the total sqft of the house. We determined our model for price had an average error of about $233,193. This means that it is poor at predicting the actual price of a house, but could be useful for determining value over a high threshold.

We became interested in 2 other variables that may explain some variation in the price when used alongside the sqft. We suspected that a home on the waterfront was more likely to be a $1 million home. We also thought that having the best view would directly impact $1 million status. To evaluate these hypotheses, we created a logistic regression model using sqft of the home, the view, and being on the waterfront as predictors. We found that homes on the waterfront have 4.67 times the odds of being a $1 million home. These odds indicate that being on the waterfront does improve the chances a home is worth $1 million. Homes with the best views have 10.19 times the odds of being a $1 million home over homes with poor views. Once again, the model indicates that a better view does indicate a better chance of the home being worth $1 million.

One way our team validated our model, was by calculating the area under the curve(AUC) of the receiver operating characteristic curve. This is simply a value from 0 to 1, that tells us how our model compares to randomly guessing, which has an AUC of 0.5. Our model has an AUC of 0.939, which indicates that our model is much better than guessing if a house is worth $1 million.

Section 2: Variables

Existing variables

- id - Unique ID for each home sold
- date - Date of the home sale
- price - Price of each home sold
- 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. 0 stands for not on the waterfront while 1 stands for on the waterfront.
- 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 house's last renovation
- zipcode - What zipcode area the house is in
- lat - Latitude
- long – Longitude
- 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

Created Variables

- above_million – if the price is above $1,000,000 then 1, otherwise 0. This is used to determine if a home sold for over one million dollars.
- train - the portion of the dataset used to create the model
- test - the portion of the dataset used to test the model's accuracy
- grade(mutated) – grade was turned into 2 categories, low if the grade is below 8.5 and high if it is above 8.5

Section 3: Questions of Interest

Question 1: What variables are most influential to the price of the house? The response for this question is price while the predictors initially include all usable variables provided. This question warrants investigation because understanding the relationship between house features and their prices provides essential insights for all real estate market stakeholders. It assists potential buyers in evaluating the value they receive in comparison to the features of a house. Simultaneously, it aids sellers in setting appropriate prices for their properties by aligning with market trends. This understanding facilitates informed decision-making and fosters a transparent and efficient marketplace. By identifying which variables are most influential, we can then determine which variables are worth further investigation.

Question 2: How does being on the waterfront, the quality of view, and the sqft_living influence the likelihood of a house being sold above $1 million? The response for this question is whether or not the house sold for more than $1 million. The predictors selected are waterfront, view, and sqft_living. We selected the variable with the strongest relationship to price, sqft_living, from part 1. We selected the other 2 due to our interest in their effect on $1 million status. We chose waterfront since we expected more homes on the waterfront to be $1 million homes. View was included for a similar reason, we suspected that homes with the best views were also more likely to be $1 million homes.

Section 4: Question 1 Visualizations

The correlation matrix highlights key relationships between house features and pricing. A prominent correlation of 0.7 between sqft_living and price suggests larger homes command higher prices. Meanwhile, waterfront and view features mildly correlate with price, indicating a modest impact on value. Strong correlations between 'sqft_above', 'grade', and 'sqft_living' emphasize the importance of living space in property valuation. This matrix informs us where to look further to develop our model.

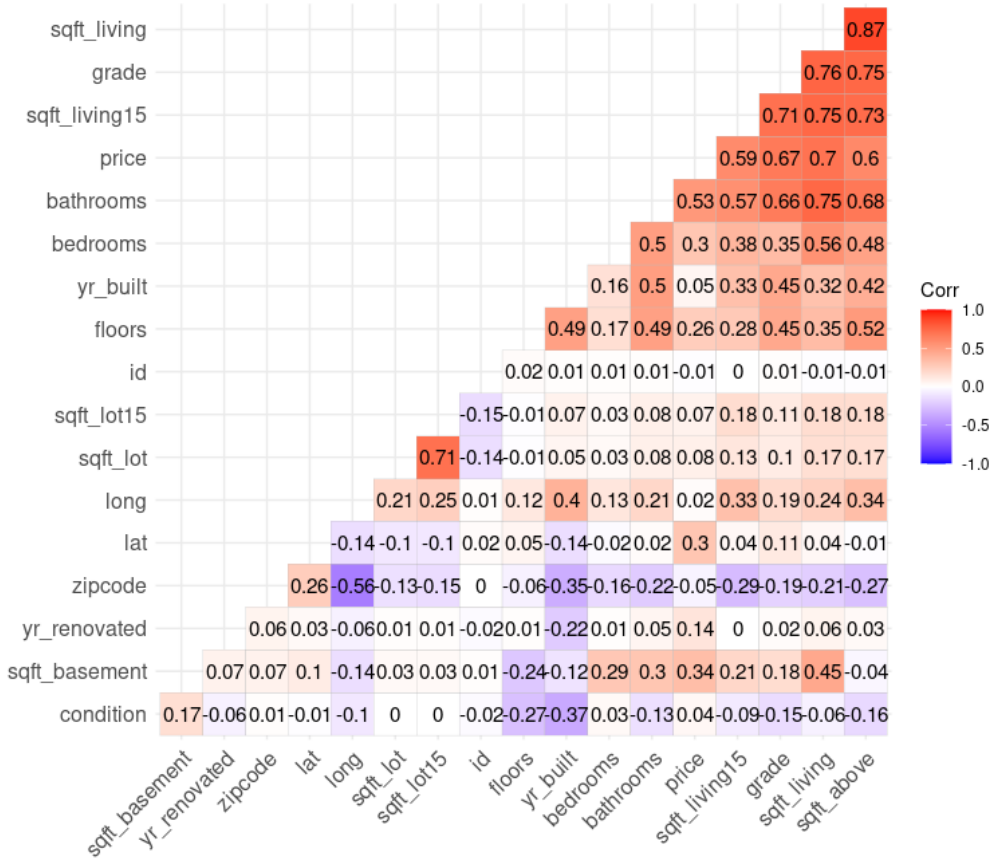

Figure 1

From the boxplot of Distribution of housing price by Grade, we observe that as housing grade increases, the housing price also increases. We have a left skewed boxplot distribution of housing grade above 8.5 with a median housing price around $750,000, which is higher compared to the left skewed boxplot distribution of housing grade below 8.5. Which has a median housing price around $375000. We have outliers in both boxplots, which indicate unusual observations that may warrant further investigation since they might indicate measurement errors, data entry errors, distribution characteristics (long tails), and sampling issues (the data point actually belongs to another population).

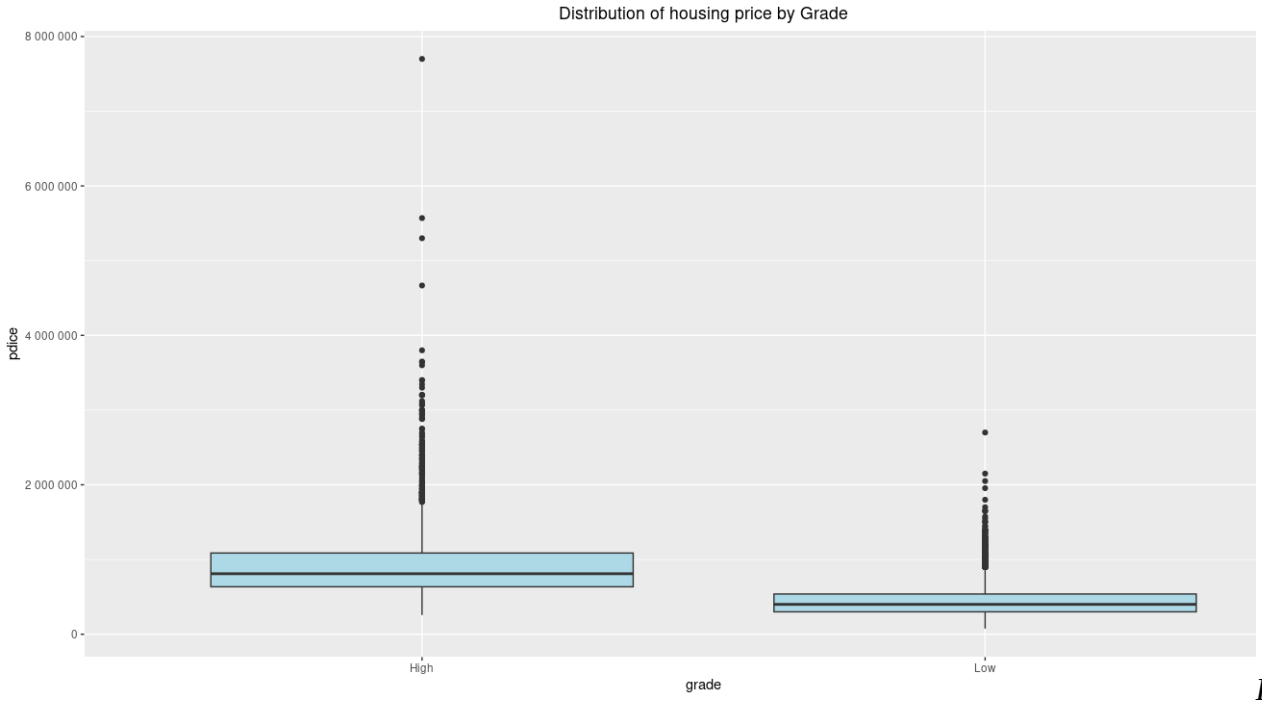

Figure 2

This scatter plot illustrates the relationship between the square footage of living space and the prices of homes. A dense cluster of data points shows that as the living square footage increases, there is a general upward trend in price. However, the spread of data points becomes wider with larger living spaces, indicating a greater variability in price for larger properties. This trend highlights the square footage as a significant factor in housing prices, with larger homes tending to be more expensive, yet with a varied price range as size increases. Additionally, the dispersion of price points suggests that factors beyond size, such as location, quality of construction, level of luxury, etc., also play a significant role in determining the price.

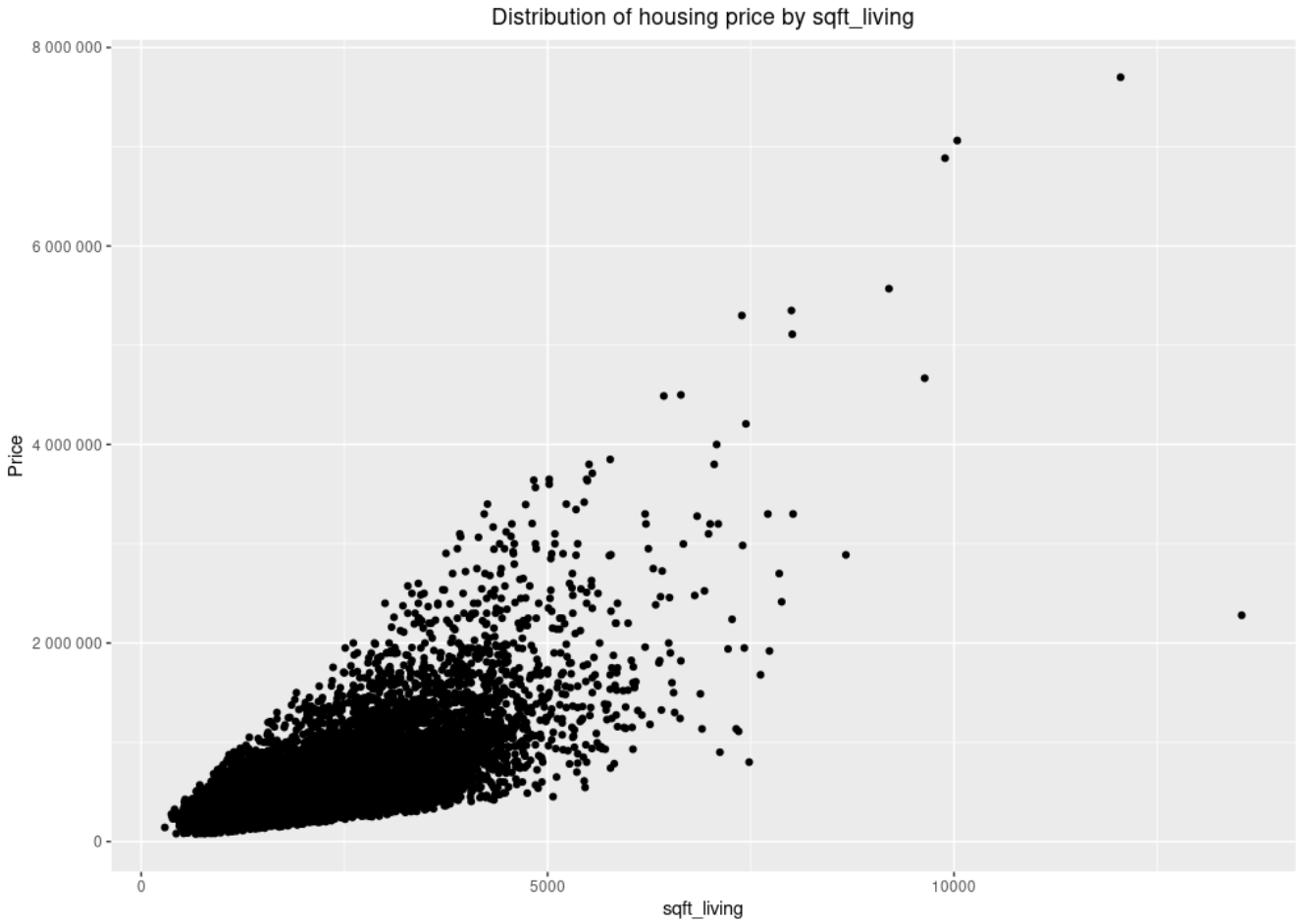

Figure 3

This scatter plot demonstrates the relationship between the housing square feet above and the price of houses. A dense cluster of data points shows that as the housing square feet above increases, there is a general upward trend in price. However, the spread of data points becomes wider with larger square feet above, indicating a greater variability in price for larger properties. This trend highlights the square feet above as a significant factor in housing prices, with larger homes tending to be more expensive, yet with a varied price range as size increases. One major reason we chose not to include both of these variables for further analysis in question 2 was due to their clear collinearity. The correlation obtained from the matrix for sqft_living and sqft_above is 0.87.

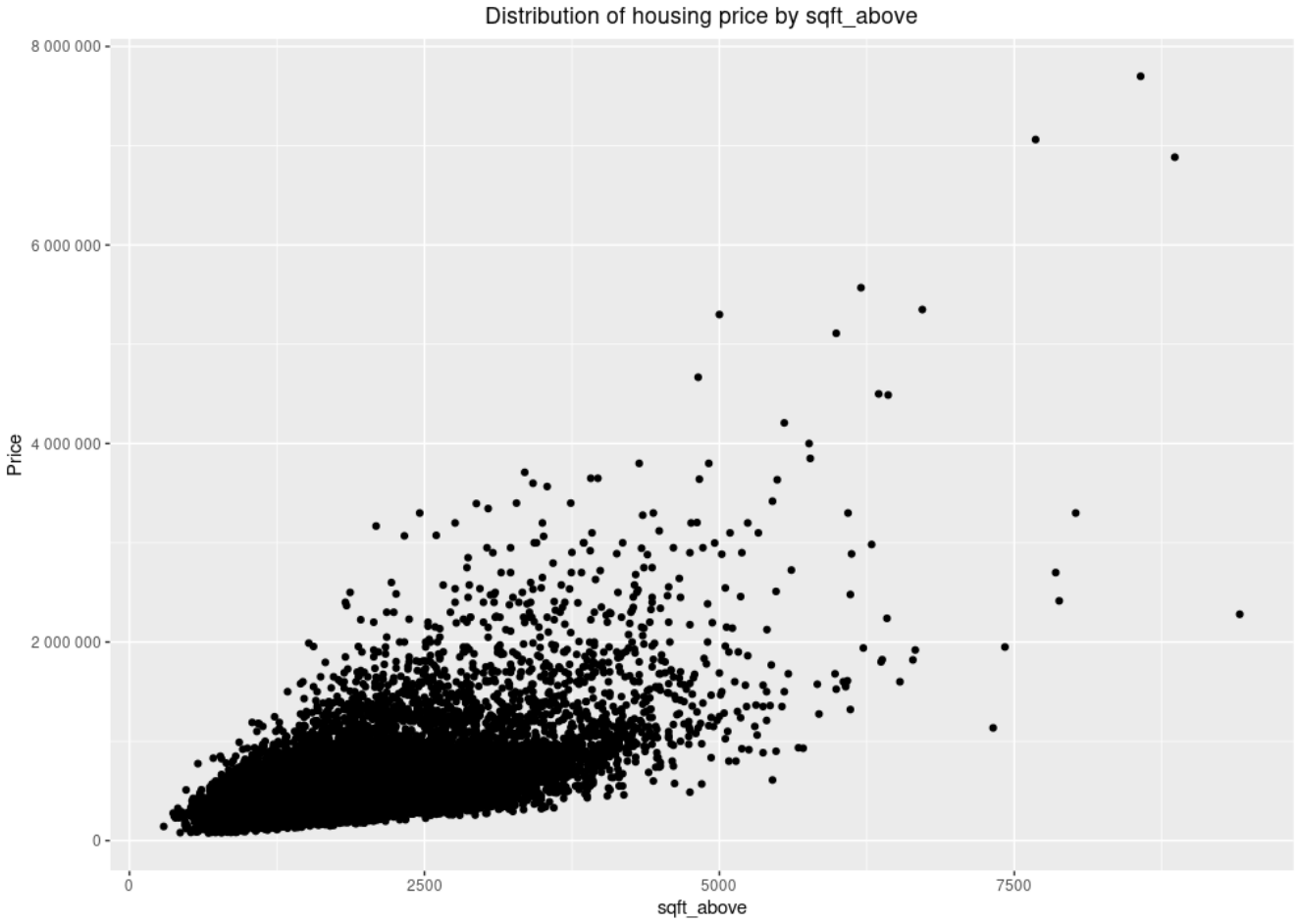

Figure 4

Section 5: Question 1 Model

For question 1, since we wanted to compare many predictors to price, a multiple linear regression felt right for this question. We first removed the variables that are not quantitative or we feel are not relevant to price based on the visualizations. We created a multiple linear regression that compared price with: "bedrooms", "bathrooms", "sqft_living", "sqft_lot", "floors", "sqft_above", "yr_built", "yr_renovated", "zipcode", "lat", "long", "sqft_living15", "sqft_lot15".

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

Residuals:
     Min       1Q   Median       3Q      Max
-1008890   -88396    -6845    73930  4475919

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -7.379e+06  2.577e+06  -2.864  0.00419 **
bedrooms       -3.292e+04  1.689e+03 -19.495  < 2e-16 ***
bathrooms       4.116e+04  2.910e+03  14.143  < 2e-16 ***
sqft_living     1.484e+02  3.862e+00  38.421  < 2e-16 ***
sqft_lot        1.974e-01  4.301e-02   4.590 4.46e-06 ***
floors          2.467e+04  3.212e+03   7.682 1.63e-14 ***
sqft_above      2.036e+01  3.817e+00   5.334 9.69e-08 ***
yr_built       -1.688e+03  6.040e+01 -27.950  < 2e-16 ***
yr_renovated    1.645e+01  3.227e+00   5.097 3.49e-07 ***
zipcode        -3.908e+02  2.936e+01 -13.312  < 2e-16 ***
lat             5.694e+05  9.477e+03  60.083  < 2e-16 ***
long           -1.799e+05  1.173e+04 -15.334  < 2e-16 ***
sqft_living15   5.417e+01  2.928e+00  18.501  < 2e-16 ***
sqft_lot15     -2.607e-01  6.580e-02  -3.961 7.48e-05 ***
above_million1  6.354e+05  5.914e+03 107.437  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 180800 on 21598 degrees of freedom
Multiple R-squared:  0.7576,    Adjusted R-squared:  0.7575
F-statistic:  4822 on 14 and 21598 DF,  p-value: < 2.2e-16
```

Figure 5

What we found with our model:

Based on the summary of the model, we can see that the r^2 is 0.6278, meaning that 67.7% of the variance in price can be explained by these variables.

This model also shows that all the predictors are statistically significant since they all have p-values less than 0.05. The t-values provide us with more information on which predictors are most influential. It seems like lat, sqft_living, bathrooms, and sqft_living15 are the most influential predictors for price.

Since the the p-value is less than 2.2e-16, we can conclude that the model is statistically significant at the 0.05 level.

To determine if this model is useful we created null and alternative hypotheses and performed an ANOVA test:

```
Analysis of Variance Table

Response: price
                 Df     Sum Sq    Mean Sq    F value    Pr(>F)
bedrooms          1 2.7696e+14 2.7696e+14  8471.9173 < 2.2e-16 ***
bathrooms         1 5.3190e+14 5.3190e+14 16270.3055 < 2.2e-16 ***
sqft_living       1 6.6776e+14 6.6776e+14 20426.1232 < 2.2e-16 ***
sqft_lot          1 5.0788e+12 5.0788e+12   155.3546 < 2.2e-16 ***
floors            1 1.4364e+10 1.4364e+10     0.4394  0.507430
sqft_above        1 2.6019e+12 2.6019e+12    79.5910 < 2.2e-16 ***
yr_built          1 1.3640e+14 1.3640e+14  4172.4761 < 2.2e-16 ***
yr_renovated      1 1.2404e+12 1.2404e+12    37.9425 7.415e-10 ***
zipcode           1 3.4424e+11 3.4424e+11    10.5301  0.001176 **
lat               1 1.5126e+14 1.5126e+14  4626.8203 < 2.2e-16 ***
long              1 2.5517e+13 2.5517e+13   780.5323 < 2.2e-16 ***
sqft_living15     1 2.9217e+13 2.9217e+13   893.7228 < 2.2e-16 ***
sqft_lot15        1 1.2117e+12 1.2117e+12    37.0658 1.161e-09 ***
above_million     1 3.7735e+14 3.7735e+14 11542.7144 < 2.2e-16 ***
Residuals     21598 7.0607e+14 3.2691e+10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Our null hypothesis is that all regression coefficients are equal to 0 while our alternative is that regression coefficients are not equal to 0. All predictors except floors have p-values less than 0.05, which suggests that all predictors except floors have an effect on price. We can see that the majority of the F values except floors are greater than 1, meaning that we can reject the null hypothesis for all predictors except floors. Since we reject the null hypothesis we can see that regression coefficients are not equal to 0 and the model is accurate.

How we improve the model:

Since floors was not a great predictor we updated the model to remove floors.

Since this is a multiple linear regression, It is good to check for multicollinearity This will tell us which of the predictors are highly correlated with each other. One quick way to check is to see if the standard error is large for any of the predictors.

With the new model, we found that quite a few standard errors are large such as bedrooms, bathrooms, lat and long. This means we have strong multicollinearity. To better look we calculated the VIFs

```
bedrooms      bathrooms    sqft_living       sqft_lot     sqft_above       yr_built   yr_renovated        zipcode
1.627312       3.084822       7.598861       2.097568       5.151807       1.924852       1.105766       1.614849
     lat           long  sqft_living15     sqft_lot15 above_million1
1.128685       1.785963       2.643757       2.130822       1.461105
```

Figure 6

The largest VIF is with sqft_living with 7.314256 and sqft_above with 5.148665. VIFs above 5 indicate a moderate degree of multicollinearity, while VIFs above 10 indicate a strong degree of multicollinearity.

To summarize what we have seen:

- The ANOVA F test is significant, and a lot of the t tests are significant.
- We see huge standard errors for the estimated coefficients.
- The largest VIF is 7.314256

Collectively, there is a high degree of multicollinearity in this model.

Since there was a high degree of multicollinearity, especially with sqft_living and sqft_above, We modified the model to include these two predictors as well as latitude, longitude, and bathrooms since those seem to be the most influential in our beginning model.

To assess the predictive ability of the new model:

After some experimentation and trial and error we realized that since there is extremely high multicollinearity, We are able to get a lower mse with all the predictors than just the most influential, however, the difference is small. The root mse we got with all variables is 233193 while the root mse for the most influential predictors is 243695 to 264416.

This means that given that the price variable ranges from $75,000 to $7,700,000, the RMSE value suggests that on average, our model's predictions are off by about $233,193, showing the model is moderately accurate.

Conclusion:

From our model, we found sqft_living to be the most influential predictor of price. Bathrooms, sqft_above, and location also seem to have a big impact. We tested the model's accuracy using an anova test and found it to be highly accurate in determining which predictors are influential. To better improve the model and see how much of an impact these found influential predictors have on price, we checked for multicollinearity and assessed the predictive ability of the new model with the most influential predictor. Even though the predictive ability is not as accurate as we want it to be, it tells us that it is difficult to get very accurate price predictions since there are many predictors that influence each other and have a high influence on price. However, we can see that sqft_living + sqft_above + bathrooms seems to cause the most volatility to price changes.

Section 6: Question 2 Visualizations

This visualization illustrates the distribution of homes based on their "view" categories, with a distinction between homes that sell for over $1 million and those that do not. The majority of homes fall into category 0 for View, where most homes are selling below $1 million. As the view quality improves, the proportion of homes selling for more than $1 million significantly increases, particularly in categories 3 and 4. This suggests that higher-quality, or more desirable, views may correlate with higher sales prices. This highlights a potential premium on properties with better views, which is a valuable insight for real estate pricing strategies.

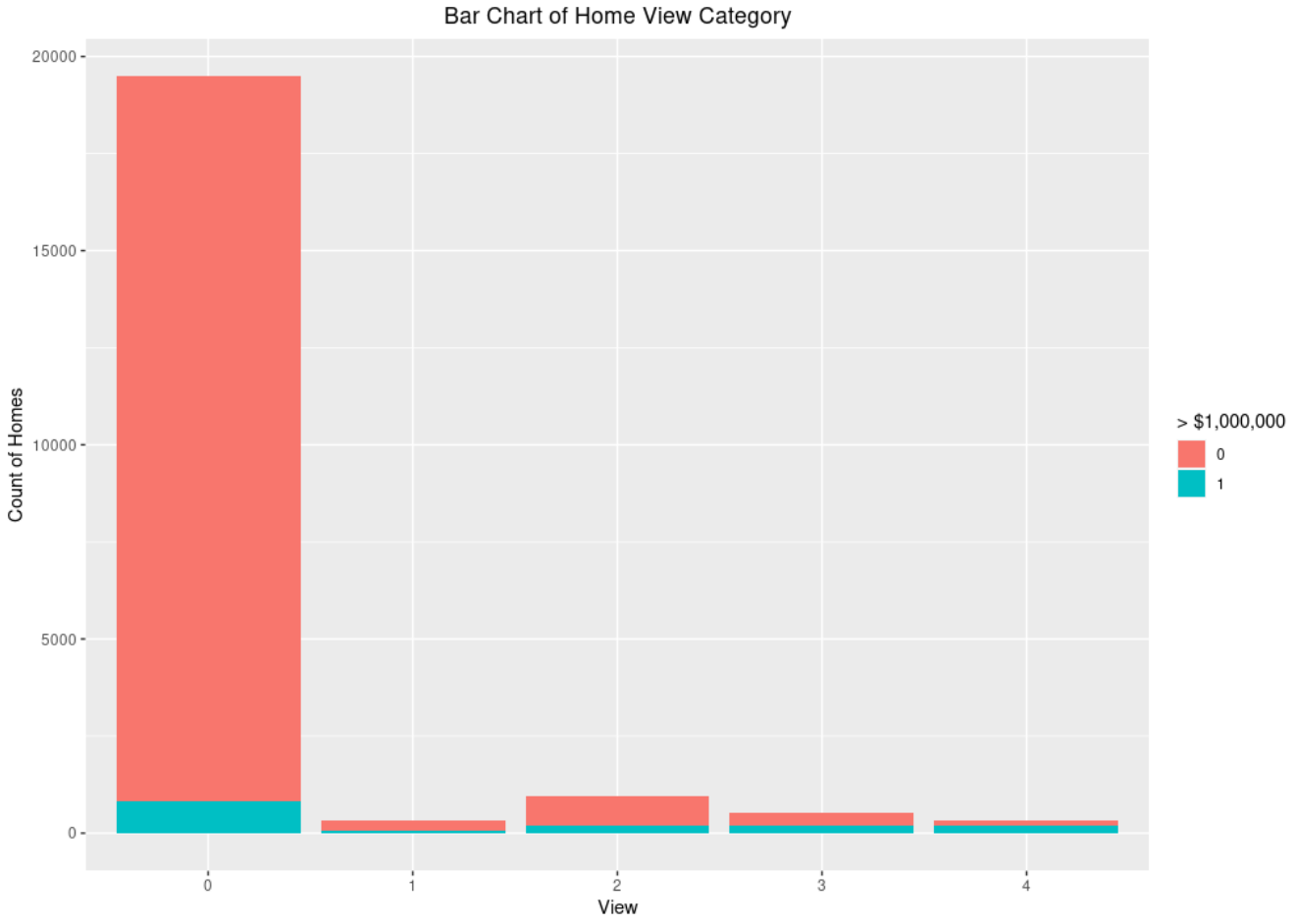

Figure 7

This visualization illustrates the distribution of home prices relative to their waterfront status, with properties that are on the waterfront being designated as 1. Generally, non-waterfront homes have a broader distribution of prices that are typically lower, with a more pronounced peat at the lower price range. In contrast, waterfront homes have a much narrower distribution, indicating less variability in price, but significantly higher overall prices with a sharp peak at higher values. This clearly demonstrates the premium attached to waterfront properties, which command higher prices.

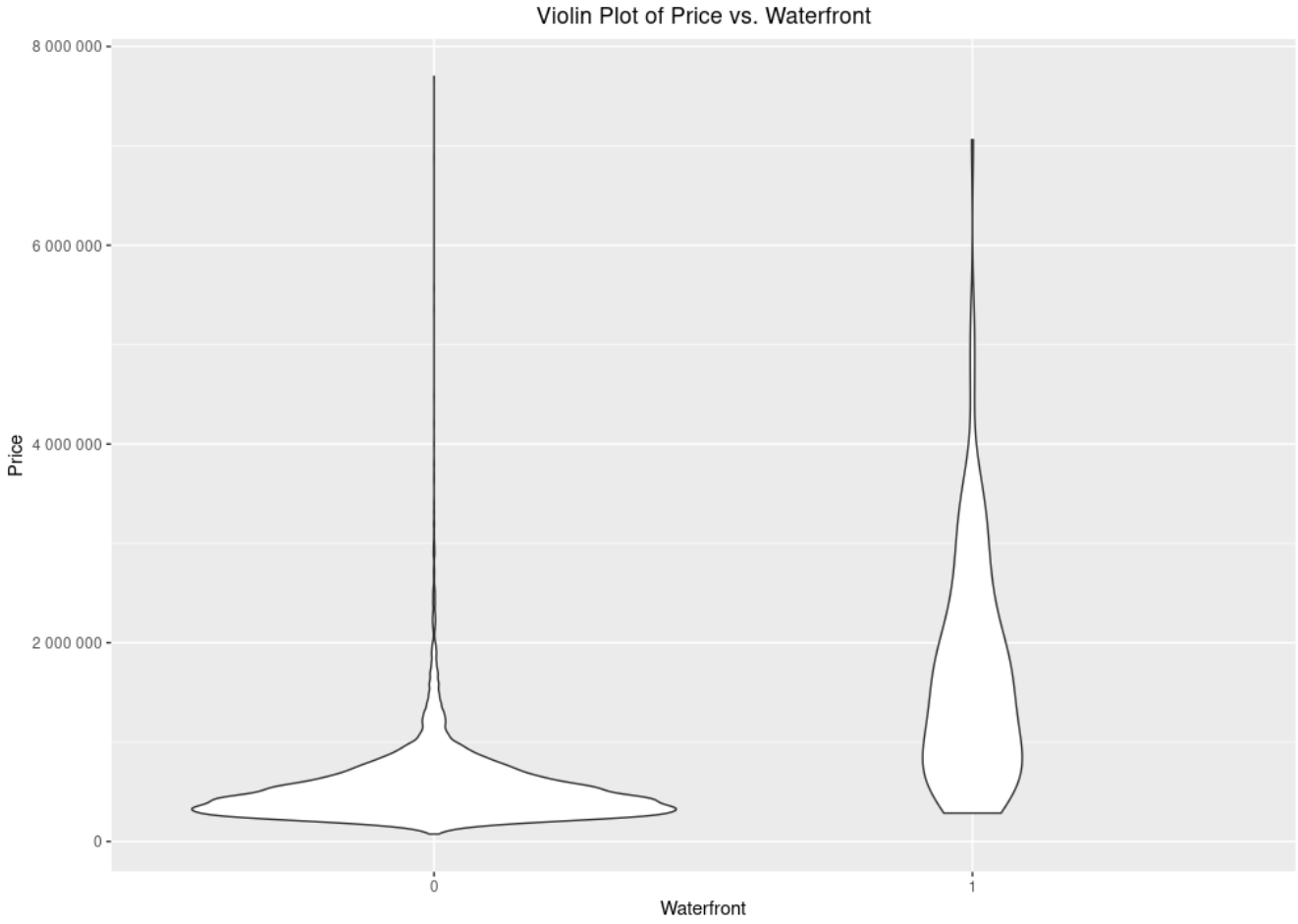

Figure 8

Building from the scatterplot of sqft_living vs price in question 1, we reproduced the plot with color and size representations of waterfront and view. Notably, properties on the waterfront, as indicated by the larger dots, command higher prices across similar living areas compared to non-waterfront properties. This premium for waterfront homes is visible across all view categories. Furthermore, homes with higher view ratings, particularly those categorized as 3 and 4, are positioned towards the higher end of the price range, regardless of their living space. This suggests that superior views enhance property value significantly, a trend that is even more pronounced for waterfront properties. The concentration of lower-priced homes largely falls within the view category of 0, indicating that a lack of a significant view leads to lower pricing, despite the living space size. This plot effectively highlights the additional impact of having an excellent view on the waterfront.

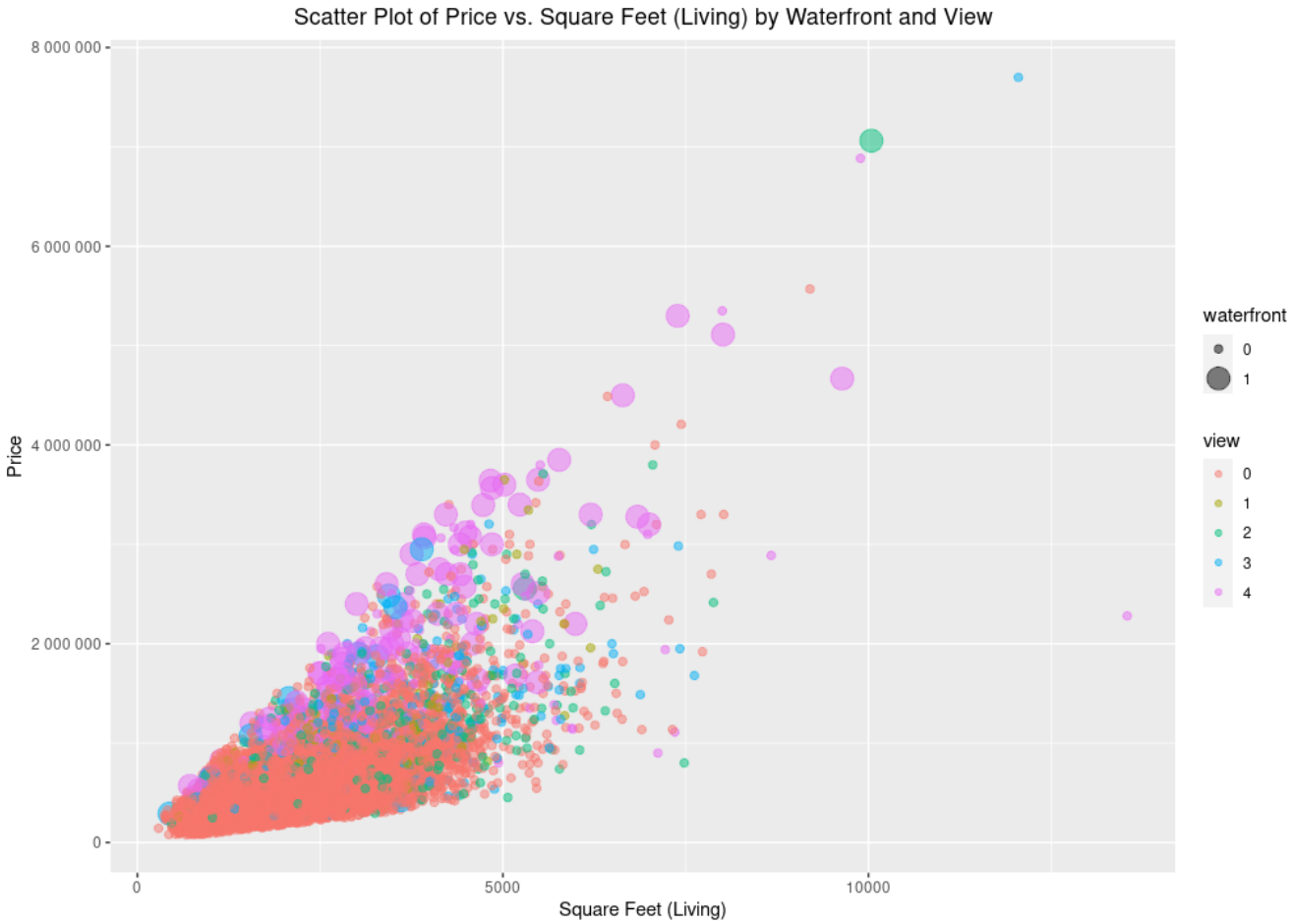

Figure 9

Section 7: Question 2 Model

```
Call:
glm(formula = above_million ~ waterfront + view + sqft_living,
    family = binomial, data = train)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.3712139  0.2068990 -40.460  < 2e-16 ***
waterfront1  1.5419769  0.3683452   4.186 2.84e-05 ***
view1        1.2799197  0.2451613   5.221 1.78e-07 ***
view2        1.1197982  0.1619749   6.913 4.73e-12 ***
view3        1.6391407  0.1818632   9.013  < 2e-16 ***
view4        2.3179823  0.2523292   9.186  < 2e-16 ***
sqft_living  0.0019499  0.0000621  31.402  < 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: 5364.9  on 10805  degrees of freedom
Residual deviance: 2888.3  on 10799  degrees of freedom
AIC: 2902.3

Number of Fisher Scoring iterations: 7
```

Our Logistic Regression Model output, which includes "view," "waterfront," and "sqft_living" as predictors, shows that all included variables are statistically significant predictors of whether a house sells for more than $1 million. With the intercept being approximately -8.37, our model indicates that the baseline category (e.g., houses not on the waterfront with the lowest view rating and a smaller sqft_living) has a low probability of exceeding the $1 million sales price threshold.

The Waterfront (waterfront1) coefficient of approximately 1.54 suggests that houses on the waterfront are more likely to sell for more than $1 million, holding all other variables constant. The odds ratio can be calculated as exp(1.54), which is 4.67. This indicates that the odds of selling above the designated sales price threshold are about 4.67 times higher for waterfront properties than for non-waterfront properties.

The View coefficients for the different view categories (view1 to view4) generally increase as the view quality improves, suggesting a strong positive relationship between the quality of the view and the likelihood of a house selling for more than $1 million. For instance, view4 with a coefficient of approximately 2.32 shows a strong positive impact. The odds ratio for view4 can be calculated as exp(2.32), which is 10.19. This is significantly higher than the odds compared to houses with the lowest view quality.

The square footage of the apartment interior living space (sqft_living) coefficient of 0.0019499 indicates that each additional square foot of living space increases the odds of selling a house above $1 million by a factor of 1.002, which seems insignificant but can accumulate to a substantial effect over hundreds of square feet.

This model appears robust and meaningful, especially for practical real estate evaluations where waterfront properties and panoramic views are highly valued. Waterfront locations and higher view ratings substantially increase the probability of a house selling for more than $1 million. The square footage of the living area also positively impacts the selling price, however the effect per square foot is smaller when compared to the categorical attributes included in the model.

ROC Curve

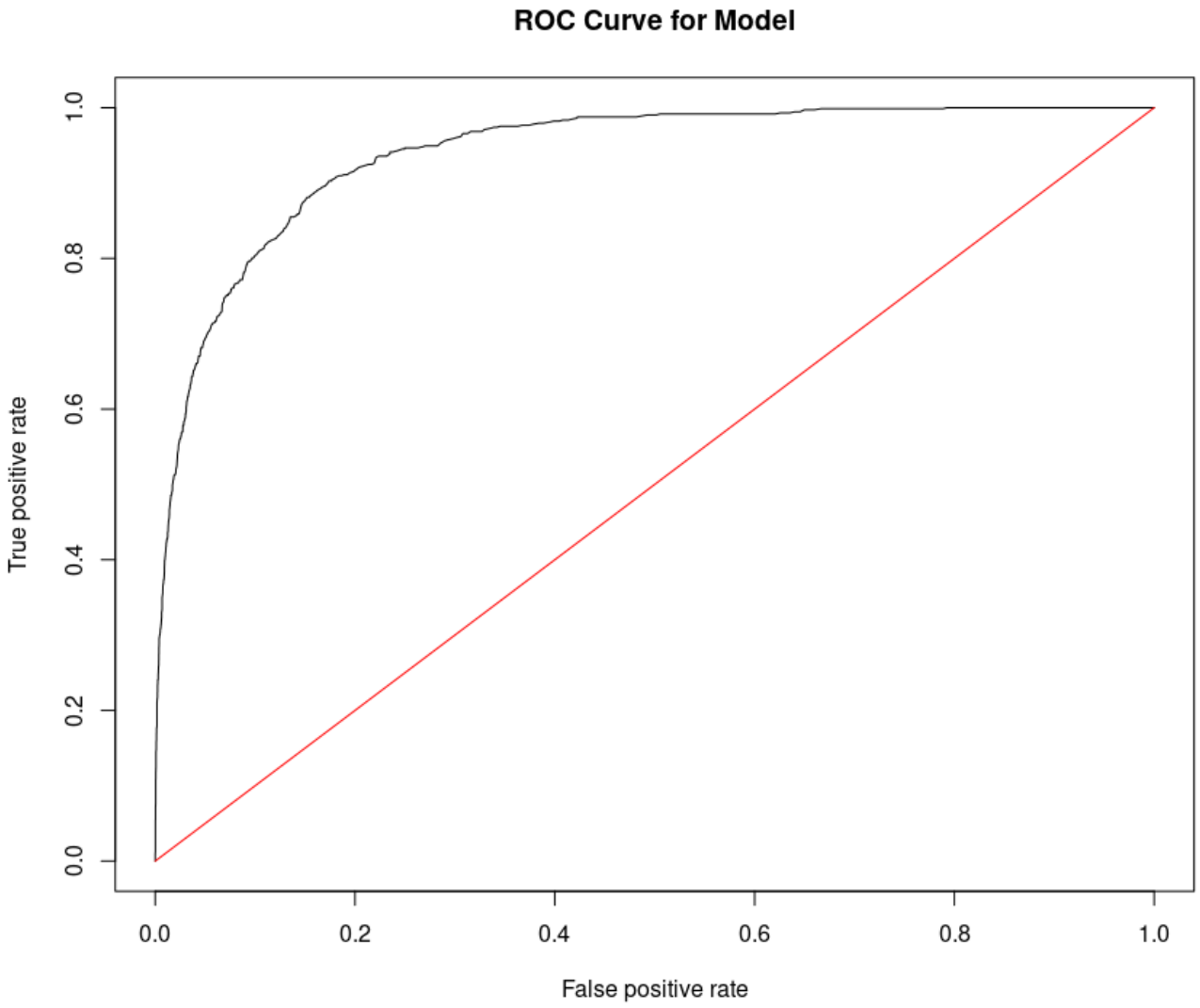

Figure 10

The Receiver Operating Characteristic (ROC) Curve is a graphical representation used to assess the performance of binary classification models at various threshold settings. The curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold levels. Our ROC Curve shows a very good model performance as it significantly bows towards the top left corner. This suggests a high TPR and a low FPR across many threshold settings. Essentially, our Logistic Regression Model performs well in distinguishing between houses that sell for more than $1 million and those that do not and is better than random guessing.

AUC

A perfect classifier will have an AUC of 1. A classifier that randomly guesses will have an AUC of 0.5. Thus, AUCs closer to 1 are desirable. The AUC of our model is 0.9393754, which means our Logistic Regression Model does better than random guessing.

Confusion Matrix

```
   FALSE TRUE
0  9964  112
1   422  309
```

According to the Confusion Matrix: 9964 observations were correctly predicted as not selling above $1 million (True Negative), 112 observations were incorrectly predicted as selling above $1 million (False Positive), 422 observations that sold above $1 million were incorrectly predicted as not doing so (False Negative), and 309 observations were correctly predicted as selling above $1 million (True Positive).

```
$Accuracy
[1] 0.9505876

$Sensitivity
[1] 0.4227086

$Specificity
[1] 0.9888845
```

According to the accuracy, our model correctly predicts the outcome approximately 95.06% of the time. The sensitivity of our model states that we identify approximately 42.27% of the houses that sell for more than $1 million correctly. The specificity of our model states the proportion of actual negatives that were correctly identified, meaning the model correctly predicts approximately 98.89% of the houses that do not sell for more than $1 million.