**THE PROFESSOR PROPOSES**

## Table of Contents

## Executive Summary

A Professor is shopping for engagement ring and quickly learns that the **pricing diamonds** is not so easy. He collects data from wholesalers on 440 different diamonds. Now he has information on prices of these diamonds along with their characteristic features such as **color, clarity, cut and carat** (famous 4Cs) besides **polish, symmetry and certification.** These characteristics are considered important features in determining the price, but he was not sure how important are each of these features really are. He has identified a diamond which interests him. He decides to build a regression model using the acquired data to estimate the price of the diamond he is interested in. He wants to ensure that the price he pays for the diamond is fair.

He builds a **multiple regression model** which suggests that the diamond he is interested in is overpriced but he also needs to account for the ring and the cost of retailing.

## Issues and Challenges

Immediate challenge facing the professor is how to price the diamond of his interest given its characteristics. Advanced multiple regression is a potential approach to solve the pricing challenge. It involves the use of regular metric variables like Carat over a specified range and grouping of data using dummy variables for cut, color, clarity and others.

## Exploratory Data Analysis

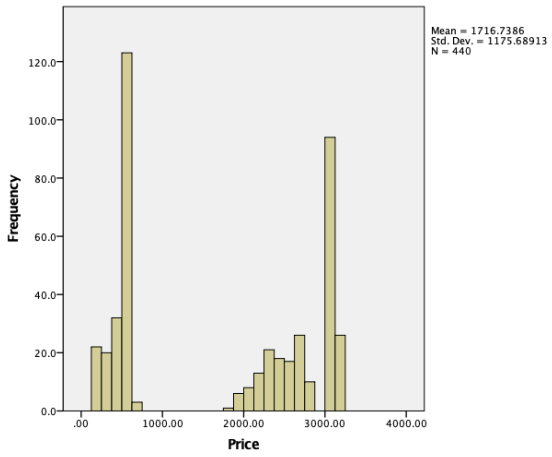
The exploratory data analysis involved univariate analysis and feature engineering.

### Univariate Analysis / Feature Engineering

#### Metric Data

*Price*

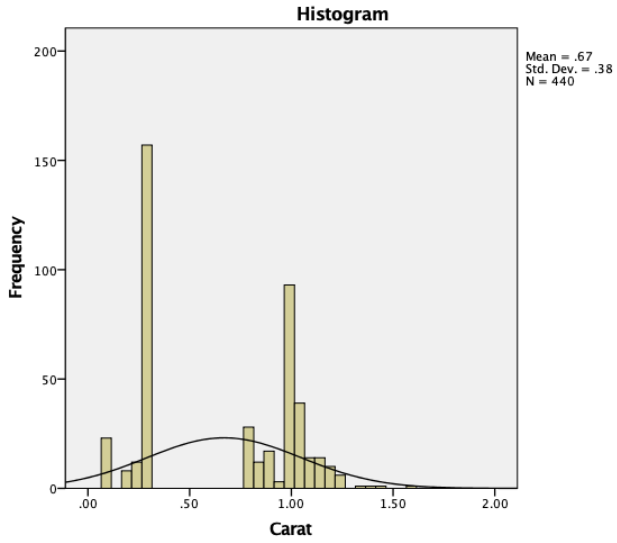
| Price | |
|---|---|
| Valid | 440 |
| Missing | 0 |
| Mean | 1716.74 |
| Median | 2169.0 |
| Std. Deviation | 1175.69 |
| Range | 2985 |
| Minimum | 160 |
| Maximum | 3145 |




The Price ranged from 160 to 3145 and had bimodal distribution. The diamonds appear to be in two clusters less than $1000 and higher than $1000.

*Carat*

| Carat | |
|---|---|
| Valid | 440 |
| Missing | 0 |
| Mean | .67 |
| Median | .81 |
| Std. Deviation | .38 |
| Range | 1.49 |
| Minimum | .09 |
| Maximum | 1.58 |




The Carat ranged from 0.09 to 1.58 and had bimodal distribution. The diamonds appear to be in two clusters 0.09 to 0.30 and 0.81 to 1.58 carats.

*Clarity*

| Clarity | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| I1 | 82 | 18.6 | 18.6 | 18.6 |
| I2 | 28 | 6.4 | 6.4 | 25.0 |
| SI1 | 116 | 26.4 | 26.4 | 51.4 |
| SI2 | 110 | 25.0 | 25.0 | 76.4 |
| SI3 | 26 | 5.9 | 5.9 | 82.3 |
| VS1 | 30 | 6.8 | 6.8 | 89.1 |
| VS2 | 41 | 9.3 | 9.3 | 98.4 |
| VVS1 | 2 | .5 | .5 | 98.9 |
| VVS2 | 5 | 1.1 | 1.1 | 100.0 |
| Total | 440 | 100.0 | 100.0 | |

This ordinally scaled data is potentially a case for being approximated as interval-scaled data. For treatment as nominal data, the VVS scale can be merged with VS since it has very low representation.

| Clarity as non-metric | N | % | Valid % | Cum % |
|---|---|---|---|---|
| Very few inclusions visible | 82 | 18.6 | 18.6 | 18.6 |
| Few inclusions visible | 28 | 6.4 | 6.4 | 25.0 |
| Very Very Few inclusions visible at 10X | 116 | 26.4 | 26.4 | 51.4 |
| Very Few inclusions visible at 10X | 110 | 25.0 | 25.0 | 76.4 |
| Several inclusions visible at 10X | 26 | 5.9 | 5.9 | 82.3 |
| Few inclusions at 30X | 30 | 6.8 | 6.8 | 89.1 |
| Several inclusions at 30X | 41 | 9.3 | 9.3 | 98.4 |
| Very very few inclusions at 30X | 2 | .5 | .5 | 98.9 |
| Very few inclusions at 30X | 5 | 1.1 | 1.1 | 100.0 |
| Total | 440 | 100.0 | 100.0 | |

| Clarity as Interval scaled | Count | % | Valid % | Cum % |
|---|---|---|---|---|
| 1 | 82 | 18.6 | 18.6 | 18.6 |
| 2 | 28 | 6.4 | 6.4 | 25.0 |
| 3 | 116 | 26.4 | 26.4 | 51.4 |
| 4 | 110 | 25.0 | 25.0 | 76.4 |
| 5 | 26 | 5.9 | 5.9 | 82.3 |
| 6 | 30 | 6.8 | 6.8 | 89.1 |
| 7 | 41 | 9.3 | 9.3 | 98.4 |
| 8 | 2 | .5 | .5 | 98.9 |
| 9 | 5 | 1.1 | 1.1 | 100.0 |
| Total | 440 | 100.0 | 100.0 | |

Non Metric Data

*Color*

| Colour | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| D | 20 | 4.5 | 4.5 | 4.5 |
| E | 54 | 12.3 | 12.3 | 16.8 |
| F | 58 | 13.2 | 13.2 | 30.0 |
| G | 43 | 9.8 | 9.8 | 39.8 |
| H | 71 | 16.1 | 16.1 | 55.9 |
| I | 79 | 18.0 | 18.0 | 73.9 |
| J | 72 | 16.4 | 16.4 | 90.2 |
| K | 31 | 7.0 | 7.0 | 97.3 |
| L | 12 | 2.7 | 2.7 | 100.0 |
| Total | 440 | 100.0 | 100.0 | |

Using the data available we will classify them into four color classes.

| ColourN | Count | % | Valid % | Cum % |
|---|---|---|---|---|
| Colorless | 132 | 30.0 | 30.0 | 30.0 |
| Near Colorless | 193 | 43.9 | 43.9 | 73.9 |
| Faint Yellow | 103 | 23.4 | 23.4 | 97.3 |
| Very Light Yellow | 12 | 2.7 | 2.7 | 100.0 |
| Total | 440 | 100.0 | 100.0 | |

We can consider using Colorless and Yellow as another potential clubbing of the data.

| Colour2C | Count | % | Valid % | Cum % |
|---|---|---|---|---|
| Colorless | 325 | 73.9 | 73.9 | 73.9 |
| Yellow | 115 | 26.1 | 26.1 | 100.0 |
| Total | 440 | 100.0 | 100.0 | |

*Cut*

| Cut | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| F | 59 | 13.4 | 13.4 | 13.4 |
| G | 49 | 11.1 | 11.1 | 24.5 |
| I | 86 | 19.5 | 19.5 | 44.1 |
| V | 97 | 22.0 | 22.0 | 66.1 |
| X | 149 | 33.9 | 33.9 | 100.0 |
| Total | 440 | 100.0 | 100.0 | |

Cut has reasonably balanced representation of each class.

| CutN | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| Fair | 59 | 13.6 | 13.6 | 13.4 |
| Good | 49 | 11.1 | 11.1 | 24.5 |
| Very Good | 97 | 22.0 | 22.0 | 46.6 |
| Excellent | 149 | 33.9 | 33.9 | 80.5 |
| Ideal | 86 | 19.5 | 19.5 | 100.0 |
| Total | 440 | 100.0 | 100.0 | |

*Certification*

| Certification | Frequency | Percent | Valid % | Cum % |
|---|---|---|---|---|
| AGS | 12 | 2.9 | 2.9 | 2.9 |
| DOW | 1 | .2 | .2 | 3.2 |
| EGL | 119 | 27.0 | 27.0 | 30.2 |
| GIA | 265 | 60.1 | 60.1 | 90.2 |
| IGI | 43 | 9.8 | 9.8 | 100.0 |
| Total | 441 | 100.0 | 100.0 | |

We can club certification in two classes as established and small labs

| CertG | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| AGS - GIA | 277 | 63.0 | 63.0 | 63.0 |
| Rest | 163 | 37.0 | 37.0 | 100.0 |
| Total | 440 | 100.0 | 100.0 | |

*Polish*

| Polish | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| F | 5 | 1.1 | 1.1 | 1.1 |
| G | 165 | 37.5 | 37.5 | 38.6 |
| I | 5 | 1.1 | 1.1 | 39.8 |
| V | 204 | 46.4 | 46.4 | 86.1 |
| X | 61 | 13.9 | 13.9 | 100.0 |
| Total | 440 | 100.0 | 100.0 | |

Fair and Ideal has very less representation in the data.

| PolishN | Frequency | Percent | Valid % | Cum % |
|---|---|---|---|---|
| Fair | 5 | 1.1 | 1.1 | 1.1 |
| Good | 165 | 37.6 | 37.6 | 38.7 |
| Very Good | 203 | 46.2 | 46.2 | 85.0 |
| Excellent | 61 | 13.9 | 13.9 | 98.9 |
| Ideal | 5 | 1.1 | 1.1 | 100.0 |
| Total | 440 | 100.0 | 100.0 | |

We will club Polish in two categories viz. Fair - Good and Very Good – Ideal.

| Polish2C | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| Fair - Good | 170 | 38.6 | 38.6 | 38.6 |
| Very Good - Ideal | 270 | 61.4 | 61.4 | 100.0 |
| Total | 440 | 100.0 | 100.0 | |

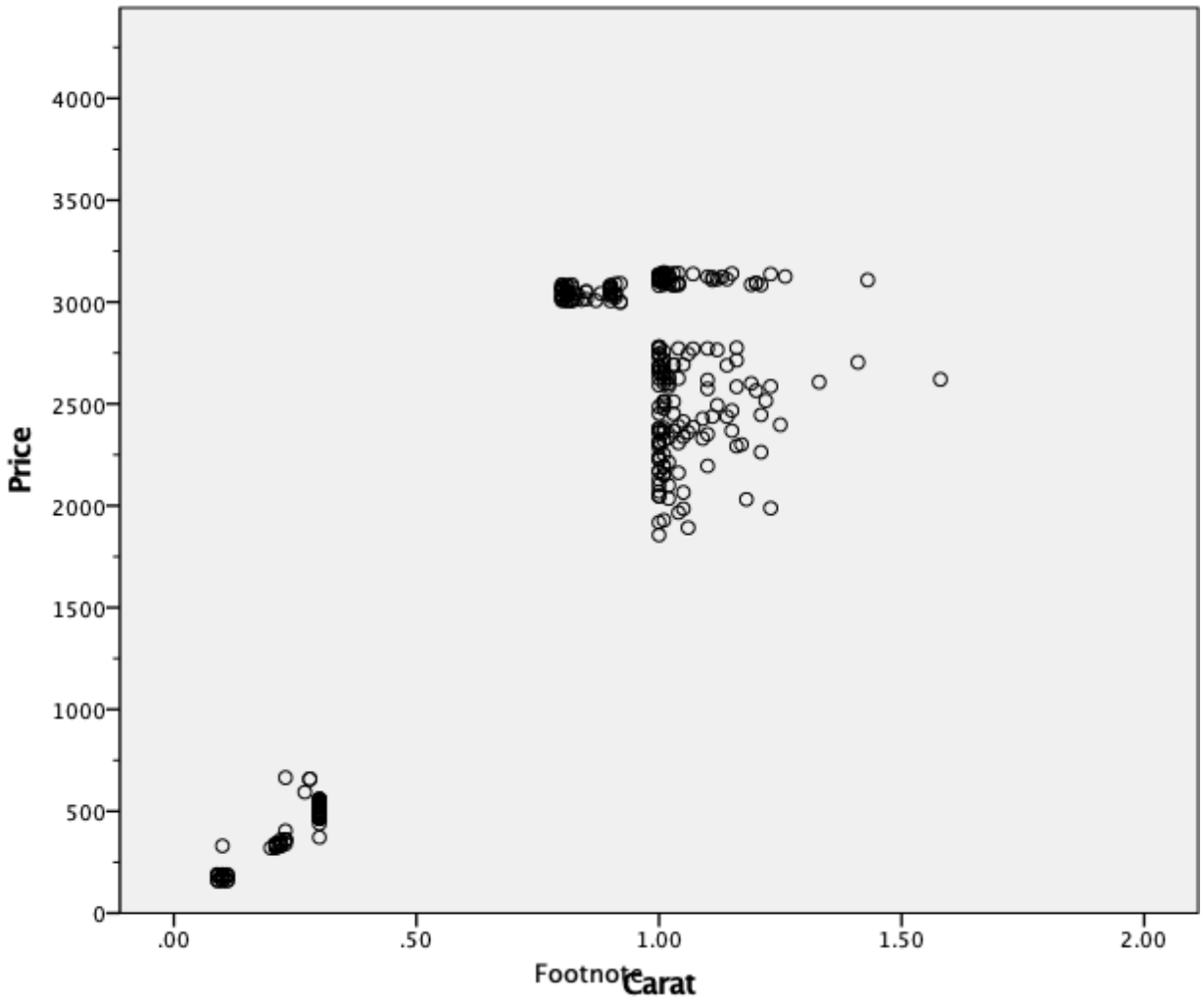
*Symmetry*

**Symmetry**

| | | Frequency | % | Valid % | Cumulative % |
|---|---|---|---|---|---|
| | F | 21 | 5.0 | 5.0 | 5.0 |
| | G | 157 | 35.6 | 35.6 | 40.6 |
| | I | 5 | 1.1 | 1.1 | 41.7 |
| | V | 206 | 46.7 | 46.7 | 88.4 |
| | X | 51 | 11.6 | 11.6 | 100.0 |
| | Total | 441 | 100.0 | 100.0 | |

| Symmetry | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| Fair | 21 | 4.8 | 4.8 | 4.8 |
| Good | 157 | 35.7 | 35.7 | 40.5 |
| Very Good | 206 | 46.8 | 46.8 | 87.3 |
| Excellent | 51 | 11.6 | 11.6 | 98.9 |
| Ideal | 5 | 1.1 | 1.1 | 100.0 |
| Total | 440 | 100.0 | 100.0 | |

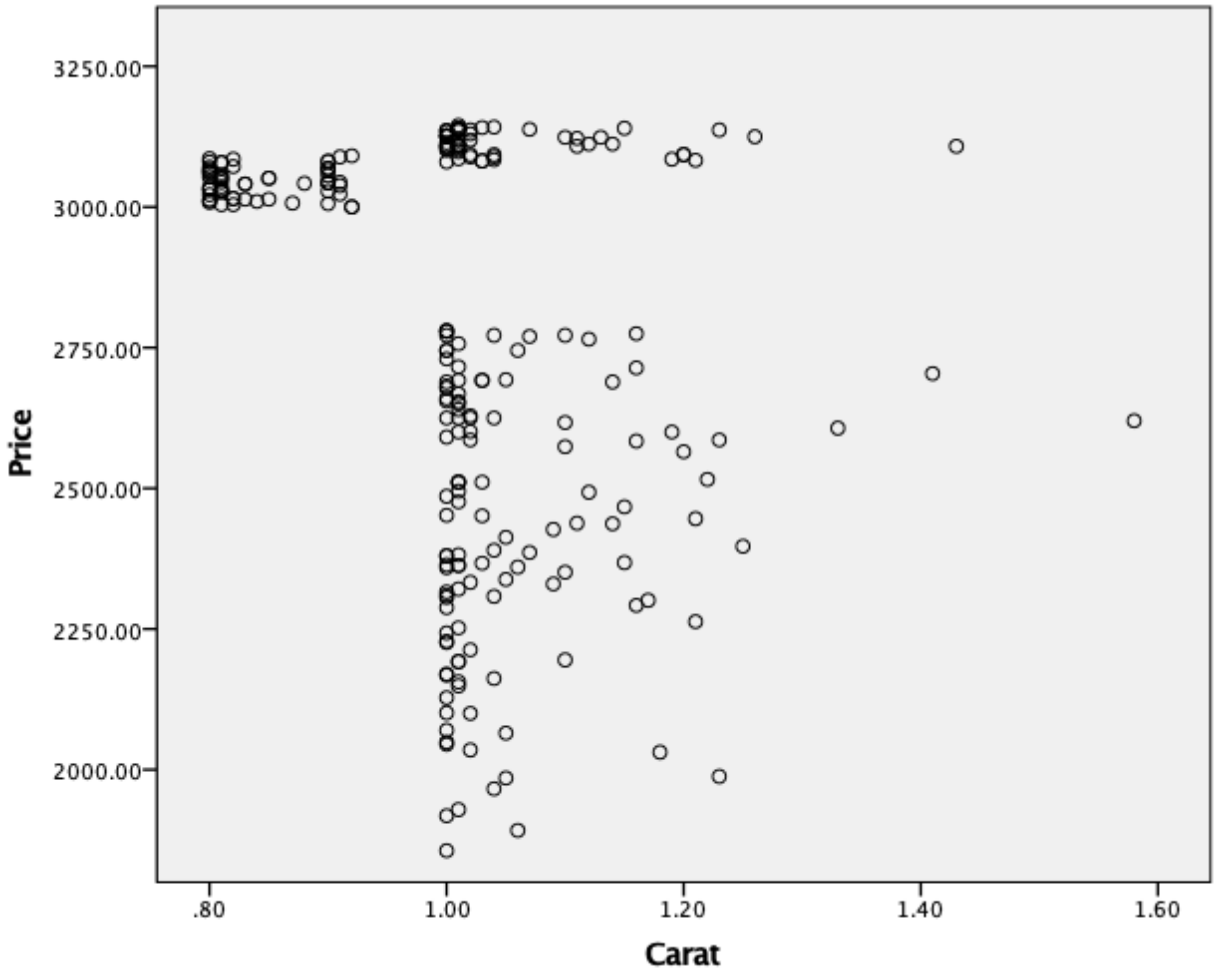
## Bivariate Analysis - Metric




We observe that there are two distinct clusters with low carat (0.09 – 0.30) diamonds having low prices and higher carat diamonds having higher prices.

We will build two models for these two clusters. Besides we will focus on the cluster containing Professors diamond.

High Carat High Price Data Set



Regression with Carat

| Model | R | $R^2$ | Adjusted $R^2$ | SE of the Estimate | Durbin-Watson |
|---|---|---|---|---|---|
| 1 | .327[a] | .107 | .103 | 347.86786 | .534 |

a. Predictors: (Constant), Carat

b. Dependent Variable: Price

| ANOVA | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 3456697.034 | 1 | 3456697.034 | 28.565 | .000[b] |
| Residual | 28800866.820 | 238 | 121012.045 | | |
| Total | 32257563.850 | 239 | | | |

| Coefficient | B | Std. Error | Beta | t | Sig. |
|---|---|---|---|---|---|
| (Constant) | 3740.984 | 185.421 | | 20.176 | .000 |
| Carat | -980.604 | 183.475 | -.327 | -5.345 | .000 |

$R^2$ of 10.7% indicates somewhat correlation between carats and price. The negative regression coefficient is surprising. Somehow heavier diamonds are not valued as much. We will need to consider other factors which can possibly explain lower prices for higher carat diamonds. For other variables we will use dummy variables.

Regression with Clarity

Case 1: Several permutations give at best a regression coefficient of 10%.

Case 2: Rating scale I3=1 to FL=12 gives a regression coefficient of 0.333.

| Model | R | $R^2$ | Adj $R^2$ | SE of the Estimate | Durbin-Watson |
|---|---|---|---|---|---|
| 1 | .577[a] | .333 | .330 | 300.64525 | 1.249 |

a. Predictors: (Constant), ClarityN
b. Dependent Variable: Price

| ANOVA | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 10745322.800 | 1 | 10745322.800 | 118.881 | .000[b] |
| Residual | 21512241.050 | 238 | 90387.567 | | |
| Total | 32257563.850 | 239 | | | |

| Model | B | Std. Error | Beta | t | Sig. |
|---|---|---|---|---|---|
| (Constant) | 2176.256 | 56.712 | | 38.374 | .000 |
| ClarityN | 139.864 | 12.828 | .577 | 10.903 | .000 |

# Bivariate Analysis – Non Metric

## Regression with Certification

We consider certification by GIA and AGS as one category and rest as other. It can be noticed that better $R^2$ can be achieved by reducing number of attributes of categorical variables.

| Model | R | $R^2$ | Adj $R^2$ | SE of the Estimate | Durbin-Watson |
|---|---|---|---|---|---|
| 1 | .231[a] | .054 | .050 | 358.16032 | 1.007 |

a. Predictors: (Constant), Cert2G
b. Dependent Variable: Price

| ANOVA | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 1727206.667 | 1 | 1727206.667 | 13.464 | .000[b] |
| Residual | 30530357.180 | 238 | 128278.812 | | |
| Total | 32257563.850 | 239 | | | |

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | 2672.442 | 32.695 | | 81.738 | .000 |
| | Cert2G | 169.667 | 46.238 | .231 | 3.669 | .000 |

a. Dependent Variable: Price

Though 5% R2 is not high but it is significant.

Regression using Cut

Regression with Cut produces 9% R square. It makes it an important feature.

**Model Summary[b]**

| Model | R | $R^2$ | Adj $R^2$ | SE of the Estimate | Durbin-Watson |
|---|---|---|---|---|---|
| 1 | .297[a] | .088 | .084 | 351.52887 | .958 |

a. Predictors: (Constant), Cut2C

b. Dependent Variable: Price

| ANOVA | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 2847297.693 | 1 | 2847297.693 | 23.042 | .000[b] |
| Residual | 29410266.160 | 238 | 123572.547 | | |
| Total | 32257563.850 | 239 | | | |

| Model | B | Std. Error | Beta | t | Sig. |
|---|---|---|---|---|---|
| (Constant) | 2559.839 | 46.975 | | 54.494 | .000 |
| Cut (Very Good – Ideal) | 257.525 | 53.649 | .297 | 4.800 | .000 |

Regression using Polish

Regression with Polish produces 11% R square. It makes it an important feature.

**Model Summary[b]**

| Model | R | $R^2$ | Adj $R^2$ | SE of the Estimate | Durbin-Watson |
|---|---|---|---|---|---|
| 1 | .339[a] | .115 | .111 | 346.37077 | .945 |

a. Predictors: (Constant), Polish2C

b. Dependent Variable: Price

| ANOVA | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 3704058.788 | 1 | 3704058.788 | 30.874 | .000[b] |
| Residual | 28553505.060 | 238 | 119972.710 | | |
| Total | 32257563.850 | 239 | | | |

| Model | B | Std. Error | Beta | t | Sig. |
|---|---|---|---|---|---|
| (Constant) | 2629.897 | 32.022 | | 82.128 | .000 |
| Polish (Very Good – Ideal) | 248.542 | 44.730 | .339 | 5.556 | .000 |

## Regression using Symmetry

Regression with Symmetry produces 7.5% R square. The p value is significant, but the t value and F stats are relatively low compared to other.

| Model | R | $R^2$ | Adj $R^2$ | SE of the Estimate | Durbin-Watson |
|---|---|---|---|---|---|
| 1 | .275[a] | .075 | .071 | 354.01005 | .908 |

a. Predictors: (Constant), Symm2C

b. Dependent Variable: Price

| ANOVA | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 2430661.646 | 1 | 2430661.646 | 19.395 | .000[b] |
| Residual | 29826902.200 | 238 | 125323.119 | | |
| Total | 32257563.850 | 239 | | | |

| Model | B | Std. Error | Beta | t | Sig. |
|---|---|---|---|---|---|
| (Constant) | 2432.286 | 77.251 | | 31.485 | .000 |
| Symmetry (Very Good – Ideal) | 356.153 | 80.870 | .275 | 4.404 | .000 |

Regression using Color

Regression with Color produces 2.2% R square. It is significant but exhibit weak relation. We also know that color is part of 4 Cs so we will investigate it further.

**Model Summary[b]**

| Model | R | $R^2$ | Adj $R^2$ | SE of the Estimate | Durbin-Watson |
|---|---|---|---|---|---|
| 1 | .148[a] | .022 | .018 | 364.10515 | .816 |

a. Predictors: (Constant), Color2C

b. Dependent Variable: Price

| ANOVA | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 705293.813 | 1 | 705293.813 | 5.320 | .022[b] |
| Residual | 31552270.040 | 238 | 132572.563 | | |
| Total | 32257563.850 | 239 | | | |

| Model | B | Std. Error | Beta | t | Sig. |
|---|---|---|---|---|---|
| (Constant) | 2910.611 | 70.511 | | 41.279 | .000 |
| Cut (Very Good – Ideal) | -115.362 | 50.016 | -.148 | -2.307 | .022 |

## Multiple Regression – High Priced Diamond

After several permutation combinations, we found a stable model which included all 4 Cs and gave a Rsquare of 0.497 with all variables significant. All 4Cs had the lowest p-values compared to Polish and Certification.

**Model Summary[b]**

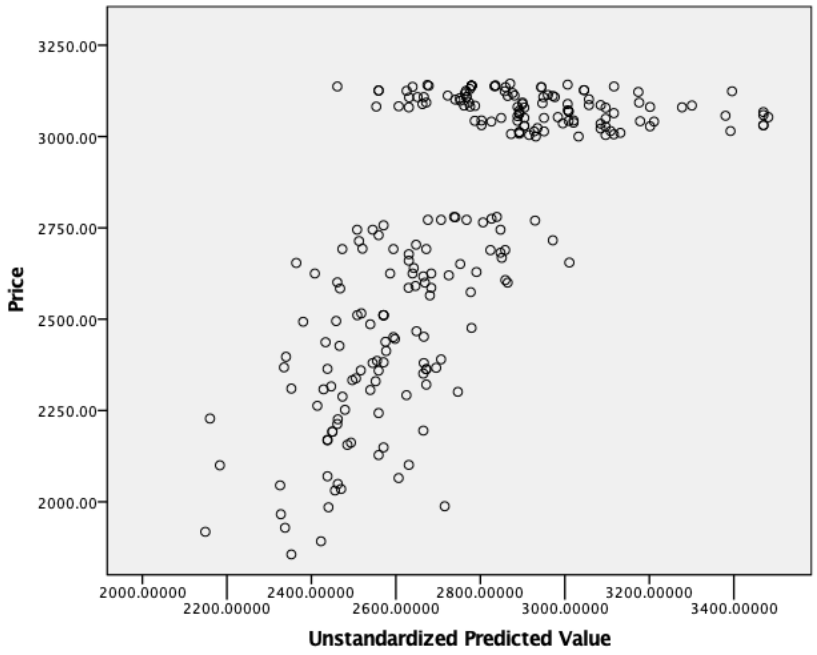
| Model | R | $R^2$ | Adj $R^2$ | SE of the Estimate | Durbin-Watson |
|---|---|---|---|---|---|
| 1 | .705[a] | .497 | .484 | 263.82079 | 1.097 |

a. Predictors: (Constant), Cut2C, Color2C, ClarityN, Cert2G, Polish2C, Carat

b. Dependent Variable: Price

| ANOVA | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 16040435.600 | 6 | 2673405.933 | 38.410 | .000[b] |
| Residual | 16217128.250 | 233 | 69601.409 | | |
| Total | 32257563.850 | 239 | | | |

| Model | B | Std. Error | Beta | t | Sig. |
|---|---|---|---|---|---|
| (Constant) | 414.207 | 294.317 | | 1.407 | .161 |
| Carat | 1168.089 | 218.618 | .390 | 5.343 | .000 |
| Color2C | 278.097 | 44.033 | .356 | 6.316 | .000 |
| ClarityIS | 192.558 | 16.863 | .795 | 11.419 | .000 |
| Cut2CF | 120.963 | 42.924 | .140 | 2.818 | .005 |
| Polish2C | 101.215 | 37.450 | .138 | 2.703 | .007 |
| Cert2G | 80.389 | 36.606 | .110 | 2.196 | .029 |




Using the parameters of the Professors Diamond We can compute the predicted price:

| Feature | Attributes | Coefficient | Value |
|---|---|---|---|
| Constant | 1 | 414.207 | 414.207 |
| Carat | 0.9 | 1168.089 | 1051.28 |
| Color | 0 | 278.097 | 0.0 |
| Clarity | 5 | 192.558 | 962.79 |
| Cut | 1 | 120.963 | 120.963 |
| Polish | 0 | 101.215 | 0.0 |
| Certification | 1 | 80.389 | 80.389 |
| Total | | | 2629.629 |

The 95% confidence interval value of price is $ 2099.98 to $ 3159.88. Comparatively, the price is about $470 more than predicted price.

## Multiple Regression of Low-priced Diamond

After several permutation combinations, we found a stable model which included only 3 Cs (excluding Cut) and gave a $R^2$ of 0.944 with all variables significant.

**Model Summary[b]**

| Model | R | $R^2$ | Adj $R^2$ | SE of the Estimate | Durbin-Watson |
|---|---|---|---|---|---|
| 1 | .972[a] | .944 | .943 | 28.67556 | 1.489 |

a. Predictors: (Constant), Color2C, ClarityN, Cert2G, Polish2C, Carat

b. Dependent Variable: Price

| ANOVA | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 2714425.380 | 5 | 542885.076 | 660.213 | .000[b] |
| Residual | 159523.815 | 194 | 822.288 | | |
| Total | 2873949.195 | 199 | | | |

| Model | B | Std. Error | Beta | t | Sig. |
|---|---|---|---|---|---|
| (Constant) | -208.674 | 22.886 | | -9.118 | .000 |
| Carat | 1485.660 | 72.934 | .812 | 20.370 | .000 |
| Color2C | 72.924 | 6.516 | .234 | 11.192 | .000 |
| ClarityIS | 25.298 | 2.186 | .241 | 11.571 | .000 |
| Polish2C | 11.512 | 5.262 | .042 | 2.188 | .030 |
| Cert2G | 65.990 | 11.871 | .226 | 5.559 | .000 |

It is surprising to note that the Carat alone had almost 85% effect on determining prices. The Polish did not matter much in determining the price.

## Observations and Comments

1. Carat, Clarity and Color were the primary determinant of the price.
2. It was followed by Cut, Certification and Polish.
3. Symmetry did not contribute in determining prices.
4. The price quoted to the Professor for the diamond appeared to be on higher side but it did not include the cost of retailing and ring itself. These were the prices of diamond alone.
5. This model is valid only for the diamonds within the range of observed value only.
6. Clarity could have been treated as nominal variable but in that case the regression coefficient would have been lower.
7. Professor is going to buy diamond only once and not regularly. Considering the one-time expense Professor can offer a price of around $2900 - $3000 to buy the diamond ring.