Project 1: Diamonds

Executive Summary:

We are tasked with analyzing a dataset containing 210,638 unique observations of diamonds from Blue Nile, one of the leading diamond specialists in the e-commerce space. Each observation includes details of a diamond's carat, clarity, cut, color, and price. As a hypothetical buyer on behalf of Blue Nile, our objective is to create a model that will evaluate the price of a diamond given the aforementioned details in order to determine where the price or "willingness to pay" is less than the amount for which we can purchase the diamond wholesale, in order to make a profit.

This report describes the observation parameters within the dataset, illustrates our methodology on how we created our model, and provides an example of how our model can be used with data from a diamond distributor, Israel Diamonds. We develop a model that relates the log(price) of the diamond to the log carat, indicators for grouped clarities, indicators for grouped cuts, indicators for color, as well as interactions between log(carat) and color. This model satisfies all necessary assumptions, and utilizing it, we recommend diamonds to purchase from a given wholesaler to make a profit.

Introduction and Problem Statement:

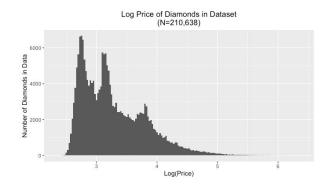
Blue Nile is one of the leading players in the diamond marketplace specializing in engagement rings, fine jewelry, and their in-house brand, "Astor by Blue NileTM" diamonds." Their catalog of diamonds is one of the largest on the internet. We have been tasked to analyze a large subset of data from Blue Nile (featuring 210,638 different diamonds) in order to draw some practical and interesting for our target customer. We would like to create a model that will predict the price of a diamond given the features associated with diamonds that are in the common vernacular, the 4 C's carat, clarity, color, and cut. With our model, we would like to evaluate the price of diamonds in new datasets in order to determine when the price or "willingness to pay" for our consumer is less than the amount for which we can purchase the diamond, in order to make a profit. Essentially we seek to answer the following questions: What diamonds should we buy from our wholesaler? How do we know if we are getting a good deal?

Descriptions of Data:

We employ a dataset containing 210,638 unique observations of diamonds, containing each diamond's carat, clarity, color, cut, and price. To gain an understanding of the types of diamonds in the data, we conducted basic exploratory data analysis, prior to developing an in-depth predictive model:

Price

Given the nature of diamonds as a luxury good, one of the most distinguishing factors in determining the sector of the diamond market that we would be analyzing is understanding the distribution of price. The median price of diamonds in the dataset is \$1,432, with a strong right-tailed skew, as can be seen in Figure 1, depicting the distribution (appearing trimodal) of log price of the diamonds.



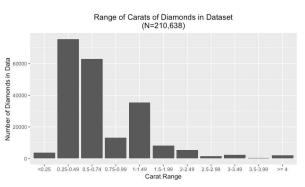


Figure 1: Distribution of log price

Figure 2: Distribution range of carats.

Carat

The term "carat" refers to the metric of how much a diamond weighs. A metric "carat" is equivalent to 200 milligrams. It is a misnomer to equate the overall size of a diamond to the carat -- any two diamonds may have the same diameter, but the cut may cause the two to have completely different carat measurements. Carat values are commonly identified as a cutoff for price, i.e. when increasing to a range of different half or whole carats (from the 0.5-0.99 range to the 1 -1.49 range). To obtain an understanding of where our carat values lie, we generate the chart in Figure 2, which tells us that the vast majority of diamonds in our analysis set are below 1.5 carats:.

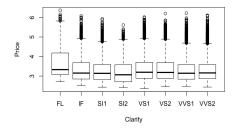
Additionally, there is a heavy bias towards diamonds of 1 carat versus those of 0.75 - 0.99 carats, which could indicate either a supply-side bias towards producing 1 carat diamonds or a demand-side bias towards people who want to purchase a diamond of at least 1 carat or larger.

Clarity

Clarity refers to the qualitative measure of the overall visual aesthetic of the diamond. The grades assigned to the diamonds in this specific dataset are:

FL	Flawless	VS1	Very Small Inclusions 1
IF	Internally Flawless	VS2	Very Small Inclusions 2
VVS1	Very Very Small Inclusions 1	SI1	Small Inclusions 1
VVS2	Very Very Small Inclusions 2	SI2	Small Inclusions 2

Depending on the lab and/or distributor, additional grades may be included, or some of the above grades may be omitted. For this data set, we assumed that the clarity grades were not evenly distributed.



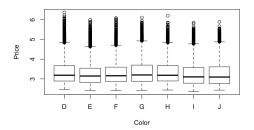


Figure 3: Distribution of clarity vs. log price

Figure 4: Distribution of color vs. log price

From Figure 3, it becomes obvious that Flawless diamonds have a different distribution with regards to the log price compared to other clarity levels.

Color

Although diamonds and other semi-precious stones come in a variety of colors, in this data set, we are specifically looking at white diamonds. In the case of white diamonds, "Color" refers to the presence (or ideally) absence of yellow discoloration in a diamond. The color grades assigned to the diamonds in this dataset are:

D, E, F Colorless - virtually colorless to the untrained eye E, G, I, J Near-colorless - nearly colorless

Additional grades K, L, and M (faint yellow), and N-Z (increasing grades of yellow) will not be used in this set. Figure 4 illustrates the distribution of log price compared to color levels.

A hypothesis we had prior to analyzing the data was that, as the carat of the diamond increases, the color of the diamond would have a heavier impact on the overall price; i.e. the bigger the diamond, the easier it is to pick out differences in color to the naked eye.¹

Cut

Counterintuitively, a diamond's cut does not refer to its overall shape in a setting. The measure refers to the overall quality of the diamond factoring in proportions (diameter-to-weight ratio), clarity and polish. Because of the melding of factors, cut is commonly held as one of the premiere factors in evaluating the value of a diamond. The grades that Blue Nile assigns their diamonds include:

Astor Idea - highest grade Ideal Very Good Good

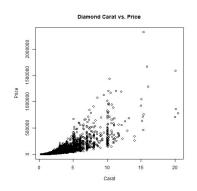
Price & Carat

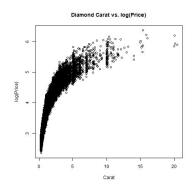
Plotting carat versus the price of a diamond for each diamond in the dataset produces the scatterplot in Figure 5. Examining the relationship between Carat and Price, demonstrates that for every unit increase in carat, price increases by an increasing amount. In an attempt to see a linear relationship between carat and price, we plot the log of price versus carat, displayed in Figure 6. Now, for every unit increase in log(price), carat is increasing at a decreasing rate. Our next step is to

¹ Lumera Diamonds - <u>Diamond Color Chart</u>

transform carat by using the log function as well and plotting the result of log(carat) versus log(price) as displayed in Figure 7.

After transforming carat and price with the log function, we finally see a linear relationship between these two entities. Our next step is to visually inspect any relationships we might see when we color each point by each of our categorical variables: color, clarity, and cut.





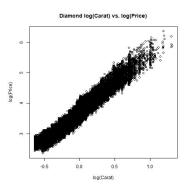


Figure 5: Plot of carat vs. price

Figure 6: Plot of carat vs. log price

Figure 7: Plot of log carat vs. log price

Price & Carat vs. Color

Figure 8 shows that given the same carat, that more desirable colors (i.e. D, E, and F) typically fetch higher prices than those of less desirable colors. This relationship holds true even as carat increases. Since color is the most empirically measured feature between color, clarity, and cut, it makes sense that there is a high level of stratification between these levels since there is little room for interpretation with regards to a particular diamond's color. Interestingly, as the carat increases, the level of stratification between colors becomes increasingly apparent. We will keep this effect in mind as we start to create and iterate upon our pricing model.

Price & Carat vs. Clarity

Figure 9 shows that given the same carat of diamond, that more desirable clarities (i.e. FL, IF, and VVS1/2) typically fetch higher prices than those of less desirable clarities. This relationship holds true even as carat increases. Clarity is less empirically measured than color, as clarity is determined by expert diamond graders, but nonetheless there are very strict guidelines and rules surrounding the grading of clarities of diamonds. This is reflected in the chart above, as we once again see strong stratification between the different levels of clarity with respect to price at a given carat.

Price & Carat vs. Cut

Figure 10 shows that given the same carat, that more desirable cuts (i.e. Astor Ideal and Ideal) generally fetch higher prices than those of less desirable cuts. However, unlike color and clarity, the impact that color has on price at a given carat is not as obvious. Blue Nile's cut rating

system² deviates from the standard GIA cut rating system³, where diamond cuts are graded on a scale of Excellent to Poor. Given the increased room for interpretation when it comes to grading cuts under Blue Nile's grading system, one should take the impact that this data set's cut grade has on the diamond's price with a grain of salt.

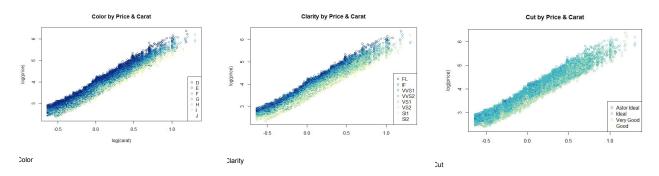


Figure 8: Plot of color vs. log price & log carat

Figure 9: Plot of clarity vs. log price & log carat

Figure 10: Plot of cut vs. log price & log carat

Analytic Approach:

Given our goal of building a model that would best predict the price of a diamond using each of the 4 C's as predictive factors, we took an iterative approach, documenting our assumptions and findings along the way. We started with a baseline model using each of the 4 C's as main effects to predict the price without transformations on the response variable or any of the predictor variables:

Model 1: Main effects, no transformations

$$y = \beta_0 + \beta_1 x_1 + \beta_{cut} x_{cut} + \beta_{clar} x_{clar} + \beta_{color} x_{color}$$

where $y = price$
 x_{cut} is a vector of dummy variables for cuts

 x_1 is carat

 x_{clar} is a vector of dummy variables for clarities

 x_{color} is a vector of dummy variables for colors

Model 1: Performance

R-squared: 0.583 All variables significant to $\alpha = 0.05$

Model 1: Diagnostics

The Q-Q plot in Figure 11 shows that our residuals do not follow a normal distribution. Both tails are "fat" relative to the normal distribution, and there is a heavy skew in the right-side tail. Since this violates a base assumption that our linear regression model has normally distributed residuals, we decided to look for ways to transform our response and/or predictors in order to ensure this assumption is not violated.

² Blue Nile - Blue Nile Cut Ratings

³ Gemological Institute of America - <u>Diamond Cut Grades</u>

Model 2: Main effects on log(price)

$$log(y) = \beta_0 + \beta_1 x_1 + \beta_{cut} x_{cut} + \beta_{clar} x_{clar} + \beta_{color} x_{color} + \varepsilon$$

where $log(y) = log(price)$
 x_{cut} is a vector of dummy variables for cuts

 x_1 is carat

 x_{clar} is a vector of dummy variables for clarities

 x_{color} is a vector of dummy variables for colors

Model 2: Performance

R-squared: 0.743 All variables significant to $\alpha = 0.05$

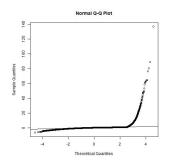
Model 2: Diagnostics

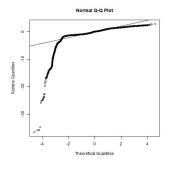
As seen in model 1, model 2's residuals, as displayed in Figure 12, still do not resemble a normal distribution, with heavy skew to the left-side tail. Given the observation of an apparent linear relationship between the log(price) and the log(carat), our next model looks to transform both price and carat using the log function.

To validate that transforming carat using the log function would be a worthwhile approach, we performed a Box-Cox transformation on price using the log(carat) and remaining main effects terms. The resulting optimal lambda after performing the Box-Cox transformation was -0.02 as displayed in Figure 13.

$$\lim_{\lambda o 0} rac{y^{\lambda} - 1}{\lambda} = \lim_{\lambda o 0} rac{\exp[\lambda \log y] - 1}{\lambda} = \lim_{\lambda o 0} \exp[\lambda \log y] \log y = \log y < \infty$$

Knowing that the limit of the Box--Cox transformation as lambda approach zero is the log transform, we felt confident that transforming both price and carat with the log transformation would do an ideal job of creating a linear relationship between these two attributes.





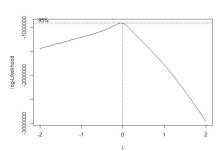


Figure 11: Q-Q Plot (normality of residuals) for Model 1

Figure 12: Q-Q Plot (normality of residuals) for Model 2

Figure 13: Box-Cox transformation to identify the optimal lambda

Model 3: Main effects (log(carat) instead of carat) on log(price)

$$log(y) = \beta_0 + \beta_1 log(x_1) + \beta_{cut} x_{cut} + \beta_{clar} x_{clar} + \beta_{color} x_{color} + \varepsilon$$

where log(y) = log(price)

 $log(x_1) = log(carat)$

 ϵ is the error term

Model 3: Performance

R-squared: 0.981

Model 3: Diagnostics

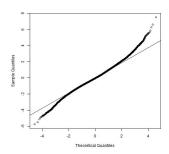


Figure 14: Q-Q Plot (normality of residuals) for Model 3

 x_{cut} is a vector of dummy variables for cuts

 x_{clar} is a vector of dummy variables for clarities

 x_{color} is a vector of dummy variables for colors

All variables significant to $\alpha = 0.05$

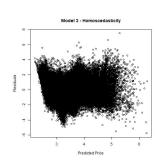


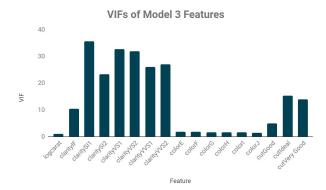
Figure 15: Test result for heteroskedasticity (residuals vs. fitted values) for Model 3

As illustrated in Figure 14, while the tails are a little fatter than those of a normal distribution, this is a significant improvement in the normality of our model's residuals over our previous attempts. Additionally, this distribution is fairly symmetric about the mean.

Residuals vs. Fitted Values (test for heteroskedasticity):

In Figure 15, we see that as we vary our fitted values, the variance of the residuals of our model remains fairly constant, and does not show signs of heteroskedasticity within our model.

Variance Inflation Factors (VIF) (test for Multicollinearity):



With a VIF greater than ten usually demonstrating multicollinearity within a linear regression model, Model 3 shows that the variance of the predicted Betas for our Clarity and Cut features is significantly above that threshold. When this behaviour is observed, our options typically are to:

- 1. Remove the variable whose predicted Beta has a high VIF
- 2. Transform the variable whose predicted Beta has a high VIF

Given that our research demonstrates that Clarity and Cut do play a prominent factor in determining the price of a diamond, we must look for ways to transform the Clarity and Cut variables to reduce multicollinearity.

Model 4: Main effects (grouped_cuts and grouped_clarity instead of cut and clarity) on log(price)

$$log(y) = \beta_0 + \beta_1 log(x_1) + \beta_{cut_g} x_{cut_g} + \beta_{clar_g} x_{clar_g} + \beta_{color} x_{color} + \varepsilon$$

$$where log(y) = log(price)$$

$$x_{cut_g} \text{ is a vector of dummy variables for grouped cuts}$$

$$log(x_1) = log(carat)$$

$$x_{clar_g} \text{ is a vector of dummy variables for grouped clarities}$$

$$\varepsilon \text{ is the error term}$$

$$x_{color} \text{ is a vector of dummy variables for colors}$$

$$levels \text{ of grouped cuts : Astor Ideal & Ideal, Very Good & Good}$$

$$levels \text{ of grouped clarities : FL & IF, VVS1 & VVS2, VS1 & VS2, SI1 & SI2}$$

Model 4: Performance

R-squared: 0.98

All variables significant to $\alpha = 0.05$

Model 4: Diagnostics

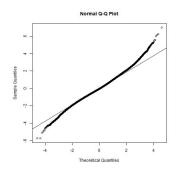


Figure 16: Q-Q Plot (normality of residuals) for Model 4

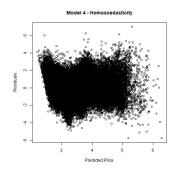


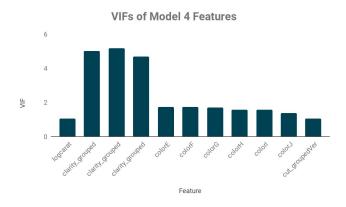
Figure 17: Test result for heteroskedasticity (residuals vs. fitted values) for Model 4

Once again, As displayed in Figure 16, the tails are slightly fat, but still do a decent job of resembling a normal distribution.

Residuals vs. Fitted Values (test for heteroskedasticity):

Like Model 3, we see that as we vary our fitted values, the variance of the residuals of our model remains fairly constant, and does not show signs of heteroskedasticity within our model. We can see this in Figure 17.

Variance Inflation Factors (VIF) (test for Multicollinearity):



Unlike what we observed from Model 3, Model 4 does not have any predicted Betas whose VIF is greater than 10. Furthermore, only clarity_groupedVS1/2 has a VIF greater than 5. This result seems sufficient to quell any doubts about multicollinearity within our model.

Model 4: Room for Improvement

As stated earlier, our research indicates that there is a direct relationship between the carat and the color of a diamond with respect to its price. The reasoning here is that as diamonds increase in carat, the color of said diamond becomes increasingly apparent to the human eye. Thus, we believe that there is an interaction between the carat and the color of a diamond when modeling on price. Model 5 will add an interaction term between color and carat and we will perform a partial F-test to determine if the addition of this variable is significant as it pertains to the model.

Model 5: Main effects plus interaction between log(carat) and color on log(price)

$$log(y) = \beta_0 + \beta_1 log(x_1) + \beta_{cutg} x_{cutg} + \beta_{clarg} x_{clarg} + \beta_{color} x_{color} + \beta_{logcarat-color} x_{logcarat-color} + \varepsilon$$

$$where \ log(y) = log(price) \qquad x_{clar_g} \ is \ a \ vector \ of \ dummy \ variables \ for \ grouped \ clarities$$

$$log(x_1) = log(carat) \qquad x_{color} \ is \ a \ vector \ of \ dummy \ variables \ for \ colors$$

$$x_{cut_g} \ is \ a \ vector \ of \ dummy \ variables \ for \ the \ interaction \ terms \ between \ logcarat \ and \ color$$

$$\varepsilon \ is \ the \ error \ term$$

$$levels \ of \ grouped \ cuts : Astor \ Ideal \& \ Ideal, \ V \ ery \ Good \& \ Good$$

$$levels \ of \ grouped \ clarities : FL \& IF, \ VVS1 \& VVS2, \ VS1 \& VS2, \ SI1 \& SI2$$

Model 5 performance:

Model 5 partial F-test:

	Res.Df	RSS	Df	Sum of Sq	\mathbf{F}	Pr(>F)
1	210626	1190.13				
2	210620	1143.88	6	46.25	1419.44	0.0000

The Analysis of Variance (ANOVA) table demonstrates that the addition of the interaction term between log(carat) and color had a significant impact on the larger model. Given this result, we will elect to keep the interaction term in our model moving forward.

Model 5 Diagnostics

Like Models 3 and 4, the tails are slightly fat, but do a decent job of resembling a normal distribution as illustrated in Figure 18.

Residuals vs. Fitted Values (test for heteroskedasticity):

Like Models 3 & 4, we see that as we vary our fitted values, the variance of the residuals of our model remains fairly constant, and does not show signs of heteroskedasticity within our model.

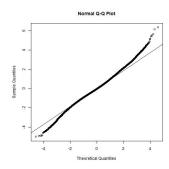


Figure 18: Q-Q Plot (normality of residuals) for Model 5

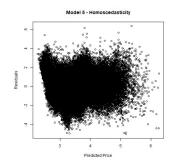


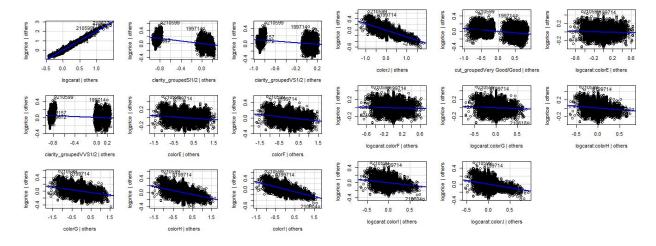
Figure 19: Test result for heteroskedasticity (residuals vs. fitted values) for Model 5

Variance Inflation Factors (VIF) (test for Multicollinearity):

Like Model 4, Model 5 does not have any predicted Betas whose VIF is greater than 10, so we can continue without much fear of multicollinearity.

Added Variable Plots

The variable plots (AV Plots) in Figure 20 show that each of the features, when regressed against all of the other features of model 5, have a significant contribution to the model. In particular, if one examines the AV Plots for the interaction terms of Model 5, you can observe that as color worsens from D to J, the slope of the line becomes increasingly negative. The impact that increasing the carat of the diamond of color J has a significantly stronger impact on the log(price) than a diamond of color E. This helps evidence the claim that changes in color have a stronger impact on price as the carat of the diamond increases.



Results:

As demonstrated by the above diagnostic tests, our model has met the necessary assumptions to be a reasonable predictor, and is able to explain roughly 98% of the variation in pricing. The ability of the model to accurately predict prices offers a significant value added, as we now have the ability to consult a diamond dealer on an expected range of selling prices for a diamond with certain characteristics. Based on diamonds that are on the wholesale market, we can predict price ranges for Blue Nile of these given diamonds and determine where the price range is significantly higher than the cost on the wholesale market, offering an opportunity for profit.

Take, for example, the following listings from Israel Diamond, a wholesale diamond retailer. Uncertainty exists as to the collection time of the data that we formulated our model based on, therefore, in order to obtain reliable predictions, we must assume that the data were taken at a time very similar to now:

Israel Diamond has the following wholesale diamonds for sale (a selected sample), and the following section describes the amount that our model predicts for price (the prediction interval):

Listing	4									Predicted Price (\$)	Lower Bound (\$)	Upper Bound (\$)
	Shape Round	Weight \$ 0.30	Color \$	Clarity \$	Cut G	Polish VG	Sym G	Report	Price \$ \$279.00	383.65	275.10	535.03
	Shape Oval	Weight \$	Color	Clarity SI2	Cu VG		Syr		Price \$ \$769.00	986.40	707.30	1,375.73
	Shape Round	Weight \$ 0.85	Color ‡	Clarity \$	Cut G	Polish VG		Report	Price \$ \$1,856.00	2,960.51	2,122.86	4,128.68
	Shape Radiant	Weight 	Color \$		Cut VG		Sym G	Report GIA	Price \$ \$5,071.00	6,908.39	4,953.63	9,634.92

Using this predictive model, we can make recommendations to our client regarding the diamonds to buy from the wholesaler in order to have the best mark-up potential and maximize profit. For example, if presented with the four above choices, we would recommend to our client the purchase of the third diamond, as 95% prediction interval for price, as determined by our model is above the price that the diamond could be purchased from the wholesaler, identifying a profit opportunity for Blue Nile.

Next Steps:

- How do the cut scores for lab-grown diamonds compare to natural diamonds? Are their scores positively or negatively impacted in any way? Is there any bias?
- Would a step function for carat do a better job of predicting price than log(carat)?⁵
- Is the additional value from the "Astor Ideal" grade derived because of brand marketing?
- Looking for high-leverage points within this dataset and determining their impact on the model

⁴ Listings and prices as of 7/17/2019, from <u>Israel Diamond's Loose Diamond Collection</u>

⁵ CreditDonkey - <u>Diamond Prices</u>