Project 2 Proposal

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

2023-11-07

Section 1: Questions of Interest

Question 1: Have houses gotten bigger and better over time? That is, is there a linear relationship between year built and square footage, grade, and the square footage of the nearest 15 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 square footage. We are interested in seeing if square footage has increased over time. 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 properties.

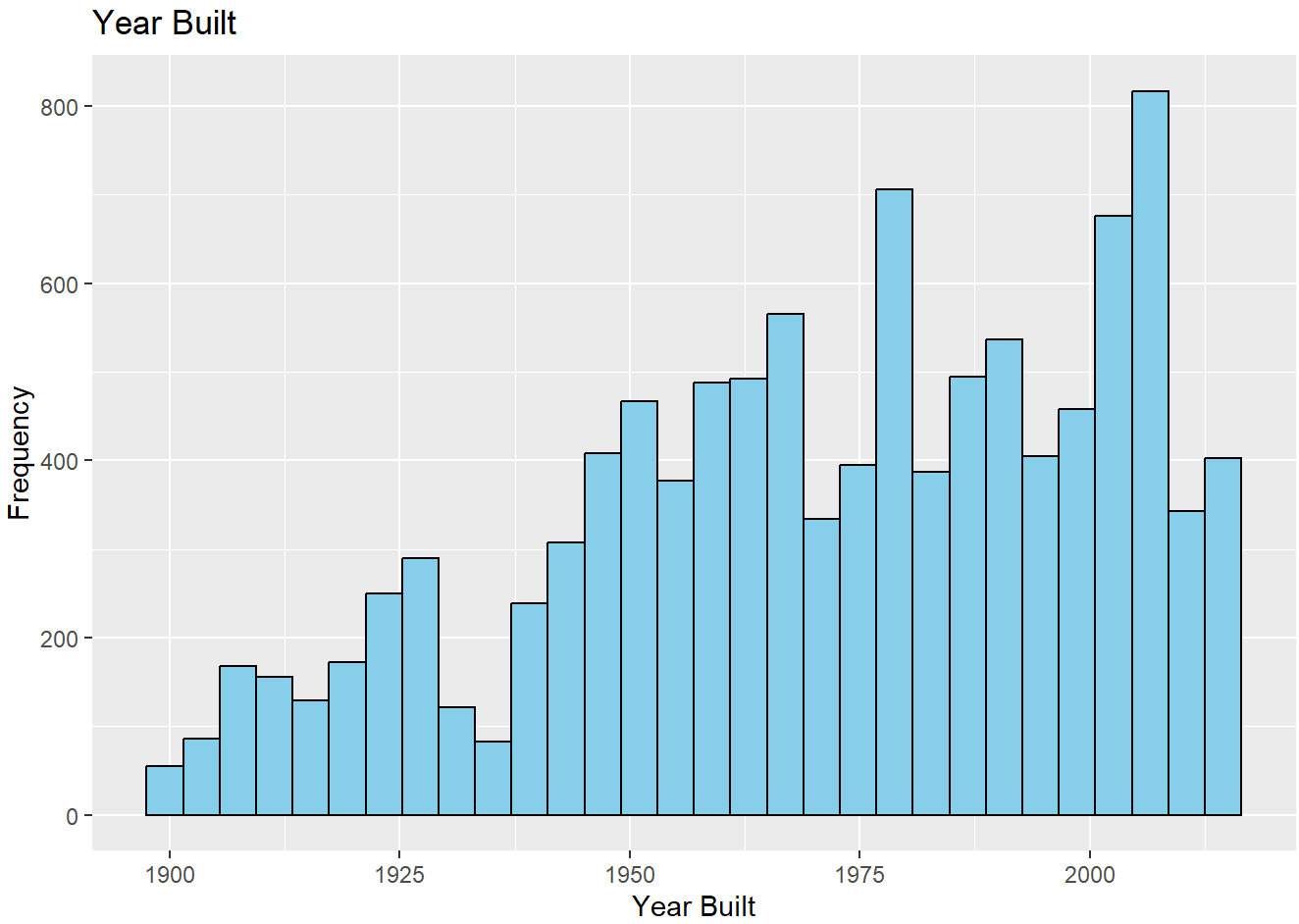
We are interested in the first question because we have noticed that houses are very expensive nowadays. Even so, when we go to historic places and see the types of houses regular people used to live in, it's like what we would call a shack in modern times. Houses have become much more extensive and also include more amenities, such as central heat, air conditioning, and indoor plumbing. We are interested in looking at if square footage, grade, and neighborhoods with larger houses can be attributed as some of the factors in the higher prices we see today, by seeing how it has changed over time. Our hypothesis is that these characteristics will have increased to a statistically significant degree and are a factor in high housing prices.

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 be of lower condition because of the proximity to water and the risk of flooding.

Section 2: Data visualizations

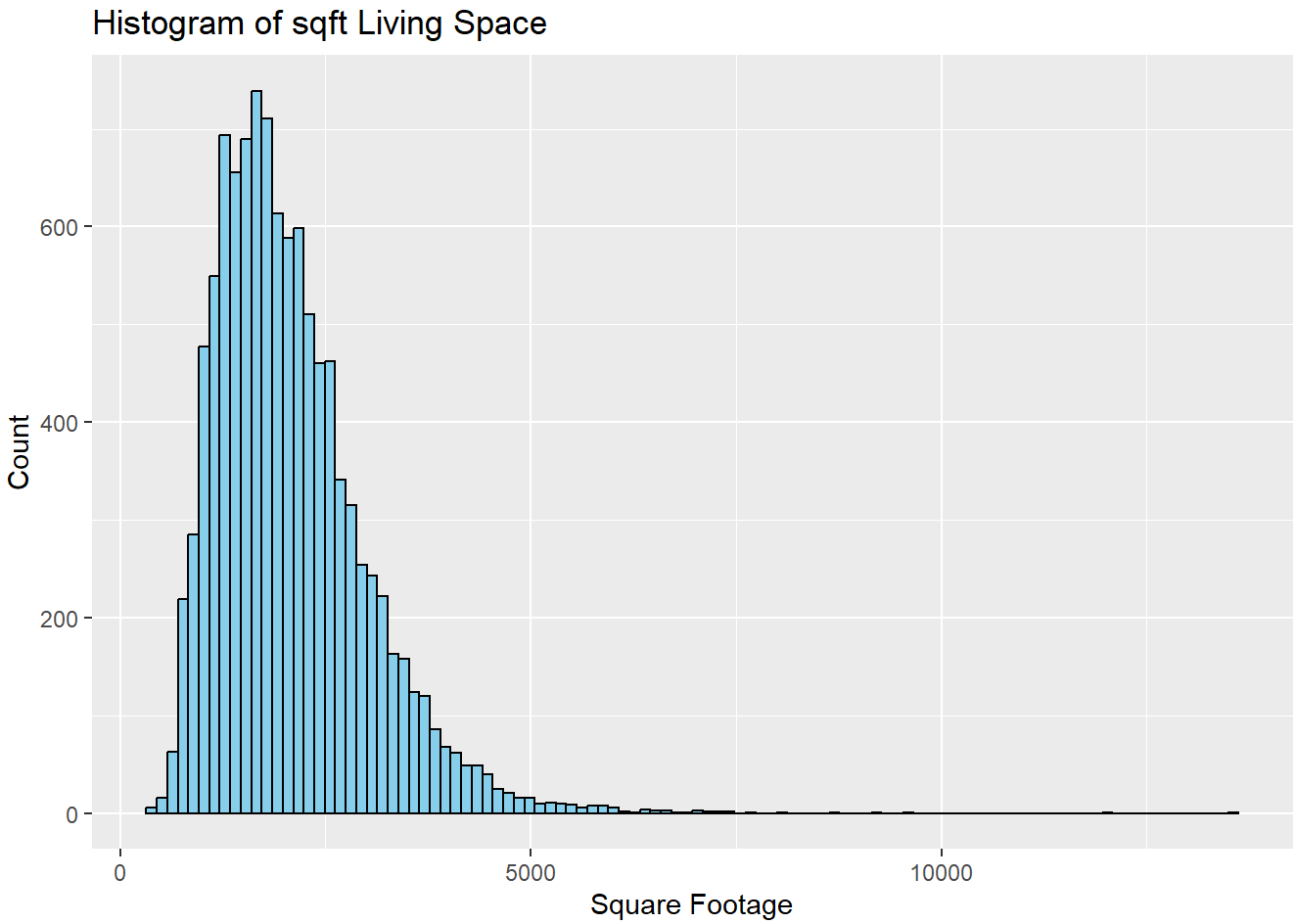
Question 1: Have houses gotten bigger and better over time? That is, is there a linear relationship between year built and square footage, grade, and the square footage of the nearest 15 neighbors?

.

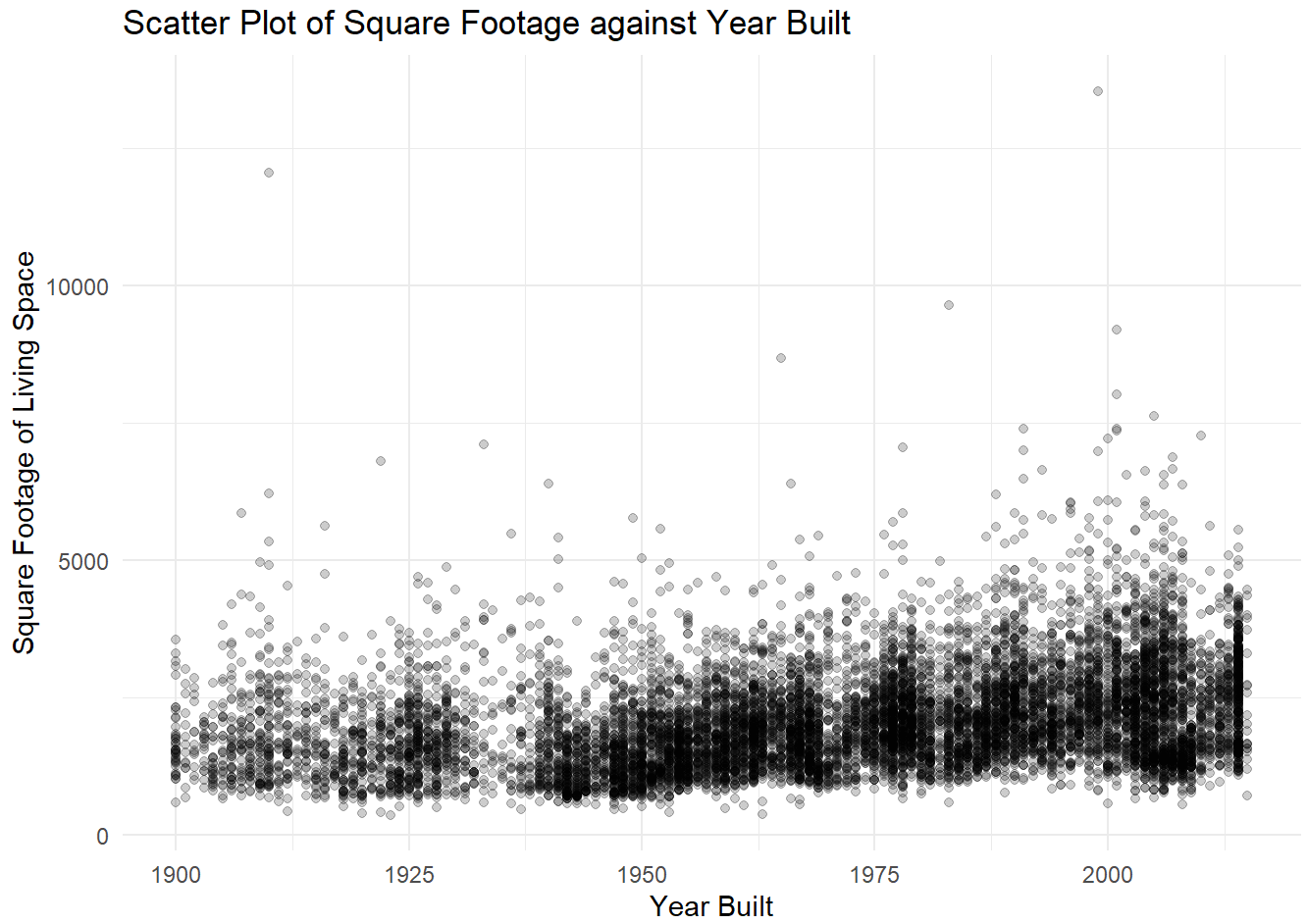


The histogram visualization notes a right skew in the data. There appears to be a positive trend between year built and homes built. This implies that there is an upward trend in home construction throughout the years represented in the dataset. This could be due to an increase in housing construction due to population growth of the area.

This visualization relates to the above question as from this positive trend we can then note if the square footage has increased over the years as well, perhaps due to larger home popularity throughout the modern residential construction age, or perhaps other factors.



The histogram displays a distribution for the squarefoot living space of the houses within the training data set. Our distribution is right-skewed, indicating potentially to be on the lookout for outliers on the extreme end that will impact the linearity of the relationship when we conduct our testing. Additionally, if we are to describe our data set we would expect the mean value to be higher than the median and may want to use the median value as a descriptor when discussing how to evaluate the central point of housing living space as a better representation of the data set.

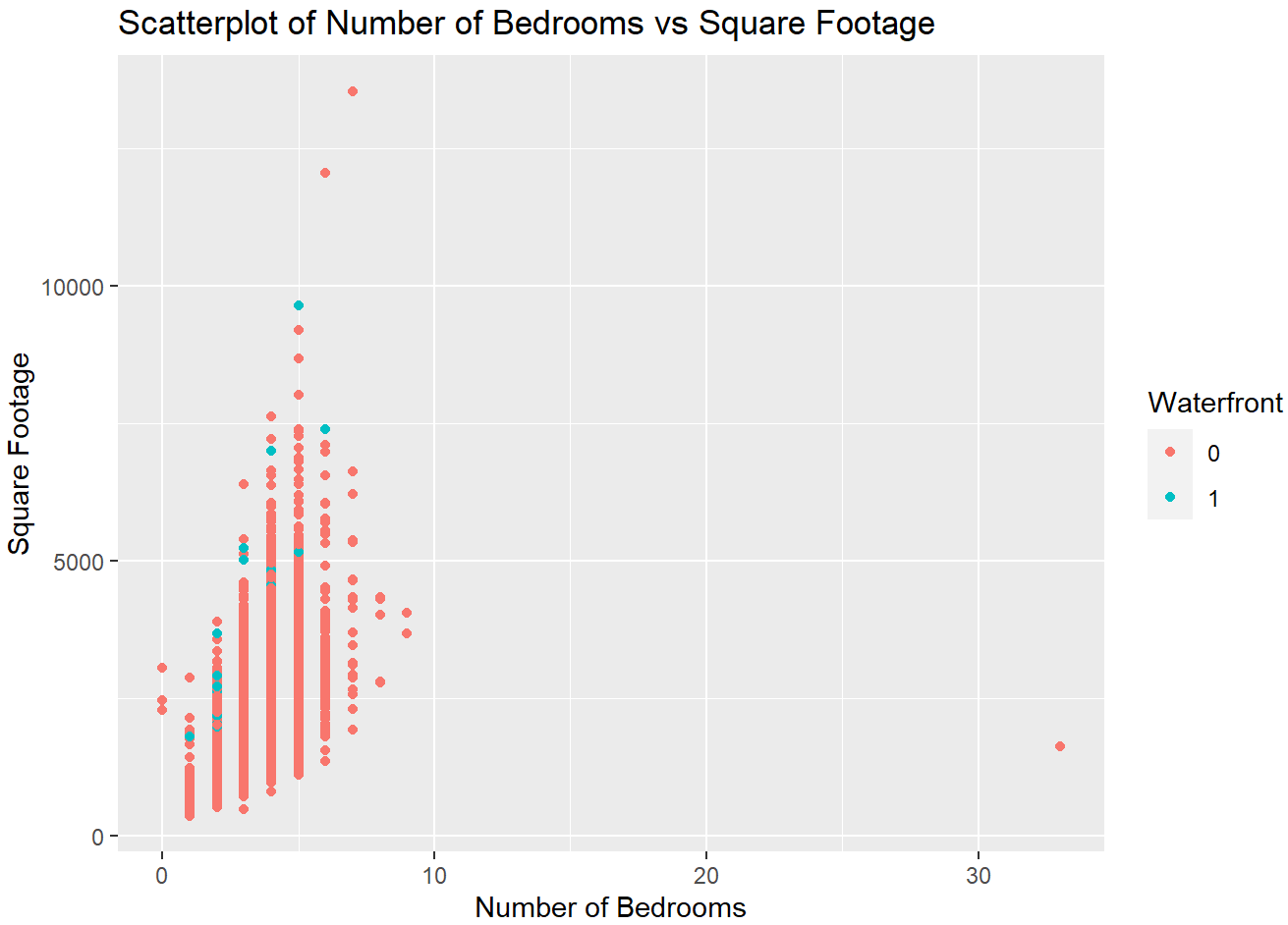


The points are spread across the entire range of years, from the early 1900s to the 2000s, showing that the dataset includes a wide range of house ages. There doesn't appear to be a clear trend indicating a strong relationship between the year built and the size of the living space. The distribution of square footage is fairly consistent across the years, with a large concentration of homes between 1000 and 3000 square feet regardless of the year built. Towards the more recent years (post-1975), there is a visible increase in the number of homes with larger square footage, which could suggest a trend in building larger homes in more recent decades.

There are a few outliers, particularly homes with exceptionally large square footage, which stand out above the main cluster of data points. These could represent mansions or unusually large properties. For the early 1900s, the data points are more sparse, which could be due to fewer homes from that era being included in the dataset, or it could reflect less consistency in house sizes during those years.

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

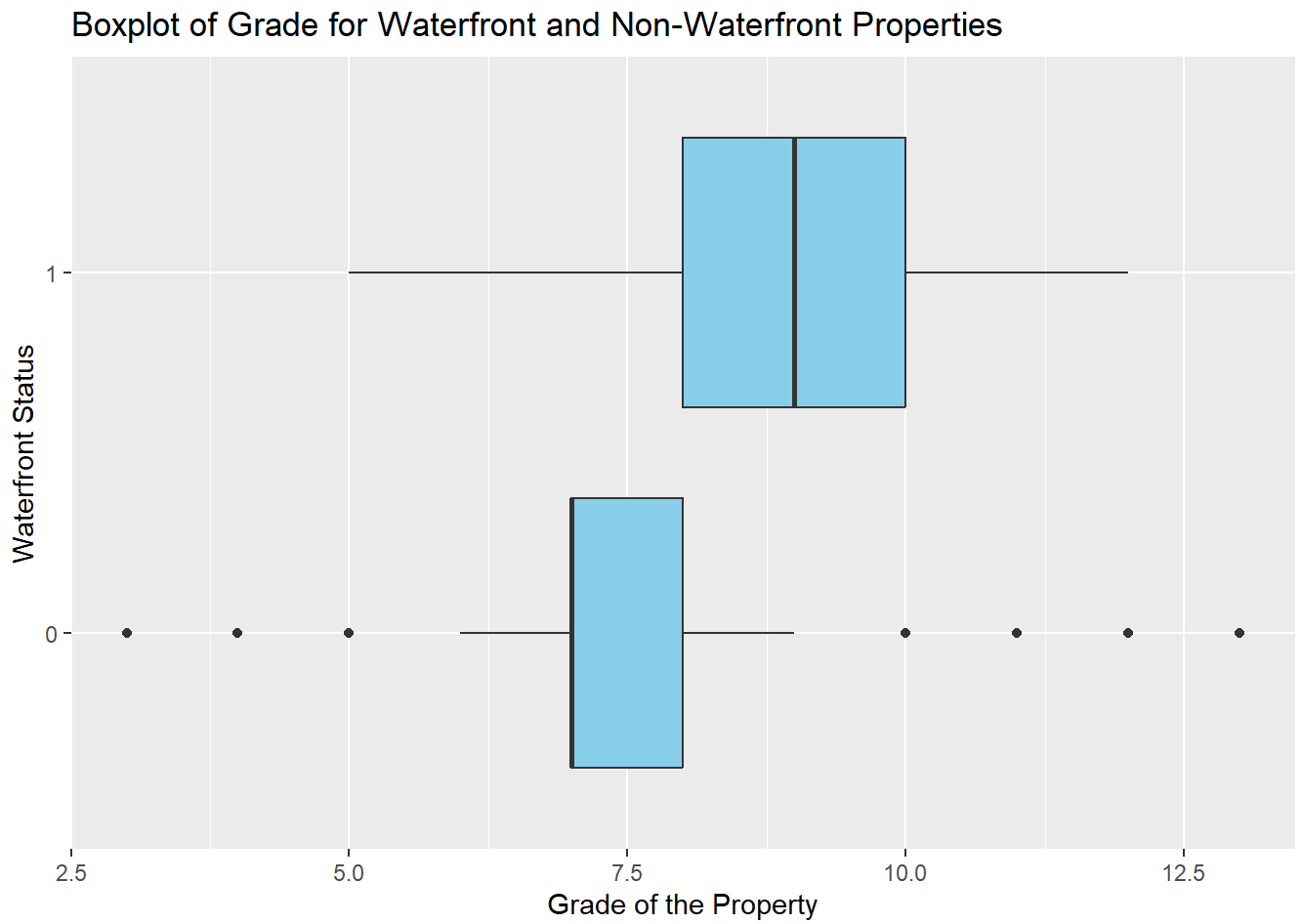
.



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.

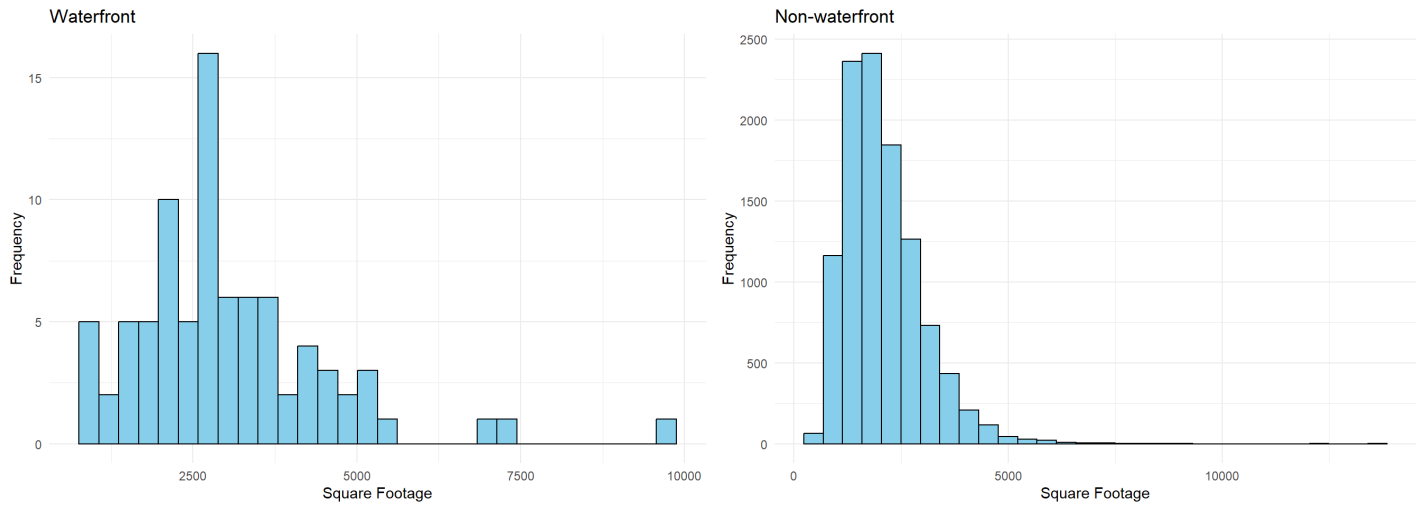
One other important thing to note from this visualization is the outlier with more than 30 bedrooms. This seems very unusual, and may be an error in input, especially because the square footage is relatively low. We may have to consider dropping this data point.



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.