The initial full model presents a problem of multicollinearity between predictors. The predictor sqft\_living the the square footage of the entire house. The predictors sqft\_above and sqft\_basement are the square footage of the house above the ground and the square footage basement respectively. This means the three predictors are not independent of each other because sqft\_living = sqft\_above + sqft\_basement.

Several small linear models will be built to see which predictors are better for the full model. The first model is one that just has sqft\_living has a predictor of price. Building this model using R yields the following.

|  |  |
| --- | --- |
| Summary Output |  |
| Regression Statistics |  |
| Multiple R | 0.4929 |
| Adjusted R-squared | 0.4928 |
| Standard Error | 261500 |
| Observations | 21612 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | df | SS | MS | F | Significance F |
| sqft\_living | 1 | 1.44E+15 | 1.44E+15 | 21002 | 2.20E-16 |
| Residual | 21611 | 1.48E+15 | 6.84E+10 |  |  |
| Total | 21612 | 2.91E+15 |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | coeffic. | Std. Err. | t Stat | P-value |
| Intercept | -43580.743 | 4402.69 | -9.899 | 2.00E-16 |
| sqft\_living | 280.624 | 1.936 | 144.92 | 2.00E-16 |

The next small model built using R is one with sqft\_above and sqft\_below as a predictor of price. Doing this had the following results.

|  |  |
| --- | --- |
| Summary Output |  |
| Regression Statistics |  |
| Multiple R | 0.4933 |
| Adjusted R-squared | 0.4932 |
| Standard Error | 261400 |
| Observations | 21612 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | df | SS | MS | F | Significance F |
| sqft\_above | 1 | 1.07E+15 | 1.07E+15 | 15638.8 | 2.20E-16 |
| sqft\_basment | 1 | 3.69E+14 | 3.69E+14 | 5397.2 | 2.20E-16 |
| Residual | 21610 | 1.48E+15 | 6.83E+10 |  |  |
| Total | 21612 | 2.91E+15 |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | coeffic. | Std. Err. | t Stat | P-value |
| Intercept | -40859.961 | 4447.95 | -9.186 | 2.00E-16 |
| sqft\_above | 276.677 | 2.15 | 128.702 | 2.00E-16 |
| sqft\_basment | 295.504 | 4.022 | 73.466 | 2.00E-16 |

We can consider the extra sum of squares to see if introducing one of the predictors has an impact on the model. Doing this presents a different kind of problem. For example, SSR(sqft\_above|sqft\_living) is a number so large, it’s impossible to tell if there is an impact of not. Thus we should compare the two Adjusted R-squared values and choose the model with the greater value. It is easy to see that the model with sqft\_above and sqft\_basement has a larger value that the model with sqft\_living. 0.4933 > 0.4929. It is safe to drop sqft\_living from the model and use sqft\_above and sqft\_basement in its place.

Next, is the model building process. Using R to expedite the process, three different selection processes will be used. Those are Forward Selection, Backward Elimination and Stepwise Regression. All three selection processes yield the same Adjusted R-squared value of 0.6995

The model is

price = grade + yr\_built + bathrooms + waterfront + lat + sqft\_above + sqft\_basment + view + bedrooms + zipcode + long + condition + sqft\_living15 + yr\_renovated + sqft\_lot15 + sqft\_lot + floors

For checking purposes, these same process were repeated using sqft\_living instead of sqft\_above and sqft\_basement. The result of this was a model with an Adjusted R-squared value of 0.6988. Which means dropping sqft\_living was the right choice