**Model Selection**

After doing feature engineering for all the independent variables which would affect the pricing model of diamond, following steps are carried out to build a good regression model for diamond’s price.

Step 1 : A multiple linear regression model (Figure XX) is constructed based on the variables shown in Table XX. These variables are gathered from the feature engineering process. By choosing a significance level of 0.05, we can see that Cut and Polish variables are not significant at the chosen level.

|  |  |
| --- | --- |
| Variable | Condition |
| Carat | - |
| Colour2 | 1 if Colour is between D and I  0 if not |
| Clarity2 | 1 if Clarity is SI1, SI2 or SI3  0 if not |
| Clarity3 | 1 if Clarity is VS1, VS2, VVS1 or VVS2  0 if not |
| Cut3 | 1 if Cut is Good  0 if not |
| Cut4 | 1 if Cut is Very Good  0 if not |
| Cut5 | 1 if Cut is Excellent  0 if not |
| Cut6 | 1 if Cut is Ideal  0 if not |
| Certification2 | 1 if Certification is AGS or GIA  0 if not |
| Polish2 | 1 if Polish is Very Good  0 if not |
| Polish3 | 1 if Polish is Excellent or Ideal  0 if not |
| Symmetry2 | 1 if Symmetry is Good  0 if not |
| Symmetry3 | 1 if Symmetry is Very Good, Excellent or Ideal  0 if not |

Table XX: Regrouped variables based on feature engineering

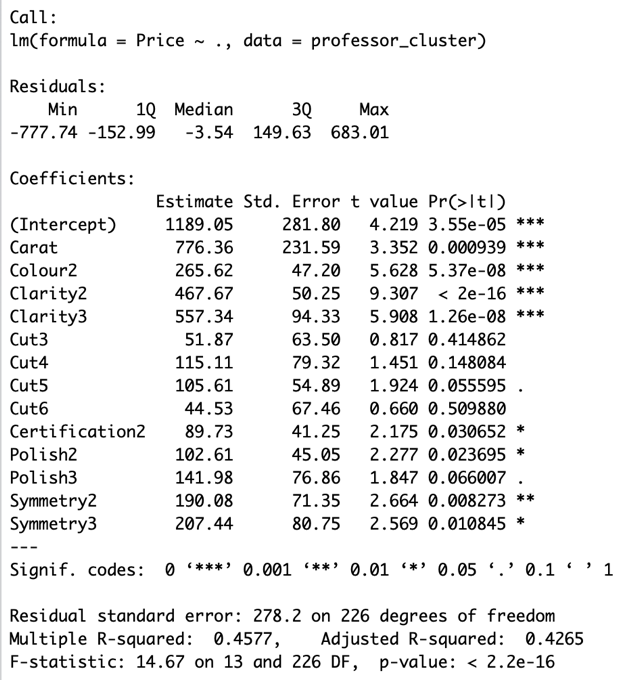


Figure XX: Model constructed based on Table XX’s variables

Step 2 : As noticed in previous step, the P value of cut is largest, so we regroup Cut into 2 groups: Fair and Good in group 1, Very Good, Excellent and Ideal in group 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Fair | Good | Very Good | Excellent | Ideal |
| 23% | 14% | 11% | 32% | 19% |

Table XX: Cut Proportions before regrouping

A regression model is built based on the new groups of Cut and others remain in the same proportions. Adjusted multiple correlation coefficient (Adjusted R2) reflects both the number of independent variables and the sample size. It may change when an independent variable is added or dropped, thus providing an indication of the value of adding and removing independent variables in the model. From this scenario, we decreased the number of variables by regrouping Cut variable into 2 groups and noticed that the adjusted R2 is slightly increased from 42.65% to 42.88%. From the model in Figure XX, the Cut variable is still not significant, however, we will continue reconstruct our model by examining Polish variable because it is having the largest p-value in the existing model and exceeds the chosen alpha level of 0.05.

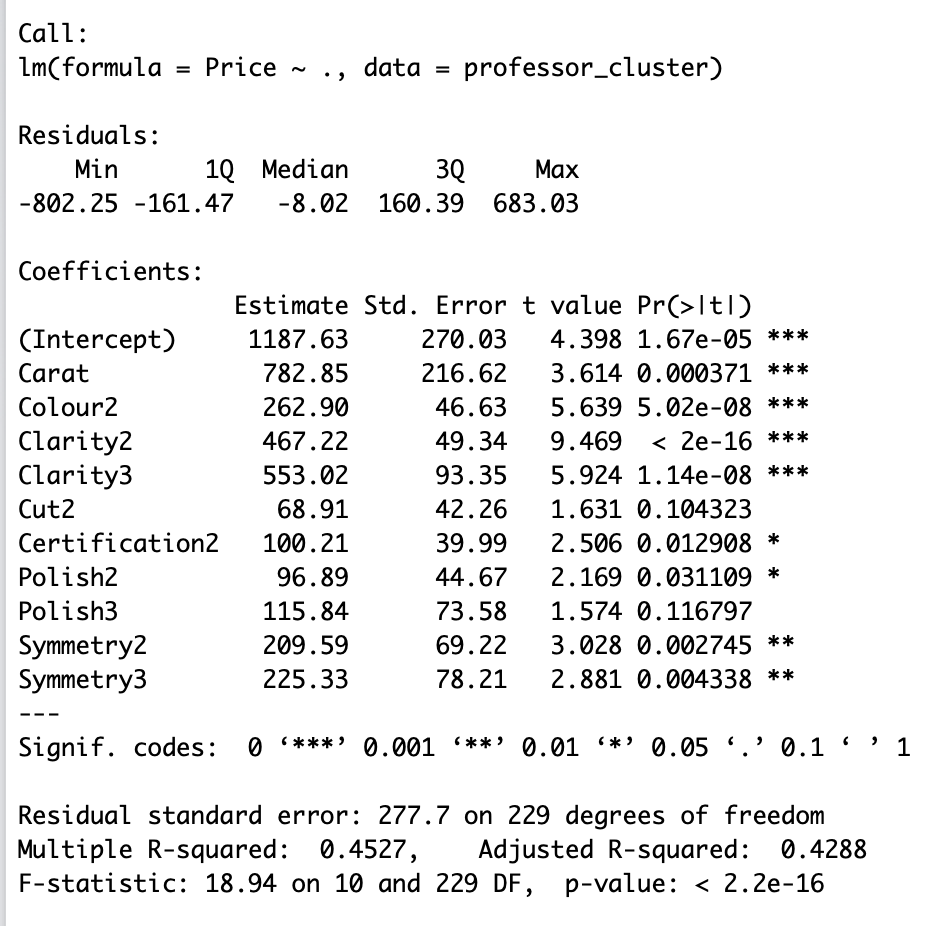


Figure 2: Model constructed after regrouping Cut variables into 2 groups

Step 3 : Polish is regrouped into 2 groups: group 1 is Fair and Good, and group 2 is Very Good, Excellent and Ideal. Below is the table provided for Polish variable that is regrouped in feature engineering. After a new regroup of Polish is done, the regression model is re- built again and adjusted R2 is examined to see if the model has improved. From Figure XX, polish variable become significant to the model after regrouping and the adjusted R2 is increased from 42.88% 43.11% which indicates that the model has improved by removing the number of variables. Nevertheless, Cut variable is still not significant to the model. A further action is needed to do for building a significant model for diamond’s price.

|  |  |  |
| --- | --- | --- |
| Fair & Good | Very Good | Excellent & Ideal |
| 49% | 40% | 11% |

Table X: Polish proportions after doing feature engineering

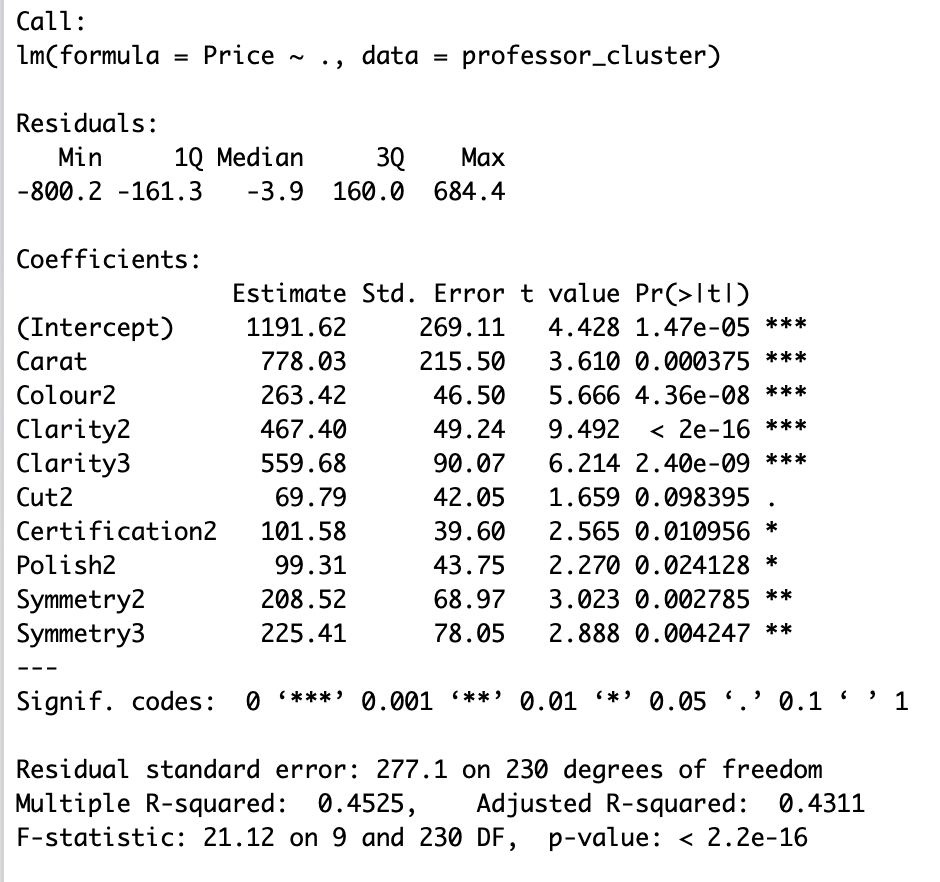


Figure 2: Model constructed after regrouping Polish variables into 2 groups

Step 3 : This situation can potentially tell us that there might be a multicollinearity in our model which means that 2 or more independent variables contain same information and are correlated with one another and can predict each other better than the dependent variable. Going back to the case, cut represents to both the shape and the proportions of the diamond. The performance of cut in a diamond is determined by its light reflective properties and same goes for symmetry, a diamond having a good symmetrical facet is depend on the light reflectivity of the diamond. Since cut is one of the main characteristics of determining the diamond pricing, we will drop Symmetry variable from the model and perform a new model again. A new model result is shown in Figure X. All variables are finally significant in this model but still the adjusted R2 decreased from 43.11% to 41.23%. This indicates us that the strength of association between the dependent and independent variables is decreased. A further step has to be done to perform a better result of the regression model.

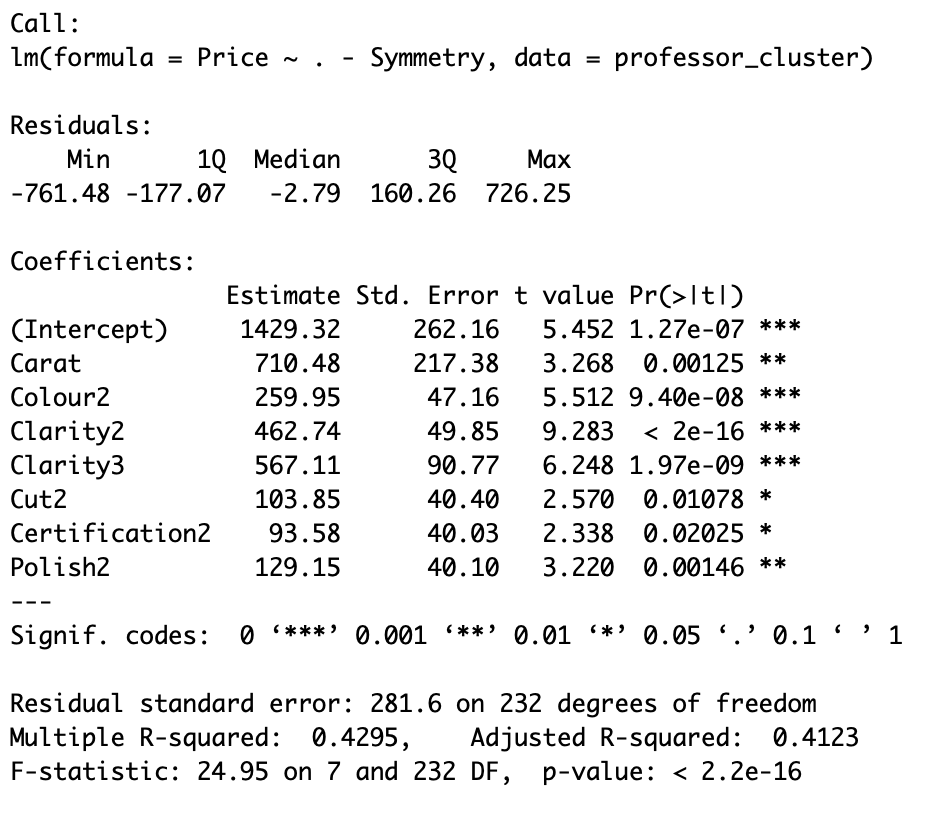


Figure X: Model constructed after removing Symmetry variables

Step 4 : To improve the performance of the previous model, we divided Color variable into 4 groups instead of 2 groups because Color is one of the significant characteristics that determine the value of a diamond, as a result, a fewer group of this variable may lead to biased results to the coefficients of the model. Table X shows the new grouping category for diamond. A model is then constructed, and having all variables are significant and most importantly, the adjusted R2 is increased from 41.23% to 46.35% which shows stronger association between dependent variables and the independent variables. This model will be our final model for the diamond pricing as it is a significant model with highest adjusted R2.

|  |  |  |  |
| --- | --- | --- | --- |
| D-F | G-I | J-K | L-N |
| 27% | 40% | 28% | 5% |

Table X: Colour Proportions

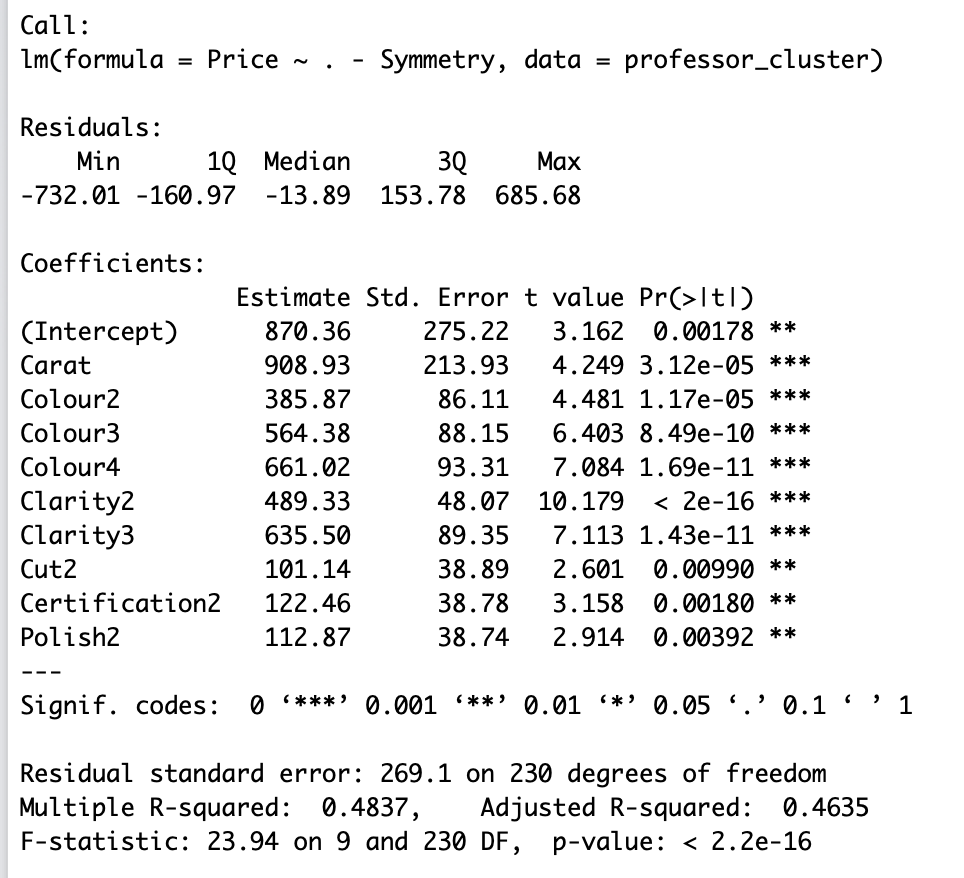


Figure 2: Model constructed after ungrouping Colour variables

**Model Summary**

The final model for the diamond pricing is:

where

Colour2 = 1 if it is J-K and 0 if not

Colour3 = 1 if it is G-I and 0 if not

Colour4 = 1 if it is D-F and 0 if not

Clarity2 = 1 if it is SI1, SI2, SI3 and 0 if not

Clarity3 = 1 if it is VS1, VS2, VVS1, VVS2 and 0 if not

Cut2 = if it is Very Good, Excellent, Ideal and 0 if not

Certification2 = 1 if it is AGS, GIA and 0 if not

Polish2 = 1 if it is Very Good, Excellent, Ideal and 0 if not

**Coefficients Interpretation**

1. Intercept : The regression intercept (y-intercept) is 870.36, which means when all independent variables are equal to 0, the base value of the diamond would be $870.36 which is still greater than 0.
2. Carat : The coefficient for Carat is 908.93, that is increase in one unit on Carat will result $908.93 increase in the value of the diamond.
3. Colour2 : The coefficient for Colour2 is 385.87, that is if the Colour of diamond in range from J to K, it will increase the value of diamond by $385.87.
4. Colour3 : The coefficient for Colour3 is 564.38 that is if the Colour of diamond in range from G to I, it will increase the value of diamond by $564.38.
5. Colour4 : The coefficient for Colour4 is 661.02 that is if the Colour of diamond in range from D to F, it will increase the value of diamond by $661.02.
6. Clarity2 : The coefficient for Colour4 is 489.33 that is if the Clarity of diamond is SI1, SI2 or SI3, it will increase the value of diamond by $489.33.
7. Clarity3 : The coefficient for Clarity3 is 635.50 that is if the Clarity of diamond is VS1, VS2, VVS1 or VVS2, it will increase the value of diamond by $634.50.
8. Cut2 : The coefficient for Cut2 is 101.14 that is if the Cut of diamond is Very Good, Excellent or Ideal, it will increase the value of diamond by $101.14.
9. Certification2 : The coefficient for Certification2 is 122.46 that is if the Certification of the diamond is from AGS or GIA, the price of diamond would be increased by $122.46.
10. Polish2 : The coefficient for Polish2 is 112.87 that is if the Polish of the diamond is Very Good, Excellent or Ideal, the price of diamond would be increased by $112.87.

**Disadvantage of the Model**

We have a set of large numbers of independent variables, and when we want to build a multiple linear regression model from this set of variables, there are potential number of possible models resulted. It is overwhelming and difficult to remove the insignificant variables effectively and develop the best regression model from the set of significant variables. As a result, our model might not be the best model that is developed by using the systematic approach.

**Conclusion**

The diamond that the professor was looking for has following requirements:

* Carat Weight : 0.9
* Cut : Very Good
* Color : J (Slightly Yellow)
* Clarity : SI2 (Slightly included: very few inclusions at 10x)
* Polish : Good
* Symmetry : Very Good
* Certification : GIA

The professor was quoted $3,100 for the diamond ring but when the following regression model is used then,

Price Calculated based on Model: $2,787.20

Therefore, the final value comes out to be $2,787.20 and the difference between the quoted price and the price calculated based on the model is $312.80.