Real Estate Statistical Analysis Technical Presentation

Presented by Zach Summy
February 8th, 2021

A government agency wants to explore opportunities in King County, WA. They will look at all homes sold in the county between May, 2014 – May, 2015 and nineteen variables ranging from price and year built to the square footage of the nearest fifteen neighbors.

Introduction

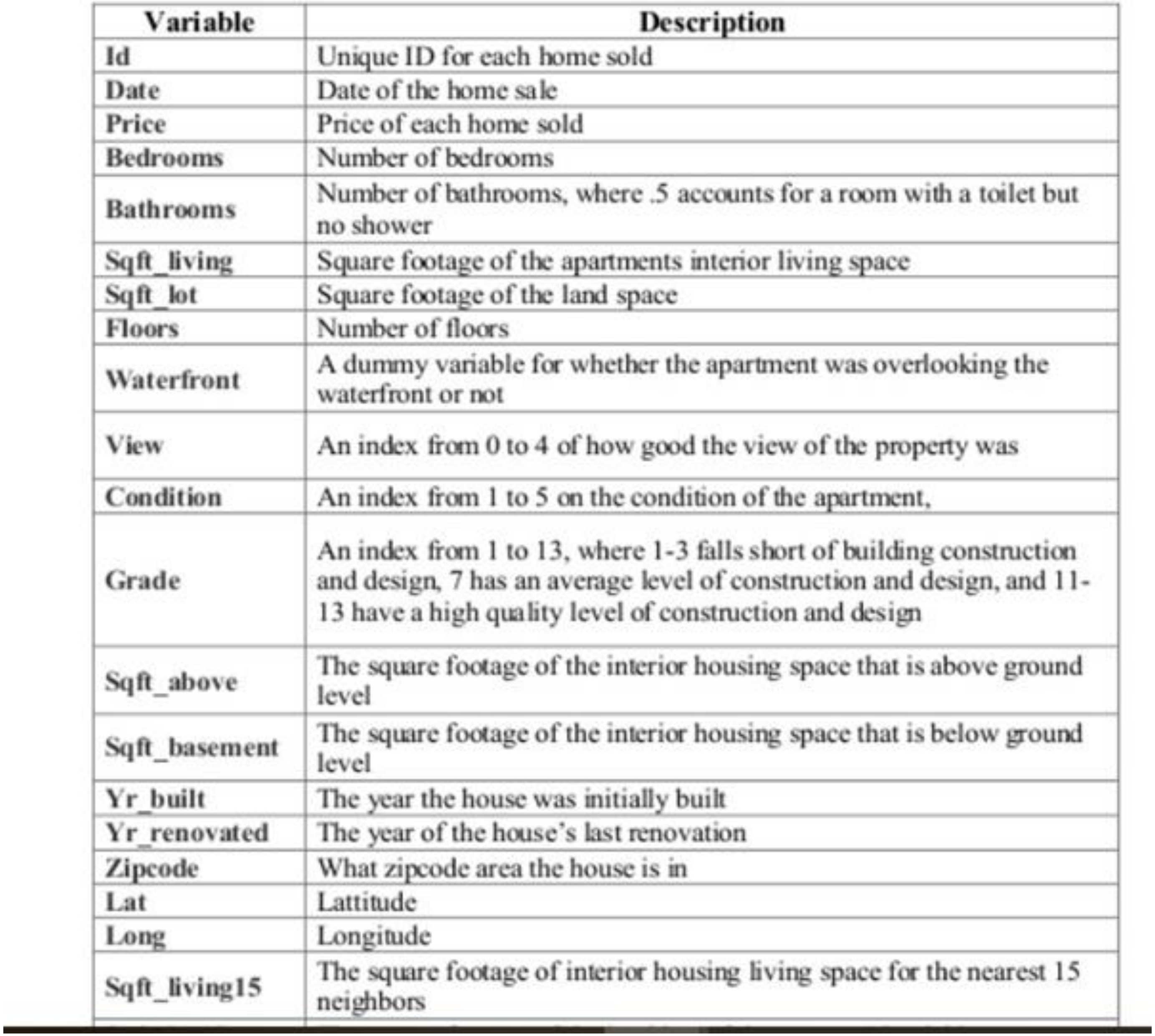
To assist the government agency we sought to answer three questions:

1. Which variables are significant in predicting the real estate price of a home?

2. How do variations in those variables affect price?

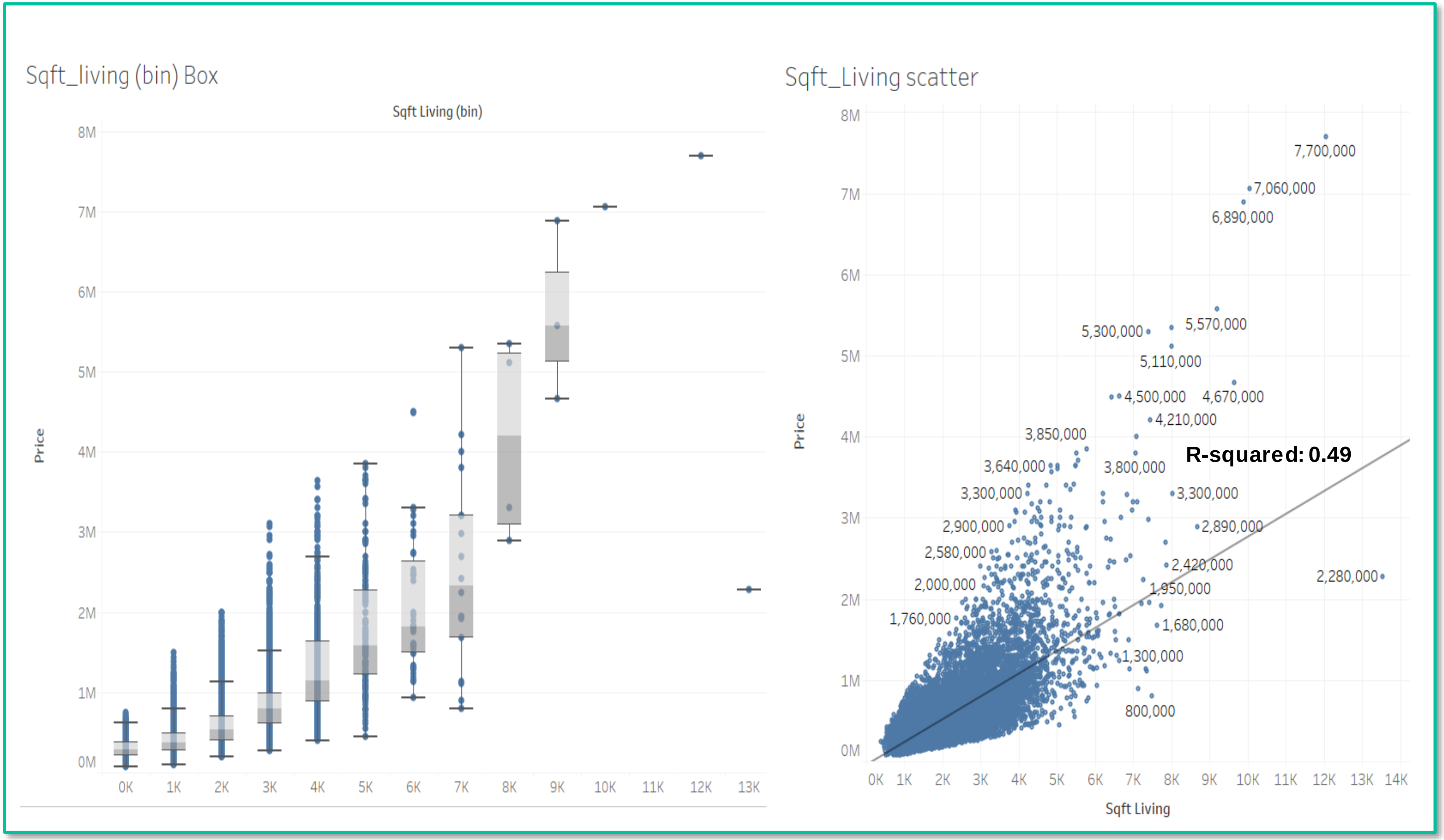
3. Can a predictive model be built, and if so, what is it?

The Variables



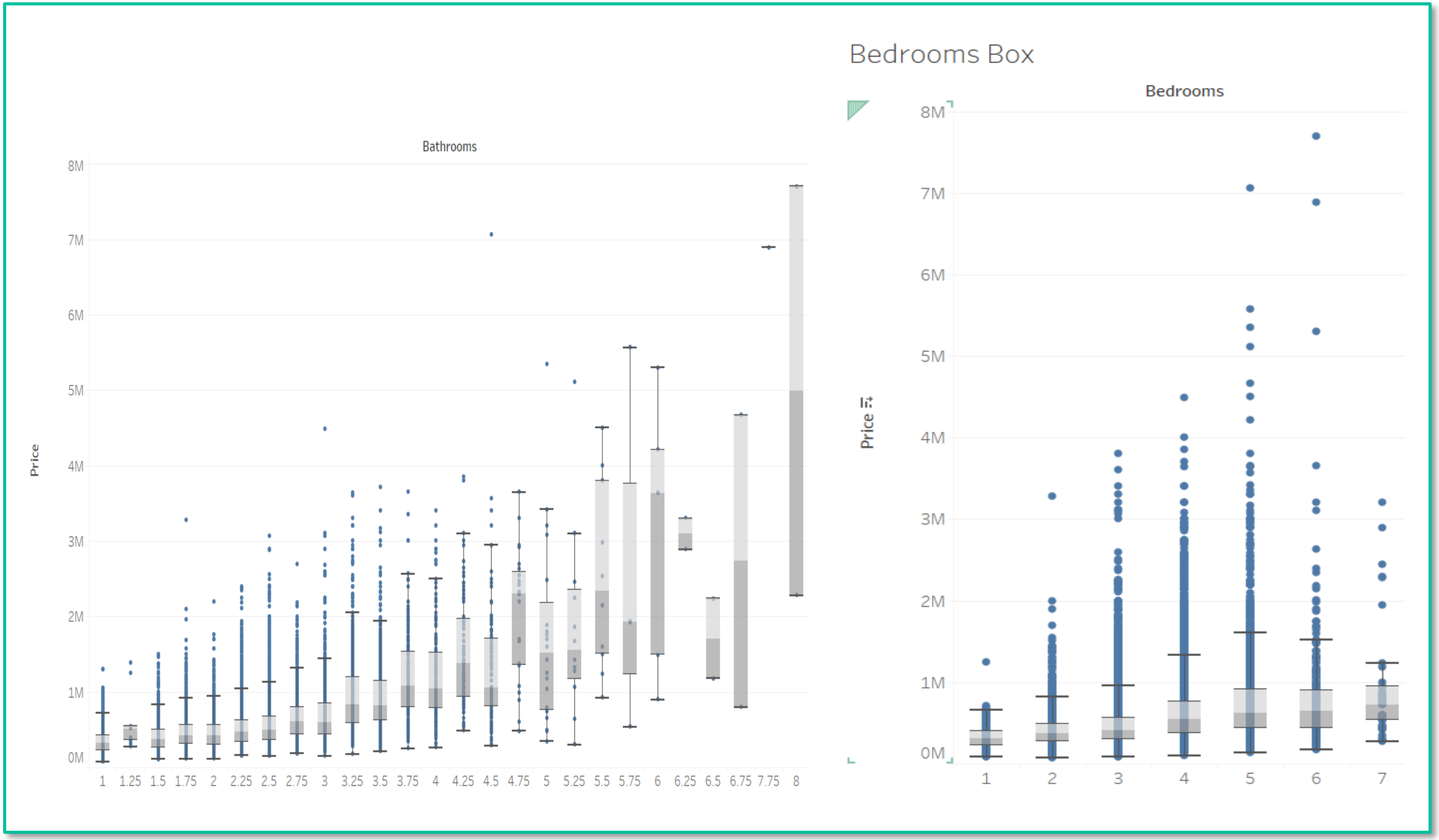
Variable	Description
Id	Unique ID for each home sold
Date	Date of the home sale
Price	Price of each home sold
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms, where .5 accounts for a room with a toilet but no shower
Sqft_living	Square footage of the apartments interior living space
Sqft_lot	Square footage of the land space
Floors	Number of floors
Waterfront	A dummy variable for whether the apartment was overlooking the waterfront or not
View	An index from 0 to 4 of how good the view of the property was
Condition	An index from 1 to 5 on the condition of the apartment,
Grade	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design
Sqft_above	The square footage of the interior housing space that is above ground level
Sqft_basement	The square footage of the interior housing space that is below ground level
Yr_built	The year the house was initially built
Yr_renovated	The year of the house's last renovation
Zipcode	What zipcode area the house is in
Lat	Lattitude
Long	Longitude
Sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors

There is a high* positive relationship between Sqft_living and price on the boxplot below. The higher the square footage the higher the median price and the greater range, too. The scatter plot shows a trend with R-squared 0.49.

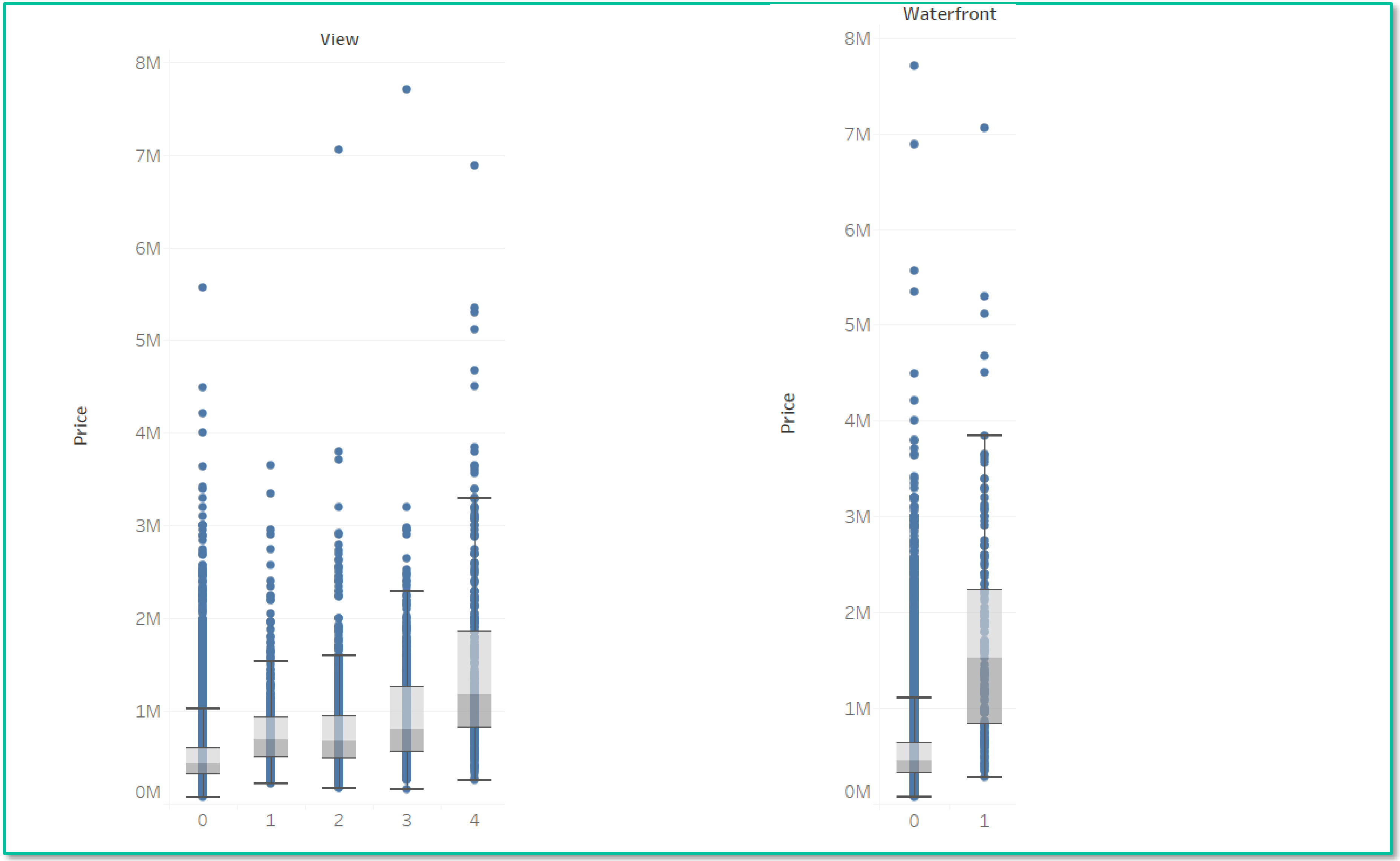


*Scale for correlation: Low< 0.4; 0.4 <= Moderate < 0.7; High >= 0.7

As number of bathrooms increases the price and range of price steadily increases. The boxplots of bedrooms is not as dramatic as bathrooms but the median price for houses with 1 and 7 bedrooms slowly increases from $310k to ~$730k, respectively. The largest increase is ~33% from 3 bathrooms to 4 bathrooms.

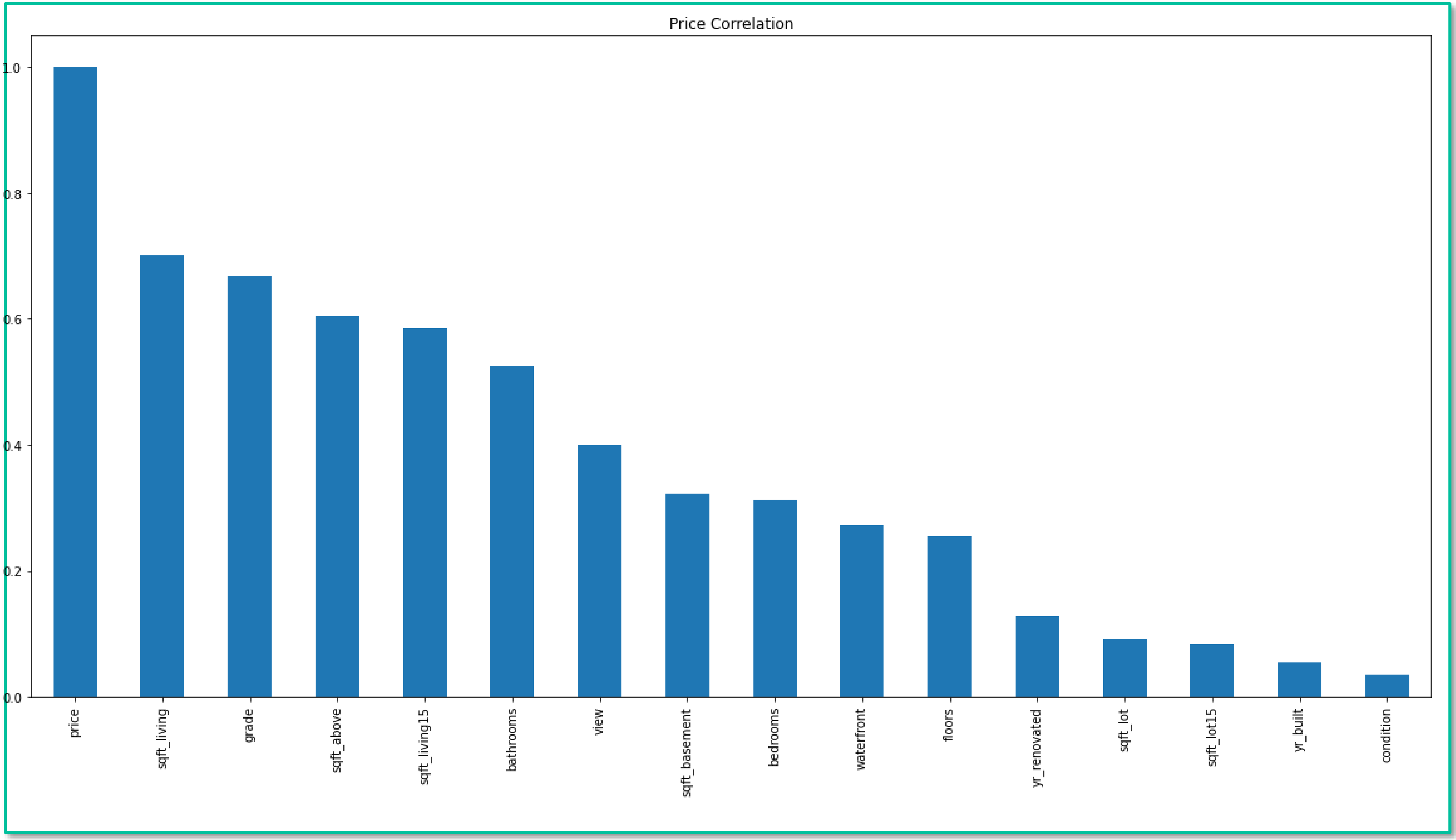


View is moderately*, positively correlated with price. The largest is a 48% median increase between a 3 and 4 rating. View and Waterfront are themselves correlated 0.4. A waterfront property has a median price 2.3 times that of a property not on the water.

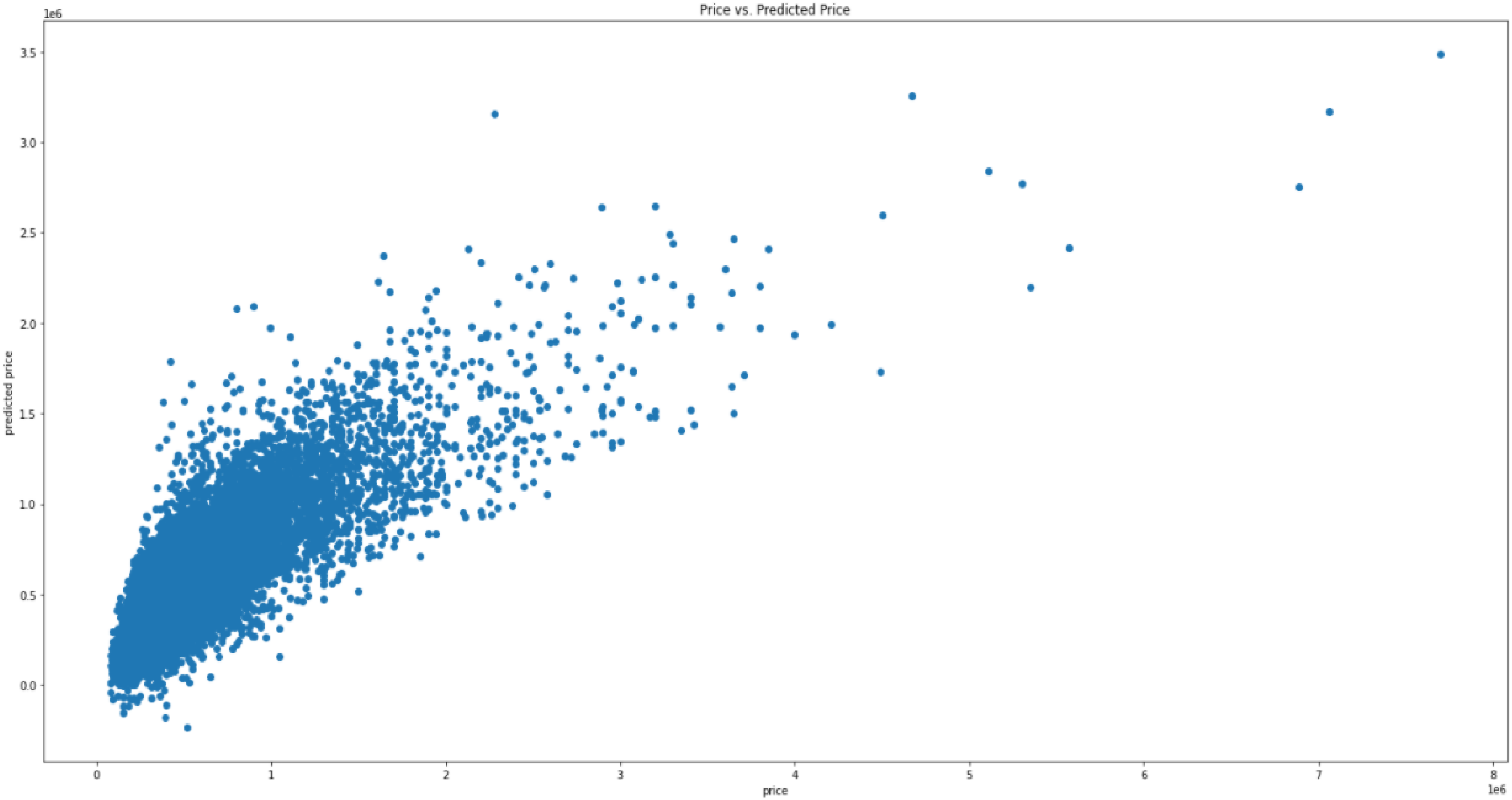


*Scale for correlation: Low< 0.4; 0.4 <= Moderate < 0.7; High >= 0.7

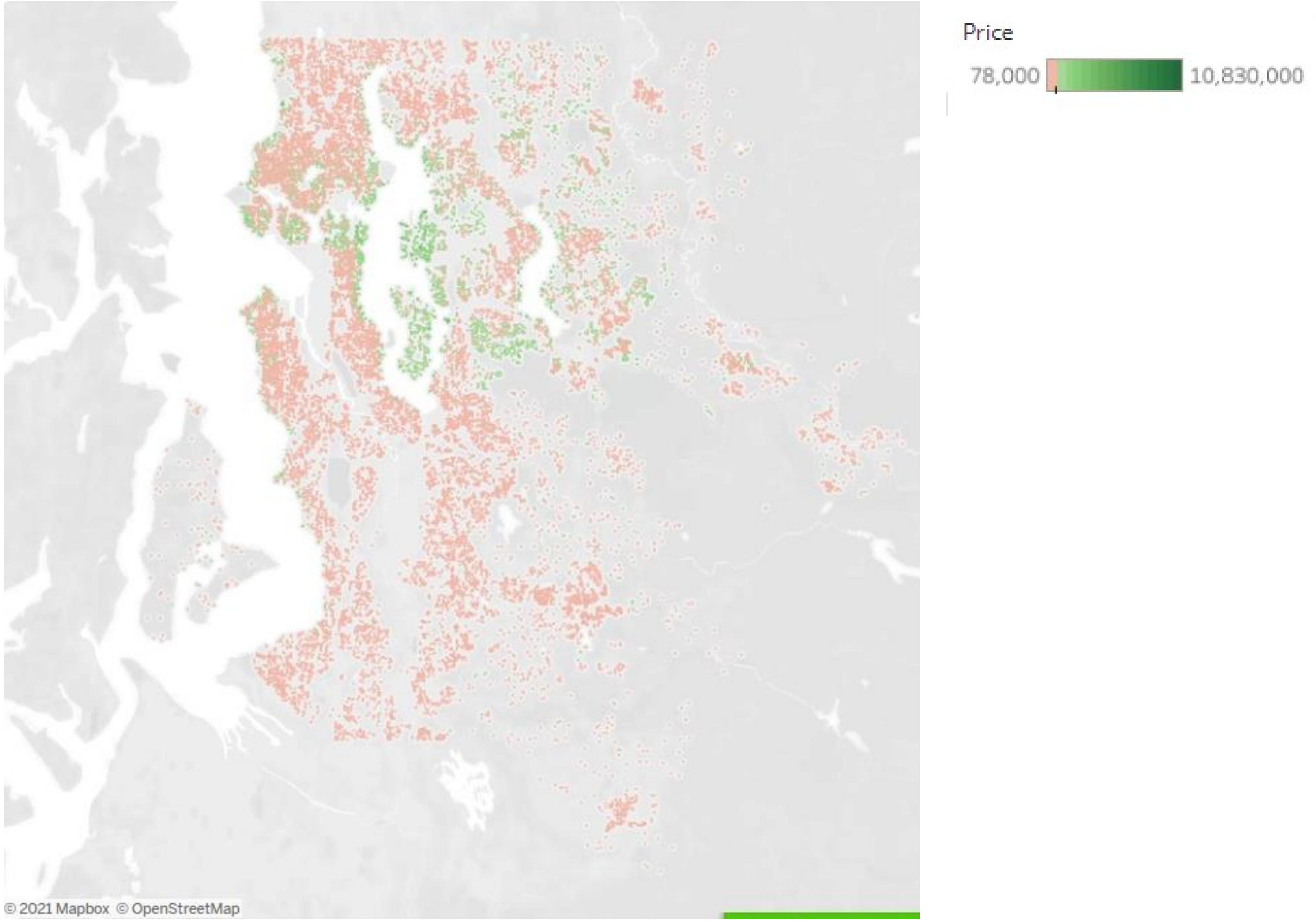
The most important factor corresponding with price is Sqft_living, Grade, Bathrooms, View, Bedrooms, and Waterfront; Sq Ft Above, Sqt Ft of Living 15, Sq Ft of Basement are not counted since they wind up being a different way of expressing Sqft_Living which we do not want to double count



We eliminated the cross-correlations and smaller variables: id, date, lat, long, zipcode, sqft_above, sqft_living15. The variables remaining are sqft_living, grade, bathrooms, view, sqft_basement, bedrooms, waterfront, floors, yr_renovated, sqft_lot, condition, sqft_lot15. The plot predicts price with a high accuracy, R-squared: 0.655.



Out of curiosity we can see the location and price of homes based on latitude and longitude: the higher priced homes are near the water (Lake Washington and the Bay).



Conclusion

- Sqft_living, bedrooms, bathrooms, view, and waterfront are all positively affected by price

- After removing cross-correlations the highest six correlated variables with price are Sqft_living, Grade, Bathrooms, View, Bedrooms, and Waterfront

- Sqft_living by itself has an R-squared value of 0.49 with price

- Adding the variables sqft_basement, floors, yr_renovated, sqft_lot, condition, and sqft_lot15 we can model and predict the price with high accuracy and a R-squared of 0.655