**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: 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.

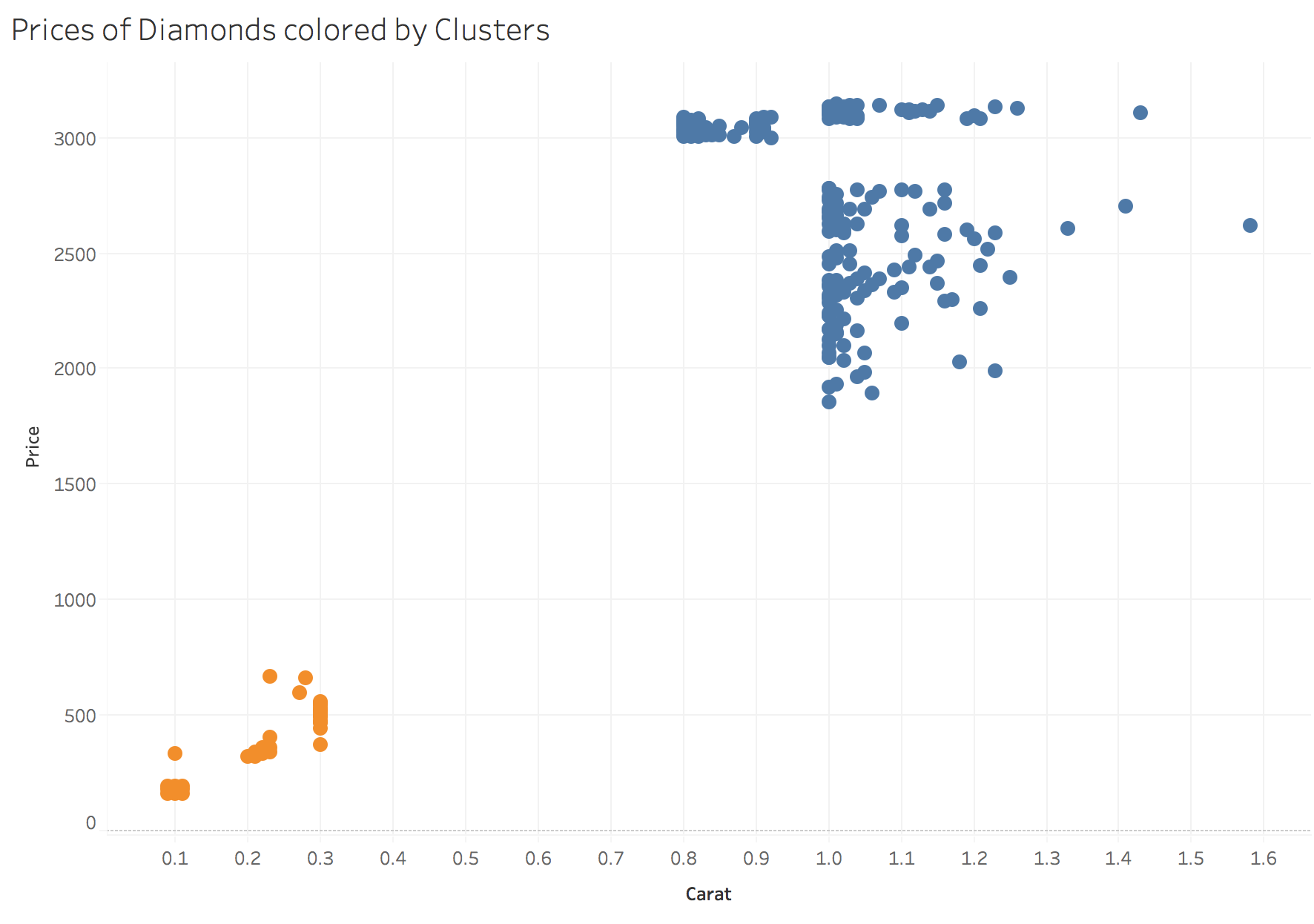


Figure: Carat vs. Price colored by cluster

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.

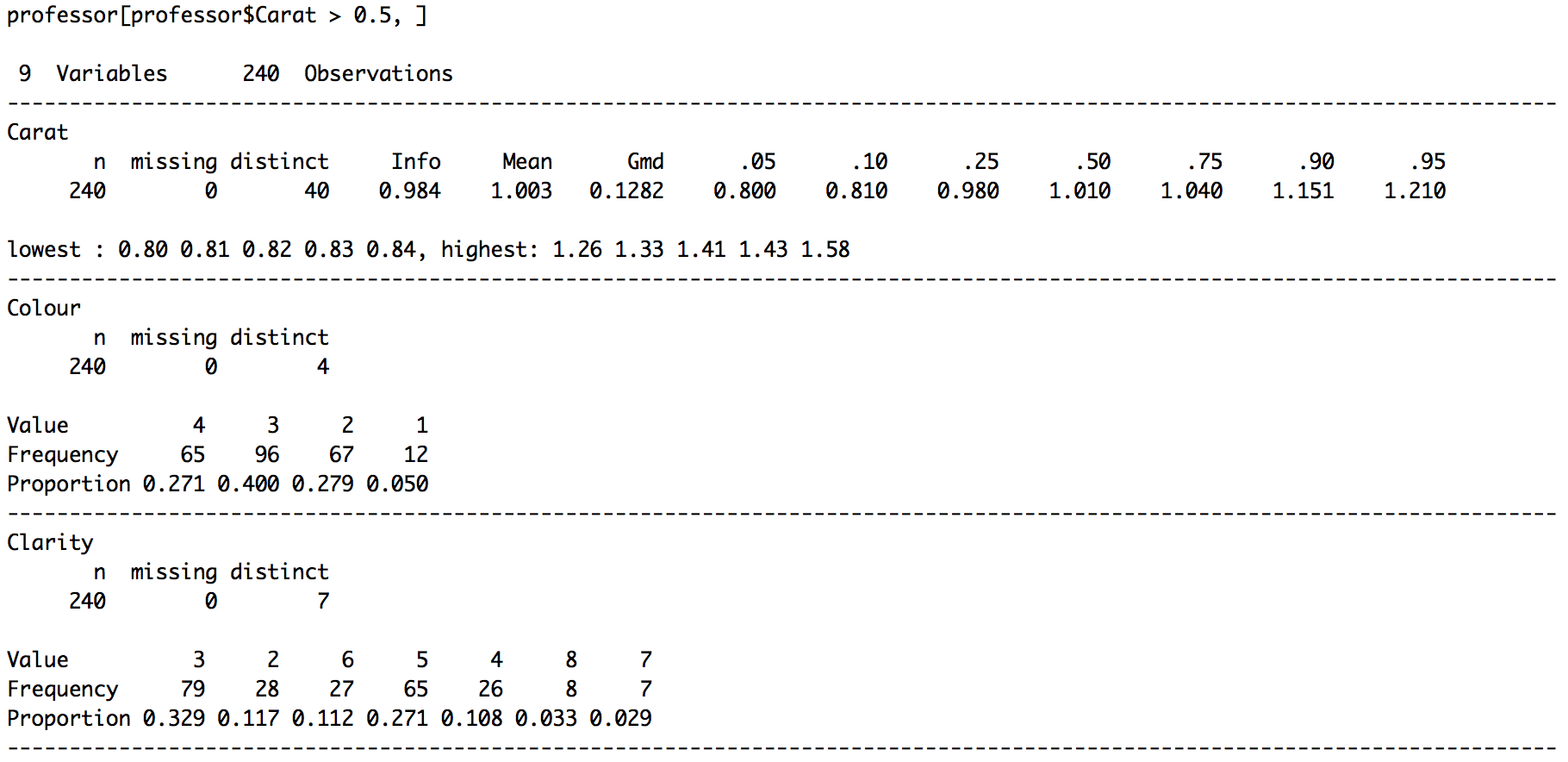
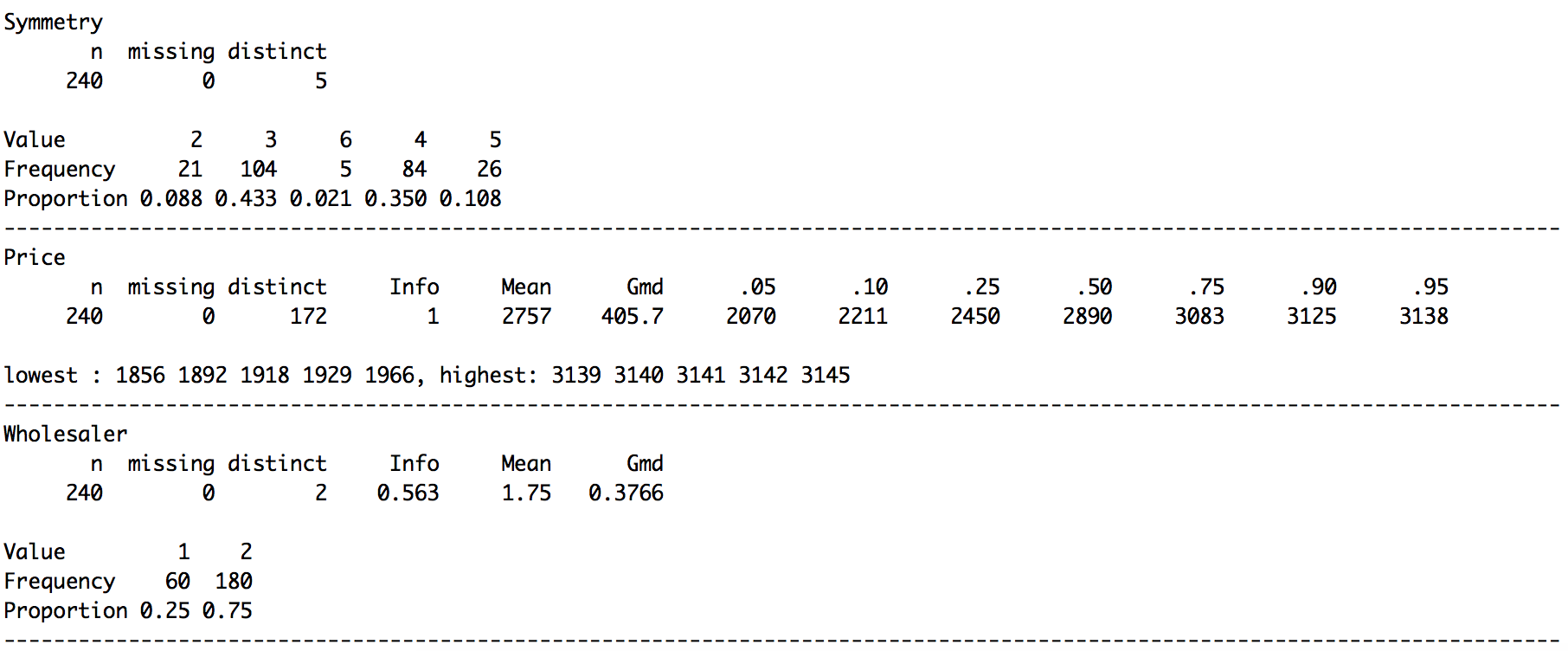
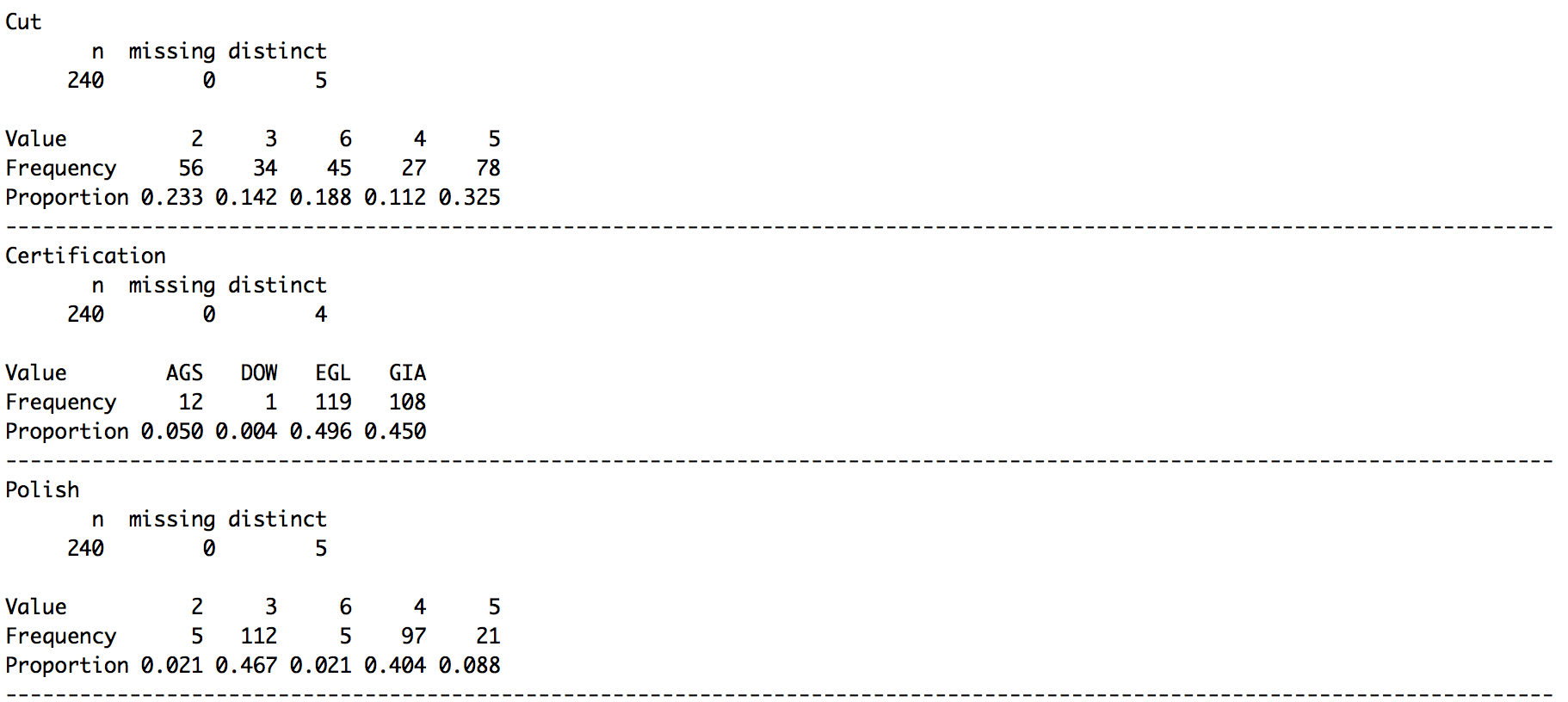
Descriptive statistics for the selected cluster are as follows;****

Figure: 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’s 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: 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.

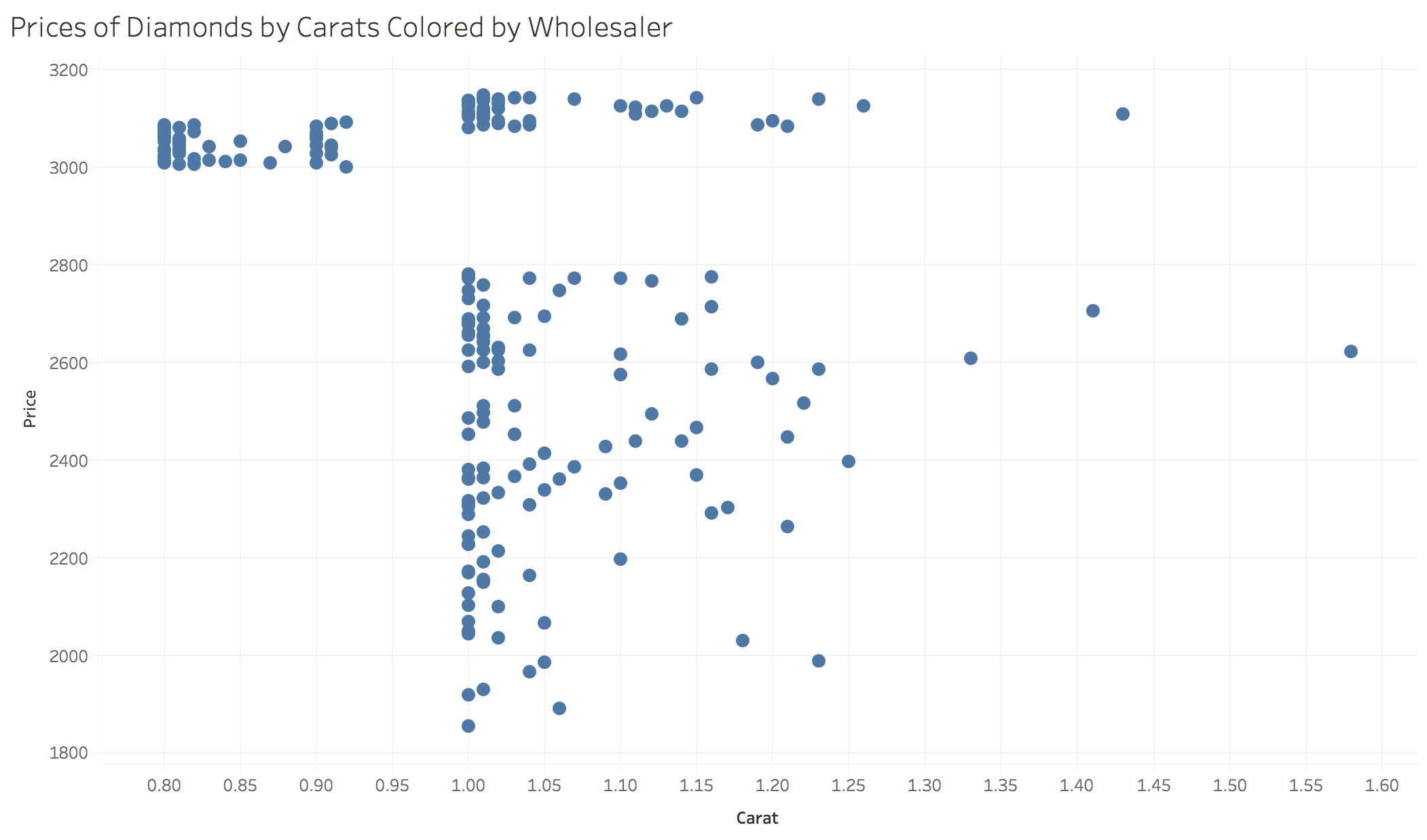
**APPENDIX  
**

Figure: Cluster where professors ring belongs

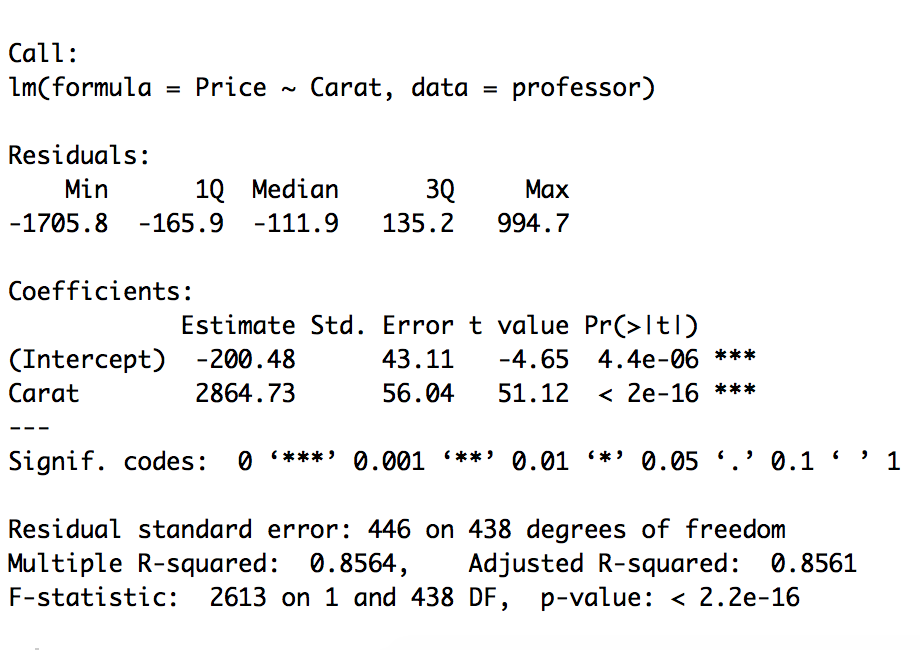
****

Figure: Carat vs Price

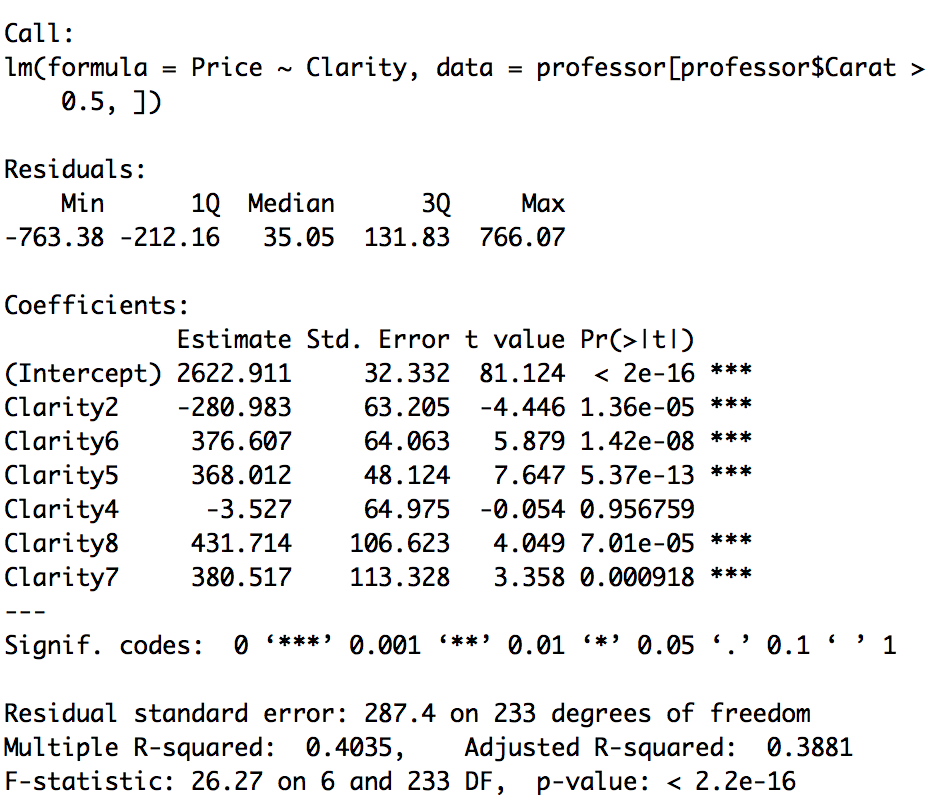
****

Figure: Price vs Clarity all levels

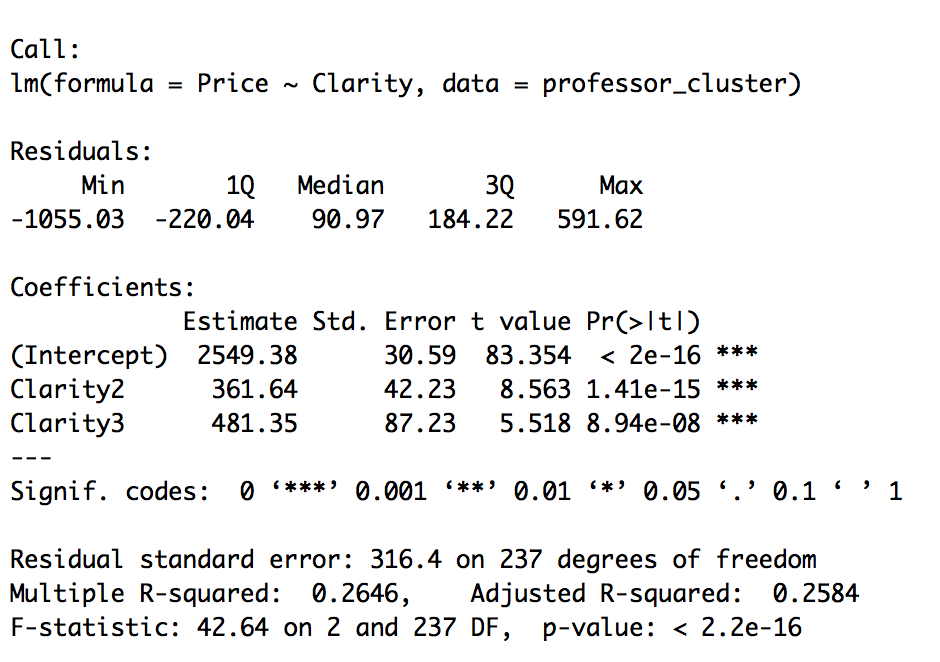
****

Figure: Price vs Clarity 3 levels

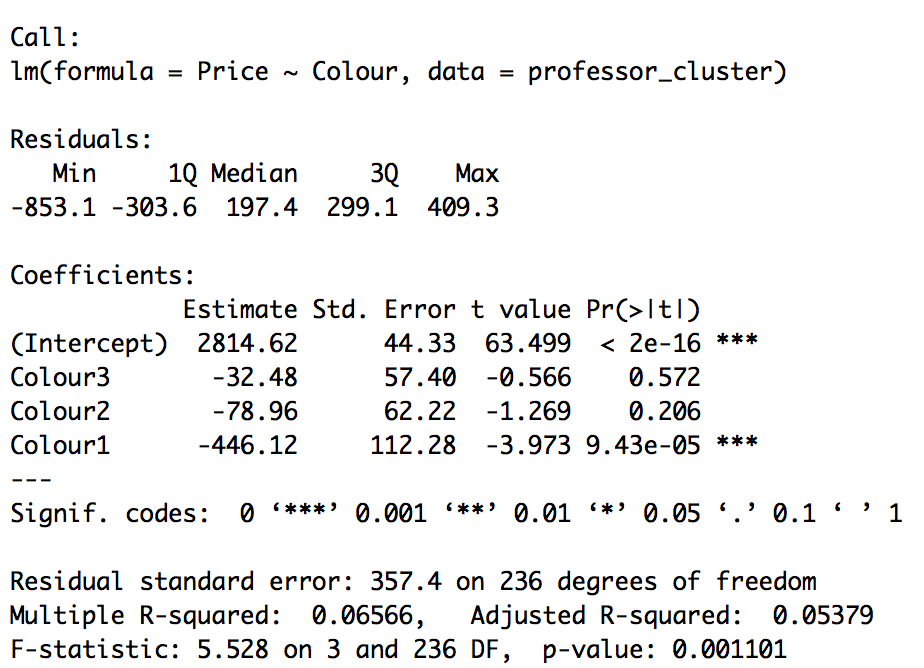
****

Figure: Color vs Price 4 categories

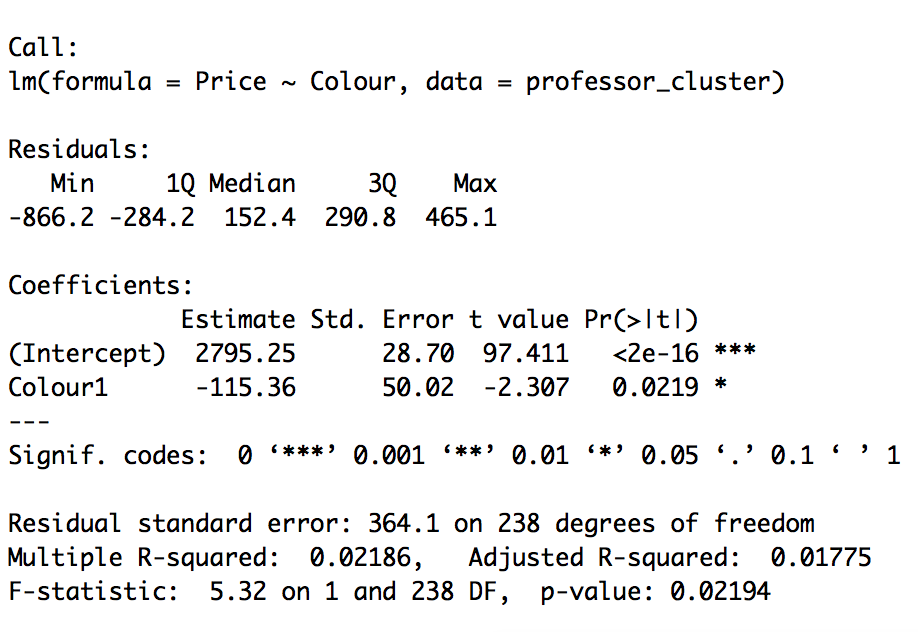


Figure: Color vs Price 2 categories

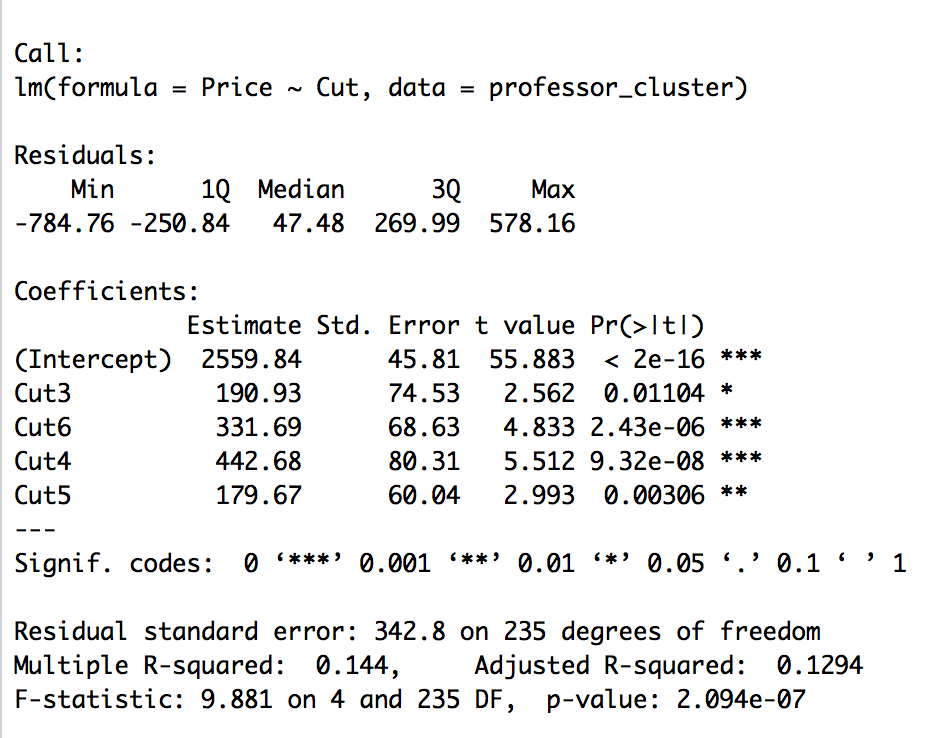
****

Figure: Price vs Cut

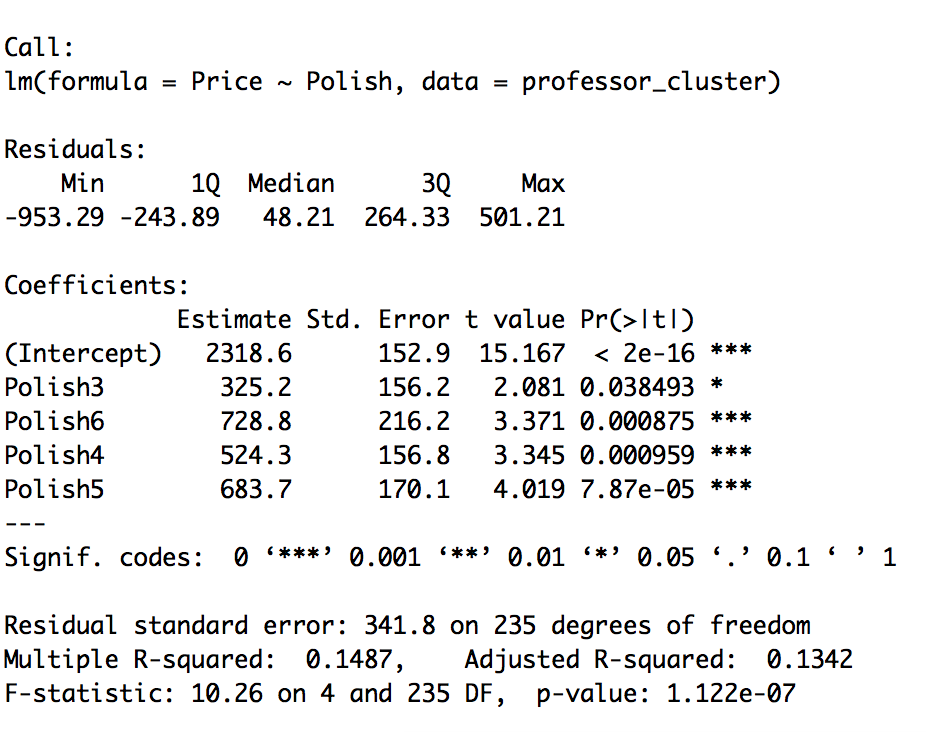
****

Figure: Price vs Polish all categories

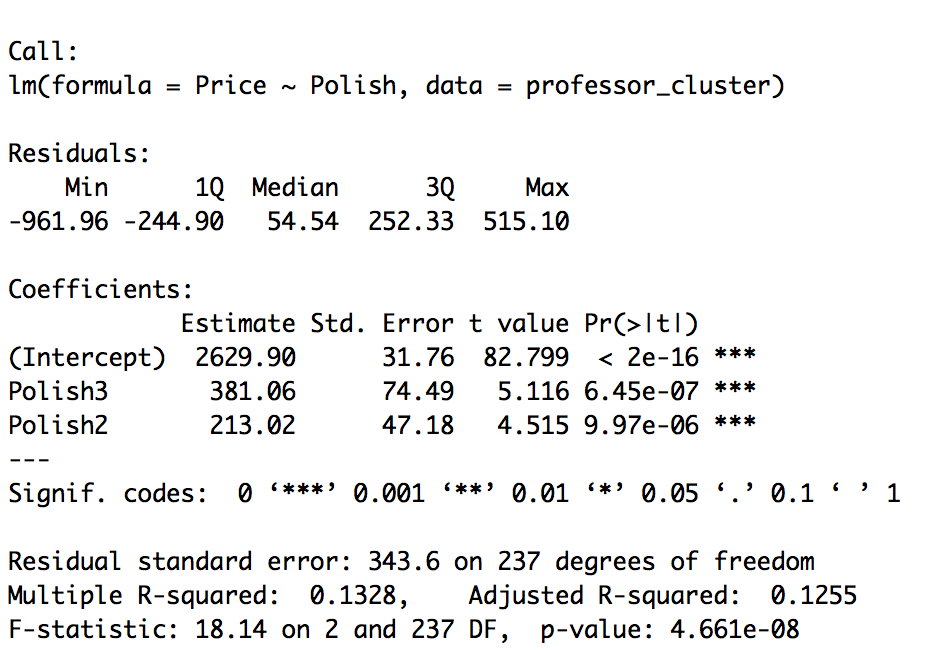


Figure: Price vs Polish 3 categories

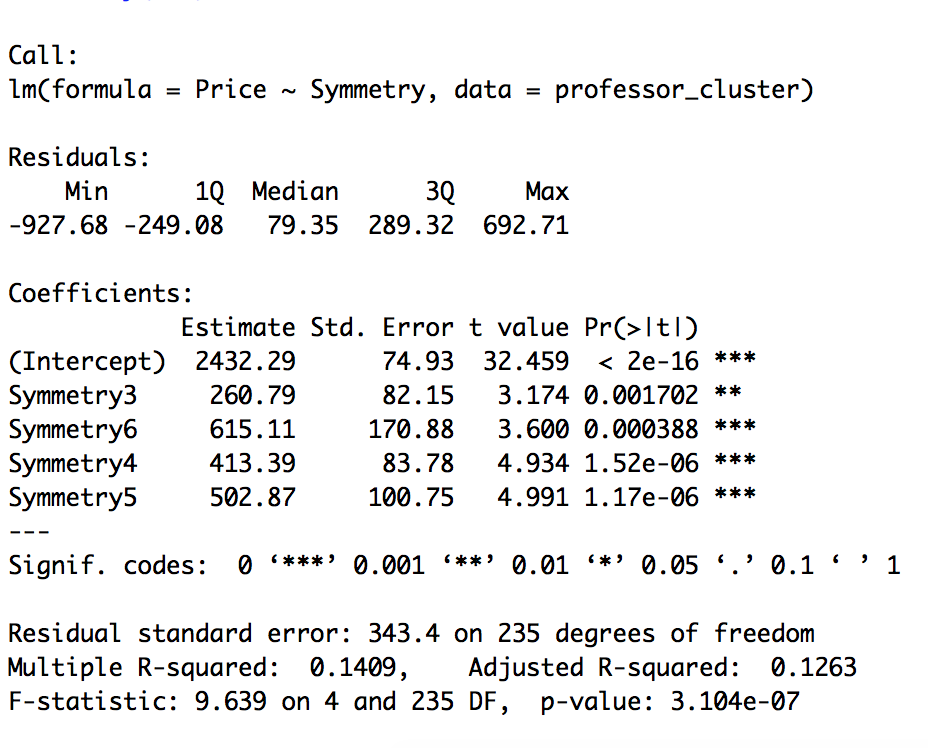


Figure: Price vs Symmetry all categories

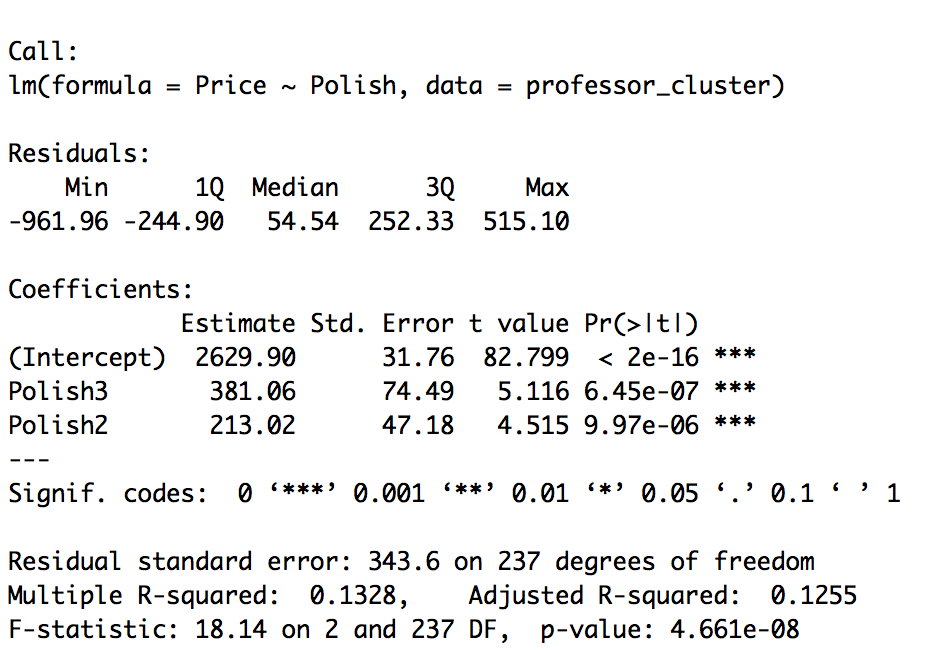
****

Figure: Price vs Symmetry 3 categories

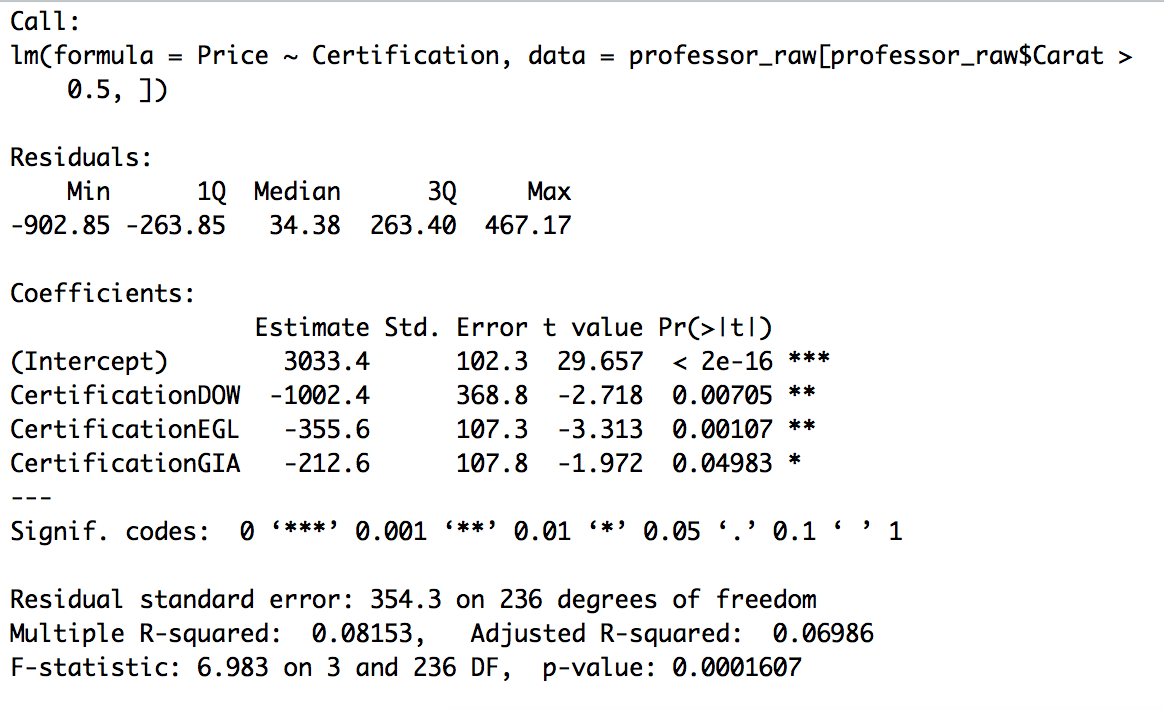
****

Figure: Price vs Certification Initial Categories

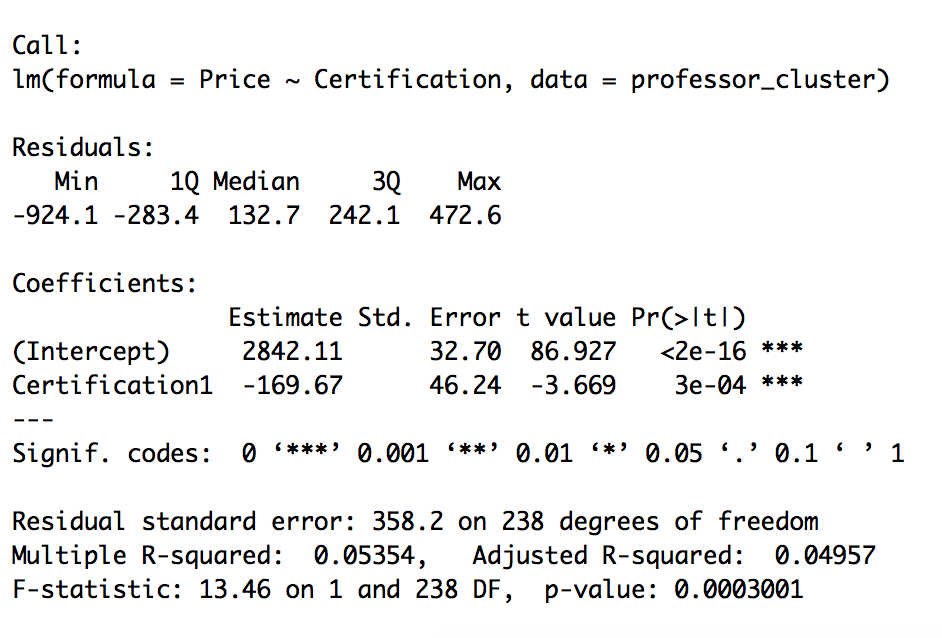


Figure: Price vs Certification Most Respected Labs vs Others