Introduction

STATISTICS & IT’S APPLICATION IN BUSINESS 5520

FINAL PROJECT

PROFESSOR PROPOSES



Submitted by:

Caner Adil Irfanoglu – A00425840

Tom Tong – A00000000

Diven Sambhwani – A00425915

Madeleine Leong – A00430926

Sreeraj Punnoli – A00000000

Submitted to:

Dr. Michael Zhang

Submitted on:

Dec 9, 2018

Table of Contents

[Executive Summary 3](#_Toc532129555)

[Introduction 4](#_Toc532129556)

[Problem Statement 5](#_Toc532129557)

[Descriptive Analysis 6](#_Toc532129558)

[Data Types of the Variables 6](#_Toc532129559)

[Distribution of Independent Variables 7](#_Toc532129560)

[Density Distribution 8](#_Toc532129561)

[Density Distribution of Price 8](#_Toc532129562)

[Density Distribution of Carat 9](#_Toc532129563)

[Factorial Anova & Multicollinearity 10](#_Toc532129564)

[Feature Engineering 12](#_Toc532129565)

[Methodology 12](#_Toc532129566)

[Data Selection for Regression 13](#_Toc532129567)

[Regrouping the Independent Variables 15](#_Toc532129568)

[Model Selection 17](#_Toc532129569)

[Model Summary 22](#_Toc532129570)

[Coefficients Interpretation 23](#_Toc532129571)

[Disadvantages of the Model 23](#_Toc532129572)

[Conclusion 24](#_Toc532129573)

[APPENDIX 25](#_Toc532129574)

# Executive Summary

The purpose of the project is to determine whether the quote given to the professor for his engagement ring is fair or not. To answer this question first descriptive analysis of the variables are made. Then, the interrelation of decision variables is examined. Based on the findings in the descriptive analyses, feature engineering is applied to optimize individual regression performance of the variables. In conclusion, final multivariate regression model is selected after 5 steps and a suggestion made based on the findings.

# Introduction

Some Filler

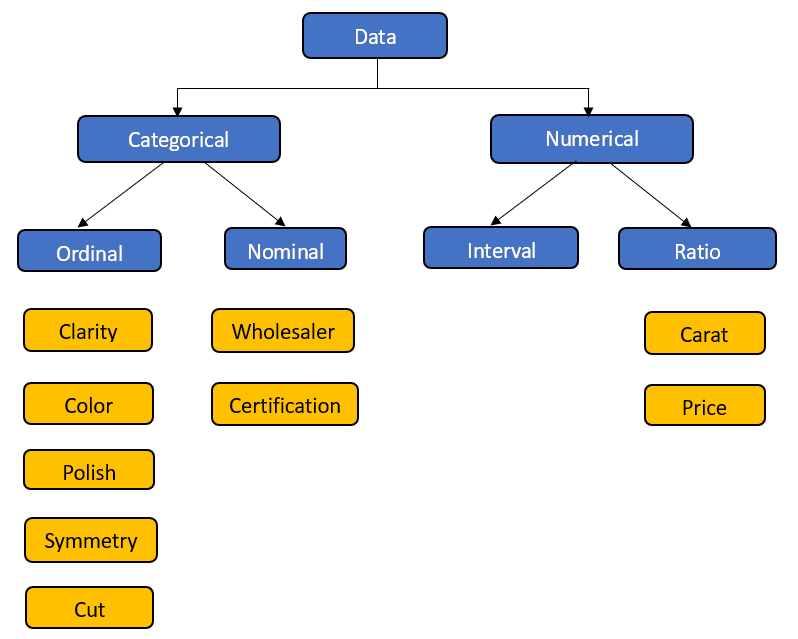
# Problem Statement

Some more filler

# **Descriptive Analysis**

## Data Types of the Variables

There are 9 variables given in the case study. Seven of them are categorical and two are numerical variables. Within the categorical variables 5 of them are ordinal variables. Also, numerical variables carat and price are ratio variables within the numerical group. To gain an intuitive understanding of the hierarchy of these levels, each ordinal variable will be renamed from worst the best by following an increasing integer scale. Details of the method will be covered thoroughly in the feature engineering section of the report.

 Figure 1: Data Types of Variables in the Dataset

## Distribution of Independent Variables

Dataset has 414 records with no missing values. The categories with low frequency on each independent variable are taken into consideration for feature engineering.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Table 1: Frequency Distribution of Categorical Variables

## Density Distribution

Density distributions are helpful for gaining familiarity with the frequencies of values within a variable. To get more understanding on the Price, density distribution of price and carat, the highest significant variable in the dataset, is taken into consideration. Density distribution also allows us to understand the mean of the distribution and the required specification of the diamond.

### Density Distribution of Price

Below shows the density distribution of price. Blue line indicates the mean price and the red line shows the quote Professor offered.

****

Figure 2: Density Distribution of Diamond Prices

### Density Distribution of Carat

Density distribution of Carat is shown below. It is evident that the mean (blue line) is lower than the Carat specification of the Professor.

****

Figure 3: Density Distribution of Diamond Carats

# **Factorial Anova & Multicollinearity**

We started by checking all the independent variables whether the levels are significant for determining the price. As all the variables p-value is less than 0.05, we concluded that all the variables are significant enough to be used in the model.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable v/s Price | F-Value | Critical Value | P-Value |
| Cut | 17.94 | 2.41 | 0.000 |
| Color | 5.32 | 3.88 | 0.021 |
| Symmetry | 17.89 | 3.03 | 0.000 |
| Polish | 18.14 | 3.03 | 0.000 |
| Clarity | 42.64 | 3.03 | 0.000 |
| Certification | 4.08 | 3.88 | 0.044 |

Table 2: p-values & F-values for independent variables against critical values

Firstly, categorical variables are converted into type numeric for checking the correlation. Then the collinearity plot is created. Since, this practice is not truly reliable for categorical variables, we decide to consider it as an indicator and do not totally rely on it. In the correlation plot the values are ranged from 1 to -1. Blue color states positive correlation and red color stated negative correlation

Bigger the circle more the correlation between those variables

*Carat and Clarity*: Clarity of diamond will decrease as per the increase of carat in diamond as both the variables are inversely proportional and it shows negative correlation.

*Polish and Symmetry*: The more polished the diamond is, the more symmetrical it becomes. Same applies with Cut and Symmetry. This explains positive correlation between these variables.



Figure 4: Multicollinearity of Independent Variables

# **Feature Engineering**

## Methodology

Integer values are assigned to levels for independent variables. Levels are relabeled for each variable from worst to best as follows:

|  |  |
| --- | --- |
| Variable Name | Mapping |
| Clarity | "SI2" = 5, "SI1" = 6, "SI3" = 4, "VS2" = 7,  "VS1" = 8, "I1" = 3, "I2" = 2 |
| Color | "L"=1, "J"=2, "K" = 2, "G" = 3, "H" = 3, "I" = 3, "F" = 4, "D" = 4, "E" = 4 |
| Cut | "P"=1, "F"=2, "G" = 3, "V" = 4, "X" = 5, "I" = 6 |
| Symmetry | "P"=1, "F"=2, "G" = 3, "V" = 4, "X" = 5, "I" = 6 |
| Polish | "P"=1, "F"=2, "G" = 3, "V" = 4, "X" = 5, "I" = 6 |
| Certification | "AGS"=2, "DOW"=1, "EGL" = 1, "GIA" = 2, "IGI" = 1 |

Table 3: Number Codes for Renaming Independent Variables

## Data Selection for Regression

After examining the bi-variates, we clearly see that the dataset is divided between two clusters. For each independent variable, there exist two price clusters. If the regression model is created based on overall data, the performance of the model will be questionable. The reason for the difference is that, the price versus carats are not linear for the whole carat scale. Also, diamond characteristics will change based on these clusters. Since, the professors diamond belongs to the blue cluster below, regression will be modelled based on the diamonds corresponding to this cluster. This is achieved by filtering out the diamonds having lower than 0.5 carats.

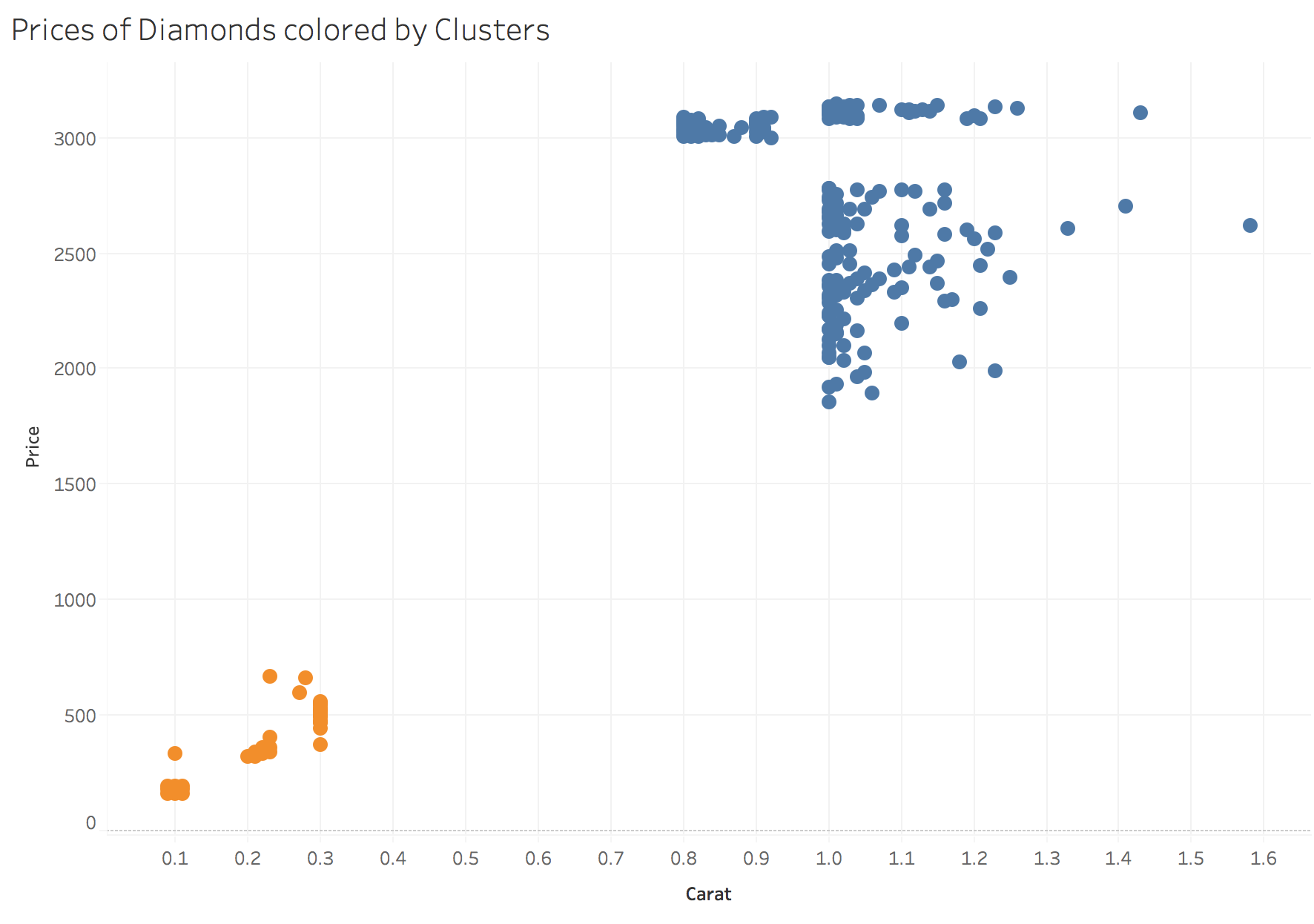


Figure 5: Carat vs. Price colored by cluster

Most of the data for blue cluster lies between 0.8 - 1.3 carats, which is a relatively small interval. Blue cluster is more favorable under linearity assumption as it is indicated in the problem statement.

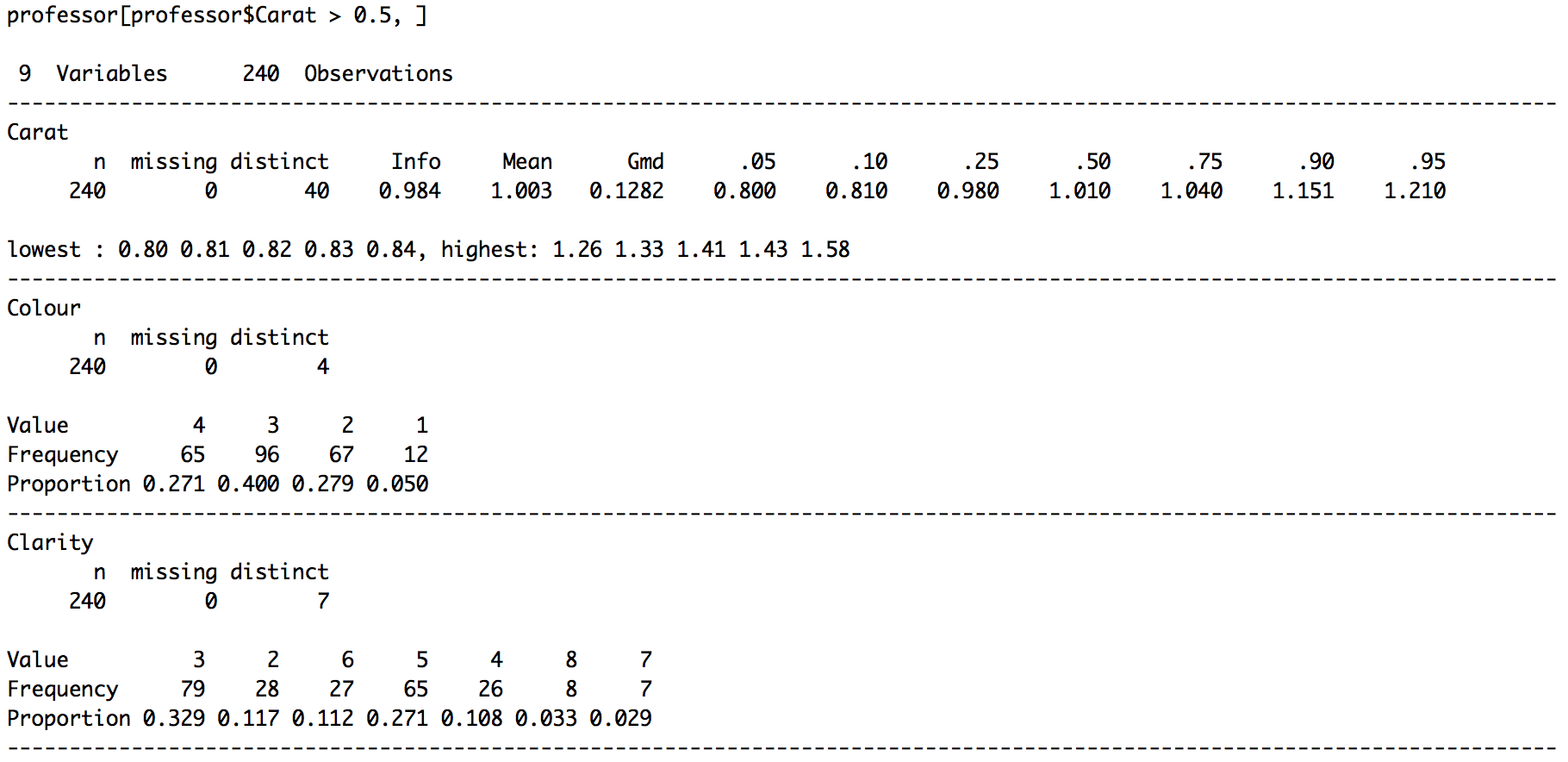
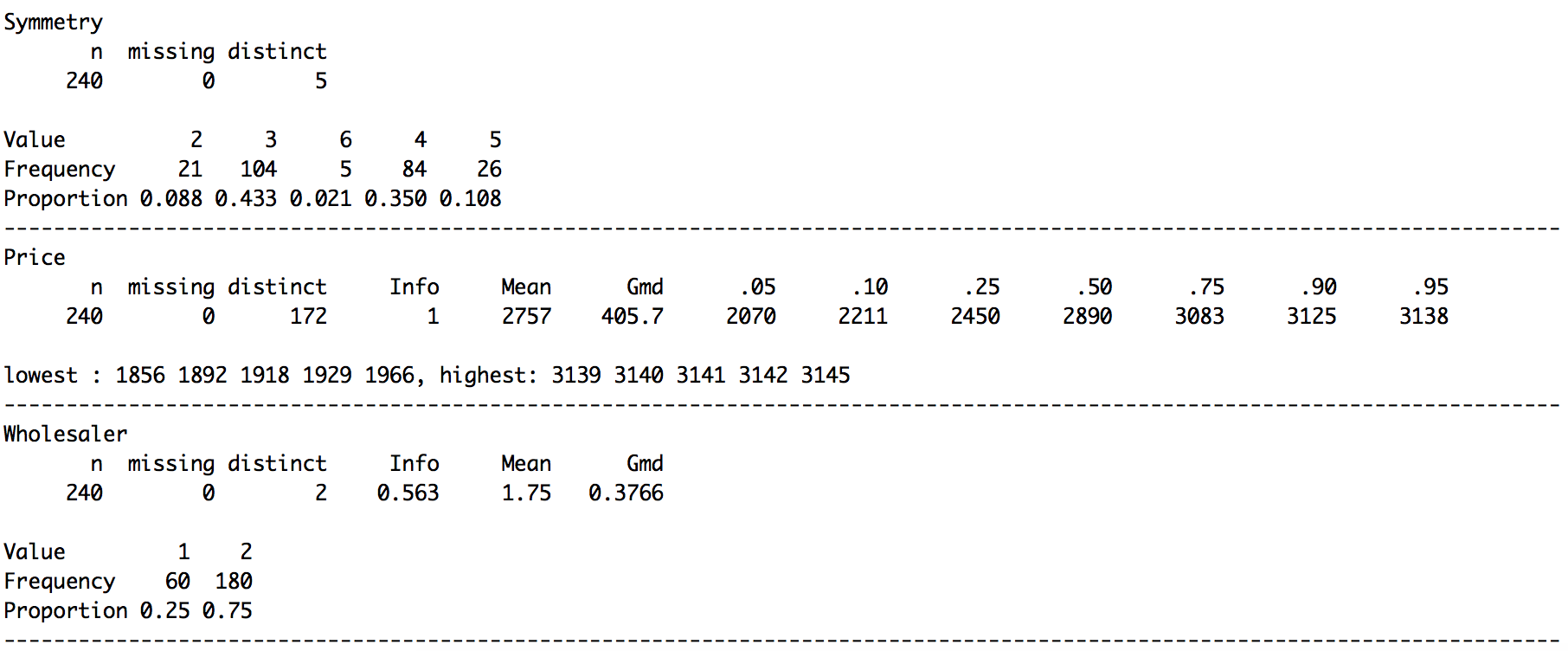
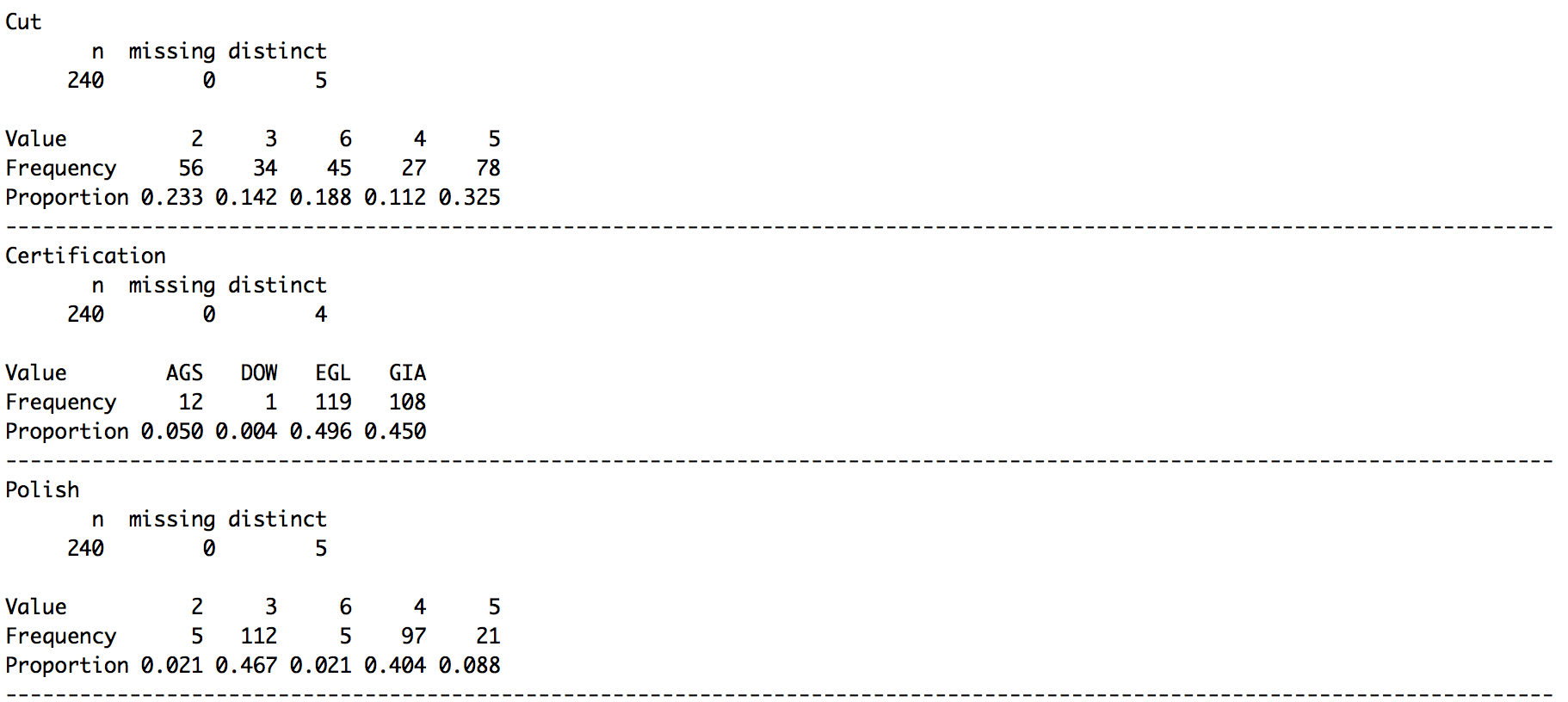
Descriptive statistics for the selected cluster are as follows;****

Figure 6: Descriptive statistics for regression data

## Regrouping the Independent Variables

In this section, all predictors will be grouped mainly based on their significance in determination of the price. The second grouping criteria is the bin size for a given level. If a level hold less than 5% of the total observations, it will be merged with it’s closest neighbor having the same statistical properties. When those criteria are satisfied the maximum number of possible groups will be used for having a higher R-Squared, meaning higher contribution to the overall model. Regression summaries for each variable before and after grouping can be found in the appendix section of the report.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable Name | Original Levels | Final Levels | Levelling Criteria(s) | Original  R-squared | Final R-squared |
| Clarity | * I2 * I1 * SI3 * SI2 * SI1 * VS2 * VS1 | - Flawed Naked Eye  - 10x Zoom Flaws  - 30x Zoom Flaws | Some Levels Insignificant for price | 0.403 | 0.265 |
| Color | * L * J,K * G,H,I * F,D,E | - Near Colorless  - Lightly Yellow | Some Levels Insignificant for price | 0.065 | 0.021 |
| Polish | * F * G * V * X * I | - F + G  - V  - X + I | Small sample size for F and I | 0.149 | 0.133 |
| Symmetry | * F * G * V * X * I | - F  - G  - V + X + I | - Small sample size for I  - Low predictive ability difference between V-X | 0.141 | 0.133 |
| Cut | * F * G * V * X * I | - F  - G  - V  - X  - I | - All levels distinct  - Bin sizes large enough | 0.144 | 0.144 |
| Certification | * AGS * GIA * EGL * DOW * IGI | - AGS + GIA  - EGL + DOW + IGI | Two most respected labs vs. others | 0.082 | 0.054 |

Table 4: Original and After Feature Engineering Levels

Since, wholesaler is not a diamond characteristic, it is excluded from the model and not shown in the table above.

# **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 7) is constructed based on the variables shown in Table 5. 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 5: Regrouped variables based on feature engineering

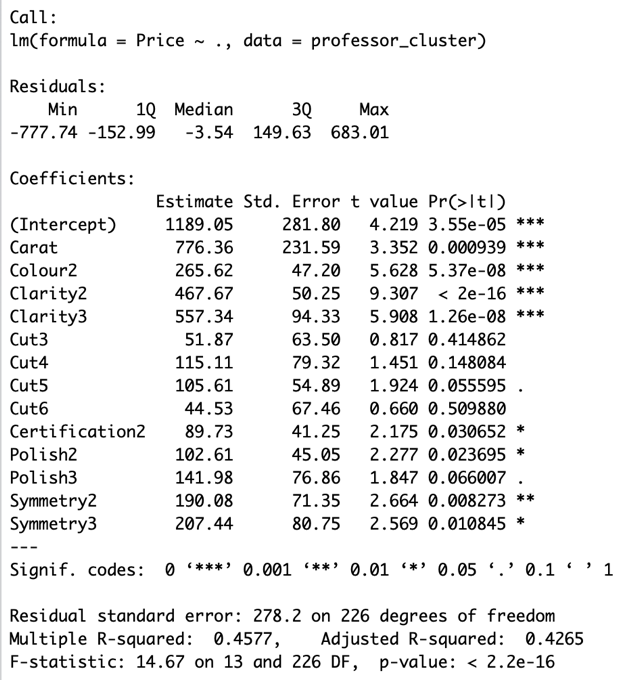


Figure 7: Model constructed based on Table 5’s variables

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 6: 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 7, 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.

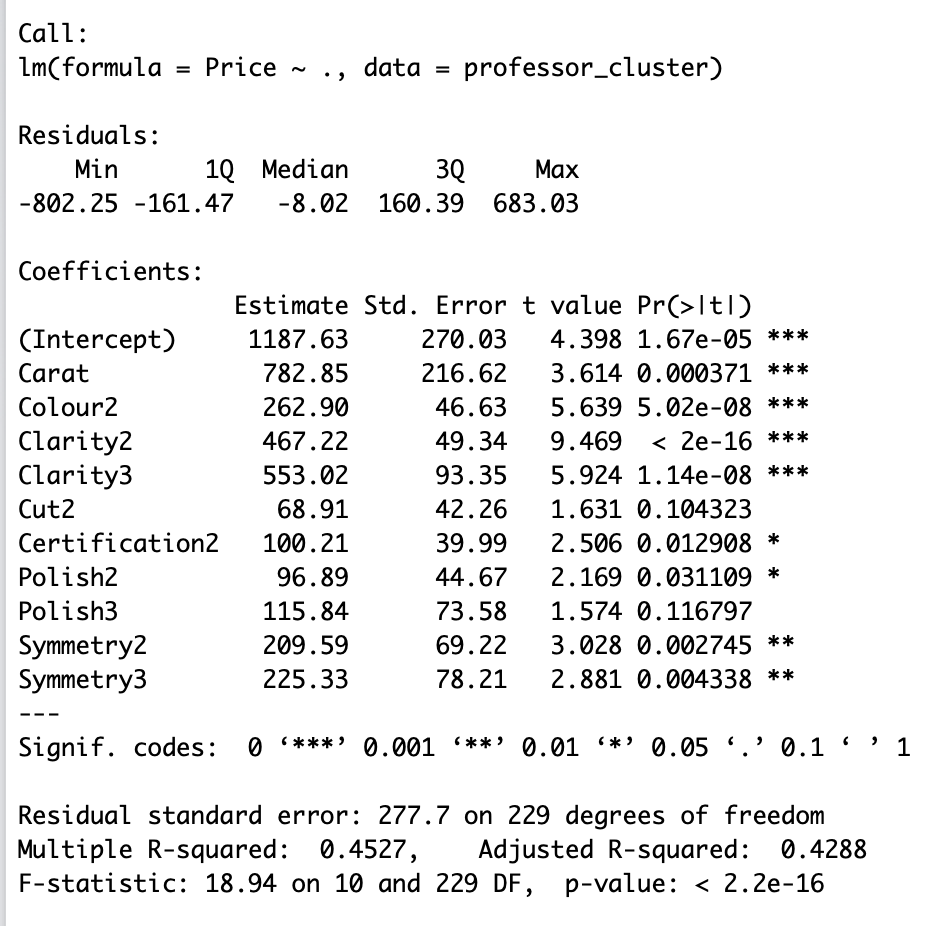


Figure 8: Model constructed after regrouping Cut variables into 2 groups

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 8, 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 7: Polish proportions after doing feature engineering

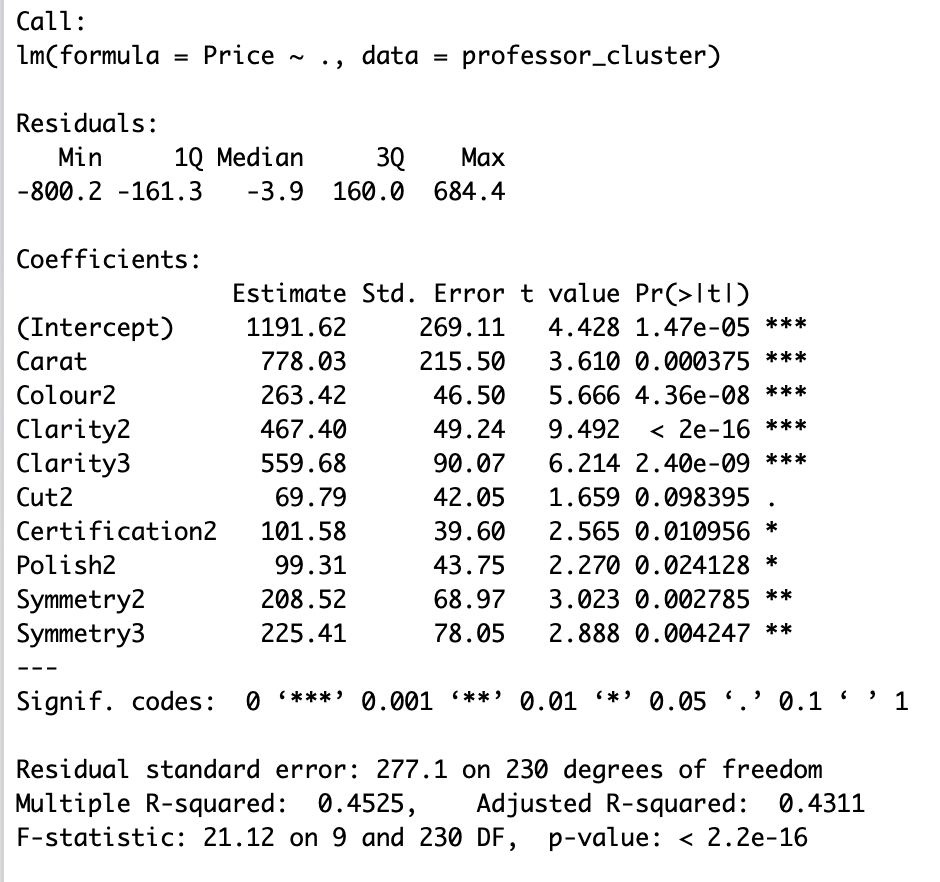


Figure 9: Model constructed after regrouping Polish variables into 2 groups

Step 4: 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 9. 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 should be done to perform a better result of the regression model.

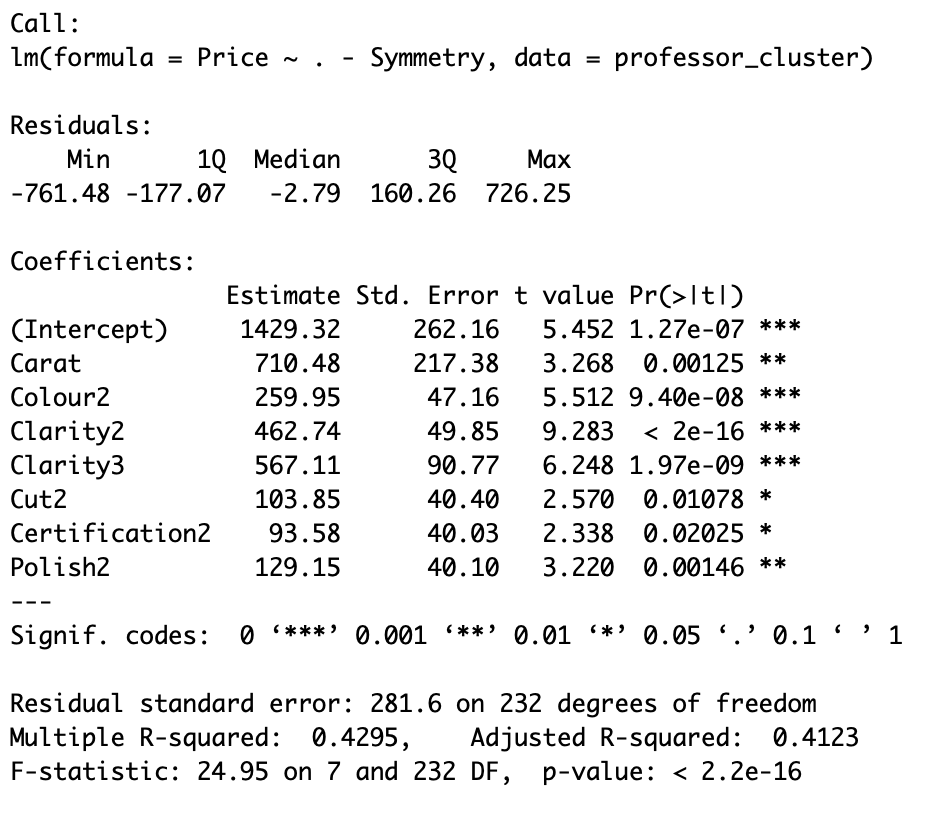


Figure 10: Model constructed after removing Symmetry variables

Step 5: 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 8 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 8: Color Proportions

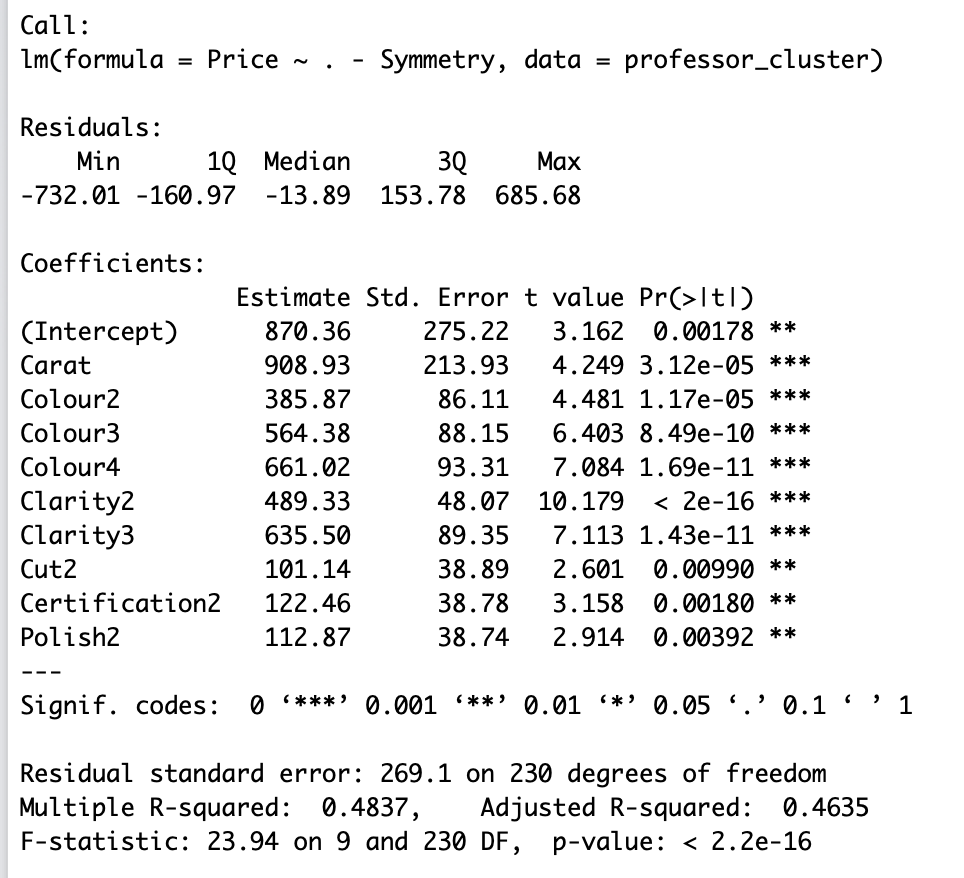


Figure 11: Model constructed after ungrouping Colour variables

## 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 Color 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 Color 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 Color 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.

## Disadvantages 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. Final suggestion for the professor is to consider the ring price as a factor before deciding. If the ring attached to diamond is more expensive than $312.80 the quote given is fair. Otherwise, the quote is more expensive than the combination of the diamond price calculated by our model plus the price of the ring.

APPENDIX

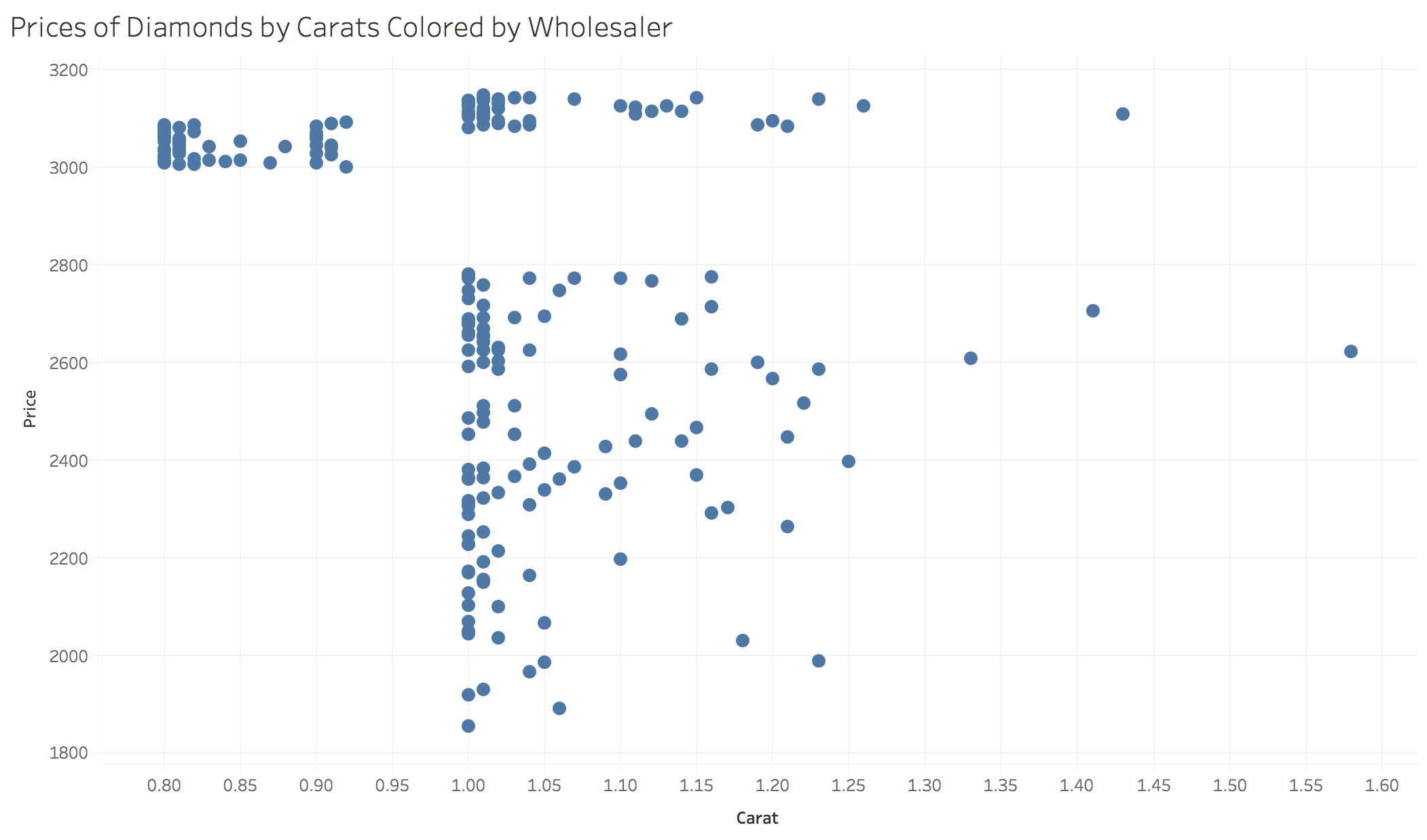
****

Figure 12: Cluster where professors ring belongs

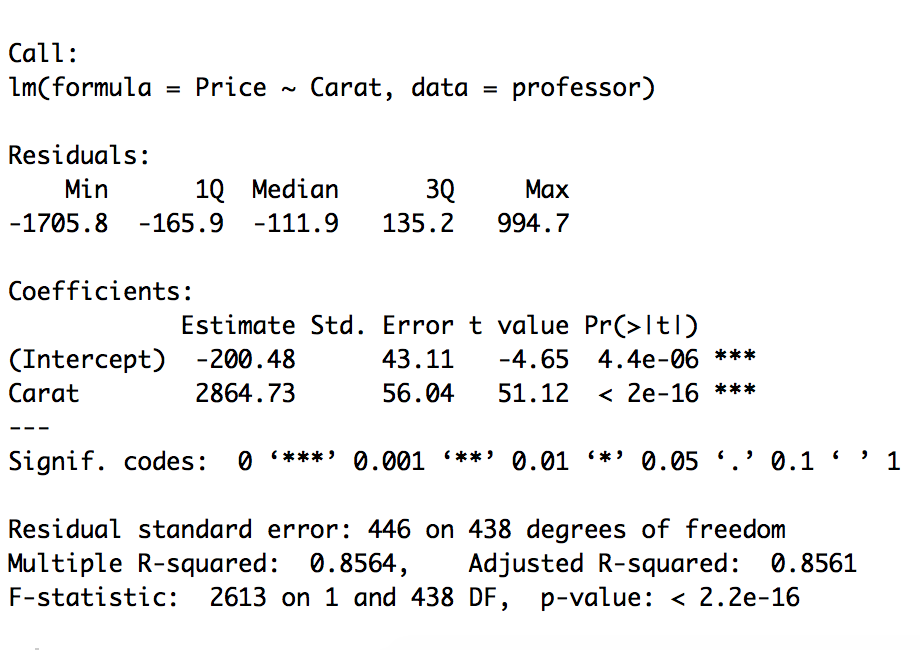
****

Figure 13: Carat vs Price

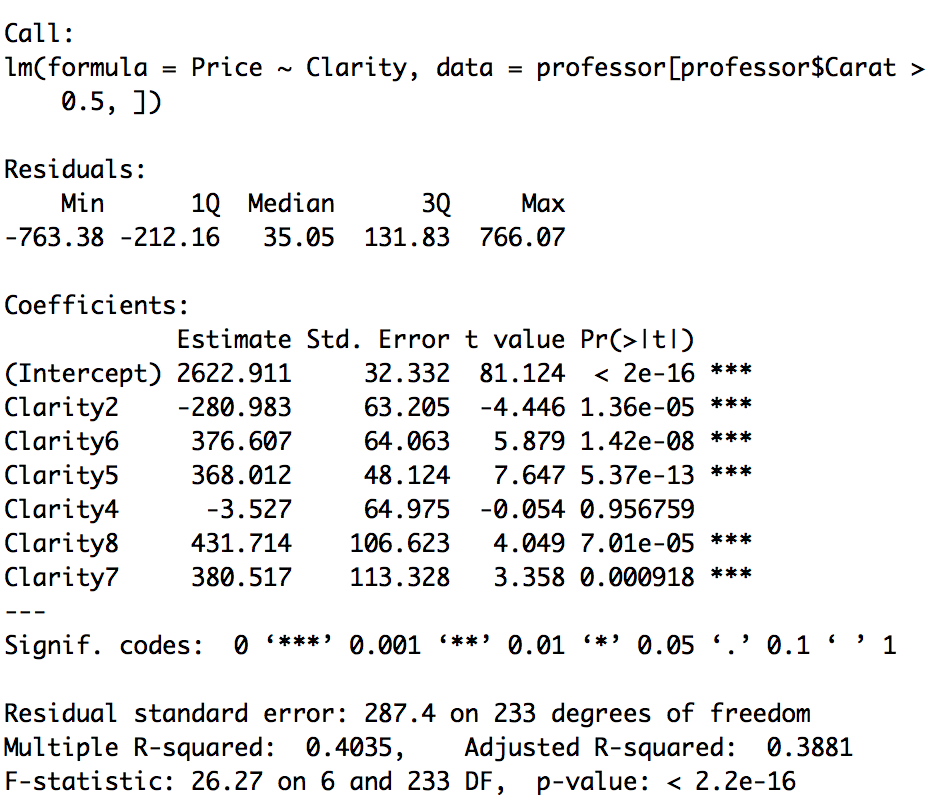
****

Figure 14: Price vs Clarity all levels

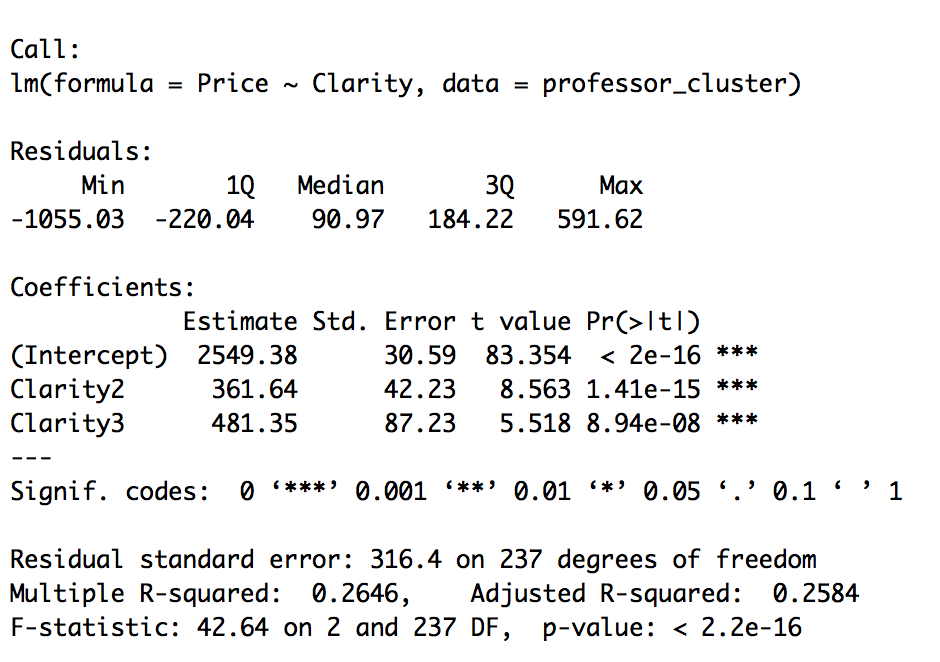
****

Figure 15: Price vs Clarity 3 levels

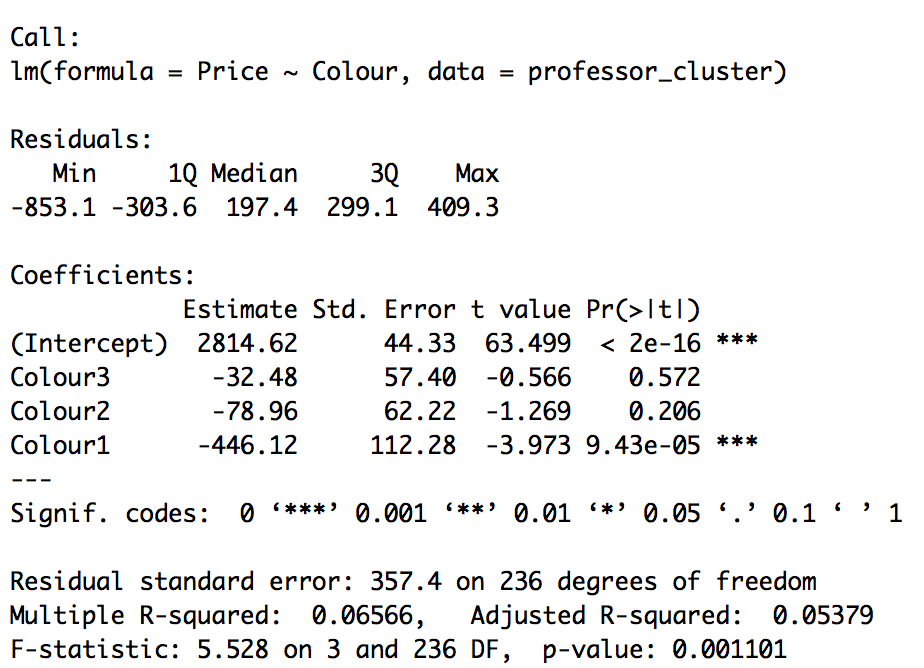
****

Figure 16: Color vs Price 4 categories

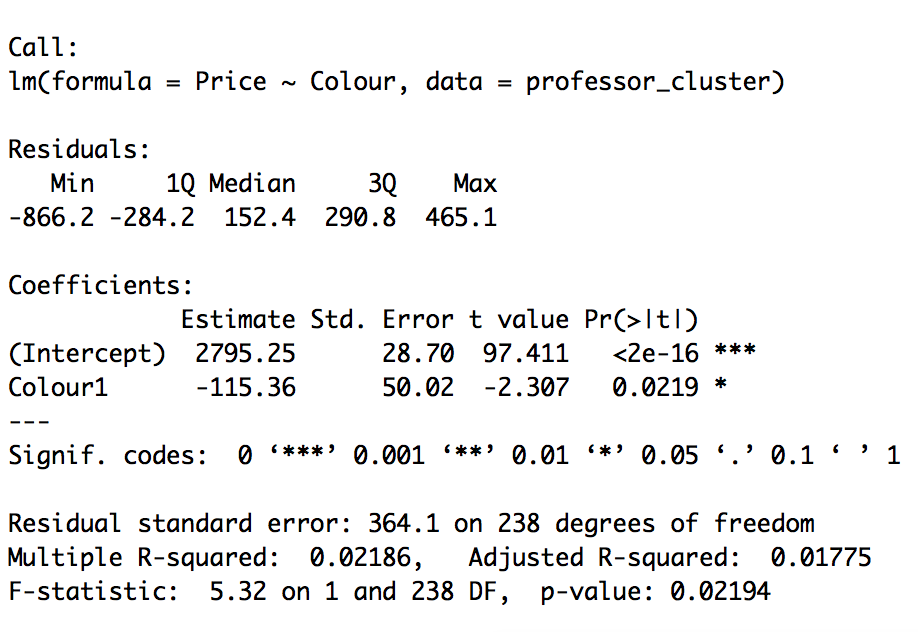


Figure 17: Color vs Price 2 categories

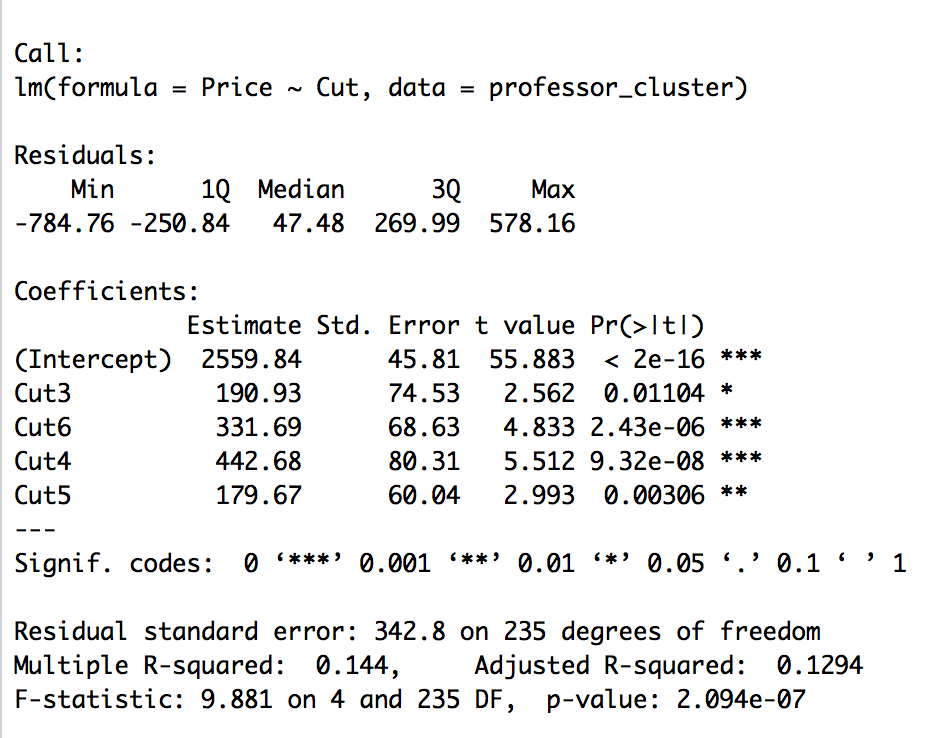
****

Figure 18: Price vs Cut

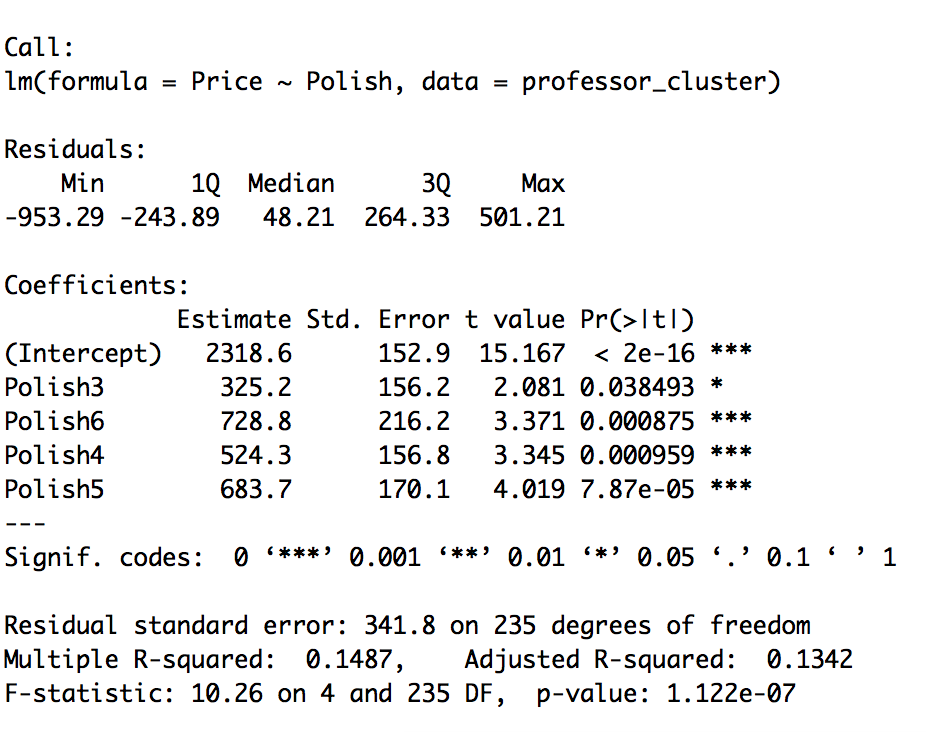
****

Figure 19: Price vs Polish all categories

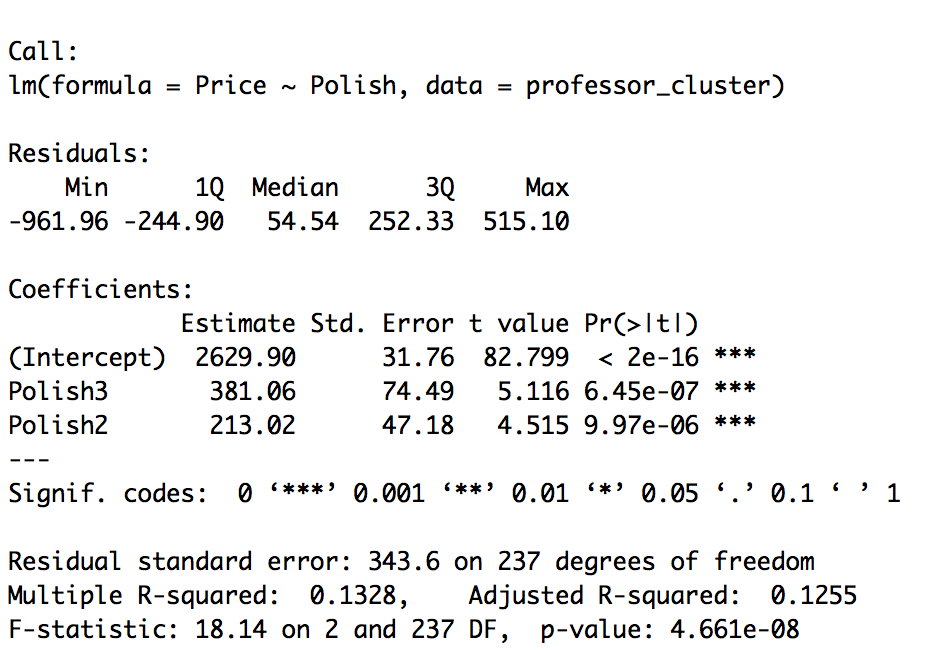


Figure 20: Price vs Polish 3 categories

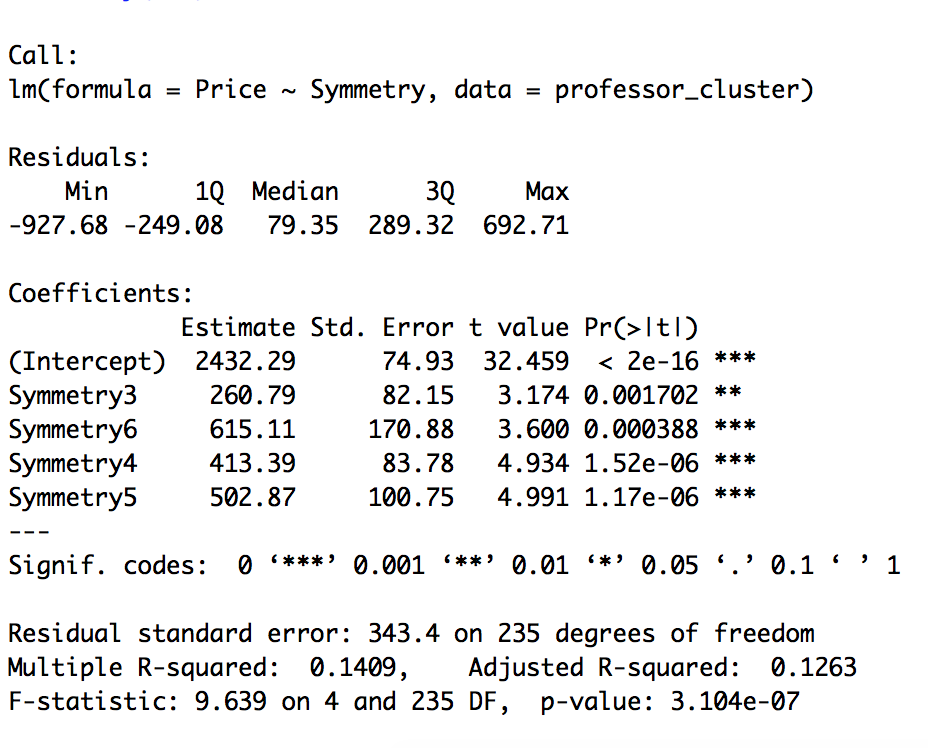


Figure 21: Price vs Symmetry all categories

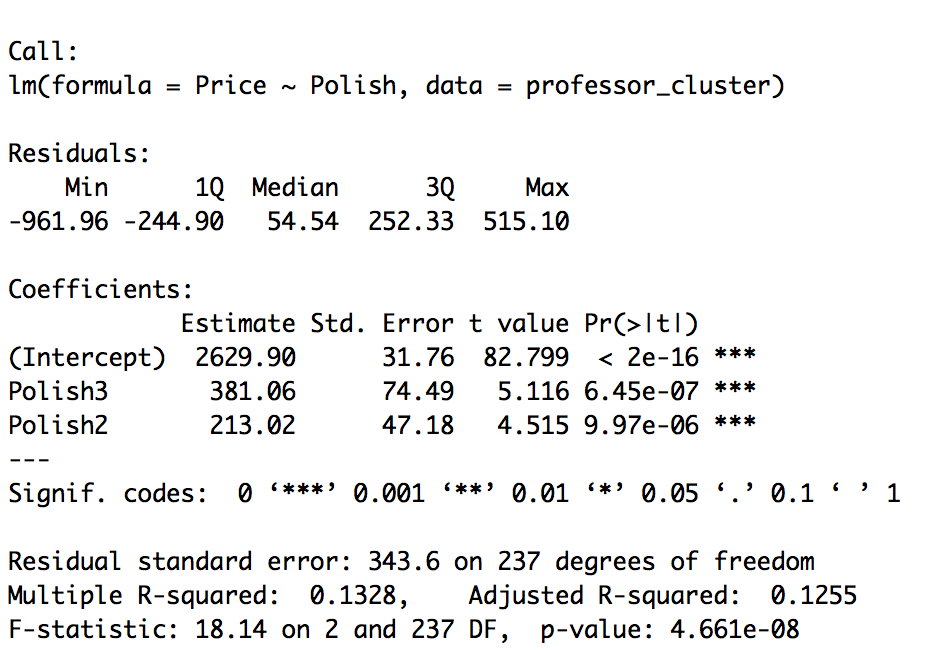
****

Figure 22: Price vs Symmetry 3 categories

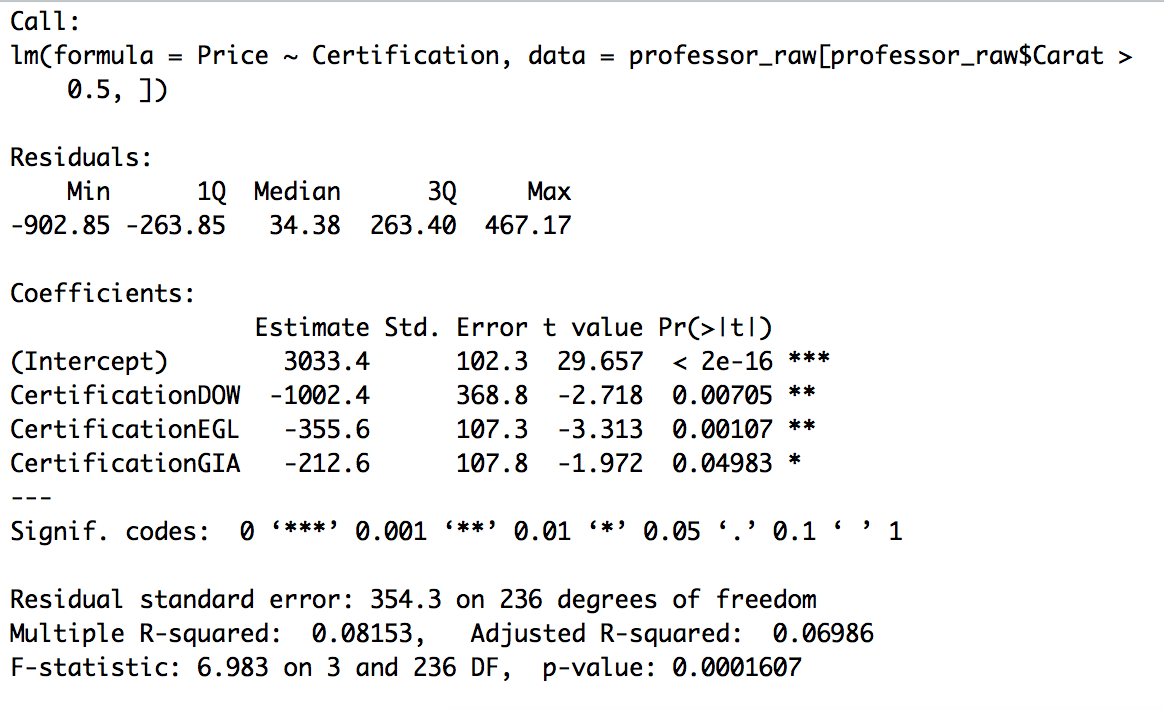
****

Figure 23: Price vs Certification Initial Categories

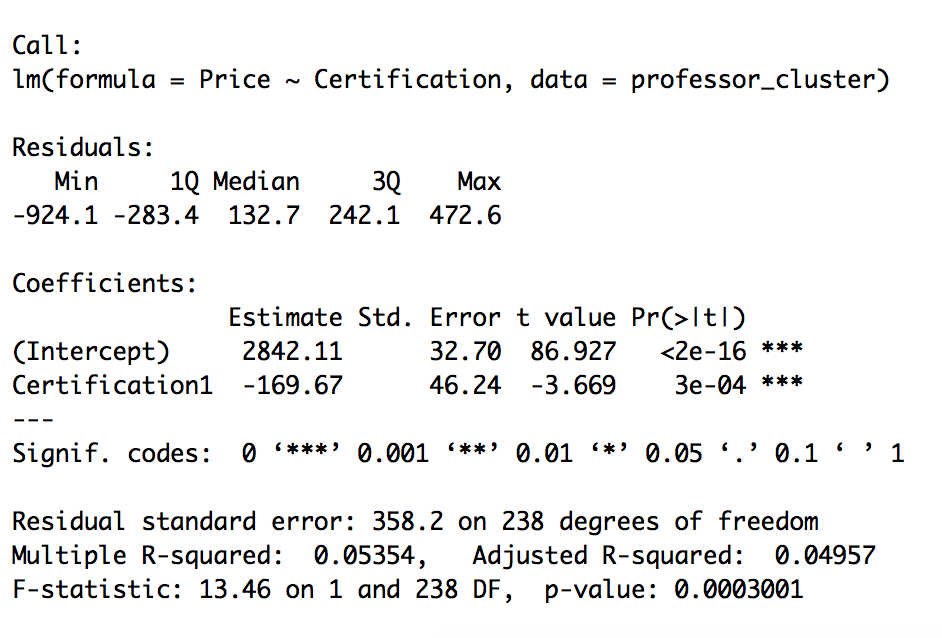


Figure 24: Price vs Certification Most Respected Labs vs Others

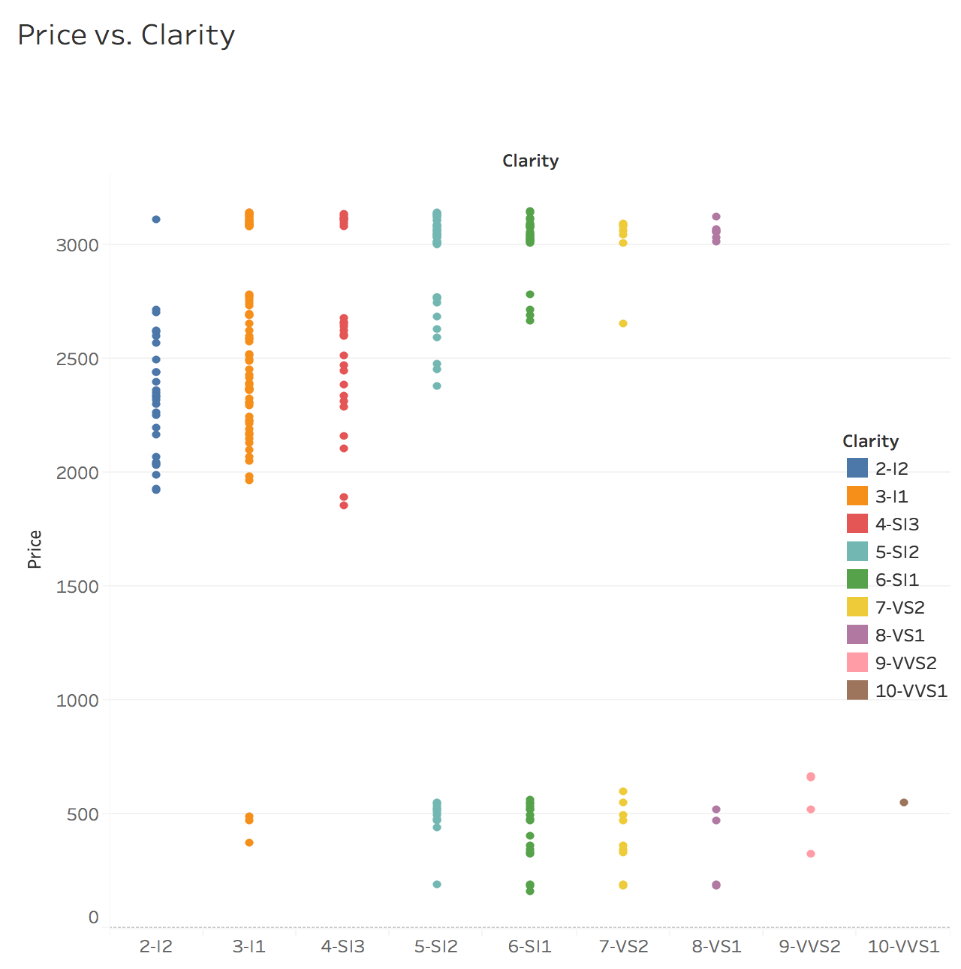
****

Figure 25: Price vs Clarity Bivariate

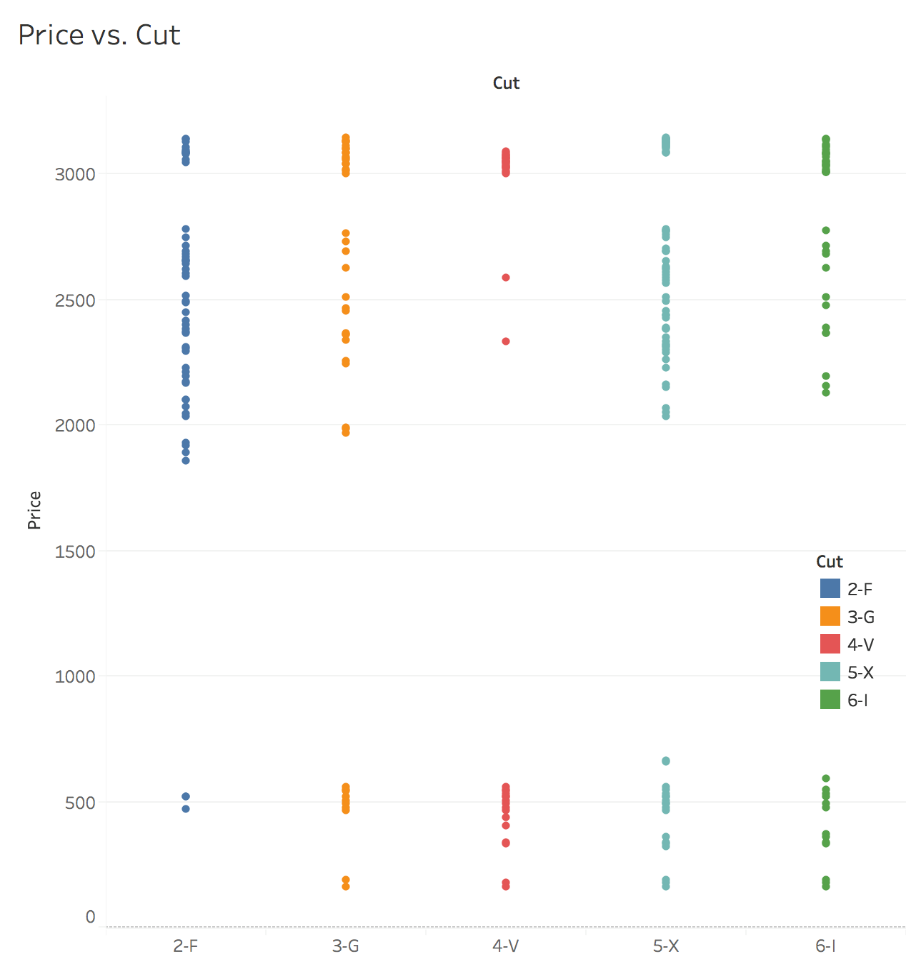


Figure 26: Price vs Cut Bivariate

****

Figure 27: Price vs Polish Bivariate

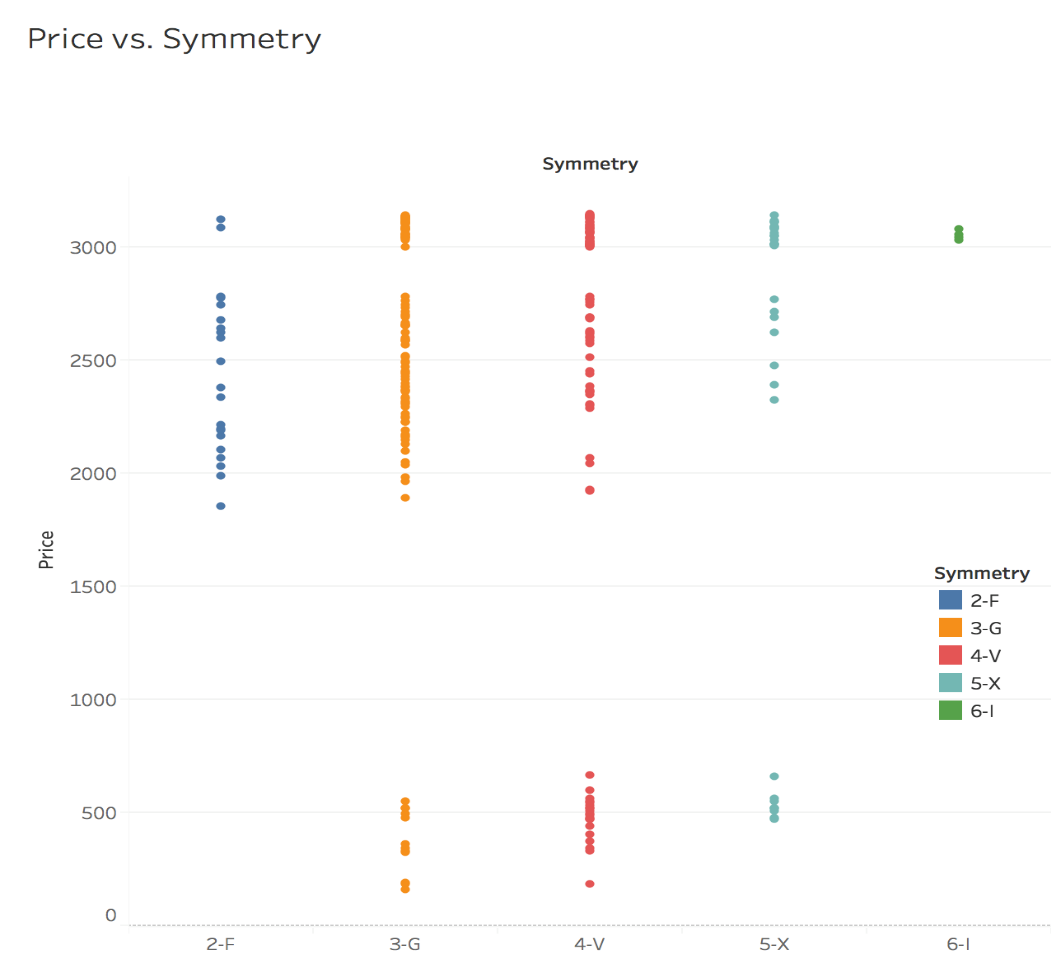
****

Figure 28: Price vs Symmetry Bivariate

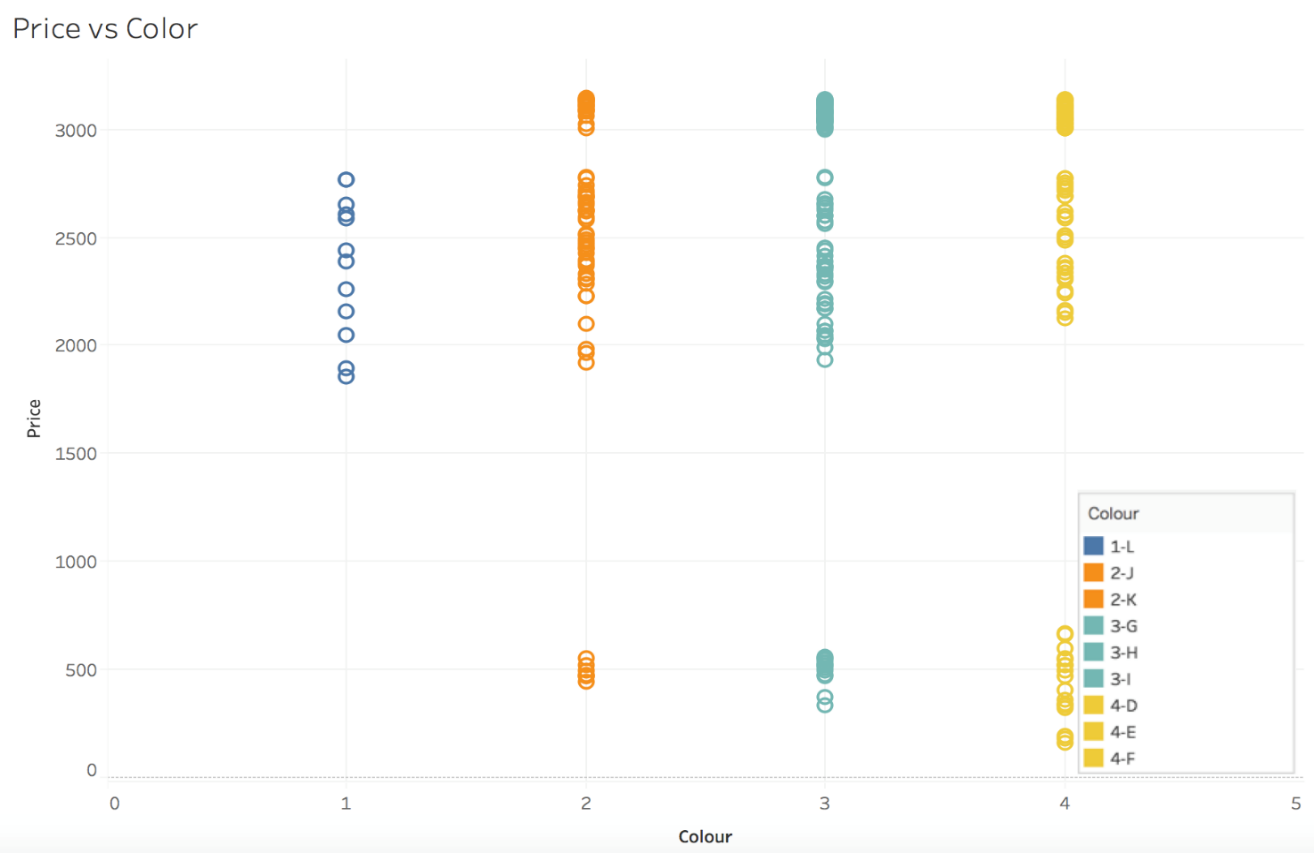
****

Figure 29: Price vs Color Bivariate