RANDOM FOREST NOTES

* Read the introduction and take notes of the factors left unused and the reason why
* The Table of Contents lists the steps taken to build the model
* Use a 3:1 train:test ratio
* The features used are 'bedrooms', 'bathrooms', 'sqft\_lot', 'waterfront', 'view', 'condition', 'grade', 'sqft\_above', 'sqft\_basement', 'yr\_built'
* r^2: Coefficient of determination or the amount of variance in the response variable (price) that can be explained by variance in one or multiple independent variable(s)
* Results in “Test & Measure Model” section:

**Train: Random Forest r^2 Accounted: 0.9634**

**Comment:** Total r^2 is close to 1, so when using that many factors, the

model is very accurate at predicting the price

**Test: Random Forest r^2 Accounted: 0.7259**

**Comment:** On test data (that it has not already seen), unsurprisingly, the model predicts the price less accurately. But a coefficient of determination of 0.7 is moderately high. For that coefficient, the model is useable in real industry, especially if more features were to be input for the coefficient to be even higher.

Graph:

**Comment:** The grade or construction quality of the home is the most significant influencer of the price, accounting for almost half of the variance. In contrast, the size of the property accounts for less than a measly 10%. It appears for this region, ignoring potential biases in how the dataset was collected, quality matters more than quantity. The least significant feature is the number of bedrooms, which mattered less than the number of bathrooms.

Suggestion for Improvement: Add more features to the model. As this dataset covers only a single year, the influence of the economy and market are unlikely to have affected the price (response variable).