Stat 6021: Project 2

Group 5: Etienne Jimenez, Bardia Nikpour, Christian Ollen, Venkat Viswanathan

Section 1

Question 1

Homeownership is one way millions of Americans build wealth. Homes are the foundations for millions of families, but what makes one home more valuable than another? When one first purchases a home, the first things a person considers are the size of the home, the number of bedrooms, and the number of bathrooms. But are these the most essential internal housing factors that drive up the price of a home? Our team conducted a study to determine what internal housing factors matter most within King County, Washington.

Our team analyzed the following internal housing factors: bedrooms, bathrooms, sqft_living, floors, condition, grade, sqft_above, and sqft_basement to see what had the most significant influence on housing prices. Initially, we were able to narrow the focus of our study to just four predictor variables:

- The number of bathrooms
- The square footage of the home
- The square footage above the ground of a house
- The construction grade and design of a home

We further reduced our model by choosing a subset of our square footage variables to only include the overall square footage, leaving our model with three predictor variables. The number of bathrooms negatively correlated with the price, which may seem counterintuitive. From our personal experiences of buying and selling homes, more bathrooms are typically desirable, so we further reduced our model to two predictor variables.

Our study's findings reveal that the two most influential variables on housing prices in King County, Washington are the square footage of the home and the construction grade of the home. These findings are crucial for understanding the dynamics of the local housing market and can greatly assist in making informed decisions related to real estate.

Question 2

Section 2

The dataset used for this analysis is obtained from Kaggle. It contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. Out of all 21 variables, we have decided to use 13 of them, and create 6 new variables based on some existing ones.

1. `price`: Sale price of the house.
2. `bedrooms`: Number of bedrooms in the house.
3. `bathrooms`: Number of bathrooms in the house.
4. `floors`: Number of floors (levels) in the house.
5. `sqft_living`: Square footage of the interior living space of the house. The bigger the living space or a house, the higher we may expect the price to be.

6. `sqft_above`: Square footage of the interior living space above ground level.
7. `sqft_basement`: Square footage of the interior living space below ground level.
8. `waterfront`: The waterfront variable is an indicator variable, and is determining whether the house in question is located near the waterfront or not. A house near the waterfront is labeled as 1, and one which is not near the waterfront is labeled as 0. We are inclined to believe this could be a relevant predictor because houses near water are expected to be more expensive.
9. `view`: The view variable rates on a scale of 0 to 4 the quality of the view the property was. The median value for this variable is 0, which may indicate that only about half of the apartments had a somewhat favorable view.
10. `view_med`: An indicator variable pointing out whether a house's view is above the median (indicated in the previous variable's description as being 0). If the view rating is above 0, it is labeled 1; otherwise, it is labeled 0.
11. `condition`: Condition is a categorical variable, indexed from 1 to 5, which represents the overall condition of the house. Nicer apartments may be sold at a higher price than apartments which have lower ratings.
12. `condition_med`: An indicator variable of condition, pointing out whether a house's condition is above the median of 3. It is labeled 1 if the house's condition is 4 or 5, and 0 otherwise.
13. `grade`: Grade is highlighting the quality level of the building's construction and design. It is indexed from 1 to 13, and houses with ratings 11-13 are considered to be of the highest quality.
14. `grade_med`: An indicator variable which is labeled 1 if a house's grade rating is of the highest quality, that is, which has a rating of 11, 12, or 13. It is labeled 0 if its grade rating is 10 or lower.
15. `yr_built`: Year the house was built. Houses in King's County were built as early as the year 1900, and the latest were built in 2015.
16. `decade_built`: To make a more general approach towards visualizing the relationship of a house's price and the year it was built, we group houses by decade of being built. The earliest decade is 1900, while the most recent decade is of the 2010s.
17. `yr_renovated`: Year the house was last renovated. We acknowledge that houses that were renovated recently could be priced highly in comparison with those who were never renovated at all. Houses that have not been renovated are marked with the value 0.
18. `decade_renovated`: Similar to decade_built, this variable groups houses by the decades they were last renovated. This helps us make better visualizations, and to find any relationships with the results obtained from visualizations in decade_built.
19. `above_median`: A simple indicator variable, labeled 1 if a house is above median price, and 0 otherwise. Note that the median price of a house, as obtained from the training data set, was found to be $450,000.

Section 3

Question 1 : What internal housing factors influence the price of homes in King County from May 2014 to May 2015?

- *Response Variable*: `price`

- Motivation: This question aims to understand the relationship between various internal features of a house (such as bedrooms, bathrooms, square footage, floors, condition, grade, and others) and its sale price. Investigating these factors can provide insights into the determinants of housing prices in the King County area during the specified time.

Question 2: Does the presence of one or more of Waterfront, Condition, View, Grade, Yr Built, and Yr Renovated cause the house to be sold above the median price?

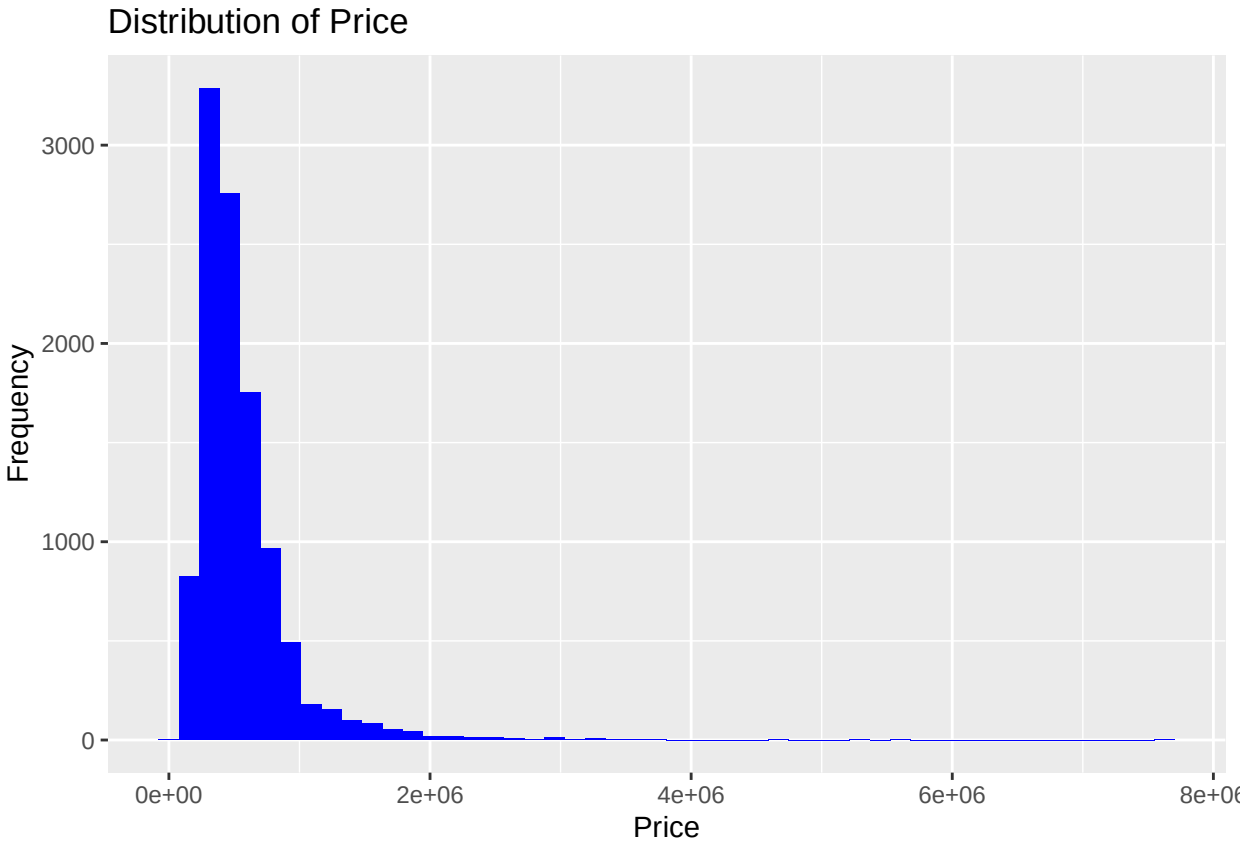
- *Response Variable*: `above_median`

- Motivation: This question seeks to find whether any of the variables mentioned above impact determining the qualities that could make a house worth more than the median housing prices in the King County region.

We are using variables that identify categorical aspects of a house. Therefore, any significant results may lead homeowners to focus not only on the size of their property but also on its design and internal properties to increase their potential earnings.

Section 4

Distribution of Price



The distribution is highly right-skewed, with a large frequency of lower-priced items and very few high-priced ones. The majority of prices of the houses fall close to the lower end of the price spectrum, which means that higher prices are outliers in this particular dataset.

Distribution of Bedrooms

```
##   bedrooms    n
## 1        0    3
## 2        1  103
## 3        2 1387
## 4        3 4907
## 5        4 3437
```

```
## 6          5  806
## 7          6  133
## 8          7   22
## 9          8    5
## 10         9    2
## 11        33    1
```

The distribution of bedrooms in the dataset shows that three-bedroom houses are the most common, followed by those with four bedrooms. Houses with 33 bedrooms is only one, along with those with 11 and 10 bedrooms, appear to be outliers. To maintain data integrity, we will exclude these outliers from further analysis.

Distribution of Bathrooms

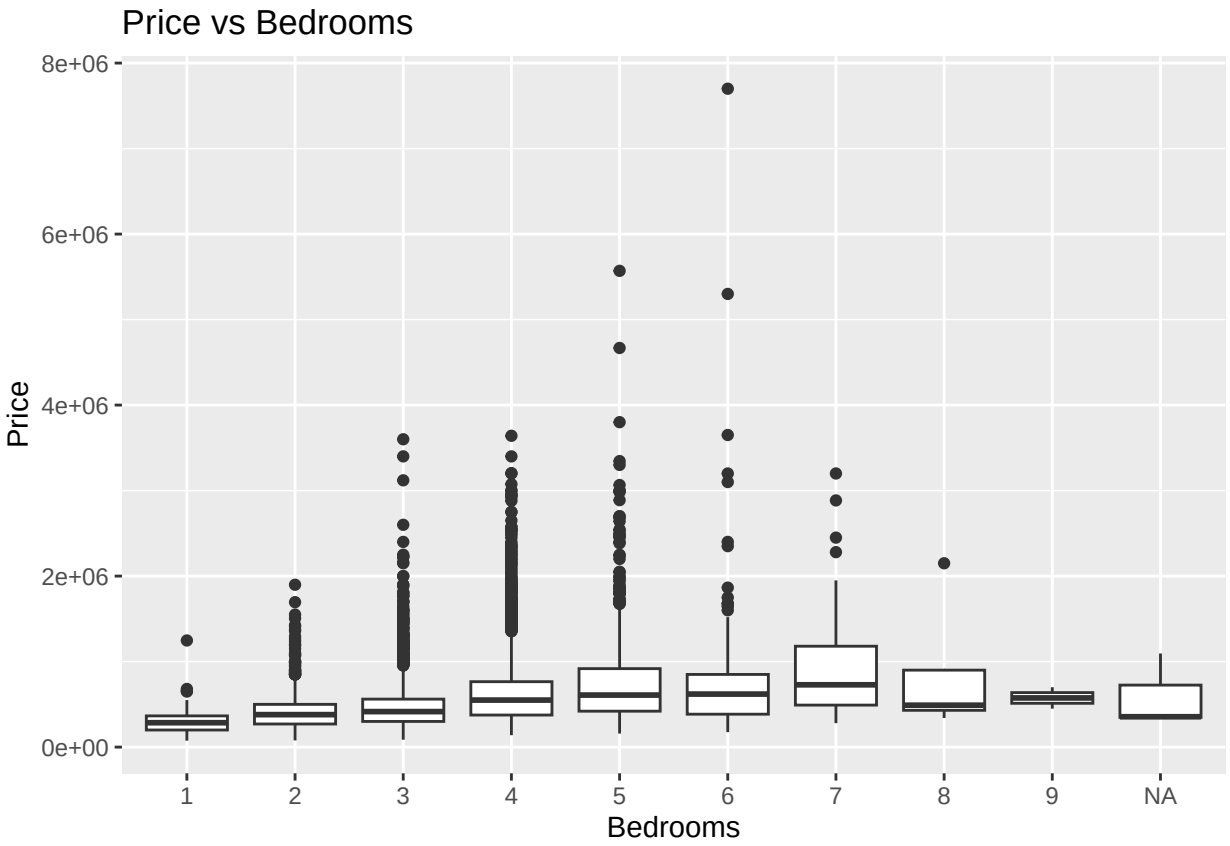
```
##    bathrooms    n
## 1       0.00    4
## 2       0.50    3
## 3       0.75   44
## 4       1.00 1925
## 5       1.25    5
## 6       1.50  680
## 7       1.75 1546
## 8       2.00  958
## 9       2.25 1036
## 10      2.50 2646
## 11      2.75  609
## 12      3.00  382
## 13      3.25  302
## 14      3.50  384
## 15      3.75   81
## 16      4.00   63
## 17      4.25   38
## 18      4.50   51
## 19      4.75   10
## 20      5.00   13
## 21      5.25    6
## 22      5.50    6
## 23      5.75    3
## 24      6.00    3
## 25      6.25    2
## 26      6.50    1
## 27      6.75    1
## 28      7.50    1
## 29      8.00    2
```

The distribution of bathrooms in the dataset shows that three-bathroom houses are the most common, followed by those with 3.5 bathrooms Houses with 8 bathrooms house listing is 2 one

Price vs Square feet Living

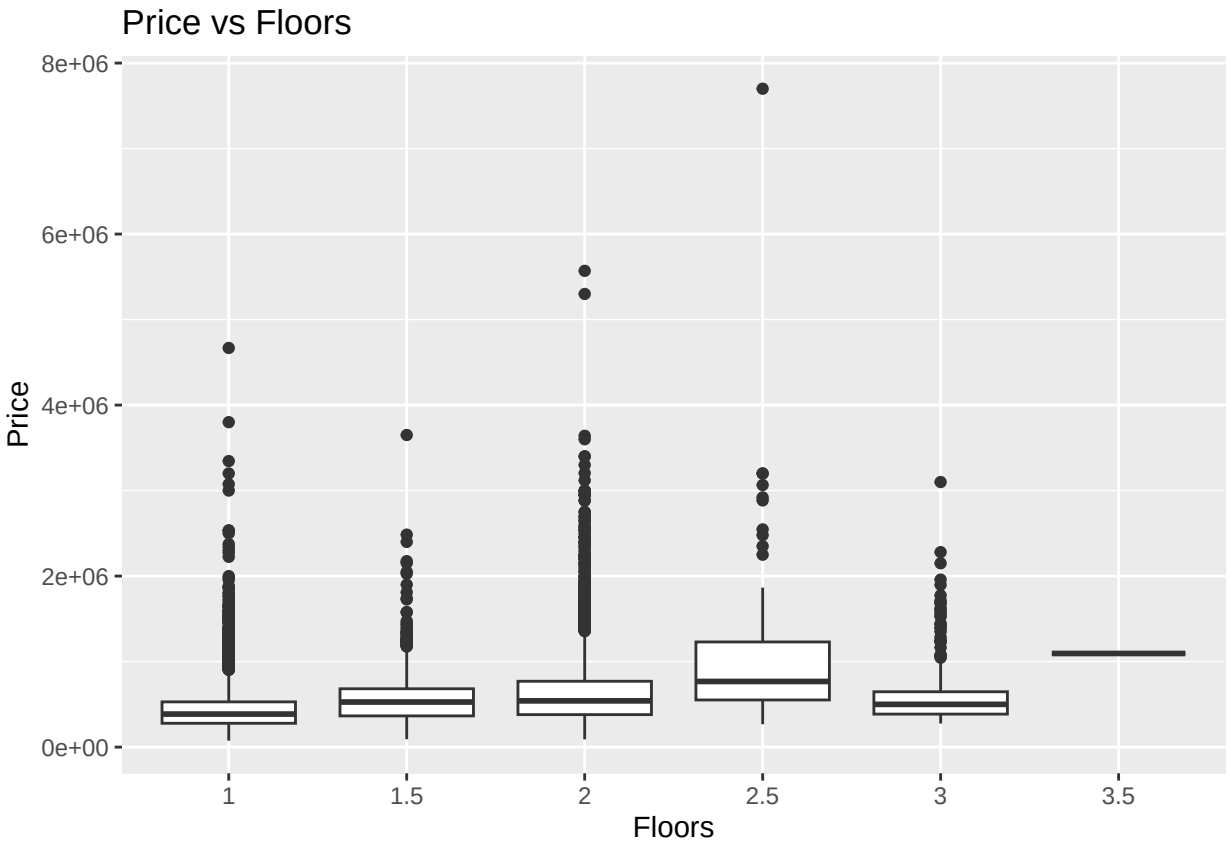
The scatter plot shows the correlation between square feet of living space and the cost of the apartments. The fact that we have an upward trend, revealing that the increase in price is proportional to the rise in the area, while it's a bit different from linear, is another proof of this cost-effectiveness. On the flip-side, there are some data points with big living area that is drastically more expensive than the rest of the listing thereby, showing these could be luxury or premium properties. These values are plotted far from the core of data and form a series of clusters that are anchored at specific points.

Prices vs Bedrooms



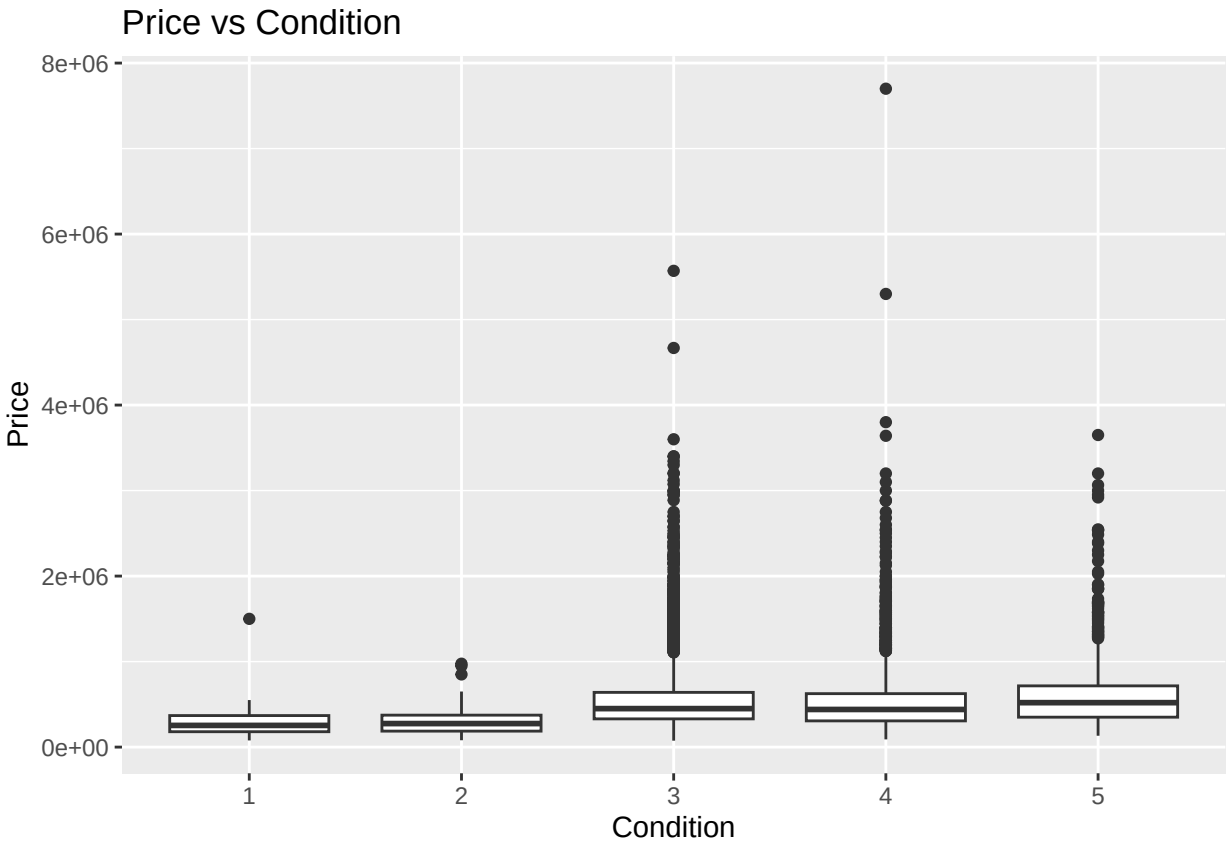
The boxplot represent the property prices distribution in line with the quantity of the bedrooms. The mu price continuously rises with the quality up to a number of bedrooms, after which it exactly fluctuates. The scatter plot of average prices sorted by category of bedrooms is such that the spread of each category's prices increase as the number of bedrooms increases, as shown by the longer boxes and whiskers, pointing to a bigger spread of property values. Particularly there are a great big number of outliers, especially in the upper area of rooms per property, which perhaps means that there are some properties, if many bedrooms have them, that are crushed against the mean of category, if it's in rooms.

Price vs Floors

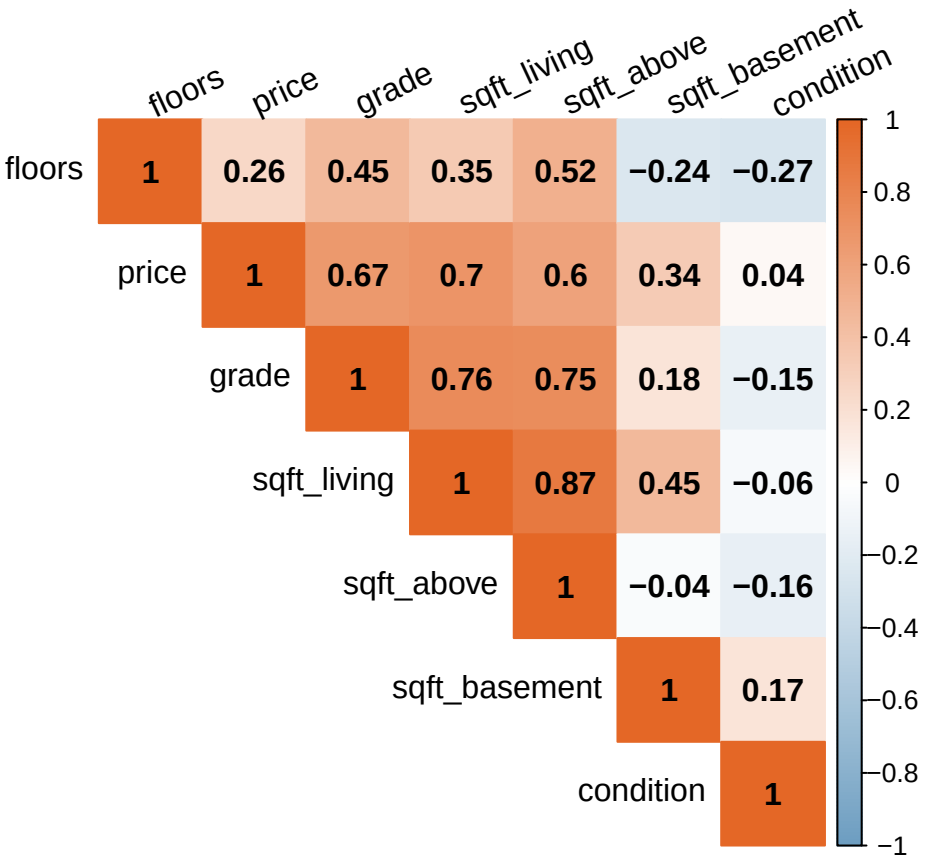


There are a significant number of community properties with 1 floor and the total price for them normally falls within a wide range. However, the median price for these types of properties is the cheapest among all the categories. Two-story boxes on the balancing beam point to the significantly larger median price, and a large part of the tale is the numerous outliers-obviously expensive homes. Residential units with one-and-half, two-and-half, and three floors have very small representation of the dataset and their price spectral is a bit uneven, but in particular, there are a few outliers among the 3-floor home residences. The scarce type of 3.5 floor homes have the highest median prices. The spread range here is narrow which implies that these house types have a pricing consistency. Outliers are present in all categories of floors, suggesting that rare or special geo-spatial attributes may be contributing to higher pricing.

Condition vs Price



Properties in conditions rated as 1 and 2 have a narrower range of prices with lower medians, suggesting they are generally less expensive. The median prices seem to increase slightly for properties rated in condition 3, and there's a wider spread of prices, with many outliers indicating some high-priced properties. Condition 4 and 5 properties also show a higher median price compared to lower-rated conditions, with condition 5 showing the most significant spread in prices, though not necessarily the highest median price. This could imply that while good condition may contribute to a higher price, other factors like location or size might also play a significant role in determining a property's value.



The Correlation Matrix, a data map created from your information, shows us the relationships between various objects. The strength of the relationship between two items is indicated by each number on this map. A value around one indicates that those variables typically move in together, i.e., as one increases, the other does too. It appears as though they move in different directions when it is close to -1. We begin to see patterns as we examine the map, such as the significant correlation between cost and size—larger houses typically cost more. However, there are also subtle connections that we can miss initially. This map allows us to identify instances where two items are overly similar, which aids in selecting the most relevant data for our investigation. There is a weak positive association between latitude and price, as seen by the correlation between latitude and price of 0.31 when we examine the relationship between latitude, price, and long price. Price and longitude have a weak link, as indicated by the 0.02 correlation between long and price. Comparable to a treasure map that leads us through our data, indicating which avenues to explore and which to keep clear of, enabling us to make informed decisions and reveal what's contained inside.

Section 5

The question is to what extent do internal housing factor affect the price of houses in King County, Washington? We are going to use linear regression to answer this question.

Our target variable is price and we are going to use all the variables in used to find the correlation between the price and other variables.

- Price Distribution: The broke in the price of housing is right skew, meaning that the many houses are distributed around lower price range, where the rest of expensive houses are located in the tail of the distribution. Premium homes, as shown within this data set, are outliers inferring.
- Bedroom Distribution: 3-bedroom houses are the biggest ones, while the number of houses is decreasing as the number of bedrooms is increasing in another one. Beyond enormous houses with more 8 bedrooms are not in usual cases and this is considered as an anomaly.

- Living Space and Price Correlation: A good correlation can be drawn between the living area size and the house price, however, such a relationship doesn't perfectly fit any specific pattern.
- Waterfront Rarity: Water frontage properties are indicating shortage; only 0.1% of the dataset being such, which might also justify higher price due to their rarity as a special commodity.
- Floor Preferences: Of the two predominant styles of homes, either single-story or two-story are typically the most prevalent. The fewer-floor homes of 3.5 storeys are almost as rare as hen's teeth, and this proves that many citizens like living in skyscrapers.
- View Quality: Its observed that although the number of home without any valuable view approximately equals the number of average(rated 0) views, the outcome of homes possessing higher-leveled (rated 1-2) view are far less.
- Condition and Price: Houses in better conditions are prone to satisfy the median price, however this rule is not always the case throughout the condition hierarchy.
- Year Built and Price: No very a strong correlation is present between property age and price which indicates that the more dominant factors might are some different other ones.

Data Transformation and Initial Model Fitting

Here, the house prices (`price`) are log-transformed to normalize the data, which is a common practice when the response variable is skewed. This helps improve the accuracy of linear regression models. An initial linear model is then fitted using a variety of predictors such as the number of bedrooms, bathrooms, square footage, and more. This step is aimed at understanding how these variables impact the log of house prices.

Regression Equation

$$price = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$$

```
##
## Call:
## lm(formula = log_price ~ bedrooms + bathrooms + sqft_living +
##     floors + condition + grade + sqft_above + sqft_basement,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77269 -0.23910  0.01061  0.22419  1.43403
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.119e+01  2.947e-02 379.822  < 2e-16 ***
## bedrooms      -3.506e-02  4.701e-03  -7.457 9.51e-14 ***
## bathrooms     -1.022e-02  7.256e-03  -1.409    0.159
## sqft_living    3.113e-04  9.627e-06  32.339  < 2e-16 ***
## floors         3.336e-02  4.097e-03   8.143 4.29e-16 ***
## condition      1.023e-01  5.323e-03  19.225  < 2e-16 ***
## grade          2.038e-01  4.652e-03  43.802  < 2e-16 ***
## sqft_above    -1.256e-04  9.612e-06 -13.064  < 2e-16 ***
## sqft_basement         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3417 on 10794 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.5783, Adjusted R-squared:  0.578
## F-statistic:  2115 on 7 and 10794 DF,  p-value: < 2.2e-16
```

Identifying and Removing Aliased Coefficients

Aliasing occurs when two or more variables in the model are linearly dependent or highly correlated, making it impossible to determine their individual effects. This code identifies any aliased coefficients in the initial model, helping to identify redundant predictors that should be removed to avoid multicollinearity issues.

```
## Aliased variables: sqft_basement
```

Remove Aliased Variables and Fit a Refined Model

Using the information from the previous step, this code creates a formula excluding the aliased variables and fits a refined model. This refined model should be more robust and provide clearer insights, as any linear dependencies among predictors have been removed.

```
##
## Call:
## lm(formula = log_price ~ bedrooms + bathrooms + sqft_living +
##     floors + condition + grade + sqft_above + sqft_basement,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77269 -0.23910  0.01061  0.22419  1.43403
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.119e+01  2.947e-02 379.822  < 2e-16 ***
## bedrooms      -3.506e-02  4.701e-03  -7.457 9.51e-14 ***
## bathrooms     -1.022e-02  7.256e-03  -1.409    0.159
## sqft_living    3.113e-04  9.627e-06  32.339  < 2e-16 ***
## floors         3.336e-02  4.097e-03   8.143 4.29e-16 ***
## condition      1.023e-01  5.323e-03  19.225  < 2e-16 ***
## grade          2.038e-01  4.652e-03  43.802  < 2e-16 ***
## sqft_above    -1.256e-04  9.612e-06 -13.064  < 2e-16 ***
## sqft_basement         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3417 on 10794 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.5783, Adjusted R-squared:  0.578
## F-statistic:  2115 on 7 and 10794 DF,  p-value: < 2.2e-16
```

Check for Multicollinearity Using VIF

Variance Inflation Factor (VIF) scores are calculated to assess multicollinearity, which occurs when predictor variables are highly correlated. High VIF scores indicate potential issues which can negatively impact the model's performance and interpretability.

```
## VIF Scores after resolving aliasing:
```

```
##    bedrooms   bathrooms sqft_living      floors   condition       grade
##    1.654165    2.939733    7.335675    1.818348    1.103454    2.780039
##  sqft_above
##    5.808483
```

Model Refinement - Interaction Terms and Removing Non-Significant Predictors

In this step, the model is updated to include an interaction term between `bedrooms` and `bathrooms`, while removing the individual `bedrooms` and `floors` predictors. Interaction terms allow the model to capture complex relationships between variables while removing non-significant predictors helps streamline the model.

Regression Equation

$$price = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

```
## Model diagnostics:

##
## Call:
## lm(formula = log_price ~ bathrooms + sqft_living + condition +
##     grade + sqft_above + bedrooms:bathrooms, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48430 -0.23790  0.00936  0.22552  1.38730
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.112e+01  2.751e-02 404.065  < 2e-16 ***
## bathrooms            4.696e-02  1.038e-02   4.525 6.10e-06 ***
## sqft_living          2.785e-04  9.048e-06  30.776  < 2e-16 ***
## condition            9.447e-02  5.281e-03  17.888  < 2e-16 ***
## grade                2.083e-01  4.704e-03  44.274  < 2e-16 ***
## sqft_above          -9.222e-05  8.747e-06 -10.544  < 2e-16 ***
## bathrooms:bedrooms -1.138e-02  1.946e-03  -5.850 5.06e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3432 on 10795 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.5747, Adjusted R-squared:  0.5744
## F-statistic:  2431 on 6 and 10795 DF,  p-value: < 2.2e-16
```

Checking Model Diagnostics

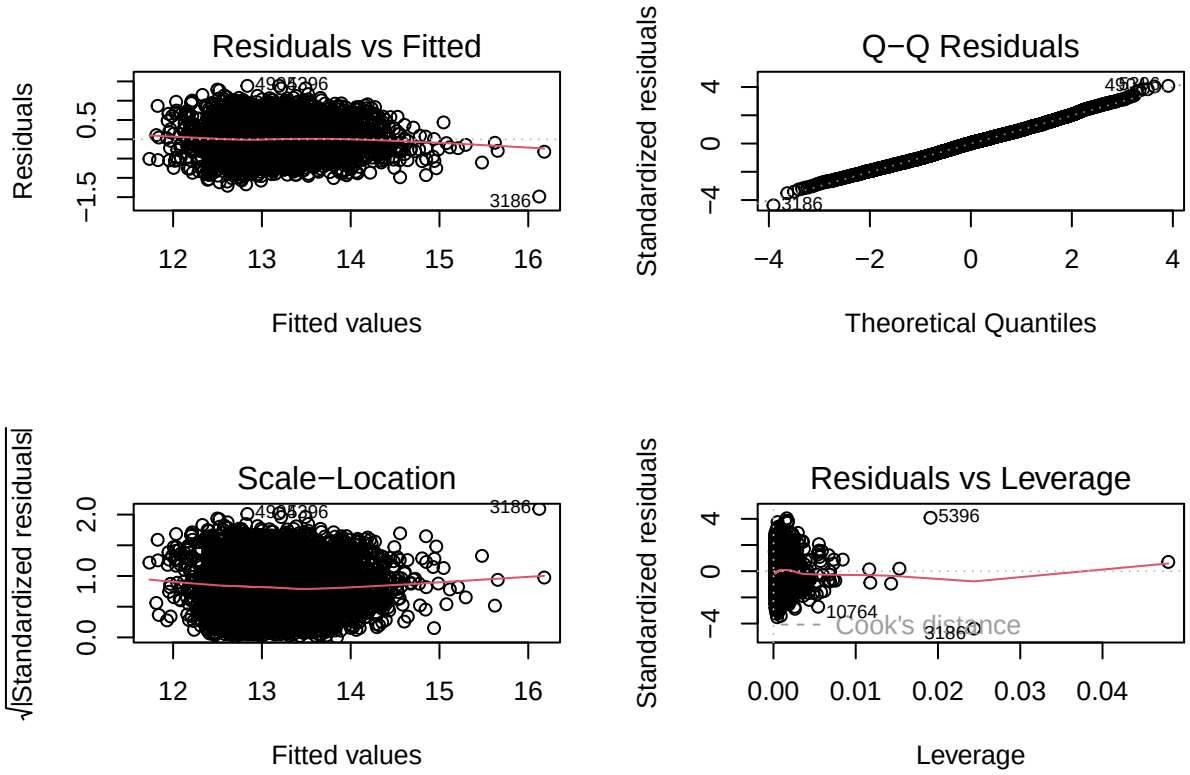
This code generates diagnostic plots for the refined model to evaluate its performance. The plots typically include residuals vs. fitted values, normal Q-Q plots, scale-location plots, and residuals vs. leverage plots. These plots help in assessing the assumptions of linear regression and identifying potential issues like non-linearity, heteroscedasticity, and influential points.

Residuals vs Fitted: This plot shows the residuals on the y-axis and the fitted values on the x-axis. It helps to check for non-linearity, unequal error variances, and outliers. If the points form a horizontal band, it indicates that the linearity assumption holds, while patterns or trends suggest model issues.

Q-Q (Quantile-Quantile) Plot: This plot compares the standardized residuals against a theoretical normal distribution. If the residuals follow a straight line, this suggests that they are normally distributed. Deviations from the line indicate departures from normality.

Scale-Location (or Spread-Location): This plot displays the square root of the standardized residuals against the fitted values. It helps to check for homoscedasticity (constant variance). A horizontal line indicates equal spread, while a fan or funnel shape suggests changing variance.

Residuals vs Leverage: This plot highlights influential data points. It shows the standardized residuals against leverage, a measure of how far each data point is from the average predictor value. Points with high leverage and high residuals can overly influence the model, indicated by Cook's distance lines.



Checking for Influential Observations, High Leverage Observations, and Outliers

This code identifies and removes influential observations that can skew the model's results. High leverage points and outliers are identified based on the "hat" values, which measure how far each observation's predictor values are from the average predictor values. Removing such influential points helps in achieving a more reliable model.

Re-Fitting the Model Without Influential Observations

After removing the influential observations, the model is refitted to the refined dataset. This step ensures that the model is not biased by any problematic data points.

```
##
## Call:
## lm(formula = log_price ~ bathrooms + sqft_living + condition +
##     grade + sqft_above + bedrooms:bathrooms, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.20247 -0.23852  0.00917  0.22617  1.37917
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)         1.112e+01  2.928e-02 379.791  < 2e-16 ***
## bathrooms           3.551e-02  1.112e-02   3.194  0.00141 **
## sqft_living         2.893e-04  9.816e-06  29.469  < 2e-16 ***
## condition           9.217e-02  5.573e-03  16.539  < 2e-16 ***
## grade               2.096e-01  5.009e-03  41.846  < 2e-16 ***
## sqft_above         -9.769e-05  9.328e-06 -10.472  < 2e-16 ***
## bathrooms:bedrooms -1.033e-02  2.152e-03  -4.800 1.61e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3433 on 10012 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.5587, Adjusted R-squared:  0.5584
## F-statistic:  2113 on 6 and 10012 DF,  p-value: < 2.2e-16
```

Fit the Final Model on Training Data

The final model is fitted using only the training data. This step ensures that the model learns the relationships between predictors and the response variable without being biased by the testing data.

Regression Equation

$$price = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

```
## Final model diagnostics:
```

```
##
## Call:
## lm(formula = log_price ~ sqft_living + grade + condition, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.19568 -0.24409  0.01076  0.22976  1.39019
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.112e+01  2.902e-02  383.36   <2e-16 ***
## sqft_living 2.065e-04  5.972e-06   34.57   <2e-16 ***
## grade       2.015e-01  4.652e-03   43.31   <2e-16 ***
## condition   1.026e-01  5.481e-03   18.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3456 on 10018 degrees of freedom
## Multiple R-squared:  0.5526, Adjusted R-squared:  0.5524
## F-statistic:  4124 on 3 and 10018 DF,  p-value: < 2.2e-16
```

Model Diagnostics on Final Model

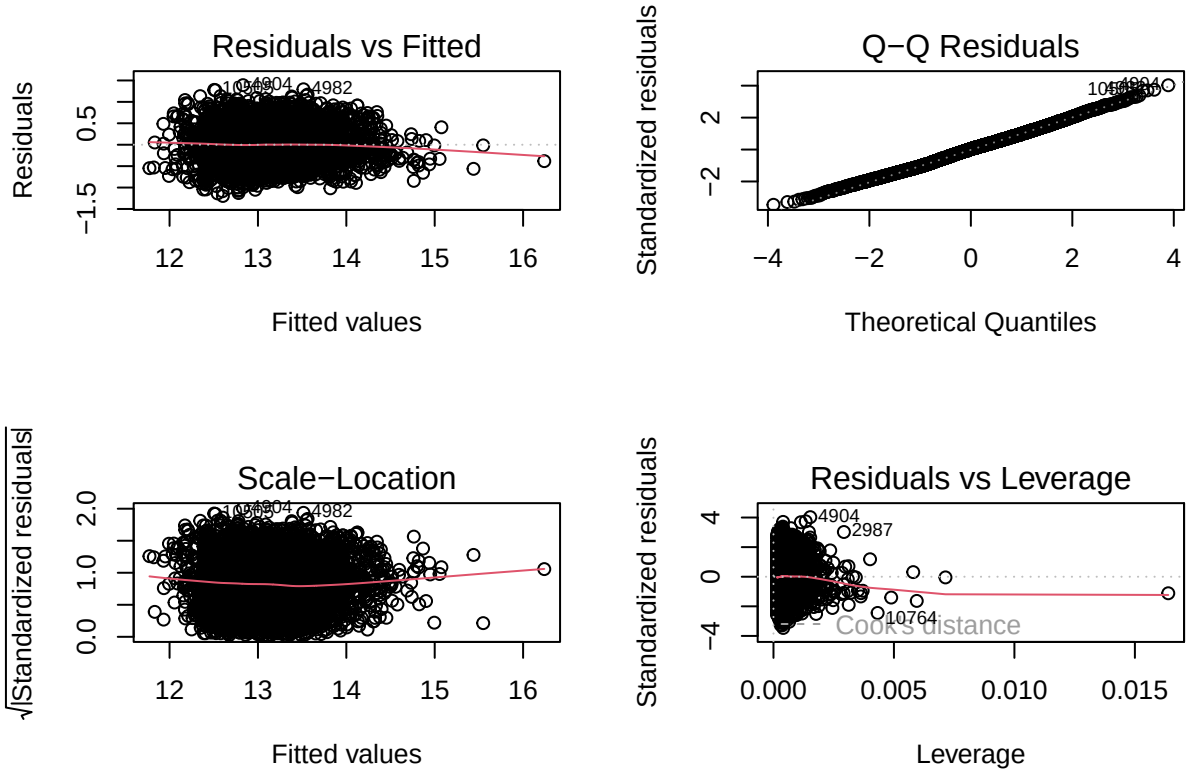
The final model's diagnostic plots and summary statistics are generated to evaluate its performance on the training data. These diagnostics help identify potential issues before moving to the testing phase.

Residuals vs Fitted: The plot looks acceptable, with no clear patterns or trends. The points are evenly distributed around zero, indicating that the model's assumptions are generally met. There's an outlier labeled "707306" that might be worth investigating.

Q-Q (Quantile-Quantile) Plot: The residuals align well with the diagonal, indicating that they are approximately normally distributed. This suggests that the normality assumption holds, although the point labeled "707306" again stands out slightly, hinting at a possible outlier.

Scale-Location (or Spread-Location): The plot is relatively horizontal, suggesting consistent variance across fitted values, although there is a slight trend visible. The same outlier is visible here as well.

Residuals vs Leverage: The plot shows that most points have low leverage. These might be influential points, given their positions near the Cook's distance threshold.



Evaluate Model Performance on Test Data

The final step involves evaluating the model's performance on the testing data. The root mean squared error (RMSE) and R-squared values are calculated to gauge the model's predictive accuracy and explanatory power. These metrics provide a clear picture of how well the model generalizes to new data.

RMSE on Test Data

The Root Mean Squared Error (RMSE) calculated on the test data is printed on this line. The average discrepancy between the expected and actual values is measured by RMSE. A model with lower RMSE values performs better in terms of prediction.

###R-squared The value of R-squared, calculated using the test data, is printed on this line. R-squared shows how much of the variance in the dependent variable 'log_price' can be accounted for by the model's independent variables (predictors). Greater R-squared values signify an improved model-data fit.

```
## RMSE on test data: NaN
```

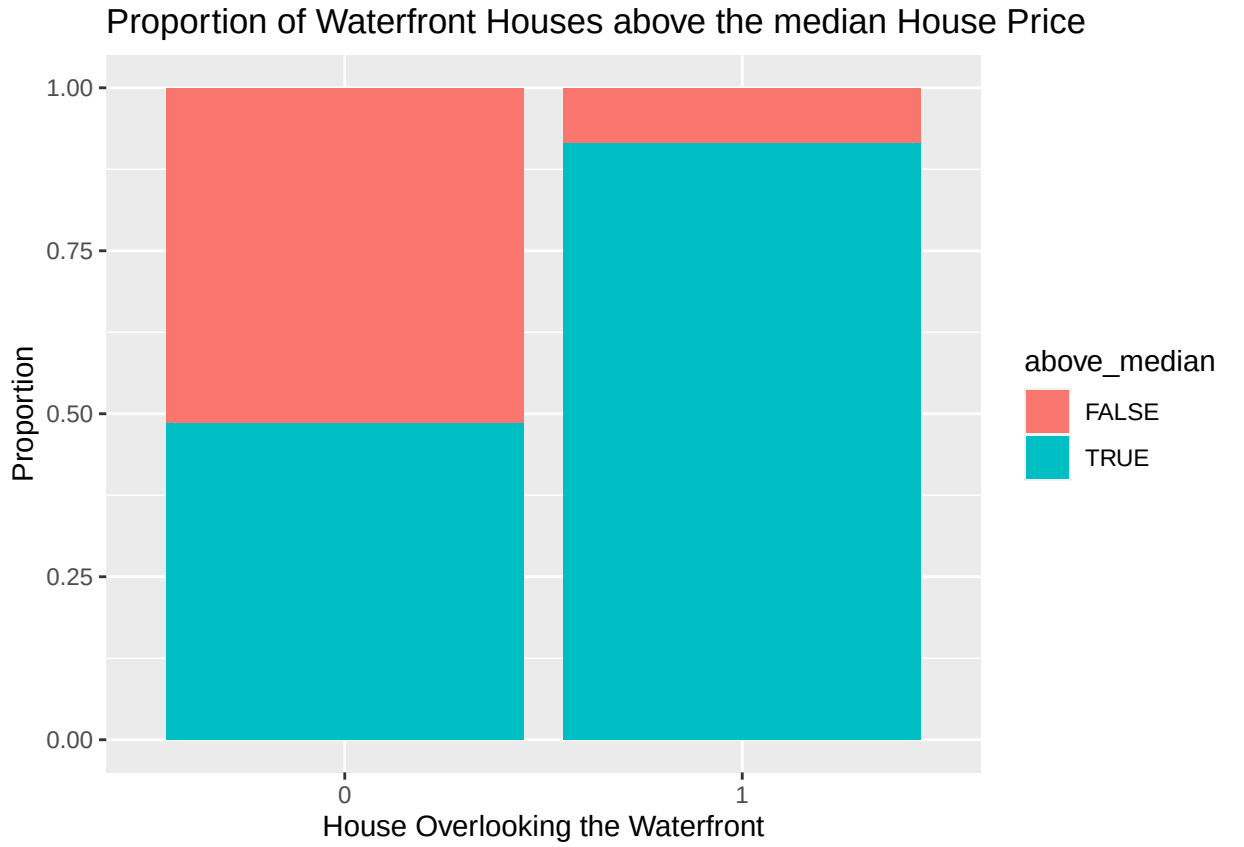
```
## R-squared: 0.552572
```

Section 6

Description of Variables

Waterfront

The waterfront variable is determining whether the house in question is located near the waterfront or not. We are inclined to believe this could be a relevant predictor because houses near water are expected to be more expensive. Consider the following visualization



There is a considerable proportion of houses overlooking the waterfront which are also above the median house price. To have a better idea of how many such houses we have in all, we also show a two-way table with the counts of houses belonging to each category

```
##
##     FALSE TRUE
##   0  5126 4826
##   1     6   64
```

Out of almost 11,000 entries of houses, it appears only 84 of them are overlooking the waterfront after all, which may lead to this variable not being too influential in the final model.

Condition

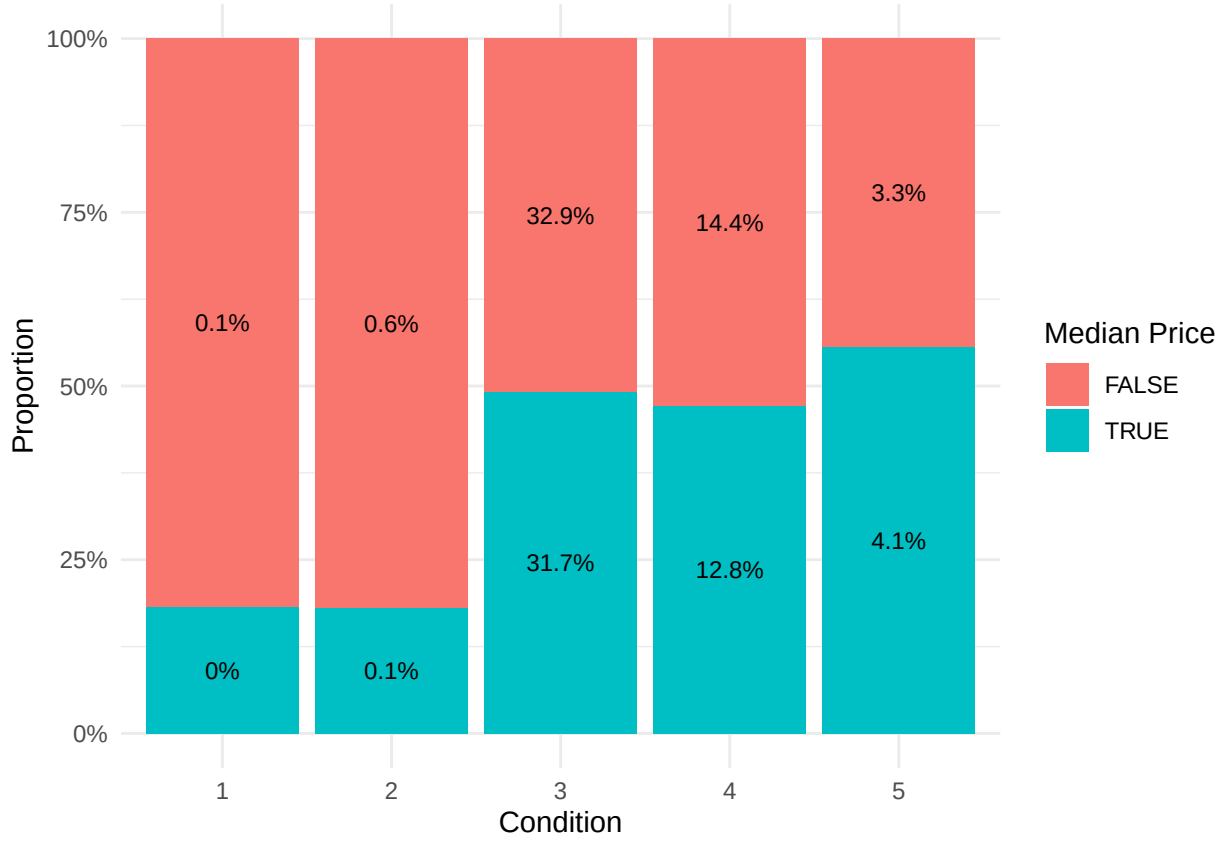
Condition is a self-explanatory categorical variable, in which it indicates the condition of an apartment in a scale of 1 to 5. Nicer apartments may be sold at a higher price than apartments which have lower ratings.

We wish to have an idea of the total number of houses for each condition rating. The following tables present a detailed summary of the number of entries for each condition level and their proportion

```
## # A tibble: 5 x 3
##   condition count percentage
##       <dbl> <int>      <dbl>
## 1         1    11      0.110
## 2         2    72      0.718
## 3         3  6477     64.6
## 4         4  2723     27.2
## 5         5   739      7.37
```

We see that very few houses have lower ratings, and most houses are in the median, 3, or above. These numbers express a strong relationship between the condition and a house being priced above the median.

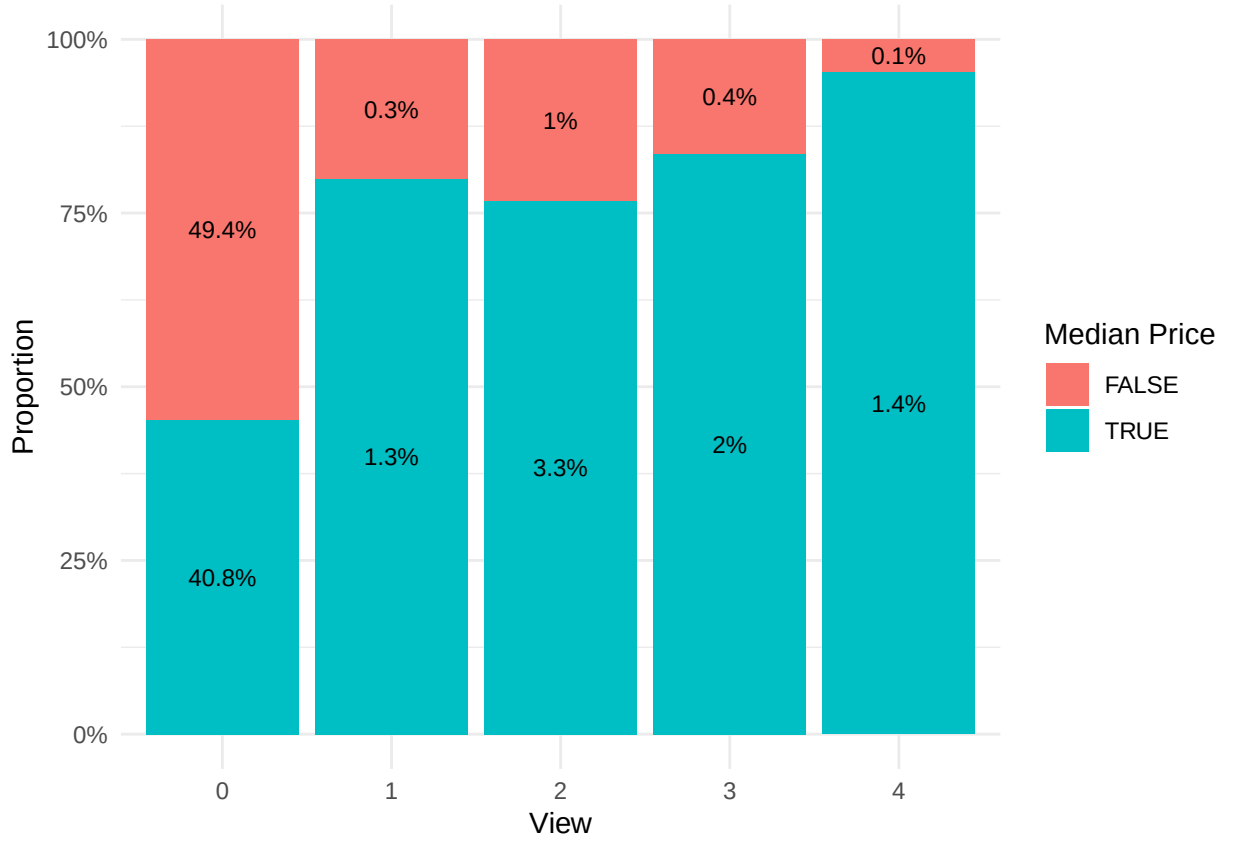
```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Indeed, we see a connection between the condition of a house and the house being above median price for the higher ratings. The proportion of houses which exceed the median increase as the rating increases.

View

The view variable rates on a scale of 0 to 4 how good the view of the property was. The median value for this variable is 0, which may indicate that only half of all apartments had a somewhat favorable view. We see a distribution of the proportions for each rating below



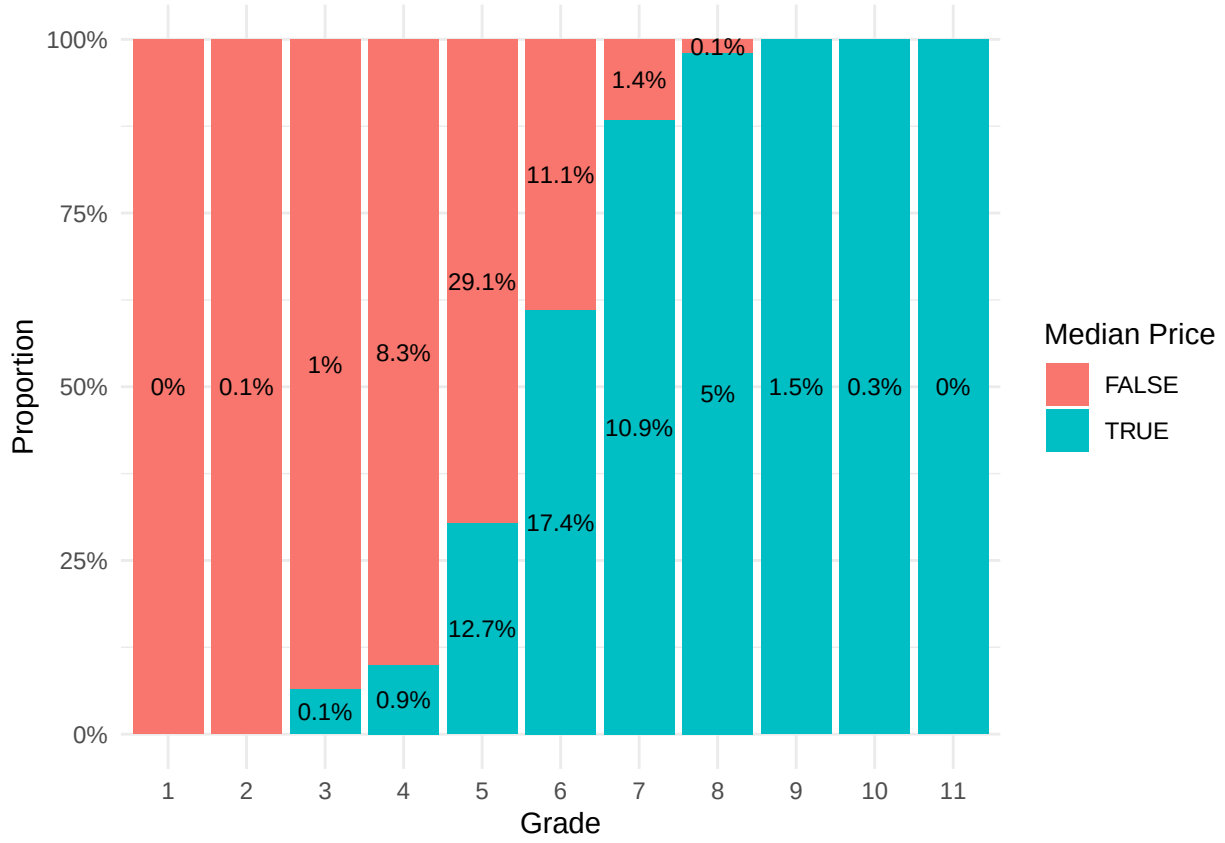
The proportion of houses above median value appears to increase as the view gets improved, and it still shows a high percentage when having a 0 rating after all. We present the counts from the table as follows

```
## # A tibble: 5 x 3
##    view count percentage
##   <int> <int>      <dbl>
## 1     0  9039      90.2
## 2     1   169       1.69
## 3     2   430       4.29
## 4     3   236       2.35
## 5     4   148       1.48
```

As mentioned before, a vast majority of houses have a view rating of 0, as we have almost 10,000 of the almost 11,000 entries having this rating. However, out of all houses with a higher view rating, over 75% of them were priced above the median across all four levels.

Grade

For the last categorical variable, grade is highlighting the quality level of the building's construction and design. It is indexed from 1 to 13, and houses with ratings 11-13 are considered to be of the highest quality. We see the proportion of houses being above or below the median price by grade.



We observe that there are no houses with grade levels 1 or 2. And that beginning from grade level 8, all subsequent levels have over 50% of the houses being above the median price. More importantly, this bar chart appears to highlight a positive relationship between the grade and being above the median price for houses.

Again, we present a two-way table to highlight the total amounts of houses over each grade level.

```
## # A tibble: 11 x 3
##    grade count percentage
##    <dbl> <int>      <dbl>
##  1     1     1    0.00998
##  2     2    13    0.130
##  3     3   109    1.09
##  4     4   925    9.23
##  5     5  4190   41.8
##  6     6  2851   28.4
##  7     7  1234   12.3
##  8     8   510    5.09
##  9     9   151    1.51
## 10    10    33    0.329
## 11    11     5    0.0499
```

We see that the vast majority of houses (it's over 9,000) have been given ratings between 6 and 9, which means that they have an average level of construction and design. Very few of them were given the highest quality level ratings.

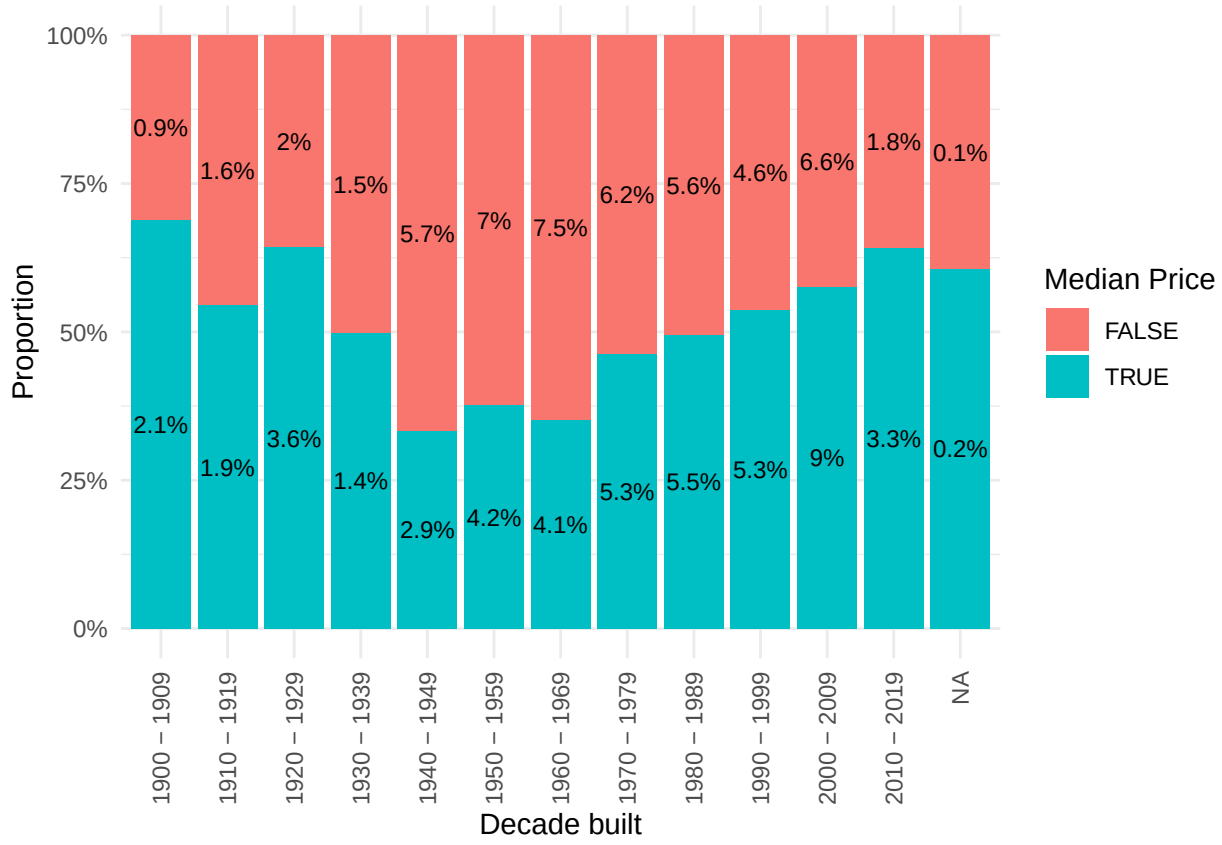
Year Built

We now present the first of two numerical predictor variables, and that is the year the house was first built. Given that this data set includes houses built from the early 1900s, we will split the houses into sets of

decades, starting with the 1900s, then the 1910s, and so on.

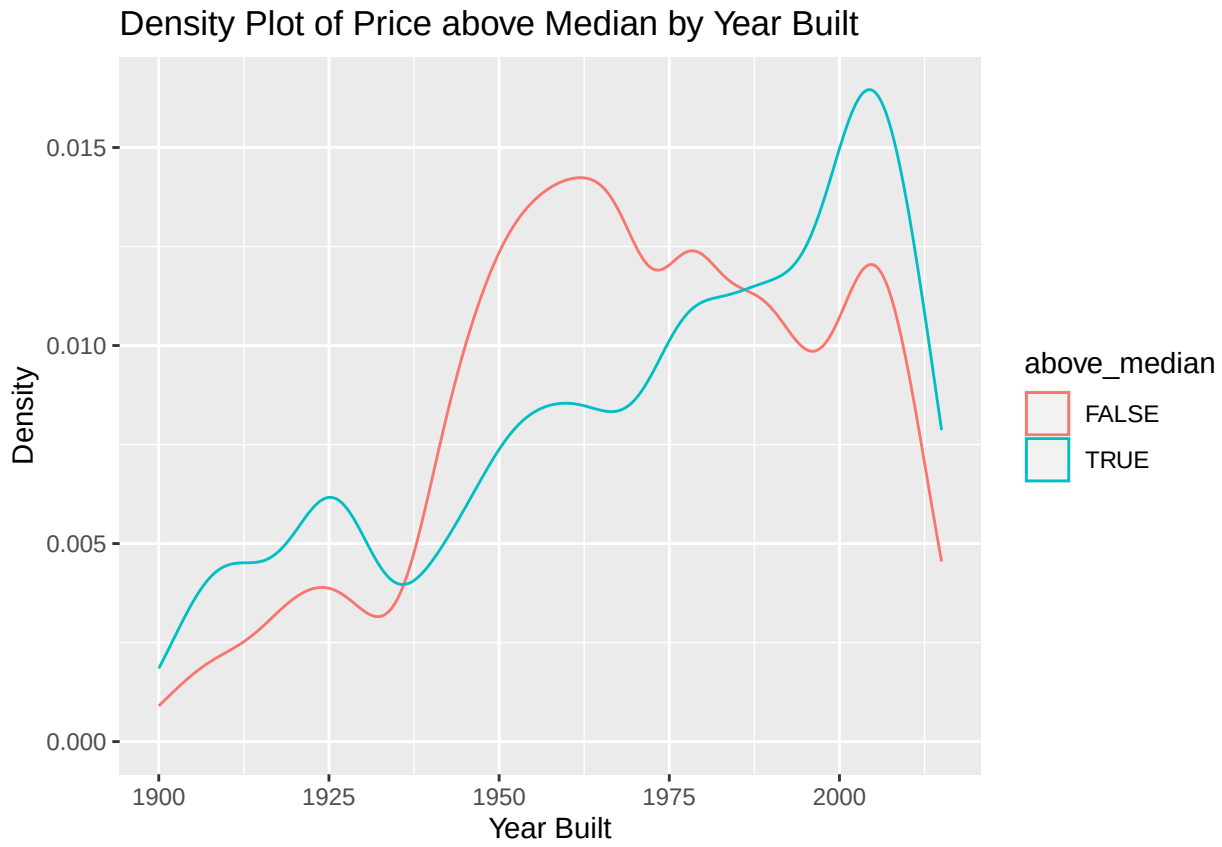
```
## # A tibble: 13 x 3
##    decade_built count percentage
##    <fct>        <int>      <dbl>
##  1 1900 - 1909    302      3.01
##  2 1910 - 1919    345      3.44
##  3 1920 - 1929    562      5.61
##  4 1930 - 1939    291      2.90
##  5 1940 - 1949    859      8.57
##  6 1950 - 1959   1120     11.2
##  7 1960 - 1969   1167     11.6
##  8 1970 - 1979   1151     11.5
##  9 1980 - 1989   1118     11.2
## 10 1990 - 1999   1000      9.98
## 11 2000 - 2009   1562     15.6
## 12 2010 - 2019    512      5.11
## 13 <NA>            33      0.329
```

From the above table, we can observe that a majority of the houses were built between the decades of 1950 and 2000, but otherwise, the decades the houses were built are very evenly spread, consisting of a fair proportion from every decade.



We can see a good proportion of house prices remain steadily above median until we enter the 1950s decade, where we experience a sudden drop on the total houses above the median from that era. Beginning from the 1950s, more and more house prices have been slowly making their way up above the median, but it is still unclear whether a linear relationship exists at all.

Finally, we will consider a density plot that considers the houses being above median price by year built.



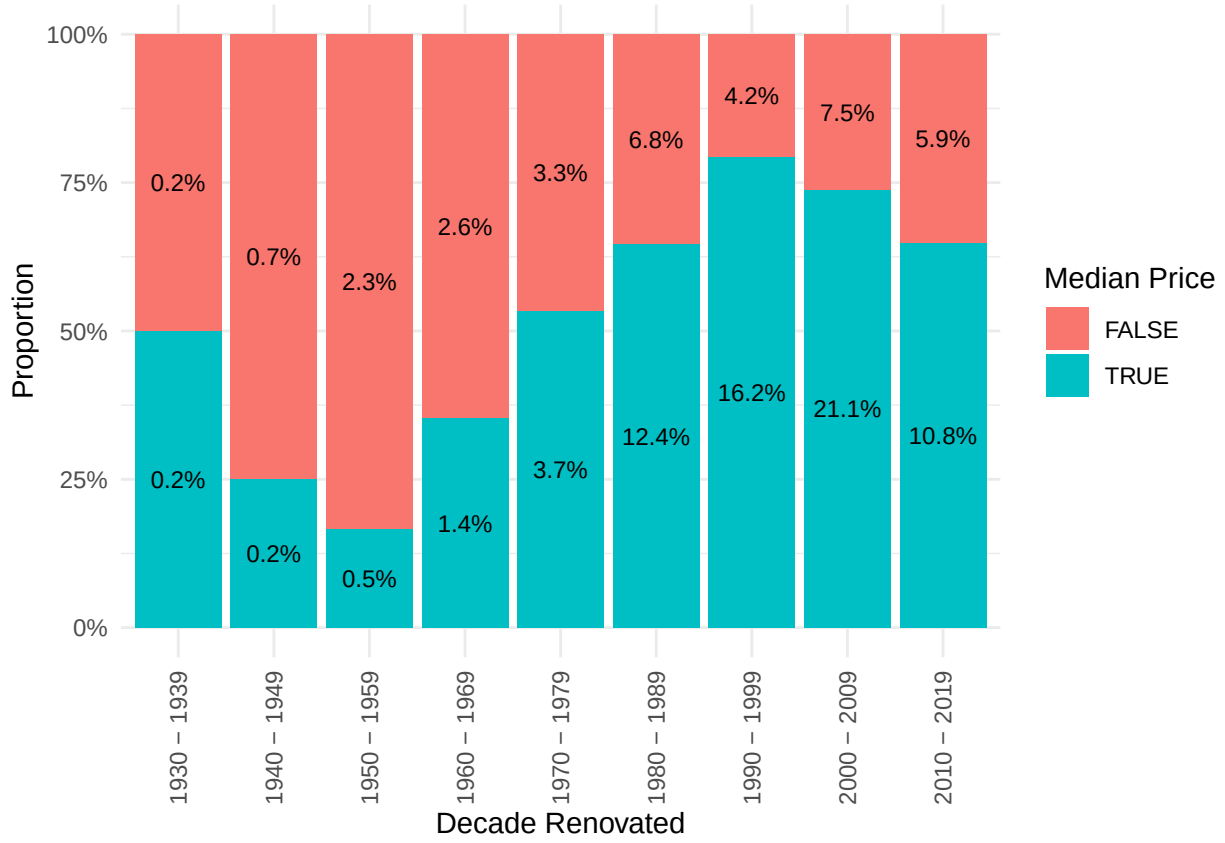
From the above graph, we see that most of the houses which are below median price were built between the 1940s and the 1970s. For houses which are above the median, they are found mostly around the 1920s decade, as well as the present day, with a majority found from 1980 to 2010.

Year Renovated

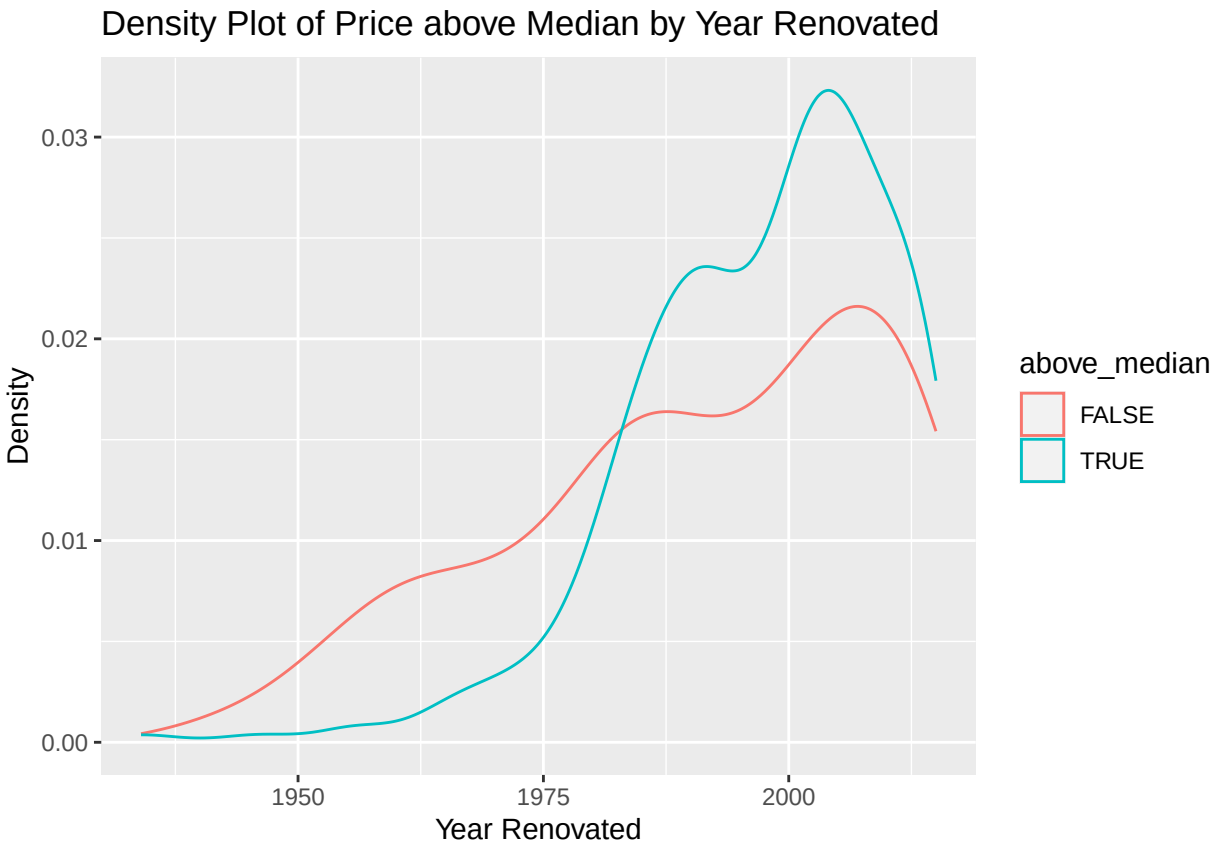
```
## # A tibble: 10 x 3
##    decade_renovated count percentage
##    <fct>            <int>      <dbl>
##  1 1930 - 1939          2     0.0200
##  2 1940 - 1949          4     0.0399
##  3 1950 - 1959         12     0.120
##  4 1960 - 1969         17     0.170
##  5 1970 - 1979         30     0.299
##  6 1980 - 1989         82     0.818
##  7 1990 - 1999         87     0.868
##  8 2000 - 2009        122     1.22
##  9 2010 - 2019         71     0.708
## 10 <NA>              9595    95.7
```

The table indicates that a staggering 95% of all houses in this data set have not been renovated yet, with less that 500 out of 10,806 houses having undergone this process. We are yet to determine whether this will prove to be a strong enough indicator during our analysis.

Also, out of the homes that were actually renovated, a good chunk of them were renovated from 1980 onwards.



After removing the values which are zero (indicating the house has never been renovated), we notice that, unlike the bar chart for year built, this bar chart provides a better look at how a house being renovated can positively impact its chances of being priced above median.



As mentioned earlier, we see that the majority of houses above the median price were renovated after the 1980s. Houses which are below the median follow a steady growth along the decades, with some of them peaking after the 2000s decade.

Section 7

Before we can start to work in the regression formula, we will define the response variable as a binary indicator. Specifically, our response's name will be `above_median`, and it will determine whether a house's price is above the median or not. The houses that are above the median will be labeled as $\hat{y} = 1$, and those which are not will be labeled $\hat{y} = 0$.

We want to assess our response variable with both numerical and categorical predictor variables. The numerical variables for this model are year built and year renovated, while the categorical variables we will use include waterfront, condition, view, and grade. Waterfront is already a binary indicator variable, but the latter three of these are labeled using indices ranging from 1 to 5, 0 to 4, and 1 to 13, respectively.

Given the nature of the categorical variables, there is no quantifiable way to measure the change from, say, a 9 to a 10 for the grade variable. Thus, instead of using the numbers from their given range, we will transform all these variables into indicator variables, where we determine whether the values of the data set are above or below the set value.

First, we find the median value of the condition variable. We know that it is 3, so variables who have a condition of 4 or 5 are going to be labeled as 1, while those which have a condition level of 1, 2, or 3 will be labeled as 0.

We will also compare the view variable, and it has a given range of 0 to 4. However, the median of this variable is 0, and as we saw before, over 89% of all houses were given a view rating of 0. So again the value of the indicator variable will be 1 if the view is greater than 0, and 0 otherwise.

Finally, we will transform the grade variable. According to the Kaggle website where we draw our data, the indices 11-13 suggest a high quality level for the design of a house. Taking this as the standard level, we reserve the value of 1 to those houses with grade level in this range, with 0 if they fail to attain it.

Our converted model will therefore contain four indicator variables, as well as two numerical variables, in order to determine whether a house is above or below the median price.

Regression Equation is as follows

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -11.36 + 0.005616x_1 + 0.000433x_2 + 0.5762I_1 + 0.1676I_2 + 1.563I_3 + 16.06I_4$$

We can interpret this logistic equation formula by considering each predictor variable's regression coefficient

(i) For every single year increase, the estimated probability that the house is above the median price is multiplied by $e^{0.005616} = 1.005632$, when controlling for all the other variables. This means that the newer the house is, the more likely it is to be above the median price.

(ii) For each additional year a house was last renovated, the estimated probability that the house will be above median price is multiplied by $e^{0.000433} = 1.000433$, when controlling for all the other variables. The more recent a house was renovated, the more likely it will be to be worth more.

(iii) The estimated probability that a house overlooking the waterfront is above the median price is $e^{0.5762} = 1.779264$ times the probability for houses that are not near the waterfront, when controlling for all the other variables. There may exist a link after all between houses near the waterfront and them having a higher price.

- We can now assess the coefficient for Waterfront using Wald test :
  - H_0 is β_1=0. H_a is $\beta_1 \neq 0$
  - Test statistic is Z = $\frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{0.2296}{0.4139} = 0.5547$
  - P value is 0.5791 which is more than test statistic and also much more than 0.05. This suggests we could drop the Waterfront predictor from our model, in presence of condition, grade, view, year built and year renovated.

(iv) The estimated probability that a house with above-median condition is above the median price is $e^{0.1676} = 1.182464$ times the probability for houses that are below the median in condition level, when controlling for all the other variables. The inside condition of a house determines a role in its price, as expected.

(v) The estimated probability that a house with above-median view is above the median price is $e^{1.563} = 4.773119$ times the probability for houses that are below the median in view level, when controlling for all the other variables. The view index of a house determines a role in its price.

(vi) Finally, according to the regression model, the estimated probability that a house with great grade level is above the median price is $e^{16.06} = 9435597$ times more likely than houses which are below the median, when controlling for all the other variables. This number appears to be very questionable.

We also notice that, since the standard error is very high, the predictor is not reliable and it might be recommended to drop the predictor in presence of all the others.

Model Assessment using Likelihood Ratio test

We now compare our model to determine whether it is useful in predicting whether a house's price is above the median better than random sampling. Let the null hypothesis be that $\beta_j = 0$ for $j = 1, 2, 3, 4, 5, 6$, and the alternative hypothesis be that at least one of the $\beta_j \neq 0$.

Consider a model with no predictors. Then we will compute the test statistic by finding the difference of the deviances between our full model and the no-predictor model.

$$\Delta G^2 = D(R) - D(F) = 14978.96 - 14054.81 = 924.1545$$

The p value from this test statistic is exactly 0. So we reject the null. The data supports our model over an intercept-only model.

Next, we want to check the individual predictor variables, and whether we can remove some of them if they are not contributing enough to the full model. We will start by considering whether we can remove the waterfront and the grade variables.

Let the null hypothesis be that $\beta_3 = \beta_6 = 0$; that is, that both predictors are not contributing any additional information in presence of the other variables. The alternative hypothesis is that at least one of β_3 or β_6 are not equal to 0. By carrying out the likelihood ratio test, we find that

$$\Delta G^2 = D(R) - D(F) = 14300.03 - 14054.81 = 245.2246$$

This has also a corresponding p-value of 0. So we reject the null. This means we cannot drop both waterfront and grade from our model. However, we can try to drop only the grade predictor, since it is the one that poses the more questionable results in the regression formula. We do this by using the Wald test.

Let $H_0 \colon \beta_6 = 0; H_a \colon \beta_6 \neq 0$.

$$Z = \frac{\hat{\beta}_6 - 0}{se(\hat{\beta}_6)} = \frac{16.06}{139.4} = 0.115$$

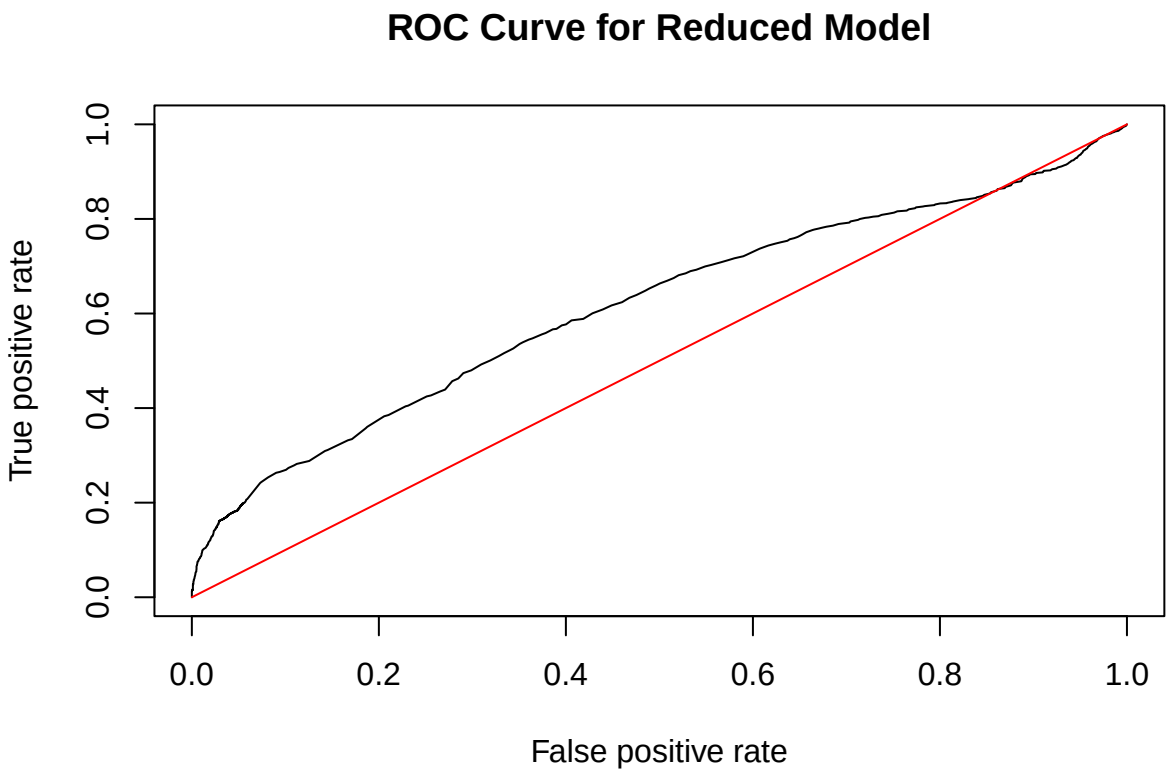
The corresponding p-value is 0.9084451. This means that we fail to reject the null hypothesis, and the data supports removing the grade predictor from the model. Our new logistic regression equation will be given by

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -13.15 + 0.006537x_1 + 0.000441x_2 + 0.1585I_1 + 1.653I_2 + 0.6518I_3$$

We check as well for any multicollinearity in our model by using the Variance Inflation Factors (VIF).

```
##      yr_built  yr_renovated condition_med      view_med    waterfront
##      1.272375      1.085757      1.212971      1.040974      1.037805
```

When we run the tests, we find that all the remaining predictor variables in our model are well below 5, with all of them being between 1 and 1.3. This indicates that there is little correlation between all predictors present.



The ROC curve above is above the diagonal for almost all values of the false positive rate, so it does better job than random guessing. We can also see that the point which maximizes the True Positive rate while minimizing the False Positive rate is around 0.3 for the FPR indicator.

Next, we take a look at the confusion matrix.

```
##
##          FALSE TRUE
##   FALSE   5092  309
##   TRUE    4313 1093
```

The sample size of our data is $n = 10807$. From the above table, we can find the following values:

The **error rate** is $\frac{536+3937}{10807} = 0.4138984$

The **accuracy** is $\frac{4865+1469}{10807} = 0.5861016$

The **false positive rate** is $\frac{536}{536+4865} = 0.09924088$

The **false negative rate** is $\frac{3937}{3937+1469} = 0.7282649$

The **true positive rate** is $\frac{1469}{3937+1469} = 0.2717351$

The **true positive rate** is $\frac{4865}{4865+536} = 0.9007591$

The **precision** is $\frac{1469}{1469+536} = 0.7326683$

We will now compute the Area Under the Curve (AUC), which also assesses whether our model is better at predicting the value response than just random sampling.

```
## [[1]]
## [1] 0.6149994
```

The AUC of our ROC curve is 0.613807, which means our logistic regression does better than random guessing.

We can now carry model diagnostic procedures by considering the three main approaches to model selection: forward selection, backward selection, and stepwise regression.