## **1. Describe your research scenario and question(s).**

This project looks at what drives housing prices in the Seattle area. The dataset that we are analyzing from Kaggle is a fairly large and comprehensive dataset that includes a wide range of information about homes in Seattle and the surrounding area. Columns in the dataset include information about a home’s size, their characteristics, price, and condition. These details give us a solid foundation to explore what really impacts the price of a house.

The main goal is to answer a straightforward but important question:

What factors have an impact on housing prices, and how well can we predict those prices using a linear regression model?

To answer this question, we developed and compared several models. Throughout the process, we focused on the following key questions:

* Which variables are statistically significant in predicting housing prices?
* Does including additional variables improve or hurt our model?
* Is a linear regression model appropriate, and do we meet all of its assumptions?

By exploring these questions, we aimed to not only build a model that predicts housing prices well, but also analyze if our model is effective or not.

## **2. Describe the data set.**

Data source: <https://www.kaggle.com/datasets/sukhmandeepsinghbrar/housing-price-dataset>

• **id**: A unique identifier for each property.

• **date**: The date the property was listed for sale.

• **price**: The property’s selling price (this is the target variable for prediction).

• **bedrooms**: Total number of bedrooms.

• **bathrooms**: Total number of bathrooms.

• **sqft\_living**: The interior living space of the house, measured in square feet.

• **sqft\_lot**: The total area of the land the property occupies.

• **floors**: Number of floors in the house.

• **waterfront**: Whether the property has a waterfront view (1 = Yes, 0 = No).

• **view**: A rating from 0 to 4 that reflects the quality of the view.

• **condition**: A rating from 1 to 5 indicating the overall condition of the property.

• **grade**: An overall rating (1 to 13) that combines construction quality and design.

• **sqft\_above**: The square footage of the house excluding the basement.

• **sqft\_basement**: The square footage of the basement area.

• **yr\_built**: The year the house was originally constructed.

• **yr\_renovated**: The year the house was last renovated (0 if it hasn’t been renovated).

• **zipcode**: The zip code indicating the location of the property.

• **lat**: The latitude coordinate of the property.

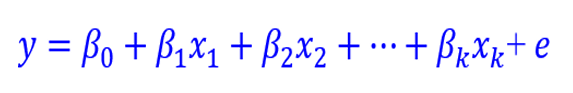
• **long**: The longitude coordinate of the property.

• **sqft\_living15**: The average living area (in sq ft) of the 15 nearest homes.

• **sqft\_lot15**: The average lot size (in sq ft) of the 15 nearest homes.

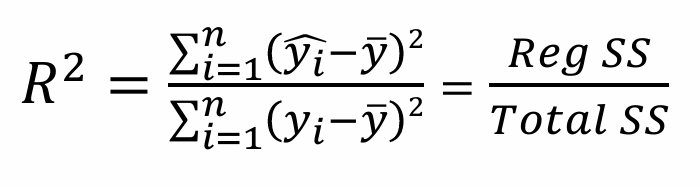
## **3. Describe the statistical methods you plan to use.**

**Multiple linear regression**

****

The slope parameter (𝛽𝑖) in a MLR gives the expected or average change in the response variable (𝑦) for a one-unit increase in the independent variable (𝑥𝑖) after controlling for the other independent variables

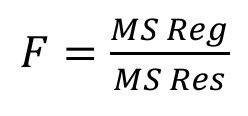
**Coefficient of Determination (R^2)**: Proportion of variance in y explained by the model. In MLR, it is not simply the square of a single correlation but is computed as



**Global F test:**

H0​: β1​ = β2​ = ⋯ = βk​ = 0

H1​: at least one βi ≠ 0. A significant F-statistic indicates that the model has explanatory power beyond a constant-only model.

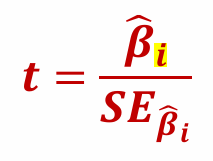


Reject H0 if 𝐹 ≥ 𝐹critical value. Also reject if p-value is less than the alpha value

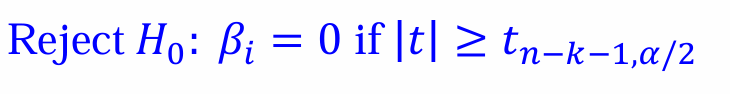
**MLR Inference t-test**

If the overall model is significant, then the significance could be attributed to any one of the independent variables. Then we should perform testing on each parameter to identify each independent variable's relative contribution.

In order to test each parameter after controlling for other independent variables in the model, we use the t-statistic.



And the decision rule is:



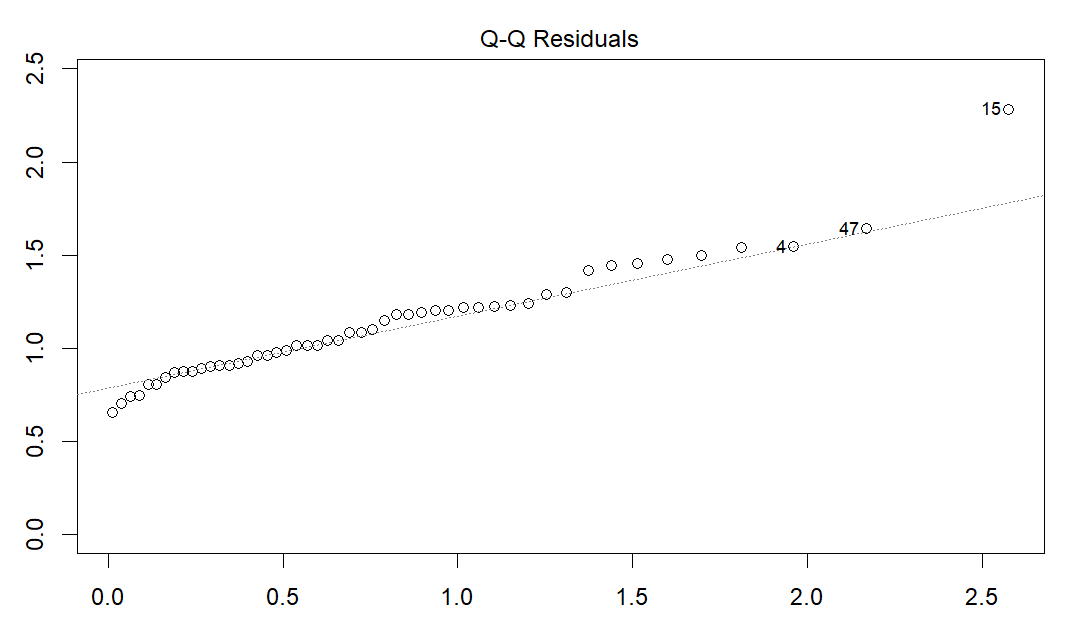
Where critical value is the value from the t-distribution with 𝑛 − 𝑘 − 1 degrees of freedom and associated with a right-hand tail probability of 𝛼/2.

**Linear Diagnosis: LINE method**

**Linearity:** Generate a residual scatterplot, with the trend expected to be roughly linear between the factors.

**Independence:** If we don’t have the same sample twice in the dataset, we can assume that it is independent. We are assuming that the selection of one sample does not influence the selection of another sample.

**Normal:** If all the data points is around the central line in our Q-Q plot, then the normality is checked



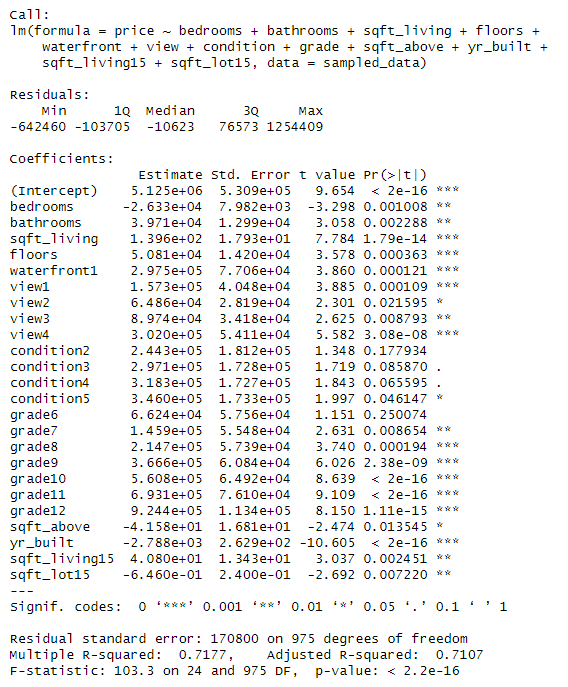
**Equal Variance:** The residual plot should show approximately the same amount of scatter from left to right (no fan shape).

## **4. Report your results.**

As an initial step before we did any work with our data, we cleaned up the data. We first examined the data for any unusual or null values. We found that one of the columns (yr\_renovated) was mostly 0s, since most of the houses had not been renovated, and made the decision to drop that particular column. Then, we dropped a number of variables that we were not interested in (zip codes, IDs, longitude, and latitude). We also converted a number of categorical columns from numeric into factors (or dummy variables). For example, one of the variables we converted was property condition. All of the properties had been given a grading of 1-5 with 1 being poor and 5 being very good. We converted this column into factors using 1 as the reference variable. After we finished all of the initial data cleaning, we took a random sample of 1000 rows (as per project instructions) and used this sample to construct our model.

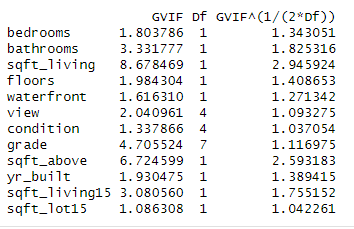
The first thing that we did was to use stepwise regression to select the variables in our model. Since there are a lot of variables in our dataset, we wanted to use an automated predictor selection method as an initial step. Stepwise regression works by either adding or removing predictors one at a time based on how much they improve the model, usually measured by AIC (Akaike Information Criterion). AIC balances model fit and model simplicity — it rewards models that fit the data well but also penalizes them for being too complex. In each step, the algorithm checks whether adding or removing a variable lowers the AIC; if it does, the change is made. This process helped us quickly narrow down to a set of variables that are most useful for predicting housing prices. In forward or backward operations, variables are added or removed one by one; and in stepwise (what we used), it does both — adding or removing predictors at each step depending on which action best improves the model fit. This helped us quickly narrow down to a set of variables that are most useful for predicting housing prices.

After the step function ran, our model ended up being:



As you can see, our adjusted R-squared value is 0.7107, indicating that approximately 71.07% of the variability in the house prices can be explained by the predictors included in the model, after adjusting for the number of predictors used. We also see that at an alpha = 0.05 level, the overall F-test indicates that the model as a whole is statistically significant, suggesting that at least one predictor provides meaningful explanatory power. When we look at the t-test values, we can see exactly which predictors were statistically significant.

However, after we test for multicollinearity using VIF, we get:

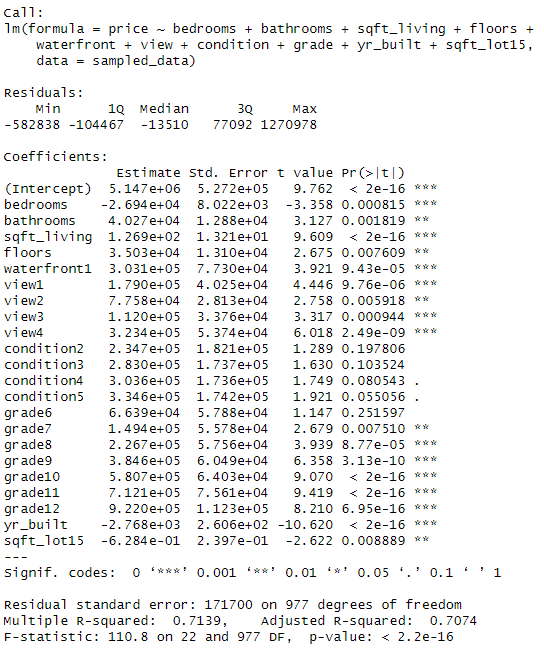


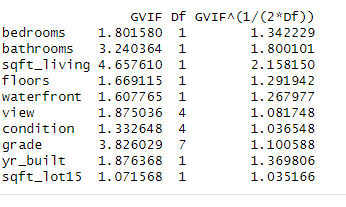
We see that *sqft\_living* and *sqft\_above* have high VIF values, and *sqft\_living15* has a moderately high VIF, indicating that there is high collinearity between them and some of the other predictor variables. As a result, we decided to drop *sqft\_above* and *sqft\_living15* since based on domain knowledge, *sqft\_above* is probably contained in *sqft\_living*, and *sqft\_living15* is probably correlated since neighborhoods tend to have houses of the same size, and *sqft\_living* might be part of the calculation to determine *sqft\_living15*.

As a result, our new model looks like:

model\_2 = lm(formula = price ~ bedrooms + bathrooms + sqft\_living + floors +

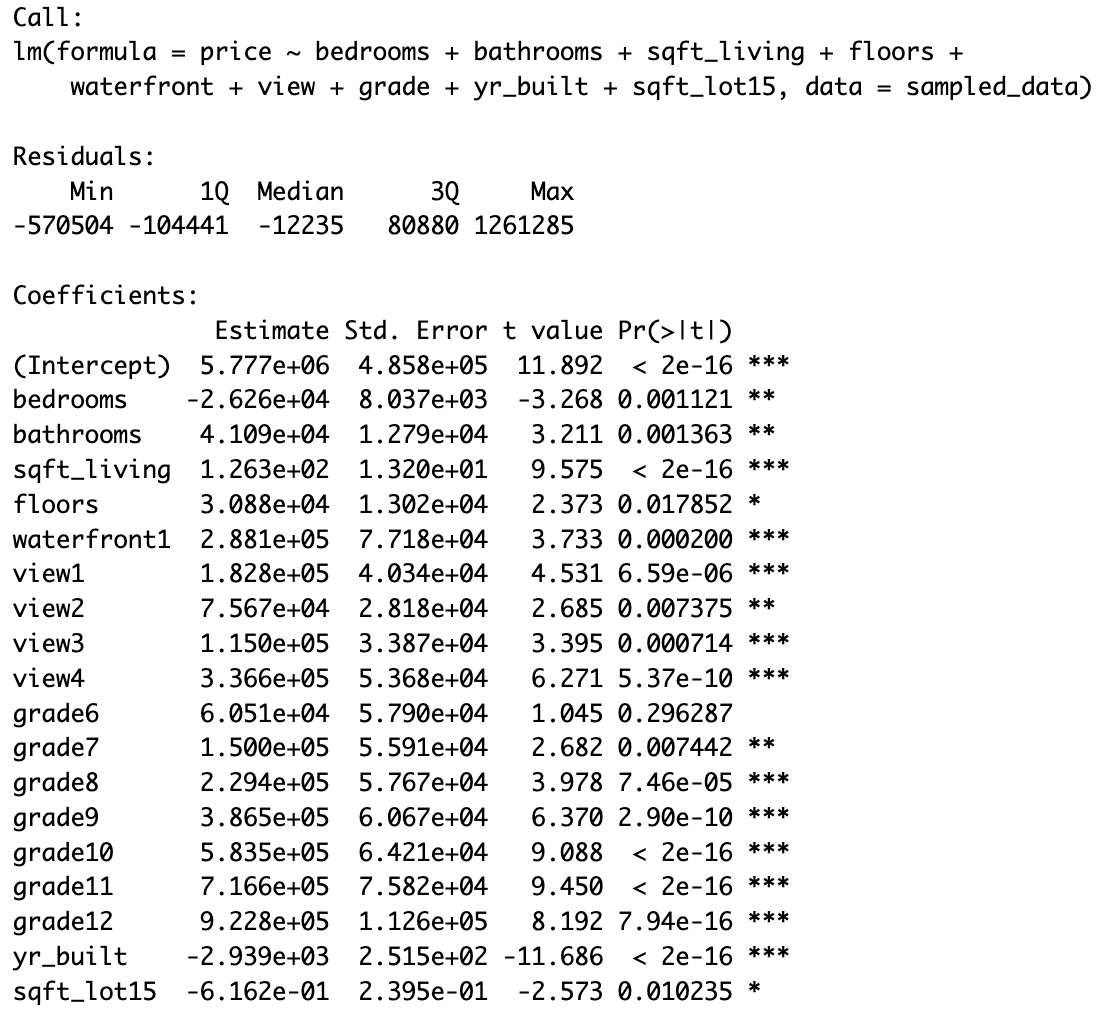
waterfront + view + condition + grade + yr\_built + sqft\_lot15, data = sampled\_data)

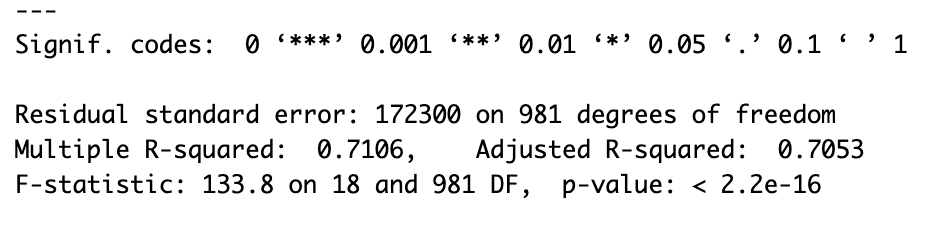


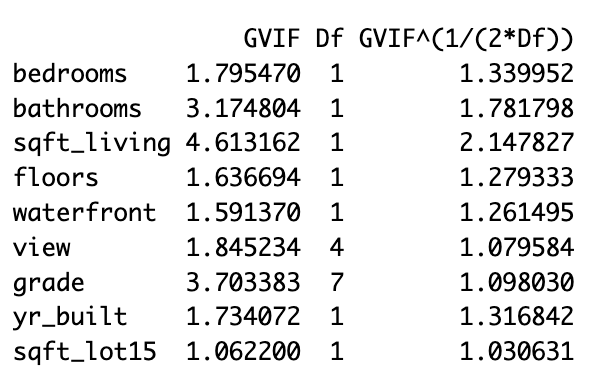


After rerunning the Variance Inflation Factor (VIF) tests on the linear regression model, all VIF values were found to be below 5, which is generally accepted as the threshold indicating the presence of moderate multicollinearity. However, *sqft\_living* and *grade* exhibit higher VIF values. This is likely attributable to moderate correlations with *bedrooms* and *bathrooms*, as an increase in the number of bedrooms and bathrooms typically corresponds with a larger living area (*sqft\_living*) and a higher overall property *grade*. Despite removing two predictors from the model, we still end up with an adjusted r-squared value of 0.7074. However, we see that at the alpha = 0.05 level, none of the *condition* variables are statistically significant. As a result, in our final model, we decided to drop them.

Final Model:

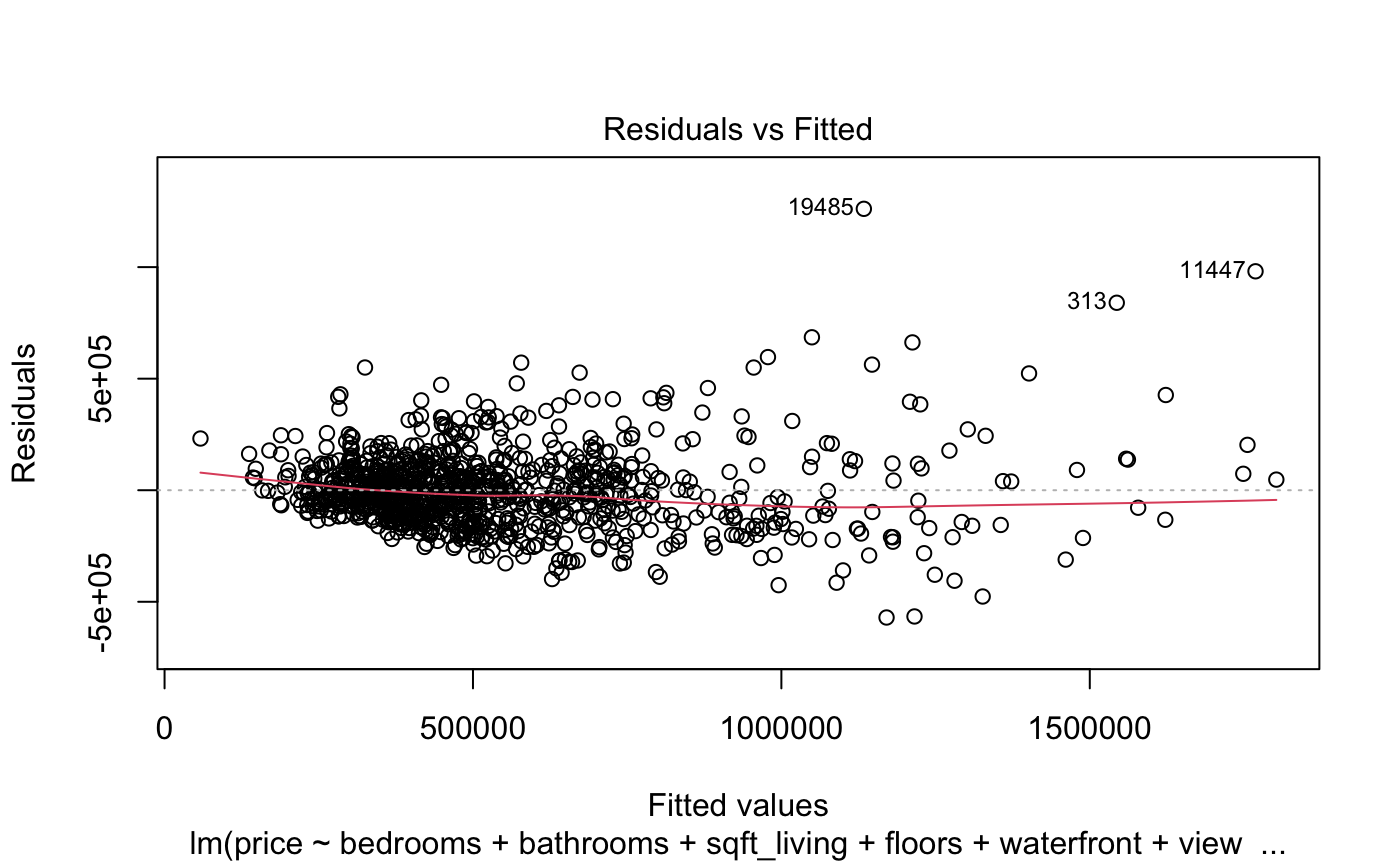




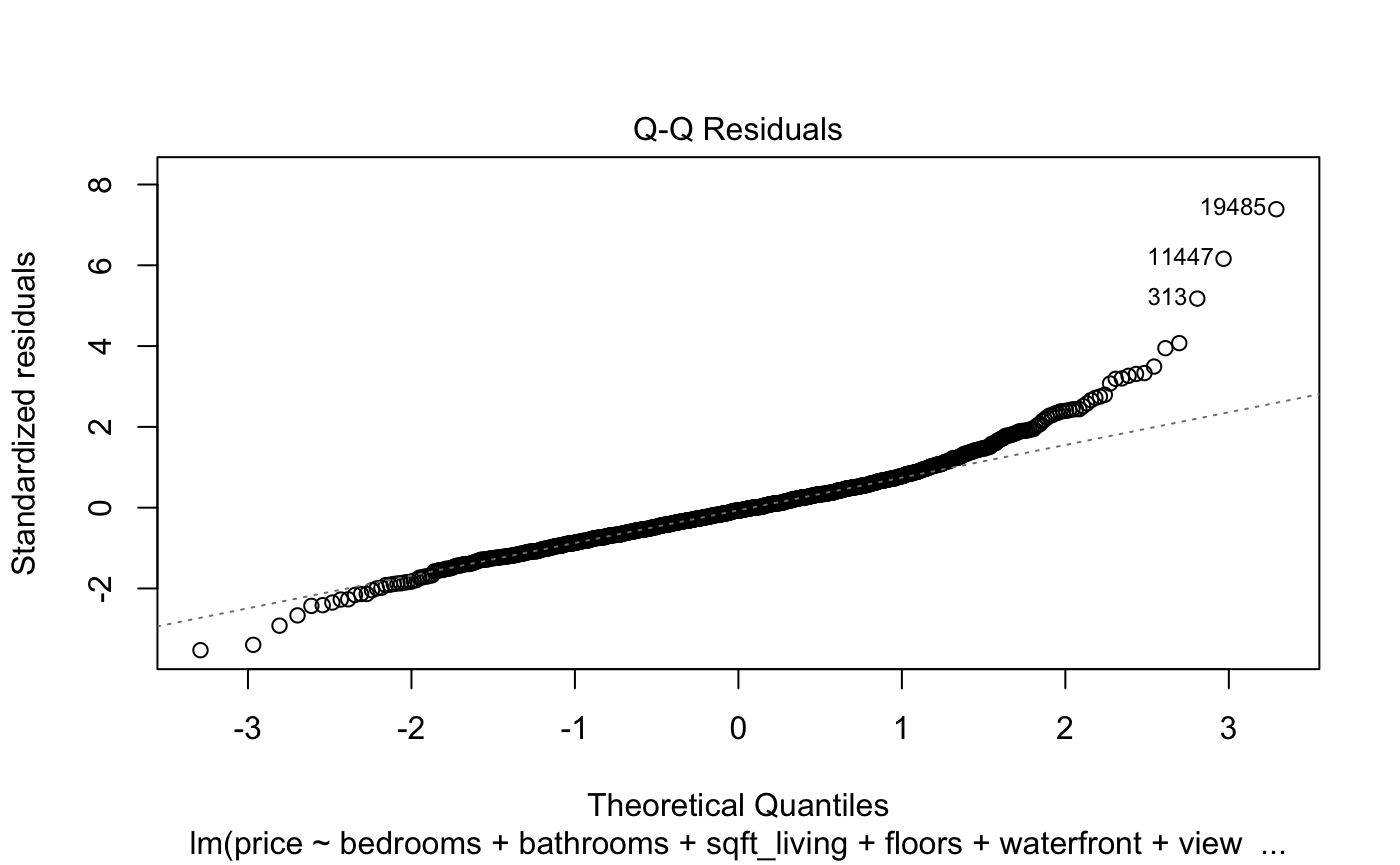


In the final model, our adjusted R-squared value is 0.7053, indicating that approximately 70.53% of the variability in the house prices can be explained by the predictors included in the final model, after adjusting for the number of predictors used. We see that at an alpha = 0.05 level, the overall F-test indicates that the model as a whole is statistically significant, suggesting that at least one predictor provides meaningful explanatory power. When we look at the t-test values, we can see exactly which predictors were statistically significant, and in the final model, almost all of the variables (excluding *grade6*) were statistically significant.

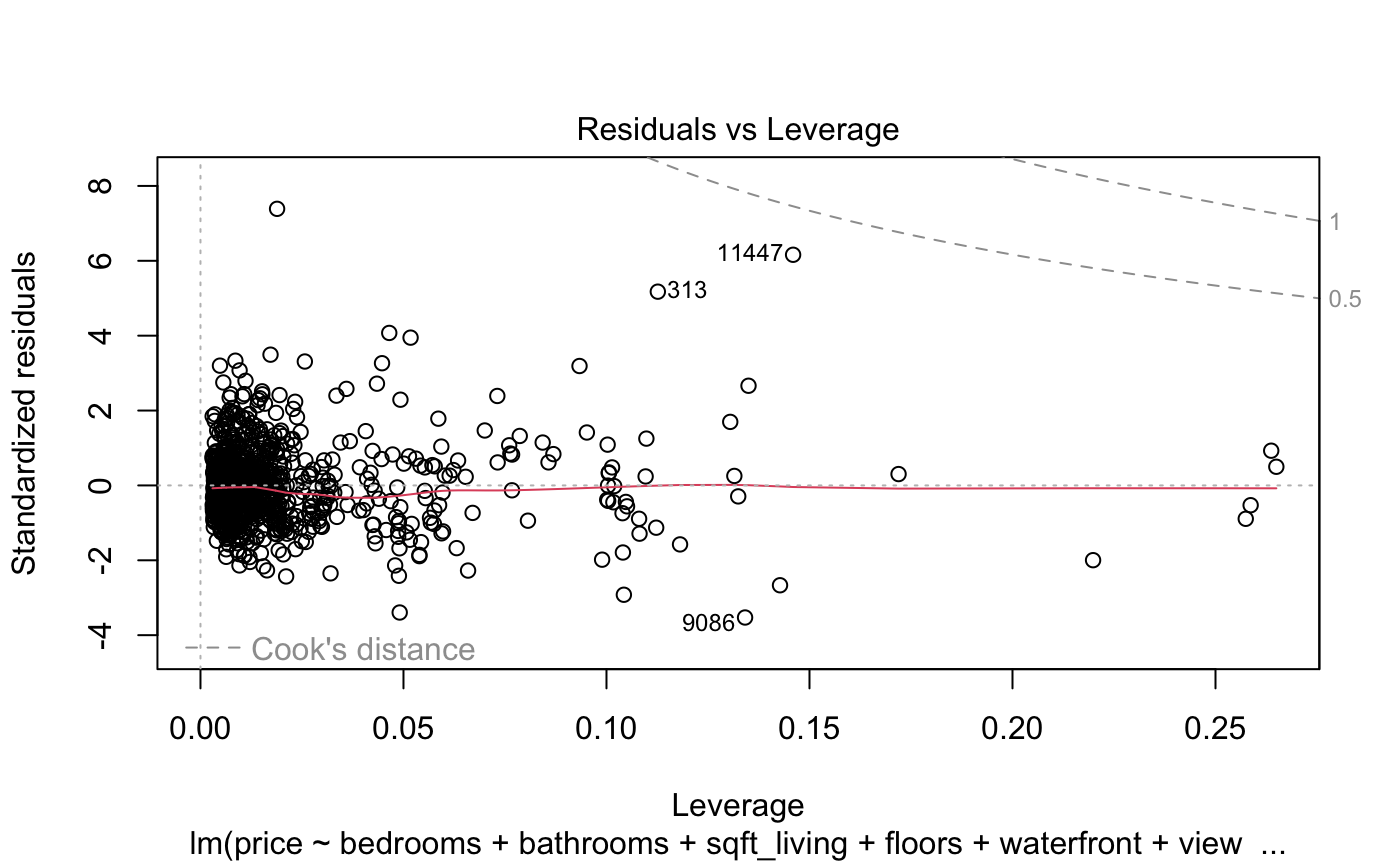
After creating our final linear regression model, we made residual plots to see if our linear regression meets the LINE assumptions of linearity, independence, normality, and equal variance.



If we look at the residual vs fitted plot, we see that the red line of fit appears to be mostly linear, indicating that a linear regression model is correct for this dataset, thus the condition of linearity is met. We also see that there doesn’t appear to be a severe fan shape with this plot, indicating that the condition of equal variance is also met.



The QQ-plot shows that most of the residuals lie along the diagonal line, suggesting that the residuals are approximately normally distributed. There is a slight upward curvature on the right end of the plot, which may indicate mild right skewness. However, since the majority of the points closely follow the diagonal, the normality assumption appears to be reasonably met.



Looking at the residuals vs. leverage plot, we observe a few potentially influential points (outliers). These points are within the Cook’s distance threshold, but are still worth nothing. Despite this, the model appears to be good as no points exceed the Cook’s distance cutoff for high influence, and there are only 3 points out of 1000.

Additionally, we can assume the independence assumption holds, as each data point represents a distinct house, and the selection of one house does not influence the selection of another.

## **5. State your conclusions and discuss any limitations.**

* Which variables are statistically significant in predicting housing prices?
* Does including additional variables improve or hurt our model?
* Is a linear regression model appropriate, and do we meet all of its assumptions?

State the conclusion so that a non-statistician can understand. Discuss any potential limitations of your analysis. For example, are you suspicious that the assumptions of your test may not hold? Do you feel the analysis may have limitations for any other reasons?

To answer our research question, we see that at the alpha = 0.05 level, *bedrooms, bathrooms, sqft\_living, floors, waterfront1, view* (all levels)*, grade7, grade8, grade9, grade10, grade11, grade12, yr\_built,* and *sqft\_lot15* were significant in predicting housing prices, while *grade6* was not statistically significant.

Additionally, we observed that including additional variables did not necessarily improve our model. We saw that by including all of the variables from stepwise selection, we started having multicollinearity issues, since some of the variables were extremely correlated to each other.

By examining the regression plots, we found that all of the **LINE** assumptions (Linearity, Independence, Normality, and Equal variance) were met. This means the linear regression model is appropriate for the data, and the results are reliable for making predictions and drawing conclusions.

We can conclude that we’ve built a solid model for predicting housing prices based on certain characteristics of the houses, such as square footage, number of bedrooms, and so on. We can predict about 70% of the variation in house prices, which is a fairly good prediction value.

However, it's important to note a potential limitation: the dataset we used only includes houses in the Seattle area. Housing prices and market dynamics can vary significantly across different regions, so this model may not be as accurate if applied to other locations. For example, housing prices in rural Alabama are likely to be much lower than in Seattle, and the factors that influence prices may differ as well. Additionally, variables such as how inspectors categorize a "grade 13" house could vary from state to state. Therefore, while this model might work well for Seattle, it may not directly apply to other regions without adjustments.