# STAT 6021 Project 2

Group 3: Christina Land, Xinnie Mai, Kaya Oguz, Pranav Sridhar

2025-04-28

## 1. Summary of Findings

In the analysis of the King County, Washington housing dataset, statistical models were used to examine the factors that influence home prices and home quality in the region. The objective was to develop tools to support buyers, sellers, and real estate professionals in making more informed decisions.

In this project, the factors that make homes more valuable and of high quality in King County, Washington, were explored. Using the statistical models, it was identified that certain significant factors drive home prices and distinguish what is considered a good-quality home from others. A multiple linear regression model predicted home prices based on features such as home size, construction quality, renovation status, scenic views, waterfront presence, and location. It was found that larger, recently renovated homes with higher construction grades consistently sold for higher prices. Properties with waterfront access or high scenic view ratings also commanded significant price premiums.

Importantly, location played a significant role: homes situated further north and east in King County tended to be more expensive. After preparing the data, including removing outliers, correcting errors, and applying a log transformation to account for the skewed price distribution, the final model explained approximately 73% of the variation in home prices. This means that the model effectively captures most of the key factors influencing house values and can be a reliable tool for predicting price trends. Buyers can use these findings to prioritize home features that offer higher value for investment. Sellers can make more strategic renovation choices to increase property value. Real estate can use these data-driven insights to more effectively price and market homes based on their features and locations.

A logistic regression model was created to predict whether a home could be classified as "good quality". In the study, a good-quality home was defined as one with both a condition score above average and a higher construction grade, reflecting homes that are both well-maintained and well-built. The analysis revealed that good-quality homes were significantly more likely to have higher sale prices compared to recently renovated or newly built homes (after 1950), offered larger living spaces, provided scenic views or waterfront access, and were located further north in King County, especially near premium residential neighborhoods.

Larger living spaces and better views increased the likelihood of being classified as good-quality; the number of bedrooms alone was not a strong indicator of quality, demonstrating that simply adding more rooms does not guarantee better home quality. The findings indicate that no single feature can guarantee a home's high quality. Instead, home quality results from a combination of factors, including strong structural condition, a desirable location, larger living space, and appealing aesthetic features such as views or waterfront access.

Neighborhood characteristics, such as the typical size of surrounding homes, also played an essential role in influencing both home prices and perceived quality. Homes located in neighborhoods with larger average living spaces tend to be more expensive and are more likely to be classified as good-quality, indicating that community and surrounding property features also contribute to individual home value.

The logistic regression model performed exceptionally well, achieving an Area Under the Curve (AUC) score of 0.902, indicating that it could correctly distinguish between good-quality homes and others more than 90% of the time. Therefore, buyers looking for high-quality homes can focus their searches on recently renovated, scenic, waterfront properties in key northern neighborhoods. Sellers should emphasize these features when marketing their homes. Builders and developers can also use these insights to design new homes that meet the high-quality standards most valued in this region.

Overall, log transformations significantly improved the model's ability to explain house prices. This confirmed that real estate prices tend to grow at a nonlinear rate. High multicollinearity was detected between variables such as living area and neighborhood size, highlighting the importance of careful variable selection for stable predictions. Outliers, such as large mansions or waterfront estates, heavily influenced early models until the data was refined, highlighting the importance of thoughtful data cleaning in real-world housing analysis. Not all expensive homes were necessarily of high quality, reinforcing that structural condition and build quality still matter, not just price.

Together, the models provide data-driven tools to support informed decision-making across all aspects of the housing market in King County. Homebuyers can better understand what drives pricing and quality, helping them make more informed offers. Sellers can focus investments on renovations or improvements that truly add market value.

# 2. Data Description

The dataset consists of home sales from May 2014 to May 2015 in King County, Washington, covering 21,613 observations. After cleaning the dataset, including removing outliers and incomplete records, the final dataset was refined to improve reliability.

Below is a detailed description of key variables in the final model and their relevance to the house price analysis:

***Original Variables***

- Price (Numeric): Price of house sale in USD
- Bedrooms (Numeric): Total number of bedrooms
- Bathrooms (Numeric): Total number of bathrooms
- Sqft_living (Numeric): Square footage of total living area
- Sqft_lot (Numeric): Square footage of lot size
- Floors (Numeric): Number of floors
- Waterfront (Factor/Categorical): Whether the property has a waterfront view (0 = no, 1 = yes)
- View (Factor/Categorical): Rating of view quality (0 - 4 scale)
- Condition (Factor/Categorical): House condition scale (1 - 5 scale)
- Grade (Factor/Categorical): House grading scale (3 - 13 scale)
- Yr_built (Numeric): Year of house construction
- Yr_renovated (Numeric): Year the house was renovated (NA if never renovated)

- Lat (Numeric): Geographic latitude
- Long (Numeric): Geographic longitude
- Sqft_living15 (Numeric): Average size of interior housing living space for nearest 15 neighbors

***Created Variables***

- Effective_yr_built (Numeric): Represents the most recent year of significant construction or renovation a home underwent and uses either the original year built or renovation year if the property has undergone renovation. This effectively captures the current age or recency of structural improvements of the home.
- home_quality (Factor/Categorical): Defines a home to be good quality if its condition is greater than 3 and grade is greater than 7

***Eliminated Variables***

- Id (Numeric): Served only as a unique identifier with no predictive power
- Sqft_above (Numeric): This would be a subset of the sqft_living (total living area), using it would simultaneously cause multicollinearity
- Sqft_basement (Numeric): This would be a subset of the sqft_living (total living area), using it would simultaneously cause multicollinearity
- Sqft_lot15 (Numeric): Weaker influence compared to sqft_living15, as it is the average size of lots for the nearest 15 neighbors and may cause redundancy with sqft_living15
- Zipcode (Numeric): Did not seem intuitive to be a quantitative predictor and there were too many distinct values for it to be an effective factor/categorical variable
- Date (Numeric): Hard to interpret

# 3. Data Issues: Observation and Variable Issues

Upon ingesting the King County Housing data, several suspicious or invalid observations were identified:

- Date formatting error: The date variable was initially recorded as a timestamp character string (e.g., "20140502T000000"), which is not in a standard date format and unsuitable for time-based analysis or visualization. This likely resulted from a data entry issue.

- Unrealistic number of bedrooms and bathrooms: Several entries recorded zero bedrooms or zero bathrooms, which is highly unlikely for residential listings. These likely reflect missing values coded incorrectly as zeros.

- Extreme outliers in price and bedrooms: Some homes were listed with more than 10 bedrooms, and several properties had sale prices above the 99th percentile (e.g., over $2,000,000). While some may represent true luxury listings, these outliers are highly atypical and could skew statistical models.

To address these issues and clean the data:

- The date variable was converted from a timestamp string to a proper Date object (YYYY-MM-DD format) to support time-based analysis and visualization.

- Observations with unrealistic numbers of bedrooms or bathrooms were removed.

- Extreme outliers were removed, specifically homes with more than 10 bedrooms and those with sale prices above the 99th percentile.

Several variables also required restructuring or removal due to data type issues or redundancy:

- Misuse of yr_renovated: The yr_renovated variable used 0 to indicate no renovation, which could be misinterpreted as an actual year (year 0) in regression models.
- Categorical variables stored as integers: Variables such as waterfront, view, condition, grade, and zipcode were stored as integers, incorrectly suggesting a continuous or ordinal relationship.
- Redundant or non-predictive variables: Variables like id (a unique identifier) provided no predictive value, while others like sqft_above, sqft_basement (components of sqft_living), and sqft_lot15 (highly correlated with sqft_lot) introduced redundancy.

To finalize preparation of the dataset for modeling:

- A new variable, effective_yr_built, was created by replacing 0 values in yr_renovated with NA and setting it to yr_renovated when available or yr_built otherwise. This better represents the most recent structural update of each home.
- Categorical variables (zipcode, view, condition, grade, and waterfront) were converted from integers to factors to ensure correct handling in regression models.
- The date variable was reformatted to a standard Date type, resolving issues with the original timestamp string.
- Redundant and non-predictive variables (id, sqft_above, sqft_basement, and sqft_lot15) were removed to reduce dimensionality and prevent multicollinearity.

# 4. Price and Influential Factors

To explore how various factors influence home prices, data visualizations were provided across four key categories: Property Size and Structure, Property Quality and Condition, Location and Neighborhood, and Other Features such as date of sale and seasonal trends. These visualizations allowed us to assess relationships between home prices and other variables.

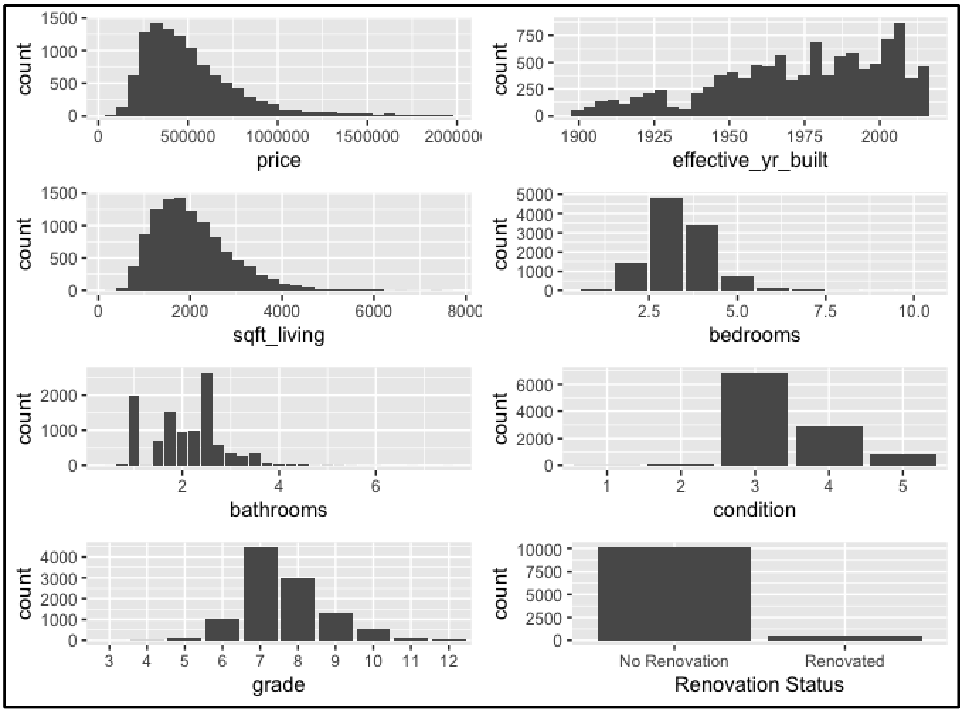


***Figure 1:*** *Distribution of Key Housing Features*

The histograms reveal the distributions of several key housing attributes—bedrooms, bathrooms, condition, grade, and renovation status—offering insights into the overall composition and quality of the homes in the dataset. Most homes feature 3 or 4 bedrooms, with 3 bedrooms being the most common. Large homes with 8 or more bedrooms are rare. In terms of bathrooms, the majority of properties have between 1.5 and 2.5 bathrooms, consistent with typical mid-sized family homes.

Most homes are rated in the middle range for condition, with the majority falling between a condition rating of 3 or 4, suggesting an average to good state of maintenance. Homes rated in the extremes (either very poor or excellent) are relatively uncommon. Similarly, the majority of homes fall within grade 6 to 8, indicating mid-range quality and construction, while homes with either very low or very high grades are rare, highlighting a market dominated by moderately high-quality properties.

Homes with more than 4 bathrooms are uncommon, and the majority of homes have not undergone renovations, suggesting that many remain in their original condition. A small number of properties show evidence of renovation, which could potentially influence their market value.

Most homes feature 3 or 4 bedrooms, with 3 being the most common, while large homes with 8 or more bedrooms are rare. In terms of bathrooms, the majority of homes have between 1.5 and 2.5, which aligns with typical mid-sized family homes.

The condition ratings for most homes fall between 3 and 4, suggesting that the majority are in average to good condition, with very poor or excellent condition ratings being relatively uncommon. Similarly, most homes fall within grade 6 to 8, indicating mid-range construction and quality, with few homes rated either very low or very high, suggesting a housing market that is primarily of moderate to high quality.

Homes with more than 4 bathrooms are rare, and the vast majority of homes have not been renovated, indicating they are in their original condition. Only a small percentage of homes show evidence of renovation, which may influence their market value in some cases.

Overall, the dataset consists primarily of mid-sized homes with 3–4 bedrooms, 1.5–2.5 bathrooms, and rated average to good in both condition and grade, with renovations being a relatively rare feature. Also, the dataset is dominated by mid-sized homes (3–4 bedrooms, 1.5–2.5 bathrooms), with most properties rated average to good in both condition and grade, and renovated homes being relatively rare.

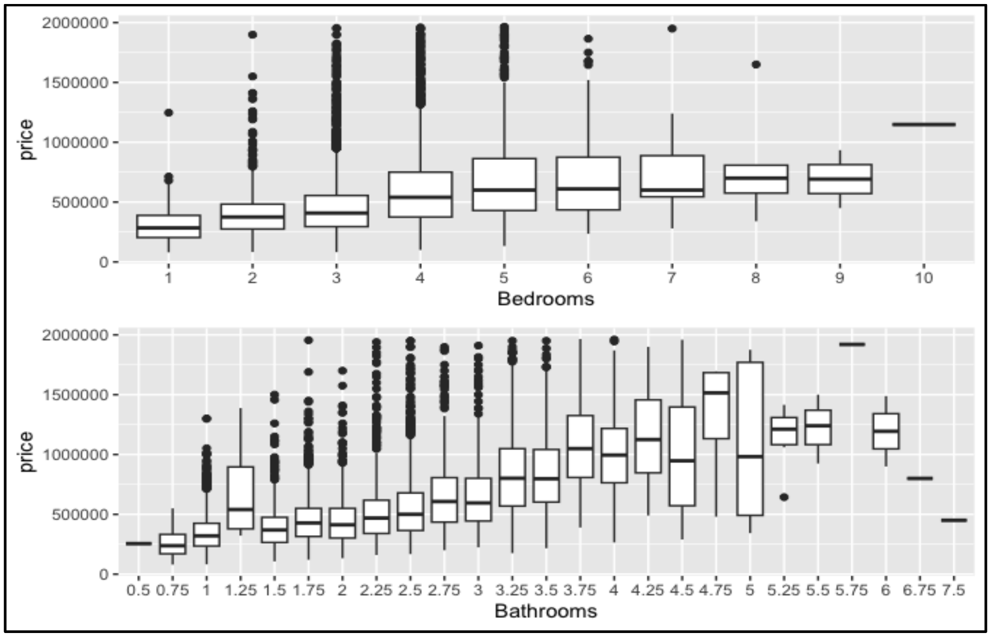


***Figure 2:*** *House Pricing by Number of Bedrooms and Bathrooms*

The above boxplots examine how bedroom and bathroom counts relate to home prices. It can be seen that there is a slight increase in median price with more bedrooms, but the trend is inconsistent beyond five bedrooms, and price variability remains high. In contrast when observing the boxplot for bathrooms, it is revealed that there is a more consistent rise in median price as the bathroom count increases from 1 to 4, with fewer low-price outliers in higher bathroom categories. Overall, the bathroom count is a stronger and more reliable predictor of home price than the bedroom count.

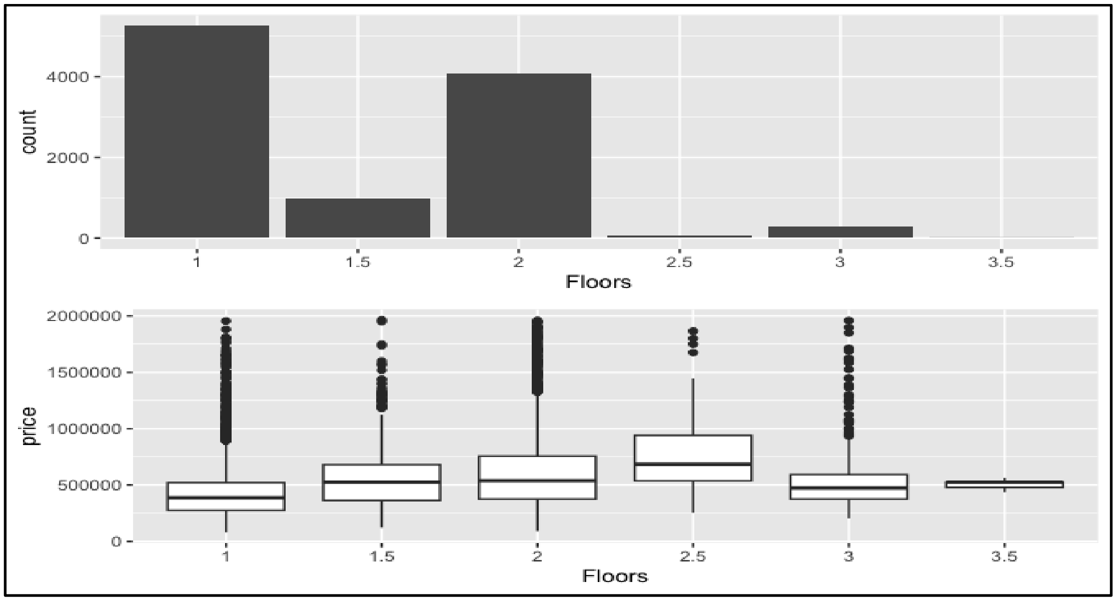


***Figure 3:*** *House Pricing by Number of Floors*

These two visualizations above analyze the distribution of homes by the number of floors and how floor count relates to home prices. Most homes have 1 or 2 floors, while multi-level homes (e.g., 1.5, 2.5, 3 floors) are uncommon. The boxplot shows no clear linear relationship between floor count and price. Although homes with 2 and 2.5 floors have slightly higher median prices, high-priced outliers appear across all floor categories. This suggests that floor count has a limited and nonlinear impact on home price.

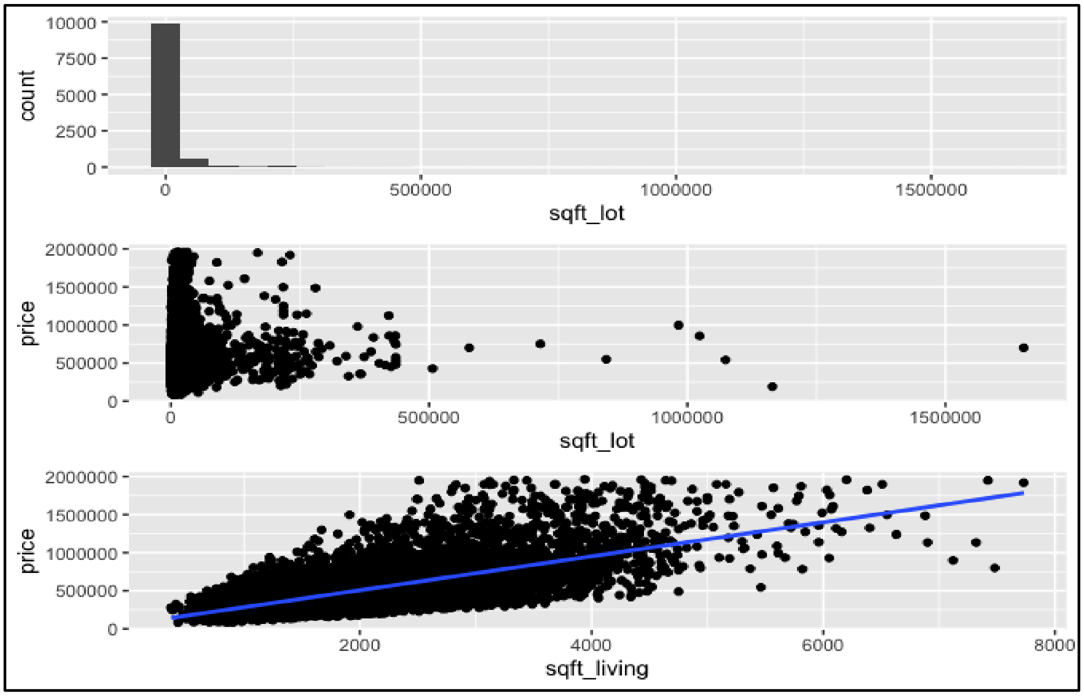


***Figure 4:*** *Impact of Lot Size and Living Square Footage on Price*

These plots analyze the relationship between lot size and home price, adding valuable insight into how this variable impacts property value. The first plot shows a right-skewed distribution, with most lots under 20,000 sqft and a few significant outliers distorting from the scale. The second plot shows no clear linear relationship between lot size and price, where many high-priced homes have modest-sized lots, which means that lot size alone is not a strong predictor of value. While some large-lot properties are expensive, the lot size on price appears limited, which means that other factors play a more significant role in determining home value. In the scatterplot of price against living square footage, a strong positive linear relationship is shown. As living space increases, home prices tend to rise significantly. The fitted regression line emphasizes this. While the trend is generally going upward, there's noticeable dispersion. This is particularly true for larger homes, indicating that they are beyond a specific size. Overall, the living area is a strong predictor of home price.

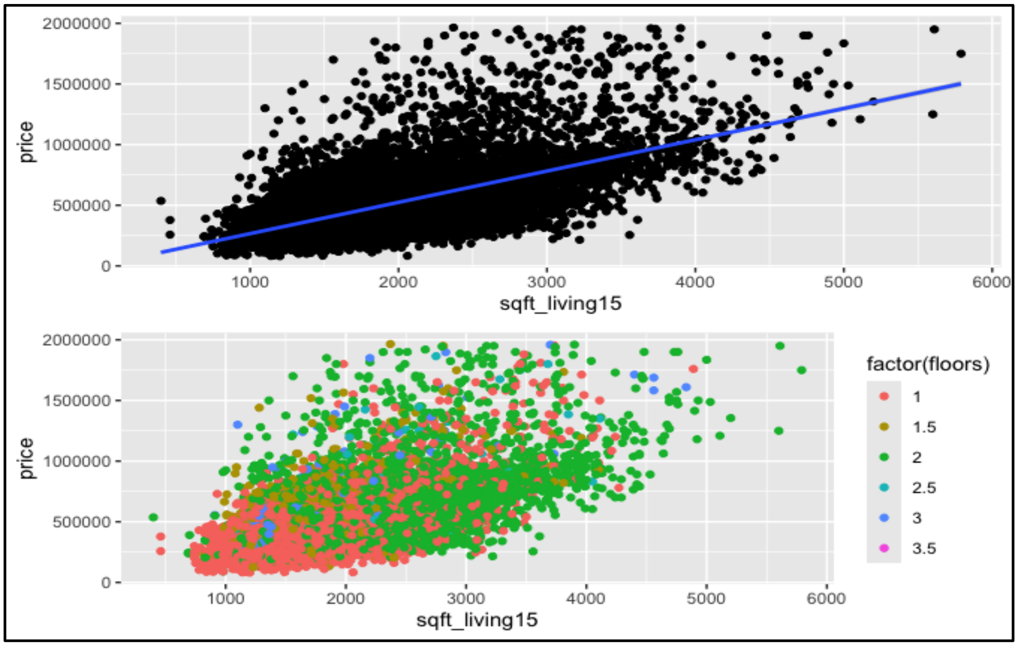


***Figure 5:*** *Influence of Neighborhood Size and Floor Count on House Pricing*

These two scatter plots demonstrate a positive correlation between neighborhood average living space (sqft_living15) and home prices. It indicates that homes in areas with larger average home sizes tend to be more expensive. Floor count adds a stratifying layer to this relationship, where most homes have 1 or 2 floors, with 2-floor homes (green), which dominates across all price levels, in particular the higher range. In contrast, 1-floor homes (red) are concentrated in the lower-price, smaller-size segment, while homes with more than 2 floors are less common but tend to align with higher prices.

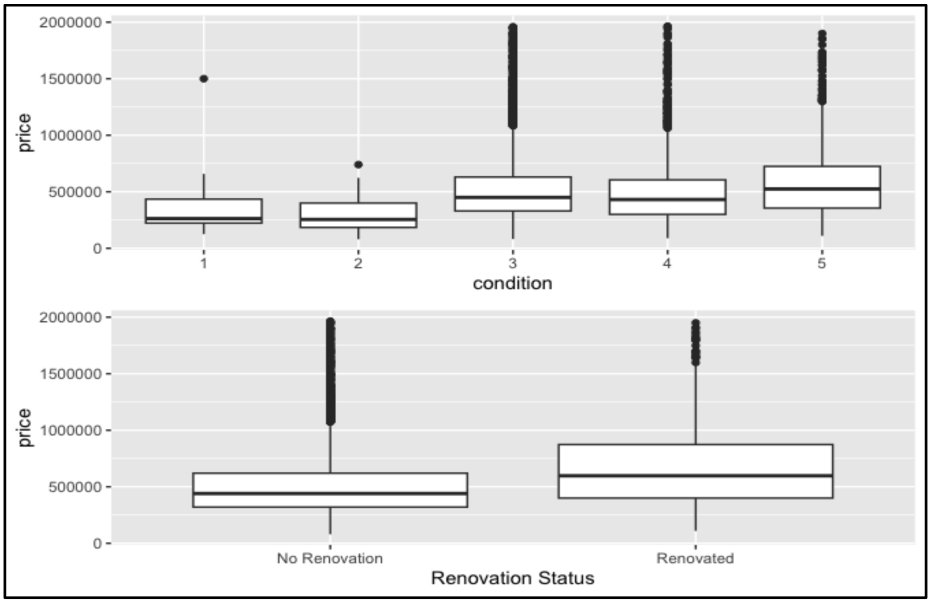


***Figure 6:*** *House Condition and Renovation Status*

The two boxplots presented here demonstrate the impact of how house conditions and renovation status influence home prices. In the first boxplot, median prices increase slightly with homes that have a better condition rating. Still, price variability is high across all levels, which indicates that condition alone is not a strong influencer of value. In contrast, the second boxplot shows that renovated homes consistently have higher median prices and broader upper price ranges, indicating that renovation contributes more significantly to a home's market value than condition alone.

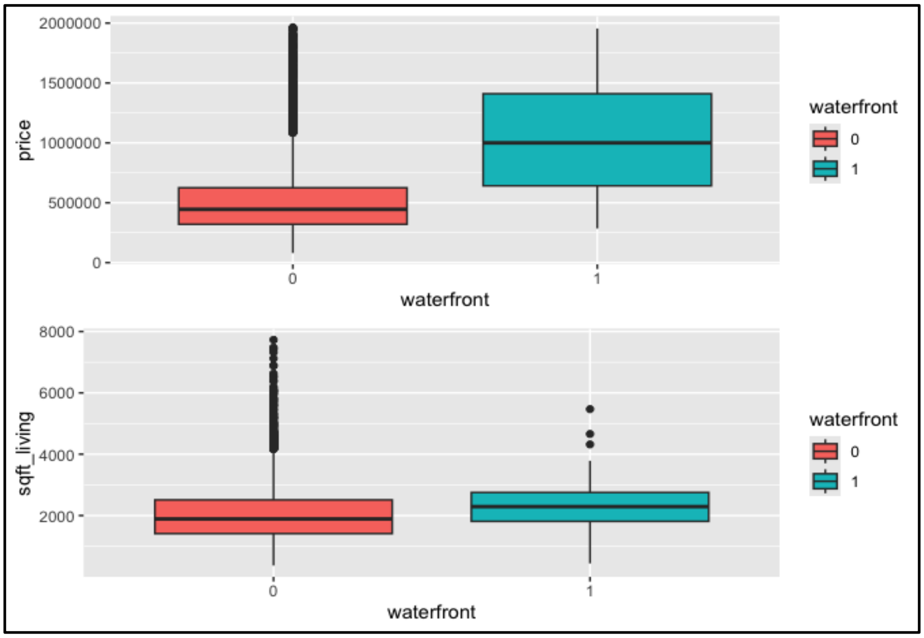


***Figure 7:*** *Waterfront House Location and Price*

The data reveals the impact of waterfront status on house prices and size. Homes with a waterfront location generally show significantly higher median prices and a wider price range compared to non-waterfront properties. Additionally, waterfront homes tend to have slightly larger living areas on average. However, the price premium for waterfront homes appears to be more driven by their location rather than their size, emphasizing the value of scenic or exclusive positioning. This highlights the substantial role that location and scenic amenities play in real estate value, positioning waterfront status as a key factor in high-end home pricing.

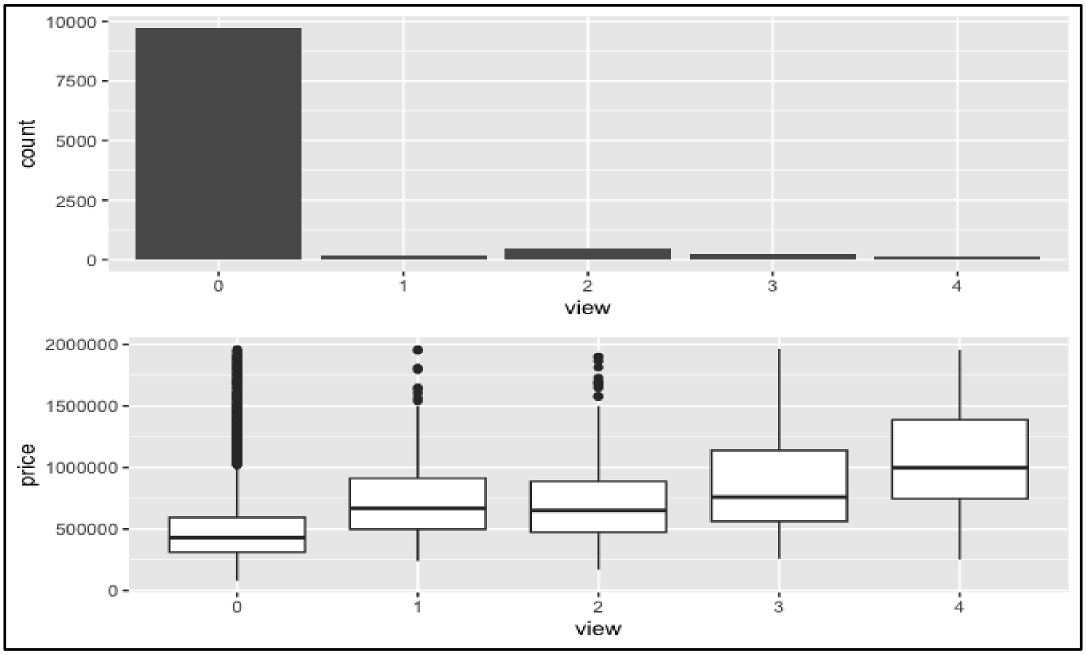


***Figure 8:*** *Impact of View Rating on Property Value*

The distribution and pricing impact of scenic views are evident in the data. Most homes have a view rating of 0, indicating no notable view, while homes with higher ratings (1–4) are relatively rare. As view ratings increase, there is a clear upward trend in median price, with the highest-rated homes (view 4) commanding significantly higher prices and wider price ranges. Even homes with modest views (ratings 1–2) show a noticeable increase in value. These findings highlight that scenic views significantly enhance property value, with their rarity further amplifying their impact on pricing.

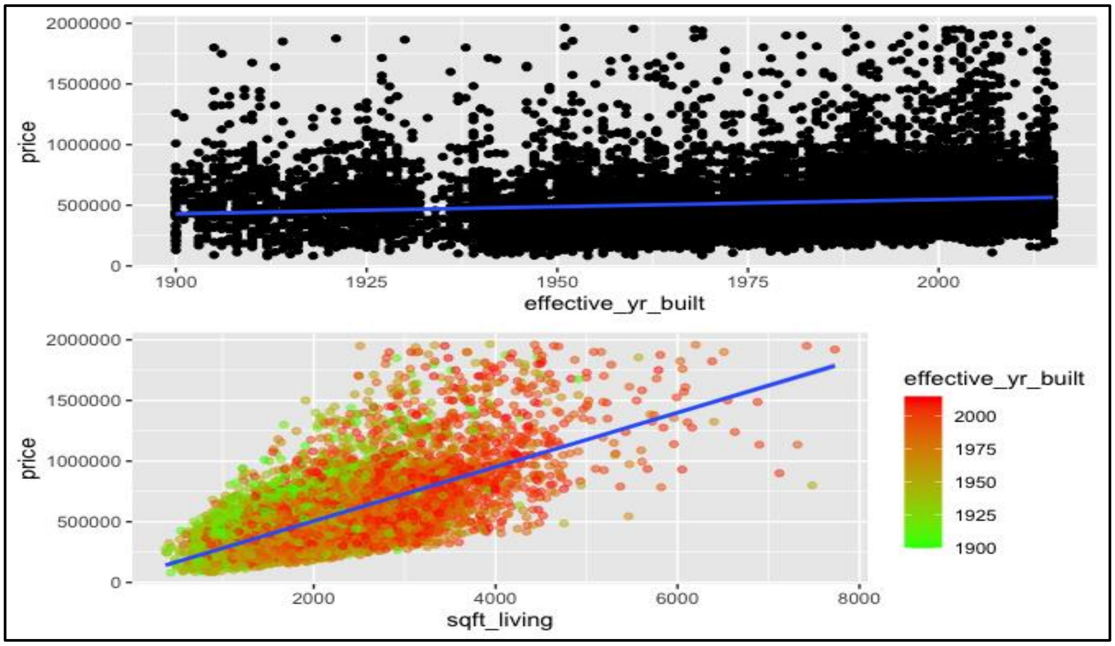


***Figure 9:*** *Impact of Effective Year Built and Living Area on Price*

The data highlights the influence of construction year and home size on price trends, as well as the relationship between home prices, effective year built, and living area, illustrating how these factors affect property value. There is a slight upward trend in prices over time, with newer homes (built post-2000) generally priced higher, despite some variability across decades. Additionally, a strong positive correlation is observed between living area and price, with newer homes typically being both larger and more expensive. These findings suggest that newer construction and larger home size are significant drivers of higher property value.

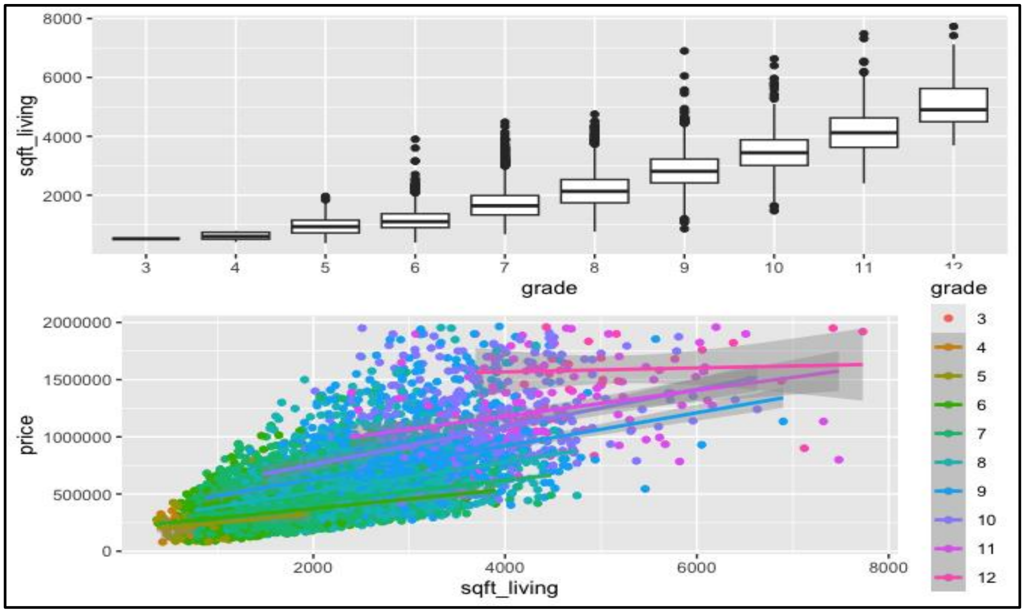


***Figure 10:*** *Relationship between Construction Quality (Grade) and Size*

The visualization illustrates the combined influence of construction quality and size on home prices. Higher-grade homes are generally larger, with greater size variation as grade increases. Additionally, while price rises with size across all grades, the rate of price increase is more pronounced for higher-grade homes. This indicates that premium construction enhances the price impact of additional square footage, further emphasizing the value of high-quality construction in pricing.

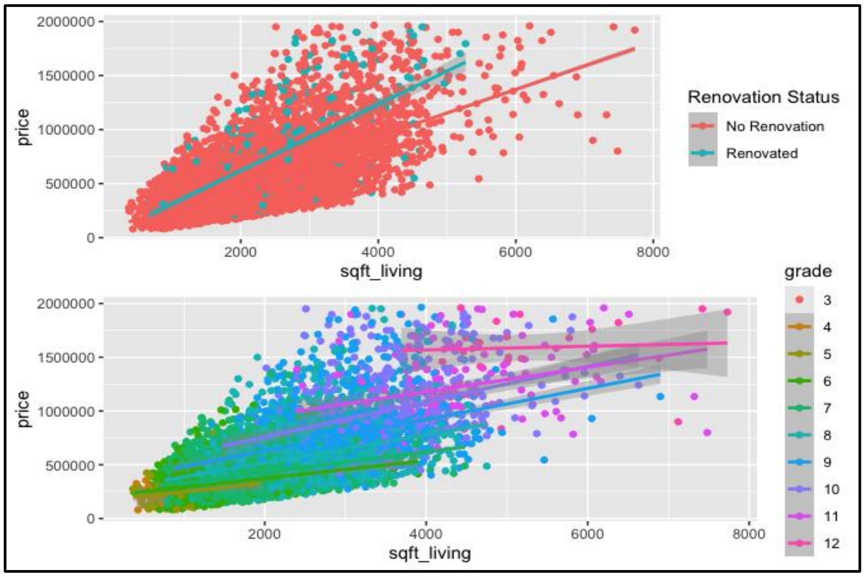


***Figure 11:*** *Renovation Status, Living Area, and Construction Quality (Grade) impacts on Price*

The above plots compare the effects of renovation and construction quality on price per square foot, illustrating how these factors interact with living area to influence home prices. Renovation status and construction quality play significant roles in determining the value of a property. Renovated homes generally command higher prices at any given square footage compared to non-renovated homes, with a more pronounced price trend as size increases, suggesting that renovations add considerable value. Additionally, construction quality (grade) impacts the price per square foot, with higher-grade homes not only tending to be larger but also experiencing steeper price increases per square foot. This highlights that both home size and construction quality work together to significantly increase property value, particularly in higher-end homes.

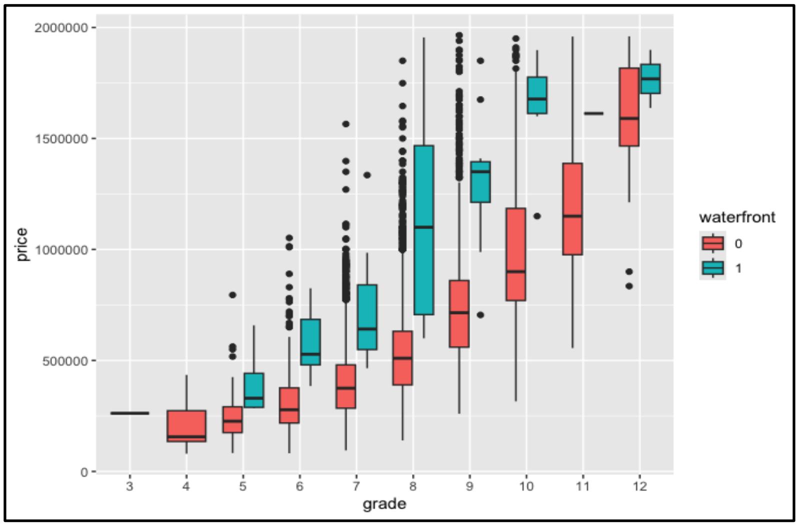


***Figure 12:*** *Impact of Construction Quality (Grade) on Price for Waterfront and Non-Waterfront Homes*

The above boxplot showing how house grade and waterfront status affect home prices. Prices rise consistently with higher grades, demonstrating the value of reconstruction. Waterfront homes (blue) have higher median prices across nearly all grade levels compared to non-waterfront homes (red), with the price gap widening at higher grades. This shows that waterfront access can further amplify the value of already high-quality homes, especially at the upper end of the market.

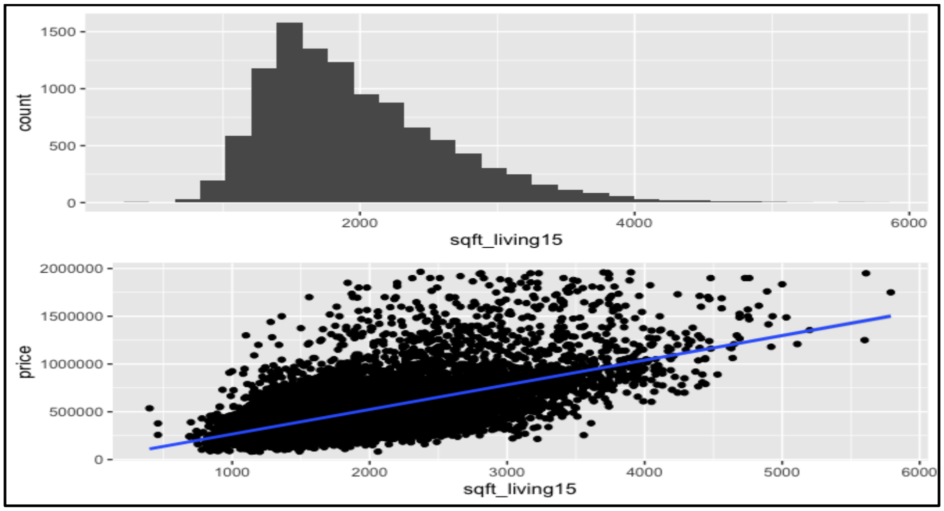


***Figure 13:*** *Neighborhood Size impact on Price*

The above two visuals present the distribution of neighborhood average living area (sqft_living15), demonstrating insight into typical home sizes within local communities. The distribution is right-skewed, with most neighborhood averages clustered between 1,200 and 2,500 square feet. The most common average size falls just below 2,000 sqft, representing that mid-sized homes dominate the market. A small number of neighborhoods show much larger average sizes (over 4,000 sqft). Most neighborhoods have featured moderately sized homes, with average living areas below 2,500 sqft. The skew toward larger averages in a few neighborhoods indicates an influence on local property values and pricing trends.

The relationship between the neighborhood average living area (sqft_living15) and home price in the above figure also highlights how neighborhood characteristics can influence property value. There is a strong positive correlation between average neighborhood home size and individual home price. The upward-sloping regression line confirms that homes in neighborhoods with larger average home sizes tend to be more expensive. Despite some scatter, the trend indicates that neighborhood context, especially the size norms, plays a significant role in shaping home values. Homes in neighborhoods with larger average living areas and higher prices can reflect the influence of surrounding property characteristics on perceived value.

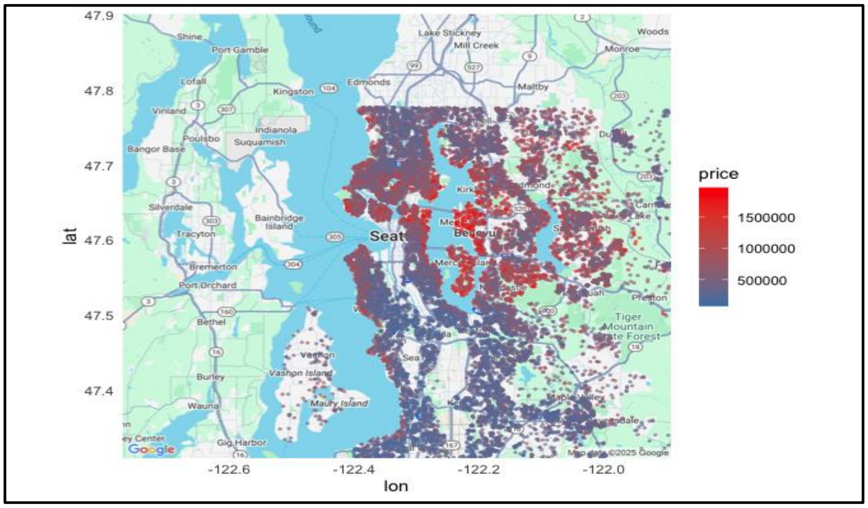
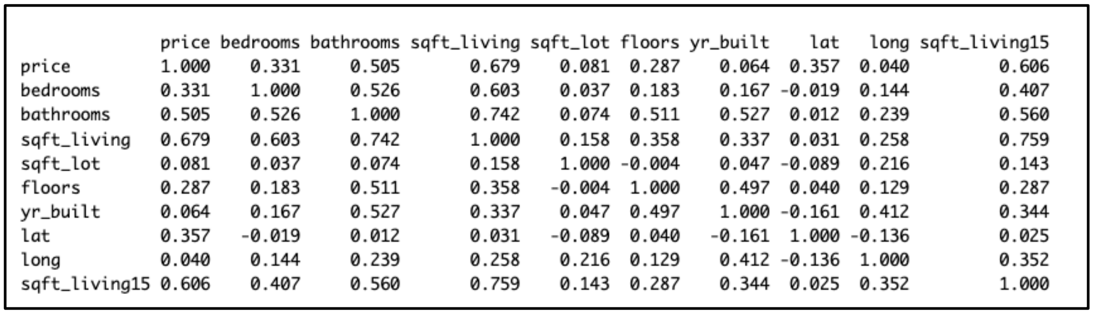


***Figure 14:*** *Impact of Geographic Location (Latitude and Longitude) on Price*

Lastly, the above map shows a visualization of home prices across King County, Washington, using color to represent price levels. Darker red points indicate higher-priced homes, while lighter blue points represent lower-priced homes. A clear spatial pattern emerges: higher home prices are concentrated in the northeast quadrant of the map. In contrast, southern and southwestern areas display consistently lower prices.

# 5. Multiple Linear Regression (MLR) for Price

From the previous section, there appear to be a multitude of factors that can influence house pricing. As a result, a multiple linear regression model was constructed to identify key predictors (quantitative and categorical) to best predict a house's price. Examining univariate and multivariate visualizations of the relationship between variables and price gave a starting point for variables to consider in the initial model.



```
              price bedrooms bathrooms sqft_living sqft_lot floors yr_built    lat   long sqft_living15
price         1.000    0.331     0.505       0.679    0.081  0.287    0.064  0.357  0.040         0.606
bedrooms      0.331    1.000     0.526       0.603    0.037  0.183    0.167 -0.019  0.144         0.407
bathrooms     0.505    0.526     1.000       0.742    0.074  0.511    0.527  0.012  0.239         0.560
sqft_living   0.679    0.603     0.742       1.000    0.158  0.358    0.337  0.031  0.258         0.759
sqft_lot      0.081    0.037     0.074       0.158    1.000 -0.004    0.047 -0.089  0.216         0.143
floors        0.287    0.183     0.511       0.358   -0.004  1.000    0.497  0.040  0.129         0.287
yr_built      0.064    0.167     0.527       0.337    0.047  0.497    1.000 -0.161  0.412         0.344
lat           0.357   -0.019     0.012       0.031   -0.089  0.040   -0.161  1.000 -0.136         0.025
long          0.040    0.144     0.239       0.258    0.216  0.129    0.412 -0.136  1.000         0.352
sqft_living15 0.606    0.407     0.560       0.759    0.143  0.287    0.344  0.025  0.352         1.000
```

***Figure 15(a):*** *Correlation Matrix of Quantitative Predictors with Price*

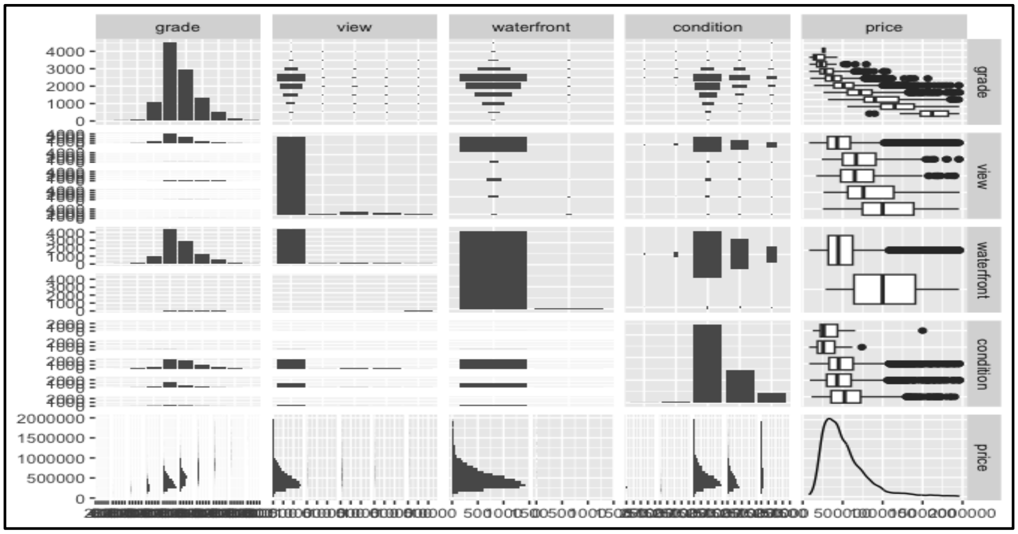


***Figure 15(b):*** *Correlation Matrix of Qualitative (Categorical) Predictors with Price*

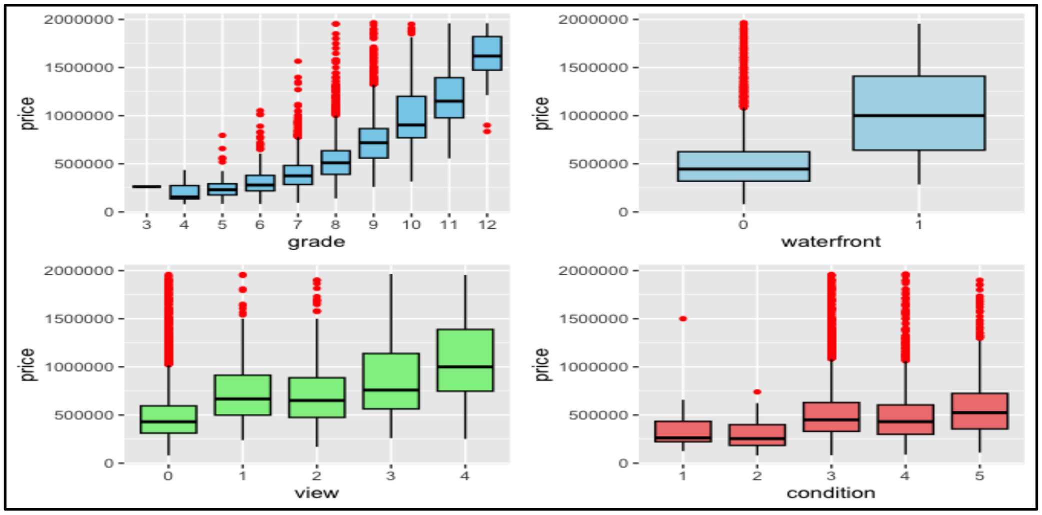


***Figure 16:*** *Box Plots for Price versus Categorical Variables*

A correlation matrix was used to select numerical predictors for the initial regression model by identifying variables with strong linear relationships with price (high correlation values). As shown in Figure 15a above, these numerical predictors included bedrooms, bathrooms, sqft_living, floors, lat, and sqft_living15. After examining the numerical predictors, a correlation matrix was employed to explore the relationship between categorical predictors and price (Figure 15b). This helped identify several variables that were strongly linked to the response including grade, view, waterfront, and condition. To further assess their linearity with price, individual scatter plots were created for these categorical predictors (Figure 16). As shown by the box plots of price against the categorical predictors, a general positive linear relationship appears between each predictor.. This reaffirmed the inclusion of these variables in the regression model. After the key numerical predictors and categorical variables were identified, they were fitted into an initial multiple linear regression model.

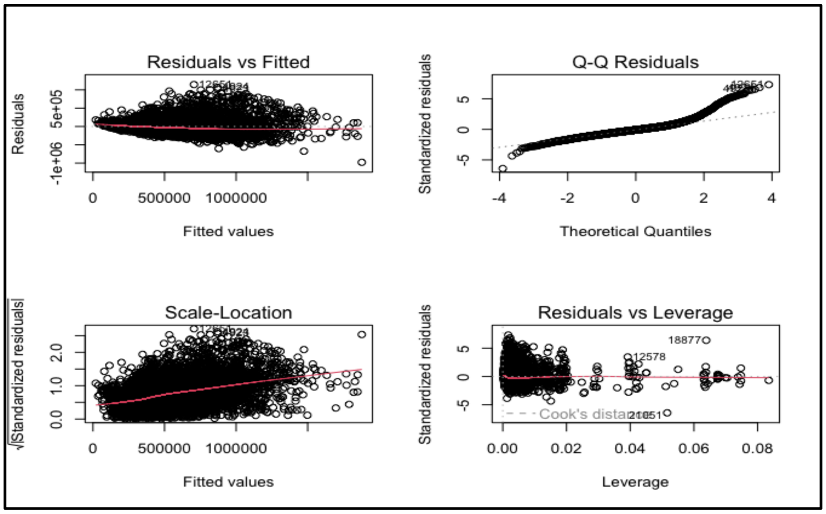


***Figure 17:*** *Residual plot and associated plots for initial linear regression model*

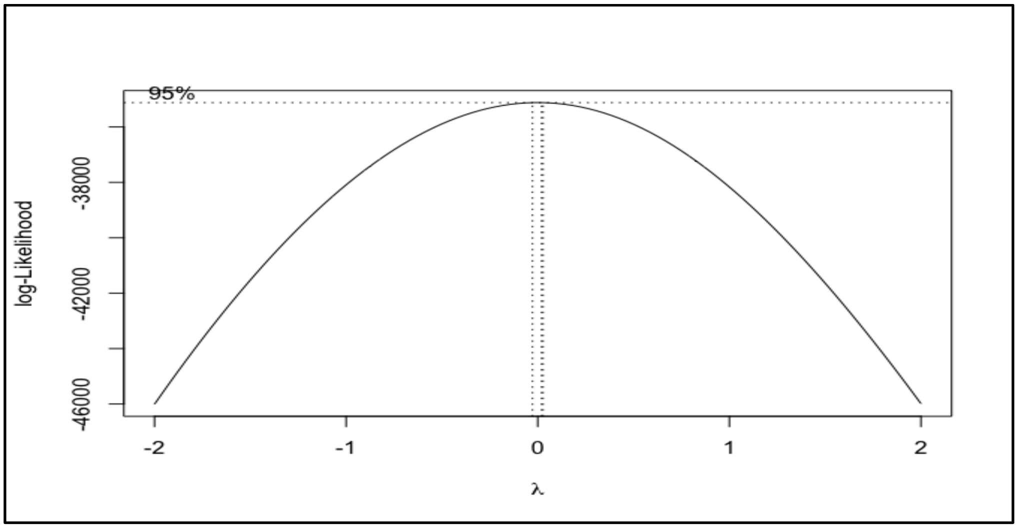


***Figure 18:*** *Residual Plots and Boxcox Plot*

After the initial model was fitted, the regression assumptions, including homoscedasticity, normality of residuals, and linearity, were carefully examined in the figure above. Heteroscedasticity was observed in the residuals. To address this, a Box-Cox transformation was applied as shown in the figure above, revealing that a log transformation of the dependent variable was necessary to improve the model's fit. After the log transformation was applied to the response variable, the model was refitted with log price as the dependent variable, and the regression assumptions were met.

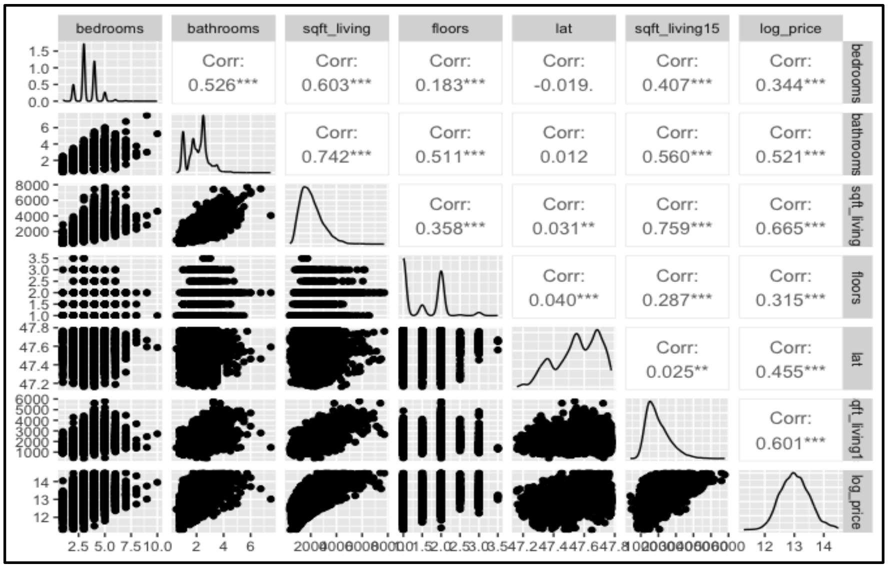


***Figure 19:*** *Scatterplot matrix for quantitative predictors*

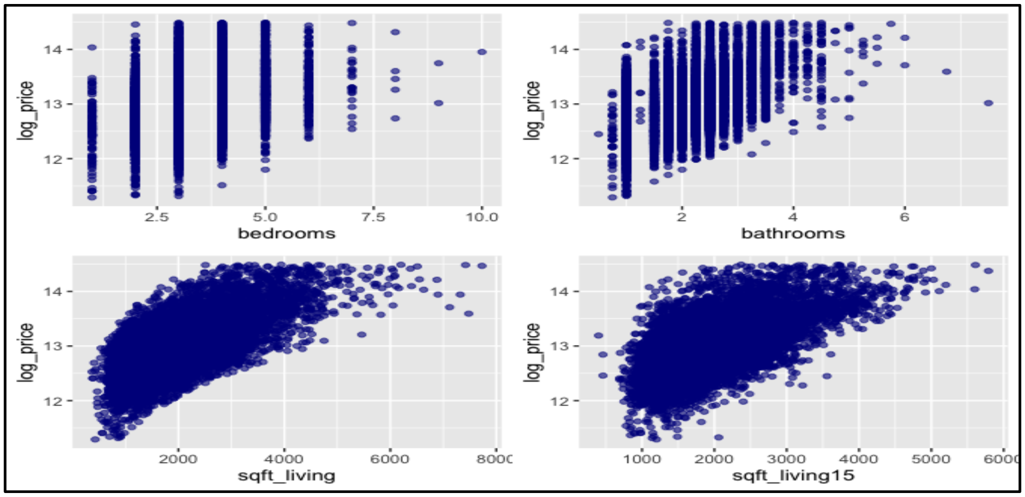
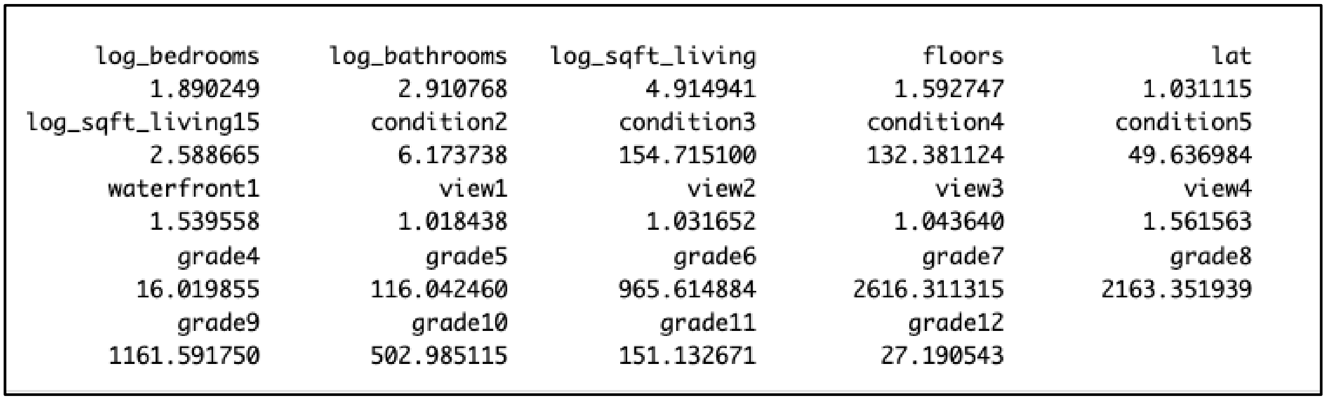


***Figure 20:*** *Scatterplot of log(price) against various predictor variables*

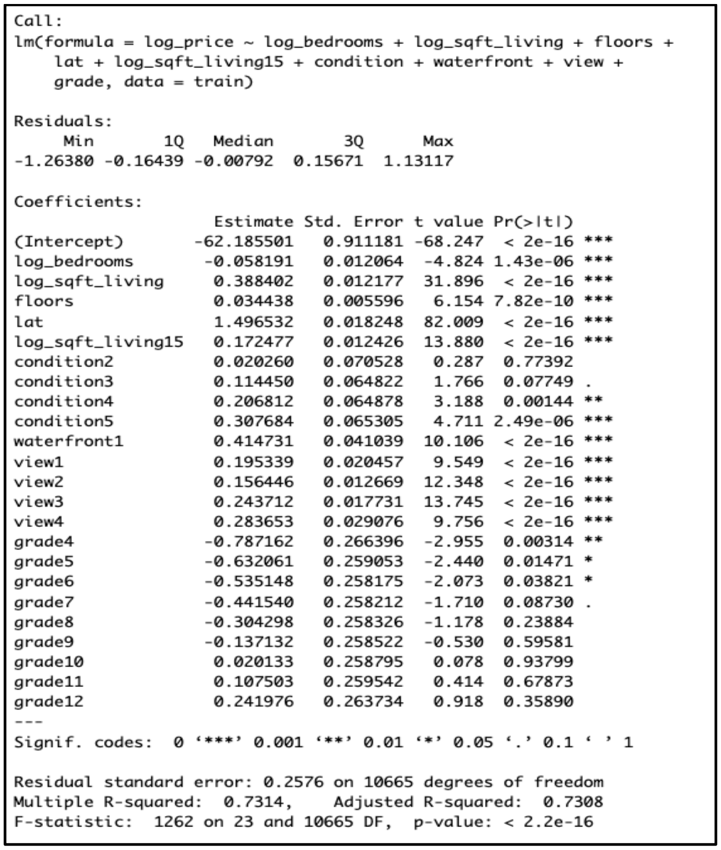
Then, a scatterplot matrix of the predictors included in the model was created to visually inspect their relationships with log_price to assess the need for any transformations of predictors, checking for any logarithmic relationships with log_price (Figure 19). Bedrooms, bathrooms, sqft_living, and sqft_living15 appeared to be logarithmically associated with log_price. Individual scatterplots of these predictors against log_price was created to further confirm these relationships (Figure 20). Once these were identified, the predictors were log-transformed and the model was refitted. Following the refitting of the model, the regression assumptions were reassessed and met with no signs of heteroscedasticity or nonlinearity. Next, VIF's of the predictor variables were examined for multicollinearity in the figure below. The results showed that the only variables with high VIF values were grade and condition, which was expected, thus, no changes were made in terms of multicollinearity.



```
    log_bedrooms    log_bathrooms   log_sqft_living          floors             lat
        1.890249         2.910768          4.914941        1.592747        1.031115
log_sqft_living15      condition2        condition3      condition4      condition5
        2.588665         6.173738        154.715100      132.381124       49.636984
     waterfront1            view1             view2           view3           view4
        1.539558         1.018438          1.031652        1.043640        1.561563
          grade4           grade5            grade6          grade7          grade8
       16.019855       116.042460        965.614884     2616.311315     2163.351939
          grade9          grade10           grade11         grade12
     1161.591750       502.985115        151.132671       27.190543
```

***Figure 21(a):*** *VIF Output from Initial Linear Regression Model*

After multicollinearity was addressed, backward elimination was performed to refine the model. During this process, the variable log_bathrooms was dropped, as it was not contributing significantly to the model. To compare the models before and after backward elimination, an ANOVA test was conducted. The p-value from the ANOVA test was greater than 0.05, so the log_bathrooms predictor didn't contribute significantly to the model. The output of the regression results is shown in the figure below. Therefore, the model from backward elimination was chosen.



```
Call:
lm(formula = log_price ~ log_bedrooms + log_sqft_living + floors +
    lat + log_sqft_living15 + condition + waterfront + view +
    grade, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.26380 -0.16439 -0.00792  0.15671  1.13117

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -62.185501   0.911181 -68.247  < 2e-16 ***
log_bedrooms      -0.058191   0.012064  -4.824 1.43e-06 ***
log_sqft_living    0.388402   0.012177  31.896  < 2e-16 ***
floors             0.034438   0.005596   6.154 7.82e-10 ***
lat                1.496532   0.018248  82.009  < 2e-16 ***
log_sqft_living15  0.172477   0.012426  13.880  < 2e-16 ***
condition2         0.020260   0.070528   0.287  0.77392
condition3         0.114450   0.064822   1.766  0.07749 .
condition4         0.206812   0.064878   3.188  0.00144 **
condition5         0.307684   0.065305   4.711 2.49e-06 ***
waterfront1        0.414731   0.041039  10.106  < 2e-16 ***
view1              0.195339   0.020457   9.549  < 2e-16 ***
view2              0.156446   0.012669  12.348  < 2e-16 ***
view3              0.243712   0.017731  13.745  < 2e-16 ***
view4              0.283653   0.029076   9.756  < 2e-16 ***
grade4            -0.787162   0.266396  -2.955  0.00314 **
grade5            -0.632061   0.259053  -2.440  0.01471 *
grade6            -0.535148   0.258175  -2.073  0.03821 *
grade7            -0.441540   0.258212  -1.710  0.08730 .
grade8            -0.304298   0.258326  -1.178  0.23884
grade9            -0.137132   0.258522  -0.530  0.59581
grade10            0.020133   0.258795   0.078  0.93799
grade11            0.107503   0.259542   0.414  0.67873
grade12            0.241976   0.263734   0.918  0.35890
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2576 on 10665 degrees of freedom
Multiple R-squared:  0.7314,    Adjusted R-squared:  0.7308
F-statistic:  1262 on 23 and 10665 DF,  p-value: < 2.2e-16
```

***Figure 21(b):*** *Regression output from backward selection model*

The logistic regression equation is:

$$\begin{aligned} log(price) &= -62.1 - 0.0582log(bedrooms) + 0.3884log(sqft\_living) + 0.0344floors \\ &+ 1.4965latitude + 0.1725log(sqft\_living15) + 0.3077condition5 \\ &+ 0.4147waterfront + 0.2837view4 + 0.2420grade12 \end{aligned}$$

The regression equation of the model shows how these variables impact the outcome of price. Significant predictors like log_sqft_living and lat suggest that features like size and location strongly influence the outcome. Conversely, predictors like conditions don't really have much of an impact on log(price). The intercept, -62.1, represents the baseline log(price) when all predictors are at zero, though this value is not practically meaningful since it's impossible for a home to have zero bedrooms, square footage, or any of the other predictors.

From the equation one can determine that:

- The log(bedrooms) coefficient indicates that a 1% increase in the number of bedrooms is associated with a 0.0582 decrease in log(price), which corresponds to about a 5.82% decrease in sale price, holding other predictors constant.
- The log(sqft_living) coefficient shows that a 1% increase in square footage results in a 0.3884 increase in log(price), or approximately a 38.84% increase in price, holding other predictors constant.
- The floors coefficient suggests that each additional floor is associated with a 0.0344 increase in log(price), which corresponds to about a 3.44% increase in price, holding other predictors constant.
- The latitude coefficient indicates that each one-unit increase in latitude leads to a 1.4965 increase in log(price), which corresponds to a 349% increase in price, holding other predictors constant.
- The log(sqft_living15) coefficient shows that a 1% increase in the square footage of neighboring homes is associated with a 0.1725 increase in log(price), or about a 17.25% increase in price, holding other predictors constant.
- Homes in the best condition (condition 5) are associated with a 0.3077 increase in log(price), translating to a 30.77% higher sale price compared to homes in condition 1, holding other predictors constant.
- Homes with a waterfront view have a 0.4147 increase in log(price), or approximately a 41.47% higher price compared to homes without, holding other predictors constant.
- Homes with better views (view4) are associated with a 0.2837 increase in log(price), which corresponds to a 28.37% higher sale price, holding other predictors constant.
- Homes in grade 12 have a 0.2420 increase in log(price), translating to a 24.20% higher sale price compared to homes in grade 1, holding other predictors constant.

These results demonstrate that factors such as location, condition, and features like views and waterfront access are crucial in determining home prices in King County during 2014–2015.

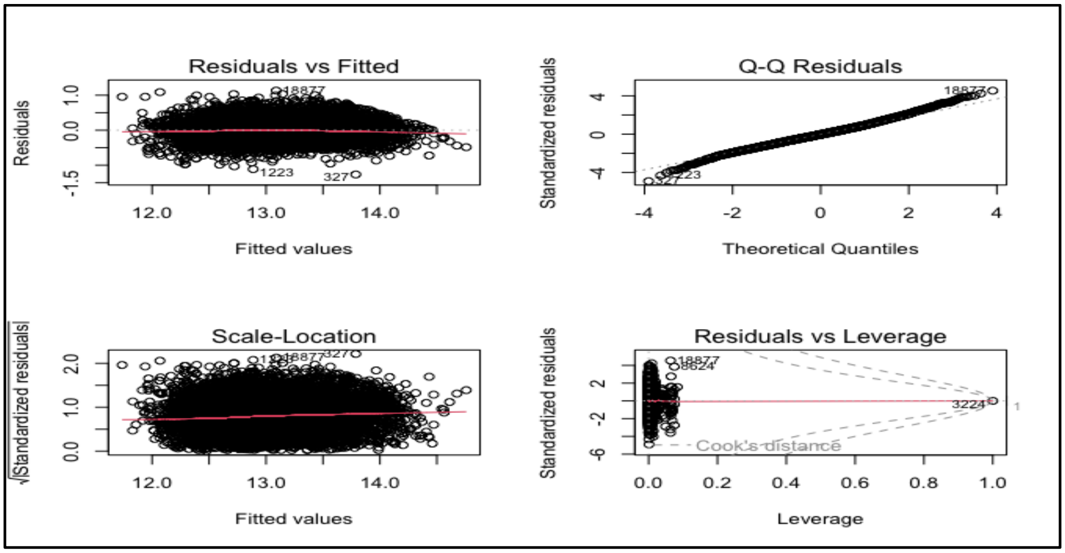


***Figure 21(c):*** *Residual plot of final linear regression model*

After the model was finalized, potential outliers, high leverage, and influential observations were examined. The residual plot for the final linear regression model was created (Figure 21c) for final confirmation of regression assumptions and to examine distinct values based on Cook's distance. No distinct values were shown based on the residual plot. Then, the dataset was examined using Cook's distance threshold, but none of the observations had extreme Cook's distance values so there were no highly influential points present in the dataset. This is most likely due to the fact that in a larger dataset, the influence of an observation on the model is not very large.

To examine potential high leverage points, the top 100 observations with the highest leverage values, accounting for approximately 1% of the dataset, were focused on. Upon reviewing these observations, an observation stood out with a leverage value of 1.000. This observation was possibly a unique property that did not follow the general trend of larger homes in more populated areas, with fewer bedrooms and bathrooms. It could be a home with a large lot compensating for the smaller living area, which sets it apart from the other data points. Given the clear deviation of the high leverage observation of 1.000, this was removed. The other high leverage values were within a reasonable range for the data columns. These observations were possibly larger homes with more square footage, so they were not removed from the dataset. They could be used for further discussion of homes that might be larger in size and contain more square footage, so they weren't removed from the dataset.

To investigate potential outliers, the studentized residuals were first examined, flagging those with absolute values greater than 3 for closer inspection. Residuals greater than 4 were then given further attention. After reviewing, it was determined that the outliers, while noticeable, should remain in the dataset. High price homes could be outliers because they might be in upscale neighborhoods. On the other hand, lower-priced homes might be undervalued. These factors that drive these outliers provide meaningful insight into the impact of houses on price. These are important values that can be looked at,  providing interesting case studies on what might be driving higher or lower priced houses.

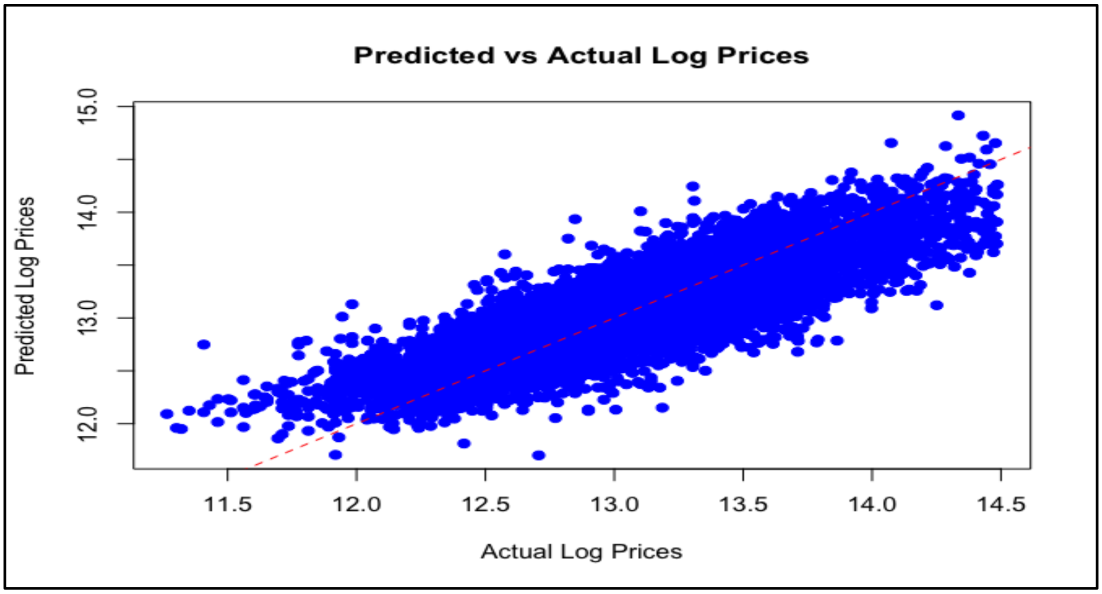


***Figure 22:*** *Predicted versus Actual log(prices)*

After the model was fitted, it was tested on the test dataset to assess its performance. Key metrics were calculated, including the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. The MAE was found to be 0.203, meaning that, on average, the model's predictions differed from the actual values by about 0.203 units. The RMSE, which was 0.262, suggested that the model's errors were somewhat variable, with larger discrepancies being penalized more. The R-squared value of 0.726 indicated that the model explained about 72.6% of the variance in the log-transformed house prices, showing a good fit. These metrics showed  the model's predictive accuracy and how well it generalizes to new data. A scatterplot is shown to better visualize the predictive performance of the final model (Figure 22).

Examining the predictors in the model, its structure was unsurprising overall, but a few aspects stood out. Square footage, condition, and the number of floors are common predictors for house prices, so their significant roles in the model were expected. What was particularly noteworthy, however, was the necessity of log-transforming certain variables, such as price, square footage, and bedrooms. The log transformations highlighted how these variables do not increase in a linear fashion—larger houses, for instance, do not necessarily cost exponentially more. This aligns with what the real estate industry would look like. The grade of the house, which represents its quality, was also key in predicting price. This made sense too, since newer, high-quality homes are often priced much higher than older, lower-quality ones.

In conclusion, the model demonstrates strong accuracy in predicting home prices in King County, Washington. With an R-squared value of 0.726, the model explains approximately 72.6% of the variability in house prices, indicating it captures most of the key factors influencing home values. The low Mean Absolute Error shows that the predictions are close to actual values on average, while the low Root Mean Squared Error suggests that any larger errors are not significant enough to diminish its overall effectiveness. While there is still some unexplained variability, the model's strong performance suggests it is a useful tool for accurately estimating home prices in the region.

# 6. Good Quality Homes and Influential Factors

Another goal of this analysis is to explore which characteristics are most associated with good quality homes in King County, Washington, where "good quality" is defined as homes with a condition score greater than 3 and a grade value greater than 7. Through a series of data visualizations, it was examined how features such as price, size, layout, construction year, location, and aesthetic elements like views and waterfront access differ between "Good" and "Other" homes.

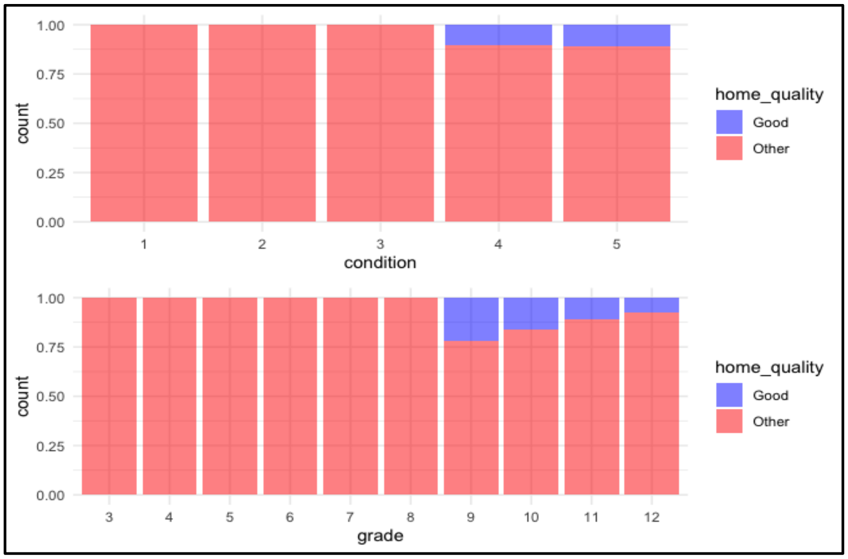


***Figure 23:*** *Home Quality and Condition*

Based on the dataset and the criteria for a good-quality home—defined as having a condition score greater than 3 and a grade value above 7—there is a noticeable imbalance between high-quality and average or lower-quality homes in the region. The left chart confirms that only homes with a condition score above 3 meet the “Good” quality threshold, supporting the definition. The right chart further illustrates that these homes represent only a small portion of the dataset, with the majority falling under the “Other” category. This highlights a significant class imbalance in home quality across King County, Seattle.

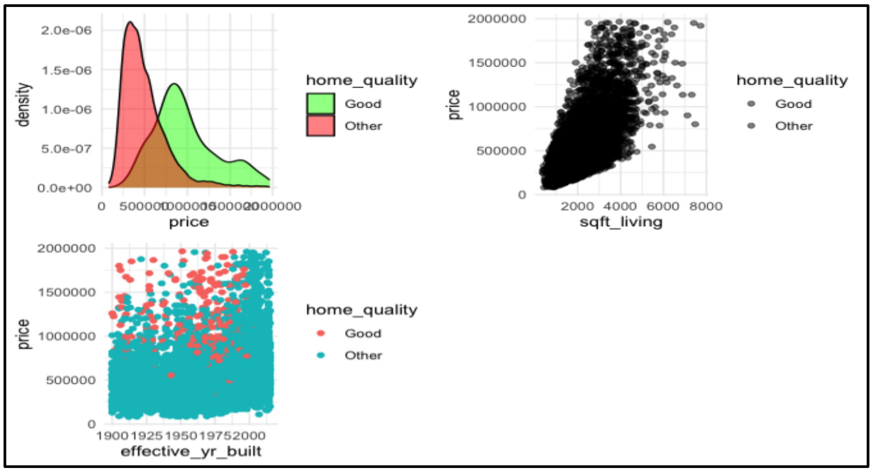


***Figure 24:*** *Price, Effective Year Built, and Living Square Footage impacts on Home Quality*

In these visualizations, the price trends and structural characteristics of good-quality homes are presented. These visualizations demonstrate that good-quality homes are generally associated with higher property values. The price distribution shows that good homes tend to occupy the higher end of the market. Additionally, there is a clear positive relationship between living space and price, with good homes often having both larger square footage and higher sale prices. Lastly, while good homes exist across a wide range of construction years, they tend to be concentrated among higher-priced homes built more recently, particularly after 1950.

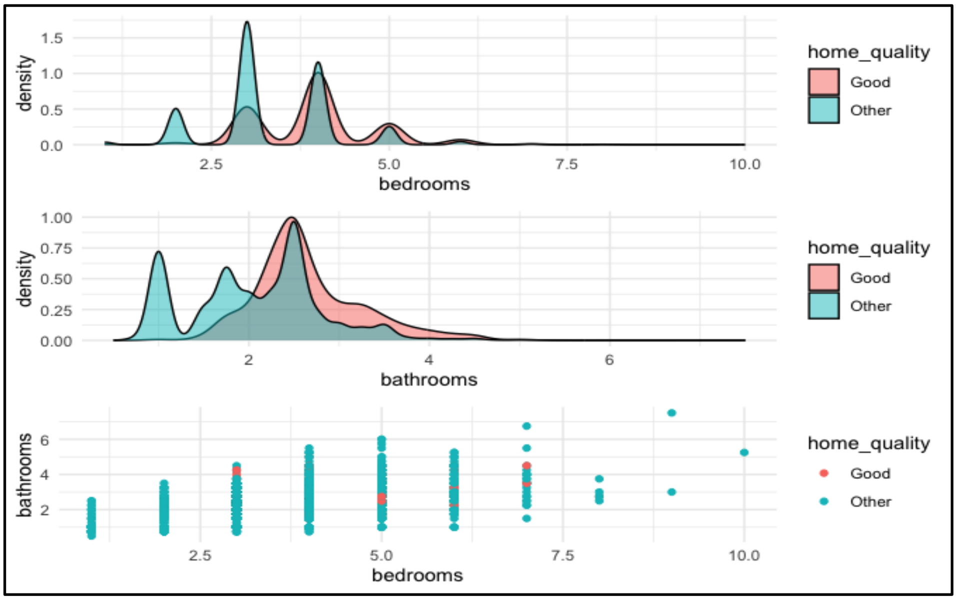


***Figure 25:*** *Home Quality against the Number of Bedrooms and Bathrooms*

These density plots show that good-quality homes generally have moderate bathrooms and bedrooms compared to other homes. The number of bathrooms in good homes tends to peak between 2.5 to 3.5, while other homes are more concentrated around 1.5 to 2.5. Good homes are more evenly spread across 3 to 5 bedrooms, whereas other homes show a sharp peak around 2 to 3 bedrooms. Together, these plots show that slightly larger homes are more likely to be classified as "Good", rather than those with extreme size differences.

The joint distribution of bedrooms and bathrooms in homes is also presented, colored by quality classification. While most homes, both good and otherwise, exhibit a positive relationship (more bedrooms generally come with more bathrooms), high-quality homes are more concentrated around combinations of 4–5 bedrooms and 2.5–4.5 bathrooms. However, they are relatively sparse across the full distribution, indicating that even among large homes, not all are classified as "Good." This demonstrates that while bed/bath layout may contribute to quality, it is not the sole determinant.

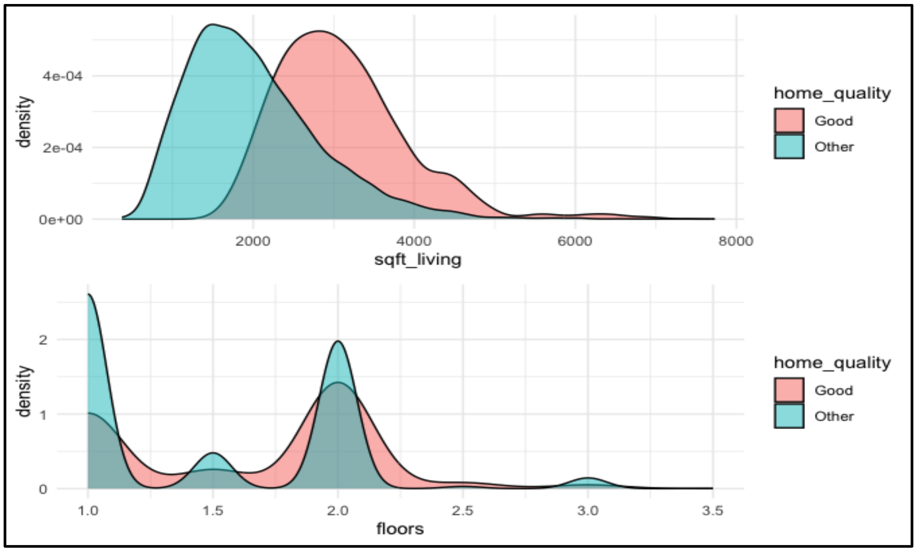


***Figure 26:*** *Distribution of Living Square Footage and Number of Floors coded by Home Quality*

The density plots presented above demonstrate the living area distribution by home quality. As shown, this density plot shows that good quality homes typically have larger living spaces, most commonly between 2,000 and 4,000 square feet, whereas other homes tend to peak below 2,000 sqft. The rightward shift of the "Good" home distribution indicates that larger living area is a key characteristic associated with higher home quality, supporting the idea that space is a significant factor in quality classification.

As shown in the floor count distribution by home quality, it shows that most homes, regardless of quality, have either one or two floors, but good quality homes are more evenly distributed between 1.5 to 2 floors, with a slightly higher density at 2 floors. In contrast, lower-quality homes show a sharper peak at single-floor structures.

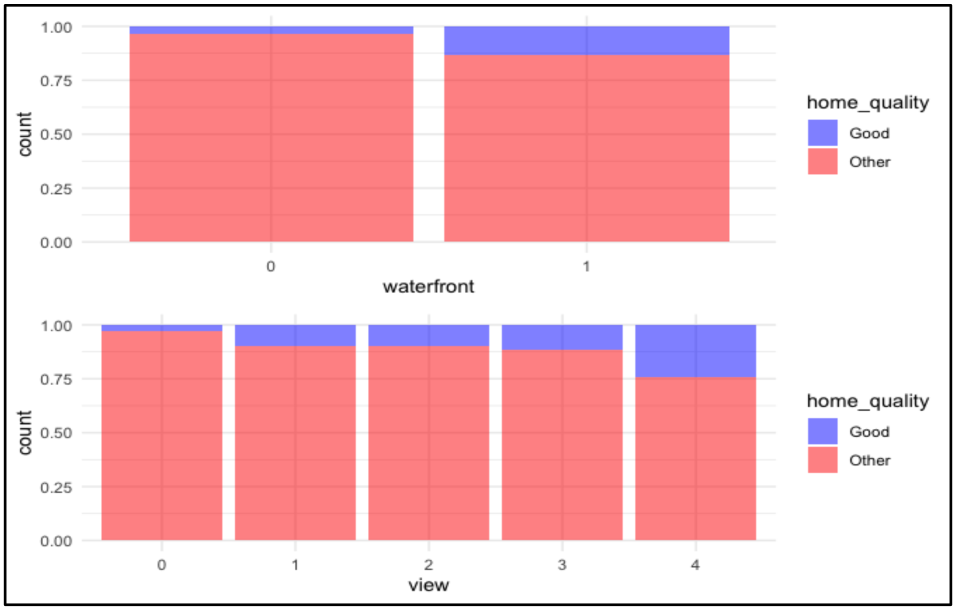


***Figure 27:*** *Proportion of Good and Other Quality homes by View and Waterfront Presence*

These plots demonstrate the aesthetic features and their relationships to home quality. These plots illustrate the relationship between aesthetic features, such as view ratings and waterfront access, and home quality classification. Homes with higher view scores (especially levels 3 and 4) exhibit a more significant proportion of "Good" classifications, suggesting that scenic views positively contribute to perceived quality. Similarly, waterfront homes are significantly more likely to be classified as "Good" compared to non-waterfront properties, suggesting that proximity to water is a strong indicator of premium home quality. Overall, these aesthetic features play an important role in distinguishing higher-quality homes within the data set.

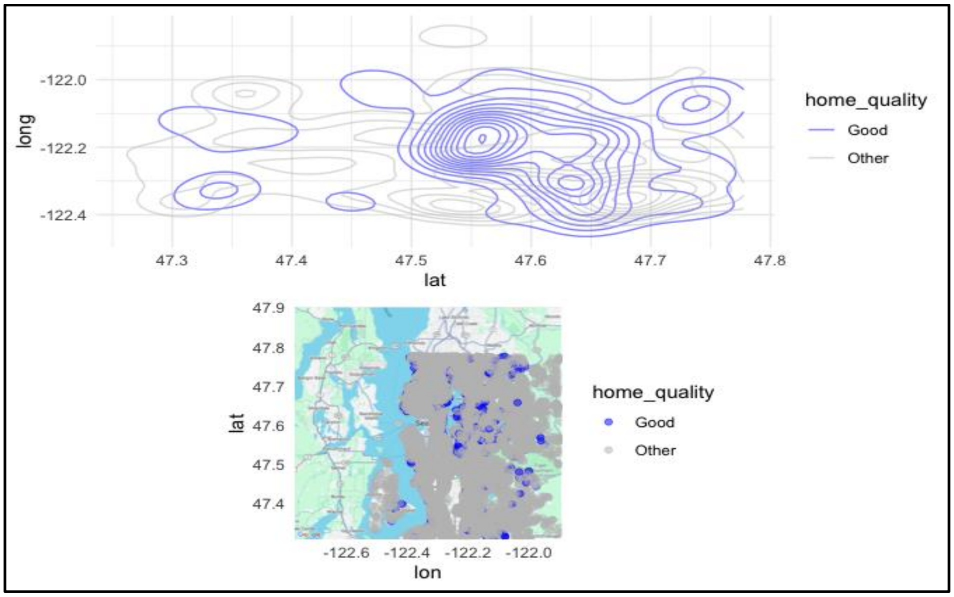


***Figure 28:*** *Home quality shown by geographic distribution*

The map shown above highlights the geographic concentration of good quality homes in King County, Seattle. The contour density map (left) shows that good homes cluster most densely between latitudes 47.55–47.7 and longitudes -122.3 to -122.1, aligning with northeast Seattle and adjacent neighborhoods. The base map with plotted points (right) provides a more intuitive geographic reference, confirming that good homes are predominantly located in northern and eastern residential areas. Together, these maps suggest that location plays a significant role in determining home quality, with better-rated homes typically situated in more desirable and higher-value areas of the region. Both sets of visuals reinforce the conclusion that good quality homes tend to cluster in the northeast and east-central portions of King County, where location likely correlates with higher grade, condition, and aesthetic appeal.

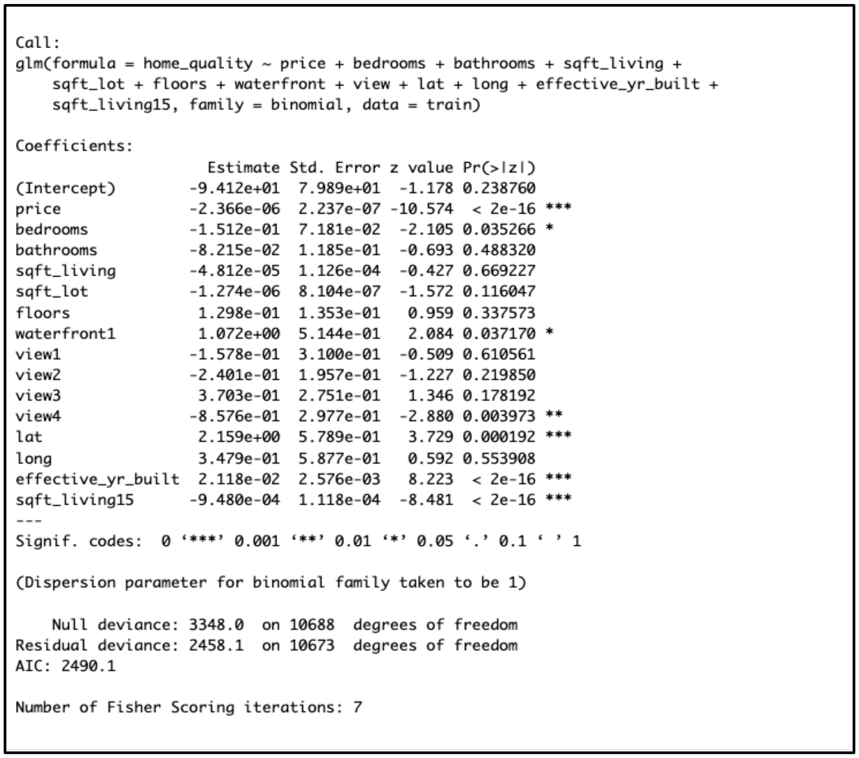
The overall findings from these visualizations indicate that a combination of structural features, aesthetic appeal, and geographic location significantly influences home quality in King County. Good-quality homes are not just newer or more expensive; they tend to demonstrate distinct traits that increase their value. Structurally, these homes typically offer larger living areas, more bedrooms and bathrooms, and are more likely to have two floors, being both spacious and functional. Aesthetically, good homes are often those with higher view ratings or waterfront access, highlighting the role of scenic and environmental desirability in determining quality. While newer construction alone does not guarantee high quality, homes built or renovated between 1965 and 2000 are more often of good quality, suggesting enduring value in certain building eras. Altogether, the data illustrate that home quality is multifaceted, with no single factor being sufficient. Geography also plays a significant and consistent role in distinguishing good-quality homes. The latitude and longitude density plots, as well as spatial contour and base map visualizations, all point to a clear geographic pattern: good homes are disproportionately concentrated in the northern and eastern regions of King County. These correspond to areas such as northeast Seattle, Bellevue, Kirkland, and Redmond, which are neighborhoods characterized by higher income levels, better schools, and other desirable proximities. In contrast, "Other" homes are more evenly distributed across the region, including less centralized areas.

# 7. Logistic Regression for Good Quality Homes

Using logistic regression, a binary classification model was developed to predict whether a home in King County is of good quality. It is defined as having a condition score greater than 3 and a grade value greater than 7. This binary classification served as the response variable. The initial model included a wide range of predictors, such as price, number of bedrooms, number of bathrooms, square footage (sqft_living), lot size (sqft_lot), number of floors, waterfront presence, view rating, latitude, longitude, effective year built, and neighborhood size (sqft_living15). These were selected based on their relevance to housing market value and perceived quality, as well as the visualizations created in the prior section. The goal of the model was to assess which of these variables contributes most meaningfully to the likelihood that a home would be classified as "Good," considering factors that include structural, geographic, and aesthetic aspects, as well as higher home quality.

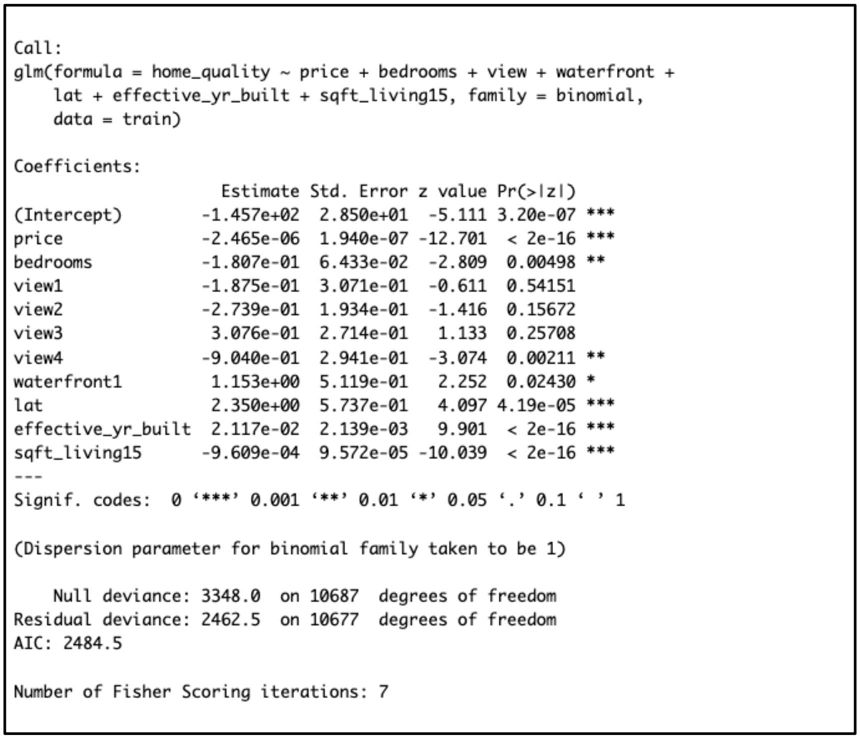
After fitting the initial model, it was evaluated for both the significance of predictors' p-values and multicollinearity VIFs. The summary output (Figure 29a) revealed several variables, including sqft_lot, bathrooms, floors, sqft_living, and long. Those were not significant at thresholds of p value > 0.05. VIF analysis showed high multicollinearity, especially among structural and location-based variables like sqft_living, lat, long, and sqft_living15.

Therefore, to simplify the model and address multicollinearity, variables such as yr_built, which was already incorporated into the custom variable effective_yr_built, were removed, along with other weak predictors. This refinement resulted in a reduced model with 7 key variables. This includes price, number of bedrooms, waterfront, view, latitude, effective year built, and neighborhood size (Figure 29b).



```
Call:
glm(formula = home_quality ~ price + bedrooms + bathrooms + sqft_living +
    sqft_lot + floors + waterfront + view + lat + long + effective_yr_built +
    sqft_living15, family = binomial, data = train)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -9.412e+01  7.989e+01  -1.178 0.238760
price             -2.366e-06  2.237e-07 -10.574  < 2e-16 ***
bedrooms          -1.512e-01  7.181e-02  -2.105 0.035266 *
bathrooms         -8.215e-02  1.185e-01  -0.693 0.488320
sqft_living       -4.812e-05  1.126e-04  -0.427 0.669227
sqft_lot          -1.274e-06  8.104e-07  -1.572 0.116047
floors             1.298e-01  1.353e-01   0.959 0.337573
waterfront1        1.072e+00  5.144e-01   2.084 0.037170 *
view1             -1.578e-01  3.100e-01  -0.509 0.610561
view2             -2.401e-01  1.957e-01  -1.227 0.219850
view3              3.703e-01  2.751e-01   1.346 0.178192
view4             -8.576e-01  2.977e-01  -2.880 0.003973 **
lat                2.159e+00  5.789e-01   3.729 0.000192 ***
long               3.479e-01  5.877e-01   0.592 0.553908
effective_yr_built  2.118e-02  2.576e-03   8.223  < 2e-16 ***
sqft_living15     -9.480e-04  1.118e-04  -8.481  < 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: 3348.0  on 10688  degrees of freedom
Residual deviance: 2458.1  on 10673  degrees of freedom
AIC: 2490.1

Number of Fisher Scoring iterations: 7
```

***Figure 29(a):*** *Initial Logistic Regression Model*



```
Call:
glm(formula = home_quality ~ price + bedrooms + view + waterfront +
    lat + effective_yr_built + sqft_living15, family = binomial,
    data = train)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.457e+02  2.850e+01  -5.111 3.20e-07 ***
price             -2.465e-06  1.940e-07 -12.701  < 2e-16 ***
bedrooms          -1.807e-01  6.433e-02  -2.809  0.00498 **
view1             -1.875e-01  3.071e-01  -0.611  0.54151
view2             -2.739e-01  1.934e-01  -1.416  0.15672
view3              3.076e-01  2.714e-01   1.133  0.25708
view4             -9.040e-01  2.941e-01  -3.074  0.00211 **
waterfront1        1.153e+00  5.119e-01   2.252  0.02430 *
lat                2.350e+00  5.737e-01   4.097 4.19e-05 ***
effective_yr_built  2.117e-02  2.139e-03   9.901  < 2e-16 ***
sqft_living15     -9.609e-04  9.572e-05 -10.039  < 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: 3348.0  on 10687  degrees of freedom
Residual deviance: 2462.5  on 10677  degrees of freedom
AIC: 2484.5

Number of Fisher Scoring iterations: 7
```

***Figure 29(b):*** *Reduced Logistic Regression Model*

The likelihood ratio test (LRT) was performed. The test statistic comparing the full and reduced models was 3.58, with a p-value of 0.733, which is far above the 0.05 significance level. This means that the data supports the reduced model that involved removing 5 predictors when attempting to classify a home as good quality or other with a smaller set of predictors.

The reduced logistic regression model retained strong predictors and demonstrated between "Good" and "Other" homes. Highly significant predictors ($p < 0.001$) included price, lat, effective_yr_built, and sqft_living15. Moderately significant predictors were identified as bedrooms (negative coefficient), waterfront, and view4.

These variables align with earlier visualizations that showed price, location, view, and age are key determinants of home quality. Although not all view levels were individually significant, view4 (very high view rating) was significant and negatively associated with good home classification. This means that there may be other anomalies that may not always correlate with high condition and grade scores.

The logistic regression equation is $logit(home\_quality) = -145.7 - 2.465e^{-6}price - 0.1807bedrooms - 0.1875I_1 - 0.2739I_2 + 0.3076I_3 - 0.9040I_4 + 1.153I_5 + 2.350lat + 0.02117effective\_yr\_built - 0.0009609sqft\_living15$, where $I_1 = 1$ when view equals 1 and 0 if other, $I_2 = 1$ when view equals 2 and 0 if other, $I_3 = 1$ when view equals 3 and 0 if other, $I_4 = 1$ when view equals 4 and 0 if other, $I_5 = 1$ when property is at a waterfront and 0 if not.

From this equation, one can determine the following:

- The odds of a property being good quality are multiplied by 0.99 for every unit increase of price when controlling for bedrooms, view, waterfront, latitude, effective year built, and square footage of nearby properties
- The odds of a property being good quality are multiplied by 0.83 for every additional unit increase of bedroom while controlling for price, view, waterfront, latitude, effective year built, and square footage of nearby properties
- The odds of a property being good quality are 0.83 times the odds for properties with views not categorized as a 1, while controlling for price, bedrooms, waterfront, latitude, effective year built, and square footage of nearby properties
- The odds of a property being good quality are 0.76 times the odds for properties with views not categorized as a 2, while controlling for price, bedrooms, waterfront, latitude, effective year built, and square footage of nearby properties
- The odds of a property being good quality are 1.4 times the odds for properties with views not categorized as a 3, while controlling for price, bedrooms, waterfront, latitude, effective year built, and square footage of nearby properties
- The odds of a property being good quality are 0.40 times the odds for properties with views not categorized as a 4, while controlling for price, bedrooms, waterfront, latitude, effective year built, and square footage of nearby properties
- The odds of a property being good quality are 3.168 times the odds for non-waterfront properties, when controlling for price, bedrooms, view, latitude, effective year built and square footage of nearby properties
- The odds of a property being good quality are multiplied by 10.5 for every additional unit increase of latitude when controlling for price, bedrooms, view, waterfront, effective year built, and square footage of nearby properties

To evaluate how well the reduced logistic regression model generalizes to unseen data, its predictive performance was assessed on the test set. Predictions were generated using the predict() function with type = "response", and a 0.5 cutoff was applied to classify homes as "Good" or "Other." From the confusion matrix, the model correctly identifies 356 out of 401 good homes (high sensitivity), while making only 50 false positives out of over 10,200 "Other" homes (high specificity).

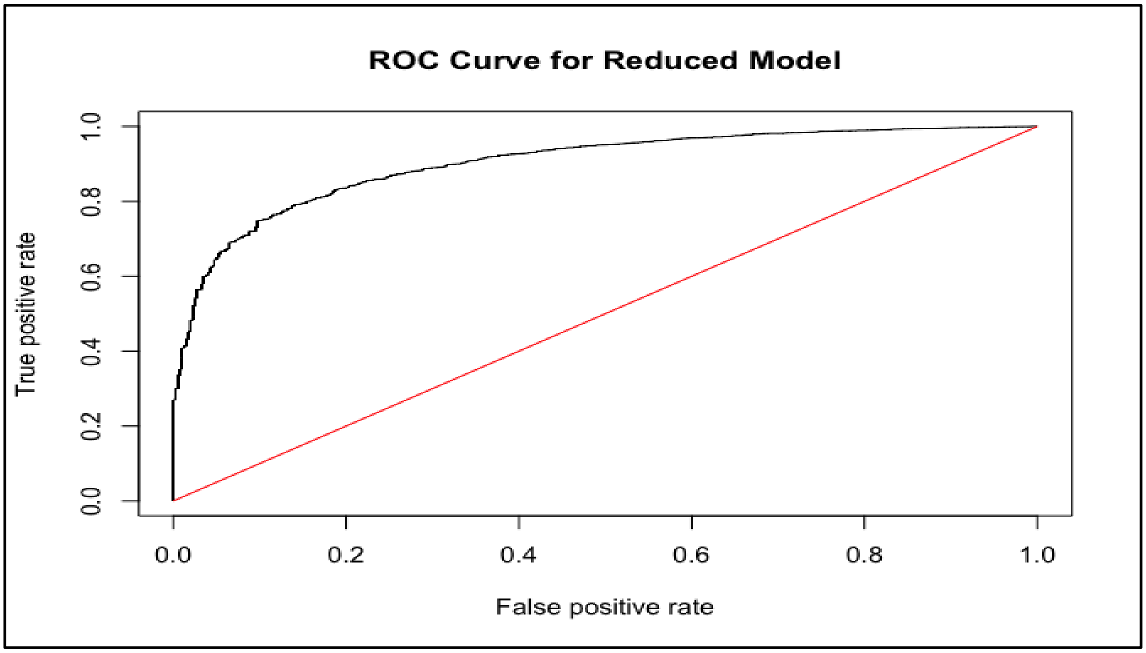


***Figure 29(c):*** *ROC curve*

To evaluate the predictive performance of the reduced logistic regression model, the model was applied to the test dataset and visualized its classification accuracy using a ROC curve (Figure 29c). As shown in the above figure, the ROC curve rises steeply toward the top-left corner, indicating a strong ability to distinguish between "Good" and "Other" homes.

The AUC was calculated to be 0.902, which falls in the range of excellent classification performance. This means that the model can correctly rank a randomly chosen "Good" home above a randomly chosen "Other" home approximately 90.2% of the time. The curve's shape demonstrates a low false positive rate and high true positive rate, affirming that the model makes accurate predictions across varying thresholds. Overall, this confirms that the reduced model strikes a strong balance between sensitivity and specificity, making it reliable for real-world applications.

The model provides several insightful takeaways regarding the factors that influence home quality. While price has a negative coefficient, this reflects the logistic regression structure, indicating that as price increases, the odds of a home being classified as "Other" decrease. This also means that higher-priced homes are more likely to be of good quality. Latitude shows a strong positive effect, suggesting that homes located further north in King County are more likely to be classified as "Good," which aligns with earlier geographic visualizations. The presence of waterfront access is also positively associated with good quality homes, highlighting the importance of aesthetic and locational advantages.

Additionally, the effective year built is a strong positive predictor, implying that newer or recently renovated homes are more likely to meet good quality standards. The number of bedrooms has a slight negative association with the quality classification, indicating that higher bedroom counts do not necessarily reflect better condition. The negative coefficient for view4 may show potential signs of outliers or unusual cases where very high view scores are not aligned with structural quality. These findings show that the quality of homes consists of a combination of structural, geographic, and aesthetic factors, rather than being influenced by any single variable.

Through the application of logistic regression, a model was developed to help predict the likelihood of a home being classified as "Good" using accessible property data. The reduced model includes only the most relevant and statistically supported predictors, making it both interpretable and effective. It shows that good quality homes in King County are typically more expensive, located further north, newer or recently renovated, and often have desirable features like waterfront access. The model captures the multifaceted nature of home quality, highlighting the value of combining structural, aesthetic, temporal, and spatial variables. Therefore, the model can be useful for predictive insights for buyers and realtors, demonstrating how high-quality homes are predicated.

# 8. Citations

- harlfoxem. (2016, May 3). *Discussion on House Sales in King County, USA* [Online forum post]. Kaggle. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction/discussion/207885