Project 2

Jack Hays, Sae'von Palmer, Varun Bhatnagar, Zach Karlovich - Group 2

2025-04-29

Section 1

Introduction

The real estate market is complex, with many factors influencing home prices and quality. This analysis examines housing data from King County, Washington, to understand what drives home prices and what makes a "good quality" home. Using data from over 21,000 homes sold between 2014 and 2015, we explore how various features - from square footage to location to renovation status - impact both the price and quality of homes.

Our analysis reveals several key insights that are valuable for both homebuyers and real estate professionals. First, we found that home prices are most strongly influenced by the total square footage, with waterfront properties and homes with better views resulting in higher prices.

These findings are particularly relevant for:

- Homebuyers looking to make informed purchasing decisions
- Real estate agents advising clients on property value
- Investors evaluating potential real estate in King County

Through statistical modeling and data visualization, we provide a comprehensive review of the King County housing market in 2014 and 2015.

Linear

From our linear model, we were able to derive a few relationships that should be useful for those looking to analyzing the King's County housing market. On a high level, our model implied many common sense relationships such as having a waterfront view or being of better quality (condition, grade, etc.) has an upwards price action. For homebuyers who are looking to estimate the price of a house given what they know about it, they can utilize this model to assist them.

Logistic

From utilizing a logistic regression, we delved into the relationship between a house's quality and how other variables impacted it. Our model implied that houses with a higher level of square feet, access to a waterfront, or a higher price were more likely to be of a higher quality while other year it was built, sold, or if the house had been renovated decreased the probability of it being

a higher quality. We believe that those looking to increase their chances of recognizing a higher quality house when it hits the market would be able to benefit from utilizing our model.

Section 2

Data Description

The dataset on House Sales in King County, USA[1] contains information about homes sold in King County, Washington from May 2014 to May 2015. The dataset in its initial state contains 21,613 observations and 21 variables.

For the purposes of this analysis, we will be focusing on `price` as the response variable and how it is impacted or related to the other variables in the dataset.

Table of Variables

The following is a table of the variables in the dataset as viewed from Kaggle[2]. The datatypes of the variables are also included.

[1] Accessible at: https://www.kaggle.com/ datasets/harlfoxem/housesalesprediction/data

[2] Accessible at: https://www.kaggle.com/datasets/harlfoxem/ housesalesprediction/discussion/207885

Variable	Description	Data Type
id	Unique ID for each home sold	Numeric
date	Date of the home sale	Character
price	Price of each home sold	Numeric
bedrooms	Number of bedrooms	Integer
bathrooms	Number of bathrooms, where .5 = toilet only	Numeric
sqft_living	Living space (square feet)	Integer
sqft_lot	Land space (square feet)	Integer
floors	Number of floors	Numeric
waterfront	Dummy variable for waterfront property	Integer
view	Rating (0-4) of view quality	Integer
condition	Rating (1-5) of apartment condition	Integer
grade	Rating (1-13) of construction/design quality	Integer
sqft_above	Square footage above ground	Integer
sqft_basement	Square footage below ground	Integer
yr_built	Year the house was built	Integer
yr_renovated	Year of last renovation	Integer
zipcode	Zipcode area	Integer
lat	Latitude	Numeric
long	Longitude	Numeric
sqft_living15	Living space of 15 nearest neighbors (sq ft)	Integer
sqft_lot15	Lot size of 15 nearest neighbors (sq ft)	Integer

Numeric, Integer, and Character Variables

The dataset consists of numeric, integer, and character data types. We can see that the `date` variable is a character, which indicates that the date of the sale is captured as a string.

Much of the data set is continuous, with the exception of the `date` variable and several of the categorical variables.

Categorical Variables

The variable `waterfront` is dummy coded, with a value of 1 indicating that the home has a waterfront view.[3]

[3]Source: King County Government Assessor's Office, https://info.kingcounty.gov/assessor/esales/Glossary.aspx

We can also see that the variables `view`, `condition`, and `grade` are all ordinal variables that are indexed according to their quality. In each case, the index is an integer value where the higher the index, the higher the quality.

View

According to King County[4], `view` is classified in the following manner:

- 0: No view
- 1: Fair
- 2: Average
- 3: Good
- 4: Excellent

Condition

According to King County[5], `condition` is classified in the following manner:

- 1: Poor; worn out
- 2: Fair; badly worn
- 3: Average; evidence of deferred maintenance
- 4: Good; no obvious maintenance issues
- 5: Very Good; well maintained condition

Grade

According to King County[6], the `grade` variable represents the overall grade given to the house based on the King County grading system:

- 1-3: Falls short of minimum building standards
- 4: Generally older, low quality construction
- 5: Low construction costs and workmanship
- 6: Lowest grade meeting building code
- 7: Average grade of construction
- 8: Just above average construction and design
- 9: Better architectural design
- 10: High quality features with better finish work
- 11: Custom design with higher quality finish work
- 12: Custom design with excellent builders and highest quality materials
- 13: Custom designed and built mansion level with highest quality finishes

Additional Variables

From the `date` variable, we can create several new variables that can be used to explore the data set.

- `yr_sold`: The year of the sale

[4]Source: King County Government Assessor's Office, https://info.kingcounty.gov/assessor/esales/Glossary.aspx
[5]Source: King County Government Assessor's Office, https://info.kingcounty.gov/assessor/esales/Glossary.aspx
[6]Source: King County Government Assessor's Office, https://info.kingcounty.gov/assessor/esales/Glossary.aspx

- `month_sold`: The month of the sale
- `day_sold`: The day of the sale
- `season_sold`: The season of the sale

We can also calculate the total square footage of the property by adding the `sqft_living` and `sqft_lot` variables to get the combined interior living space and lot size. Using this total, we can calculate the price per square foot of the entire property.

```
  yr_sold month_sold day_sold season_sold total_sqft price_per_sqft
1    2014         10       13        Fall       6830          32.49
2    2014         12        9      Winter       9812          54.83
3    2015          2       25      Winter      10770          16.71
4    2014         12        9      Winter       6960          86.78
5    2015          2       18      Winter       9760          52.25
```

Section 3

Data Errors Found

1. One house has 33 bedrooms. This is likely a data entry error and should be removed from the data set. As a result, we removed this observation from our set at the end of this section.
2. We can also see that there are a number of houses that are sold before they are built. These may be data entry errors or they may be houses that are sold before they are built. The price/quality of these houses are not representative of typical situation for houses, thus we remove them from the dataset. We clean these houses at the end of this section.
3. Looking at high leverage points, we can see a house with over 1 million square footage but is below 190,000. We remove this later in section five in order clean the linear regression model.

Data Cleaning

After examining our data structure, we can see several potential issues that need to be addressed.

1. The `date` column is currently a character vector with a format "20141013T000000" making it difficult to work with.
2. The basement and year renovated columns can be cleaned and mutated into two new categorical variables to make analysis easier.
3. The `waterfront`, `view`, `condition`, and `grade` columns are currently numeric but should be factors.

Date Format

The first correction to be made is to the `date` column which is currently a character vector with a format "20141013T000000". This should be converted to a Date data type with the format (YYYY-MM-DD). This will enable us to better handle date related operations.

```
[1] "2014-10-13"
```

Basement and Year Renovated

Next, we will create two new categorical variables `basement` and `renovated` to indicate whether a house has a basement or has been renovated. We will also set `yr_renovated` to `NA` for houses that have never been renovated (`yr_renovated = 0`).

Categorical and Ordinal Variables

We can also factor our categorical and ordinal variables `waterfront`, `view`, `condition`, and `grade`.

```
 waterfront         view           condition         grade
 No :21450   No view  :19489   Poor     :   30   7      :8981
 Yes:  163   Fair     :  332   Fair     :  172   8      :6068
             Average  :  963   Average  :14031   9      :2615
             Good     :  510   Good     : 5679   6      :2038
             Excellent:  319   Very Good: 1701   10     :1134
                                                 11     : 399
                                                 (Other): 378
```

Data Entry Error Cleaning

Looking at the summary for bedrooms we can see that there is a house with 33 bedrooms. This is likely a data entry error and should be removed from the data set.

```
    bedrooms
 Min.   : 0.000
 1st Qu.: 3.000
 Median : 3.000
 Mean   : 3.371
 3rd Qu.: 4.000
 Max.   :33.000
```

We'll perform some additional sanity checks on the data to clean up any other potential errors.

Created Variables and Cleaned Data

After cleaning the data and creating the new variables, it takes on the following structure:

```
tibble [21,612 × 29] (S3: tbl_df/tbl/data.frame)
 $ id           : num [1:21612] 7.13e+09 6.41e+09 5.63e+09 2.49e+09 1.95e+09 ...
 $ date          : Date[1:21612], format: "2014-10-13" "2014-12-09" ...
 $ price         : num [1:21612] 221900 538000 180000 604000 510000 ...
 $ bedrooms      : int [1:21612] 3 3 2 4 3 4 3 3 3 3 ...
 $ bathrooms     : num [1:21612] 1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
 $ sqft_living    : int [1:21612] 1180 2570 770 1960 1680 5420 1715 1060 1780
1890 ...
 $ sqft_lot     : int [1:21612] 5650 7242 10000 5000 8080 101930 6819 9711 7470
```

```
6560 ...
 $ floors        : num [1:21612] 1 2 1 1 1 1 2 1 1 2 ...
 $ waterfront    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ view          : Ord.factor w/ 5 levels "No view"<"Fair"<..: 1 1 1 1 1 1 1 1
1 1 ...
 $ condition     : Ord.factor w/ 5 levels "Poor"<"Fair"<..: 3 3 3 5 3 3 3 3 3
3 ...
 $ grade         : Ord.factor w/ 13 levels "1"<"2"<"3"<"4"<..: 7 7 6 7 8 11 7 7
7 7 ...
 $ sqft_above    : int [1:21612] 1180 2170 770 1050 1680 3890 1715 1060 1050
1890 ...
 $ sqft_basement : int [1:21612] 0 400 0 910 0 1530 0 0 730 0 ...
 $ yr_built      : int [1:21612] 1955 1951 1933 1965 1987 2001 1995 1963 1960
2003 ...
 $ yr_renovated  : int [1:21612] NA 1991 NA NA NA NA NA NA NA NA ...
 $ zipcode       : int [1:21612] 98178 98125 98028 98136 98074 98053 98003 98198
98146 98038 ...
 $ lat           : num [1:21612] 47.5 47.7 47.7 47.5 47.6 ...
 $ long          : num [1:21612] -122 -122 -122 -122 -122 ...
 $ sqft_living15 : int [1:21612] 1340 1690 2720 1360 1800 4760 2238 1650 1780
2390 ...
 $ sqft_lot15    : int [1:21612] 5650 7639 8062 5000 7503 101930 6819 9711 8113
7570 ...
 $ basement      : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 1 1 2 1 ...
 $ renovated     : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 1 ...
 $ yr_sold       : num [1:21612] 2014 2014 2015 2014 2015 ...
 $ month_sold    : num [1:21612] 10 12 2 12 2 5 6 1 4 3 ...
 $ day_sold      : num [1:21612] 13 9 25 9 18 12 27 15 15 12 ...
 $ season_sold   : chr [1:21612] "Fall" "Winter" "Winter" "Winter" ...
 $ total_sqft    : int [1:21612] 6830 9812 10770 6960 9760 107350 8534 10771
9250 8450 ...
 $ price_per_sqft: num [1:21612] 32.5 54.8 16.7 86.8 52.2 ...
```

Cleaning up pre-sold Houses

We can also see that there are a number of houses that are sold before they are built. These may be data entry errors or they may be houses that are sold before they are built. The price/quality of these houses are not representative of typical situation for houses, thus we remove them from the dataset.

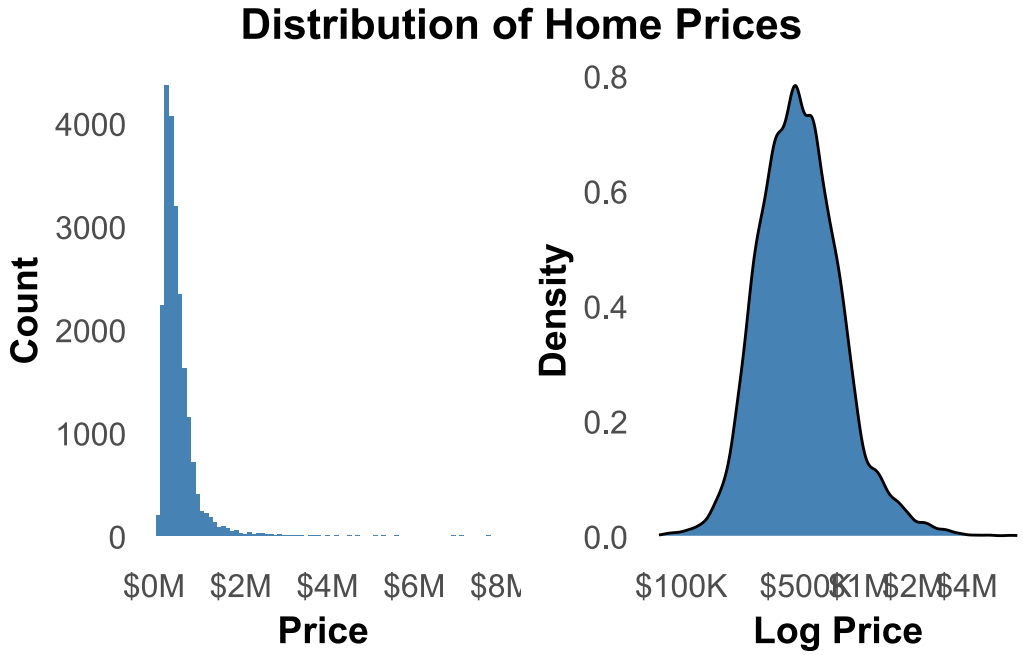
Section 4

Visualizations - Home Price vs. Other Variables

We will use visualizations to gain a deeper understanding of housing prices and their relationships with other variables in the dataset.

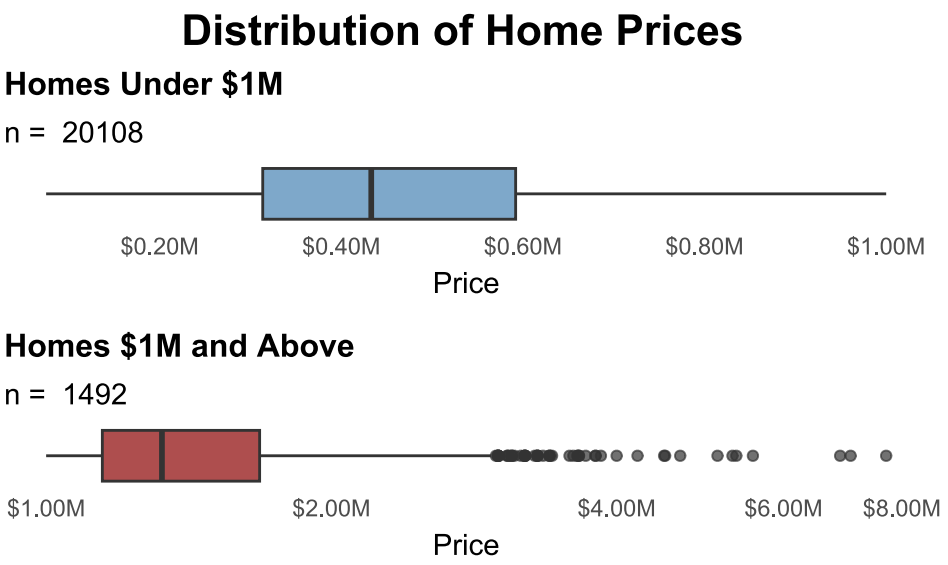
Univariate Visuals

First, we will examine the distributions of our key numeric variables. This will help us understand the range and concentration of prices and other features in the King County housing market.



The left plot shows the distribution of home prices in the dataset. The right plot shows the distribution of home log prices in the dataset. The log transformation shows a more normal distribution of home prices. This will be useful for modeling.

Given the right skewed nature of the price distribution, we're going to look at two separate box plots to get a little more structure at both price ranges. We'll first look at pricing between $0 and $1 million and houses greater than $1 million.

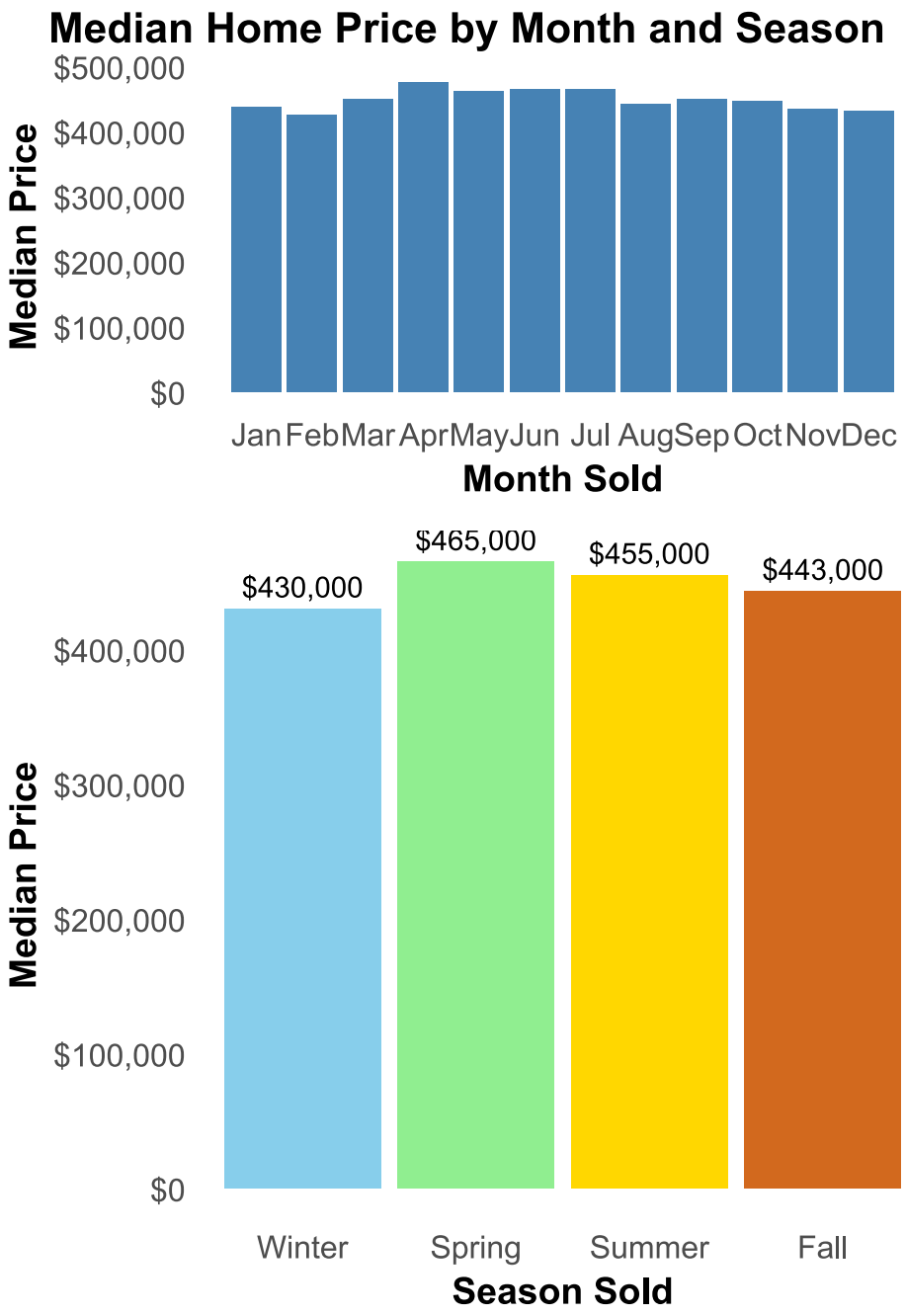


From the two plots, we can clearly see that the bulk of the homes have a median price of around $430,000, with the middle 50% roughly between $300,000 and $600,000. The box plot for homes above $1,000,000 still has a very pronounced right-hand tail of outliers.

Bivariate Visuals

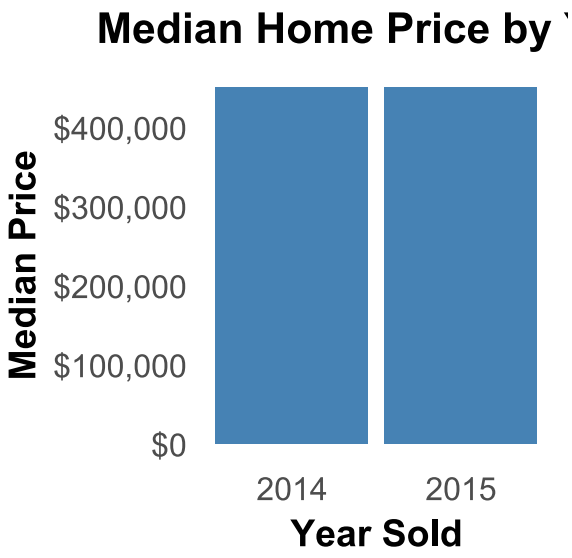
Next, we will explore relationships between home prices and individual features. These visualizations will help us understand how different characteristics of a house relate to its price.

Since we have time data on the pricing, let's look at the time series data for median prices sold by month and then median prices sold by season to see if there are any trends.

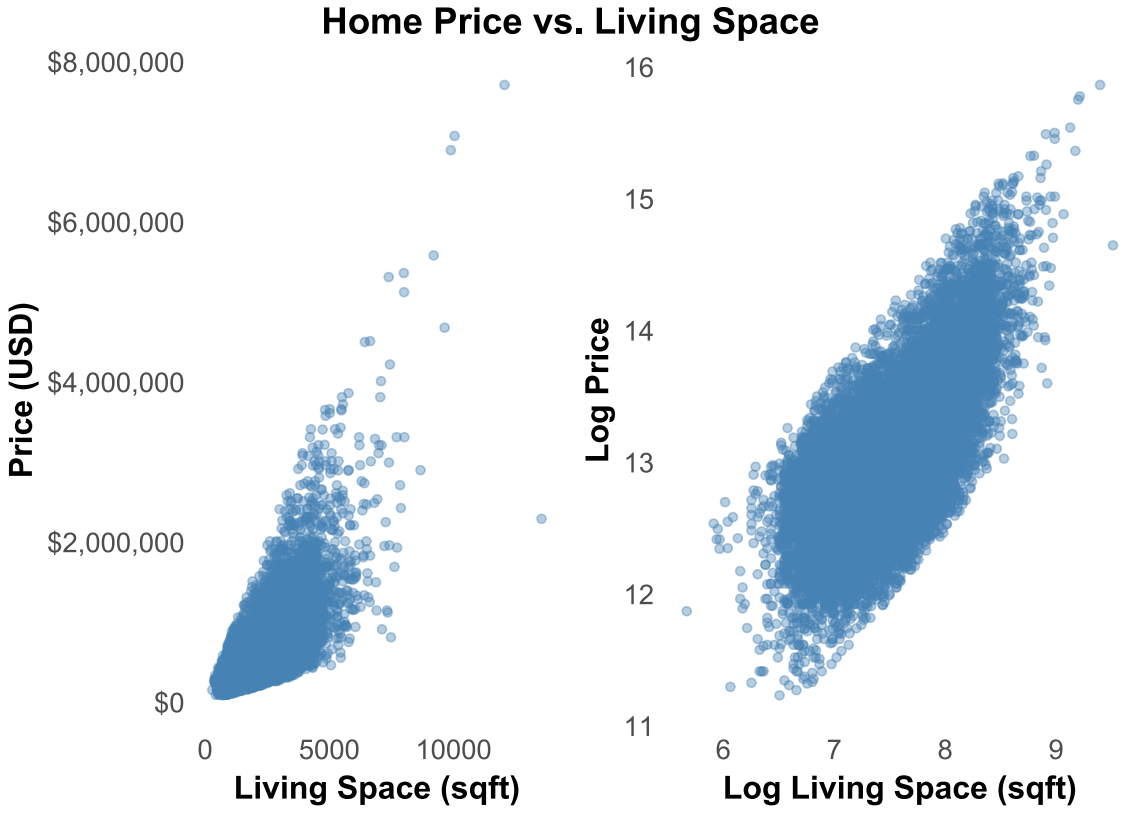


From the plots, we can see that the median price of homes is highest in the summer months and lowest in the winter months. This is likely due to the fact that people are more likely to buy homes in the summer when they may have a greater flexibility in their moving plans. The seasonal plot shows this trend more clearly, with the highest median price in the spring/summer and lowest in the fall/winter.

Next, we will look at the relationship between home prices and the year the home was sold.

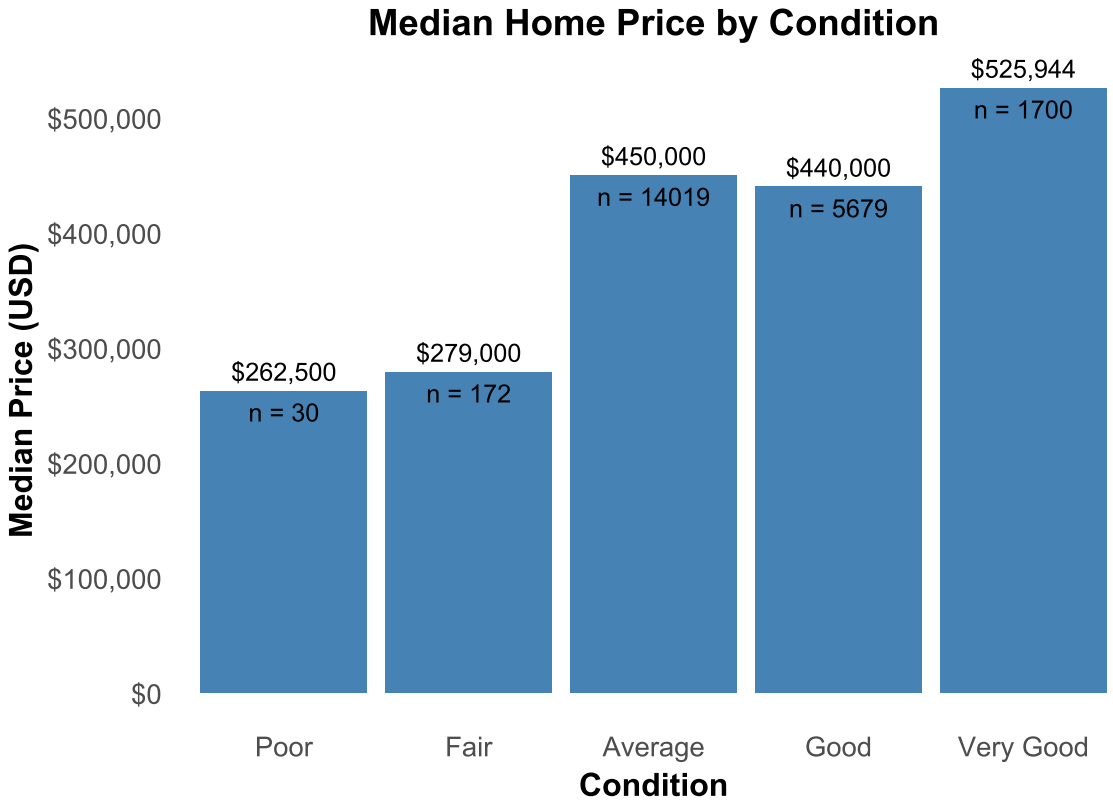


There is a minimal difference between 2014 and 2015 median home prices. One of the drivers in housing prices is the amount of square footage of the home. Let's look at the relationship between home price and the liveable square footage of the home.



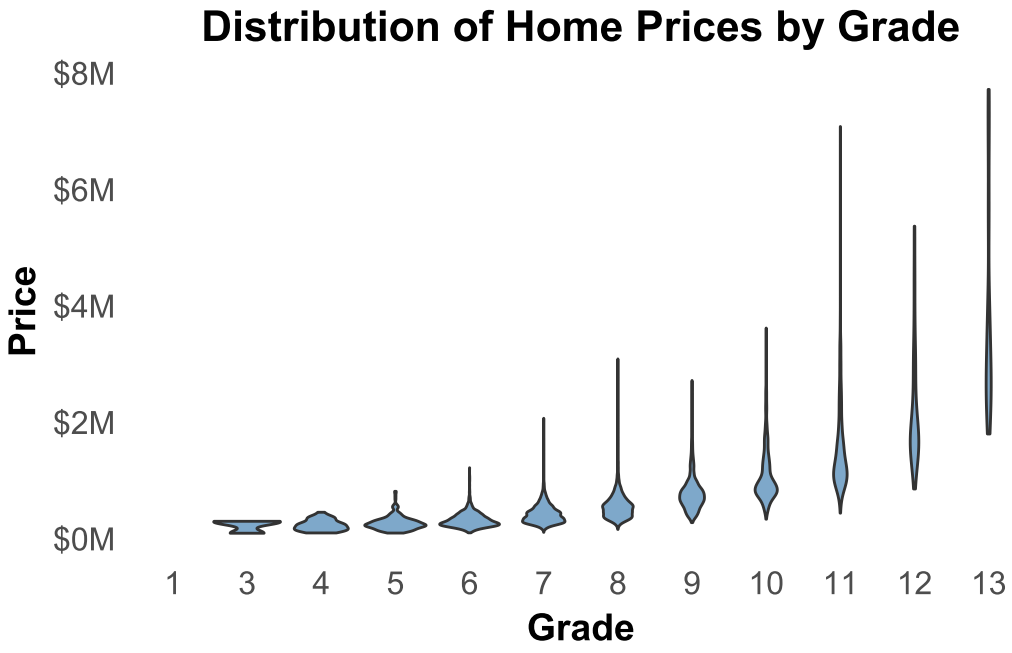
The plot on the left shows the relationship between home price and the liveable square footage of the home. The plot on the right shows the relationship between the log of home price and the liveable square footage of the home. The dual-log transformation shows a more linear relationship between home price and the liveable square footage of the home. This transformation is useful for modeling because it stabilizes the variance of the residuals and makes the relationship more interpretable.

Next lets look at how the condition of the home relates to the price of the home.



Poor and fair homes have significantly lower median prices, which should be expected, but what is really interesting is that the average condition homes actually have a $10,000 greater median price than the good-condition homes.

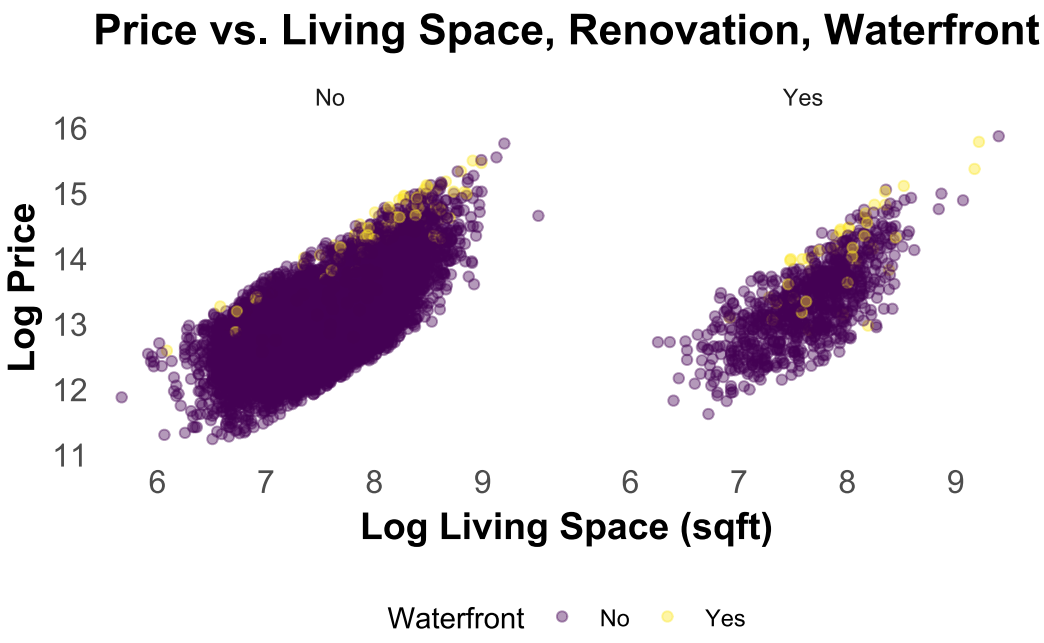
Lets explore the relationship between the grade of the home and the price of the home.



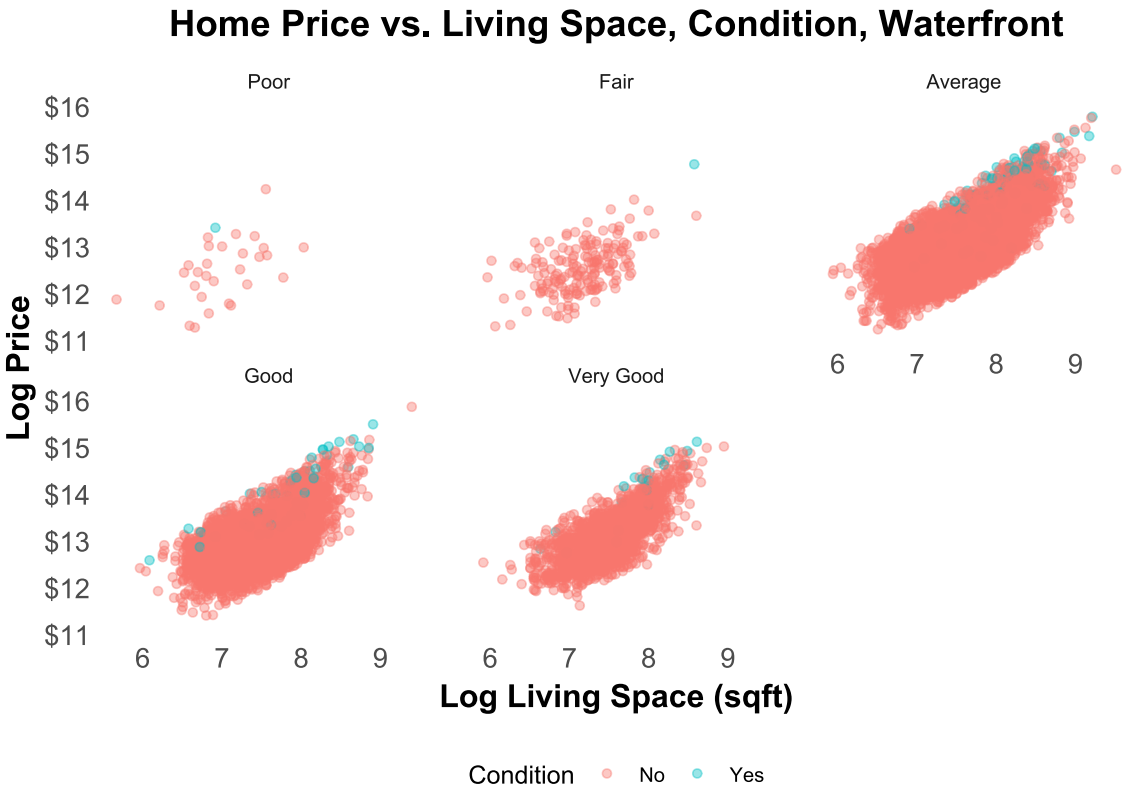
Here we can see a tendency in price to increase with grade, but we also see an increase in the variance of price as the grade increases.

Multivariate Visuals

Finally, we will examine how multiple features impact home prices. These visualizations help us understand more complex relationships in the housing market.

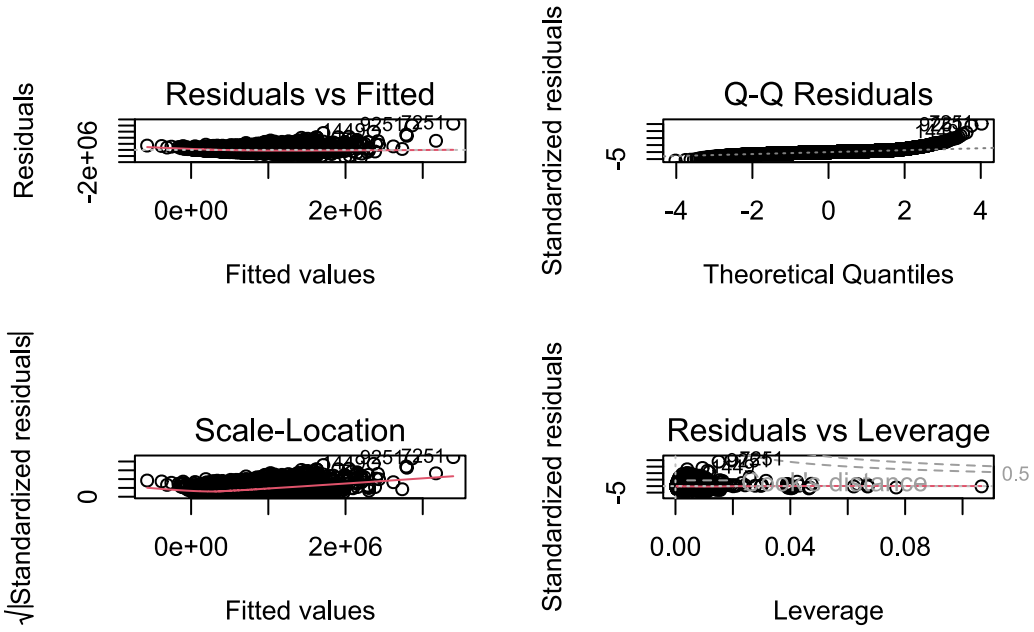


Lets look at the relationship between a home's condition, whether it has a waterfront, and it's price.



Based on our previous graphics, we would find that this one holds that as the percentage increase relationship between price and a home's living space. In addition, it is visually implied that higher condition homes (Average/Good/Very Good) have a higher proportion of Waterfront houses relative to lower quality houses (Poor/Fair).

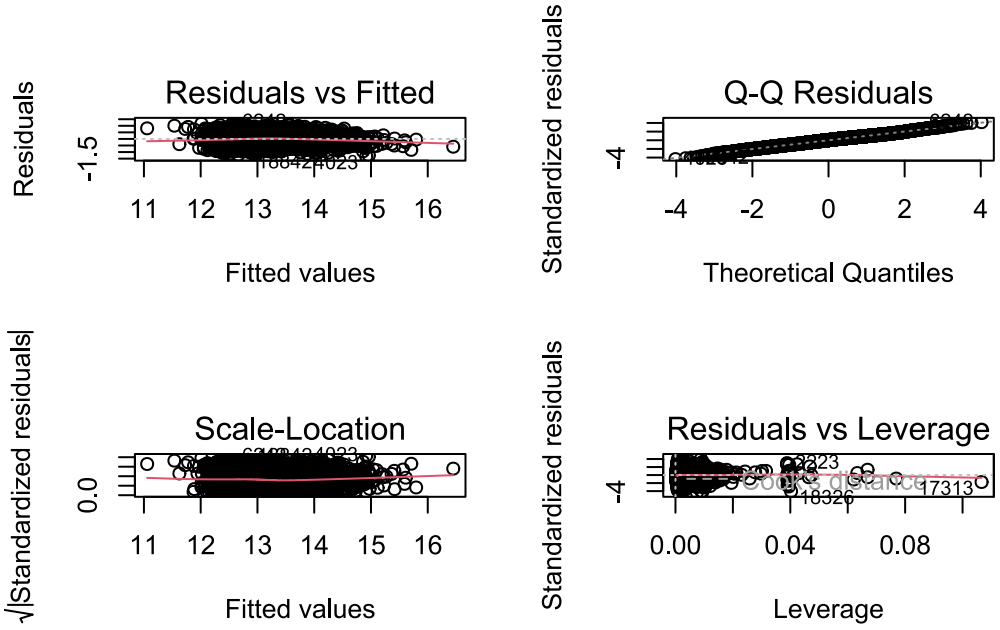
Section 5



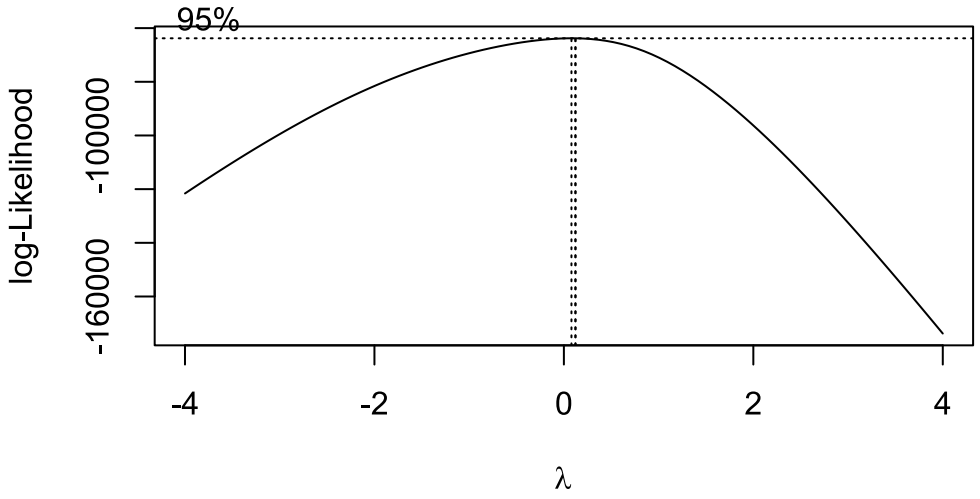
We started with a linear regression model predicting price based on all available variables. The goal was to evaluate how features such as square footage, location, condition, and time of sale affect housing prices in King County.

Initial residual plots suggested heteroscedasticity, indicating that the model assumptions were not well met. A density plot of price showed skewness, so we applied a log transformation to price, which improved the distribution.

This singular transformation improves how our model meets the assumptions of linear regression. The residuals are more evenly distributed across the horizontal axis, however there do seem to be significant outliers. Additionally, there does seem to be a slight curve in the residuals that looks to be influenced primarily by, again, the large relative outliers.



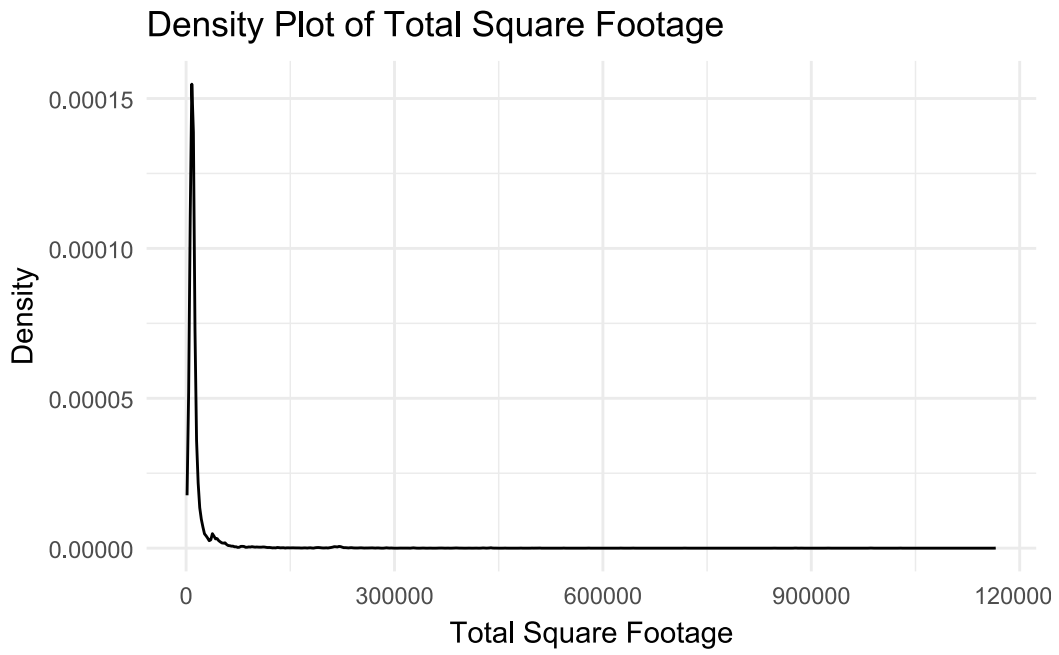
For completeness, we also checked the box-cox transformation to find the best lambda value for the transformation. While the CI for lambda does not seemingly include 0, it is very very close. We choose to go with the log transformation as it is more interpretable.



Next, we check for collinearity between the numeric variables. Correlations above 0.7 are all related to measures of square footage. We will remove all but total_sqft, as it is the most inclusive

measure of square footage. While we could use all of these variables, we also want to add interaction terms to the model which we believe would have more value than additional square footage variables, for computational efficiency, we would like to limit the amount of interactions we must add before backwards selection.

Next we will examine the distribution of total_sqft. The distribution is right-skewed, which provides evidence that we may need to transform this variable as well.

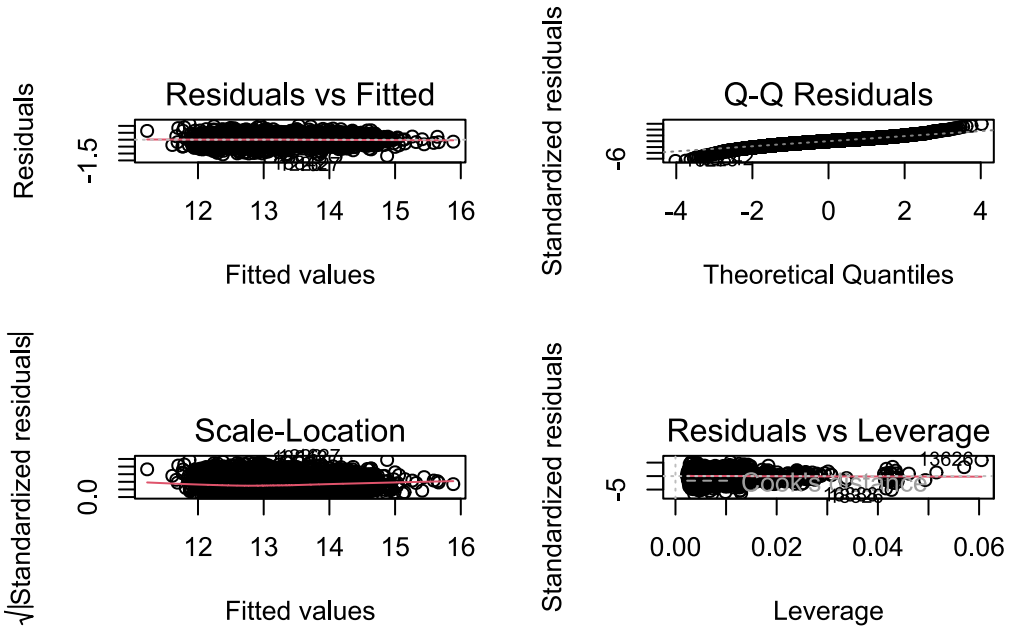


We examine the partial regression plot (omitted due to length check RMD file) of total_sqft to see how it relates to log(price) in the regression model. It is difficult to determine if the relationship is linear or not based on the plots alone, there was a slight curve in the residual plot of log(price) that may be due to the scarcity of data points in the upper range of total_sqft. We will log transform total_sqft to see if it improves the model fit.

After examining the regression without interaction terms with the various transformations, we can clearly see that the residual plots nearly match the assumptions of linear regression. The residuals exhibit no curve, however it does seem like higher priced homes may have a higher variance given the spread of residuals. There again seems to be a few outliers that are significantly effecting the model fit. We will fit a regression model with interaction terms through backwards elimination, then revisit outliers and influential observations.

```
      waterfront    view condition basement renovated season_sold
17401         No No view   Average      Yes        No      Spring
4775          No No view   Average       No       Yes      Summer
13218         No No view   Average       No        No        Fall
10539         No No view   Average       No        No      Spring
```

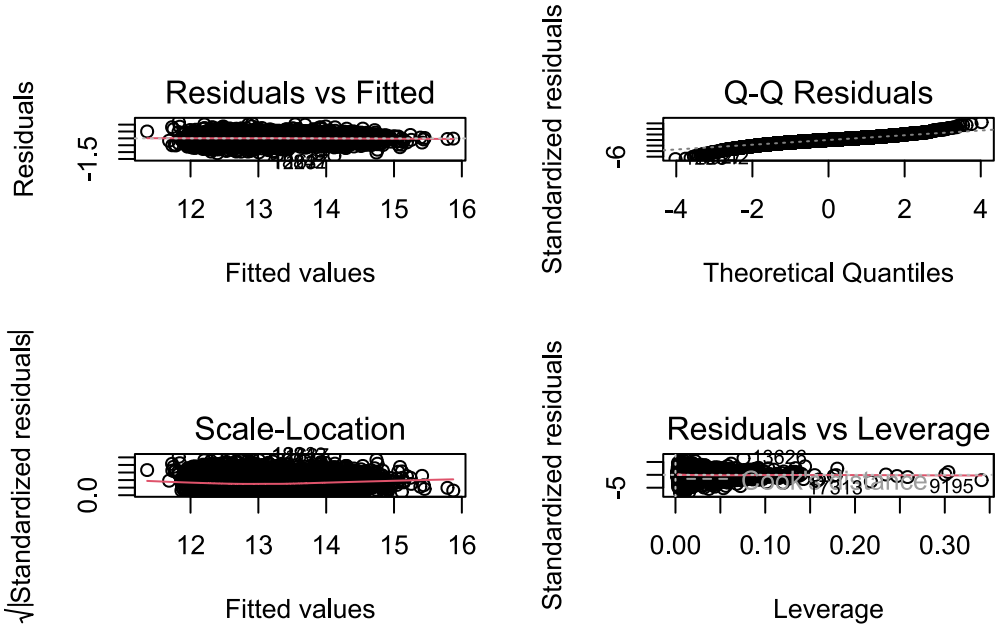
```
8462          No No view   Average       No        No      Summer
4050          No No view   Average       No        No      Winter
```



We choose to go with BIC for the backwards elimination process to create a more parsimonious model given the large number of predictors. (In the RMD file we have commented the backward elimination process out, as it takes a long time to run. We will use the output of the backwards elimination process to fit a new model with the chosen variables.)

We take the output of the backwards elimination process and fit a new model with the chosen variables. We will interpret the coefficients after we examine residuals and influential observations. (omitted due to length check RMD file)

Looking at the residual plot, we can see very little change in our previous assessment. It still seems as though a few significant outliers are impacting the model fit.



First, we will check for outliers using the studentized residuals. We find that about 1.09% of the data is flagged as outliers. The top 10 outliers are shown below. From this limited view, there does not seem to be a consistent trend in the outliers. We do however see an observation that seems to have a very high price for a low quality home, going for $1,200,000 and another going for $1,052,000. Both of these houses were built around the 1950's, there may be some covariate that we are missing which could explain the excised price, although there does not seem to be enough evidence to remove them.

Just on this view alone, we may not have cause to remove any observations. It does not seem as though the outliers represent data entry errors or issues that we do not want to measure in our study. (DF with outliers removed, check RMD file)

```
[1] 0.0109375
```

Using cooks distance, we find no values greater than 1. The highest value is 0.0201, which suggests that using this metric we find no outliers or influential points. We still examine this point as it seems to be significantly higher than the next highest value 0.131.

```
[1] "COOKS"
```

```
      13626        9195       17313        6627       18326       16888
0.020107749 0.013104769 0.011156914 0.010942326 0.009098868 0.008911371
      12423        6690         657       21040
0.007949326 0.007655858 0.007482627 0.006936463
```

This data point is also strange, a 1 floor 4 bedroom 2.75 bathroom house with 8410 sqft going for $805,000 also built in the 1950's. This house does seem to be a significant outlier, or atleast a very flat house. This is not representative of the data we are trying to model. We will remove this observation aswell.

```
       price bedrooms bathrooms floors waterfront    view condition grade
17141 805000        4      2.75      1         No No view Very Good     8
      yr_built zipcode basement renovated yr_sold month_sold day_sold
17141     1950   98107      Yes        No    2014          5       22
      season_sold total_sqft
17141      Spring       8410
```

Next we will check for influential observations using leverage. To check for influential observations and outliers, we used leverage, where observations with a leverage greater than 2p/n were flagged. The largest values were examined to identify high impact observations. We flag around 9.41% of the data seems to be influential observations. We can see an issue with the eigth highest observation, it has 1,165,504 total sqft but only has 2 bedrooms, 1 bathroom, and is going for 190,000. These prices are uncharateristic of the rest of the data, and may not be representative of the population we are trying to model. (DF with influential points removed, check RMD file)

Next we will use DFBETAs to measure the influence of each observation on the error. There are several influential observations, but there does not seem to be enough evidence to remove any of them. We see a similar situation with high amount of bedrooms but 1 floor, except in this case we have 245,020 total sqft which seems to justify the price. (DF with influential points removed due to length, check RMD file)

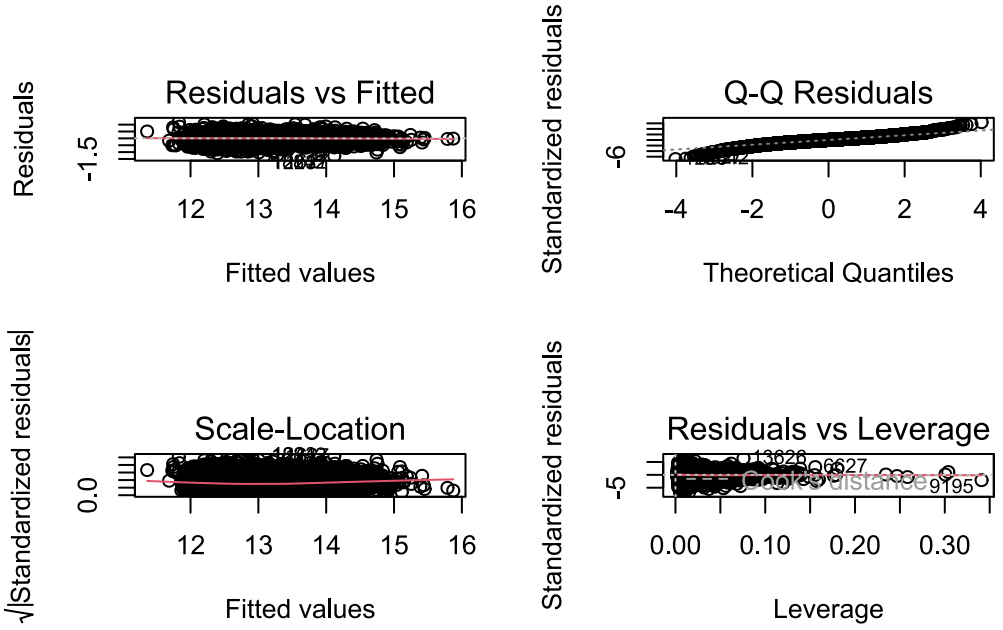
We will remove the outlier and re-fit the model moving forward with the assumption that the model is not an accurate representation of houses with high total sqft but low bedrooms and bathrooms.

Additionally, we test for if the model is useful.

H_0: The model is not useful, all coefficients equal 0. H_a: The model is useful, at least one coefficient is not equal to 0.

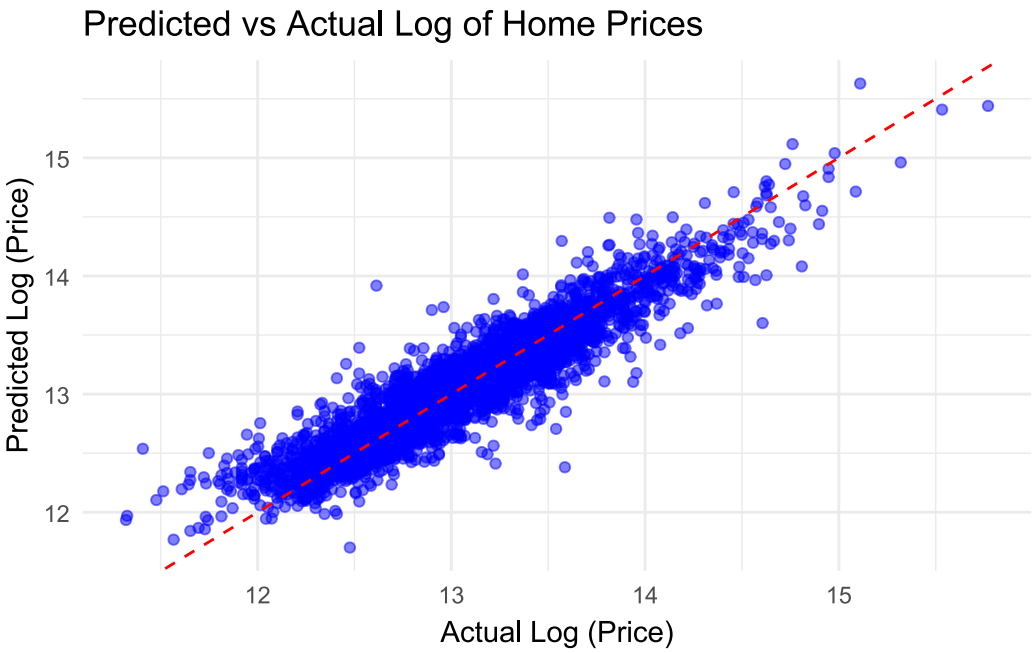
As we can see from the model summary output, the p-value of the F test statistic is approximately 0, which is less than 0.05. Therefore we reject null hypothesis with confidence, atleast one coefficient is not equal to 0. This suggests that the model is useful in predicting the log of price.

We also check for whether this model meets the assumptions of linear regression. The residuals exhibit no curve and are pretty evenly distributed across the horizontal axis, with some high leverage points closer to large values in price. For the most part, the residuals exhibit constant variance, except for again a few high leverage points. We go forward with the belief that this model fulfills the assumptions of linear regression.



We will now evaluate the model on the test data. We find that the R^2 value is 0.862, or that the model is able to explain 86.2% of the variance in the test data. This is close to the training R^2 value of 0.8708. This suggests that the model is not overfitting, and our interpretation of the coefficients may be generalizable to the whole population.

```
[1] "R^2 on test set:  0.862"
```



Interpretation:

Based on our model, we can see how the parameters can have a positive or negative relationship on the percent price estimation.

Those which are positively related are increases in the percent of total_sqft (1.435e-01), bedrooms (2.504e-02), bathrooms (8.970e-02), floors(6.355e-02), grade(1.590e-01), yr_sold (8.721e-02), month_sold (7.587e-03). Roughly speaking, as the unit values or percentages increased for these predictors as the others were held constant, it should have a positive impact on price.

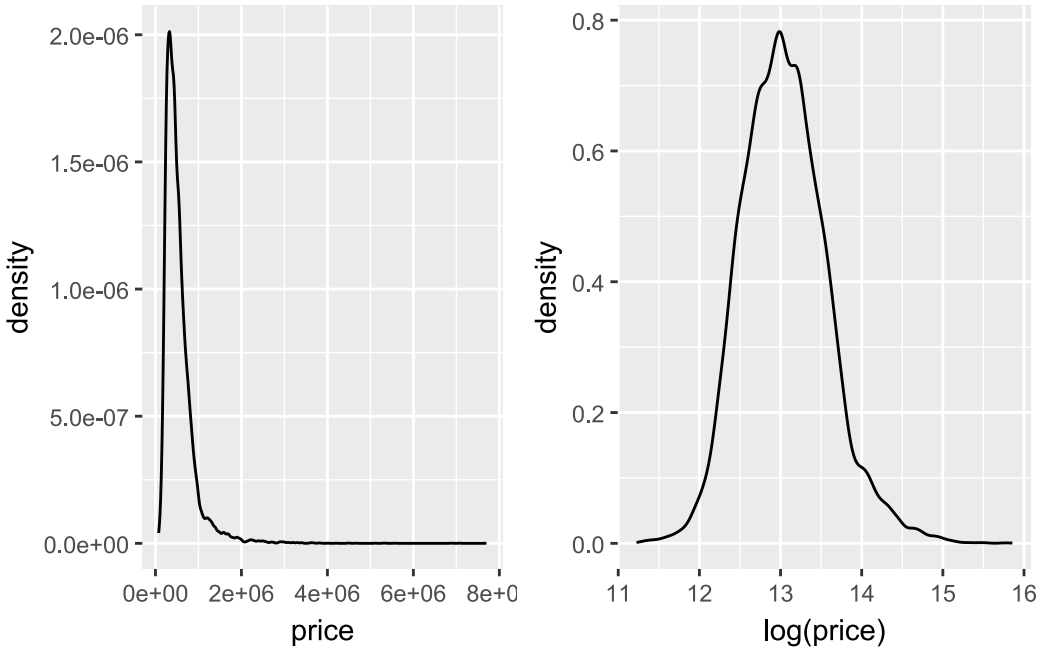
Categorically speaking, some of the positive relationships seen were with waterfront (Yes - 4.681e-01), basement (Yes - 3.147e-02),renovated (Yes - 7.307e-02), where the inclusion of one of these features would positively impact price. For season_sold, as we approached the winter, the offset would trend negative from the reference class. This is similar to condition, where the only negative offset from the reference class was houses that fell into the 'poor' category.

In short, our model provides us with a few implications. Roughly speaking, more is better for the price of a house when looking at its attributes (sqft, bathrooms, bedrooms, floors, grade). The inclusion of unique features such as a basement, waterfront view, or having been renovated, also implies the same positive impact on price. Our model also implies that the overall condition of the house can impact the price range of the house, and that the season in which the house is sold can also move the price with earlier seasons being more expensive.

Section 6

Visualizations

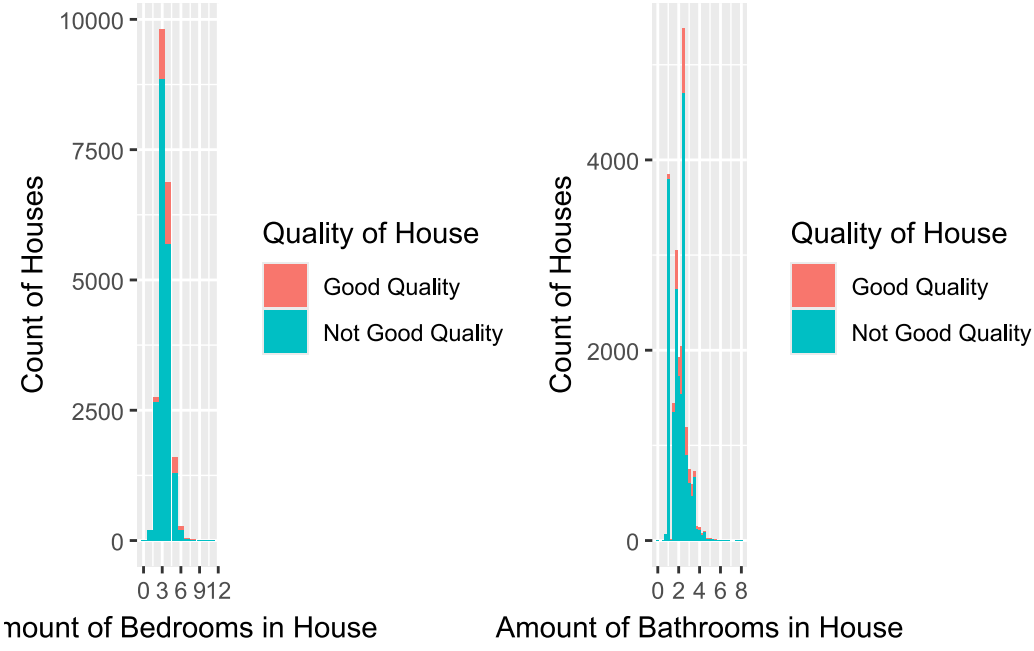
When looking at good quality homes, we wanted to first examine the relation ship between a house's quality and its price. When initially looking at price, we notice that the data is not normally distributed, so we decided to look at the price and quality relationship after putting price through a log transformation to bring it to a more normal distribution.



After this transformation, we examined the relationship by splitting the houses into two categories, good quality and not good quality. On a log-adjusted basis, we could see that a higher quality house would imply a higher price range over that of not good quality houses.

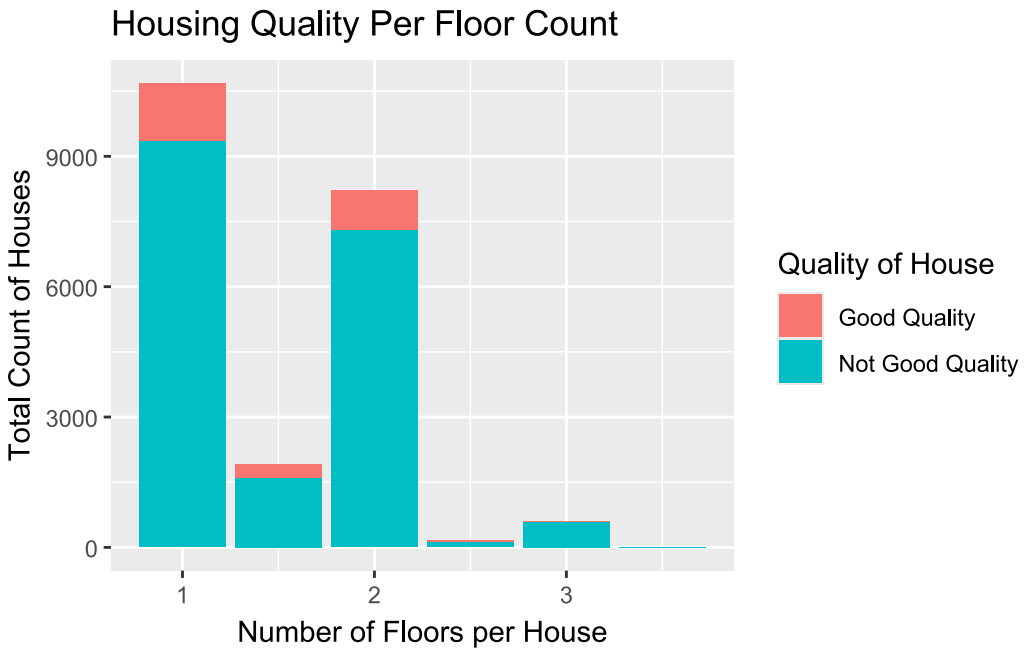
Generally speaking, we found that higher quality houses had a higher range than those of a lower quality.

Next, we wanted to explore the relationship between a houses quality and its relationship to bedrooms and bathrooms.



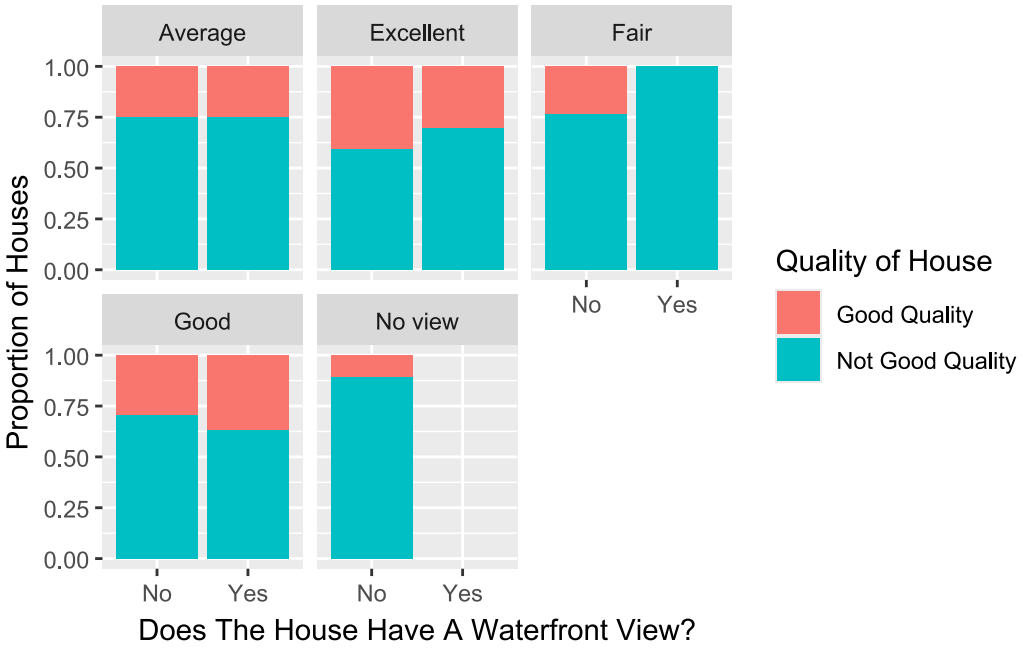
From our graphic, it looks like higher quality houses are found in the 3-6 bedroom range, and higher quality houses are spread across the bathroom.

Next, we wanted to look at the house quality and its relationship to the number of floors a house had.



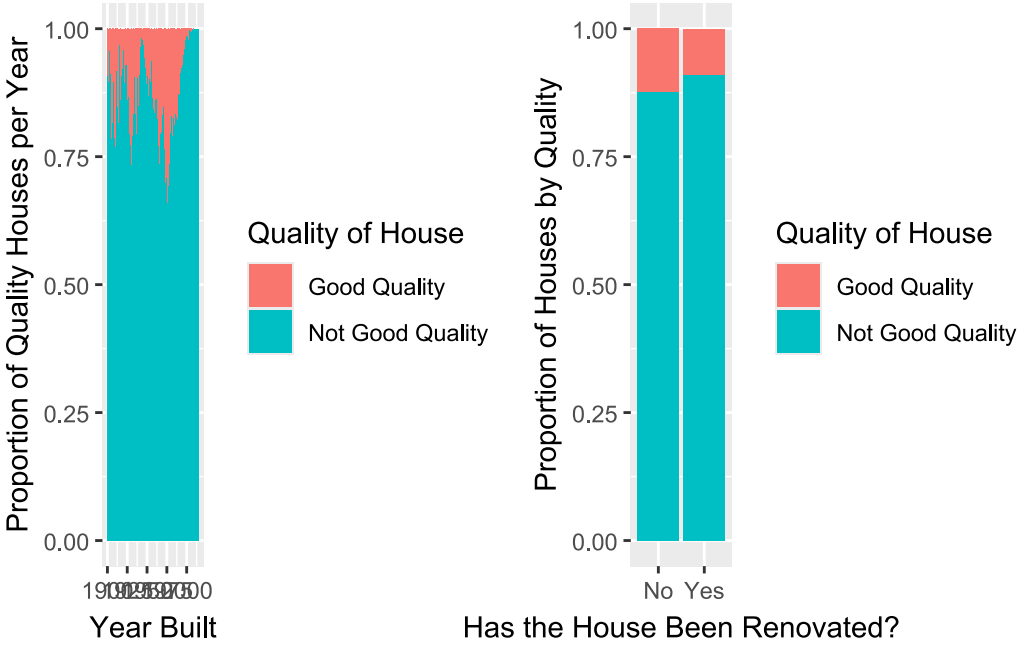
Our graphic implied that high-quality houses were relatively evenly spread proportionally across the different floor levels.

Next, we wanted to see the relationship between waterfront access and the type of view a person had access to.



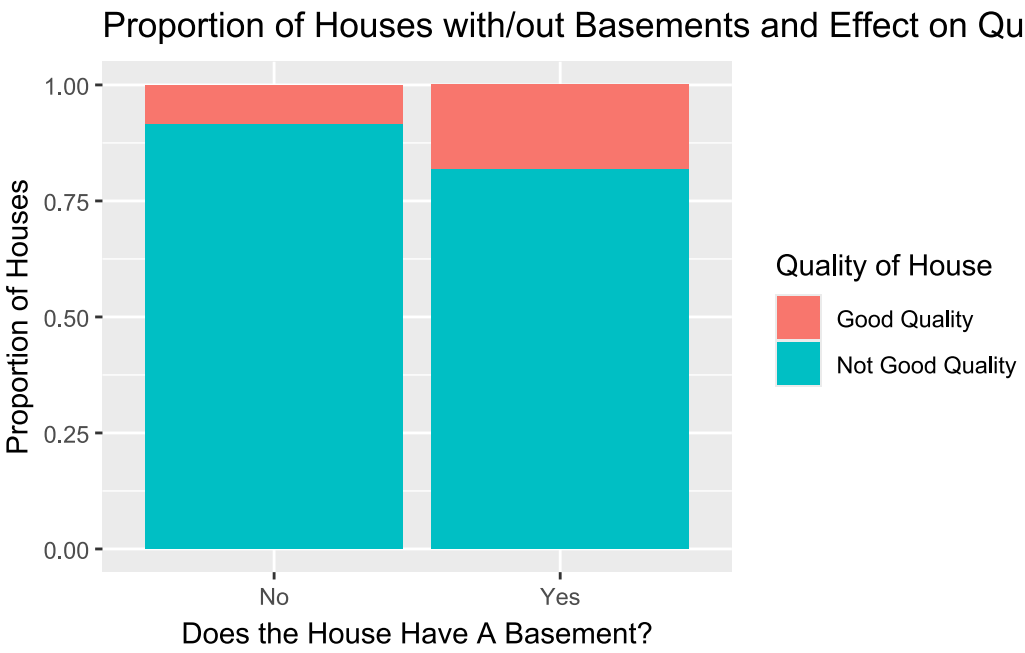
Generally speaking, there wasn't a great proportional difference between how many houses were of high quality with a waterfront view, regardless of the quality of that houses view.

Next, we wanted to see if the year in which the house was built and whether the house had been renovated had any material impact on the houses quality.



Generally speaking, based on the leftmost graphic we can see that the quality of houses and when they were built appear to have a cyclical relationship, where the housing market has built a greater a proportion of good quality houses in a given year and portion of the cycle. The right-most graphic implies that there isn't a material difference in the proportion of renovated houses when it comes to house quality.

Next, we wanted to check if the relationship between whether a house had a basement and its quality. It implied that the inclusion of a basement had only slight or negligible relationship with a house's quality.



Based on the histogram, we can see that the inclusion of a basement has a slight relationship with a house's quality. Many of the houses have a basement, but are not considered to be of good quality.

Section 7

We will use the same procedure we have in section 5 for removing square footage variables and log transforming the price and total_sqft predictors. For our justification in log transforming the price as a predictor, we can see that price is right skewed and has a long tail, effectively the same justification we have made for log transforming the total_sqft predictor.

Class Imbalance

As we can see, the classification problem is imbalanced, with only 4% of the data being classified as good quality. We will take this into account when designing an appropriate threshold for the model.

```
FALSE  TRUE
16572   708
```

Full Model

Summary of full model omitted due to length, but it is available in the R markdown file. The full model is all categorical interactions with log(total_sqft) and log(price).

Looking at the summary of the full model, we can see most of the coefficients have high p-values, indicating that they individually may not be significant. The larger issue is interpretability; the model is very complex and has several seemingly contradictory implications. For example, the VIF of full model. Again omitted for length but available in the R markdown file.

As we can see, almost all of the variables have a VIF greater than 10, indicating that they are highly correlated with each other. This is a problem for interpretability, as it makes it difficult to determine the individual effect of each variable on the response variable. This is also somewhat expected between log(price) and log(total_sqft), as we have just identified a regression relationship between the two. We will rexamine the VIF of the reduced model after backwards elimination.

Model Selection

We use BIC and backwards elimination to create a sparser model than just AIC during backwards elimination, this is mainly for interpretability, but the high VIF values of nearly all the predictors in the full model suggest that we should favor a parsimonious model. (In the RMD file we have commented the backward elimination process out, as it takes a long time to run. We will use the output of the backwards elimination process to fit a new model with the chosen variables.)

Reduced Model Interpretation

We extract the final model and perform diagnostics on it. Firstly, we examine the reduced model itself.

```
Call:
glm(formula = quality ~ log(total_sqft) + waterfront + renovated +
    yr_built + yr_sold + log(price) + waterfront:log(price),
    family = binomial, data = train)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             595.535032 192.905108   3.087  0.00202 **
log(total_sqft)           0.593538   0.047576  12.476  < 2e-16 ***
waterfrontYes            20.571289   5.753851   3.575  0.00035 ***
renovatedYes             -1.593162   0.216624  -7.354 1.92e-13 ***
yr_built                 -0.019102   0.001508 -12.664  < 2e-16 ***
yr_sold                  -0.300193   0.095773  -3.134  0.00172 **
log(price)                2.834836   0.089904  31.532  < 2e-16 ***
```

```
waterfrontYes:log(price)  -1.498871   0.401694  -3.731  0.00019 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5910.5  on 17279  degrees of freedom
Residual deviance: 4151.8  on 17272  degrees of freedom
AIC: 4167.8

Number of Fisher Scoring iterations: 7
```

The following interpretation assumes all other predictors are held constant. For every percentage increase in total square footage, the odds of a home being of good quality increase by 0.593%.

For every year increase in the year built, the odds of a home being of good quality decrease by 0.0191%. For every year increase in the year sold, the odds of a home being of good quality decrease by 0.300%.

If the home is renovated, the odds of a home being of good quality decrease by 1.59%. If the home is waterfront, the odds of a home being of good quality increase by 20.57%.

The interaction term between price and waterfront show that a home being waterfront decreases the slope of the price predictor. We can infer that it being waterfront is very important/desirable, and thus the quality of a waterfront home is less sensitive to the price of the home.

Additionally, a home being renovated decreases the odds of a home being of good quality. This is somewhat counterintuitive, as we would expect a renovated home to be of better quality. However, this may be due to the fact that homes that are renovated are often older homes that have been renovated to increase their value, but are still not of good quality. This also applies to year_built. We might expect newer houses to be of better quality, even though it is a common talking point today that newer houses are built worse. This may also be due to older homes being in better locations, if we were to further analyze this data we might want to take another look at zipcodes and their relationship with quality.

The other predictors conform much more with our biases, as we would expect a home to be of better quality if it has more square footage, is a waterfront, is more expensive, and has had less owners/been sold less recently.

Model Usefulness

We check the ANOVA of the reduced model against the full model to see if the reduced model is significantly different from the full model.

We'll perform a likelihood ratio test to compare the full model and the reduced model. The null hypothesis is to drop all predictors except those in the reduced model, ie. all coefficients of predictors not in the reduced model are equal to 0. The alternative hypothesis is that at least one of the coefficients of the predictors not in the reduced model is not equal to 0.

The chi-squared test statistic is 782.33, the critical value is 276.062. The associated p-value is approximately 0. Thus we reject the null hypothesis in favor of the alternative, the atleast one of the coefficients of predictors not in the reduced model is not equal to 0 with 95% confidence.

This is somewhat expected, we don't expect backwards elimination to select the best possible subset of predictors. We still go forward with the reduced model due to the issues with multicollinarity in relative interpretability of the reduced model.

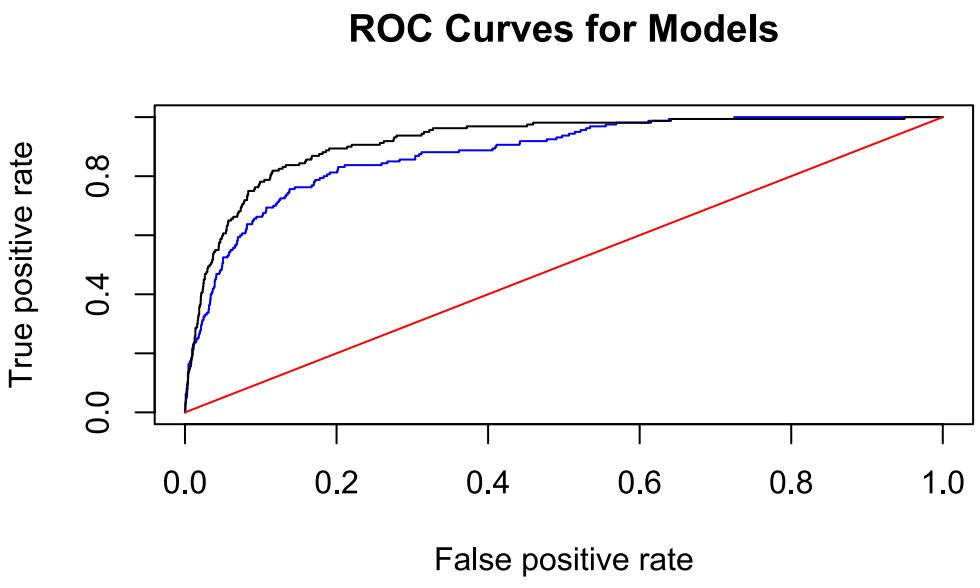
```
[1] "TS: 782.335211, p-value: 0.000000, critical_value: 276.062417"
```

Next, we will do a likelihood ratio test to compare the reduced model and the null model. The null hypothesis is that the reduced model is not significantly different from the null model, ie. all coefficients of predictors in the reduced model are equal to 0. The alternative hypothesis is that at least one of the coefficients of the predictors in the reduced model is not equal to 0. The Chi-squared test statistic is 1758.734593, the critical value is 14.067140. The associated p-value is approximately 0. Thus we reject the null hypothesis in favor of the alternative, the reduced model is useful in predicting quality with 95% confidence.

```
[1] "TS: 1758.734593, p-value: 0.000000, critical_value: 14.067140"
```

ROC Curve

We examine the ROC curve of the reduced and full models to see how well they predict the test data.



As we can see, the reduced model ROC curve is generally lower than the full model, however not by much.

AUC and Threshold

We examine the AUC next.

```
[1] "AUC of full model: 0.918837, AUC of reduced model: 0.879967"
```

As we can see, the AUC for the reduced model is slightly lower than the full model. Both are significantly better than guessing at random, however the reduced model is slightly worse than the full model at inference. Finally, we will select a threshold for the reduced model.

In the case of a home buyer, we want to minimize the false negative rate, as this would limit the amount of options the home buyer has. For a big singular purchase like a house, buyers will benefit from having more options to choose from, and the false positives may still be acceptable. For this, a threshold of around 0.05 seems reasonable, as it gives a false negative rate of 0.406250, instead of 0.843750 at 0.5.

The threshold is low due to large class imbalance and the fact we are trying to minimize the false negative rate. However, within this context the model is still useful, as it is able to identify a large number of good quality homes.

For a large multinational, we want to minimize the false positive rate as the utility of money is less important than the ability to make a decision. They can afford to have fewer options if it means they need less time to inspect homes. This may also apply to realestate recommendation engines as customers may be less likely to use the service again if they had to spend time looking at homes that were not of good quality.

In this case, the default threshold seems fine having a false positive rate of 0.004.

```
[1] "Threshold of .1 : FPR: 0.070192, FNR: 0.406250, TPR: 0.593750, ERROR:
0.082639, ACCURACY: 0.917361, PERCISION: 0.245478"
```

```
[1] "Threshold of .5 : FPR: 0.004567, FNR: 0.843750, TPR: 0.156250, ERROR:
0.035648, ACCURACY: 0.964352, PERCISION: 0.568182"
```