

Analysis of Diamonds Data: Do Consumers Pay for Sparkle?

Executive Summary

This report was chartered to examine an online price model to be used for a brick-and-mortar retailer who is extending their business to online diamond sales.

The retailer has proposed, prices in a store are influenced by the sparkle customers see radiating from a stone, but their concern is: online buyers cannot experience the sparkle. Will online customers pay more for sparkle? What affects a diamond's price for online buyers?

This research considers the asking price for 1,214 diamonds from prominent online diamond retailer, Blue Nile, against Blue Nile's interpretation of industry-standard diamond quality factors. The report reveals that online diamonds are not priced specifically for the characteristic of sparkle, but are more significantly influenced by the carat weight.

The analysis reveals that all grade factors: clarity, cut, color and carat, are influential in determining the sell price of a diamond, but carat emerges as the most significant influence. Except for extreme cases, such as a flawless diamond, the relative importance of color, cut, and clarity remains the same as carat weight increases.

Consumers appear to be willing to balance the factors of cut, color, and clarity to get a larger diamond. A good cut, that is slightly flawed, and slightly colored should be priced higher as the carat weight increases.

It is recommended that the retailer:

- Price their online diamonds with an emphasis on the carat weight, while scaling the price for the cut, color, and clarity.
- Employ a virtual assistant that suggests diamonds with higher carat weights to shoppers, but having equal color, cut, and clarity as the diamond being considered by the shopper.

Introduction

As stated in educational material on the Blue Nile website, a diamond's ability to reflect light is determined by its cut. Reflectiveness determines its shine, or 'sparkle,' and ultimately its beauty. But the raw data show that "Ideal" and "Astor Ideal" cuts have lower mean prices than do diamonds with the lesser "Good" and "Very Good" cuts. We explore whether prices increase with the quality of the cut, once we control for other factors.

Description and Exploratory Analysis

The goals of this section are to (1) describe the dataset and what the variables measure; (2) show how simple descriptive statistics relate to our question of interest, (3) assess linearity among the quantitative variables (4) identify predictors that should be considered and tested, as well as correlated predictors that could cause multicollinearity. These foundations help us set priorities for model-building and hypothesis testing.


Data

The dataset consists of 1,214 diamonds for sale on BN's online store. The variables include price and all of the "4-C's" that the gem industry uses to classify diamonds. We use the sample to make inferences about the population of diamonds sold online; Blue Nile alone currently has over 3,000 diamonds for sale on its website, and there are other vendors.

The data include two quantitative variables, price (mean=\$7,047, std. dev. = \$24,112) and carat (mean=.81 or .006 oz.), and four categorical variables. Table 1 reports the categorical variables and mean price for each category. Because these categories are essentially the industry

grading system for diamonds, they are on an (largely) ordinal scale and we have listed the categories from highest quality to lowest.¹

Table 1. Categorical Variables

	BN Category	Description	N. Obs.	%Obs in Cat.	Mean Price (\$)	Collapsed Category
CLARITY	FL	Flawless	3	0.2%	123,403	Flawless
	IF	Internally Flawless	49	4.0%	6,362	Flawless
	VVS1	Very, Very, Slightly Flawed	149	12.3%	8,667	Near Flawless
	VVS2	Very, Very, Slightly Flawed	158	13.0%	7,334	Near Flawless
	VS1	Very Slightly Flawed	233	19.2%	9,078	Very Slight Flaw
	VS2	Very Slightly Flawed	214	17.6%	7,726	Very Slight Flaw
	SI1	Slightly Flawed	243	20.0%	5,073	Slight Flaw
	SI2	Slightly Flawed	165	13.6%	2,626	Slight Flaw
COLOR	D	Colorless / Icy	207	17%	10,525.16	Perfectly clear
	E		181	15%	9,905.87	Clear
	F		223	18%	6,204.57	Clear
	G		198	16%	4,571.62	Very Slight Color
	H		148	12%	7,798.54	Very Slight Color
	I		167	14%	4,779.37	Slight Color
	J	Slightly Yellow	90	7%	3,934.09	Slight Color
CUT	Astor Ideal	Most Reflective (BN only)	20	1.6%	5,852	Astor Ideal
	Ideal	Most Reflective	739	60.9%	6,489	Ideal
	Very Good	Very Reflective	382	31.5%	7,758	Very Good
	Good	Reflective	73	6.0%	9,467	Good

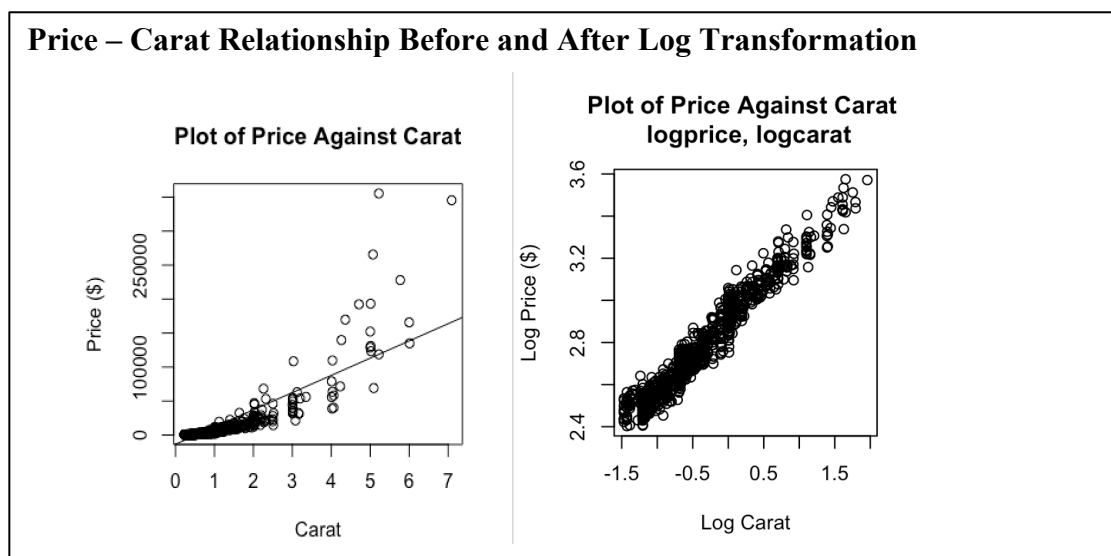
Information in the “Description” column paraphrases educational material from the BN site. The frequencies show a well-balanced dataset except for the three flawless diamonds comprising less than 1% of the data, with mean price \$123K, over 5 times the standard deviation of prices in the sample; and the Astor Ideal cut which is only 20 diamonds. To ensure indicator variables have sufficient degrees of freedom for hypothesis testing in a regression and to make the analysis of categorical variables more manageable we use alternate, collapsed levels for some of the analysis. These tend to follow industry breakpoints (e.g., the two classes each of “very slightly” and “very, very slightly” flawed diamonds are combined).

¹ VVS1 and VS1 are better than VVS2 and VS2, respectively. “Typically, similar cut, color, and carat VVS2 vs. VVS1 diamonds will differ somewhere around 25 percent. VVS1 diamonds being the more expensive of the two grades...You probably won’t be able to tell the difference between the two.” Source: With Clarity, vendor website, undated. Accessed 4/3/2021 at <https://www.withclarity.com/education/diamond-education/diamond-clarity/vvs2-vs-vvs1>

Table 1 shows inconsistent and unexpected patterns in prices central to the study question. Color is the only variable for which mean price and quality run consistently in the expected direction. Mean prices by clarity category are inconsistent, suggesting that increments to clarity below flawlessness are not significant for consumers, perhaps because flawlessness is difficult to see with the naked eye. For cut (sparkle), prices run in the opposite direction from expected, suggesting that consumers value other attributes much more.

Price-Carat Correlation

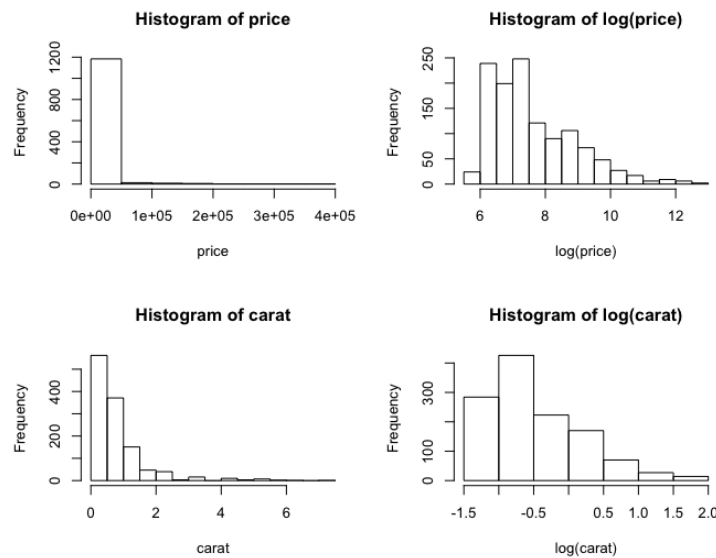
According to Blue Nile, carat is the dominant factor driving price and is the natural starting point for exploratory analysis. A scatterplot of the raw data is in the left panel of the figure below. It shows the expected positive relationship.



However, the plot suggests that the raw relationship between price and carat is non-linear and not suited for linear modeling. What is more, the spread of price around the fitted regression is not constant across the value range of the predictor, carat. Assumptions for least squares to be the Best (minimum variance) Linear Unbiased Estimator of parameters would not be met. The slope parameters estimated with this raw data are likely to be unbiased but the standard errors would be wide, resulting in inaccurate significance tests.

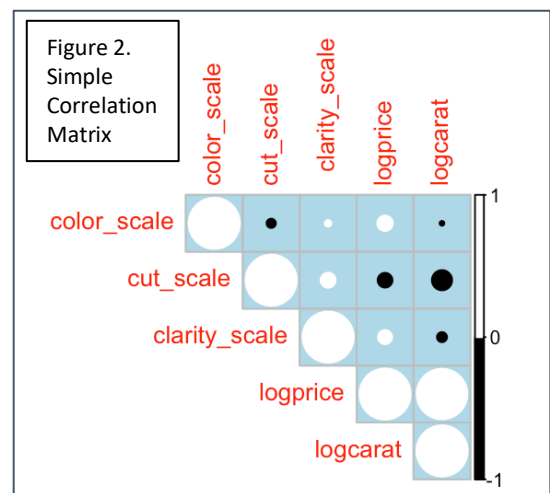
We applied variance-stabilizing natural log transformations to price and carat.

Our original carat to price plot shows an exponential distribution. A great way to linearize an exponential function is to do a log-log (natural log) transformation of the predictor and result variables.² Looking at the distribution of our variables for price and carat, you can see how the log transformation helps transform the shape closer to a normal distribution.



Other Correlations

In the multiple regression context, a correlation matrix is a good way to check for simple correlations between predictor variables. Simple correlations guide the model building process by suggesting what to include and where



² Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006). Introduction to linear regression ANALYSIS: Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining. In *Introduction to linear regression analysis: Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining* (pp. 177-178). Hoboken: Wiley-Interscience

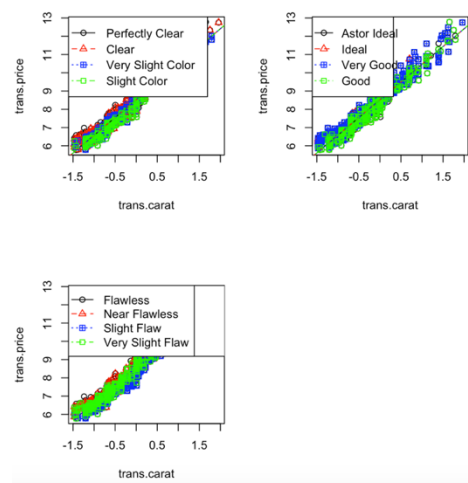
multicollinearity could be an issue. With so few variables in the dataset, we are focused less on what to include and more on multicollinearity.

We recoded the categorical variables into integer scales (highest quality mapped to highest integer). The numbers themselves are not meaningful, but a non-numeric plot showing relative magnitudes provides some useful indicators. The correlation graphic, Figure 2, suggests potential a negative relationship between cut and carat which may explain why the mean price ran in an unexpected direction across the cut categories (Table 1). If finer cuts tend to be of lower carat, it could explain the apparent negative relationship to price; BN consumers may be sacrificing sparkle to get a larger diamond resulting in the apparent negative relationship between cut and price. There is also a negative correlation between clarity and carat which could explain the inconsistent price-clarity pattern in Table 1.³

Model Building

Basic Model

In building the model, we aimed to find the best overall fit while considering the variables necessary to answer the central question. Initial tests focused on building up from $\log\text{price} = b_0 + b_1\log\text{carat}$, to testing whether the categorical variables (using the collapsed categories in Table 1) should be included.



Fitted regression equations are plotted in the figure above. Visually, the slopes are the same for each of the categories suggesting there are no major interactions between the

³ STHDA, “Visualize correlation matrix using correlogram,” Tutorial, Undated. Accessed April 3, 2021 at <http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram>

categorical variables and logcarat in determining price. There appear to be small differences in the intercept terms for each fitted line. This suggests that the price varies by category for a given carat weight; but that the relative importance of each category remains the same as carat increases along the horizontal axis.

Analysis of Variance (ANOVA) confirmed that all of the categorical variables, taken collectively, belong in the basic model.

```

Analysis of Variance Table

Response: trans.price
          Df Sum Sq Mean Sq  F value    Pr(>F)
trans.carat      1 1946.77  1946.77 59665.663 < 2.2e-16 ***
color_cat        3   19.71    6.57  201.386 < 2.2e-16 ***
diamonds4$clarity_cat  3   23.97    7.99  244.840 < 2.2e-16 ***
diamonds4$cut      3    9.47    3.16   96.773 < 2.2e-16 ***
Residuals      1203   39.25    0.03
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The ANOVA table indicates extra regression sums of squares that are added to the total sum of squares as new categories are added to the model. Logcarat (labeled transcarat) absorbs most of the of the total sum of squares, $\Sigma(\log\text{price}_i - \log\text{price_bar})^2$. All of the sequentially added terms increase explanatory power, given the inclusion of the previous terms. P-values at each stage are less than 1% indicating that we only have a 1% chance of incorrectly rejecting the null hypothesis that the terms do not belong in the model. The total sum of squares, SST, is $1947+24.95+18.75+9.47+39.25=2039.17$. The coefficient of determination, R^2 , is $1-\text{SSR}/\text{SST} = .98$.

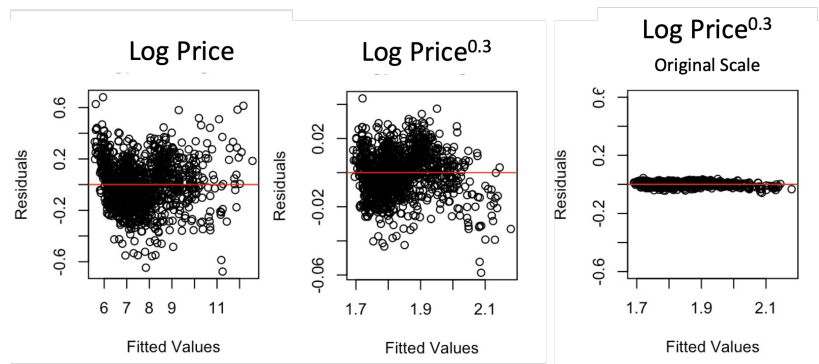
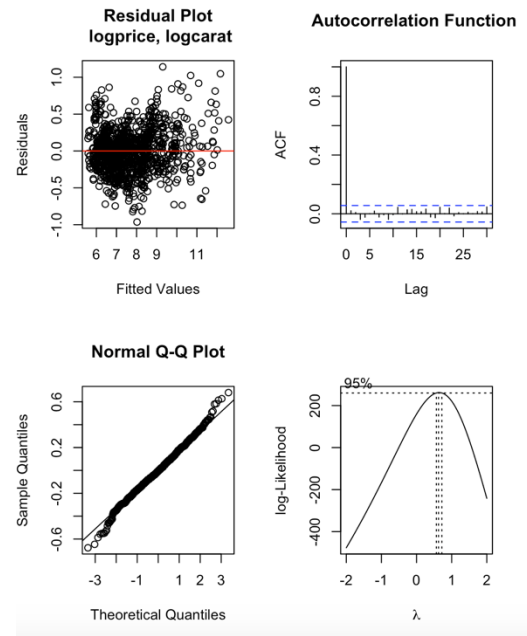
The ANOVA table and sequential F tests do not, however, indicate which combination of categorical variables best explain prices, nor do they guarantee that the model meets the criteria for least squares to be the best (minimum variance) linear unbiased estimator (BLUE).

Does the Basic Model Meet the Assumptions for Least Squares?

The residuals are evenly scattered on either side of the regression line for all fitted values.

Autocorrelation, that is, correlation between the residuals, is not an issue. The data appear to be reasonably normal although there is a slight indicator of a heavy upper tail (a greater percentage of values being in the upper end of the distribution than in the

normal) and a thin lower tail. These results support the basic model. However, the Box Cox test shows that $\lambda=1$ is outside the 95% confidence interval for minimizing the Sum of Squared Residuals (SSRes). We decided to transform the response variable further, to be the cube root of log price. At first glance the cube root transformation worsens the scatter of residuals around the zero-axis compared to the first model (Log Price), pushing them below the line at higher fitted values. However, rescaling the y-axis shows that the $\log\text{price}^{0.3}$ model greatly compresses the variance of the errors around the fitted values.



Interaction Terms

We explored the need for interaction terms between carat on one hand and cut, clarity and color on the other. Although visually, we saw no indication of separate slopes or intercept terms for the categorical variables, the effect may be small. The interacted model is:

$$\begin{aligned} \logprice^{0.3} = & b_0 + b_1 \logcarat + \sum b_{2j} Color_i + \sum b_{3j} Clarity_i + \sum b_{4j} Cut_k \\ & + \sum b_{5j} Color_j \logcarat_i + \sum b_{6j} Clarity_i \logcarat_i + \sum b_{7j} Cut_i \logcarat_i \end{aligned}$$

The reference variables are the lowest quality levels for each category: Cut - “Good”, Clarity – “Slight Flaw” and Color – “Slight Color” which collectively are absorbed in the intercept term b_0 .

The ANOVA table for the interacted model shows the regression sum of squares from adding each set of interactions to the model, in addition to the previous terms in the model. A p-value $> .05$ shows that that each set of interaction terms adds new information toward explaining the total sum of squares, beyond the basic model.

Response: tlogprice						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
trans.carat	1	9.3209	9.3209	56540.2320	< 2.2e-16	***
cut	3	0.0447	0.0149	90.3771	< 2.2e-16	***
clarity_cat	3	0.1045	0.0348	211.2926	< 2.2e-16	***
color_cat	3	0.0947	0.0316	191.5823	< 2.2e-16	***
trans.carat:cut	3	0.0027	0.0009	5.4866	0.0009593	***
trans.carat:clarity_cat	3	0.0013	0.0004	2.6508	0.0474926	*
trans.carat:color_cat	3	0.0061	0.0020	12.3642	5.748e-08	***
Residuals	1194	0.1968	0.0002			

Discussion: Final model considerations

The interacted model above is the starting point for testing variations and choosing the final model for hypothesis testing. We debated whether to continue with a response variable of $\logprice^{0.3}$ or to simply use \logprice . This was a trade-off between interpretability and minimization of standard errors. $\logprice^{0.3}$ has no interpretation, while changes in \logprice

can be interpreted as percent changes. If there is not a statistically meaningful or qualitative (i.e., ranking of predicted values) difference between the Cut terms, we would prefer to use the logprice model. We expected that the model would fit better for the $\logprice^{.3}$ model but wanted to see hands-on how it affected the results. To do so, we compared the p-values, rankings of predicted values, and prediction spreads as a percentage of the point estimate for (a) response = logprice (regression output not shown) and (b) response = $\logprice^{.3}$. The results are summarized in the table.

Issue 1 Comparison: Impact of Dependent Variable Choice on Interacted Model

	Response = $\logprice^{.3}$	Response = LogPrice
Fitted Value Ranking by Category _ - Point Estimates	Astor Ideal, Good, Very Good, Ideal	Same
Spread of Upper and Lower Predictions (as % of Point Estimate) – Astor Ideal	98.6% 101.4%	96% 104%
Can Astor ideal be ranked last within 95% CI?	Yes	Yes
P-Values for Cut Interactions	P>.05 for Astor Ideal and Very Good, <.05 Ideal	P>.05 for all

As expected, the logPrice model has a slightly higher error spread around the predicted price. Further, the logPrice model had no significant interactions with cut despite the ANOVA indicating that collectively, the interactions are significant – signifying the multicollinearity issue with these terms that would inflate the standard errors. Given this “uphill battle” the cubic root transformation was retained to improve the potential fit of the model.

Secondly, we decided to eliminate the interactions between cut and logcarat. While useful, they added complexity to our hypothesis testing and could be pursued at a later stage.

Hypothesis Testing

We will use the final model to test whether consumers pay for additional sparkle, which is dependent upon cut. Writing the final model in a way that highlights the terms of interest:

$$\text{Logprice}^0.3 = b_0 + b_1 * \text{AstorIdealCut} + b_2 * \text{IdealCut} + b_3 * \text{VeryGoodCut} + \Sigma b_5 * X,$$

Where $\Sigma b_5 * X$ is a vector of carat, color, and clarity indicators together with their interaction terms.

Final Model

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.859360	0.001878	990.176	< 2e-16 ***
trans.carat	0.132286	0.001551	85.306	< 2e-16 ***
cutAstor Ideal	0.023484	0.003275	7.171	1.30e-12 ***
cutIdeal	0.013907	0.001609	8.645	< 2e-16 ***
cutVery Good	0.002737	0.001662	1.647	0.09991 .
clarity_catflawless	0.030979	0.002485	12.465	< 2e-16 ***
clarity_catnear_flawless	0.017713	0.001233	14.368	< 2e-16 ***
clarity_catvery_slight_flaw	0.012360	0.001095	11.288	< 2e-16 ***
color_catclear	0.023174	0.001292	17.933	< 2e-16 ***
color_catperfectlyclear	0.030906	0.001496	20.657	< 2e-16 ***
color_catveryslightcolor	0.016154	0.001322	12.222	< 2e-16 ***
trans.carat:clarity_catflawless	-0.001665	0.002841	-0.586	0.55784
trans.carat:clarity_catnear_flawless	-0.004365	0.001539	-2.837	0.00463 **
trans.carat:clarity_catvery_slight_flaw	-0.003130	0.001427	-2.193	0.02847 *
trans.carat:color_catclear	0.008077	0.001610	5.017	6.04e-07 ***
trans.carat:color_catperfectlyclear	0.008837	0.001895	4.663	3.47e-06 ***
trans.carat:color_catveryslightcolor	0.008074	0.001672	4.829	1.55e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01293 on 1197 degrees of freedom

Multiple R-squared: 0.9795, Adjusted R-squared: 0.9793

F-statistic: 3581 on 16 and 1197 DF, p-value: < 2.2e-16

The coefficient values are given in the regression table, Final Model. The coefficients of interest are cutAstor Ideal, cutIdeal and cutVerygood, which represent intercept terms for each cut category. “Good” is the reference cut category. The differences indicate the average difference in the price when logcarat=0 and when other indicators – color and clarity- are set to their reference values. The reference diamond in the regression has good cut, is slightly flawed, and slightly colored (lowest quality in each dimension).

Intercept Point Estimates by Cut		
Cut	Formula	Intercept
Astor Ideal (Highest Sparkle)	$b_0 + b_1$	1.883
Ideal	$b_0 + b_2$	1.873
Very good	$b_0 + b_3$	1.862
Good (Lowest Sparkle)	b_0	1.859

Test 1 $H_0: b_1=b_2=b_3=0$. $H_a: b_1>0$ or $b_2>0$ or $b_3>0$. Test 1 merely checks if Very Good, Ideal or Astor Ideal diamonds have a higher intercept than Good diamonds, given the presence of the other predictors in the model and is a weak test. *Result:* The R output is for a two-sided test, so we must recalculate the critical value for rejecting the null hypothesis with 95% confidence, which is 1.646. We reject the null hypothesis as their t-values are at or well above the critical value (the t-value for cutVeryGood is 1.647). This result is counter to the simple means shown in Table 1 where “good” cuts commanded the top price, over \$9K on average. Given the presence of the other variables in the model, “good” cuts are less expensive than some of the others.

Test 2. $H_0: b_1=0$ or $b_2=0$ or $b_3=0$; $H_a: b_1>0$ and $b_2>0$ and $b_3>0$. Test 2 assesses if Very Good, Ideal and Astor Ideal diamonds have higher intercepts than Good diamonds, given the presence of the other predictors in the model. Unlike the first test, this is a joint test for collective significance of the family of coefficients. *Result.* The Bonferroni method tests if a family of parameters is different from the value under the null, which is zero. To apply the Bonferroni method, we identify the critical value that would ensure at least 95% confidence for all three members of the family. There are 3 parameters to test so instead of obtaining the 95% CI, we need to estimate the 98.3% CI ($=1-\alpha/3$) with $1214-3-1=1,210$ degrees of freedom. In our one-sided test, the t-value is $qt(1-.017,1210)= 2.12$. For test 2, we cannot reject the null hypothesis because the t-statistic for b_3 (Very Good) is less than the critical value.

Confidence Intervals for Differences in Intercepts. With the Bonferroni method we develop confidence intervals around differences in the cut intercept terms. The table below shows these calculations for each category. The formula for the CI are $Est \pm t*SE$. In this case, we will

apply the critical value for a 2-sided test, which is $qt(1-.017/2,1210)=2.39$. We can be 95% confident that Astor Ideal has a larger logprice intercept than Ideal (zero and negative numbers are outside the CI for the difference); and that Ideal is priced higher than Very Good. The one departure from expectation is that the 95% confidence interval of the difference in Very Good includes zero suggesting that consumers do not differentiate between Very Good and Good.

	Formula for Est.	Est.	Var or COV	Formula for SE	SE	Lower CI	Upper CI
Astor ideal - Good	b1	0.0235	0.00001073	Sqrt(Var)	0.003276	0.0157	0.0313
Ideal - Good	b2	0.0139	0.00000259		0.001609	0.0101	0.0177
Very Good - Good	b3	0.0027	0.00000276		0.001661	-0.0013	0.0067
Astor Ideal - Ideal	b1-b2	0.0096	0.00000233	SQRT(V(b1)+V(b2) - 2COV(b1,b2))	0.002943	0.0026	0.0166
Astor Ideal - VG	b1-b3	0.0208	0.00000232		0.002975	0.0137	0.0279
Ideal vs. Very Good	b2-b3	0.0112	0.00000233		0.000831	0.0092	0.0132

Conclusion

At first glance, the data suggest that consumers do not pay more for the higher quality cuts that produce the greatest sparkle. Controlling for other factors, especially carat, changes this conclusion. The estimated intercepts show consistency with the quality grading scales for cut. However, the price of sparkle appears to be overwhelmed by an inverse relationship with carat, which, as our analysis has confirmed, is the dominant factor in what consumers pay.