Multivariate Linear Regression – Predicting Housing Prices in New York

The dataset contains a random sample of 1057 houses taken from The Saratoga New York Housing Dataset. The data has the following features, **Price** being the target variable:

- Price – The price of the property (US Dollars).
- Living Area – Living area (Square feet)
- Bathrooms – Number of Bathrooms
- Bedrooms – Number of Bedrooms
- Fireplaces – Number of Fireplaces
- Lot Size – Lot size (Acres)
- Age – Age of the house (Years)
- Fireplace – Whether the house has a fireplace or not (Boolean)

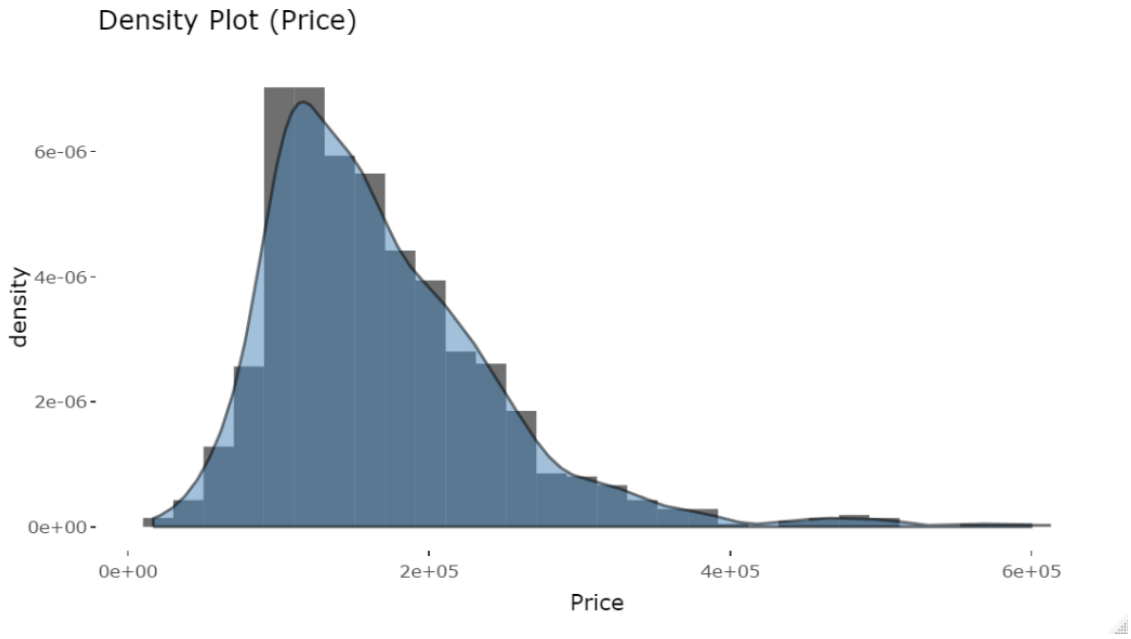
Price	Living.Area	Bathrooms	Bedrooms	Fireplaces	Lot.Size	Age	Fireplace
599701	5114	4.5	5	2	0.34	131	TRUE
578856	2472	2.5	3	1	0	6	TRUE
562546	5228	4	4	4	0.45	14	TRUE
509488	3530	3.5	4	1	2.05	0	TRUE
506149	2586	3	4	1	0.21	3	TRUE

Figure: First five rows of the dataset

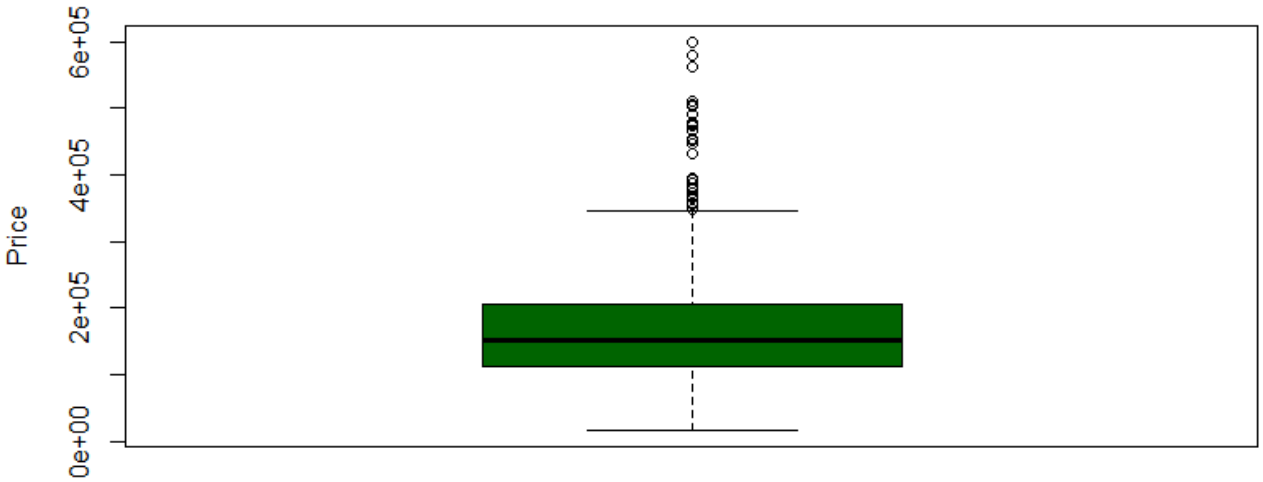
Exploration – Getting a feel for our data

Since we are going to predict the **Price ($USD)** column, let us start with it.

Min	1st Quartile	Median	Mean	3rd Quartile	Max
16858	112614	152258	167919	206512	599701



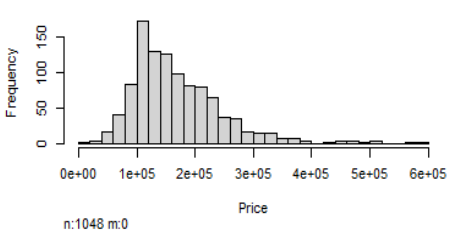
The distribution of Price is positively skewed, and we also observe outliers in the dataset. We generate a box plot to investigate further:

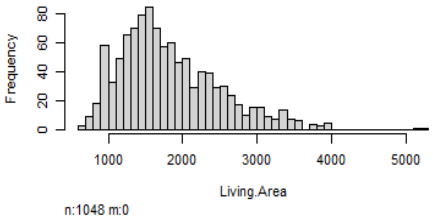


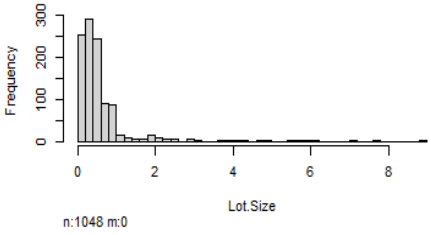
From the box plot, we observe that there are a considerable number of houses with **outlying** prices. However, as the dataset is sourced from a real-world setting, and we are building a model to simulate the real-estate market, the presence of outliers is expected. Although removing outliers

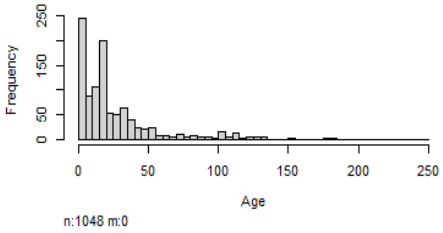
would simplify our model, we choose to keep them in order to simulate the real-world setting as closely as possible.

Next, we observe that there are two kinds of features in our dataset i.e. Continuous and Discrete. In order to explore these features before we move forward, we use suitable plotting techniques to get an idea of their individual distributions and take note of any anomalies.



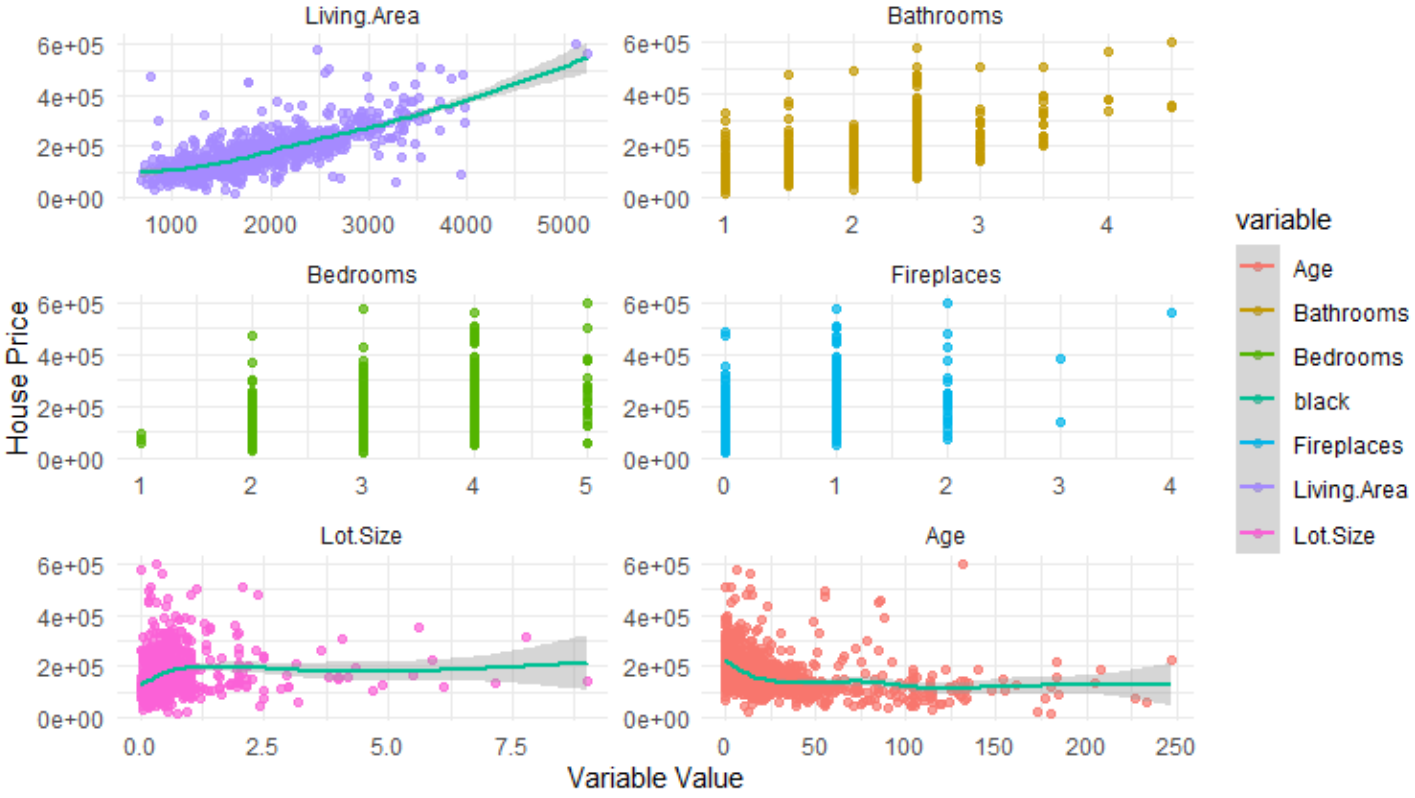






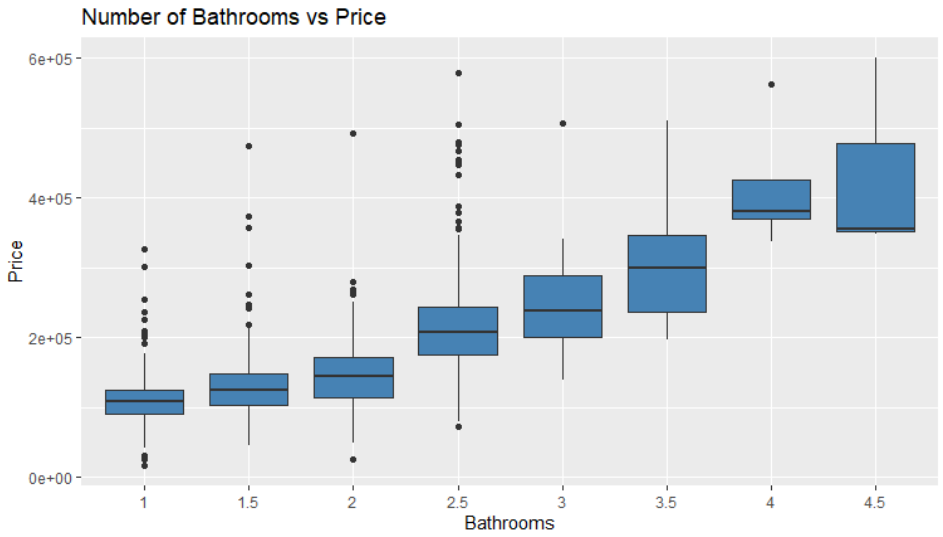
From the individual distributions, we observe that there are outliers present in the features as well. This further strengthens our earlier decision to not remove outlying prices as we now know that those prices are associated with extreme values of these features and were not made in error.

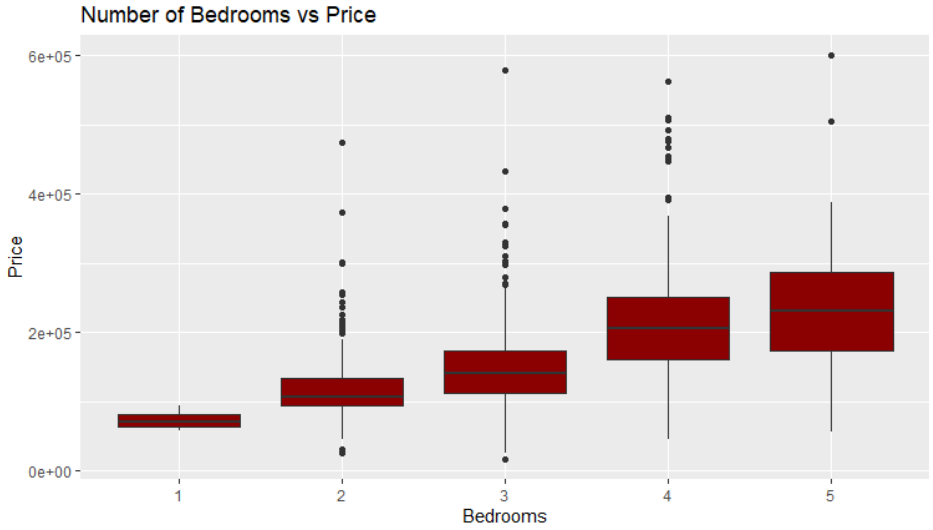
Next, in order to explore and find any relationships between the features and the Price we generate scatter plots for each feature against the Price individually.

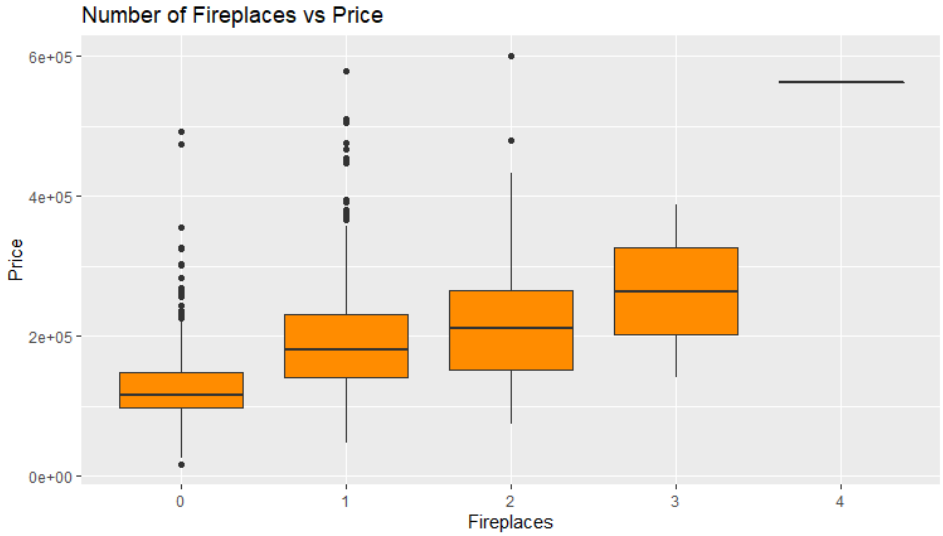


Right away, we can observe that Living Area shows a positive linear relationship with Price i.e. A larger living area generally means that the house goes for a higher price.

Lot Size and Age show little to no correlation to Price and need not be considered further. However, scatterplots did not prove suitable for the remaining features i.e. Bathrooms, Bedrooms, and Fireplaces. We will use box plots to better visualize their relationship with Price:


Number of Bathrooms vs Price
Price
6e+05
4e+05
2e+05
0e+00
1
1.5
2
2.5
3
3.5
4
4.5
Bathrooms


Number of Bedrooms vs Price
Price
6e+05
4e+05
2e+05
0e+00
1
2
3
4
5
Bedrooms


Number of Fireplaces vs Price
Price
6e+05
4e+05
2e+05
0e+00
0
1
2
3
4
Fireplaces

From the boxplots, we see that Price generally increases with the number of Bathrooms, the number of Bedrooms, and the number of Fireplaces. Each feature shows a considerable correlation to price. We keep these features under consideration for further analysis.

Next, to verify or refute our findings we need to quantify these relationships and determine which features have significant correlation to our target variable. For this purpose, we generate a correlation matrix for the dataset. Features that clear the threshold will be passed to our model and others will be discarded from consideration.

From the correlation matrix, we find that Living Area is highly correlated to our target variable and Number of Bathrooms has a moderate correlation. We take these features ahead to use them for our multi regression model.

Predicting housing prices

Preprocessing

Before we proceed to train our model, we need to normalize the features in order to change the values of the numeric variable to a common scale. For this dataset, we will use the Min-Max Scalar approach which scales the data between 0 and 1.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Training our Model

Next, the dataset is split into training and testing in the ratio 75:25. The training data is now ready for regression. Our final objective is to estimate the values of the weights in the following equation:

Price = β_0 + β_1 *(Living Area)* + β_2 *(Number of Bathrooms)* + ε

```
Call:
lm(formula = Price ~ Living.Area + Bathrooms, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-207430  -22582   -3586   19350  351460

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12709.355   5328.692  -2.385   0.0173 *
Living.Area     66.589      3.689  18.052  < 2e-16 ***
Bathrooms    30198.923   3788.875   7.970 5.58e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46670 on 783 degrees of freedom
Multiple R-squared:  0.6332,    Adjusted R-squared:  0.6323
F-statistic: 675.8 on 2 and 783 DF,  p-value: < 2.2e-16
```

The Fitted Equation is as follows:

Price = -12709.355 + 66.589 (Living Area) + 30198.923 (Number of Bathrooms)
Adjusted R-squared: 0.6323

The value found for β_0 is -12709.355. This is just the y-intercept, showing what the Price of a house would be if all the coefficients are equal to 0.

The value for β_1 is 66.589 and is interpreted as keeping the Number of Bathrooms constant, if the Living Area increase by 1 square foot then the Price will increase by 66.59 USD.

The value for β_2 is 30198.923 and is interpreted as keeping the Living Area constant, if the Number of Bathrooms increase by 1 then the Price will increase by 30198.923 USD.

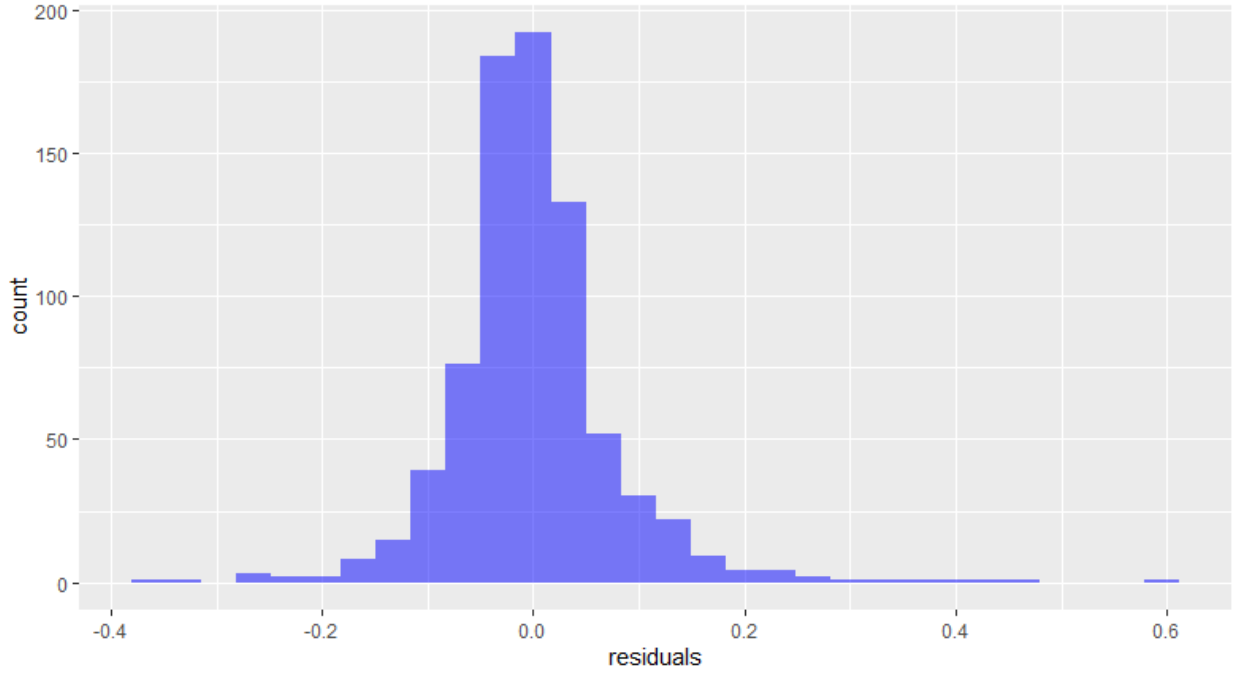
```
Anova Table (Type III tests)

Response: Price
               Sum Sq  Df  F value    Pr(>F)
(Intercept) 1.2389e+10   1   5.6886   0.01731 *
Living.Area 7.0968e+11   1 325.8567 < 2.2e-16 ***
Bathrooms   1.3836e+11   1  63.5276 5.584e-15 ***
Residuals   1.7053e+12 783
---
```

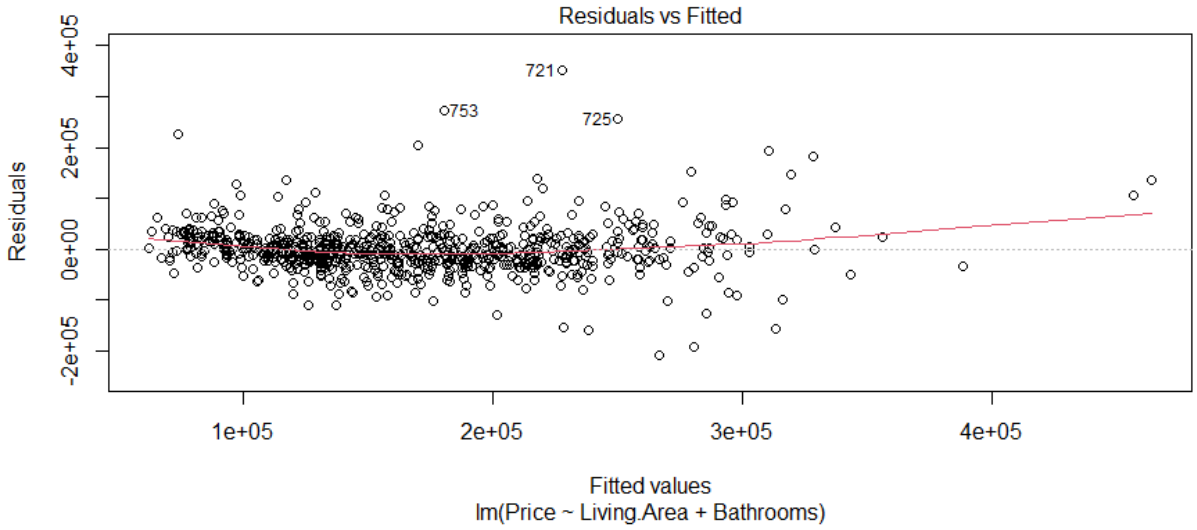
From the ANOVA Table, we note the value of Prob(F). This is the probability of null hypothesis being true (all regression coefficients are zero). As the probability is almost negligible, we can safely reject the Null Hypothesis.

Our regression equation therefore does have validity in fitting the data (i.e., the independent variables are not purely random with respect to the dependent variable).

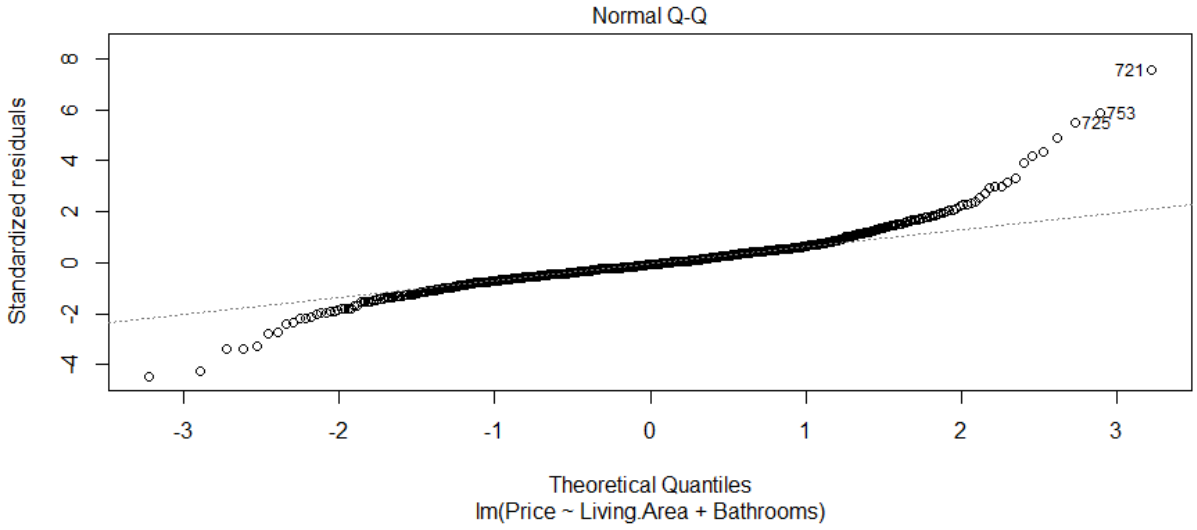
Visualizing our model



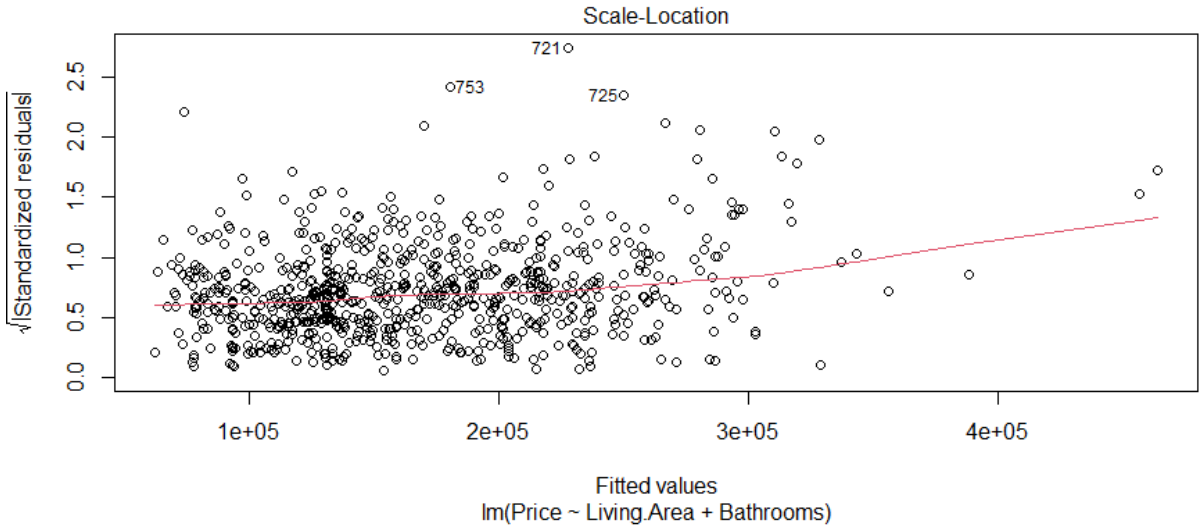
As expected, the residual errors are normally distributed. The errors are mostly clustered around zero meaning that our equation fits the data quite well.



Here, we plot the predicted Price on the x-axis, and the residual error on the y-axis. Here we see that linearity seems to hold reasonably well, as the red line is close to the dashed line. The residuals are generally clustered near the red line (except for a few outliers), but as we move towards the right, the spread of the residuals seems to be increasing.



The Q-Q (quantile-quantile) plot shows two sets of quantiles places against each other. It is a test of whether the dependent variables are normally distributed or not. Notice the points fall along a line in the middle of the graph but curve off in the extremities. This means that our data has more extreme values than would be expected if they truly came from a Normal distribution.



Here, we test the assumption of equal variance (**homoscedasticity**). The residuals are scattered randomly along the red line and do not show a pattern. The variance is equal.

Predictions

Let us test our model by predicting on our testing dataset.

Assessing our Model

The Root Mean Square Error (RMSE) for our Model is: $4163.897.