STATISTICS & IT'S APPLICATION IN BUSINESS 5520
FINAL PROJECT
PROFESSOR PROPOSES

Submitted by:
Caner Adil Irfanoglu – A00425840
Tom Tong – A00369116
Diven Sambhwani – A00425915
Madeleine Leong – A00430926
Sreeraj Punnoli – A00429404

Submitted to:
Dr. Michael Zhang

Submitted on:
Dec 9, 2018

Table of Contents

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

As the result of marketing of diamond industry, many of us believed diamond is the hardest material in the world and it is very very rare, as a good wish people also would like their relationship between their partner is also strongest in the world.

In the early 1940s, the diamond mining output increase is unprecedented, people want to figure out a way to assess a diamond's value so they can have a universal method to assess diamond from different regions and retailers. The founder of Gemological Institute of America (GIA) Robert M. Shipley has introduced the term 4Cs to help his students to remember the four factors that can describe a cut diamond: color, clarity, cut and carat weight. Slowly this 4Cs became a standard way to measure the value of the diamond.

The first "c" s the color, the color used for describing a diamond is begin with grade D which is the first letter for diamond. From grade D to Z it became a color-grading system to measures the degree of colorlessness of a diamond. D is most colorless and Z is the color trend to yellow. When the other characteristics are the same a diamond with color D will have more value than color Z. If the diamond has a pure special color such as pink, yellow or green it will contain significantly higher value.

Figure 1: Color of Diamond

The second "c" refers to clarity, this indicates how much inclusions in the internal of a diamond and how much blemishes on the external of a diamond. The number, size, relief, nature and position of those inclusions and blemishes will help determine the value of a diamond. There are total of 11 grades for clarity, they are: Flawless (FL), Internally Flawless (IF), Very Very

Slightly Included (VVS_1 and VVS_2), Very Slightly Included (VS_1 and VS_2), Slightly Included (SI_1 and SI_2) and Included (I_1, I_2 and I_3). The higher grading the higher value the diamond has.

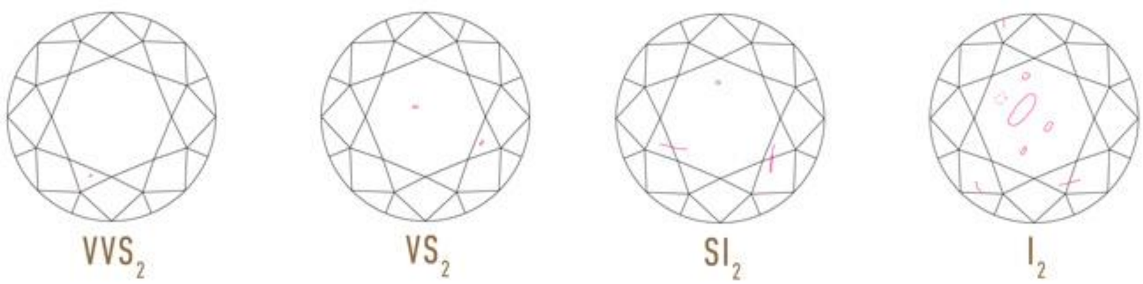

Figure 2: Clarity of Diamond

The third "c" means cut, which is the most important process in diamond manufactory. A well cut diamond will maximum the reflection of the light and make it looks brilliant. The cut is grading in the following grade: Excellent Cut, Very Good Cut, Good Cut, Fair Cut and Poor Cut. The better cutting a diamond has the higher value.

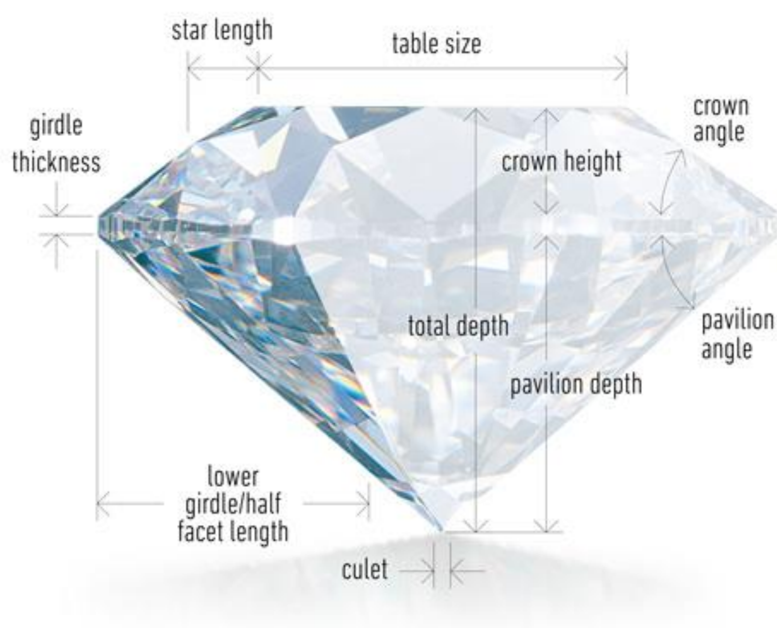

Figure 3: Cut of Diamond

The last “c” is the carat weight of the diamond, carat is the unit to describe the weight of a diamond, a carat is equal to 200 milligrams. The heavier the more expensive.

Figure 4: Carat of Diamond

There are also other characteristics may affect the value of a diamond, they are Polish, Symmetry and Fluorescence.

Problem Statement

The professor’s girlfriend had already hinted third times about marriage, under this pressure the professor finally decided for the proposal. Find a nice diamond engagement ring is the first step but is not as easy as his initially thought.

Is it fair to pay $3,100 for a diamond which has 0.9 carats in weight, J color, SI2 clarity, Very good cut, Good polish, Very Good Symmetry and come with GIA certificate in store become the biggest question in the professor’s mind.

Exhibit 2

THE PROFESSOR’S DIAMOND ENGAGEMENT RING

Price	$3,100
Carat Weight	0.9
Cut	Very Good
Color	J
Clarity	SI2
Polish	Good
Symmetry	Very Good
Certification	GIA

Figure 5: Professor’s Diamond Engagement Ring

Now the professor decided to learn the 4Cs assessment standard and collecting available information from different diamond online retailer to build a module to assess a diamond's value based on the characteristics in order to answer his question.

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.

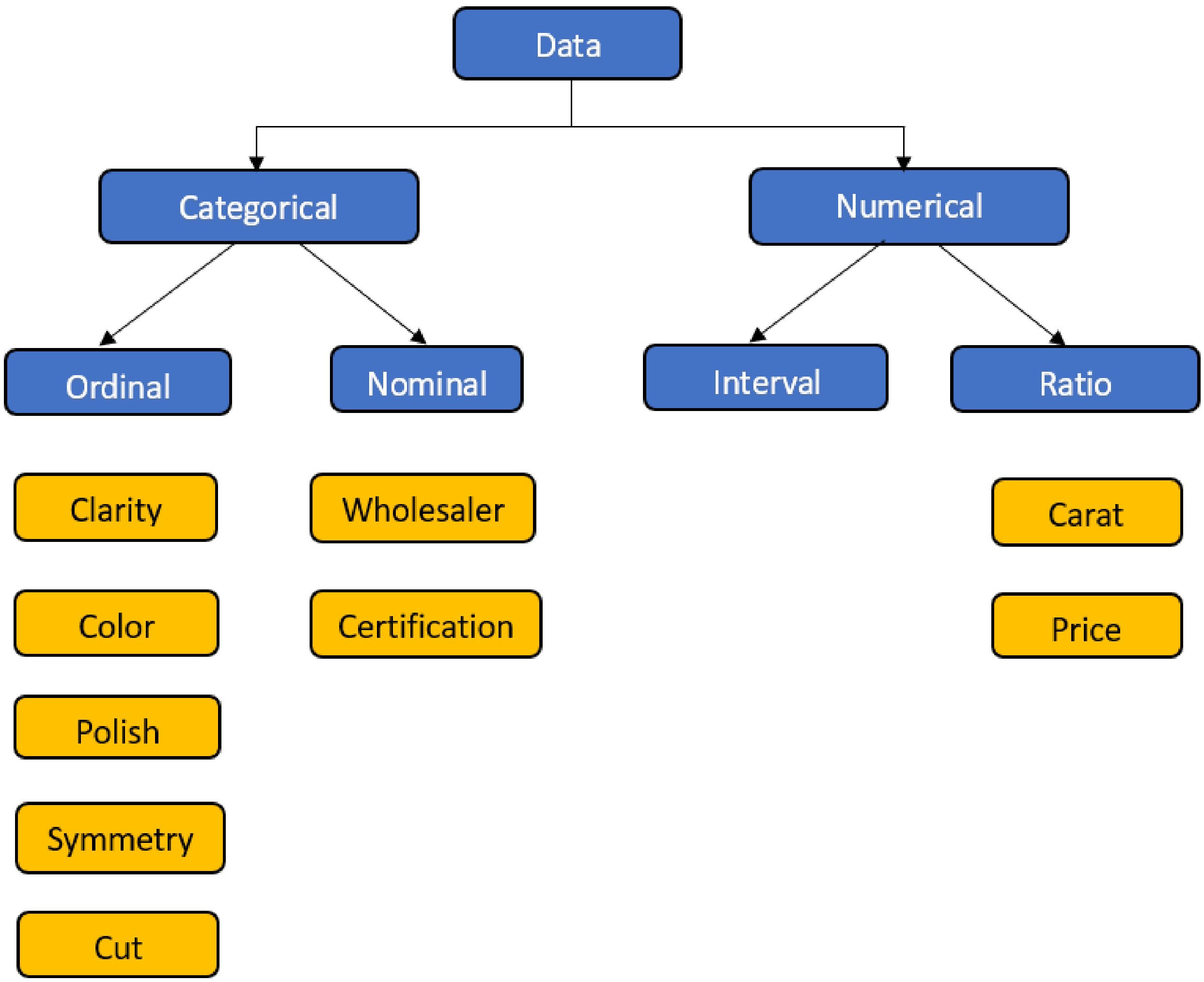

Figure 1: Data Types of Variables in the Dataset

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.

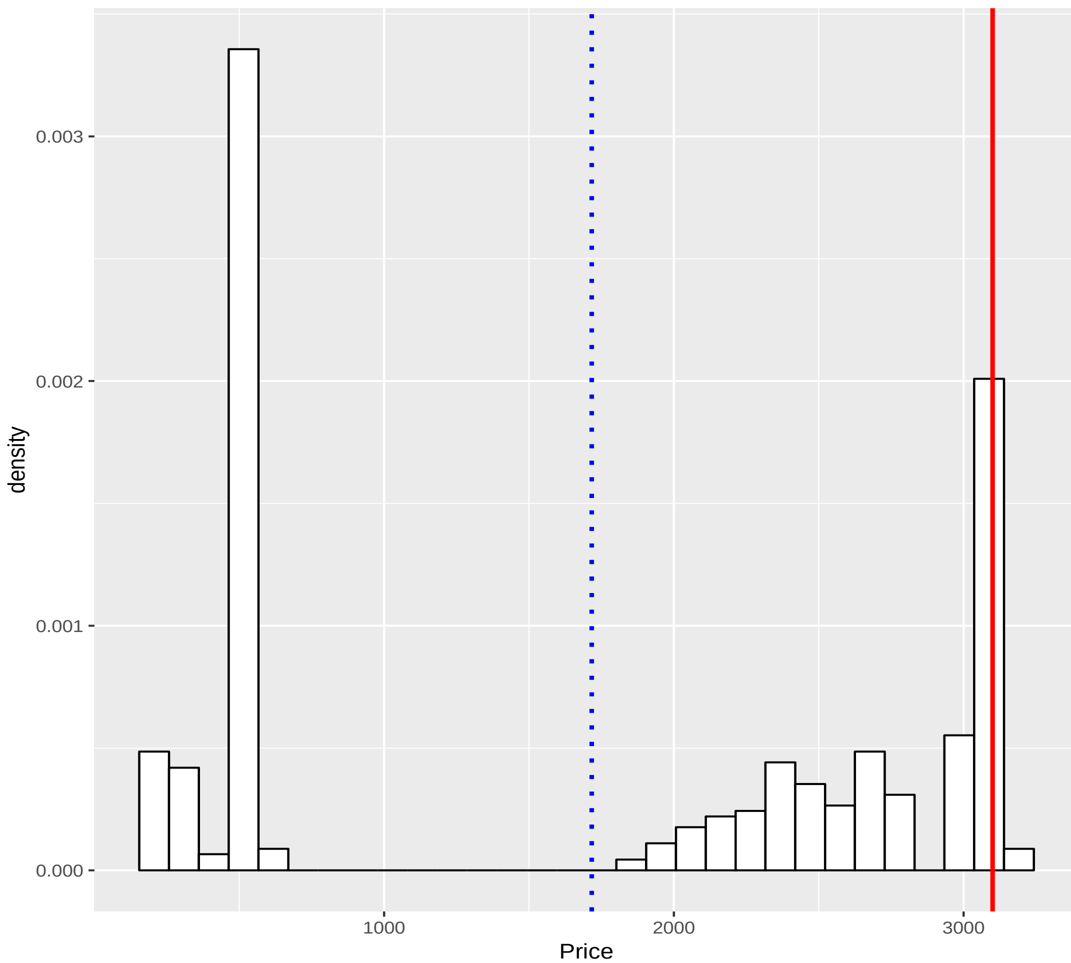

Figure 2: Density Distribution of Diamond Prices

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.

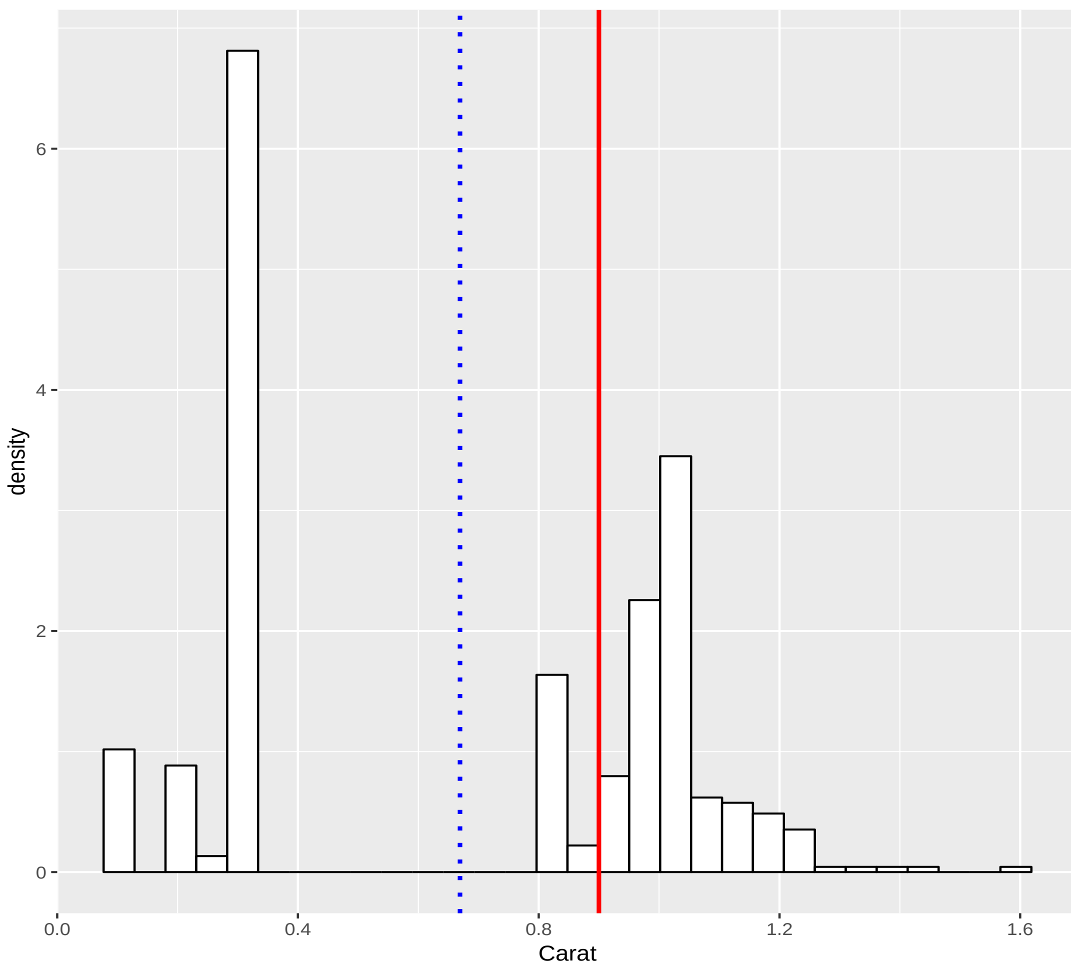

Figure 3: Density Distribution of Diamond Carats

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.

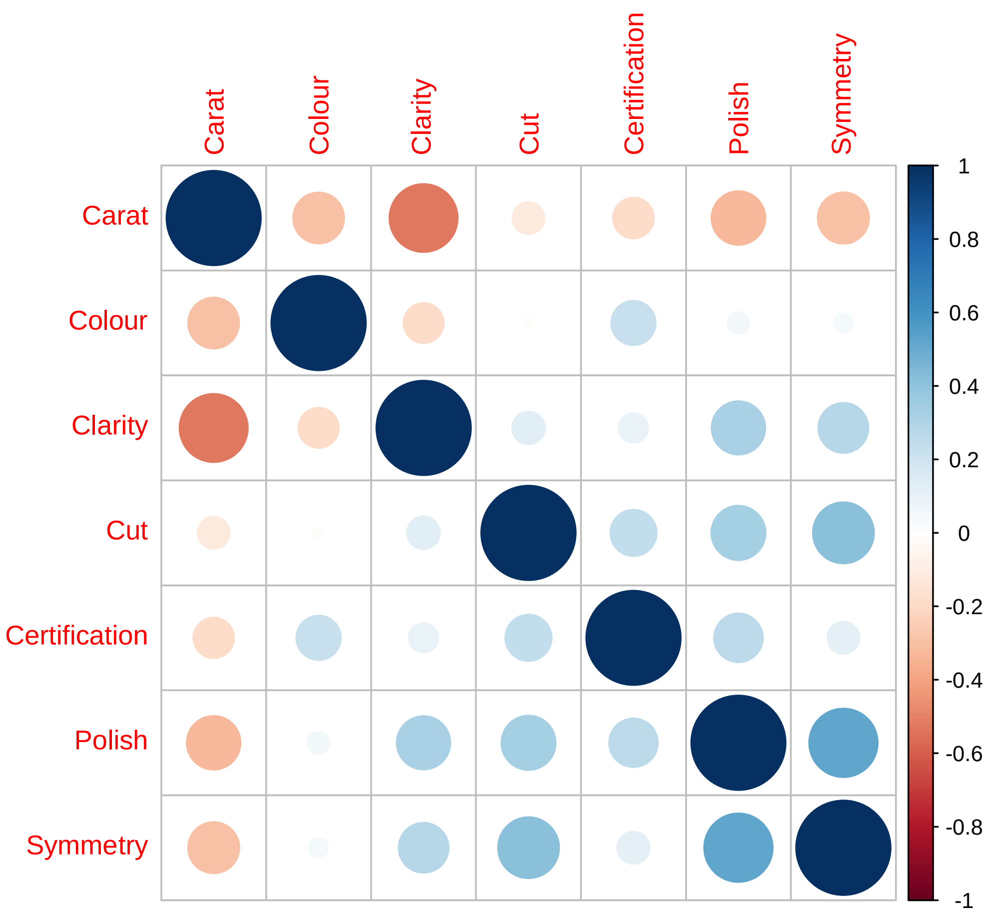

Figure 4: Multicollinearity of Independent Variables

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 modeled based on the diamonds corresponding to this cluster. This is achieved by filtering out the diamonds having lower than 0.5 carats.

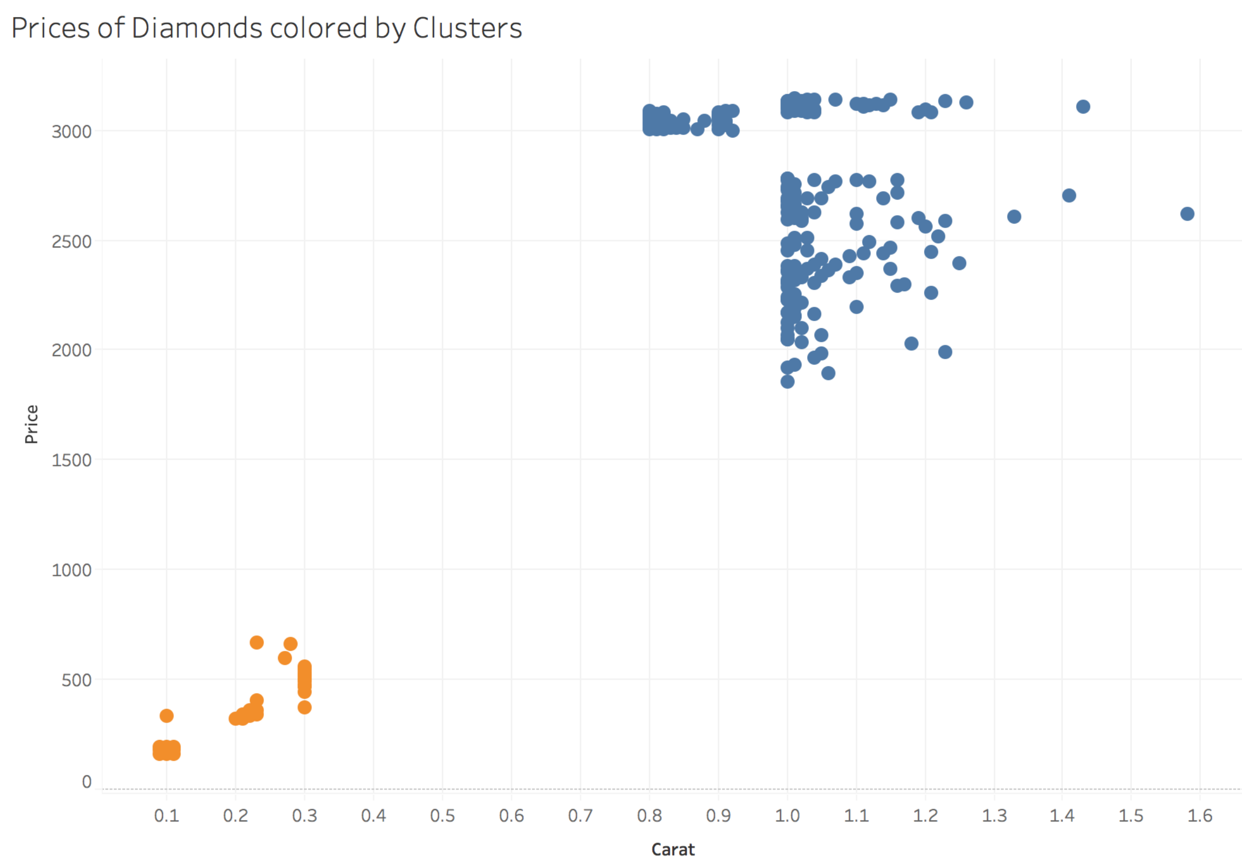

Figure 5: Carat vs. Price colored by cluster

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;

```
Cut
       n  missing distinct
     240        0        5

Value          2     3     6     4     5
Frequency     56    34    45    27    78
Proportion 0.233 0.142 0.188 0.112 0.325
--------------------------------------------------------------------------------------------------------
Certification
       n  missing distinct
     240        0        4

Value        AGS   DOW   EGL   GIA
Frequency     12     1   119   108
Proportion 0.050 0.004 0.496 0.450
--------------------------------------------------------------------------------------------------------
Polish
       n  missing distinct
     240        0        5

Value          2     3     6     4     5
Frequency      5   112     5    97    21
Proportion 0.021 0.467 0.021 0.404 0.088
--------------------------------------------------------------------------------------------------------
Symmetry
       n  missing distinct
     240        0        5

Value          2     3     6     4     5
Frequency     21   104     5    84    26
Proportion 0.088 0.433 0.021 0.350 0.108
--------------------------------------------------------------------------------------------------------
Price
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     240        0      172        1     2757    405.7     2070     2211     2450     2890     3083     3125     3138

lowest : 1856 1892 1918 1929 1966, highest: 3139 3140 3141 3142 3145
--------------------------------------------------------------------------------------------------------
Wholesaler
       n  missing distinct     Info     Mean      Gmd
     240        0        2    0.563     1.75   0.3766

Value        1    2
Frequency   60  180
Proportion 0.25 0.75
--------------------------------------------------------------------------------------------------------
professor[professor$Carat > 0.5, ]

 9  Variables      240  Observations
--------------------------------------------------------------------------------------------------------
Carat
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     240        0       40    0.984    1.003   0.1282    0.800    0.810    0.980    1.010    1.040    1.151    1.210

lowest : 0.80 0.81 0.82 0.83 0.84, highest: 1.26 1.33 1.41 1.43 1.58
--------------------------------------------------------------------------------------------------------
Colour
       n  missing distinct
     240        0        4

Value          4     3     2     1
Frequency     65    96    67    12
Proportion 0.271 0.400 0.279 0.050
--------------------------------------------------------------------------------------------------------
Clarity
       n  missing distinct
     240        0        7

Value          3     2     6     5     4     8     7
Frequency     79    28    27    65    26     8     7
Proportion 0.329 0.117 0.112 0.271 0.108 0.033 0.029
--------------------------------------------------------------------------------------------------------
```

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

```
Call:
lm(formula = Price ~ ., data = professor_cluster)

Residuals:
    Min      1Q  Median      3Q     Max
-777.74 -152.99   -3.54  149.63  683.01

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1189.05     281.80   4.219 3.55e-05 ***
Carat            776.36     231.59   3.352 0.000939 ***
Colour2          265.62      47.20   5.628 5.37e-08 ***
Clarity2         467.67      50.25   9.307  < 2e-16 ***
Clarity3         557.34      94.33   5.908 1.26e-08 ***
Cut3              51.87      63.50   0.817 0.414862
Cut4             115.11      79.32   1.451 0.148084
Cut5             105.61      54.89   1.924 0.055595 .
Cut6              44.53      67.46   0.660 0.509880
Certification2    89.73      41.25   2.175 0.030652 *
Polish2          102.61      45.05   2.277 0.023695 *
Polish3          141.98      76.86   1.847 0.066007 .
Symmetry2        190.08      71.35   2.664 0.008273 **
Symmetry3        207.44      80.75   2.569 0.010845 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 278.2 on 226 degrees of freedom
Multiple R-squared:  0.4577,	Adjusted R-squared:  0.4265
F-statistic: 14.67 on 13 and 226 DF,  p-value: < 2.2e-16
```

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 R^2) 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 R^2 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.

```
Call:
lm(formula = Price ~ ., data = professor_cluster)

Residuals:
    Min      1Q  Median      3Q     Max 
-802.25 -161.47   -8.02  160.39  683.03 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1187.63     270.03   4.398 1.67e-05 ***
Carat            782.85     216.62   3.614 0.000371 ***
Colour2          262.90      46.63   5.639 5.02e-08 ***
Clarity2         467.22      49.34   9.469  < 2e-16 ***
Clarity3         553.02      93.35   5.924 1.14e-08 ***
Cut2              68.91      42.26   1.631 0.104323    
Certification2   100.21      39.99   2.506 0.012908 *  
Polish2           96.89      44.67   2.169 0.031109 *  
Polish3          115.84      73.58   1.574 0.116797    
Symmetry2        209.59      69.22   3.028 0.002745 ** 
Symmetry3        225.33      78.21   2.881 0.004338 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 277.7 on 229 degrees of freedom
Multiple R-squared:  0.4527,	Adjusted R-squared:  0.4288 
F-statistic: 18.94 on 10 and 229 DF,  p-value: < 2.2e-16
```

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 R^2 is examined to see if the model has improved. From Figure 8, polish variable become significant to the model after regrouping and the adjusted R^2 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

```
Call:
lm(formula = Price ~ ., data = professor_cluster)

Residuals:
   Min     1Q Median     3Q    Max 
-800.2 -161.3   -3.9  160.0  684.4 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1191.62     269.11   4.428 1.47e-05 ***
Carat            778.03     215.50   3.610 0.000375 ***
Colour2          263.42      46.50   5.666 4.36e-08 ***
Clarity2         467.40      49.24   9.492  < 2e-16 ***
Clarity3         559.68      90.07   6.214 2.40e-09 ***
Cut2              69.79      42.05   1.659 0.098395 .  
Certification2   101.58      39.60   2.565 0.010956 *  
Polish2           99.31      43.75   2.270 0.024128 *  
Symmetry2        208.52      68.97   3.023 0.002785 ** 
Symmetry3        225.41      78.05   2.888 0.004247 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 277.1 on 230 degrees of freedom
Multiple R-squared:  0.4525,	Adjusted R-squared:  0.4311 
F-statistic: 21.12 on 9 and 230 DF,  p-value: < 2.2e-16
```

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 R^2 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.

```
Call:
lm(formula = Price ~ . - Symmetry, data = professor_cluster)

Residuals:
    Min      1Q  Median      3Q     Max
-761.48 -177.07   -2.79  160.26  726.25

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1429.32     262.16   5.452 1.27e-07 ***
Carat            710.48     217.38   3.268  0.00125 **
Colour2          259.95      47.16   5.512 9.40e-08 ***
Clarity2         462.74      49.85   9.283  < 2e-16 ***
Clarity3         567.11      90.77   6.248 1.97e-09 ***
Cut2             103.85      40.40   2.570  0.01078 *
Certification2    93.58      40.03   2.338  0.02025 *
Polish2          129.15      40.10   3.220  0.00146 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 281.6 on 232 degrees of freedom
Multiple R-squared:  0.4295,    Adjusted R-squared:  0.4123
F-statistic: 24.95 on 7 and 232 DF,  p-value: < 2.2e-16
```

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 R^2 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 R^2.

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

Table 8: Color Proportions

```
Call:
lm(formula = Price ~ . - Symmetry, data = professor_cluster)

Residuals:
    Min      1Q  Median      3Q     Max
-732.01 -160.97  -13.89  153.78  685.68

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      870.36     275.22   3.162  0.00178 **
Carat            908.93     213.93   4.249 3.12e-05 ***
Colour2          385.87      86.11   4.481 1.17e-05 ***
Colour3          564.38      88.15   6.403 8.49e-10 ***
Colour4          661.02      93.31   7.084 1.69e-11 ***
Clarity2         489.33      48.07  10.179  < 2e-16 ***
Clarity3         635.50      89.35   7.113 1.43e-11 ***
Cut2             101.14      38.89   2.601  0.00990 **
Certification2   122.46      38.78   3.158  0.00180 **
Polish2          112.87      38.74   2.914  0.00392 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 269.1 on 230 degrees of freedom
Multiple R-squared:  0.4837,	Adjusted R-squared:  0.4635
F-statistic: 23.94 on 9 and 230 DF,  p-value: < 2.2e-16
```

Figure 11: Model constructed after ungrouping Colour variables

Model Summary

The final model for the diamond pricing is:

$$
\begin{aligned}
\text{Price} = \; & 870.36 + 908.93 \times \text{Carat} + 385.87 \times \text{Colour2} \\
& + 564.38 \times \text{Colour3} + 661.02 \times \text{Colour4} + 489.33 \times \text{Clarity2} \\
& + 635.50 \times \text{Clarity3} + 101.14 \times \text{Cut2} \\
& + 122.46 \times \text{Certification2} + 112.87 \times \text{Polish2}
\end{aligned}
$$

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,

$$\begin{aligned}\text{Price} = \ & 870.36 + 908.93 \times \text{Carat} + 385.87 \times \text{Colour2} + 564.38 \times \text{Colour3} \\ & + 661.02 \times \text{Colour4} + 489.33 \times \text{Clarity2} + 635.50 \times \text{Clarity3} \\ & + 101.14 \times \text{Cut2} + 122.46 \times \text{Certification2} + 112.87 \times \text{Polish2}\end{aligned}$$

Price Calculated based on Model: <u>\$2,787.20</u>

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

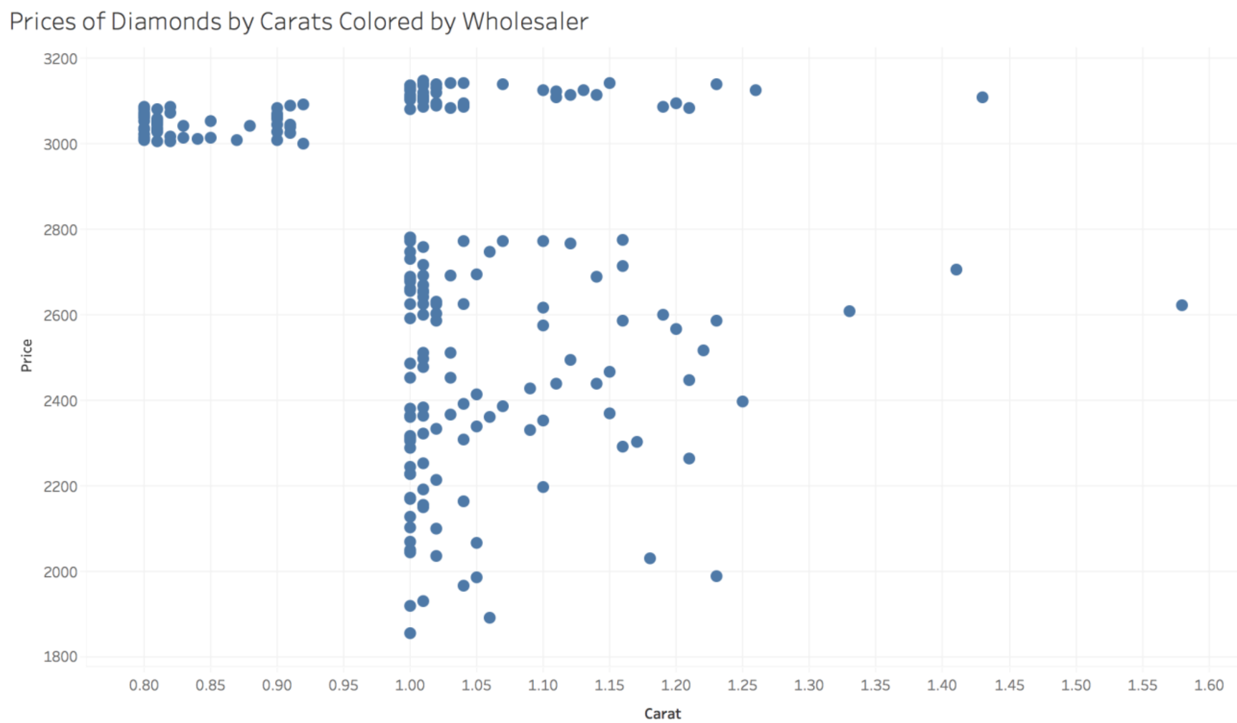

Figure 12: Cluster where professors ring belongs

```
Call:
lm(formula = Price ~ Carat, data = professor)

Residuals:
    Min      1Q  Median      3Q     Max
-1705.8  -165.9  -111.9   135.2   994.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -200.48      43.11   -4.65  4.4e-06 ***
Carat        2864.73      56.04   51.12  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 446 on 438 degrees of freedom
Multiple R-squared:  0.8564,    Adjusted R-squared:  0.8561
F-statistic:  2613 on 1 and 438 DF,  p-value: < 2.2e-16
```

Figure 13: Carat vs Price

```
Call:
lm(formula = Price ~ Clarity, data = professor[professor$Carat >
    0.5, ])

Residuals:
    Min      1Q  Median      3Q     Max 
-763.38 -212.16   35.05  131.83  766.07 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2622.911     32.332  81.124  < 2e-16 ***
Clarity2    -280.983     63.205  -4.446 1.36e-05 ***
Clarity6     376.607     64.063   5.879 1.42e-08 ***
Clarity5     368.012     48.124   7.647 5.37e-13 ***
Clarity4      -3.527     64.975  -0.054 0.956759    
Clarity8     431.714    106.623   4.049 7.01e-05 ***
Clarity7     380.517    113.328   3.358 0.000918 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 287.4 on 233 degrees of freedom
Multiple R-squared:  0.4035,	Adjusted R-squared:  0.3881 
F-statistic: 26.27 on 6 and 233 DF,  p-value: < 2.2e-16
```

Figure 14: Price vs Clarity all levels

```
Call:
lm(formula = Price ~ Clarity, data = professor_cluster)

Residuals:
     Min       1Q   Median       3Q      Max 
-1055.03  -220.04    90.97   184.22   591.62 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2549.38      30.59  83.354  < 2e-16 ***
Clarity2      361.64      42.23   8.563 1.41e-15 ***
Clarity3      481.35      87.23   5.518 8.94e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 316.4 on 237 degrees of freedom
Multiple R-squared:  0.2646,	Adjusted R-squared:  0.2584 
F-statistic: 42.64 on 2 and 237 DF,  p-value: < 2.2e-16
```

Figure 15: Price vs Clarity 3 levels

```
Call:
lm(formula = Price ~ Colour, data = professor_cluster)

Residuals:
   Min     1Q Median     3Q    Max
-853.1 -303.6  197.4  299.1  409.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2814.62      44.33  63.499  < 2e-16 ***
Colour3       -32.48      57.40  -0.566    0.572
Colour2       -78.96      62.22  -1.269    0.206
Colour1      -446.12     112.28  -3.973 9.43e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 357.4 on 236 degrees of freedom
Multiple R-squared:  0.06566,   Adjusted R-squared:  0.05379
F-statistic: 5.528 on 3 and 236 DF,  p-value: 0.001101
```

Figure 16: Color vs Price 4 categories

```
Call:
lm(formula = Price ~ Colour, data = professor_cluster)

Residuals:
   Min     1Q Median     3Q    Max
-866.2 -284.2  152.4  290.8  465.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2795.25      28.70  97.411   <2e-16 ***
Colour1      -115.36      50.02  -2.307   0.0219 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 364.1 on 238 degrees of freedom
Multiple R-squared:  0.02186,   Adjusted R-squared:  0.01775
F-statistic:  5.32 on 1 and 238 DF,  p-value: 0.02194
```

Figure 17: Color vs Price 2 categories

```
Call:
lm(formula = Price ~ Cut, data = professor_cluster)

Residuals:
    Min      1Q  Median      3Q     Max 
-784.76 -250.84   47.48  269.99  578.16 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2559.84      45.81  55.883  < 2e-16 ***
Cut3          190.93      74.53   2.562  0.01104 *  
Cut6          331.69      68.63   4.833 2.43e-06 ***
Cut4          442.68      80.31   5.512 9.32e-08 ***
Cut5          179.67      60.04   2.993  0.00306 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 342.8 on 235 degrees of freedom
Multiple R-squared:  0.144,	Adjusted R-squared:  0.1294 
F-statistic: 9.881 on 4 and 235 DF,  p-value: 2.094e-07
```

Figure 18: Price vs Cut

```
Call:
lm(formula = Price ~ Polish, data = professor_cluster)

Residuals:
    Min      1Q  Median      3Q     Max 
-953.29 -243.89   48.21  264.33  501.21 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2318.6      152.9  15.167  < 2e-16 ***
Polish3        325.2      156.2   2.081 0.038493 *  
Polish6        728.8      216.2   3.371 0.000875 ***
Polish4        524.3      156.8   3.345 0.000959 ***
Polish5        683.7      170.1   4.019 7.87e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 341.8 on 235 degrees of freedom
Multiple R-squared:  0.1487,	Adjusted R-squared:  0.1342 
F-statistic: 10.26 on 4 and 235 DF,  p-value: 1.122e-07
```

Figure 19: Price vs Polish all categories

```
Call:
lm(formula = Price ~ Polish, data = professor_cluster)

Residuals:
    Min      1Q  Median      3Q     Max
-961.96 -244.90   54.54  252.33  515.10

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2629.90      31.76  82.799  < 2e-16 ***
Polish3       381.06      74.49   5.116 6.45e-07 ***
Polish2       213.02      47.18   4.515 9.97e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 343.6 on 237 degrees of freedom
Multiple R-squared:  0.1328,    Adjusted R-squared:  0.1255
F-statistic: 18.14 on 2 and 237 DF,  p-value: 4.661e-08
```

Figure 20: Price vs Polish 3 categories

```
Call:
lm(formula = Price ~ Symmetry, data = professor_cluster)

Residuals:
    Min      1Q  Median      3Q     Max
-927.68 -249.08   79.35  289.32  692.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2432.29      74.93  32.459  < 2e-16 ***
Symmetry3     260.79      82.15   3.174 0.001702 **
Symmetry6     615.11     170.88   3.600 0.000388 ***
Symmetry4     413.39      83.78   4.934 1.52e-06 ***
Symmetry5     502.87     100.75   4.991 1.17e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 343.4 on 235 degrees of freedom
Multiple R-squared:  0.1409,    Adjusted R-squared:  0.1263
F-statistic: 9.639 on 4 and 235 DF,  p-value: 3.104e-07
```

Figure 21: Price vs Symmetry all categories

```
Call:
lm(formula = Price ~ Polish, data = professor_cluster)

Residuals:
    Min      1Q  Median      3Q     Max 
-961.96 -244.90   54.54  252.33  515.10 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2629.90      31.76  82.799  < 2e-16 ***
Polish3       381.06      74.49   5.116 6.45e-07 ***
Polish2       213.02      47.18   4.515 9.97e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 343.6 on 237 degrees of freedom
Multiple R-squared:  0.1328,	Adjusted R-squared:  0.1255 
F-statistic: 18.14 on 2 and 237 DF,  p-value: 4.661e-08
```

Figure 22: Price vs Symmetry 3 categories

```
Call:
lm(formula = Price ~ Certification, data = professor_raw[professor_raw$Carat > 
    0.5, ])

Residuals:
    Min      1Q  Median      3Q     Max 
-902.85 -263.85   34.38  263.40  467.17 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        3033.4      102.3  29.657  < 2e-16 ***
CertificationDOW  -1002.4      368.8  -2.718  0.00705 ** 
CertificationEGL   -355.6      107.3  -3.313  0.00107 ** 
CertificationGIA   -212.6      107.8  -1.972  0.04983 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 354.3 on 236 degrees of freedom
Multiple R-squared:  0.08153,	Adjusted R-squared:  0.06986 
F-statistic: 6.983 on 3 and 236 DF,  p-value: 0.0001607
```

Figure 23: Price vs Certification Initial Categories

```
Call:
lm(formula = Price ~ Certification, data = professor_cluster)

Residuals:
   Min     1Q Median     3Q    Max 
-924.1 -283.4  132.7  242.1  472.6 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2842.11      32.70  86.927   <2e-16 ***
Certification1  -169.67      46.24  -3.669    3e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 358.2 on 238 degrees of freedom
Multiple R-squared:  0.05354,	Adjusted R-squared:  0.04957 
F-statistic: 13.46 on 1 and 238 DF,  p-value: 0.0003001
```

Figure 24: Price vs Certification Most Respected Labs vs Others

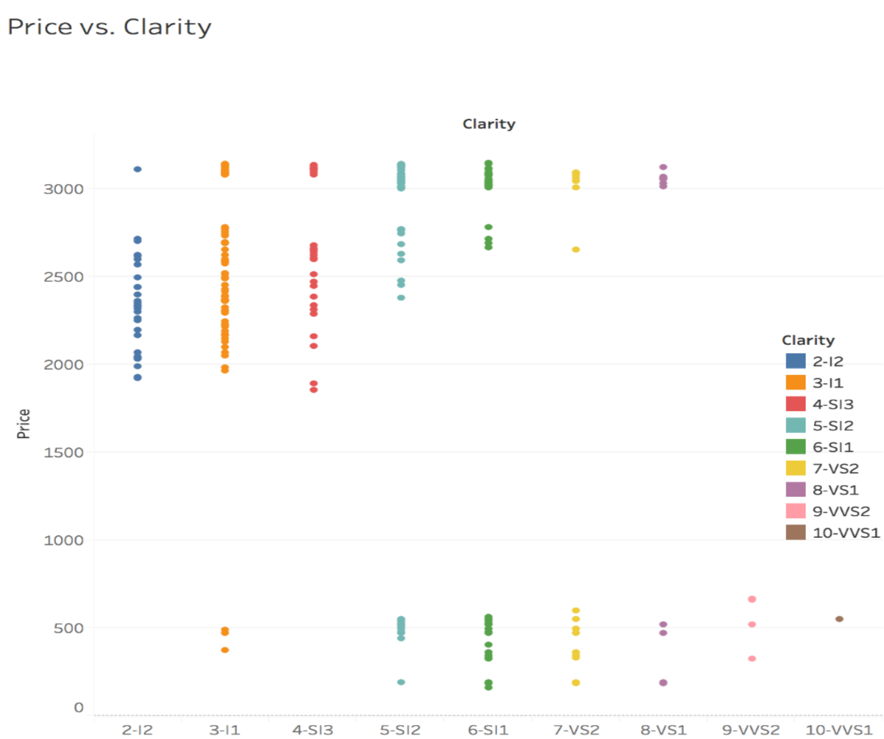

Figure 25: Price vs Clarity Bivariate

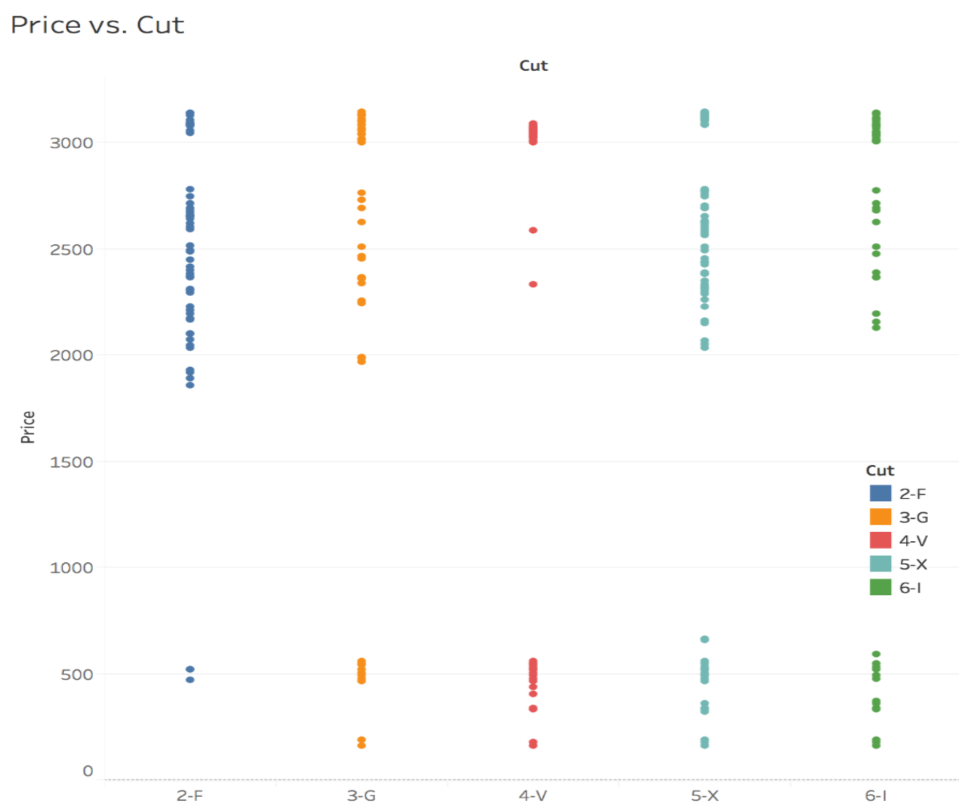

Figure 26: Price vs Cut Bivariate

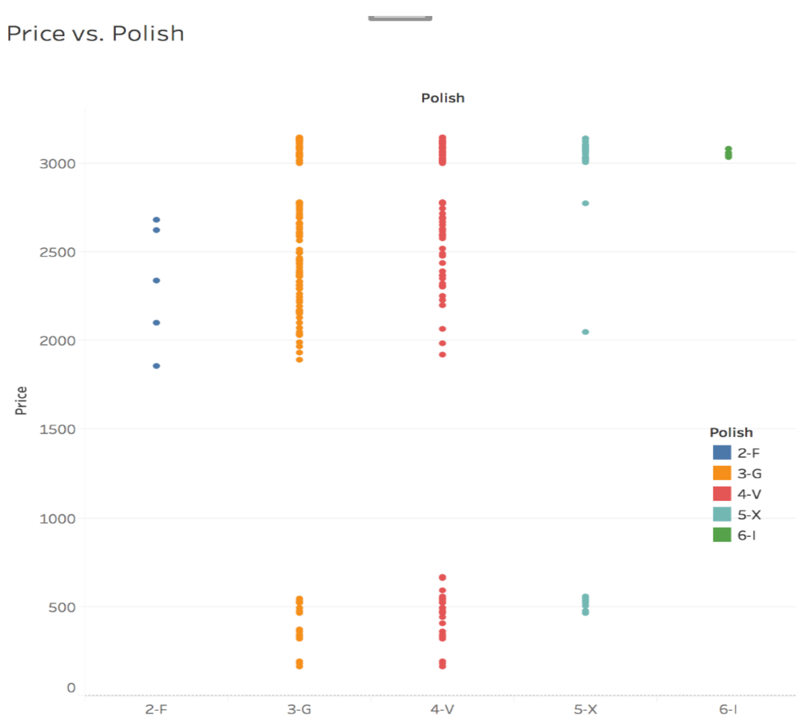

Figure 27: Price vs Polish Bivariate

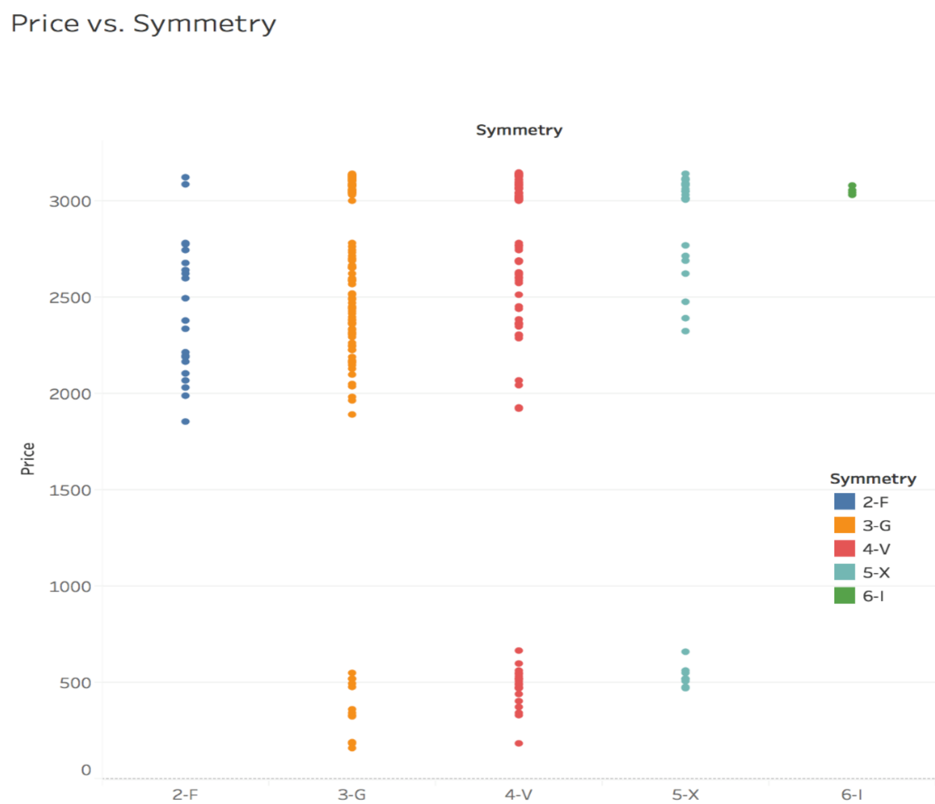

Figure 28: Price vs Symmetry Bivariate

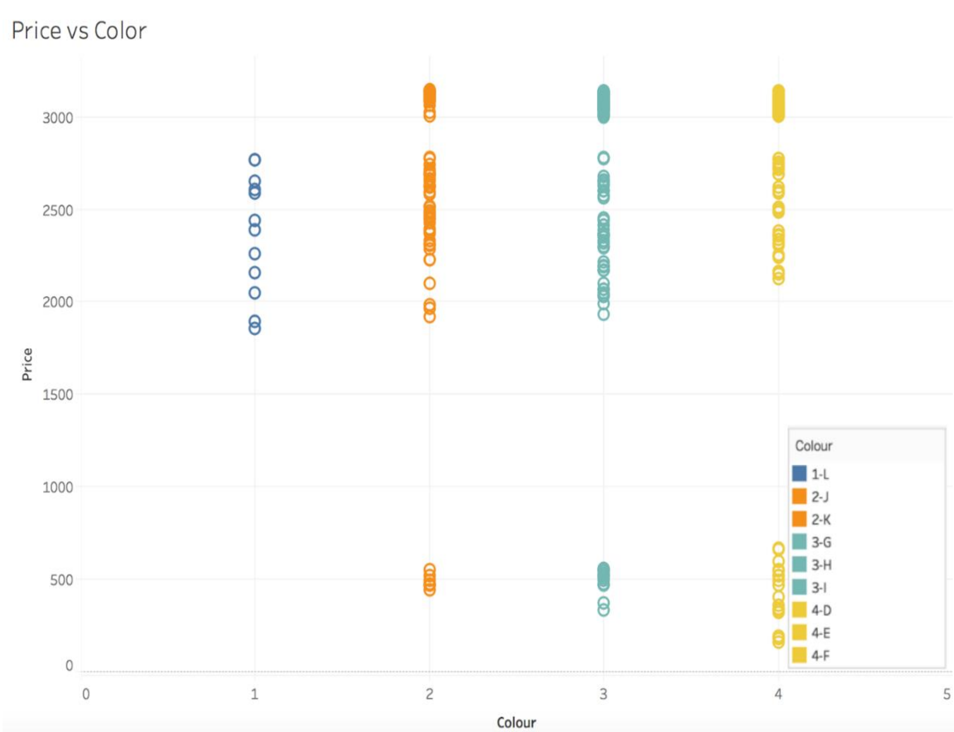

Figure 29: Price vs Color Bivariate