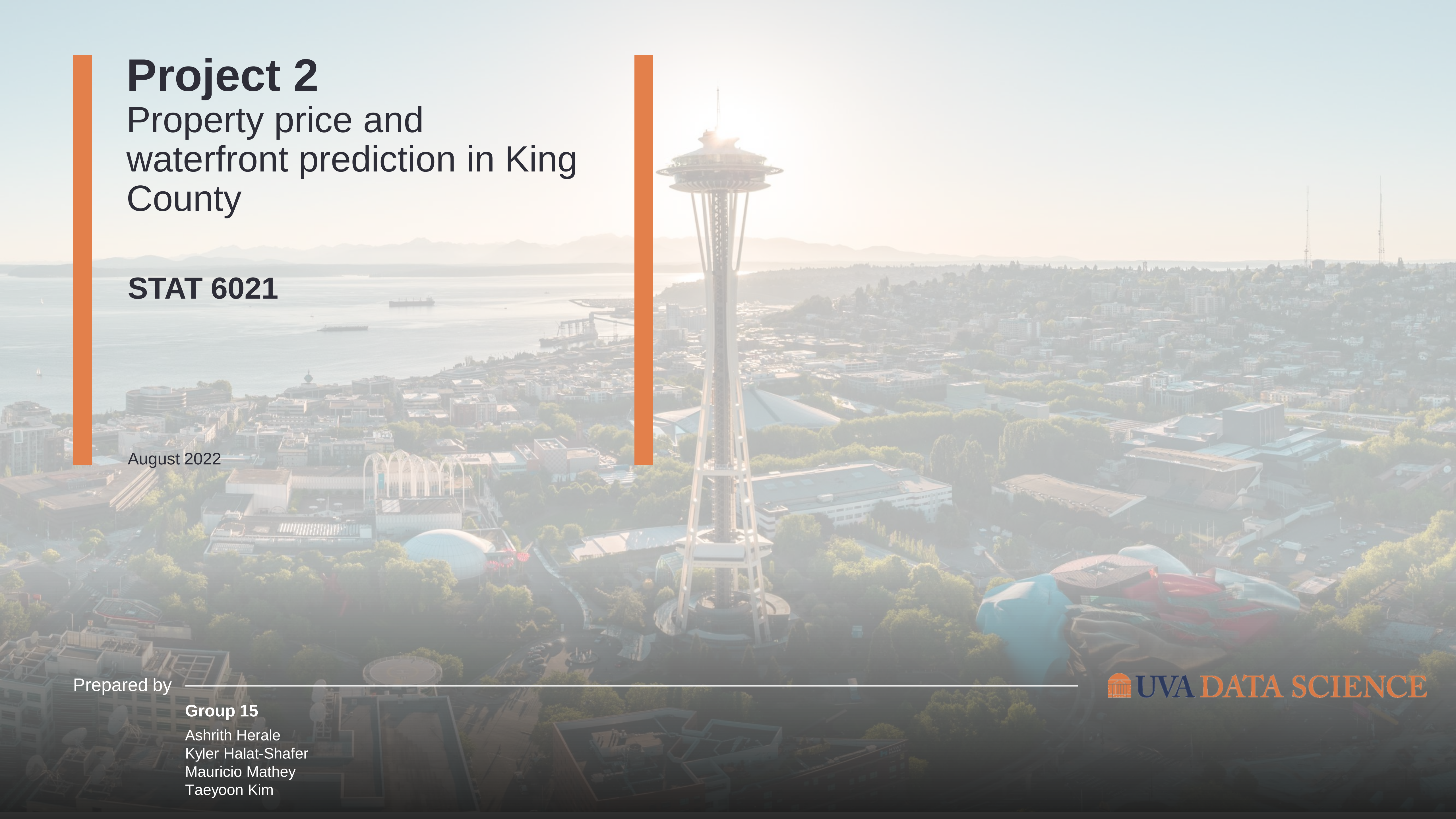


Project 2

Property price and waterfront prediction in King County

STAT 6021

August 2022

Prepared by

Group 15

Ashrith Herale
Kyler Halat-Shafer
Mauricio Mathey
Taeyoon Kim

UVA DATA SCIENCE

Agenda

- **Research questions and rationale**
- Exploratory data analysis
  - Price
  - Waterfront
- Modeling results
  - Price – Linear regression
  - Waterfront – Logistic regression

For this project we are looking to predict the property price and wether or not a property has a waterfront by using linear regression and logistic regression

Motivation for our linear regression question

- **Question**
  - What predictors have the strongest explanatory value on price?
- **Motivation**
  - In terms of a typical home, we are told 'Location, Location, Location’
  - But is this true? Will location or other factors have the largest impact on price?
- **Preliminary insights**
  - The best way to predict price is knowing the grade, square footage of the home, the income level of the zip code of the home, and whether that home is a waterfront property.

Motivation for our logistic regression question

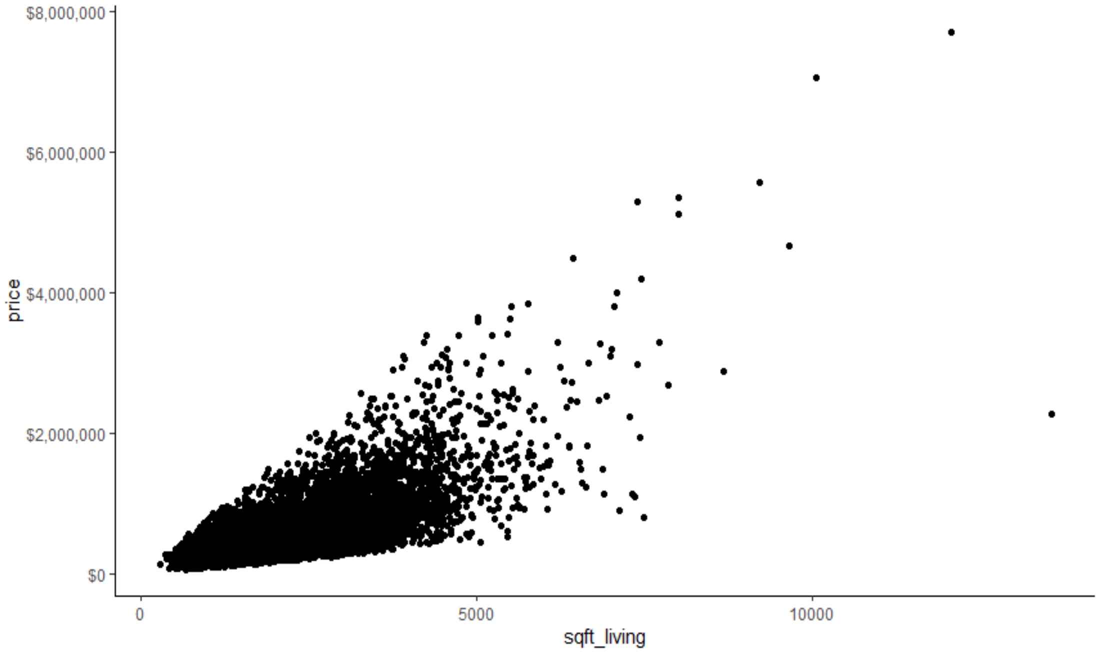
- **Question**
  - Looking at a very specific location, waterfront properties; are there characteristics that can tell you what is typically seen in a waterfront property?
- **Motivation**
  - Will price be a strong indicator that people are willing to pay more for a desirable location?
- **Preliminary insights**
  - The key elements in predicting if a property is waterfront or not are price, the square footage of the living area, the median income bracket, and the year the house was built.

Agenda

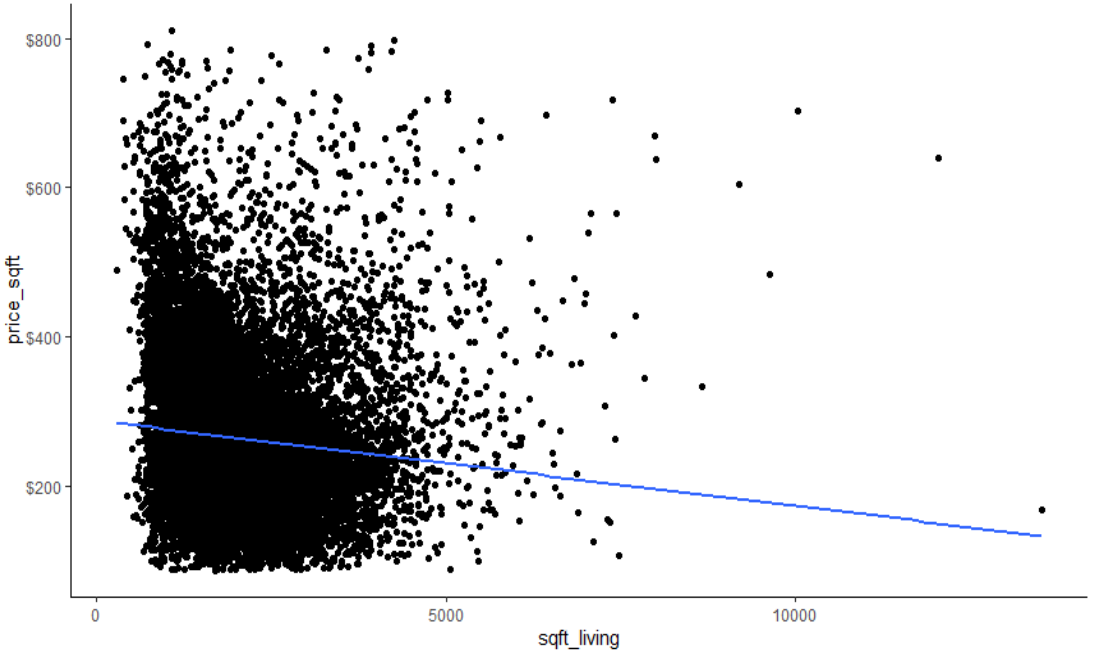
- Research questions and rationale
- **Exploratory data analysis**
  - **Price**
  - Waterfront
- Modeling results
  - Price – Linear regression
  - Waterfront – Logistic regression

As expected, there is a clear relationship between price and living area of the property and there appears to be a slight trend towards a decrease in price per square foot as living area increases

Price by living area (sq.ft.)



Price per square foot by living area (sq.ft.)

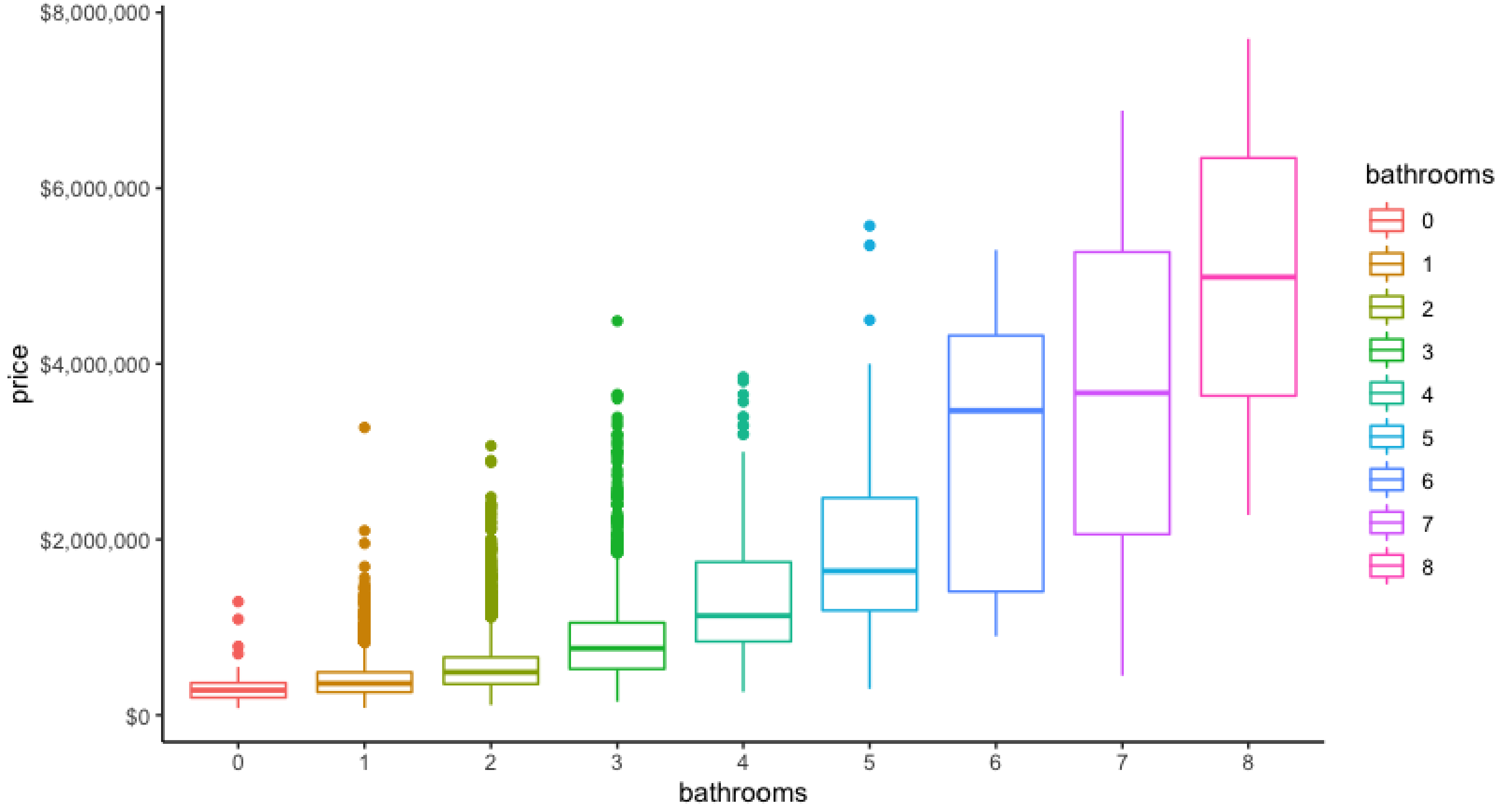


Key findings

- There is a clear relationship between price and living area; nonetheless, there appears to be a change in the variance of the data as we move towards higher price points which would indicate that a transformation on price is needed and probably (to be confirmed) a transformation on living area too
- A slight trend towards a decrease in price per square foot is observed, which means that bigger properties tend to compensate size by carrying smaller unit prices to maintain market value attractiveness despite their high price

Bathrooms appear to be a predictor for property prices, showing a clear trend between number of bathrooms and property value, the same relationship was not observed with bedrooms

Distribution of price by number of bathrooms

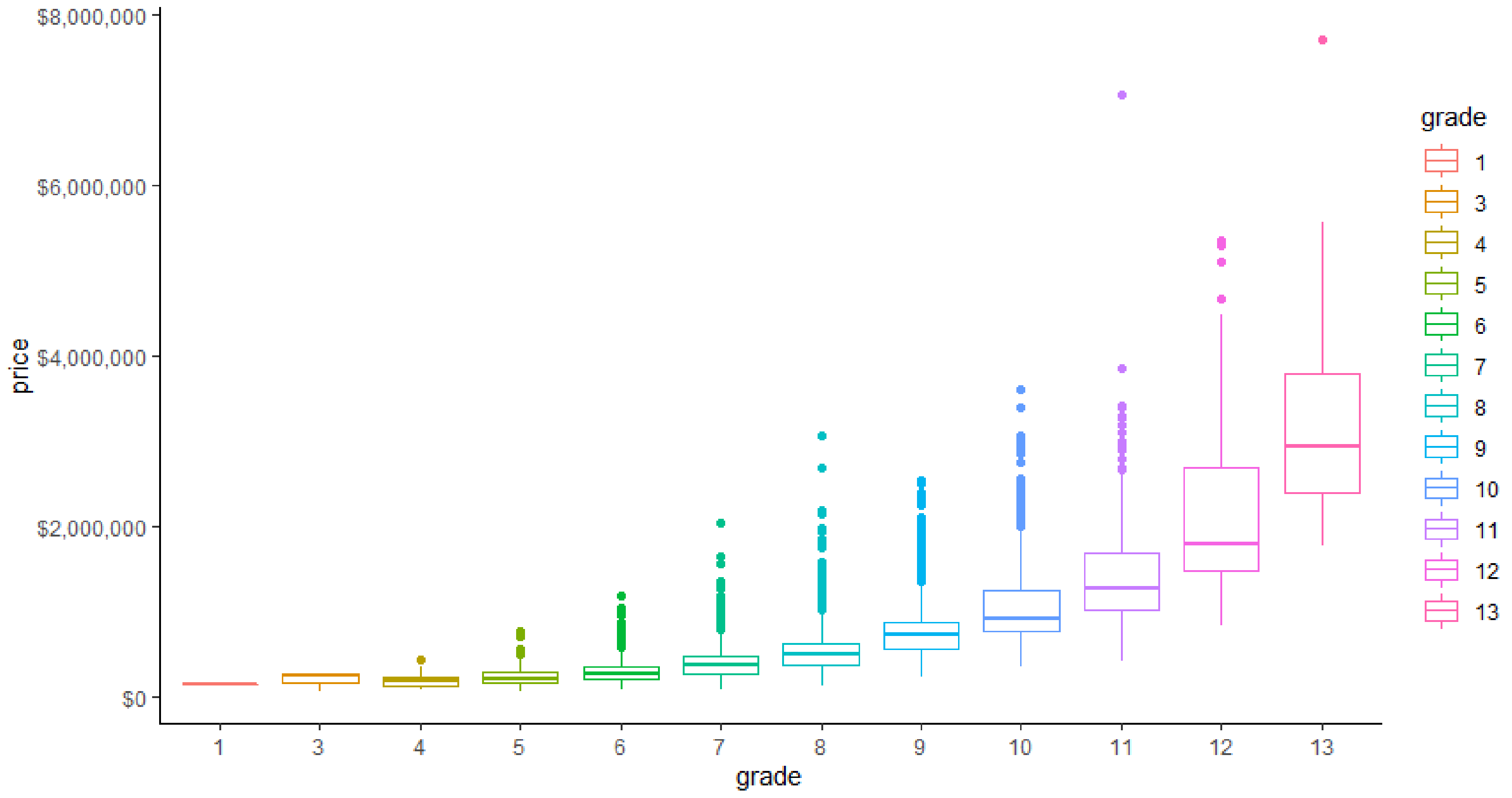


Key findings

- A relationship between number of bathrooms and price can be observed along all number of bathrooms
- 75% of properties that have 3 or 4 bathrooms exceed the value of more than 75% of the properties that have 1 bathroom
- A similar trend can be observed in properties that have 5 bathrooms, almost 75% of these properties exceed the value of properties with 4 bathrooms
- A higher dispersion can be observed in the price of properties with 6, 7, and 8 bathrooms.
- Properties with 6 and 7 bathrooms have similar median prices and distributions
- Properties with 8 bathrooms do carry a higher median price, but there is a significant overlap in prices with the properties with 7 bathrooms

A higher construction grade is associated with a higher price; nonetheless, we can observe higher dispersion in the prices as we move towards a higher grade

Distribution of price by grade

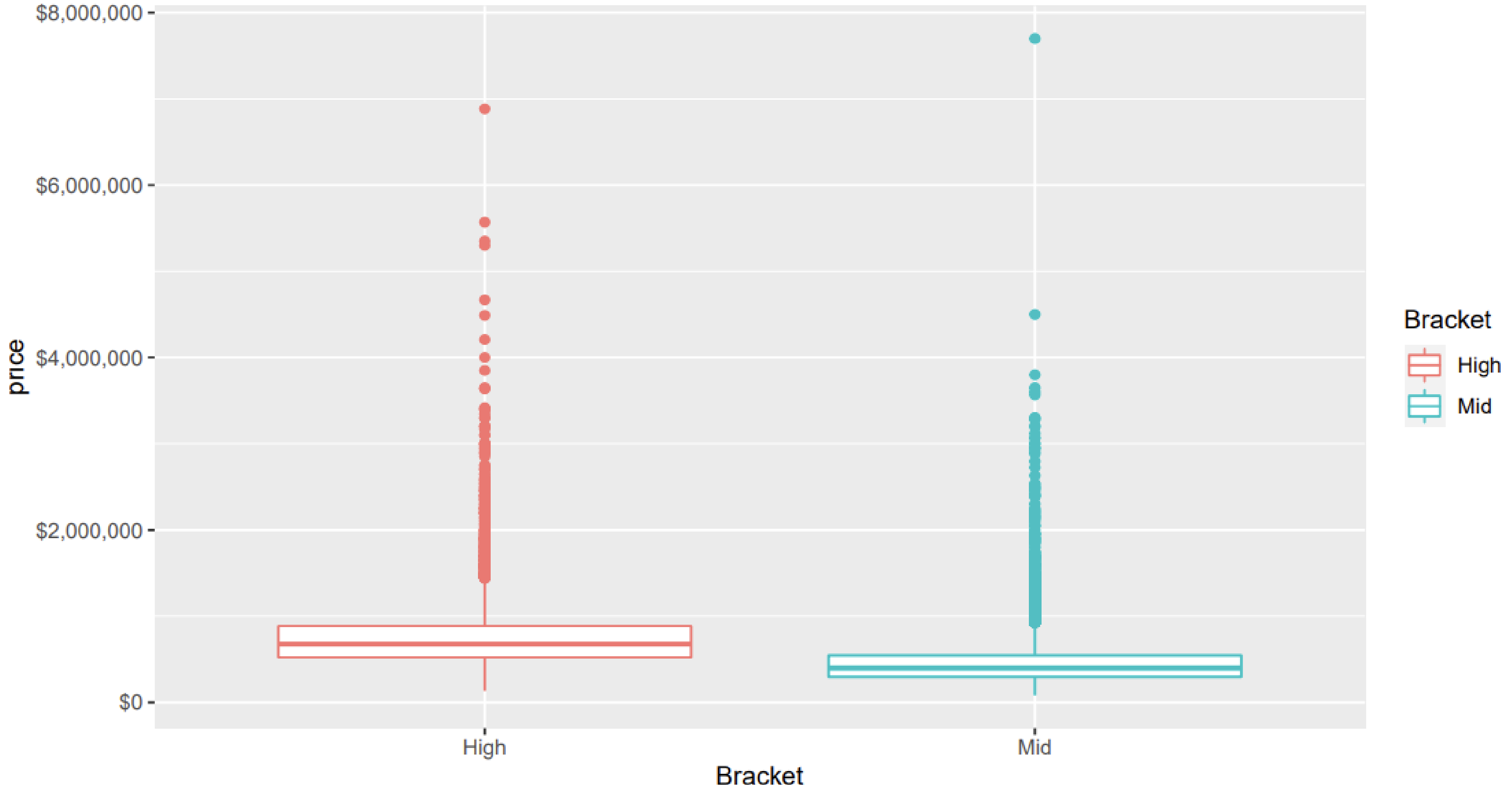


Key findings

- From grade 1 to grade 6 there appears to be no significant difference in the median price and the distribution, the only difference appears to be some outliers in 5 and 6 with higher prices
- Starting at grade 7 we start to see a significant increase in the median price, where the next grade tends to have a median price that is equal to at least the 75th percentile of the previous grade
- From grade 10 onwards, we also observe an increase in the variability of the prices of the properties

Properties located in zip codes that are considered high income carry a higher median price than properties that are in zip codes that are considered mid income

Distribution of price by income bracket



Key findings

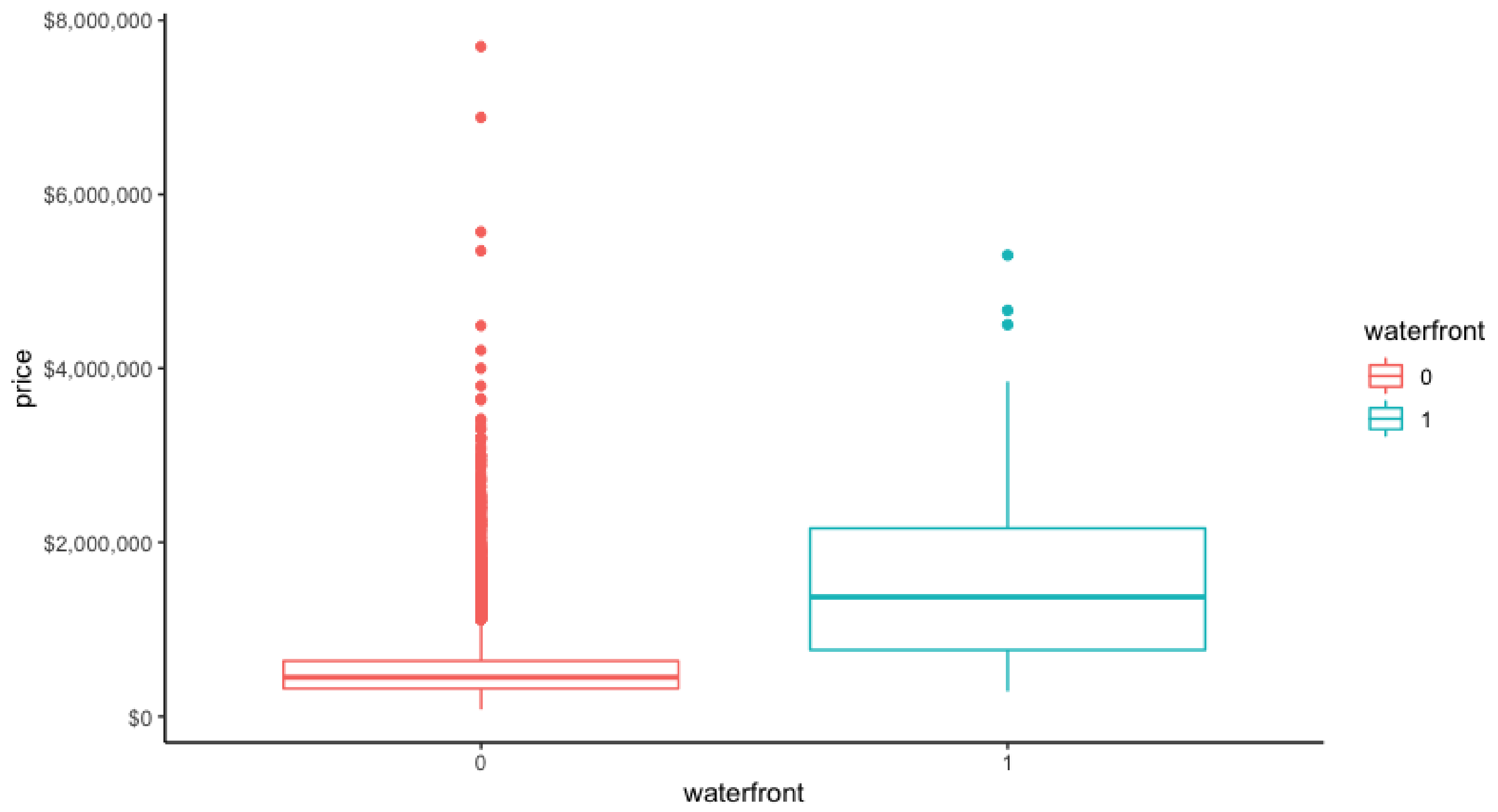
- 75% of the properties located in high income zip codes exceed the value of ~75% of the properties located in mid oncome zip codes
- Both income brackets present significant outliers, and the most expensive property is located in a mid income zip code

Agenda

- Research questions and rationale
- **Exploratory data analysis**
  - Price
  - **Waterfront**
- Modeling results
  - Price – Linear regression
  - Waterfront – Logistic regression

As explained in the previous section, waterfront properties carry a higher median price than properties that do not have a waterfront as well as a bigger interquartile range

Distribution of price by waterfront

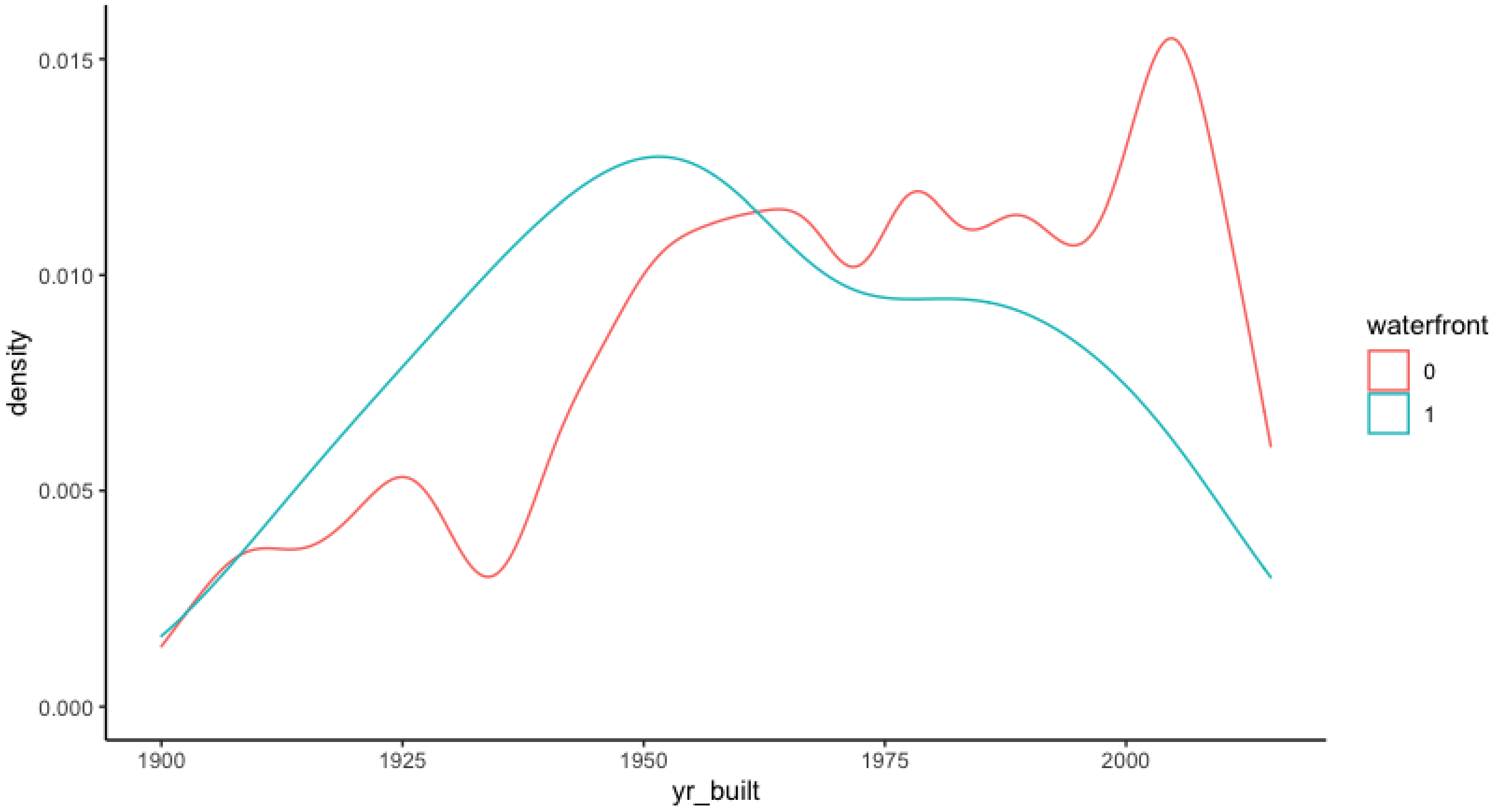


Key findings

- Properties with a waterfront carry a significant higher price point than the ones that don’t have a waterfront
- To the extent that approximately 75% of properties that have a waterfront exceed the price of 75% of the properties that do not have a waterfront
- Nonetheless, some of the non-waterfront properties present the highest prices in the study population, with a significant number of houses being considered outliers

Waterfront properties tend to have been built earlier than non-waterfront properties, which makes sense given the limited space of water facing land

Distribution of year built by waterfront



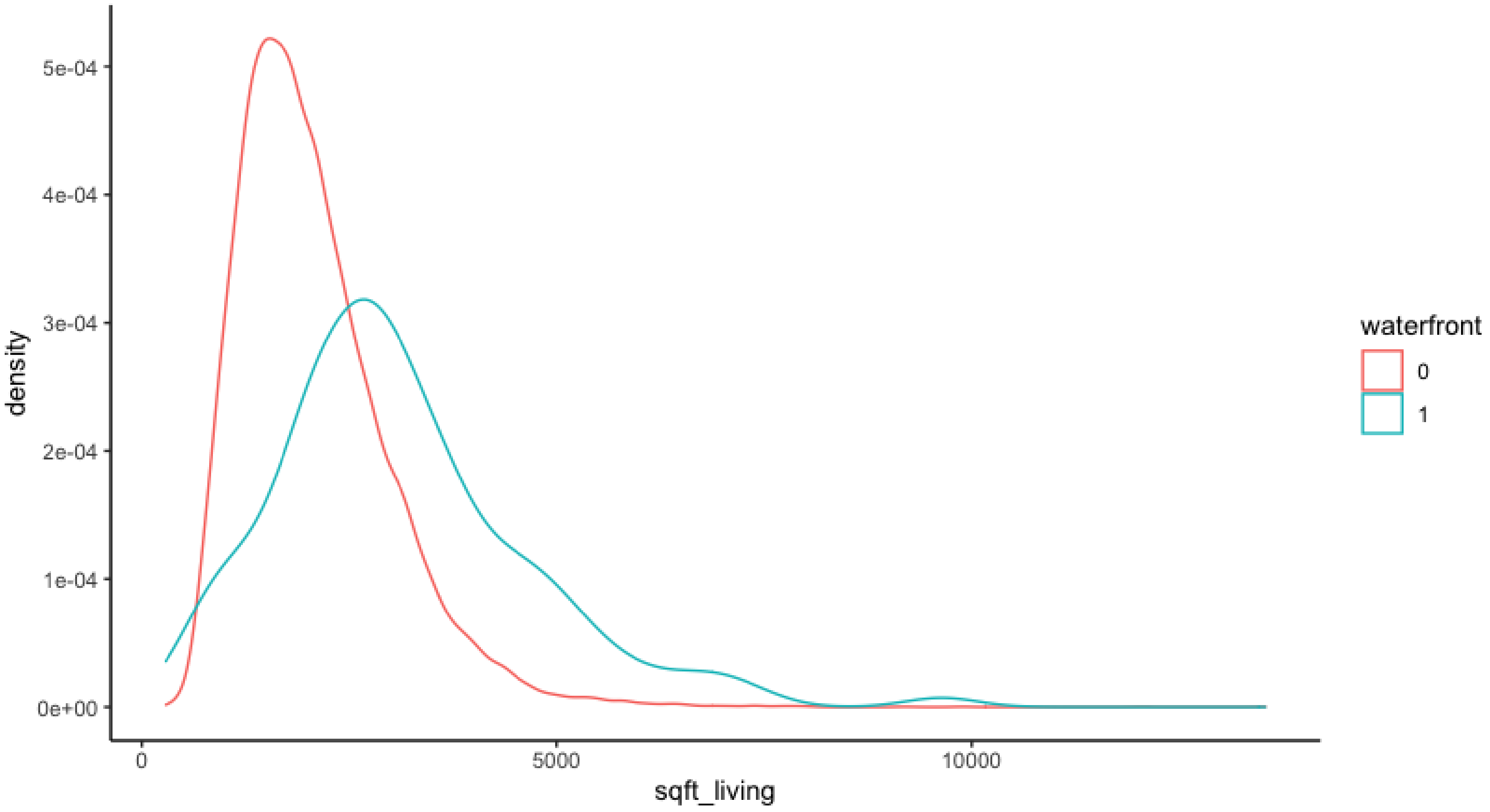
Key findings

- Up until the ~1915s there was no apparent trend in waterfront vs non-waterfront properties
- Only since ~1915 we start to see waterfront property construction maintaining a steady growth, peaking at around the 1950s and since then started a decline, with slow down around the 1975-1980 and then resuming the decline
- Non-waterfront properties tend to have been constructed mostly between the 1960s and 2000, reaching a peak in ~2005 then having a decline

Exploratory data analysis - Waterfront

While waterfront properties tend to be larger than non-waterfront ones, in both cases we see a trend towards properties below 8,000 sq.ft.

Distribution of living area (sq.ft.) by waterfront



Key findings

- Waterfront properties tend to be 50% larger than non-waterfront properties
- Waterfront properties tend to be below 8,000 sq.ft. while non-waterfront usually are below 5,000 sq.ft.
- Overall, properties with a living space of over 8,000 sq.ft. are uncommon independent of their location

Agenda

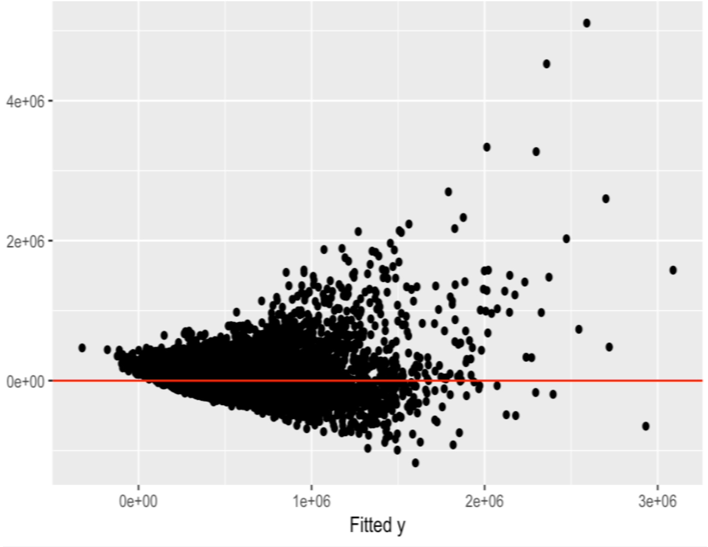
- Research questions and rationale
- Exploratory data analysis
  - Price
  - Waterfront
- **Modeling results**
  - **Price – Linear regression**
  - Waterfront – Logistic regression

With our linear regression model, we were able to predict the property price based on living area, income level, grade and waterfront indicator with an R square of 61%

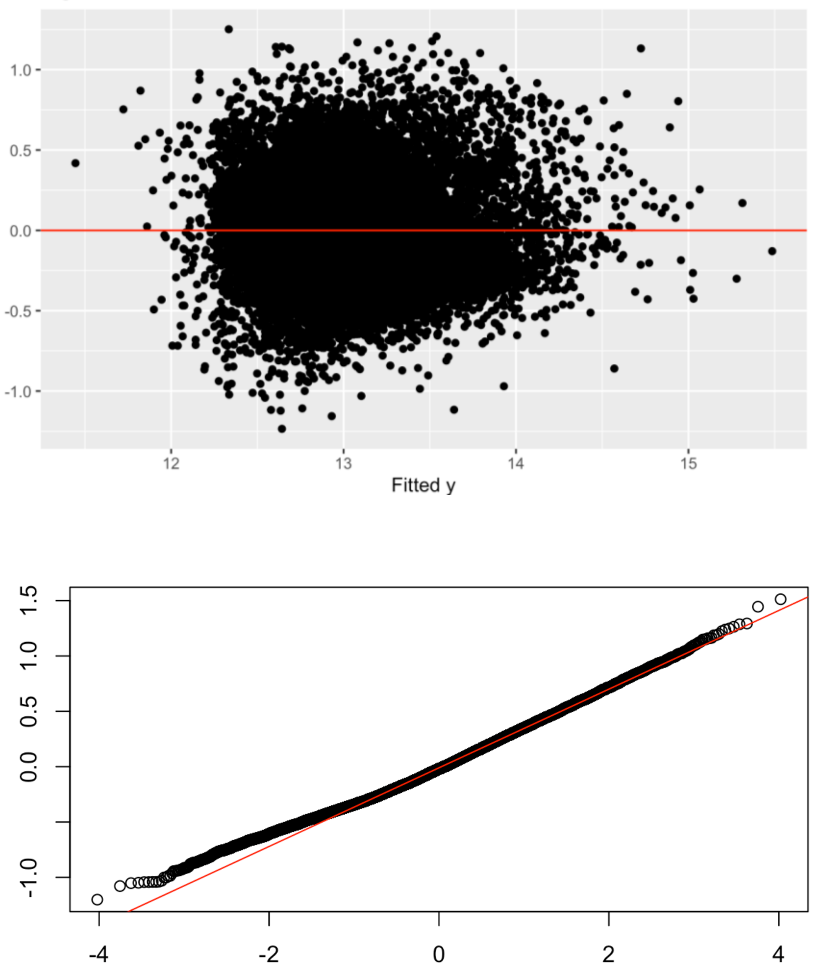
Variables

- Dependent variable
  - Log(Price)
- Predictor variables
  - Grade
  - (Living area (in sq ft))$^{1/3}$
  - Median Income level by zip code (high/ medium/ low)
  - Waterfront property (Yes/ No)

Basic model Residual Plot



Final training model residual analysis



Final model coefficients

Predictor	Coefficient	2.5%	97.5%
Grade	0.16	0.15	0.16
(Living area)$^{1/3}$	0.1	0.09	0.1
Income level :low	-0.36	-0.37	-0.34
Income level : mid	-0.18	-0.2	-0.17
Waterfront home: Yes	0.72	0.66	0.78

Model testing results

- Testing the utility of the model
  - R Square pred : 61.3 %
  - MSE (training) : 0.1
  - MSE (test) : 0.1
- As the pairs R Square and R square pred ,and Mean square of errors for the training and test datasets are nearly the same, we conclude that there is no overfitting in the finalized model
- All linear regression assumptions hold true
- Model created after removing influential, high leverage and outliers data is not significantly different

Agenda

- Research questions and rationale
- Exploratory data analysis
  - Price
  - Waterfront
- **Modeling results**
  - Price – Linear regression
  - **Waterfront – Logistic regression**

With our logistic regression model, we were able to predict based on price, living area, income bracket, and year built if the property has a waterfront with a sensitivity of 89%

Variables

- Dependent variable
  - Waterfront
  - 0 for non-waterfront
  - 1 for waterfront
- Predictor variables
  - Price
  - Living area (sq.ft.)
  - Income bracket (high or medium)
  - Year in which the property was built

Training model results

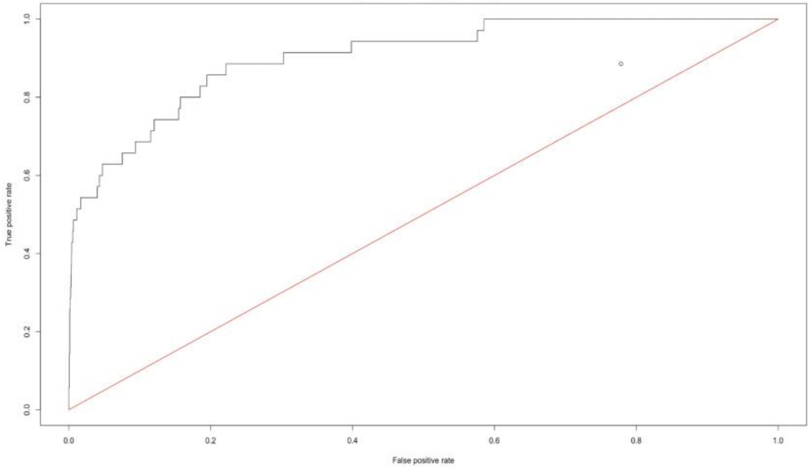
- Testing the utility of the model
  - Ho : B1 = B2 = B3 = B4 = 0
  - Ha : at least one of the coefficients in H0 is not zero
  - We received a low p-value, we reject the null hypothesis and conclude the model is useful
- Model coefficients

Predictor	Coefficient	Odds	Probability
Price	3.26e-6	1.00[1]	0.5
Area	-8.75e-4	1.00[2]	0.5
Middle Income	1.063	2.90	0.7
Year built	-6.988e-3	1.00[2]	0.5

Testing model results

- Testing the utility of the model
  - Considered a threshold of 0.006 due to the number of the waterfront properties
  - With this threshold we obtained the following test results
    - Overall error rate: 22%
    - Sensitivity: 89%
    - Specificity: 78%
- Given our selected threshold and our sensitivity and specificity, we are above the diagonal in the ROC curve, confirming the predicting capacity of our model

Figure 46: ROC Curve (Waterfront)



[1] Due to rounding to two decimals it rounds to 1 but in the model the value is slightly above 1, this is due to the scale of the number compared to the other variables
[2] Due to rounding to two decimals it rounds to 1 but in the model the value is slightly below 1, this is due to the scale of the number compared to the other variables