# Project 2 Proposal

Becky Desrosiers, Abner Casillas-colon, Alexandra Ferentinos, John Le

2023-11-07

# Section 1: Questions of Interest

## Question 1: What factors affect the price of a home? Out of all the predictors, which is the best model? Bedrooms, bathrooms, square footage of the house or the property, condition, square footage of above-floor-level space, square footage of basement space, year built, year renovated, and the square footage of the houses and properties of the 15 closest neighbors.

## Question 2: Can we predict whether a house is on a waterfront based on other qualities? Square footage, # bedrooms, # bathrooms, sqft_living15, condition, grade.

The response variable for the first question will be price. We are interested in finding which factors most influence the price. The response variable for the second data set is waterfront status: is it a waterfront property or not? We are interested in predicting whether a property is waterfront or not based on other attributes.

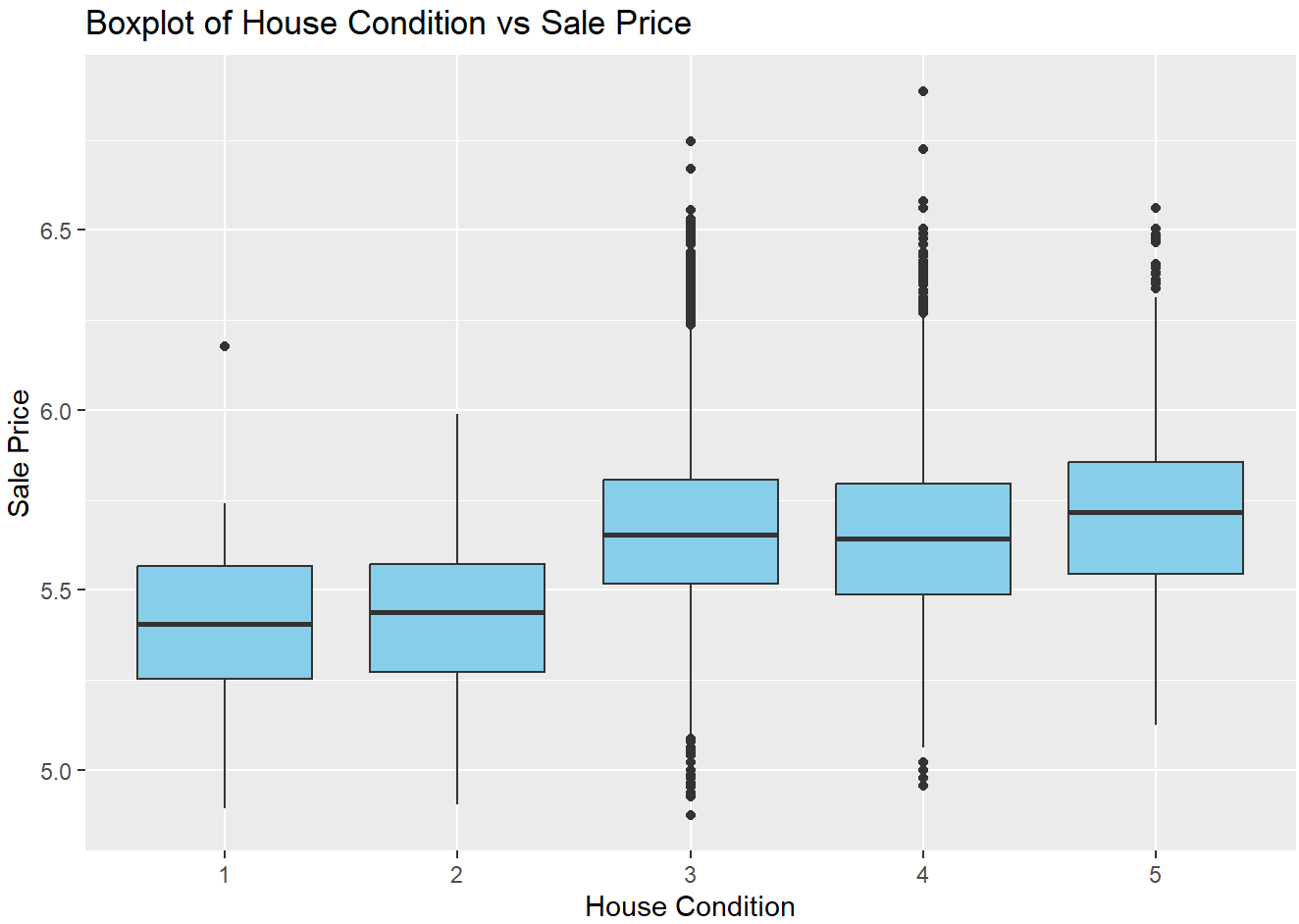
The practical application for the first question will be for a construction company to determine how best to invest their money into a project so that they can sell it for the best price. A buyer may also benefit from the analysis by discerning which attributes can be money-savers vs. with which attributes (number of bathrooms, number of floors, etc.) they may be able to gain more without spending much more.

We are interested in the second question to find out if there is a significant difference between waterfront properties and non-waterfront properties. We hypothesize that waterfront properties will be higher quality in general (higher square footage, more bedrooms and bathrooms, higher grade), and have neighbors with bigger houses, but perhaps will be in worse condition because of the proximity to water and the risk of flooding.

# Section 2: Data visualizations

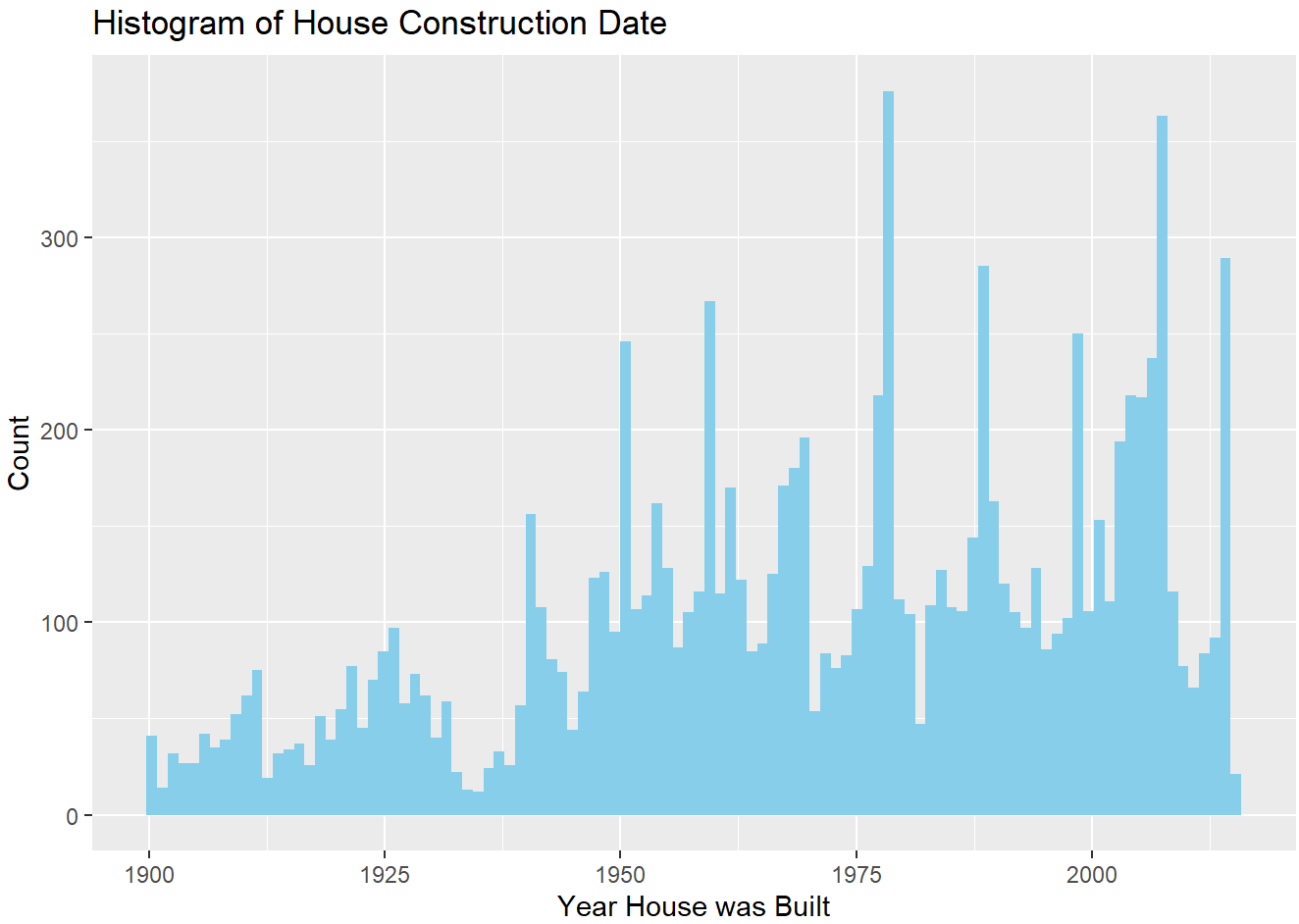
## Question 1: What factors affect the price of a home? Out of all the predictors, which is the best model? Bedrooms, bathrooms, square footage of the house or the property, condition, square footage of above-floor-level space, square footage of basement space, year built, year renovated, and the square footage of the houses and properties of the 15 closest neighbors.

.

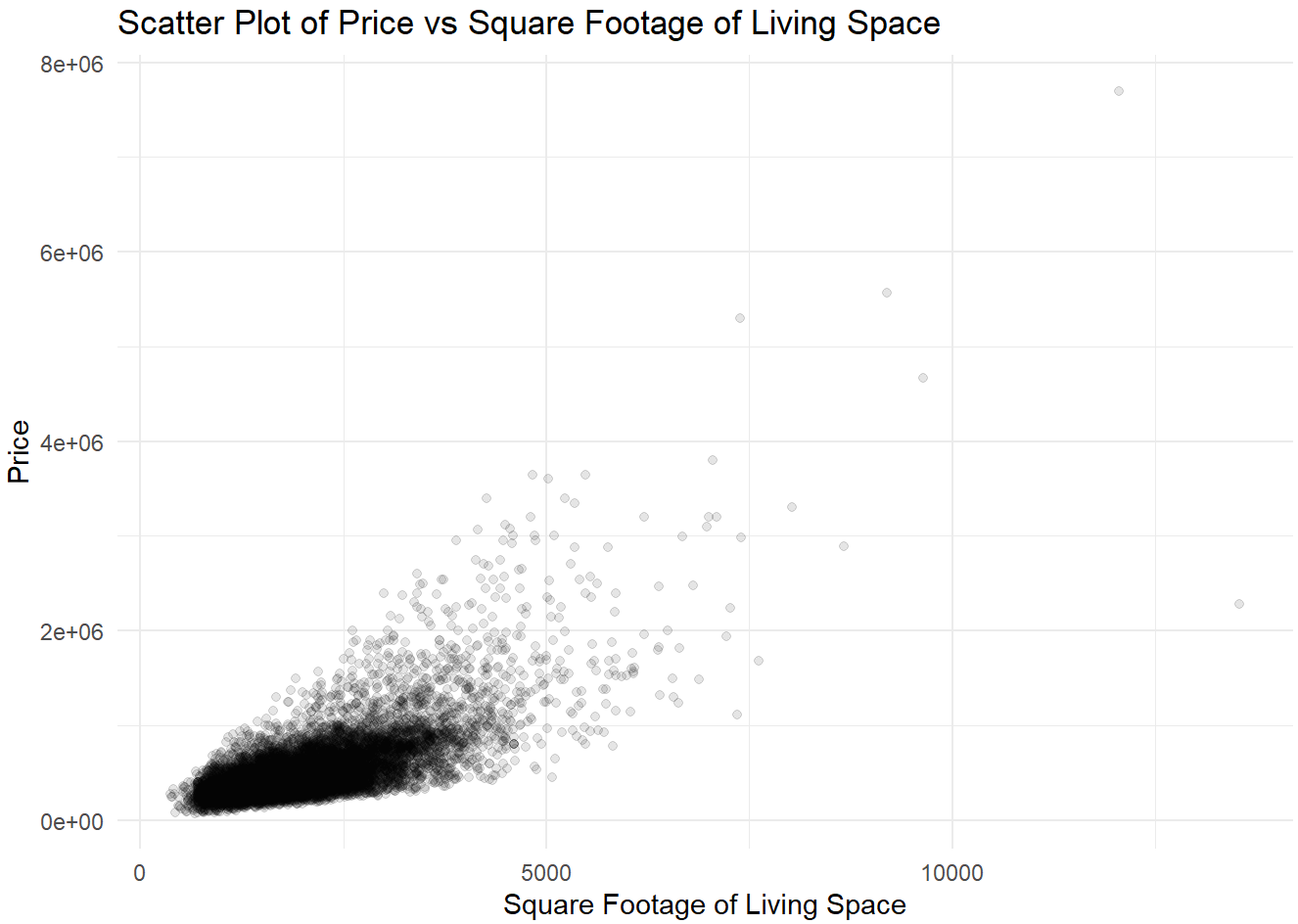



The box plot provides a visualization of sale price based on house condition, where 1 is the worst condition and 5 is the best condition. The boxplots show interesting information about the relationship between condition and sale price. Houses in better condition have more outliers. This indicates that there is a higher variance in sale prices within house conditions 3 to 5.

This visualization relates to the above question as it shows that house condition 3 to 5 has higher variability in sale price meaning there may be other predictors affecting this nuance. So more analysis of sale price, against the above predictors from the question could illuminate the cause for such outliers.




This histogram is created to evaluate the distribution of the construction date of the houses in the dataset. The distribution of the histogram is left skewed, with more observations occurring in recent years. Though the skewness that is present is not necessarily extreme it may still be prudent to examine if the older homes present outsized influence on price when fitting the model. One thing to note is that certain years exhibit high spikes of activity in regular intervals, which appear to be after every 5-10 years. When examining the relationship between price and construction date these high volume points may be worth further scrutiny in case the increased quantity coincided with decreased quality of construction that impacts price.

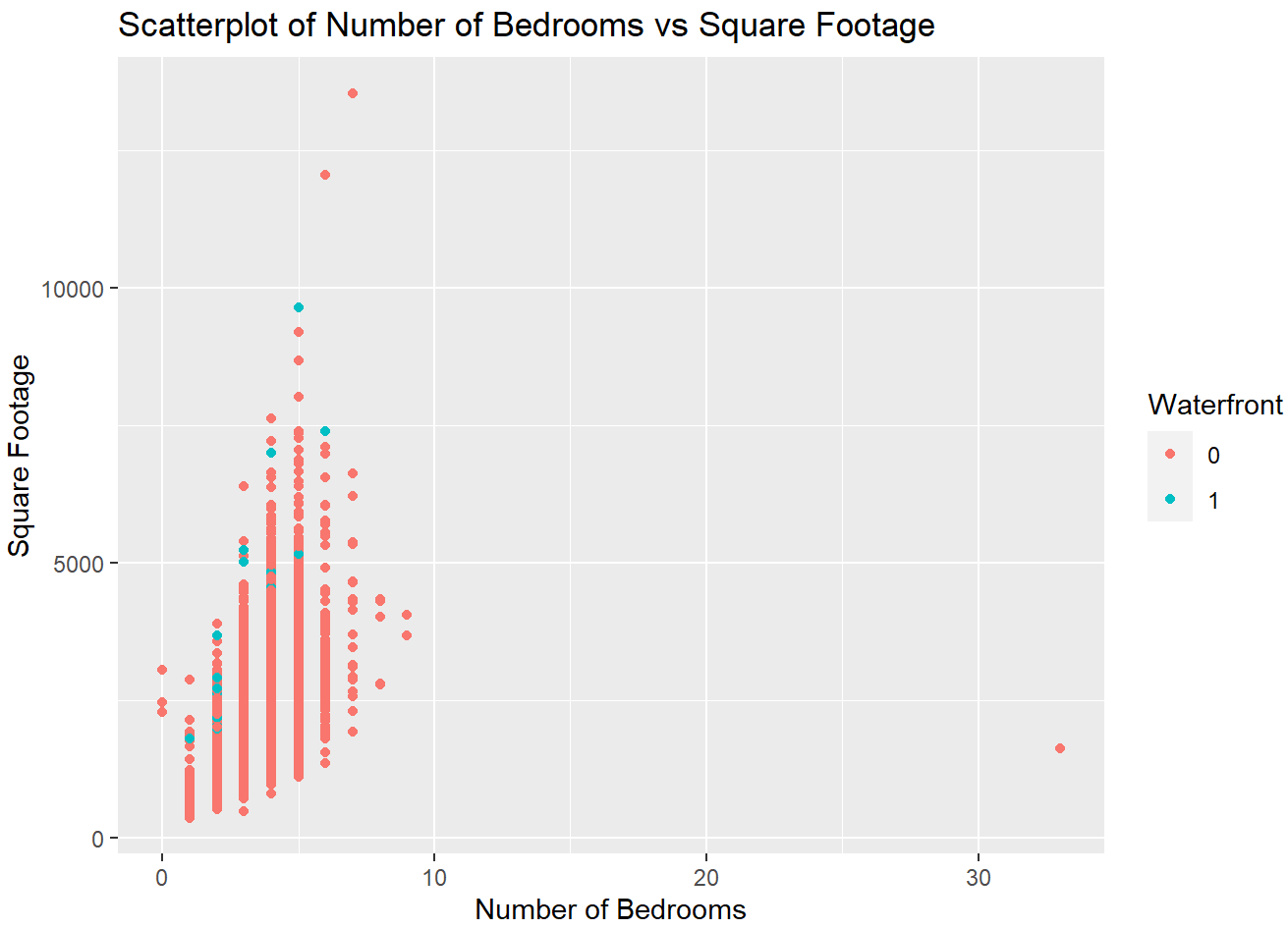



There is a clear positive correlation between the square footage of living space and the price of a house. As the square footage increases, so does the price, which is a common trend in real estate markets. Most of the data points are concentrated in the lower range of square footage and price, indicating that a majority of the houses in this dataset are moderately sized and priced. The plot suggests that as homes increase in square footage, the price does not just increase linearly but may increase at a higher rate. This could imply a premium for larger homes beyond a certain size.

The dense clustering of points at the lower square footage range could make it challenging to distinguish between the price differences among smaller homes. This might necessitate a more detailed analysis or the use of additional variables to understand the pricing structure for these homes. For any given square footage, there is quite a wide variation in price. This indicates that factors other than square footage are also affecting the price, which we will explore in our analysis. There are several data points which may have high leverage, particularly houses with a large square footage that are priced much higher than the rest. These could represent luxury homes or properties with unique features that significantly increase their value. Also of note is the one point at the high end of square footage but the low end of price, which could be influential.

# Question 2: Can we predict whether a house is on a waterfront based on other qualities? Square footage, # bedrooms, # bathrooms, sqft_living15, condition, grade.

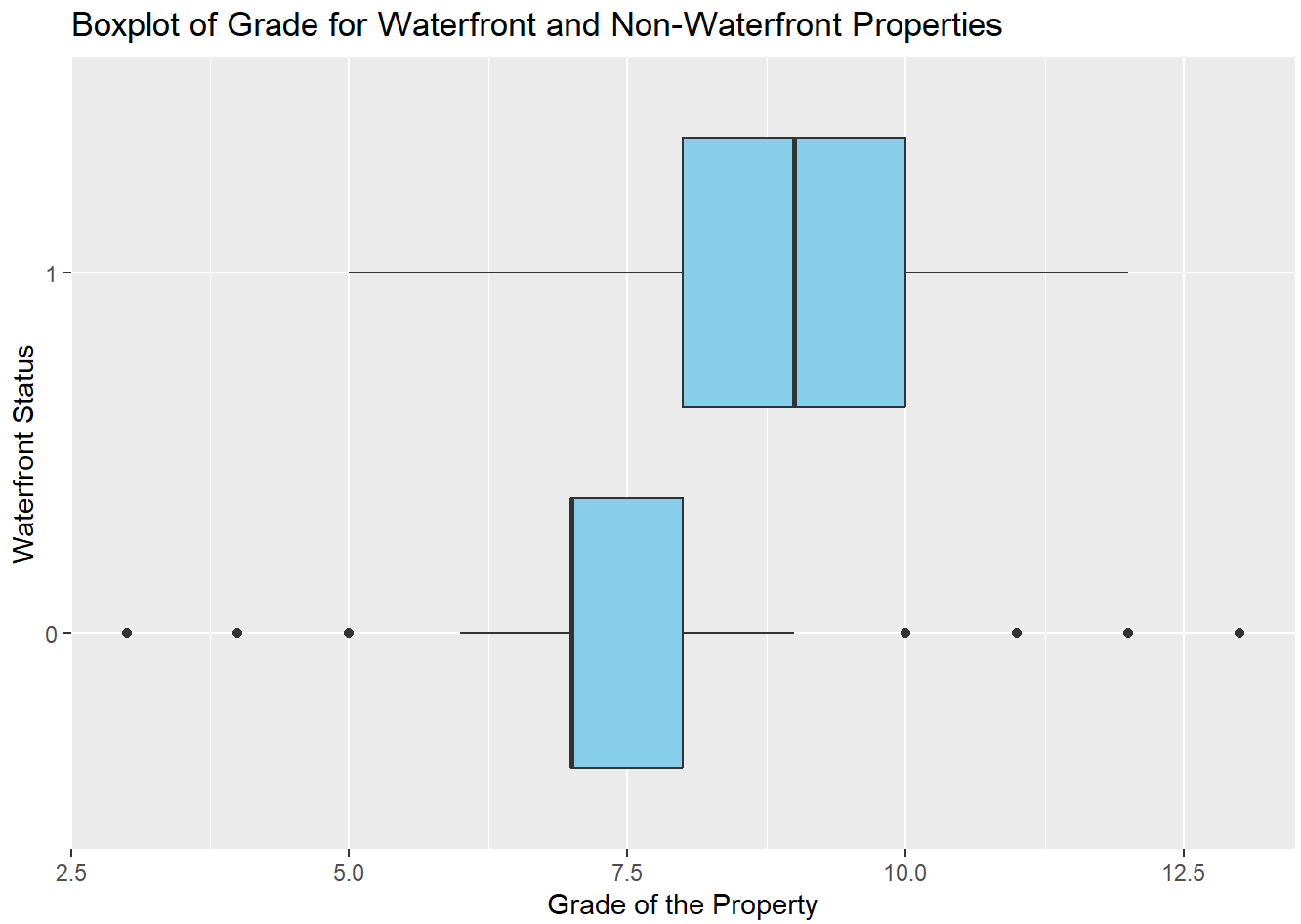
.




The scatter plot data visualization indicates a concentration of residential properties within the range of 0 to 10 bedrooms in terms of frequency. Additionally, it demonstrates that homes with higher square footage tend to be centered within 0 to 10 bedrooms, with 5 being the most common. From the plot, it appears that a the majority of homes are not waterfront homes.

It is interesting to note that the outliers in terms of the square footage of the living spaces do not seem to be distinctly influenced by the proximity of a home to a waterfront. This observation may imply that the size of the living area of a house may not necessarily correlate with its location relative to a waterfront area, this could be that other predictors are more influential in the response variable square footage. So, from the waterfront quality in consideration to square footage we cannot necessarily predict whether the house is a waterfront home. It does appear, however, that within a given number of bedrooms, waterfront properties tend to be on the higher end in terms of square footage.

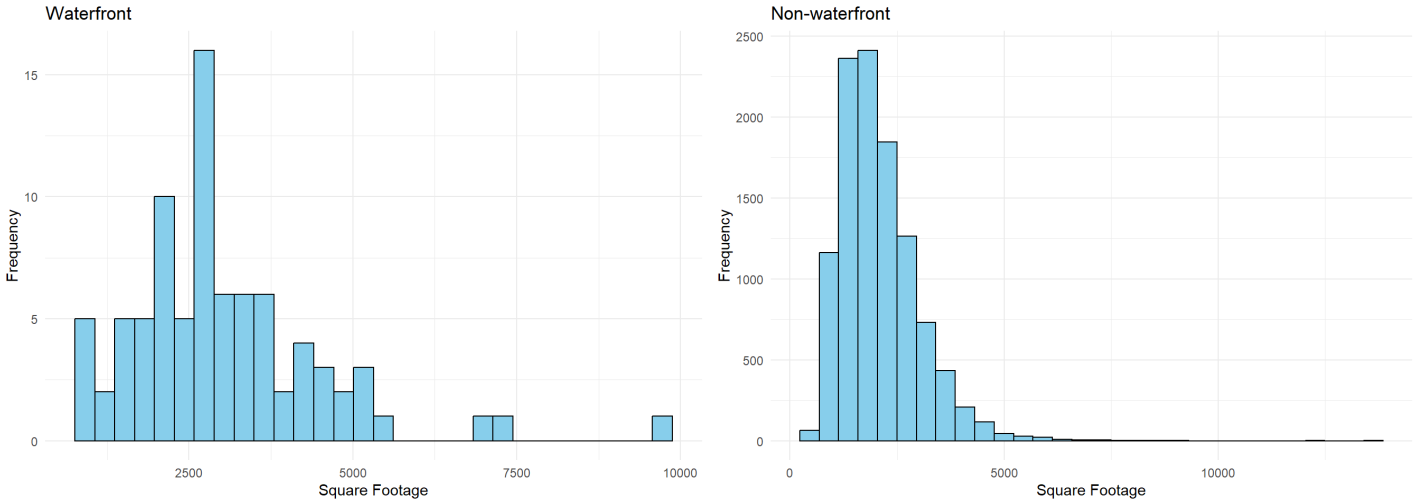
One other important thing to note from this visualization is the outlier with more than 30 bedrooms. This seems very unusual, and may be an error in input, especially because the square footage is relatively low. We may have to consider dropping this data point.




The x-axis for this visualization represents grade which for context indicates that 1-3 can be considered “poor construction” 7 has an average quality and 11-13 have a high quality. On the y-axis, the “1” group indicates that the property is waterfront while “0” indicates that it is not a waterfront property.

When examining the properties without comparing their categories we note that waterfront properties tend to be fairly soundly made averaging a score of approximately 8 and nothing within the poor construction range. The distributions appears to be fairly even around the average score of 8 as well. This is contrasted by non-waterfront properties which have 7 outliers and houses that seem to be lower grade construction averaging around 7. When we conduct the analysis we may wish to keep an eye out for the many outliers in the non-waterfront property category that may potentially skew the analysis.

## Histogram of Square Footage for Waterfront and Non-waterfront Properties




From the graphs, we can observe that the distribution of living area sizes for both waterfront and non-waterfront homes appears right-skewed, indicating that there are a larger number of homes with smaller living areas and fewer homes with larger living areas.Homes that are not on the waterfront are more numerous across the range of sizes, so non-waterfront homes make up a larger proportion of the data. Waterfront homes are less frequent, which is expected since waterfront properties are typically rarer and potentially more desirable.

The range of square footage for non-waterfront homes extends from the smallest to the largest sizes, showing a wide variety of home sizes. Waterfront homes tend to have larger living areas on average, with fewer small-sized homes compared to non-waterfront homes. There is a peak in frequency for non-waterfront homes at the smaller end of the scales. For waterfront homes, the data are more evenly distributed, although with a much lower frequency overall due to fewer waterfront homes in the dataset.