Behind the Price Tag: King County Housing Market Analysis

STAT 6021 Project 2

Alysa Pugmire, Angleo Orciuoli, Khalil Goddard, Maryam Ali
GROUP 11

Analysis of Pricing for Homes in King County Washington

This report explores the various factors that contribute to home pricing in King County Washington through data visualization and modeling. Specifically, it builds two models, one that explains how the prices of homes in King County are affected by various variables, and one that predicts if a home in King County is of good quality.

Section 1: High Level Findings

This study analyzed over 21,000 homes in the King County region of Washington with the purpose of understanding what factors affect the pricing of homes as well as what factors contribute to whether or not a home in King County is of good quality.

First, by building linear regression models to predict home prices, we gained insights into which factors influence pricing and the extent of their impact. Our final linear regression model explains approximately 76% of the variation in house prices, indicating a strong understanding of the relationship between factors and price. For example, we determined each additional square foot of living space increases the price of a home by about $155, while waterfront properties are worth approximately $584,913 more than non-waterfront ones. Homes located farther from downtown decrease in value by roughly $457,495 per unit of distance. A one-point increase in view rating adds about $45,372, and each point increase in construction grade raises price by around $96,291. Suburban homes are priced $41,909 higher than urban ones, while rural homes are $59,865 lower. Notably, more bedrooms slightly reduce price by about $25,592 per bedroom, and each additional year of age reduces a home's value by approximately $1,783.

While these findings may seem intuitive—larger homes, waterfront properties, and proximity to downtown typically command higher prices—quantifying the exact impact of each factor provides meaningful insights that can inform real-world decisions. Specifically, home buyers and agents can use these insights to better assess a property's value based on features like square footage, location, and construction quality. For example, understanding that waterfront homes command a premium of nearly $585,000 or that proximity to downtown significantly boosts value can guide more informed pricing and negotiation decisions. Similarly, developers and planners can use the identified value drivers—such as view quality and grade—to prioritize features in future construction projects that align with market demand and maximize return on investment.

Through the creation and analyses of a logistic regression model, we found that factors like where the home is located relative to the city center of King County, when it was built and when it was most recently renovated, greatly impacted the odds of a home being good quality.

For example, contrary to what one would assume, we found that homes that had been recently renovated (since 2005) had lower odds of being deemed good quality than homes that had been renovated a long time ago (before 2005) or homes that were never renovated at all. This could be explained by the fact that homes built and renovated in the 21st century have been focused on functionality and sustainability which has come at a cost to perceived quality related to materials and craftsmanship.

To expand on this idea, we found that the majority of good quality homes in our data set were built around 1975. Again, this confirms the notion that the quality of homes has collectively diminished in the King County region as the years have gone on. Someone who wishes to purchase a home in King County may want to narrow their search to homes that were not recently renovated but were built around 1975.

Another interesting conclusion is that suburban homes had the greatest odds of being deemed good quality when compared to homes located in rural or urban areas of King County. Again, someone wishing to purchase a home in King County may want to look at homes in the suburbs if they prioritize quality over closeness to downtown.

Section 2: Description of the Dataset and Variables

The dataset used in this analysis contains information on residential properties in King County, Washington. It includes data on various features of the homes such as their size, number of bedrooms and bathrooms, and location, as well as sale prices. The dataset comprises 21,613 observations, with 26 variables, including both original attributes and newly created variables that provide additional insights for the analysis. This dataset allows for a detailed exploration of factors influencing home pricing in the area.

Original dataset variables and their descriptions:

1. id: Unique identifier for each house listing.
2. date: The date the house was sold
3. price: Sale price of the house.
4. bedrooms: Number of bedrooms.
5. bathrooms: Number of bathrooms
6. floors: Number of floors/stories.
7. sqft_living: Total interior living space (in square feet).
8. sqft_lot: Total area of the land lot.
9. sqft_above: Square footage above ground (excludes basement).
10. sqft_basement: Square footage of the basement.
11. sqft_living15: Living space (sqft) of the 15 nearest neighbors.
12. sqft_lot15: Lot size of the 15 nearest neighbors.
13. waterfront: if the property is on a waterfront (0 = no, 1 = yes).
14. view: Integer rating of the quality of the view on a 0–4 scale.
15. condition: Overall condition of the house on a 1–5 scale.
16. grade: Construction and design quality rating on a 1–13 scale.
17. yr_built: Year the house was originally built.
18. yr_renovated: Year of the last renovation (0 if never renovated).
19. zipcode: Postal code of the property.
20. lat: Latitude coordinates
21. long: Longitude coordinates.

Our created variables and their descriptions:

1. year_sold: The year the house was sold, extracted from the date variable.
2. month_sold: The month the house was sold, extracted from the date variable.
3. region: A categorical variable indicating properties as 'rural', 'suburban', or 'urban' based on their zipcode.
4. renovation_group: A categorical variable indicating renovation status. Values are either never renovated, recently renovated, or renovated long ago.
5. distance_to_downtown: The euclidean distance from each property to the downtown center. The unit is decimal-degrees, calculated using lat and long coordinates.
6. good_quality: a binary variable (0 = no, 1 = yes) derived from the condition and grade variables. Classified as 1 is the condition rating is greater than 3 and the grade rating is higher than 7.

Section 3: Data Cleaning and Variable Transformation

This section discusses the steps taken to identify and fix any data entry errors and transform problematic variables in the dataset. We began by loading the King County housing dataset and taking a first look at the data. With 21,613 observations and no missing values, we were able to focus on ensuring data accuracy rather than addressing issues of imputation.

Data Validation

We began by checking for clearly incorrect values in the dataset. These included homes listed with zero or extremely high bedroom or bathroom counts (e.g., 33 bedrooms), zero square footage, or implausible price or year-built values.

Seventeen properties were flagged for investigation. Instead of removing these rows outright, we verified the entries using the King County Parcel Viewer. Based on these external records:

- Thirteen rows were corrected using verified bedroom and bathroom counts.
- One studio-style unit with 0 bedrooms and 1.5 bathrooms was verified and kept unchanged.
- Three properties could not be verified and were removed from the dataset to avoid potential skewness.

We also checked for duplicate property IDs and found 177 cases. These likely represent homes that were sold more than once during the data collection period. While these duplicates are not data entry errors per se, we note that they may introduce dependence between observations. For now, we chose to retain them, but we remain aware that their presence may influence price variance.

We then checked for additional issues:

- No homes had a square footage of zero.
- No homes were listed with a price less than or equal to zero.

- No homes were built or renovated after 2015, which is consistent with the dataset timeline.

Variable Transformations:

As we cleaned the dataset, we ran into four variables that could not be used in their original form. They either had too many categories, did not make much sense numerically, or could hurt the models' accuracy if we left them as-is.

- Zipcode → Region (3-Level Categorical)

    1. Original: 70+ distinct zip codes, too many for modeling.

    2. Transformation: Grouped into City, Suburb, and Rural based on King County boundaries.

    3. Why: Reflects real differences in urban, suburban, and rural housing markets, and reduces risk of overfitting from too many dummy variables.

    4. Result: 4,471 City, 7,267 Suburb, 9,875 Rural.

- Year Renovated → Renovation Group (3-Level Categorical)
    1. Original: yr_renovated, with 0 indicating no renovation, and positive values give the year of renovation.

    2. Transformation: Re-coded to renovation_group, a 3 level categorical variable:

        - "Never Renovated" if yr_renovated = 0

        - "Renovated Long Ago" if yr_renovated > 0 and < 2005

        - "Recently Renovated" if yr_renovated >= 2005

        We chose 2005 as a reference point for "recent" because renovations done within the last 10 years tend to retain visible and functional value.

        This is not a strict benchmark, but a practical one. A renovation from 2005 would still likely have a noticeable impact on a home's quality and appeal by the time it was sold in 2014–2015.

    3. Why: Instead of using the raw year, which assumes newer always means better, this 3-level group gives a clearer picture of whether a home was never updated, updated a long time ago, or updated more recently.

    4. Result: Out of 21,613 homes:

        - 20,696 were Never Renovated
        - 594 were Renovated Long Ago
        - 320 were Recently Renovated

- Latitude & Longitude → Distance to Downtown

  1. Original: Raw latitude and longitude.

  2. Transformation: Created distance_to_downtown as Euclidean distance from Pike Place Market (≈ 47.6062° N, 122.3321° W). We chose Pike Place Market because it sits at the heart of downtown Seattle and is the city's best known landmark.

     Note: Around Seattle, 1° ≈ 46.5 miles east-west and 69 miles north-south.

     Why: Reduces two coordinates to one number, making location and how urban or suburban it is easier to interpret.

     The summary statistics show:

     ➔ Most homes in the dataset are relatively close to downtown, with distances ranging from 0.013° to just over 1.02°. The median distance is about 0.18°.
     ➔ This spread covers city center to outer suburbs, letting the model see how locations influence price and condition.
     ➔ Closer homes generally cost more, thanks to shorter commutes and better access to amenities.

We also conducted a preliminary multicollinearity check for the square footage variables to help determine which should be used in the visualizations and analysis. The correlation results showed:

1. sqft_living and sqft_above: r = 0.88

2. sqft_living and sqft_basement: r = 0.43

Because the living square footage is the square footage above + the square footage below, we opted to retain sqft_living and disregard sqft_above and sqft_below.

After transforming the variables above, we were ready to create preliminary visualizations and build our models. We started these processes by splitting the dataset with the set.seed() function into two equal subsets, one for building and training the model and the other for testing it.

<u>Section 4: Visualizing Factors That Influence Home Price</u>

This section provides visualizations that show how different factors are related to the pricing of homes in King County Washington. The visualizations were created only using the training data set.

To begin the analysis, we first looked at the distribution of price, without any specific variables, so we could see how the values were spread across the dataset.

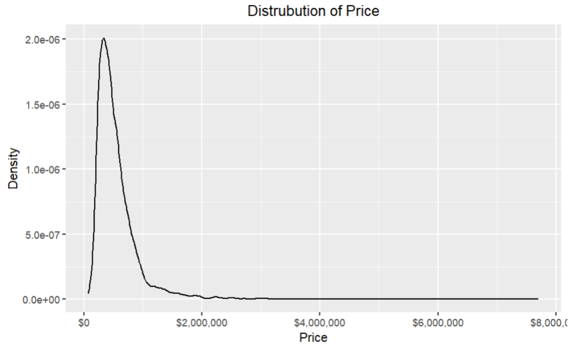


Figure 1. Density plot of price

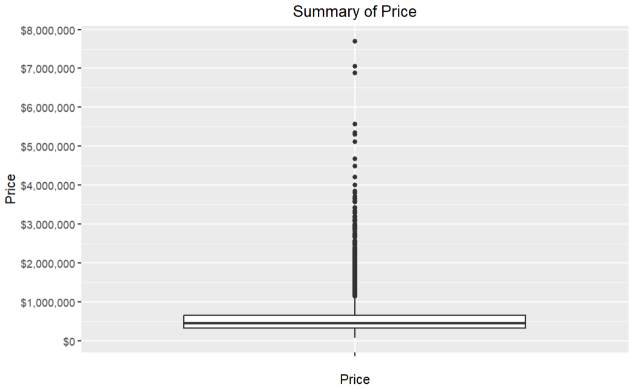
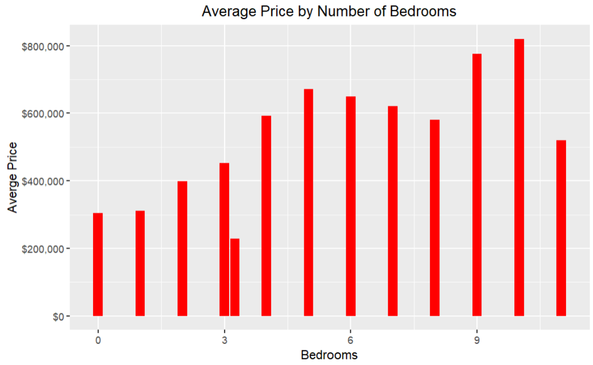
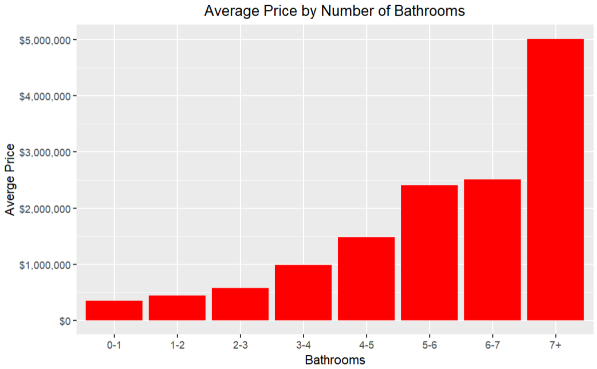


Figure 2. Box plot of price

The density plot shows the prices are extremely right skewed, with a majority of homes priced around $500,000. The box plot also shows that a bulk of the homes are priced between $325,000 and $649,000; however, there are a significant number of outliers.

We then looked at features of the home that would potentially impact the price. We created two bar graphs that show the average price by number of bedrooms and the average price by number of bathrooms, as well as a boxplot to show the price distribution according to number of bedrooms.





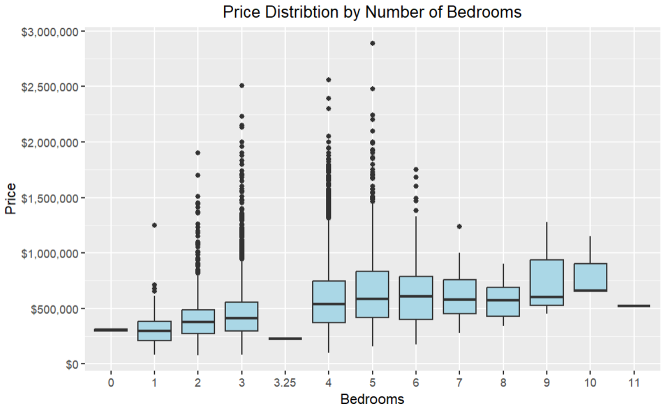


Figure 3a (top left). Bar chart of Average Price by Number of Bedrooms
Figure 3b (top right). Bar chart of Average Price by Number of Bathrooms
Figure 3c (bottom left). Box plot of Price by Number of Bedrooms

As demonstrated in both bar charts, the number of bedrooms and bathrooms a home has positively impacts the price. Homes with more bedrooms and more bathrooms tend to have higher average prices than homes with less bedrooms and less bathrooms. Interestingly, the homes with the most bedrooms do not have the highest average price, which indicates that the number of bedrooms alone does not impact the price. The boxplot of Price by Number of Bedrooms also confirms that the number of bedrooms is related to the price. As the number of bedrooms increases, so does the median price. However, houses with more bedrooms also tend to have more variability in the price. The houses with 3-6 bedrooms also have significantly more outliers than houses with 1-2 bedrooms and 7+ bedrooms.

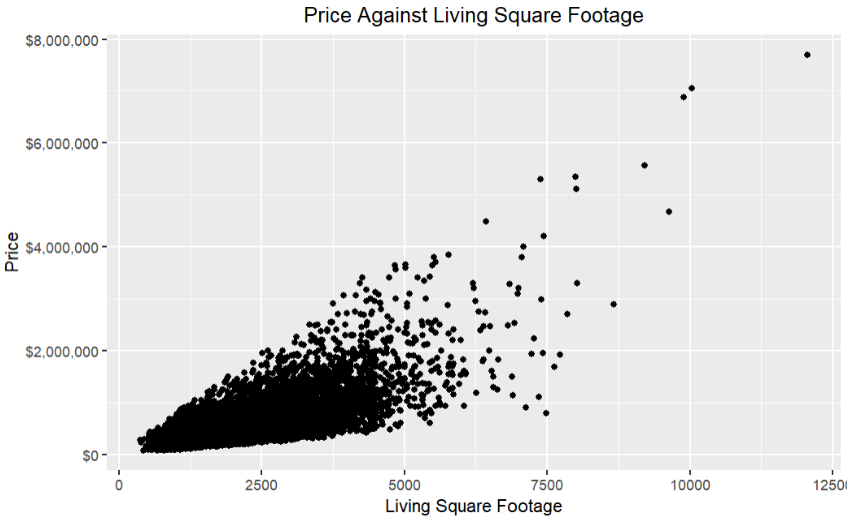


Figure 4. Scatter plot of price against square footage

Next we looked at the impact square footage has on the price using the sqft_living variable. The scatter plot of price against living square footage shows that a majority of the homes are less than 5,000 sqft and less than $2,000,000. The scatter plot also shows that there is a positive relationship between the square footage of the living space and the price. As the square footage increases, so does the price of the home. While there does not appear to be any outliers, there are potentially a few high leverage observations.

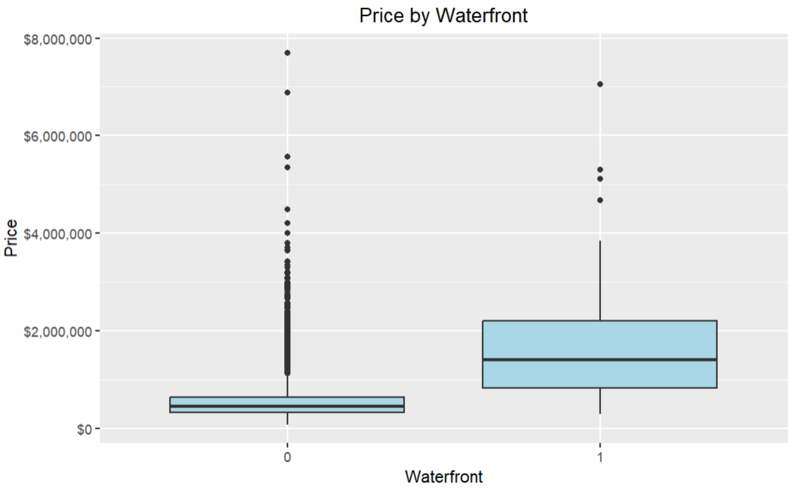


Figure 5. Comparison of price by waterfront

We used side by side boxplots to compare the prices of homes that overlooked the waterfront and homes that do not overlook the waterfront. The median price of homes that overlook the waterfront is significantly higher than the median price of homes that do not, which indicates that waterfront may be a significant predictor of price. There is more variability in the prices of homes that overlook the waterfront, compared to homes that do not. There are significantly more outliers with high prices among homes that do not overlook the waterfront than among homes that do overlook the waterfront, which indicates that overlooking the waterfront is not the sole predictor of price.

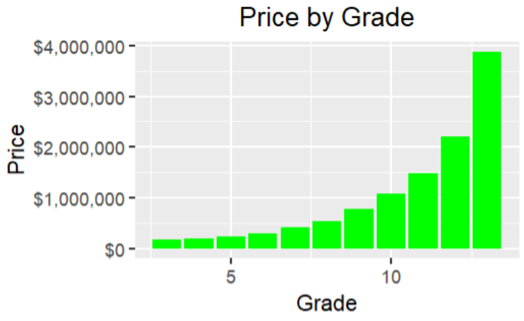


Figure 6a. Price by Grade

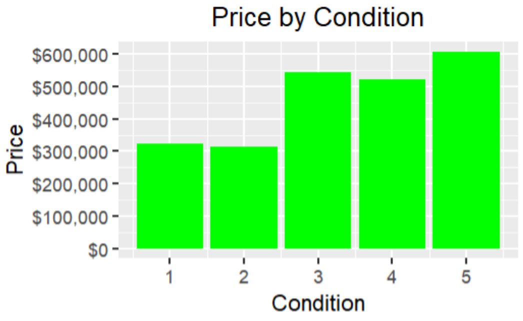
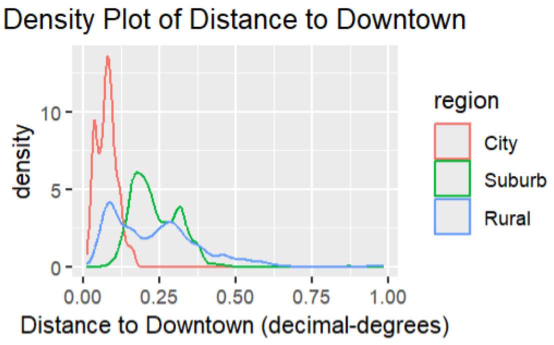


Figure 6b. Price by Condition

Finally, we looked at the relationships between the grade of the home and the average price, as well as, the condition of the home and the average price. Both bar graphs show a positive relationship between grade and condition, and price. As the grade increases, so does the average price. And similarly, as the condition increases, so does the average price. This indicates that grade and condition are likely strong predictors of price. However, homes with a condition rating of 1 have a slightly higher average price than homes with a condition rating of 2. The same is also true for homes with a condition rating of 3 and homes with a condition rating of 4. If condition were the sole predictor of price, we would not see this trend.

After looking at aspects of the home itself that could impact the price, we turned our attention to external factors, such as when the home was built and location, which could also impact the price.

To better understand the location variables, we created density plots to see the distribution of homes according to distance to downtown and year built, color coded by region.



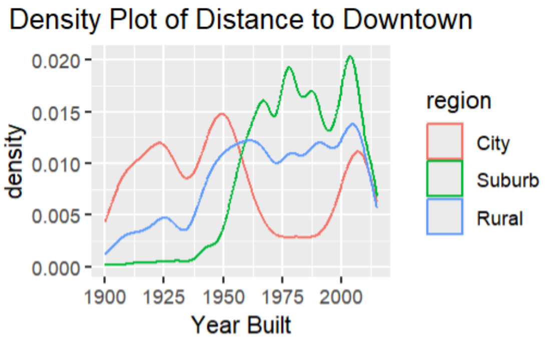


Figure 7a. (left) Distribution of Distance to Downtown by Region

Figure 7b (right). Distribution of Year Built by Region

The density plot of distance to downtown is right skewed for all regions, however a significantly higher proportion of city homes are closer to downtown compared to suburban and rural homes.

Suburban homes have the highest proportion of homes that are moderately close to downtown and rural homes have the highest proportion of homes that are further away from downtown. City homes tend to be the closest to downtown with suburban homes a little further away, and rural homes the furthest away.

The density plot of year built for city homes is right skewed, indicating that a high proportion of homes in the city were built prior to 1950. The opposite is true for Suburban homes. The density plot of year built for houses in the suburbs is left skewed, indicating a higher proportion of homes that were built more recently (1975 and on). The density plot of year built for rural homes is also slightly left skewed but shows a lower proportion of recent homes compared to homes in the suburbs and a lower proportion of older homes compared to homes in the city. Older homes tend to be in the city with newer homes in the suburbs

Next we created side by side boxplots to compare the prices of homes in the three different regions. We used the log of price for this visualization to better show the median prices and outliers.

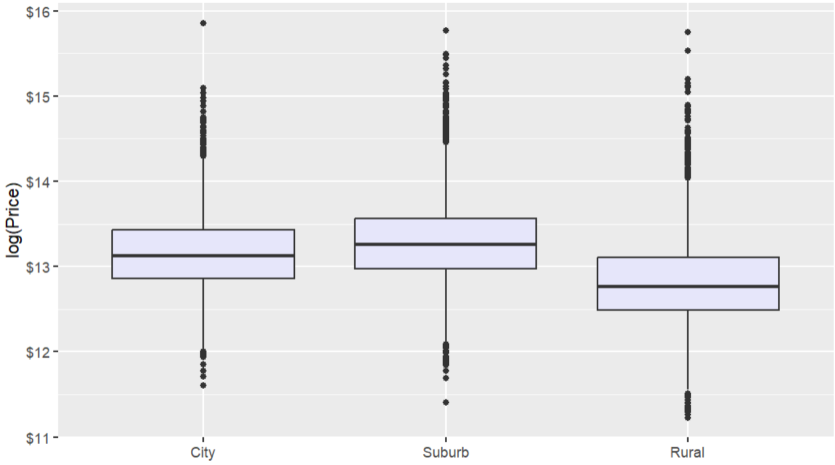


Figure 8. Box pots of Log (Price) by Region

The median prices for homes in the city and suburb are extremely close, with the median price for Suburb being slightly higher. The median price for rural homes is significantly lower than both city and suburban homes. The variability in price for each region is approximately the same and each region also has a significant number of outliers. This indicates that the region alone may not be a strong predictor of price.

We concluded our visualizations with four multivariate scatter plots to explore the impact multiple predictors could have on price. For all four visualizations we used the log of price as our response variable in an attempt to break up some of the clustering that was noted in figure 4.

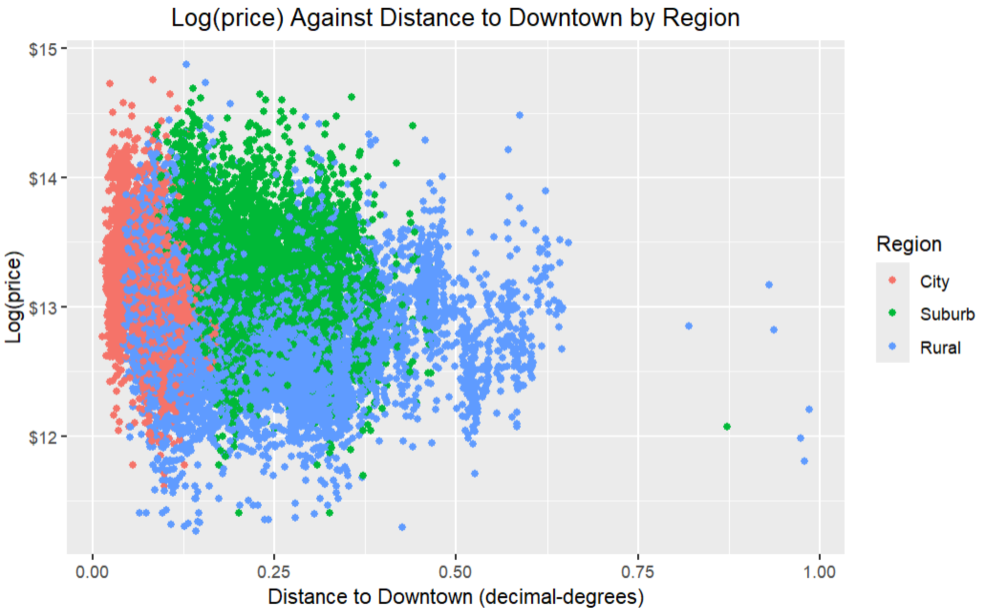


Figure 9. Scatter plot of Log (Price) Against Distance to Downtown by Region

The scatter plot of the log of price against distance to downtown by region shows interesting clustering by region. City homes and suburban homes tend to be the closest to downtown and tend to be more expensive than rural homes. City homes are closer to downtown than suburban homes, but have very similar price ranges. Rural homes have the most variability in distance to downtown and price. They tend to be less expensive than both city and rural homes, and have a mix of homes that are close to downtown and far away. Interestingly, the second most expensive home is a rural home.

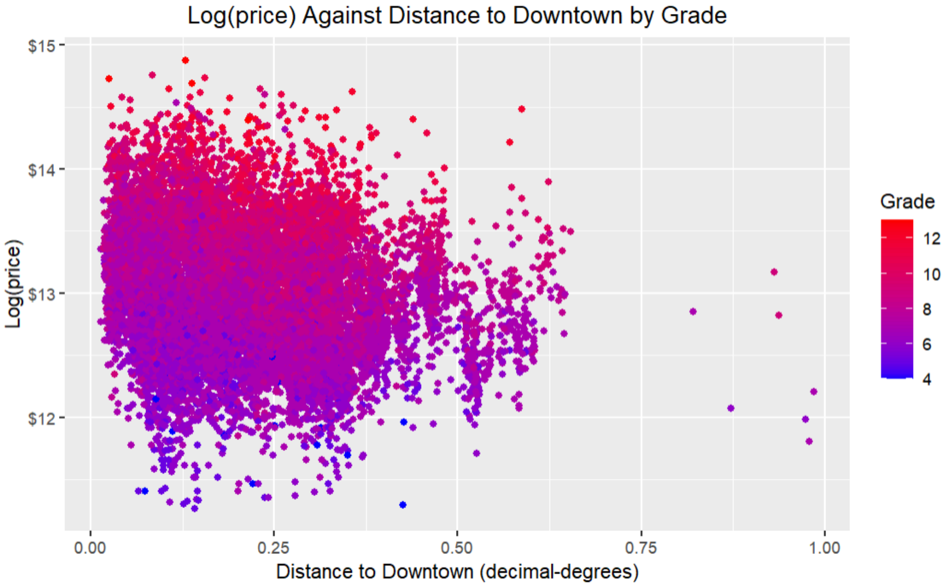


Figure 10. Scatter plot of Log (Price) Against Distance to Downtown by Grade

The scatter plot of the log of price against distance to downtown by grade shows that grade has a fairly strong relationship to price, even when considering the distance to downtown. There is variability in distance to downtown for all grades. Additionally, homes with higher grades tend to have higher prices, regardless of the distance to downtown. This indicates that the home's grade may have a stronger influence on the price than the distance to downtown.

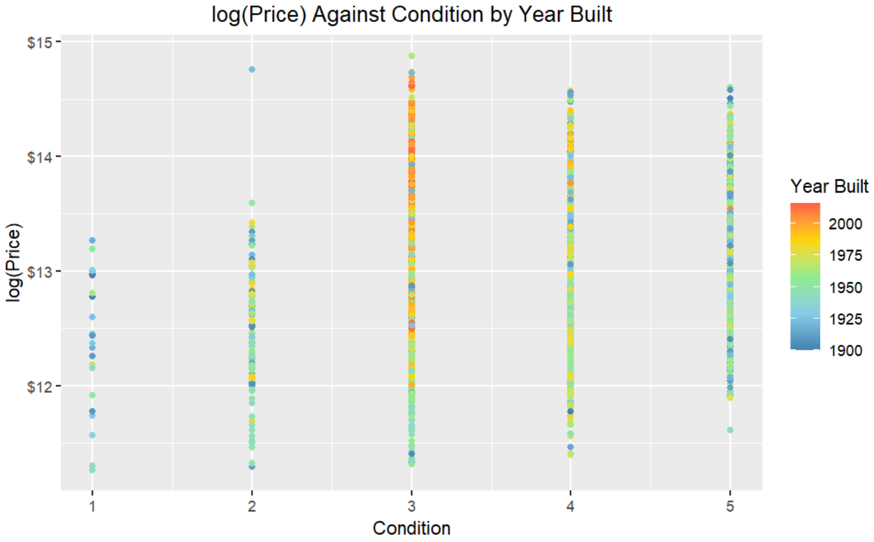


Figure 11. Scatter plot of Log (Price) Against Condition by Year Built

The scatter plot of the log of price against condition by year built shows a mix of house ages among the different conditions. Homes with the worst condition rating are almost all older homes and the average condition rating seems to have more recently built homes than the higher condition ratings. This is surprising as we would expect homes built more recently to be in better condition.The more recently built homes do tend to have higher prices regardless of the condition, though it is worth noting that there are a handful of recently built homes that are priced low.

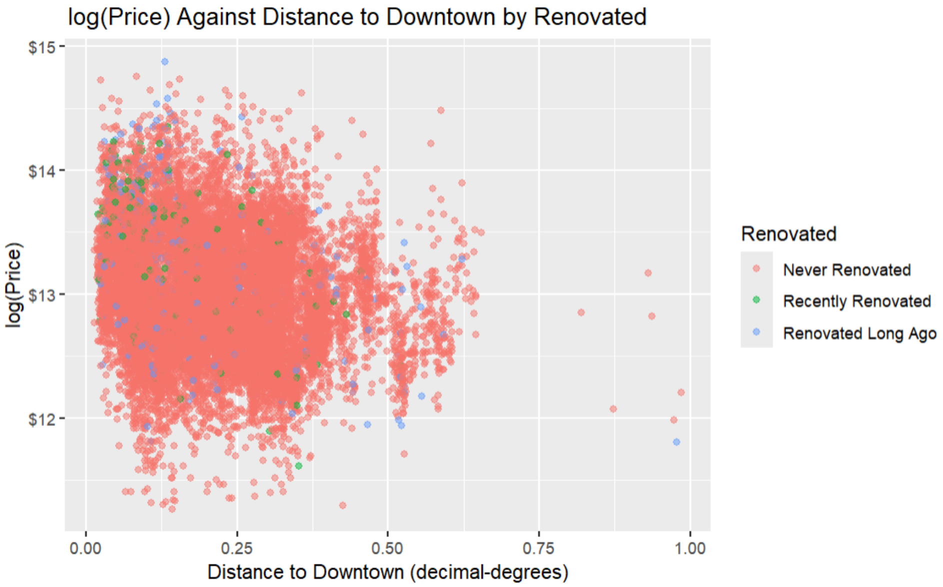


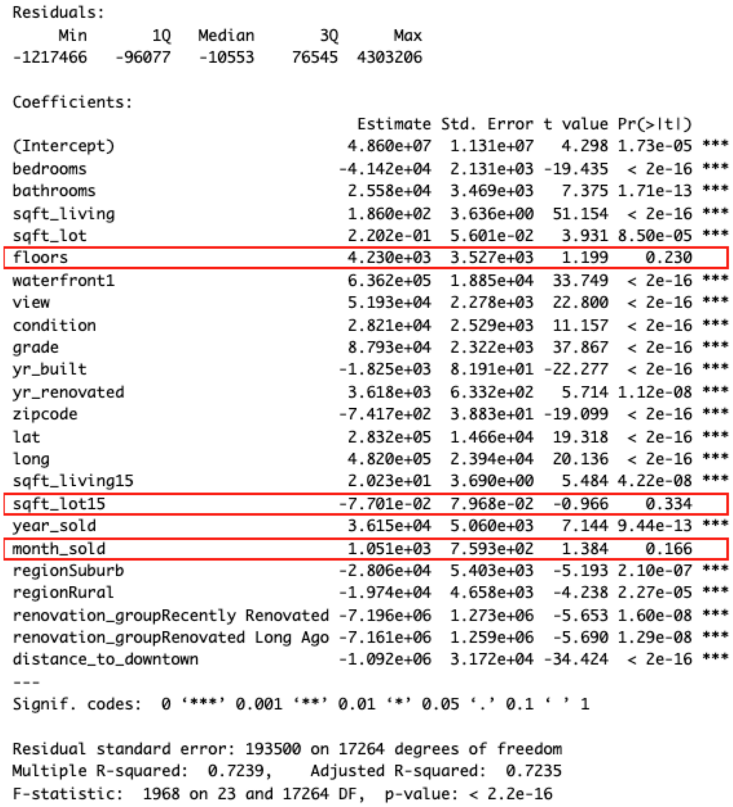
Figure 12. Scatter plot of Log (Price) by Distance to Downtown by Renovated

The scatterplot of the log of price against distance to downtown by renovated shows that a clear majority of the homes were never renovated. There does not appear to be a distinct pattern for homes that were renovated, either long ago or recently, which indicates that renovation may not be a good predictor of price. There also does not appear to be a relationship between the distance to downtown and when the home was renovated, if ever.

Section 5: Linear Regression Model

After creating preliminary visualizations to help explore the relationship between various home characteristics and their price, we continued to deepen our understanding of these relationships by building linear regression models that predict the price of homes in King County Washington. Our goal for this process was to define a regression model that accurately predicts the price of a home based on select features.

To start, we created a linear regression model that defined price as the response variable, and 21/23 of the remaining variables as predictors. Variables 'id' and 'date' were excluded from this initial model since a unique identifier inherently has no predictive value and a coefficient for every unique date is ineffective. The summary of this model indicated that 'floors', 'sqft_lot15', and 'month_sold' are not statistically significant predictors of price and therefore should be removed from the model. Additionally, the geographic coordinates ('lat' and 'long') are redundant given we previously defined the 'distance_to_downtown' variable.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-1217466   -96077   -10553    76545  4303206 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       4.860e+07  1.131e+07   4.298 1.73e-05 ***
bedrooms                         -4.142e+04  2.131e+03 -19.435  < 2e-16 ***
bathrooms                         2.558e+04  3.469e+03   7.375 1.71e-13 ***
sqft_living                       1.860e+02  3.636e+00  51.154  < 2e-16 ***
sqft_lot                          2.202e-01  5.601e-02   3.931 8.50e-05 ***
floors                            4.230e+03  3.527e+03   1.199    0.230    
waterfront1                       6.362e+05  1.885e+04  33.749  < 2e-16 ***
view                              5.193e+04  2.278e+03  22.800  < 2e-16 ***
condition                         2.821e+04  2.529e+03  11.157  < 2e-16 ***
grade                             8.793e+04  2.322e+03  37.867  < 2e-16 ***
yr_built                         -1.825e+03  8.191e+01 -22.277  < 2e-16 ***
yr_renovated                      3.618e+03  6.332e+02   5.714 1.12e-08 ***
zipcode                          -7.417e+02  3.883e+01 -19.099  < 2e-16 ***
lat                               2.832e+05  1.466e+04  19.318  < 2e-16 ***
long                              4.820e+05  2.394e+04  20.136  < 2e-16 ***
sqft_living15                     2.023e+01  3.690e+00   5.484 4.22e-08 ***
sqft_lot15                       -7.701e-02  7.968e-02  -0.966    0.334    
year_sold                         3.615e+04  5.060e+03   7.144 9.44e-13 ***
month_sold                        1.051e+03  7.593e+02   1.384    0.166    
regionSuburb                     -2.806e+04  5.403e+03  -5.193 2.10e-07 ***
regionRural                      -1.974e+04  4.658e+03  -4.238 2.27e-05 ***
renovation_groupRecently Renovated -7.196e+06  1.273e+06  -5.653 1.60e-08 ***
renovation_groupRenovated Long Ago -7.161e+06  1.259e+06  -5.690 1.29e-08 ***
distance_to_downtown             -1.092e+06  3.172e+04 -34.424  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 193500 on 17264 degrees of freedom
Multiple R-squared:  0.7239,	Adjusted R-squared:  0.7235 
F-statistic:  1968 on 23 and 17264 DF,  p-value: < 2.2e-16
```

Figure 13. R Summary Output for Full Model

Before reducing the model, we determined the influential observations, high leverages observations, and outliers in the training set. There were 592 outliers according to studentized and standardized, 1860 high leverage data points, 1162 influential points according to DFFITS, and 0 influential points according to Cook's distance. Although there is a disparity between DFFITS and Cook’s distance, this is expected, as individual data points can strongly influence their own predicted values without substantially impacting the overall model fit. At first, we determined to not remove any outliers to capture the full variability of home prices.

Based on the results from the full linear regression model, we created a slightly reduced version with 15 predictors instead of the original 21. However, we agreed that 15 variables still introduced unnecessary complexity. To further refine the model, we used the regsubsets() function to identify an optimal subset of predictors. The models with the best R², Mallows’ Cp, and BIC values all selected the same set of predictor variables: bedrooms, sqft_living, waterfront, view, grade, yr_built, region, and distance_to_downtown. However, the model's diagnostics indicated that some linear regression assumptions were violated.

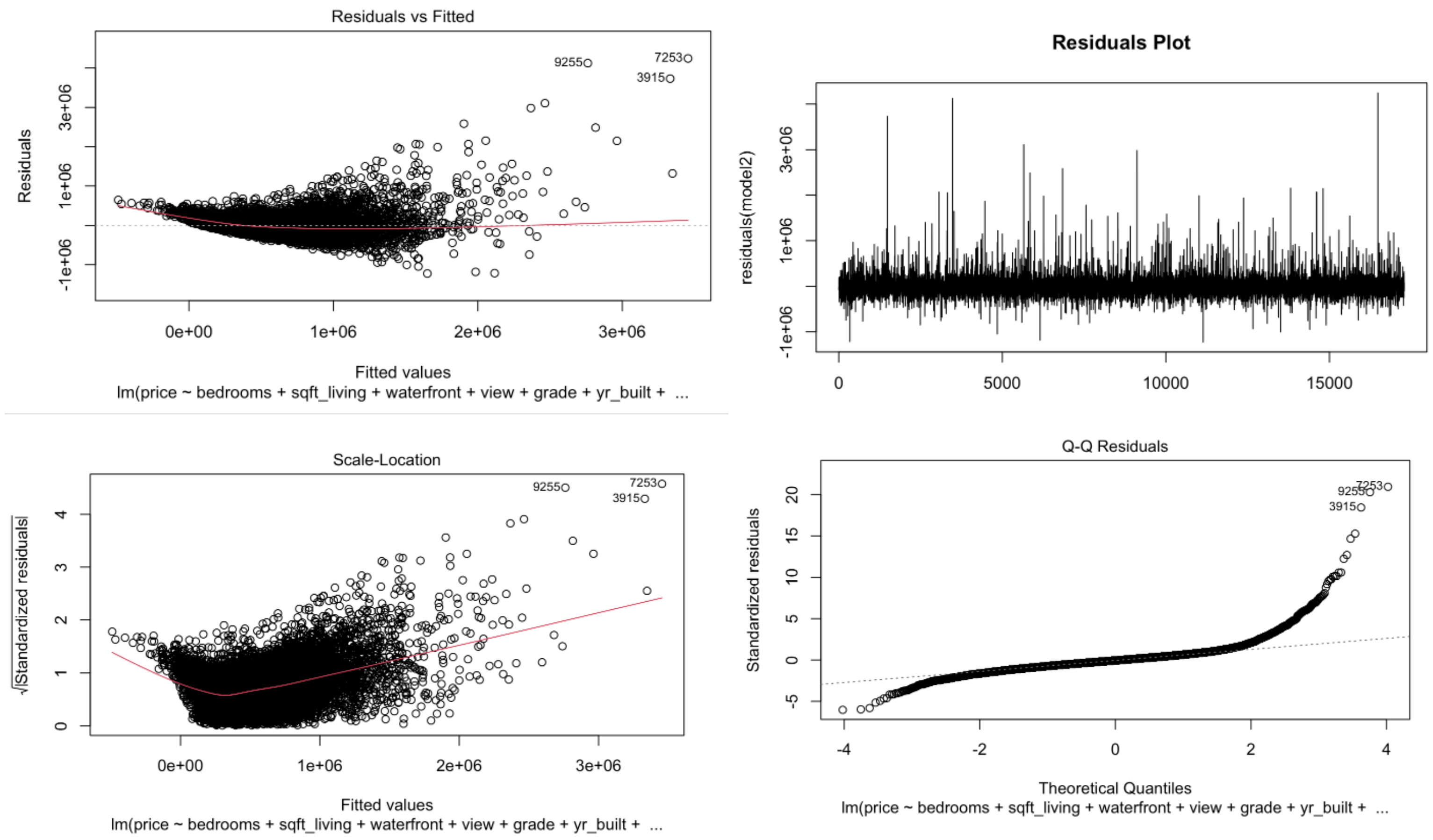
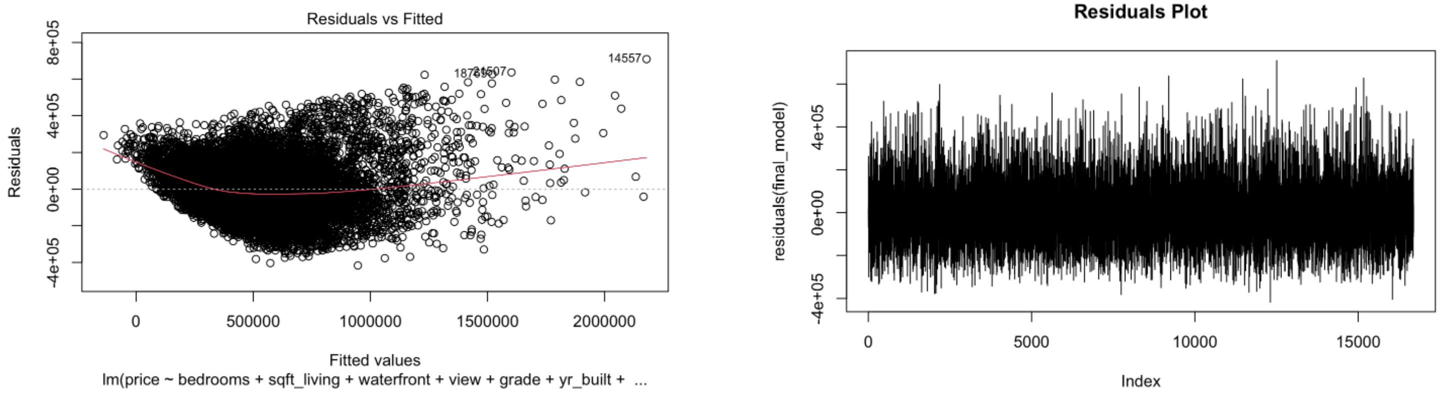


Figure 14. Model Diagnostics (before removing outliers)

The residual plots clearly indicate violations of both homoscedasticity and normality assumptions. The non-constant spread of residuals in the Scale-Location plot and the deviation of the Q-Q plot's tail from the line suggest these issues.

Rather than applying log transformations or standardizing select variables, we observed that the residual plots exhibited a heavy tail—indicating the presence of outliers that deviate substantially from the overall distribution. Therefore, we chose to remove 592 outlier observations from the training data to reduce bias and improve the model.



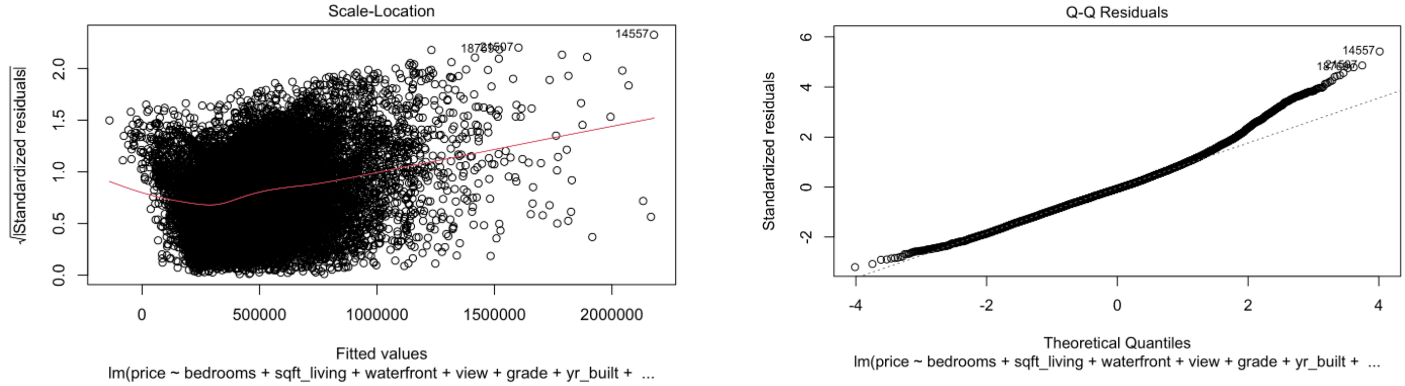


Figure 15. Model Diagnostics (after removing outliers)

Despite the tails in the residual plots being seemingly still heavy, the context of our data, where significant price variations are expected as homes can be very expensive, allows us to conclude that the linear regression assumptions are met after outlier removal. Specifically: linearity is supported by the random scatter of residuals in the Residuals vs Fitted plot; independence is indicated by the randomness in the residuals plot; homoscedasticity is satisfied by the constant spread of residuals; and normality of errors is supported by the linear alignment of points in the Q-Q plot.

Finally, we defined the equation associated with our final model and used it to predict the prices of the homes in the testing set.

Final Regression Equation:

price = 3154933.30 – 25591.98(bedrooms) + 154.65(sqft_living) + 584913.42(waterfront) + 45372.16(view) + 96291.07(grade) – 1782.64(yr_built) + 41909.01(regionSuburb) – 59865.43(regionRural) – 457494.79(distance_to_downtown)

To determine the accuracy of our model's predictions, we calculated the RMSE value to be 204831.62 and R^2 value to be 0.68. These values tell us that on average, our model's predicted house prices deviate from actual prices by $204,831.62 and the model explains about 68.3% of the variation in house prices.

In short, our linear regression model showcases how location and property features play a substantial role in determining home values in King County. For example, homes farther from downtown or in rural areas tend to be significantly less expensive, while higher construction grades and better views drive prices upward.

Section 6: Visualizing Factors That Influence Home Quality

In this section, we aimed to explore what characteristics make a home in King County, Washington, more likely to be considered 'good quality.' A home is considered 'good quality if it has both a condition rating greater than 3 and a grade rating higher than 7. This helped us analyze homes that are both structurally sound and built or finished with higher standards.

The visualizations in this section were created based on the training subset of the dataset and focus on the binary, good_quality variable.

We started by exploring the overall distributions of the two main features used to define home quality: grade and condition.

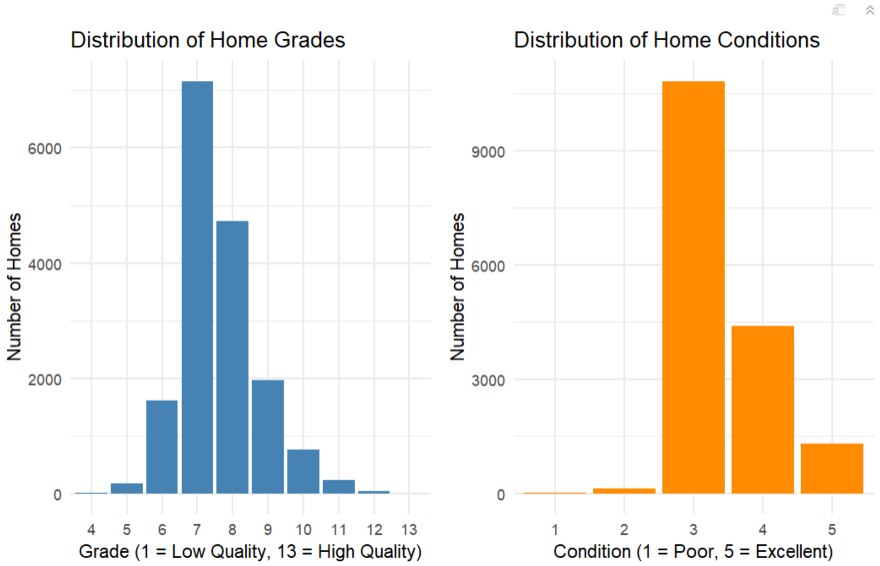


Figure 16a (left). Distribution of Home Grade

Figure 16b (right). Distribution of Home Condition

Looking at these plots, most homes fall between grades 7 and 9, with a sharp drop-off after that, which indicates that very low-end and very high-end homes (grades 1-5 and 10-13) are rare in this dataset.

For condition, the majority of homes are rated condition 3, which seems to be the baseline. That tells us the dataset is skewed toward homes in average or decent condition, and only a small portion are in poor or excellent shape.

Next, we looked at how many homes met our definition of good quality..

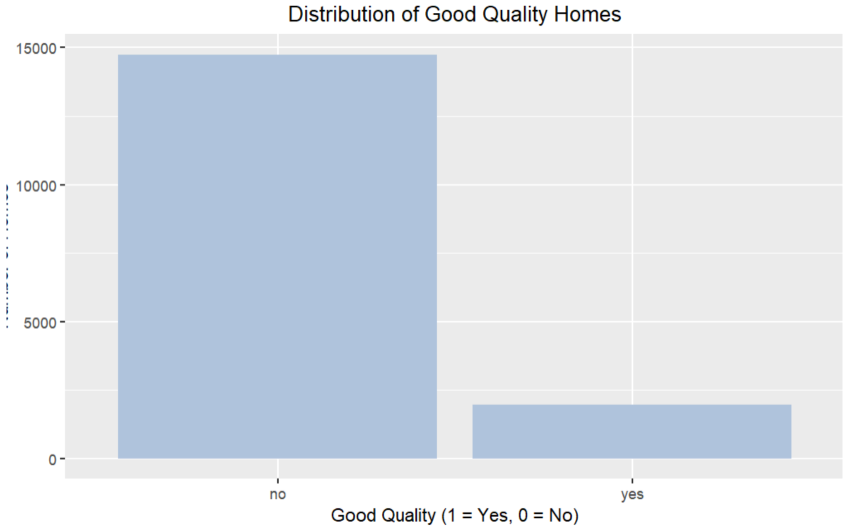
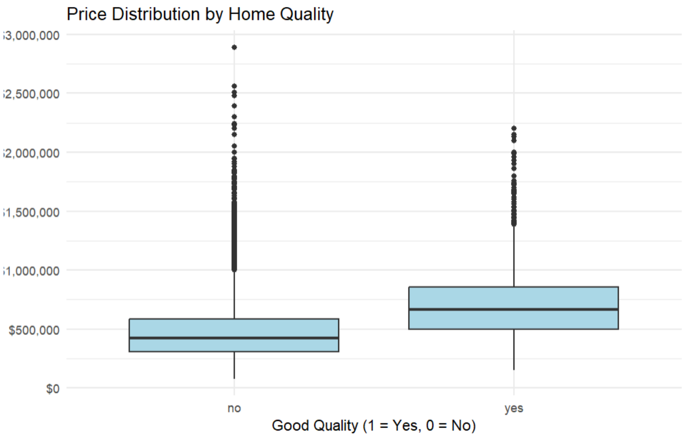


Figure 17. Proportion of good quality homes

Since we set a fairly strict threshold, requiring both high condition and high grade, it was not surprising that most homes did not meet the criteria. The majority were marked as 0 (not good quality), with a much smaller portion labeled as 1 (good quality).

To see if good quality homes actually sell for more, we compared sale prices between homes labeled as good quality and those labeled as not good quality.



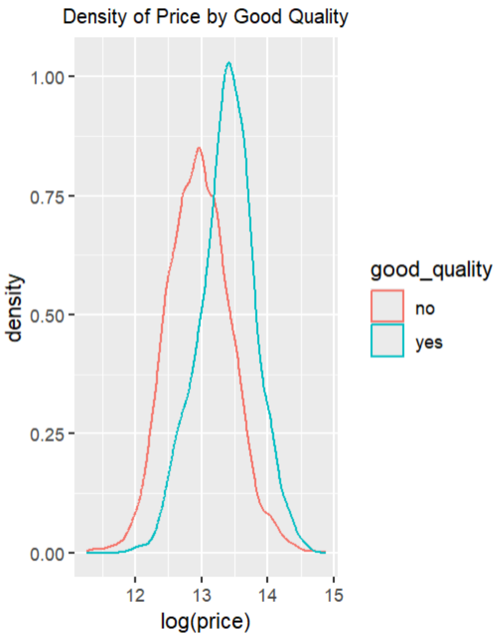


Figure 18. Boxplots of sale price by home quality

Figure 19. Density plot of price by home quality

The boxplot shows a clear difference in price between the two quality groups. Homes labeled as good quality have a noticeably higher median sale price and a wider range overall. This confirms our assumption that better condition and construction grade are tied to higher market value in King County. Both groups have some pricey outliers, but the most expensive homes are mostly in the good quality category.

This is confirmed with the density plot which shows a higher proportion of good quality homes have a greater log price than homes that are not good quality. This is to be expected as one would expect better quality to positively correlate with price.

We then looked at how various features of the home and its location related to whether it was considered good quality or not.

We created bar charts to assess the relationship between a home overlooking the waterfront and fitting the good quality criteria.

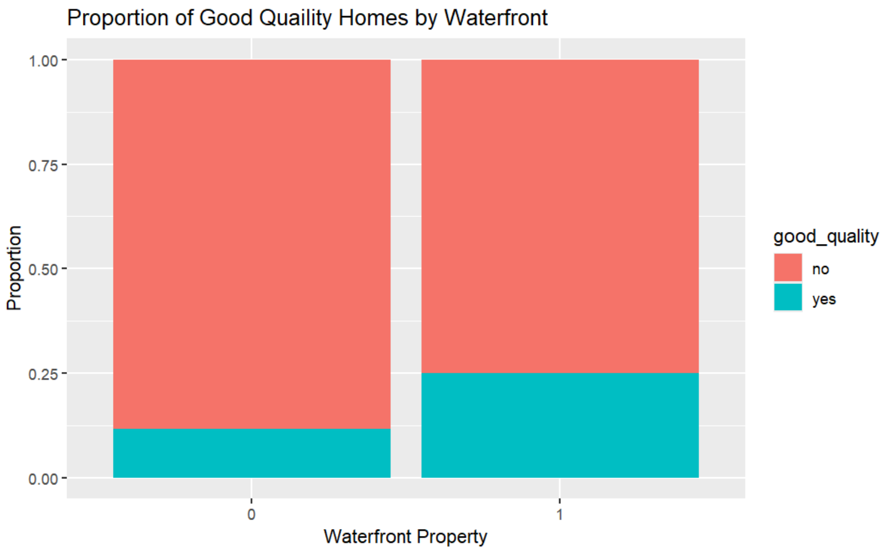


Figure 20. Proportion of good quality homes by waterfront

The plot shows that the proportion of homes that are waterfront properties and good quality is significantly higher than the proportion of homes that are non-water properties and good quality.

Next, we wanted to see if there was any connection between home quality and location. Since we already grouped zip codes into City, Suburb, and Rural regions, we used a stacked bar chart to compare how the proportion of good quality homes varies across these areas.

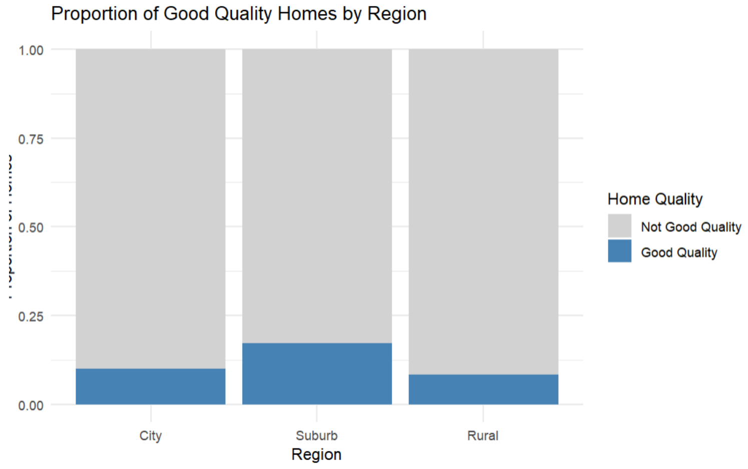
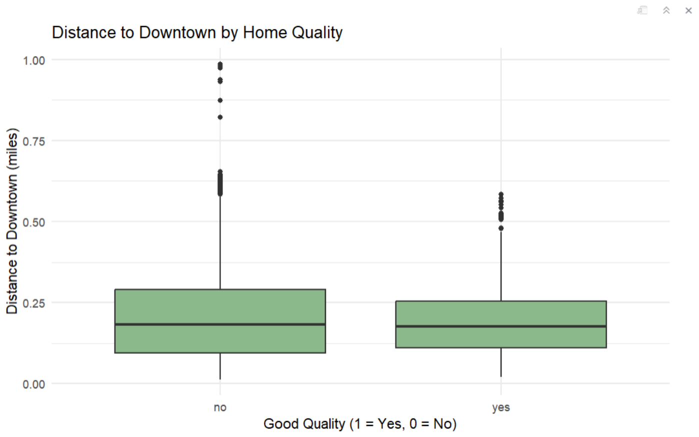


Figure 21. Proportion of good quality homes across regions

The chart shows that the Suburban area has a higher proportion of good quality homes compared to City or Rural areas. The differences are not huge, but there is a clear trend, suburban neighborhoods seem to have more newer or upgraded homes overall. That lines up with what we would expect, since suburbs usually have more planned developments and newer builds.

We also looked at whether home quality is related to how close a home is to downtown Seattle. Since homes near central areas and amenities are usually more desirable, we compared the distance to downtown for good quality homes versus the rest.



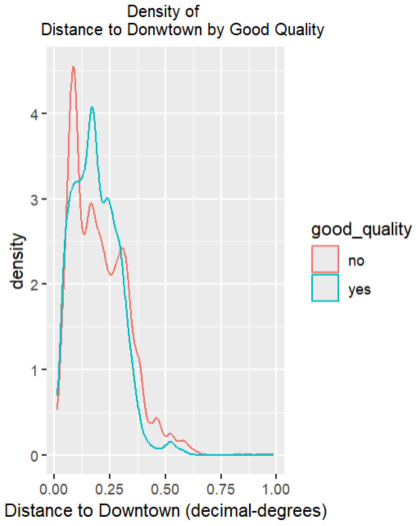


Figure 22. Boxplot of Distance to Downtown by Home Quality Group

Figure 23. Density Plot of Distance to Downtown by Home Quality Group

While the median distance to downtown is similar for both groups, good quality homes tend to be more tightly clustered around the city center. Homes labeled as not good quality show more variability in distance, suggesting they are more spread across both urban and outer suburbs.

Additionally, the density plot shows that a slightly higher proportion of not good quality homes are closer to downtown than good quality homes.

Finally, we looked at the relationship between when a house was built and the quality.

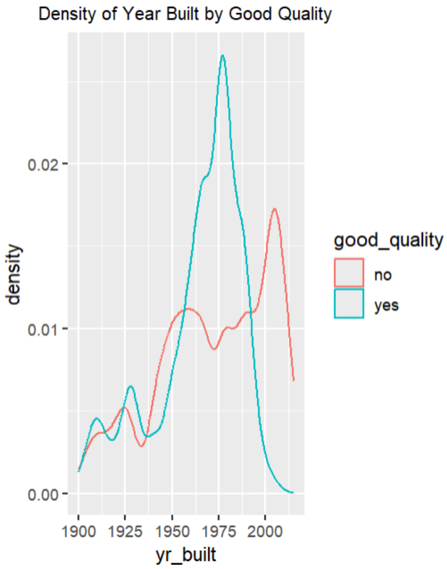


Figure 24. Density Plot of Year Built by Home Quality Group

A significantly higher proportion of good quality homes were built in the mid to late 20th century compared to homes that are not good quality, with a higher proportion of not good quality homes being built from 2000-2015. This suggests that at the turn of the century, building quality diminished possibly due to economic factors and or building practices. Furthermore, this suggests that the year a house was built may be associated with the likelihood of a home being good quality.

After creating visualizations summarizing the data and looking at one or two features of the home, we wanted to look at how multiple variables interact together with home quality. We used multivariate scatter plots to look at how factors like size, price, distance to downtown, and region interact with each other and whether a home is good quality.. By adding color or breaking the plots into regions, we were able to spot patterns and clusters that don't show up in simpler charts. These visualizations helped us see how certain combinations of traits show up more often in higher-quality homes.

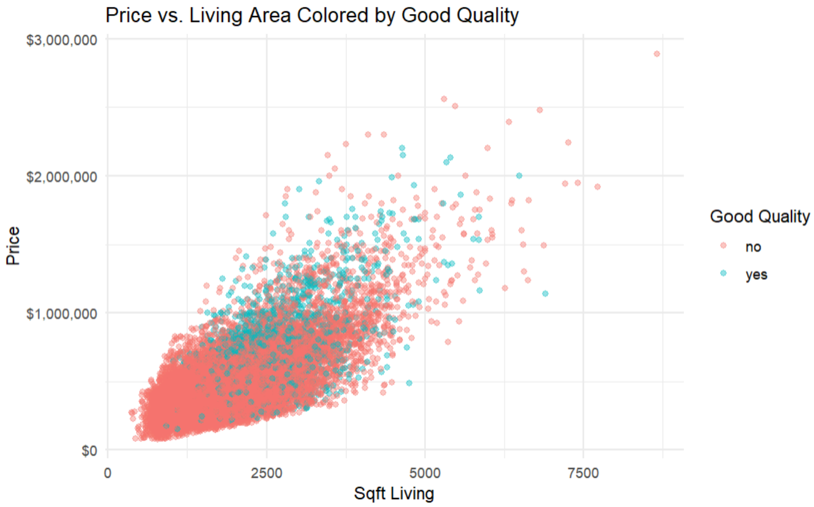


Figure 25. Scatter plot of Price Against. Living Area by Home Condition

The plot of price against living area by good quality shows a clear positive relationship between home size and price, bigger homes generally sell for more. What stands out, though, is that good quality homes are mostly clustered in the higher end of both size and price. While a few show up in the middle, they are significantly more likely than lower-quality homes to fall into the most expensive, largest category. This confirms the idea that better construction and condition are often tied to high end properties.

To see how distance from downtown and price relate to home quality, we created a scatterplot with distance on the x-axis and price on the y-axis. We colored the points based on whether the home was labeled as good quality or not, so we could spot any patterns or clusters.

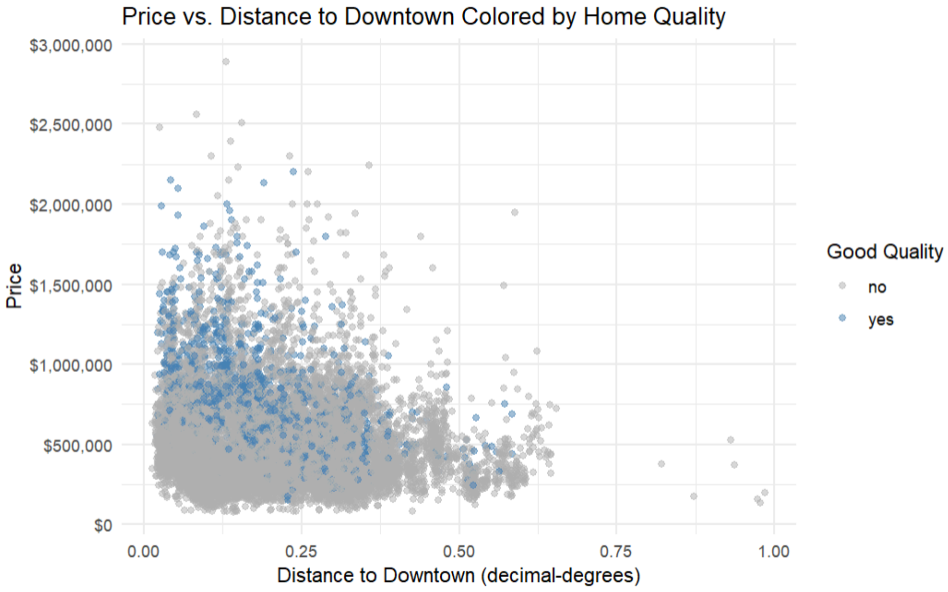


Figure 26. Scatter plot of Price Against Distance to Downtown (decimal-degrees), by Home Quality

The scatterplot shows that homes located near downtown generally sell for higher prices, with good quality homes more commonly found in these higher priced, central neighborhoods. That said, there is a heavy concentration of lower priced homes at shorter distances, which makes it harder to clearly see broader pricing trends across the area.

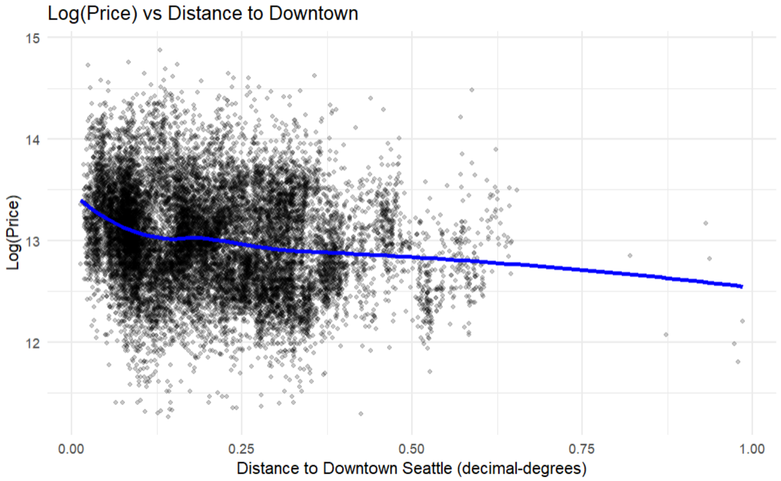


Figure 27: Scatterplot of Log(Price) vs Distance to Downtown (decimal-degrees)

To fix the clusters around areas near downtown, we applied a log transformation to the sale prices. After the transformation, the data spread out more, making it easier to see that homes closer to downtown Seattle generally sell for higher prices, while prices tend to drop as distance increases, although there's still a lot of variation.

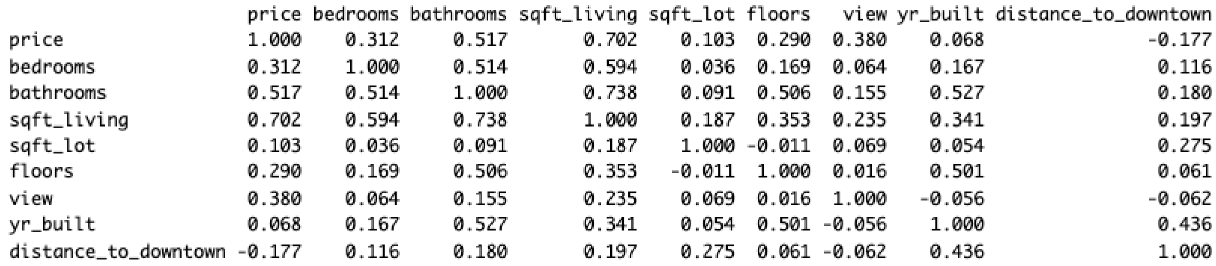
Overall, while proximity to downtown plays an important role in home prices, other factors also contribute to the variation observed in the data.

Section 7 - Using Logistic Regression to Predict Good Quality Homes

This section details the process of creating a suitable logistic regression model that predicts whether a home in King County, Washington is deemed good quality or not. As stated in the previous section, we created a new variable called 'good_quality', which marks a home as high quality if it has both a condition rating greater than 3 and a grade rating higher than 7. To begin our process for building a suitable logistic regression model, we referenced the visualizations in Section 6 to help determine which variables in the data set may have influence on our good_quality variable.

There appears to be a significant difference between good quality homes and not good quality homes when looking at whether the property overlooks the waterfront, when the home was built, the distance to downtown, and the price. This indicates that these variables may be strong predictors of whether a home is good quality. Region and renovation group did not appear to be the strongest predictors of whether a home was good quality, but could still be useful.

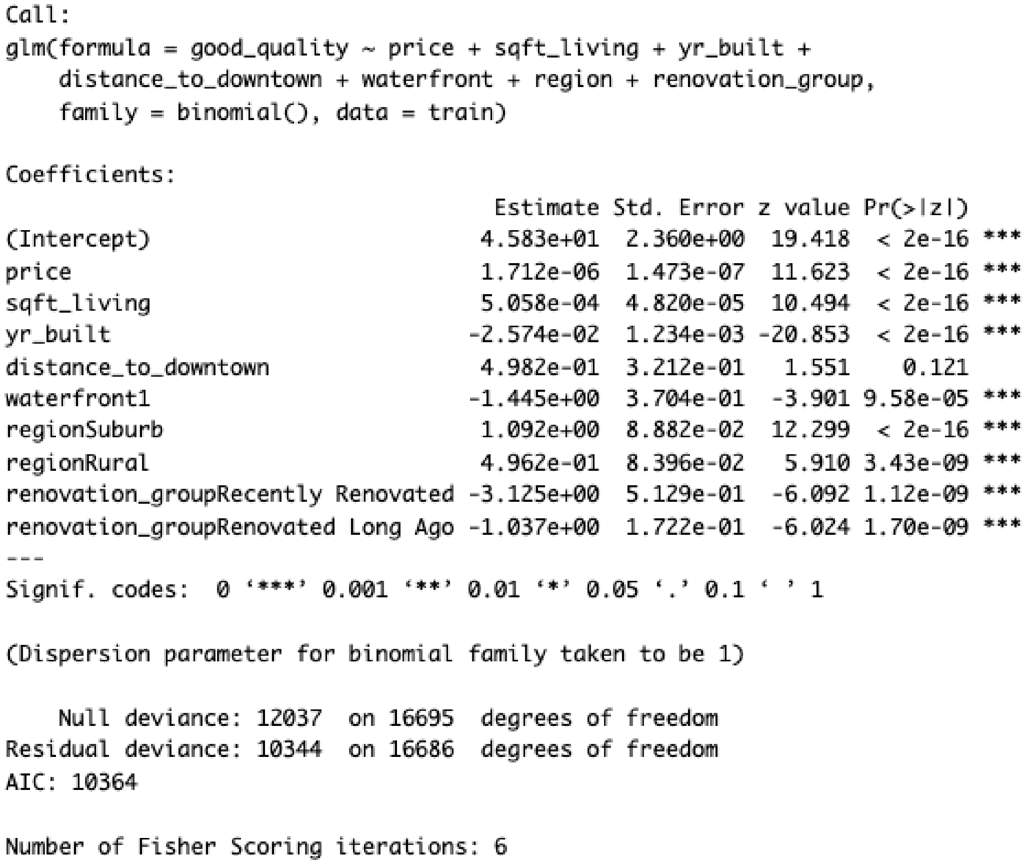
Before building our first logistic regression model we check the correlation between the quantitative predictors.



```
                     price bedrooms bathrooms sqft_living sqft_lot floors   view yr_built distance_to_downtown
price                1.000    0.312     0.517       0.702    0.103  0.290  0.380    0.068               -0.177
bedrooms             0.312    1.000     0.514       0.594    0.036  0.169  0.064    0.167                0.116
bathrooms            0.517    0.514     1.000       0.738    0.091  0.506  0.155    0.527                0.180
sqft_living          0.702    0.594     0.738       1.000    0.187  0.353  0.235    0.341                0.197
sqft_lot             0.103    0.036     0.091       0.187    1.000 -0.011  0.069    0.054                0.275
floors               0.290    0.169     0.506       0.353   -0.011  1.000  0.016    0.501                0.061
view                 0.380    0.064     0.155       0.235    0.069  0.016  1.000   -0.056               -0.062
yr_built             0.068    0.167     0.527       0.341    0.054  0.501 -0.056    1.000                0.436
distance_to_downtown -0.177   0.116     0.180       0.197    0.275  0.061 -0.062    0.436                1.000
```

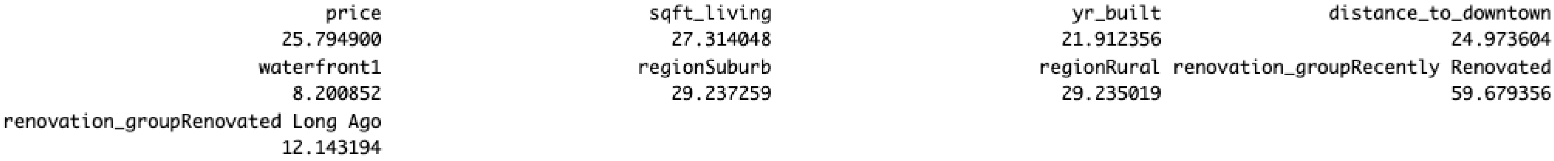
We noticed that price and sqft_living are highly correlated along with sqft_living and bathrooms and sqft_living and bedrooms.

Based on the visualizations and our prior analysis in Section 5, we suspect that waterfront, region, renovation_group, price, sqft_living, yr_built, and distance_to_downtonwn may influence our response variable. We used the glm() function in R to fit a logistic regression model using these predictors as a starting point.



```
Call:
glm(formula = good_quality ~ price + sqft_living + yr_built + 
    distance_to_downtown + waterfront + region + renovation_group, 
    family = binomial(), data = train)

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)                         4.583e+01  2.360e+00  19.418  < 2e-16 ***
price                               1.712e-06  1.473e-07  11.623  < 2e-16 ***
sqft_living                         5.058e-04  4.820e-05  10.494  < 2e-16 ***
yr_built                           -2.574e-02  1.234e-03 -20.853  < 2e-16 ***
distance_to_downtown                4.982e-01  3.212e-01   1.551    0.121    
waterfront1                        -1.445e+00  3.704e-01  -3.901 9.58e-05 ***
regionSuburb                        1.092e+00  8.882e-02  12.299  < 2e-16 ***
regionRural                         4.962e-01  8.396e-02   5.910 3.43e-09 ***
renovation_groupRecently Renovated -3.125e+00  5.129e-01  -6.092 1.12e-09 ***
renovation_groupRenovated Long Ago -1.037e+00  1.722e-01  -6.024 1.70e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12037  on 16695  degrees of freedom
Residual deviance: 10344  on 16686  degrees of freedom
AIC: 10364

Number of Fisher Scoring iterations: 6
```

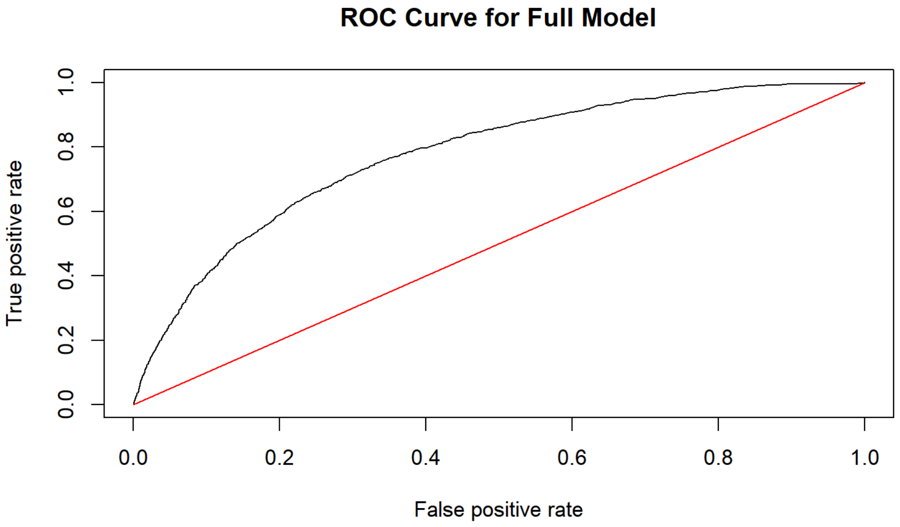
From our full model we see that all of our predictors are reported as significant except for waterfront and region. Waterfront being reported as an insignificant predictor came as a surprise to us, given the original visualizations. After fitting this initial model we wanted to assess any multicollinearity in our model. We calculated the following VIFs.



```
                             price                        sqft_living                           yr_built               distance_to_downtown
                         25.794900                          27.314048                          21.912356                          24.973604
                       waterfront1                       regionSuburb                        regionRural renovation_groupRecently Renovated
                          8.200852                          29.237259                          29.235019                          59.679356
renovation_groupRenovated Long Ago
                         12.143194
```

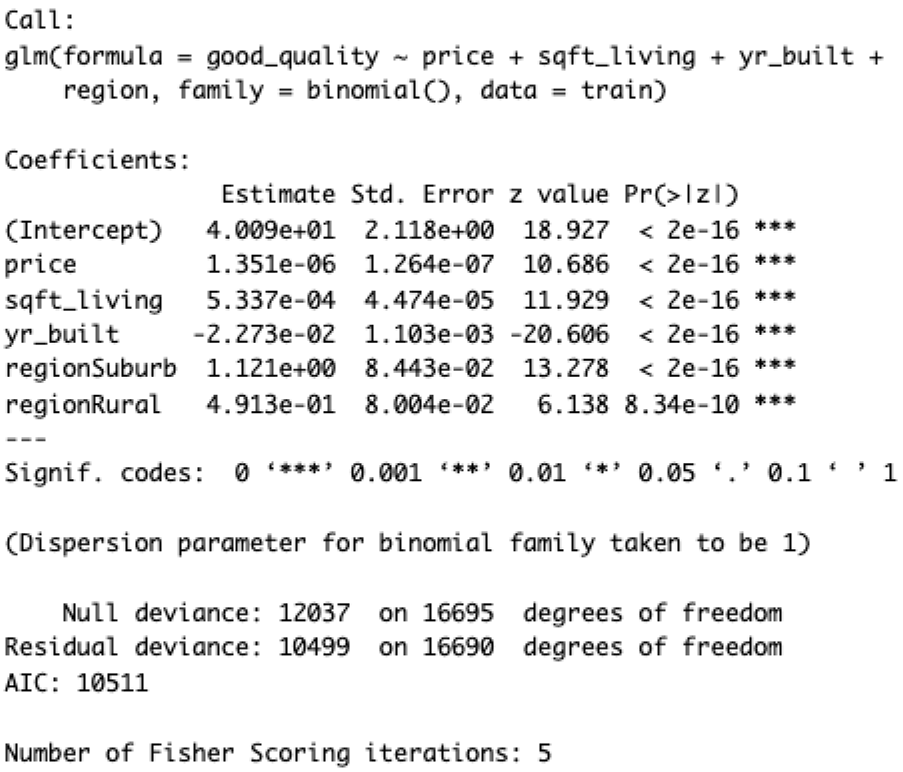
From the VIFs we see there seems to be a high level of multicollinearity amongst all predictors, but the renovation group seems to have the highest of all the predictors in our model.

To assess the predictive ability of this first logistic regression model, we created and analyzed an ROC plot and calculated the area under the curve (AUC).



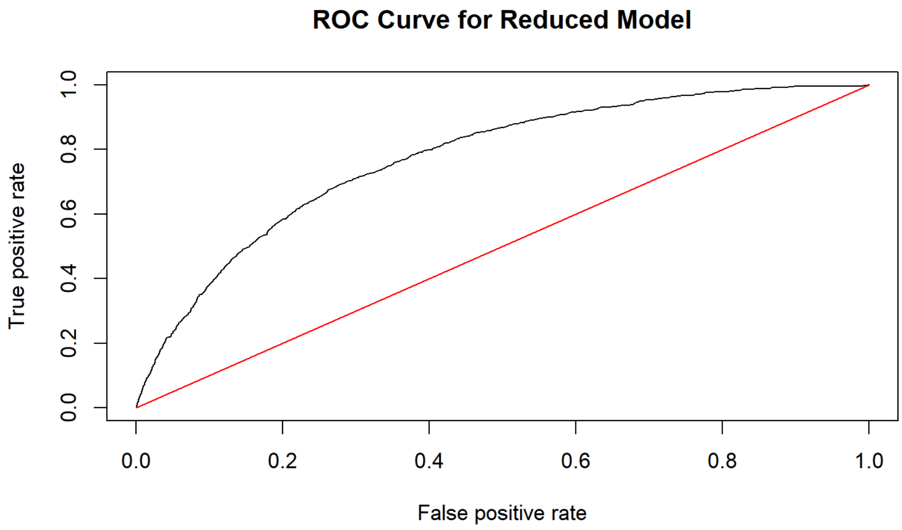
From the ROC plot we see that our initial logistic regression model performs better than random guessing and has an AUC of 0.787. We also calculated the error rate and accuracy of our logistic regression model. Our accuracy was 87.55% and our error rate was 12.45%. Here we have what appears to be good accuracy and error rate, but we must remember, in Figure 17 we saw that most of the train data set is already made up of homes that are considered not good_quality, i.e. we are dealing with an unbalanced sample size of the response variable.

In an attempt to improve upon this model, we decided to remove the reportedly insignificant predictors, waterfront and region as well as the renovation group variable and refit a reduced logistic regression model.



```
Call:
glm(formula = good_quality ~ price + sqft_living + yr_built +
    region, family = binomial(), data = train)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.009e+01  2.118e+00  18.927  < 2e-16 ***
price         1.351e-06  1.264e-07  10.686  < 2e-16 ***
sqft_living   5.337e-04  4.474e-05  11.929  < 2e-16 ***
yr_built     -2.273e-02  1.103e-03 -20.606  < 2e-16 ***
regionSuburb  1.121e+00  8.443e-02  13.278  < 2e-16 ***
regionRural   4.913e-01  8.004e-02   6.138 8.34e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12037  on 16695  degrees of freedom
Residual deviance: 10499  on 16690  degrees of freedom
AIC: 10511

Number of Fisher Scoring iterations: 5
```

After fitting the reduced logistic regression model, we again created and analyzed an ROC plot and calculated the area under the curve (AUC) to assess the predictive ability of the reduced model. For our reduced logistic regression model we see that the ROC curve is practically identical to the ROC curve for the full model and the new AUC is 0.782.



ROC Curve for Reduced Model

Since we cannot visually discern which model performs better, we decided to conduct a likelihood ratio test to determine if it makes sense to use the full or reduced model to predict good quality homes in King County, Washington.

Our null hypothesis was that β_4, β_5, and β_8 were equal to zero, and our alternative hypothesis was that at least one of the coefficients in the null are not zero. We computed a test statistic of 155 and had a critical value of 7.814; therefore we rejected the null hypothesis, meaning the additional predictors left out of the reduced model significantly improve the model. The data supports using the full model over the reduced model.

Some interesting conclusions we saw from our logistic regression model (full) were that for a one unit increase in year_built, the log odds of a home being good quality decreases by 0.0227 when holding all other predictors constant. Furthermore the log odds of a home being good quality is 1.121 higher for a home located in the suburbs than for a home located in the urban/city area of King, County Washington, when holding all other predictors constant.

In conclusion, we selected our original, full model to predict good quality homes in King County Washington. After analyzing the ROC curve we can confirm that our final model performs better than random guessing.