Final Project Report

Group 8

Authors: Sam Knisely (sck4jh), Darreion Bailey (rzu5vw), Dae Hwang (qsh9fk), Michael Amadi (mxg9xv)

Section 1

Summary of Findings

When setting the list price for a home, we often assume it comes down to the neighborhood and the perceived status of its residents. Our study reveals that apart from obvious indicators like the types of cars in the driveways and the social standing of the neighbors, many other critical factors significantly impact how a property is valued. In our report, we aimed to uncover how different features of a house affect its price, and to explore what aspects, aside from the house's size, determine whether a house is priced low or high per square foot.

We found that there are not just one or two factors that wholistically can be used to determine the price of a house, and many factors can come into play. For example, we found that the number of bedrooms, number of bathrooms, number of floors, the square living square feet of the house, the total square feet of the lot, the various grades and conditions of the house, whether or not the house is newly constructed, whether or not the house is waterfront, and whether the house is an urban or rural area all have a role to play in determining the price of a house. As can be seen, it seems like it can be a complicated process to determine the price of a house. We set out to demystify and simplify this process by crafting a model that is useful in predicting house prices.

On the other hand, we found that categorizing houses as priced high or low per square foot is a bit more nuanced. By this, we mean that many of the factors listed above are heavily related to each other when classifying houses as priced high or low per square foot. In fact, in the end we found that just knowing the condition of a house and whether it is in an urban or rural neighborhood can be useful to predict if the house will be priced high or low per square foot. However, we think that additional location information may be useful to determine whether houses are priced high or low per square foot.

Section 2

Description of pre-existing and custom variables

The Project 2 dataset explores home sales between May 2014 and May 2015 in King County, Washington, USA (including Seattle).

Quantitative Variables

price – price of each home sold

bedrooms – number of bedrooms in the sold home

bathrooms – number of bathrooms in the sold home; half bathrooms are listed as 0.5, accounting for a room with a toilet but lacking a shower setup

sqft_living – total square footage of the interior living space (not including basement or garage)

sqft_lot – total square footage of the land space, excluding the home.

floors – number of floors within the home (e.g., two story home).

sqft_above – represents the square footage of the interior housing space that is above ground level

sqft_basement – represents the square footage of the interior housing space that is below ground level

yr_built – represents the year the home was fully constructed on the lot of land

yr_renovated – represents the year of the properties last recorded renovation

sqft_living15 – represents the approximate square footage of interior living space for the homes of the nearest 15 neighbors

sqft_lot15 – represents the square footage of the land lots of the nearest 15 neighbors

view – an index from 0 to 4 of quality of the view from a property, where 0 indicates no view or an unremarkable view, and 4 represents an excellent or highly desirable view

condition – the overall condition of a property, where 1 indicates a property in poor condition and 5 signifies a property in excellent condition, requiring no repairs or renovations.

grade – an index ranging from 1 to 13 and is used to evaluate the construction and design quality of the property. Scores from 1 to 3 indicate subpar construction and design, a score of 7 represents an average level, and scores from 11 to 13 denote a high-quality level of construction and design.

lat – represents how far north or south the property is located from the equator

long – represents how far west or east the property is located from the equator

Qualitative Variables

id – ten-digit King's County parcel ID number

date – date of the home sale in public tax and accessors records

waterfront – a dummy variable indicating whether the apartment has a waterfront view or not.

zipcode – postal code derived from the United States Postal Service defining the particular region, city, or town the house is geographically located

Custom Continuous Variable

Price per square foot ($/ sq. ft)

We created a continuous variable price per square foot ($/ sq. ft) to aid in our quest to achieve a binary response in Question 2. Additionally, $/sq. ft is further constrained to the high/expensive (75% percentile = $318.49 per sq. ft) and low/cheap (25% percentile = $182.48 per sqft) parameters.

Custom Categorical Variables

Zipcode_category

The original zipcode data includes postal codes from the United States Postal Service, which specify the region, city, or town where a house is geographically located. The new variable, "zipcode_category," was transformed into a non-numerical categorical variable representing urban or rural neighborhoods. This categorization was achieved by referencing visual data from Google Maps. Google Maps provides detailed visual cues such as the density of road networks, housing clusters, and the presence of commercial or industrial areas, which help in distinguishing urban neighborhoods, characterized by higher density and infrastructure development, from rural neighborhoods, which typically show more open spaces and fewer infrastructural developments. A value of 1 represents an urban zip code and a value of 0 represents a rural zip code.

Construction_category

Secondly, we created a new variable called "construction" by categorizing the "yr_renovated" and "yr_built" variables. We set the parameters for the combined variables to assign values of 0 and 1 based on the existing data. Properties with years less than 2000 were assigned a value of 0, indicating they were outdated, while those built or renovated in 2000 or later were assigned a value of 1, representing modernized properties.

Data Preprocessing

For our analysis of King's County housing data, we have omitted longitude and latitude data because we do not want to include spatial directions as predictors in our models.

After loading the dataset in RStudio, we conducted a visual inspection and noticed that `sqft_living` variable appeared to be the exact sum of the `sqft_above` and `sqft_below` variables. Through our data discovery process, we determined that removing the `sqft_above` and `sqft_below` variables and utilizing `sqft_living` would be appropriate to minimize the number of predictor variables related to livable space within a King's County property.

Section 3

Our group is pursuing two primary research questions focused on the housing market in the greater Kings County (KC), motivated by the potential to enhance understanding and decision-making in real estate.

Question 1: How do the various features of a house influence its price?

Motivation: Exploring how different features of a house affect its market price is crucial for several economic reasons. Firstly, it assists in developing predictive models to forecast housing prices beyond King County, which is invaluable for buyers, sellers, and investors to make informed decisions. Additionally, analyzing these influences can validate or refine existing theories about what factors drive house prices, offering deeper insights into market dynamics. Lastly, understanding the impact of house features on price contributes to broader discussions on housing affordability and socioeconomic factors in urban planning or affordable housing programs, such as those managed by HUD (U.S. Department of Housing and Urban Development).

Question 2: What influence do house qualities, aside from the size of the house, have on whether a house is categorized as high price or low price per living square foot?

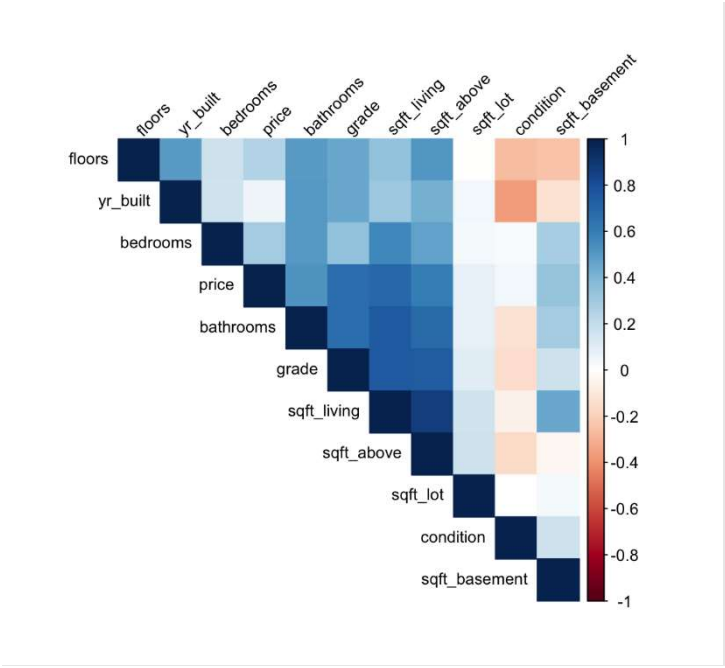
Motivation: The structure of this question delves deeper into the dynamics of the King's County housing market by examining the price per square foot as a standardized measure of value, which facilitates comparisons beyond simple square footage calculations (e.g., comparing a 50 sq. ft area to a 10-by-10 area). This approach helps identify underlying trends that influence property value based on factors other than size, such as location, age, and design quality. By categorizing the price per square foot into high/expensive (above $318.49 per sqft) and low/cheap (below $182.48 per sqft), we obtain a nuanced view of the housing market. This segmentation is especially useful in logistic regression models, where price classification can demonstrate how

non-size-related attributes, like proximity to water, affect perceived value, thus influencing buyer confidence, urban planning, investment strategies, and policy-making related to housing markets.

Together, these inquiries create a comprehensive approach to understanding real estate valuation in King County, aiming to inform the decision-making processes of real estate agents, buyers, and financial institutions by offering deeper insights into the factors that determine whether a house's value is considered high or low.

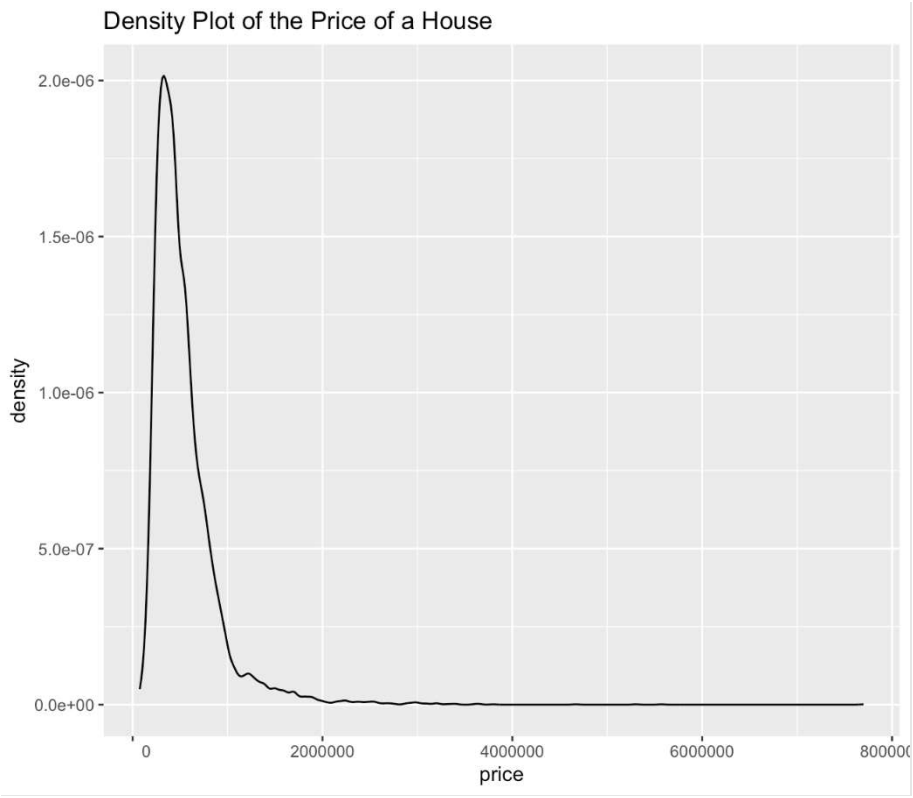
Section 4

Figure 1: Correlation Matrix



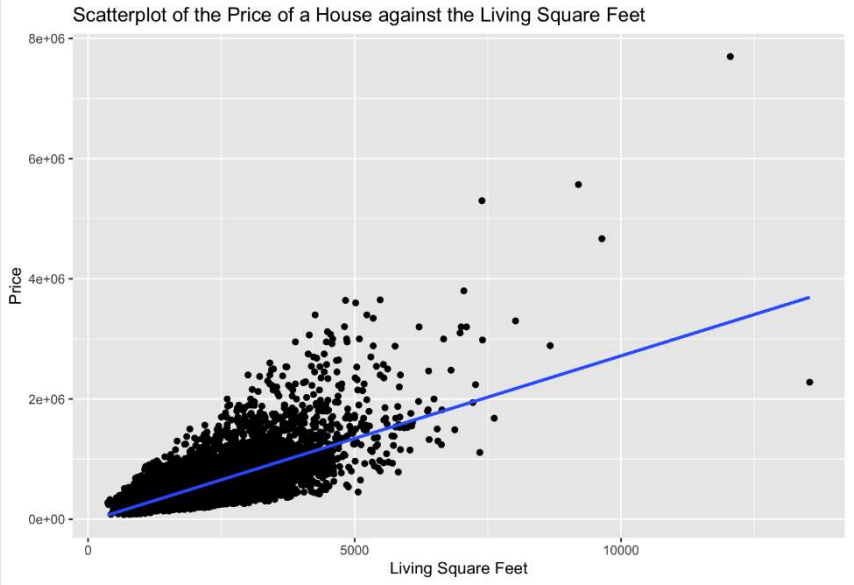
The above plot shows the correlations between many of the variables in the dataset. This plot helps give us an initial idea of which variables might be linearly associated. We observe that price, our response variable for the multiple linear regression, has high correlations with some variables such as the amount of bathrooms, the grade of the home, and the living square feet of the home. This indicates that these variables have a strong linear association with the price of the house and may be useful to include in the multiple linear regression. We will explore these relationships further in our model selection.

Figure 2: Density Plot of Price



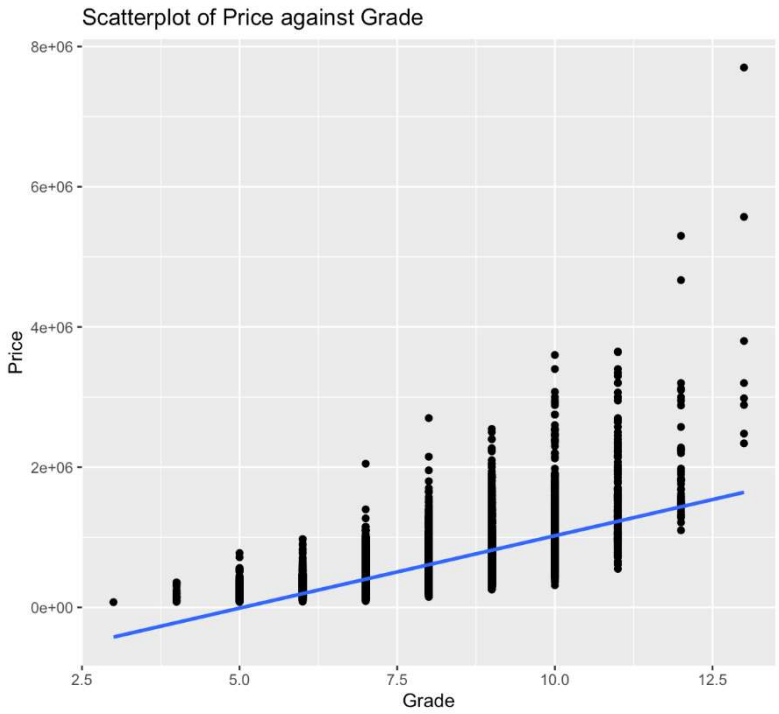
The density plot above helps give an understanding of the values of price within the dataset. We observe that most houses in the dataset are priced under $1 million. The highest concentration of house prices looks to be less than $500,000, around $300,000 or so. This helps give us some context of our response variable, price, as we move forward with our multiple linear regression.

Figure 3: Scatterplot of Price against Living Square Feet



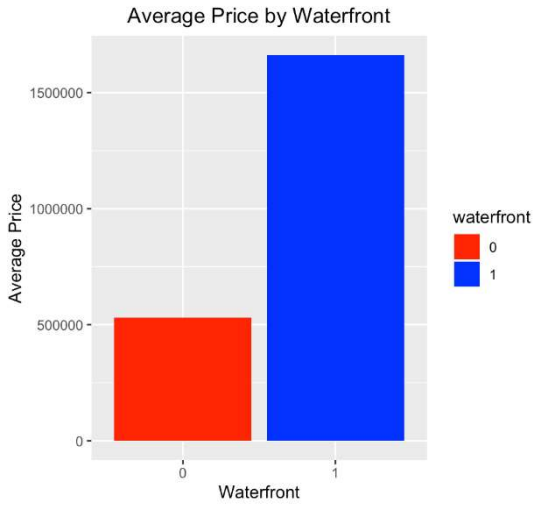
Based on the above correlation matrix, we consider exploring the relationship between price and one of the potential predictors, living square feet of the house. We observe from the scatterplot that there looks to be a linear trend between price and living square feet. We note this as we move towards model selection.

Figure 4: Scatterplot of Price against Grade



We also observe from the correlation matrix that price and grade have a high correlation. Therefore, the above plot explores the relationship between the price of a house and the grade of the property's construction and design. As can be seen, the house prices steadily increase as grade increases. This shows a clear trend and is noted as we move towards model selection.

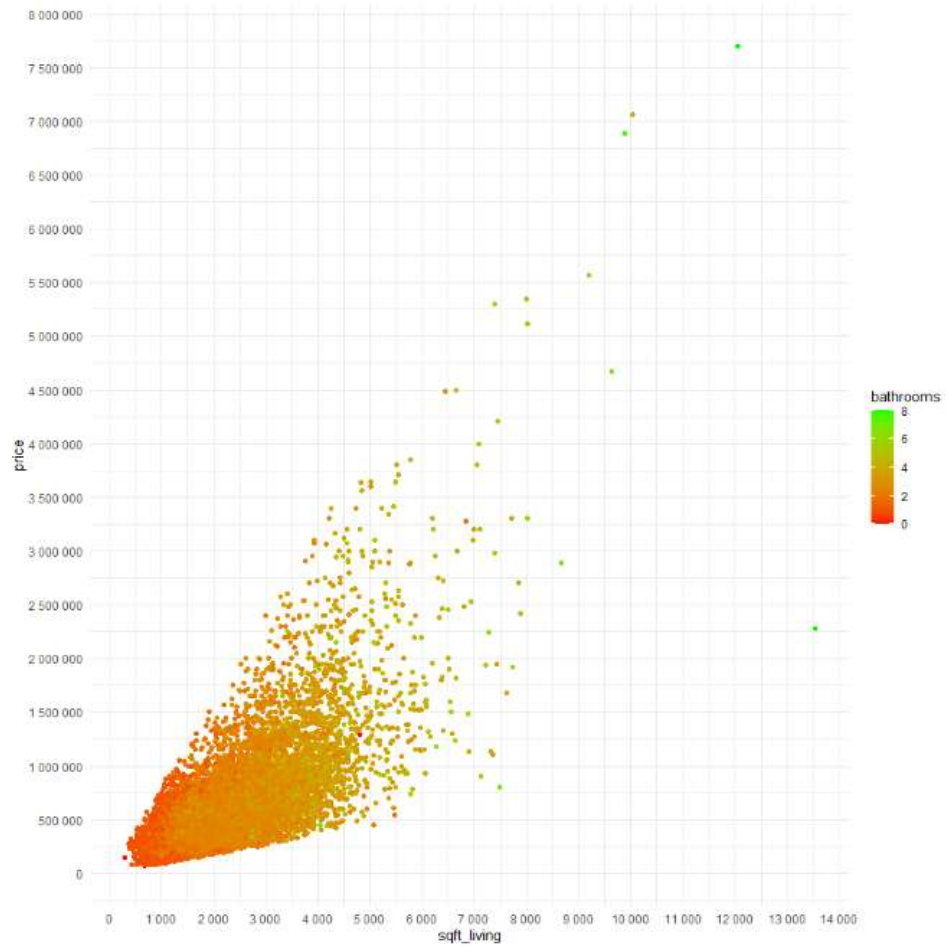
Figure 5: Bar Chart of the Average Price by Waterfront



We also see from the above chart that the price of houses indicated being waterfront is on average about three times higher than houses not indicated being waterfront. We may consider including this variable as a factor in our model and would expect there to be a positive regression

coefficient for the waterfront equal to 1 variable. We will keep this in mind as we create our multiple linear regression model.

Figure 6: Scatterplot (Sqft_living vs. Price by Bathrooms)



Additionally, we see from the correlation matrix that the price of a house and the number of bathrooms have a high correlation. Building off Figure 3, we add the number of bathrooms as the color of points to explore additional relationships. On the left side of the graph with lower prices and living square feet, there are lower amounts of bathrooms. As we move up and right along the graph, we observe that the number of bathrooms increases as price and living square feet also increase. This helps give more context on our data as we consider which variables to include in model selection.

Section 5

We now look to fit a multiple linear regression model for our first question of interest, how the features of a house influence its prices. We first start with some data cleaning procedures, as we do not want to include all the variables in the dataset as predictors for the price of a house.

We start by removing the ID variable from our set of predictors, because this variable is simply an identifier of the houses and is not fit to be a predictor. Next, we also drop sqft_above and sqft_below from our set of predictors because these variables sum up to the sqft_living variable and would introduce redundant information to the model if included. We also drop the date variable because we are not looking at our analysis over time. Furthermore, we drop the latitude and longitude variables because we do not want to include spatial directions in the model as it does not make much sense in the context of our question. Additionally, we drop year renovated and year built and instead use the new construction variable described in Section 2 above. Lastly, we drop zipcode because we instead use the categorical variable we created called zipcode_category that classifies zipcodes as either urban or rural.

Now that we have our predictors limited to what we logically think fits our question of interest, we use some automatic search procedures to get an initial model. We run forward selection, backward elimination, and stepwise regression. Our forward selection and stepwise regression procedures yield the same models and AICs of 297051.9, which was lower than the backwards elimination's AIC but still a large value. Nonetheless, we decide to move forward with this initial model from the stepwise regression and see if we can improve it. The summary output of this initial model is provided below.

Figure 7: Initial Model Summary Output

```
Call:
lm(formula = price ~ sqft_living + grade + view + zipcode_category +
    waterfront + condition + bedrooms + sqft_living15 + sqft_lot15,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1128260  -121779   -10857    93705  4646129

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -7.652e+05  2.368e+04 -32.308  < 2e-16 ***
sqft_living       1.865e+02  4.619e+00  40.370  < 2e-16 ***
grade             9.299e+04  3.045e+03  30.539  < 2e-16 ***
view              5.271e+04  3.192e+03  16.513  < 2e-16 ***
zipcode_category1 9.973e+04  5.084e+03  19.616  < 2e-16 ***
waterfront1       5.325e+05  2.705e+04  19.685  < 2e-16 ***
condition         5.624e+04  3.417e+03  16.460  < 2e-16 ***
bedrooms         -3.564e+04  2.842e+03 -12.540  < 2e-16 ***
sqft_living15     2.483e+01  5.225e+00   4.751 2.05e-06 ***
sqft_lot15       -3.808e-01  8.079e-02  -4.714 2.46e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 225300 on 10796 degrees of freedom
Multiple R-squared:  0.6167,    Adjusted R-squared:  0.6164
F-statistic:  1930 on 9 and 10796 DF,  p-value: < 2.2e-16
```

Figure 8: Initial Model Residual Plots

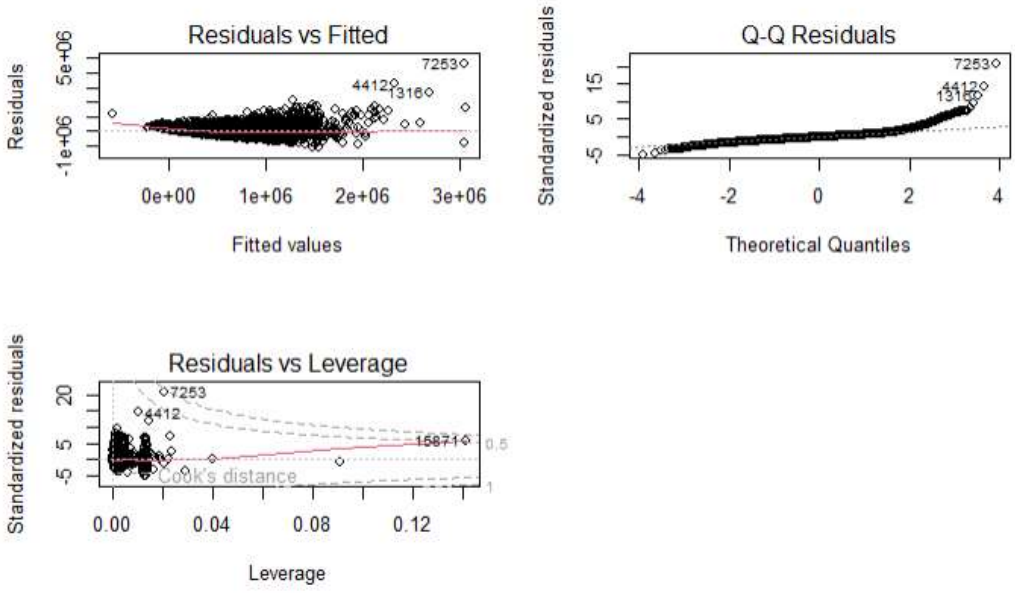
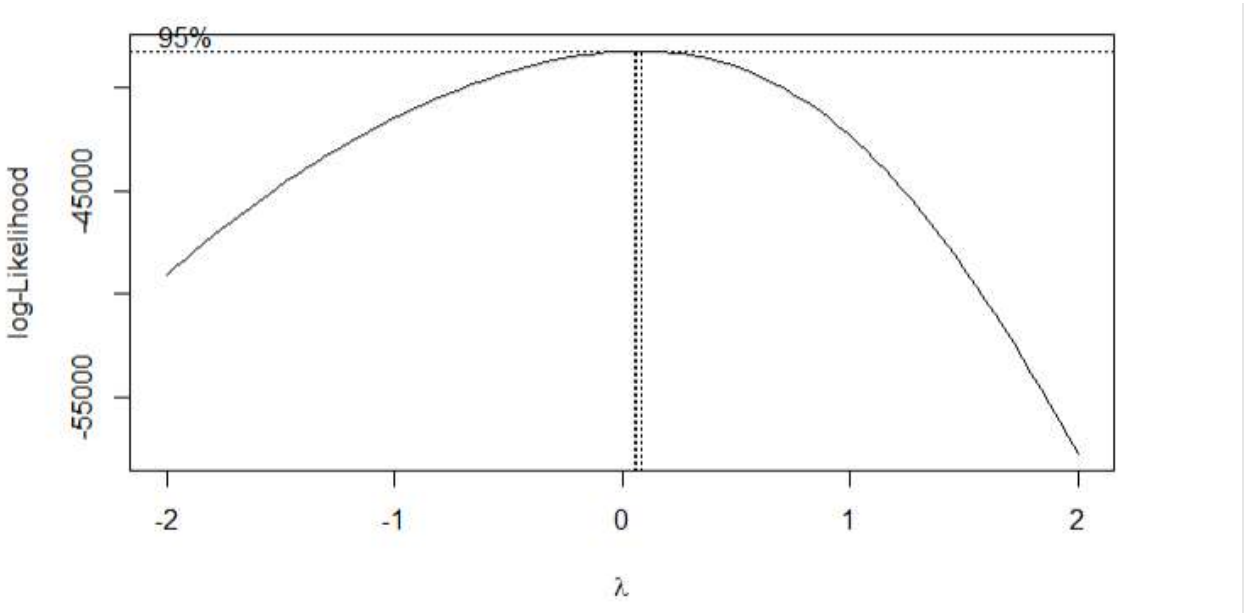


Figure 9: Box-Cox Plot



The `Residuals vs Fitted` plot above (Figure 8) shows increasing variance in the residuals in our initial model. This indicates a failure to meet regression assumption 2, that the errors have a constant variance. A failure of regression assumption 2 indicates that a transformation of the response variable may be needed. Given the potential non-linearity, and the box-cox plot above showing a lambda very near to zero, we try to log transform the price variable to see if that helps the fit.

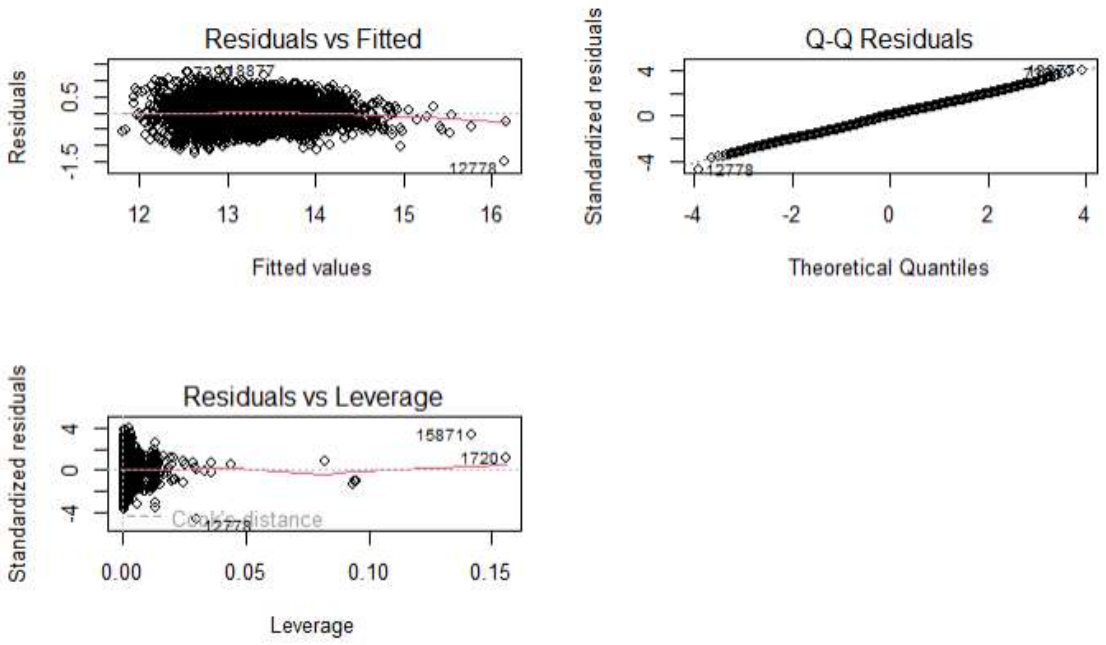
After applying a log transformation to the response variable price, we rerun the stepwise automatic search procedure to fit a new model. See the summary output below. This new model also reports a much lower AIC of 6823.531 as compared to the original model's AIC of 297051.9.

Figure 10: Log Transformed Summary Output

```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.078e+01  3.514e-02 306.596  < 2e-16 ***
grade                   1.670e-01  4.705e-03  35.496  < 2e-16 ***
sqft_living             1.793e-04  6.836e-06  26.236  < 2e-16 ***
view                    6.511e-02  4.716e-03  13.807  < 2e-16 ***
condition               1.020e-01  5.307e-03  19.225  < 2e-16 ***
zipcode_category1       1.266e-01  7.565e-03  16.732  < 2e-16 ***
sqft_living15           8.840e-05  7.744e-06  11.416  < 2e-16 ***
waterfront1             3.218e-01  3.982e-02   8.080 7.18e-16 ***
floors                  3.621e-02  7.542e-03   4.801 1.60e-06 ***
bedrooms               -2.093e-02  4.190e-03  -4.996 5.94e-07 ***
construction_category1  2.417e-02  9.316e-03   2.594  0.00949 **
sqft_lot15             -5.193e-07  1.635e-07  -3.177  0.00149 **
sqft_lot                3.186e-07  1.015e-07   3.139  0.00170 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3316 on 10793 degrees of freedom
Multiple R-squared:  0.6031,    Adjusted R-squared:  0.6027
F-statistic:  1367 on 12 and 10793 DF,  p-value: < 2.2e-16
```

Next, we check the regression assumptions again.

Figure 11: Log Transformed Model Residual Plots

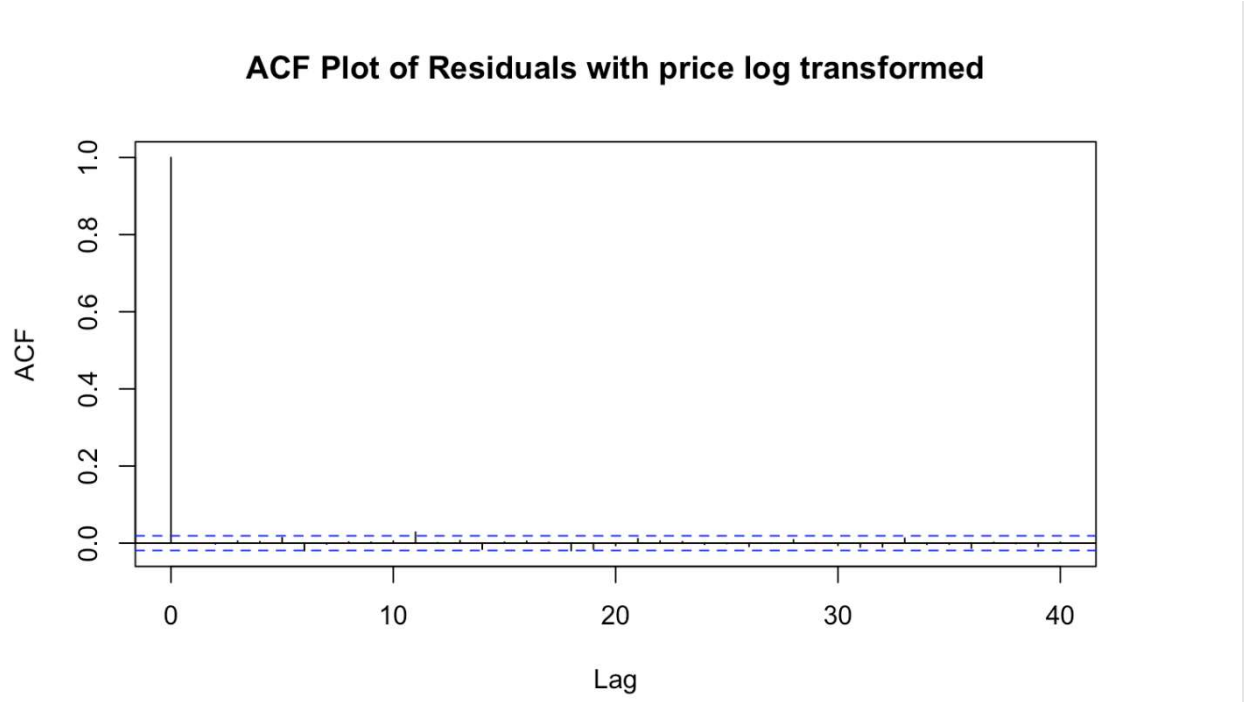


We observe that now the regression assumptions generally look to be met. Additionally, the new model reports Regression assumption 1, that the errors have a mean of zero, now looks to be met

since the red line on the residual plot is not showing much curvature. Regression assumption 2, that the errors have a constant variance, also looks to generally be met in the new model since the residual plot shows relatively constant vertical variance of the residuals when moving left to right across the plot.

Additionally, regression assumption 3, that the errors are independent from each other, also looks to be satisfied in the model with log transformed price We observe this by looking at the below ACF plot and seeing that none of the ACFs beyond lag 0 are significant, and therefore we have no evidence that the observations are dependent on each other in the model.

Figure 12: ACF Plot



Lastly, we see that regression assumption 4 also looks to be met, which is that the errors are normally distributed. We observe this by referencing the Q-Q plot (included in Figure 11) and seeing that most of the observations fall on the 45-degree line.

Next, we check for high leverage observations, influential observations using DFFITS, and outliers. We identify 633 high leverage observations, 428 influential observations, and 28 outliers. We are hesitant to remove these observations from our training data, so we take a closer look to see if we notice anything significant. We do notice that observation with the ID 2402100895 is listed as having 33 bedrooms but only 1620 living square feet. We believe this

must be a data entry error and therefore remove this observation from our training data. Other than that, we leave the rest of the observations in our analysis.

After removing this observation we identified as a data entry error, we refit the log transformed price variable against our training data using a stepwise regression to see if the model changes. Our stepwise regression yields a model with the same predictors as before removing this observation, but the AIC is slightly lower now at 6812.525.

Next, we check the variance inflation factors (VIFs) for the predictors to check if multicollinearity is present in the model. These values are provided below. We see that none of the predictors in our model yield VIFs above 5, so we do not have any strong indication of multicollinearity being present in the model.

Figure 13: Variance Inflation Factors

grade	sqft_living	view	condition	zipcode_category1
3.027	4.066	1.352	1.165	1.157
sqft_living15	waterfront1	floors	bedrooms	construction_category1
2.747	1.203	1.639	1.631	1.550
sqft_lot15	sqft_lot			
2.104	2.038			

Next, we do one last check on the regression assumptions. As can be seen below, the plots look reasonable and the regression assumptions look to be met.

Figure 14: Final Model Residual Plots

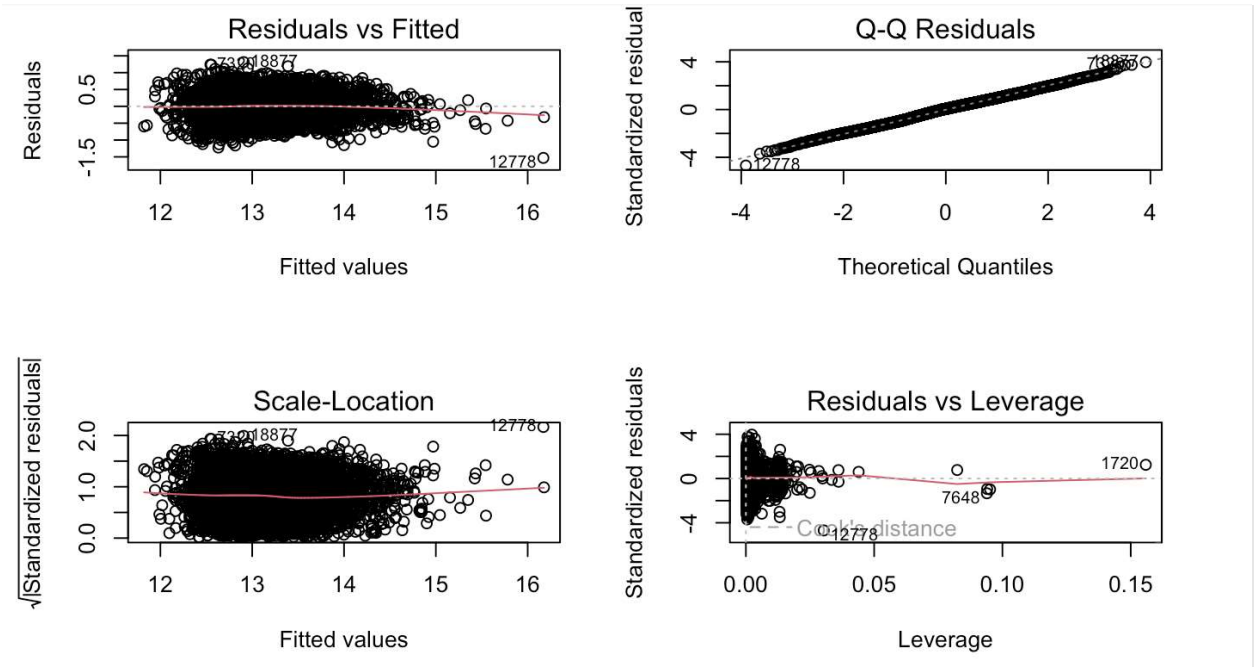
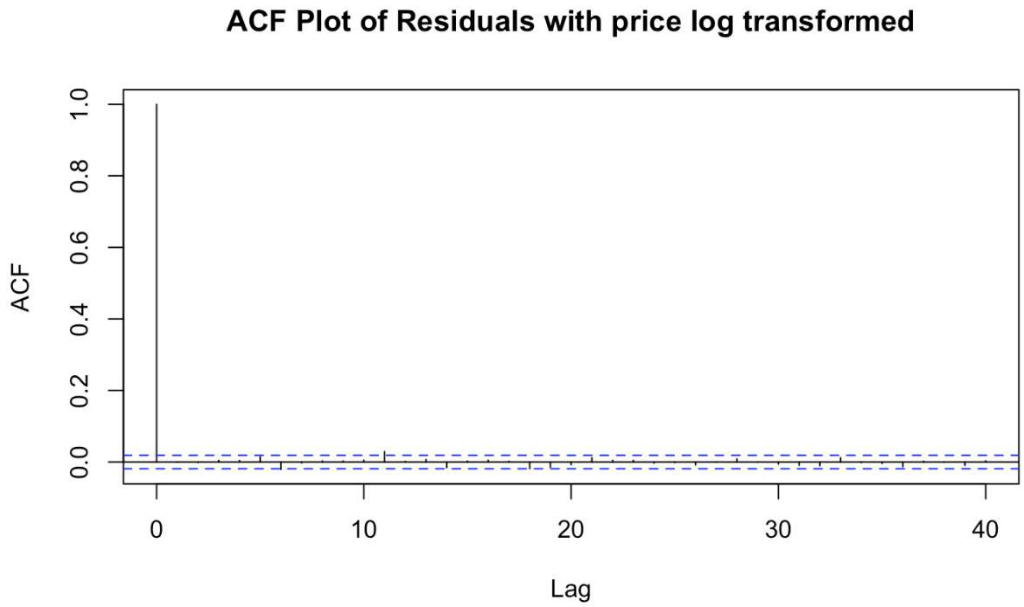


Figure 15: Final Model ACF Plot



Now, we look at the summary output and regression coefficients of our final model as can be seen in Figure 16 below.

Figure 16: Summary Output of the Final Model

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.079e+01  3.545e-02 304.377  < 2e-16 ***
grade                   1.662e-01  4.708e-03  35.300  < 2e-16 ***
sqft_living             1.836e-04  6.951e-06  26.422  < 2e-16 ***
view                    6.455e-02  4.717e-03  13.687  < 2e-16 ***
condition               1.020e-01  5.305e-03  19.219  < 2e-16 ***
zipcode_category1       1.267e-01  7.562e-03  16.756  < 2e-16 ***
sqft_living15           8.850e-05  7.740e-06  11.434  < 2e-16 ***
waterfront1             3.199e-01  3.981e-02   8.035 1.03e-15 ***
floors                  3.667e-02  7.540e-03   4.863 1.17e-06 ***
bedrooms               -2.666e-02  4.519e-03  -5.899 3.76e-09 ***
construction_category1  2.335e-02  9.315e-03   2.506  0.01222 *
sqft_lot15             -5.309e-07  1.634e-07  -3.249  0.00116 **
sqft_lot                3.159e-07  1.015e-07   3.114  0.00185 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3314 on 10792 degrees of freedom
Multiple R-squared:  0.6035,    Adjusted R-squared:  0.6031
F-statistic:  1369 on 12 and 10792 DF,  p-value: < 2.2e-16
```

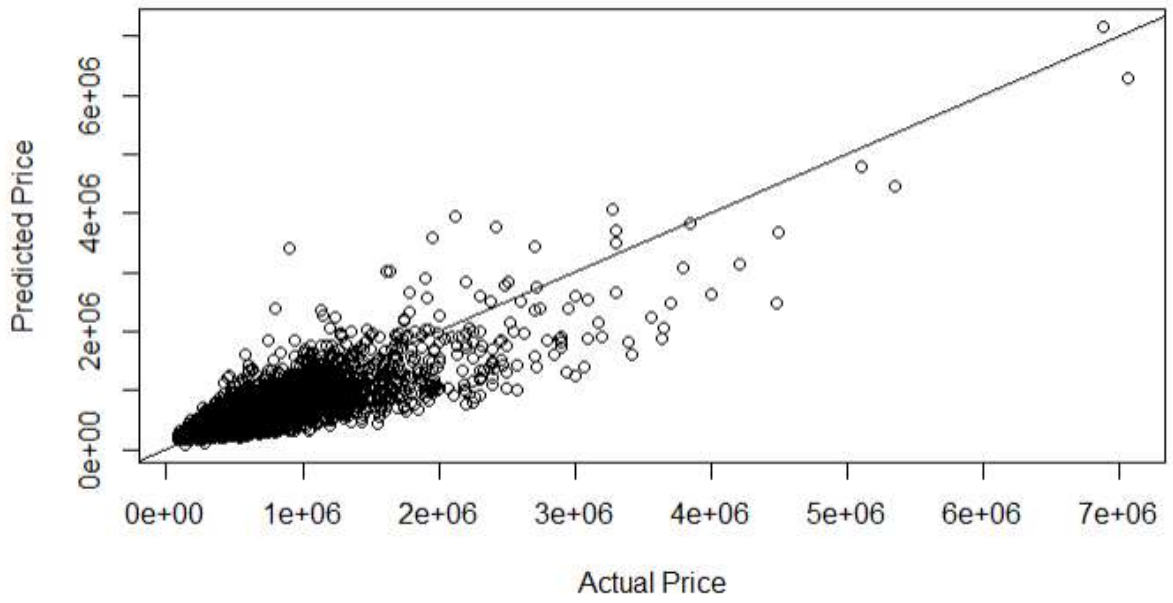
Our summary output for the final model yields the following regression equation:

y* = 10.79 + 0.1662 (grade) + .0001836 (sqft_living) + 0.06455 (view) + 0.1020 (condition) + 0.1267(zipcode_category1) + 0.00008850 (sqft_living15) + 0.3199 (waterfront1) + 0.03667 (floors) - 0.02666 (bedrooms) + 0.02335 (contruction_category1) - 0.0000005309 (sqft_lot15) + 0.0000003159 (sqft_lot)

where y* is the log transformed price.

The model output shows that the model is useful in predicting the log transformed price, since the p-value associated with the ANOVA F-test is far below 0.05. Additionally, the t-tests on each regression coefficient show highly significant p-values, so we do not have evidence to drop any of the predictors in the presence of the other predictors. Lastly, the adjusted R^2 reported in the output is 0.6031, which tells us that 60.31% of the variation in the log-transformed price can be explained by the predictors in the model.

Figure 17: Actual vs. Predicted Price – Final Model Test



From the output observed above, we notice that most of the data is below the $3,000,000 mark. Additionally, the MAE (Mean Absolute Error) and RMSE (Root Mean Square Deviation) are calculated as 138596.19010 and 209197.38999, respectively. Essentially, this means that the average price deviation is around $200,000. However, we notice that there is an extremely tight clustering around the $500,000 mark – which is the calculated mean for the entire dataset – that is close to the regression line, with tapering outward as the price increases. From this we can infer that the model is much better at predicting prices in the $0 - $1,000,000 range.

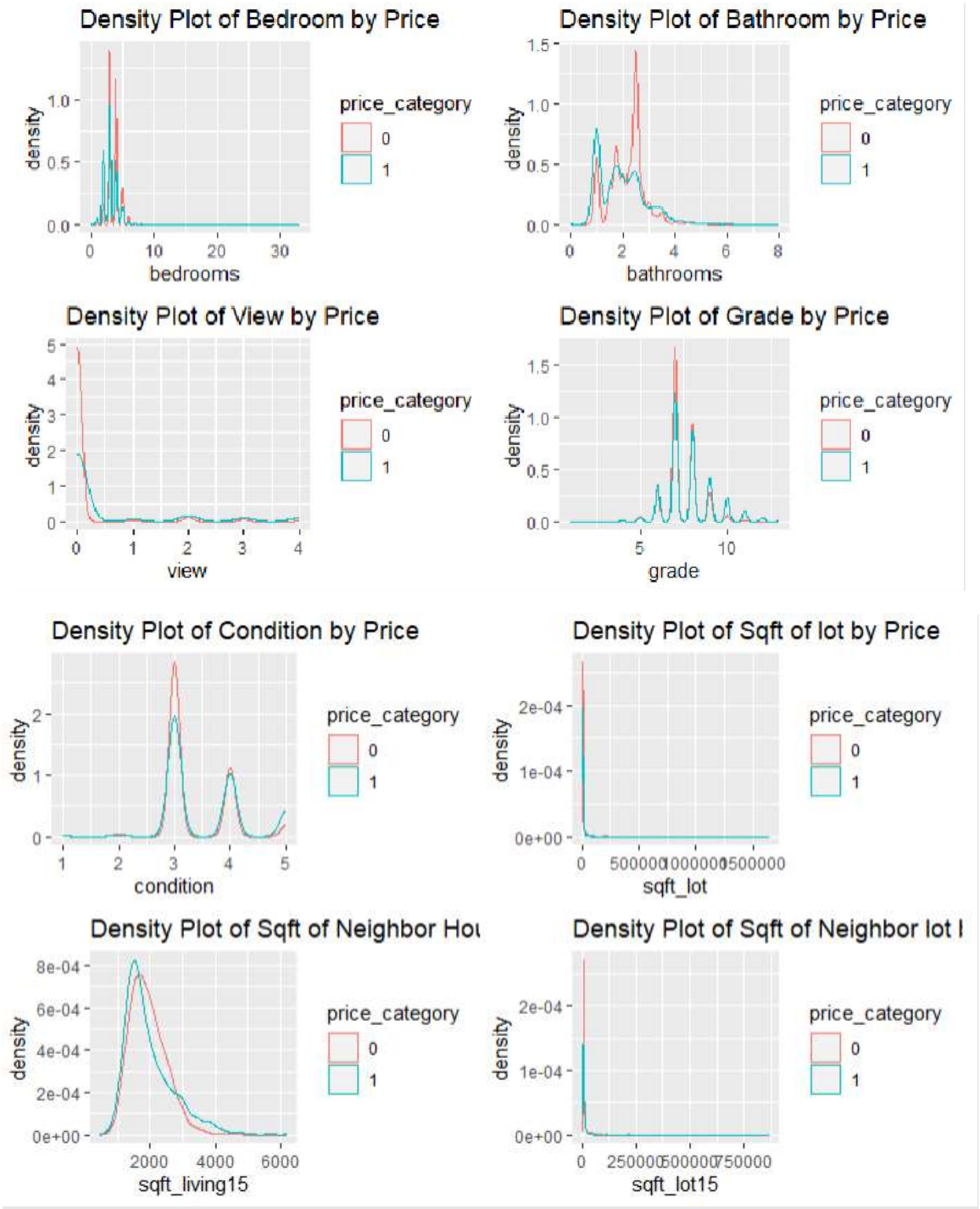
Given the scale of our range ($0 - $7,000,000), an RMSE of 209197.38999 represents about 3% of the maximum value. Therefore, we conclude given this consideration that a prediction error of 3% is reasonable for this model in the industry of real estate.

Section 6

Figure 18: Boxplot of Price per Living Square Foot

As mentioned above, our response variable for the logistic regression is high vs low price per living square foot. To determine our classifications of high or low prices per living square foot, we look at the five number summary of the price per living square foot variable, which is graphically presented in the boxplot above. 25% percentile and 75% percentile as cutoffs for our binary response variable are used. Therefore, we deem a house to be priced expensive or high per living square foot if the price per living square foot is greater than or equal to the 75% percentile ($318.32 per sqft) and we classify a house as cheap or low priced per living square foot if the price per living square foot is less than or equal to the 25% percentile ($182.29 per sqft).

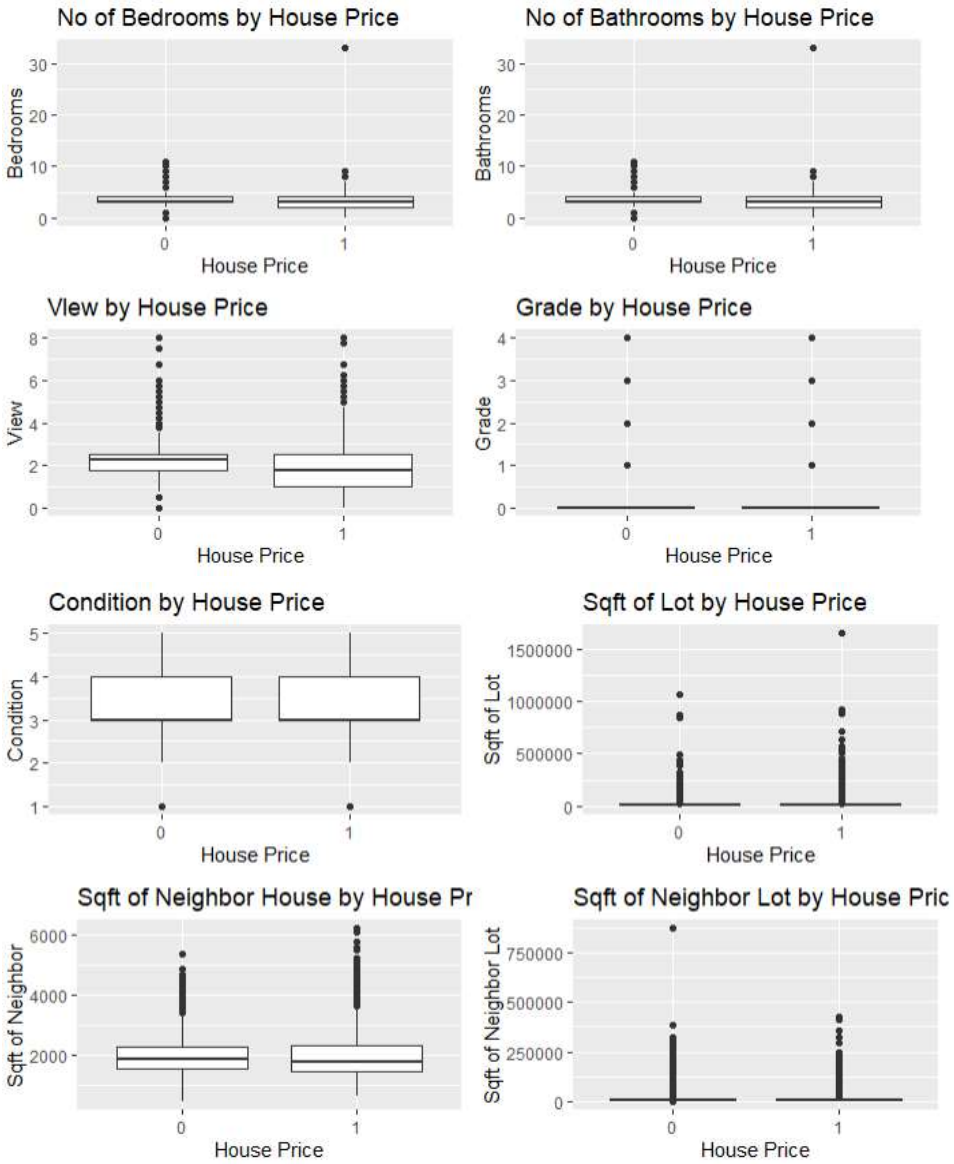
Figure 19: Density Plots of Price per Square Foot Categorizations against Number of Bedrooms, Bathrooms, View, Grade, Condition, sqft_lot, sqft_living15 and sqft_lot15.



The above density plots of bedrooms, bathrooms, view, grade and condition by price per square foot category (with 1 representing high prices per square foot) provides insight into whether the houses identified with high vs low prices per square foot have different proportions of the number of bathrooms and bedrooms. Overall, the proportions of the number of bathrooms and

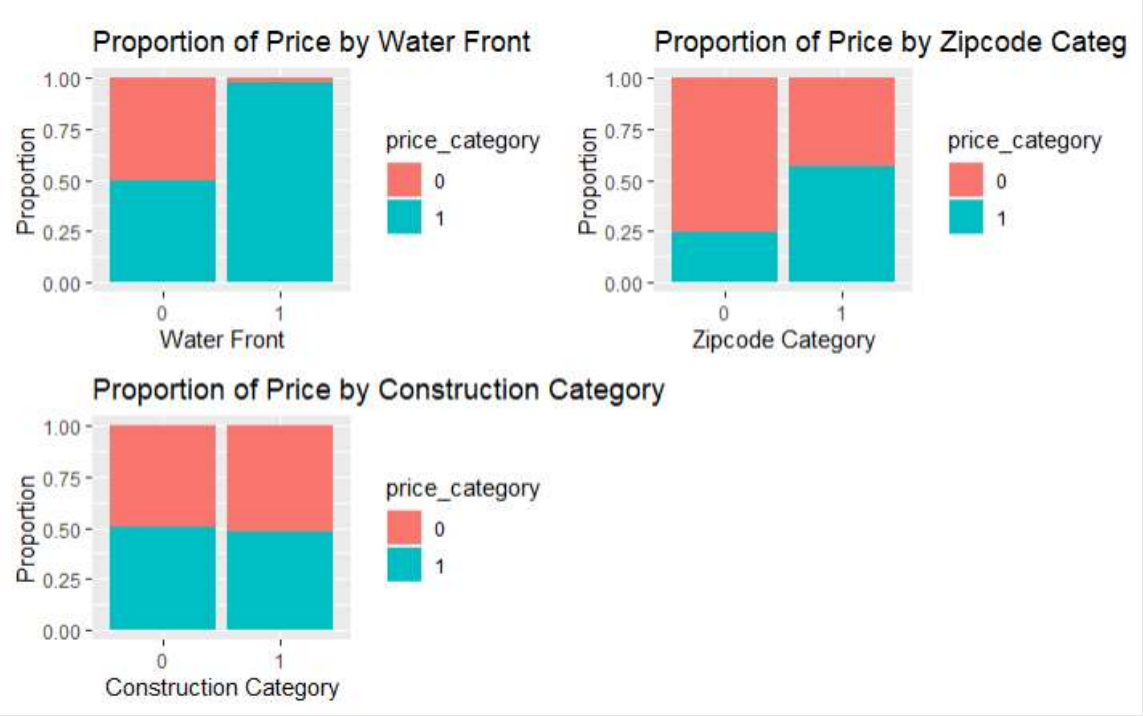
bedrooms, view, grade, and condition across price per square foot categorizations do not look much different. If anything, the density plot for bedrooms indicates a lightly higher density for houses with fewer bedrooms in the high price category. Similar to bedrooms, the plot for bathrooms shows a higher peak at lower numbers for the high price category. The plot for view shows a higher density at lower view for the low price category. The plots for condition and grade show a slight difference in the middle range between the low and high price categories, with the low price category being slightly higher. Therefore, it seems these predictor variables may not provide great insight into our response variable, and we note this as we move towards model selection.

Figure 20: Box Plots: Number of Bedrooms, Bathrooms, View, Grade, Condition, sqft_lot, sqft_living15 and sqft_lot15.



The above box plots of bedrooms, bathrooms, views, grades, conditions, sqft_lot, sqft_living15 and sqft_lot15 by price per square foot category (with 1 representing high prices per square foot) provide valuable insights into the variance between houses with high versus low prices per square foot. The number of bedrooms, number of bathrooms, view, grade, condition, sqft_lot, sqft_living15 and sqft_lot15 do not significantly differ between the two price categories. Therefore, it seems these predictor variables may not provide great insight into our response variable and we note this as we move towards model selection.

Figure 21: Bar Charts: Waterfront, Zip code, Construction Categorizations by Price per Square Foot Categorizations.



The bar charts are plotted to assess how each categorical predictor, including waterfront, zip code (rural vs urban), construction category (not recently renovated / constructed vs. recently constructed / renovated) affects the classification of a house as expensive or cheap per square foot. It appears that houses with a waterfront have a higher proportion of the high price category, and houses located in urban area have a higher proportion of the high price category. However, the plot for the construction category does not show a significant difference in the price category distribution. We note these relationships as we move forward to model selection for our logistic regression.

Section 7

We now look to fit a logistic regression model for our second question of interest, what influence do house qualities, aside from the size of the house, have on whether a house is categorized as high price or low price per living square foot ($/sqft).

We start with sqft_living, as we want to find influential variables other than house size. We also removed the ID variable from our set of predictors, as this variable is simply an identifier for the houses and is not suitable to be a predictor. Next, we also drop sqft_above and sqft_below from our set of predictors because these variables sum up to the sqft_living variable and would introduce redundant information into the model if included. We also drop the date variable because we are not looking at our analysis over time. Furthermore, we drop the latitude and longitude variables because we do not want to include spatial directions in the model as it does not make much sense in the context of our question. Additionally, we drop the year renovated and year built and instead use the new construction variable described in Section 2 above. Lastly, we drop the zipcode because we instead use the categorical variable that we created called zipcode_category that classifies zip codes as either urban or rural.

Figure 22: Wald Test with Full Model

```
Call:
glm(formula = price_category ~ bedrooms + bathrooms + sqft_lot +
    waterfront + grade + view + condition + grade + sqft_living15 +
    sqft_lot15 + zipcode_category + construction_category, family = "binomial",
    data = train)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -5.726e+00  3.369e-01 -16.996  < 2e-16 ***
bedrooms             -8.658e-01  4.726e-02 -18.320  < 2e-16 ***
bathrooms            -7.893e-01  6.824e-02 -11.566  < 2e-16 ***
sqft_lot              3.017e-06  1.332e-06   2.265 0.023501 *
waterfront1           2.164e+00  8.107e-01   2.669 0.007608 **
grade                 8.982e-01  4.892e-02  18.360  < 2e-16 ***
view                  5.816e-01  5.645e-02  10.303  < 2e-16 ***
condition             7.048e-01  5.395e-02  13.065  < 2e-16 ***
sqft_living15        -1.300e-04  7.452e-05  -1.744 0.081184 .
sqft_lot15           -1.220e-05  3.145e-06  -3.879 0.000105 ***
zipcode_category1     1.365e+00  8.848e-02  15.423  < 2e-16 ***
construction_category1  4.872e-01  9.518e-02   5.118 3.08e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7489.5  on 5402  degrees of freedom
Residual deviance: 5718.5  on 5391  degrees of freedom
AIC: 5742.5

Number of Fisher Scoring iterations: 6
```

As shown on the above figure, we conduct Wald test to evaluate the statistical significance of each predictor in our full model. All predictors in our model have small p-values, signifying that each has a statistically significant impact on determining the price category of houses.

Figure 23: Forward Selection

```
Call:  glm(formula = price_category ~ zipcode_category + bedrooms +
    grade + condition + view + bathrooms + construction_category +
    sqft_lot15 + waterfront + sqft_lot + sqft_living15, family = "binomial",
    data = train)

Coefficients:
          (Intercept)       zipcode_category1               bedrooms
            -5.726e+00               1.365e+00              -8.658e-01
                  grade              condition                   view
             8.982e-01               7.048e-01              5.816e-01
              bathrooms  construction_category1              sqft_lot15
            -7.893e-01               4.872e-01              -1.220e-05
             waterfront1               sqft_lot            sqft_living15
             2.164e+00               3.017e-06              -1.300e-04

Degrees of Freedom: 5402 Total (i.e. Null);  5391 Residual
Null Deviance:      7490
Residual Deviance: 5719         AIC: 5743
```

Figure 24: Backward Elimination

```
Call:  glm(formula = price_category ~ bedrooms + bathrooms + sqft_lot +
    waterfront + grade + view + condition + grade + sqft_living15 +
    sqft_lot15 + zipcode_category + construction_category, family = "binomial",
    data = train)

Coefficients:
          (Intercept)              bedrooms             bathrooms
            -5.726e+00            -8.658e-01            -7.893e-01
              sqft_lot            waterfront1                 grade
             3.017e-06             2.164e+00             8.982e-01
                  view             condition           sqft_living15
             5.816e-01             7.048e-01            -1.300e-04
            sqft_lot15      zipcode_category1  construction_category1
            -1.220e-05             1.365e+00             4.872e-01

Degrees of Freedom: 5402 Total (i.e. Null);  5391 Residual
Null Deviance:      7490
Residual Deviance: 5719         AIC: 5743
```

Figure 25: Stepwise Regression

```
Call:  glm(formula = price_category ~ zipcode_category + bedrooms +
    grade + condition + view + bathrooms + construction_category +
    sqft_lot15 + waterfront + sqft_lot + sqft_living15, family = "binomial",
    data = train)

Coefficients:
          (Intercept)       zipcode_category1               bedrooms
            -5.726e+00               1.365e+00              -8.658e-01
                  grade              condition                   view
             8.982e-01               7.048e-01              5.816e-01
              bathrooms  construction_category1              sqft_lot15
            -7.893e-01               4.872e-01              -1.220e-05
             waterfront1               sqft_lot            sqft_living15
             2.164e+00               3.017e-06              -1.300e-04

Degrees of Freedom: 5402 Total (i.e. Null);  5391 Residual
Null Deviance:      7490
Residual Deviance: 5719         AIC: 5743
```

As shown on the above figures, we conduct forward selection, backward elimination, and stepwise regression. Despite the different methodologies of adding and removing predictors, the results consistently showed that all predictors included in these analyses are significant. Consequently, none of the predictors were removed through these processes, indicating all predictors are important in determining house prices.

Figure 26: VIF

```
     bedrooms     bathrooms      sqft_lot    waterfront1         grade
       10.661        15.559        17.009        45.412        17.695
         view     condition sqft_living15     sqft_lot15 zipcode_category1
       12.840         6.974        12.923        31.096         7.286
construction_category1
                8.762
```

As shown in the figure above, we conducted a Variance Inflation Factor (VIF) analysis to assess multicollinearity. Waterfront displays an exceptionally high VIF of 45.412, and other variables including bedrooms, grade, view, sqft_living15, bathrooms, sqft_lot15, and sqft_lot show VIFs higher than 10, indicating significant multicollinearity.

We decide to remove waterfront from our model to see if the significantly reduces multicollinearity in our model.

Figure 27: VIF after Removing Waterfront

bedrooms	bathrooms	sqft_lot	grade	view
10.627	15.476	16.443	17.605	11.582
condition	sqft_living15	sqft_lot15	zipcode_category1	construction_category1
6.960	12.906	29.319	7.236	8.727

As shown in the figure above, the VIF is significantly reduced. Sqft_lot15 displays an exceptionally high VIF of 29.319, and other variables including bedrooms, grade, view, sqft_living15, bathrooms, sqft_lot15, and sqft_lot show VIFs higher than 10, indicating significant multicollinearity.

We decide to remove sqft_lot and sqft_lot15 from our model, as sqft_lot, sqft_lot15, and zipcode_category are variables used to predict whether an area is rural or urban. The larger the lot size, the higher the likelihood that it is located in a rural area.

Figure 28: VIF after Removing Sqft_lot & Sqft_lot15

bedrooms	bathrooms	grade	view	condition
10.518	15.450	17.467	11.303	6.906
sqft_living15	zipcode_category1	construction_category1		
12.665	6.845	8.565		

As shown in the figure above, the VIF is significantly reduced. Grade displays an exceptionally high VIF of 17.467, and other variables including bedrooms, view, sqft_living15, and bathrooms show VIFs higher than 10, indicating significant multicollinearity.

We decide to remove grade from our model to see if the significantly reduces multicollinearity in our model. We also decided to remove bathrooms and sqft_living15 from our model, as bedrooms, bathrooms and sqft_living15 are variables used to predict a house size.

Figure 29: VIF after Removing Grade, Bathrooms & Sqft_living15.

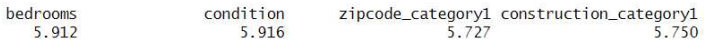
bedrooms	condition	view	zipcode_category1	construction_category1
6.606	6.281	9.364	6.075	6.099

As shown in the figure above, the VIF is significantly reduced. The highest remaining VIF is view, which is 9.364. All other variables show VIFs higher than 5, indicating moderate multicollinearity.

We conduct the likelihood test to determine if we can drop the view variable since the view can be correlated to the zipcode_cateogory and construction_category variables. The p-value, close to zero, and the test statistic of 302.07, larger than the critical value of 11.1, support dropping the view variable.

We decide to remove view from our model to see if the significantly reduces multicollinearity in our model.

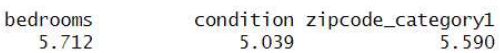
Figure 30: VIF after Removing View



All other variables show VIFs slightly higher than 5, indicating moderate multicollinearity. We decide to conduct another automated search based on bedrooms, condition, zipcode_category, and construction_category variables.

We conduct the likelihood test to determine if we can drop the construction_category variable since the bar chart indicated no relationship between the price of the house and this variable. The p-value, close to zero, and the test statistic of 106.3, larger than the critical value of 9.49, support dropping the construction_category variable.

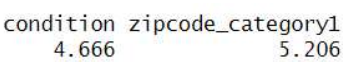
Figure 31: VIF after Removing Construction_category



VIFs of bedrooms and zipcode_category1 are still higher than 5, indicating moderate multicollinearity. We conduct another VIF to determine if we can drop the 'bedrooms' variable, as the number of bedrooms can be correlated with the zipcode_category, since the size of houses can vary based on specific locations.

We conduct the likelihood test to determine if we can drop the bedrooms variable since the bedrooms may be related to the sqft_living variable, which is incorporated in the response variable. The p-value, close to zero, and the test statistic of 402.1, larger than the critical value of 7.82, support dropping the view variable.

Figure 32: VIF after Removing Bedrooms



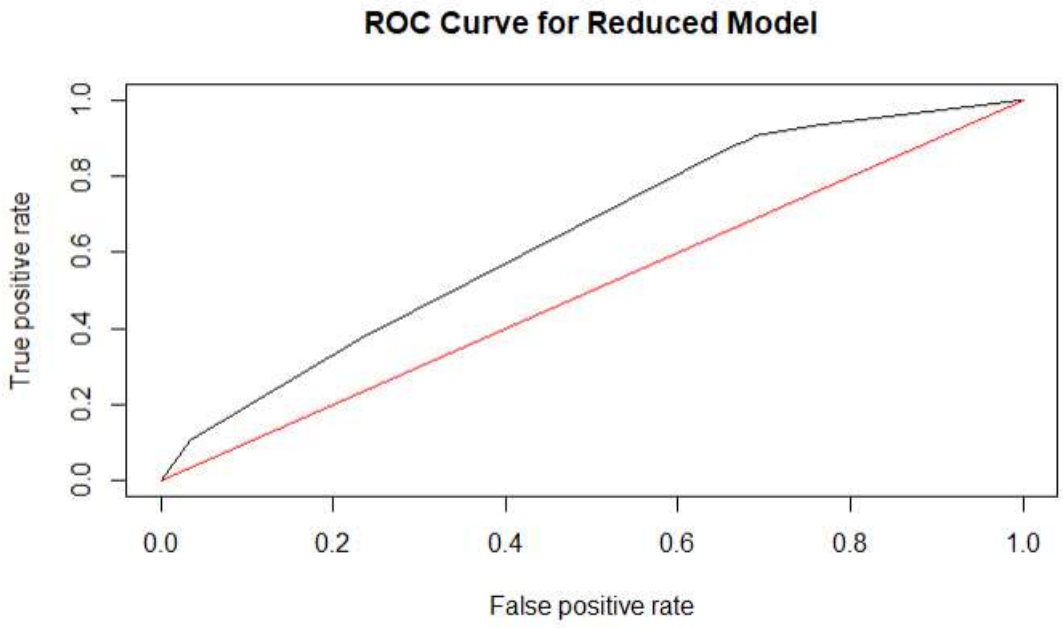
The VIF of the zipcode_category variable is only slightly higher than 5. We conduct Wald test to assess the degree of association between each variable and the response variable.

Figure 33: Wald test on Reduced Model

```
Call:
glm(formula = price_category ~ condition + zipcode_category,
    family = "binomial", data = train)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.59716    0.16387 -15.849   <2e-16 ***
condition          0.43854    0.04412   9.939   <2e-16 ***
zipcode_category1  1.40396    0.07479  18.772   <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: 7489.5  on 5402  degrees of freedom
Residual deviance: 6953.6  on 5400  degrees of freedom
AIC: 6959.6

Number of Fisher Scoring iterations: 4
```

The above result shows that each predictor variable has a high degree of association since p-value of each predictor variable is close to zero.

Figure 34: ROC Curve for Reduced Model



The above ROC curve is above the red line random classifier, suggesting the reduced model is better than the random model at predicting low / high house price category. However, the accuracy of the predictions is not very high, indicating that the variables are not perfect predictors. There may be other variables useful in predicting if a house is priced high or low per square foot. For example, variables relating to location may play a more important role in predicting whether a house in priced high or low per square foot and the generalized zipcode_category variable may not capture all of this location information.