Executive Summary

Data Description

Exploratory Data Analysis

Regressing Price on Carat

Blue Nile Diamond Price Prediction

8/1/2021

Executive Summary

The purpose of our team's analysis focused on identifying the relationship between the price of Blue Nile diamonds as dependent on the diamonds' carat size, overall clarity, color, and cut. Specifically, we divided this focus into to three core goals:

- Understand the relationship of a diamonds' price as it is associated with the carat size, clarity, color, and cut
- Assess the claims made by Blue Nile on the diamond education page
- Fit an appropriate simple linear regression model for diamond price against carat to determine the relationship

Goal 1

In initially understanding the relationship of price to the other variables, it is important to note that outliers in price are common across the data regardless of how the data is organized by attribute. Although the majority of diamond prices remain in a range of 0 - 4K dollars, these significant outliers in diamond prices skew prices to the right making averages for price considerably biased.

With this understanding of skewed price, our exploratory analysis led to examining prices based on carat as well as the price based on defined categories for cut, color, and clarity. From viewing the categories against price by varying permutations using tests of distribution, correlation, and variation, we observed a noticeable, yet weak, relationship between price and clarity followed by a similar, although slightly weaker, relationship between price and color. It is important to note that these relationships became stronger as prices skewed higher, likely resembling that rather than there being a single attribute pushing price higher that there is a interaction effect between variables.

Goal 2

Leveraging the knowledge gleaned from the Goal 1, the assessment of claims by Blue Nile's education page became more apparent. Specifically, the claims assessed included:

- Cut is the most important of the 4Cs
- Color is the second most important of the 4Cs

- clarity is least important of the 4Cs
- SI clarity is the best value
- VS clarity is the most popular

From the analysis performed of price against the 4 types of cuts (Good, Very Good, Ideal, Astor Ideal), there is not a noticeable linear relationship. Specifically the average prices dip between each of the four levels as the categories increase in value while the median prices dip between each of the initial three levels (Good to Very Good and Very Good to Ideal) before jumping to the highest at Astor Ideal. We would have assumed that a positive linear relationship would exist showing as cut increased in quality, price increased at a proportional rate. After removing skewed prices with the median statistical measure, we can see this is not the case.

Color did indeed have a relationship with price. As the recorded second most important of the 4Cs, the median prices of the seven categories primarily followed in order of color grade outside of colors J and H falling outside of the expected values. Regarding Clarity as the least important, we assessed this in turn with the fourth and fifth claims associated with SI and VS.

In assessing the most popular claim given to SV clarity diamonds, we can confirm agreement as the volume of stock happens to be far higher than for other clarity diamond groupings, assuming volume is associated with sales. The follow up claim though stating that SI clarity is the best value is far more difficult to confirm. This statement had likely been concluded initially as a result of the average and median prices of the SI2 clarity diamonds being the lowest average price of the eight clarity types, but the SI1 happens to be the second highest average and median price as well as the most popular. There would need to be a refined definition of ‘best value’ but if this is purely determined based on median price (as to remove the outlier prices), the VV clarity would be the ‘best value’ with a median price of 1,390.71 (compared to SI of 1,442.28). It is important to note that although this is a factor in determining price, we can agree with the third claim that it may have the weakest relationship with price given the variation observed.

Goal 3

Finally, our team narrowed our analysis to fitting a simple linear regression model for diamond price against carat. The purpose of this final goal was to understand what impact carat size has on diamond price in order to assess what proportion of price cannot be explained by carat but rather by variance or one of the other categorical variables. For the lower priced diamonds we found the majority of data had a well fit relationship with carat size although tailed off as carat and price increased (a result of the known outliers). This impact from the outliers led to receiving a ~68% R^2 value which tells us that 68% of price can be explained by carat size. In an effort to enhance the fit and decrease the proportion associated to variance, we developed a residual plot and a box-cox plot. The residual plot, measuring the distance between each observation and the theoretical mean, provided clarity regarding the unequal dispersion of observations - an understandable result again given the outlier prices beyond the range of 0-4K dollars with a greater variance. To cure this issue, the box-cox plot explained how the observations should be transformed, producing a calculation approaching 0 and signifying the need to use a logarithmic transformation. Transforming data using a log transformation has the effect of dispersing tight pockets of data while tightening spread-out data. By performing the transformation to both price and carat variables, in effect instituting the Power Law, the reperformed simple linear regression produced an improved R^2 of ~95%.

Data Description

The data utilized in this analysis provides information on the price as well as the 4Cs (cut, color, clarity, carat) of 1,214 Blue Nile diamonds.Please note that the selected data is a subset of the original population provided via Kaggle.

```
##   carat clarity color       cut price
## 1  0.24    VVS2     G Very Good   379
## 2  0.40     VS2     H Very Good   605
## 3  3.35      IF     F      Good 56151
## 4  0.61     VS2     G     Ideal  1947
## 5  0.40    VVS1     J     Ideal   684
## 6  0.50     VS1     E Very Good  1553
```

Variable Description

Diamond Cut

This dataset contains information on diamonds with four distinct cuts: Good, Very Good, Ideal and Astor Ideal. According to the Blue Nile website, a diamond's cut, "refers to how well-proportioned the dimensions of a diamond are, and how these surfaces, or facets, are positioned" (e.g. ratio of diameter to depth). It is based on factors such as proportions, symmetry and polish. A diamond's cut is different from a diamond's "shape" because the cut is what results in more/less light reflection: An element which contributes to the diamond's quality. The scale is defined as follows:

- **Good:** Top 25% of diamond cut quality, reflects *most* light that enters the diamond, but not as much as the "Very Good" cut.
- **Very Good:** Top 15% of diamond cut quality, reflects almost as much light as the Ideal cut but for a lower price.
- **Ideal:** Top 3% of diamond cut quality, reflects *most* light that enters the diamond
- **Astor Ideal:** Crafted to gather and reflect the most light possible

For our purposes and for visualization, we re-factored the diamond cut to follow the order of the above cut scale.

Diamond Color

According to the Blue Nile website, part of a diamond's valuation is determined by the *absence* of color. Certified grading professionals determine the diamond's "color grade" according to an alphabetical scale. Blue Nile claims that diamond prices decline in alphabetical order (e.g. color grade G is less expensive than grade J). Color grades at Blue Nile range from letters "D" to "K" (7 distinct grades) as described below:

- **Faint color diamonds:** Budget friendly (Grade K)
- **Near colorless diamonds:** Great value for quality (Grades I-J, G-I)
- **Colorless diamonds:** Rare, highest quality (Grades D-F)

For our purposes and for effective visualization, we re-factored the color grade scale to follow reverse alphabetical order above (most budget friendly, to least budget friendly).

Diamond Clarity

According to the Blue Nile website, a diamond's clarity assesses the small imperfections on the surface (called blemishes) and within the diamond (called inclusions). Clarity is determined by five factors (size, number, position, nature, color and relief). Diamond clarity from Blue Nile spans 6 categories with 11 clarity grades and Blue Nile claims that higher clarity grades are reflected in a higher price. The clarity grades which Blue Nile sells described below from least to fewest noticeable imperfections:

- **SI2:** "Slightly Included", may be detected by unaided eye
- **SI1:** "Slightly Included", not quite detected by unaided eye
- **VS2:** "Very Slightly Included", easy to see at 10x magnification
- **VS1:** "Very Slightly Included", difficult to see at 10x magnification
- **VVS2 and VVS1:** "Very, Very Slightly Included", difficult even for trained eyes to see under 10x magnification
- **IF:** Internally flawless (some small surface blemishes)
- **FL:** Flawless (< 1% of all diamonds)

According to their website, Blue Nile claims that SI diamonds are the best value on Blue Nile's website whereas VS diamonds are supposedly the most popular diamond clarity. For our purposes and for effective visualization, we additionally re-factored the clarity scale to follow what Blue Nile claims would result in lower to higher prices.

Diamond Carat

A diamond's carat refers to the diamond's weight (not size which tends to be a common belief). According to our data, Blue Nile carats range from 0.23 to 7.09 carats. Interestingly, Blue Nile's website states that larger carat weights are not always better than smaller weights as weight is not related to sparkle (a factor of a diamond's cut). A diamond's quality, according to Blue Nile, should be focused on a balance against all the 4Cs. Regardless, Blue Nile points out that carat has the biggest impact on price which is manifested mostly by the media and society: Larger carats are oftentimes associated with higher status/wealth.

Exploratory Data Analysis

In order to assist our team in the investigation of the relationship between price and associated variables, we have initially conducted a series of exploratory data analysis steps to expand our understanding of the relationships and trends that exist. Additionally, we have set out to address the claims on Blue Nile's education page as discussed within our Variable Description section (i.e. cut is most important, color is second most important, clarity is least important). Before diving into visualizations, our teem deemed it necessary to summarize how many diamonds are in each category to further understand the sample size of diamonds from each category (displayed below):

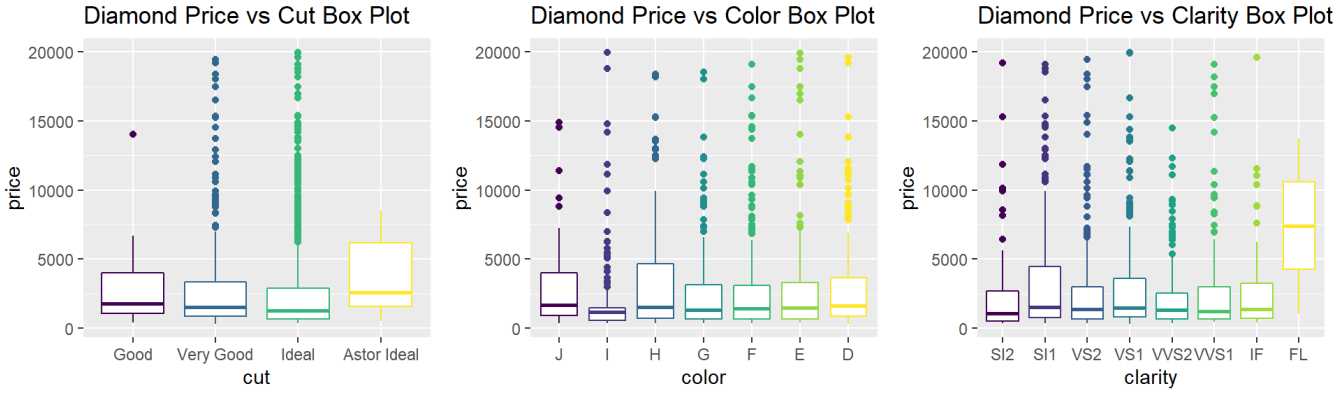
```
## # A tibble: 4 x 4
##   cut         count `average price` `median price`
##   <ord>       <int>           <dbl>          <dbl>
## 1 Good           73           9467.          1903
## 2 Very Good     382           7758.          1744.
## 3 Ideal         739           6489.          1354
## 4 Astor Ideal    20           5852.          2854
```

```
## # A tibble: 7 x 4
##   color count `average price` `median price`
##   <ord> <int>           <dbl>          <dbl>
## 1 J        90           3934.           1803
## 2 I       167           4779.           1212
## 3 H       148           7799.           1575
## 4 G       198           4572.           1374
## 5 F       223           6205.           1485
## 6 E       181           9906.           1602
## 7 D       207          10525.           1781
```

```
## # A tibble: 8 x 4
##   clarity count `average price` `median price`
##   <ord>   <int>           <dbl>          <dbl>
## 1 SI2       165           2626.           1129
## 2 SI1       243           5073.           1655
## 3 VS2       214           7726.           1472
## 4 VS1       233           9078.           1594
## 5 VVS2      158           7334.           1390.
## 6 VVS1      149           8667.           1392
## 7 IF         49           6362.           1427
## 8 FL          3         123403.          13733
```

Upon analysis of our tabular results, the claims by Blue Nile that VS diamonds are the most popular diamond clarity appear confirmed given the count volume. Strictly based on volume of diamonds, VS has the highest count per clarity category. The initial observation by Blue Nile though regarding best value associated with SI diamonds appears far less clear. By analyzing this statement purely based on average price it would seem correct although, as will be revealed in the following sections, price is right skewed due to significant outliers. By viewing price through the median price per clarity category, claiming best value associated with SI seems far less clear.

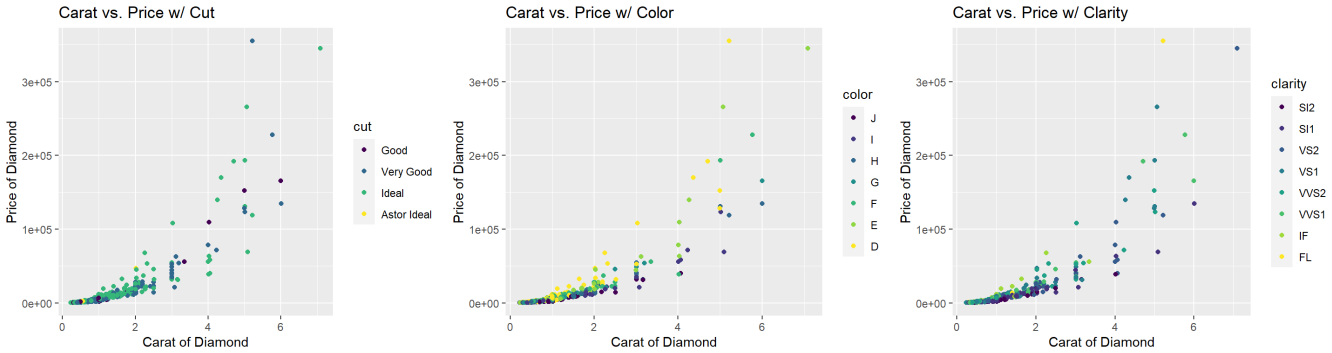
Price Variance per Attribute



In our initial view of the relationship between price and color/ clarity/ cut, there is clear variation across each which makes us less confident in making any initial claims. Understanding that cut has fewer categories, we see fewer outliers in our box plots leading to an initial assumption that perhaps a closer

relationship exists with price than with price and color/ clarity as would align with Blue Nile's 'education'. The outliers are certainly worrying given that there could be more significant steps to pricing with color and clarity that are less obvious. Please note that the price variables (y-axis) has been capped for the sake of the box plot visuals at 20K due to significant outliers which influenced the interperability of the visuals. For the purposes of the model, no data points have been removed.

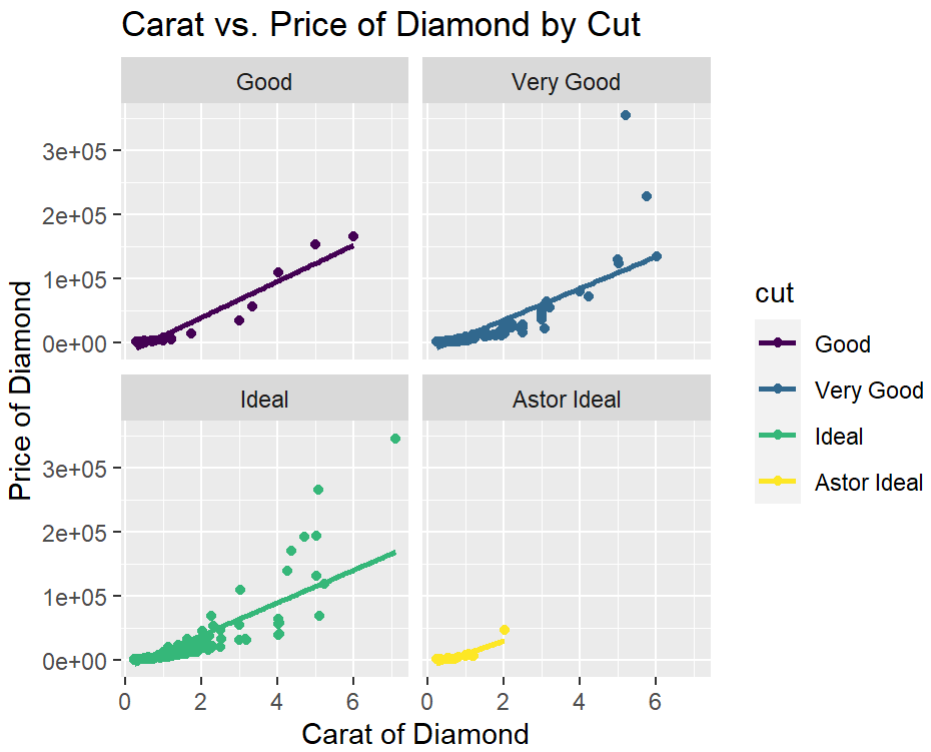
Price vs Carat per Attribute

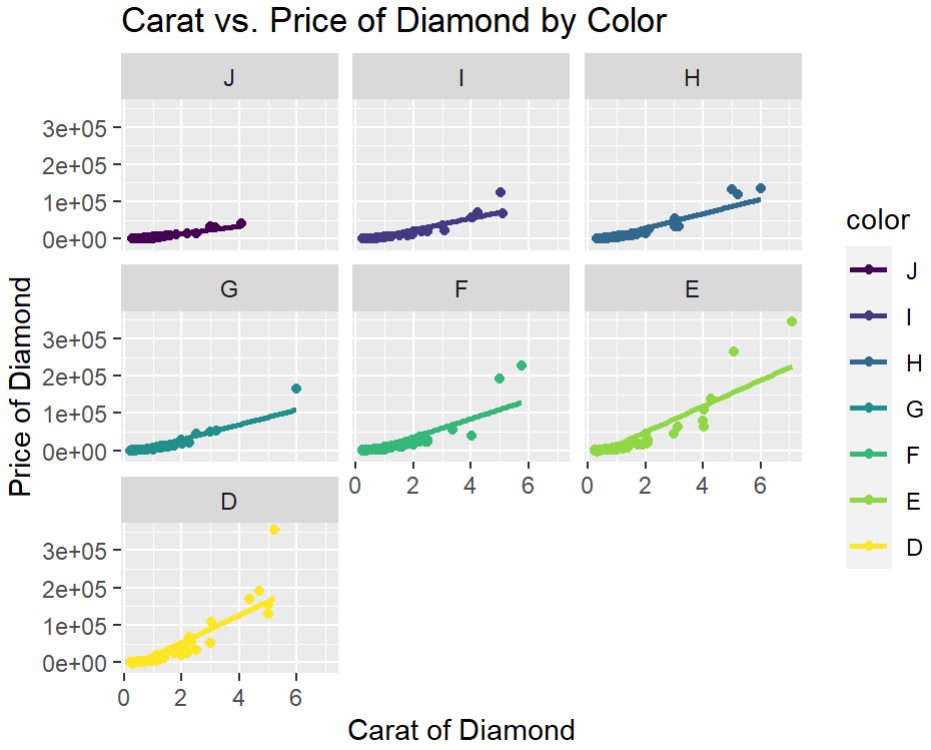


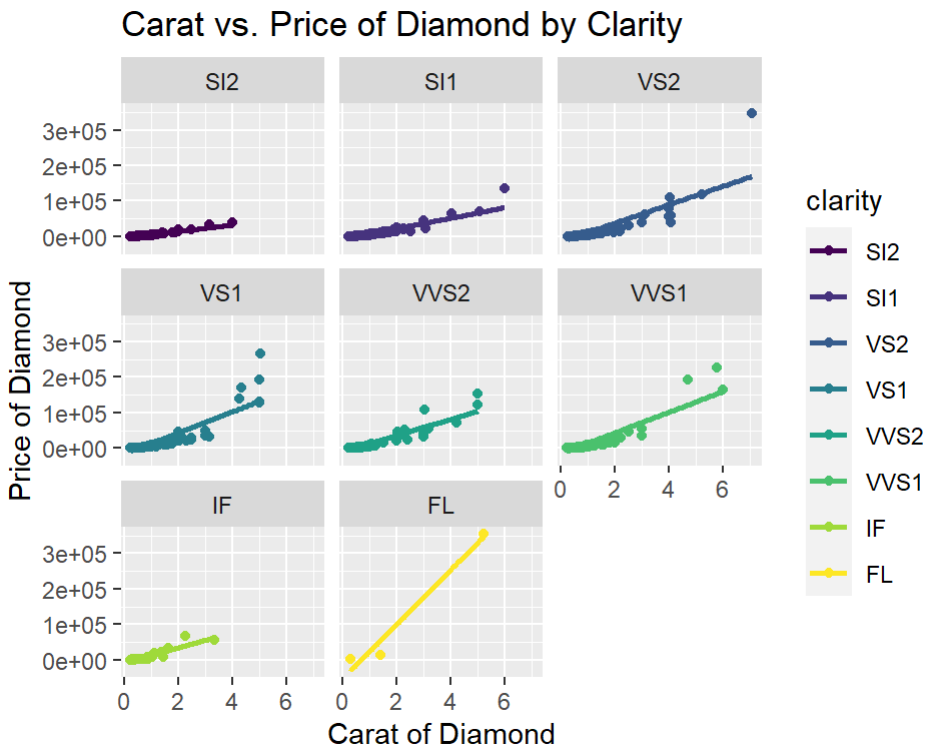
To investigate this relationship between price and carat further, our team sought to view any apparent interaction effects between a third, categorical variable. With the base of each scatter diagram based on price and carat, we can see some lighter relationships by adding in color (second most important variable). There appears to be a constant trend where those diamonds of the type 'D' (designating absolutely colorlessness) seem to consistently outperform (in terms of price) those of type 'E' or 'F', which in turn outperform those grouped in more purple to blue colors of types 'G', 'H', 'I', 'J'. This does trend appropriately based on the knowledge gleaned from Blue Nile's 'education' page, although we can still see many instances where the relationship does not hold. For instance, between carats of 1 and 3.5, there are many observations where lesser colors are priced similarly to those of greater color values. We would need to investigate further interaction effects before stating that these are indeed mispriced diamonds.

Introducing clarity and cut to the price vs. clarity visuals does not prove to provide any further clarification alone in trends at this time.

In order to further examine how cut, color, and clarity effect the relationship between carat and price our team separated the graphs above to see how each factor within the categorical variables affected the relationship between price and carat.


Carat vs. Price of Diamond by Cut
Good
Very Good
Ideal
Astor Ideal
Price of Diamond
3e+05
2e+05
1e+05
0e+00
0
2
4
6
Carat of Diamond
cut
Good
Very Good
Ideal
Astor Ideal

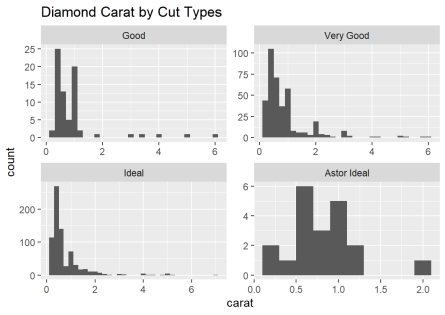

Carat vs. Price of Diamond by Color
J
I
H
G
F
E
D
Price of Diamond
3e+05
2e+05
1e+05
0e+00
0
2
4
6
Carat of Diamond
color
J
I
H
G
F
E
D

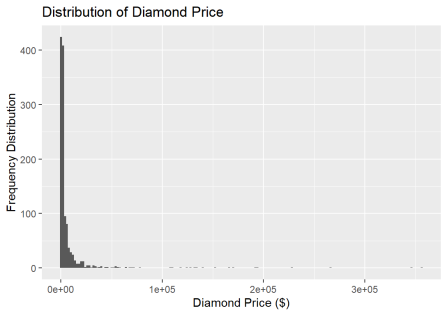


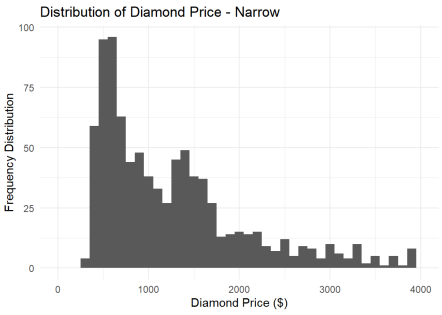
The diamond color graph divided by category confirms our findings from the previous scatterplot where the we found that the more colorless a diamond is the stronger the relationship between price and carat.

Based on the clarity graphs broken up by category we were able to see that, at least for the first few clarity grades, as clarity increases the realtionship between price and carat increases. However, this trend does not continue after the first few grades. The cut graphs still do not provide any further clarification of trends.

Distribution of Attributes







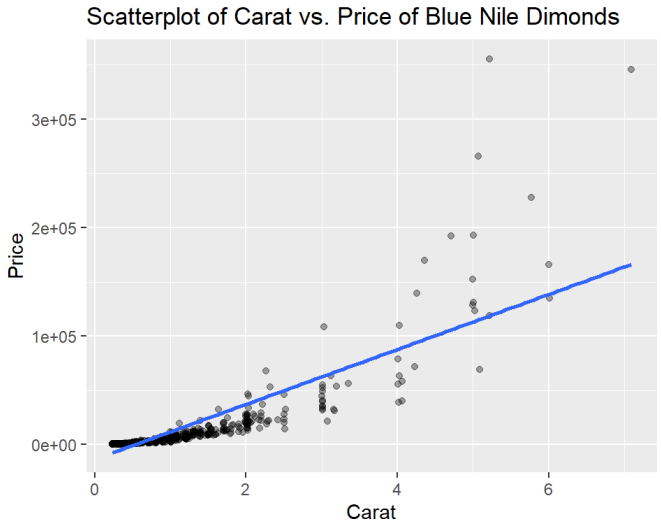
To further investigate cut from our initial EDA visual where we started to see a slight relationship with price, a histogram was developed to view distribution by type. What our team will need to take into account when looking at this relationship is the uneven distribution of cut types (far more 'Ideal' types than any other). When visualizing price vs. cut earlier, we recognized in initial observations that far fewer overall outliers existed which may mean a closer relationship to price. Now this is seeming far less certain given the uneven distribution of categories, especially since 'Very Good' and 'Ideal' had a greater number of outliers while also having a greater number of observations.

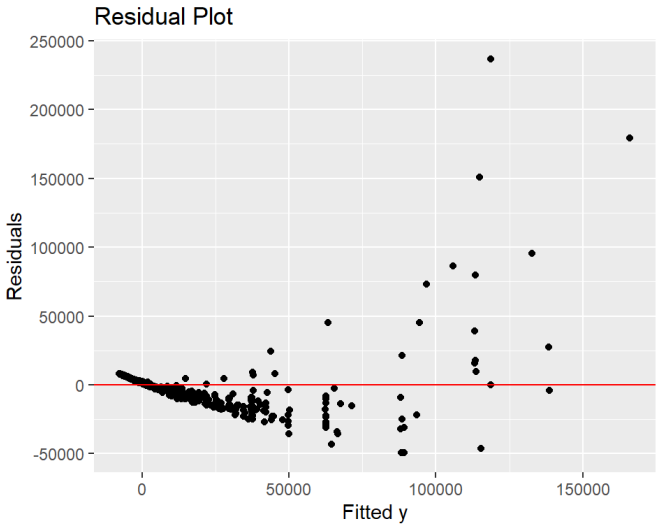
Finally, to get a macro view of price distribution, a histogram was developed and further focused to narrow in on those observations between a price of 0 and $4000. This allows us to understand that outside of those outliers of significant value, the majority of our observations within the regression will come from the lower

price end. Please note that for the initial distribution visual (labelled Diamond Carat by Cut Types), the y-axis for the four visuals have differing thresholds - great care should be taken when performing a comparison based soley on these four visuals.

Regressing Price on Carat

Examining the relationship between price and carat, our team has run a simple linear regression (SLR) between our predictive variable (carat) and our response variable (price). The first step in performing the SLR is to examine the relationship between variables on a scatter plot to confirm whether or not a linear model is appropriate to fit our data.

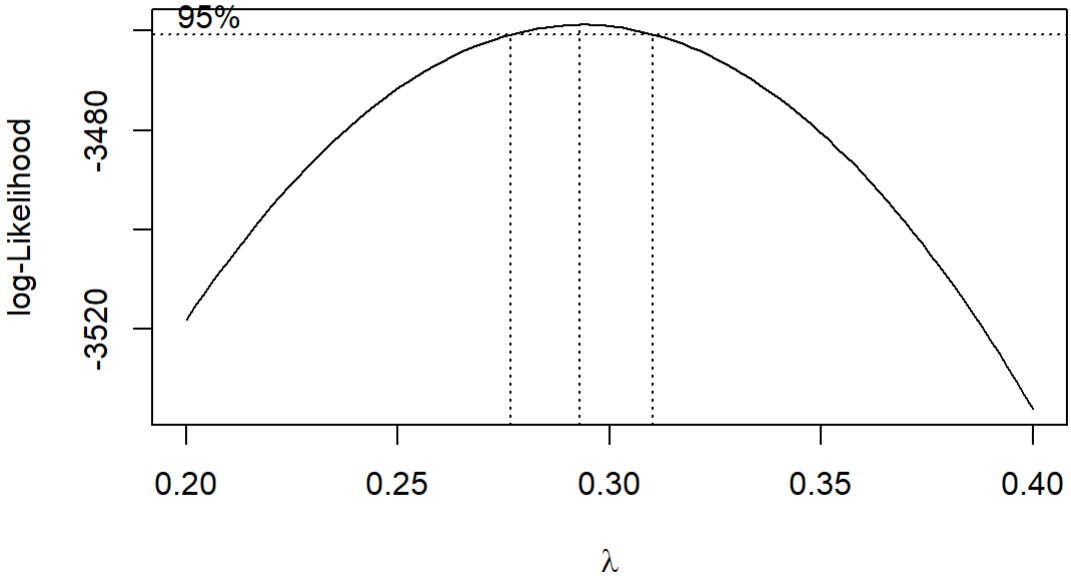




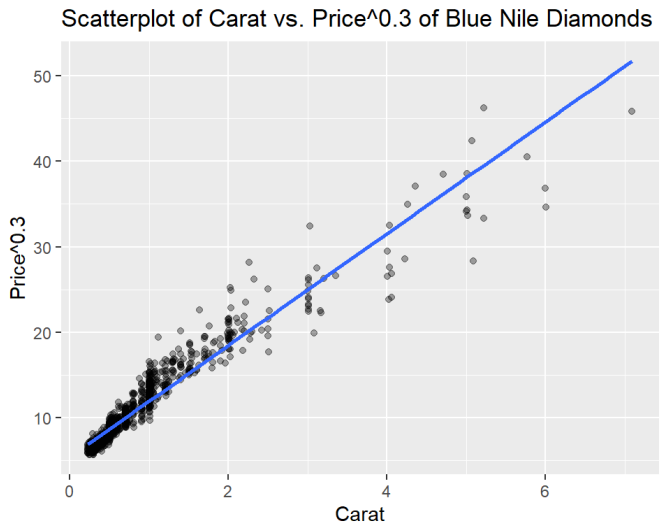
In viewing the initial scatter diagram, there is clearly a positive relationship between values which supports the prior Blue Nile ‘education’ page. The relationship starts very strong although as can be seen from the linear model line and emphasized in the residual plot, as carat size grows the relationship with price becomes far more varied. This supports the types of interaction effects we were seeing throughout our EDA visuals previously.

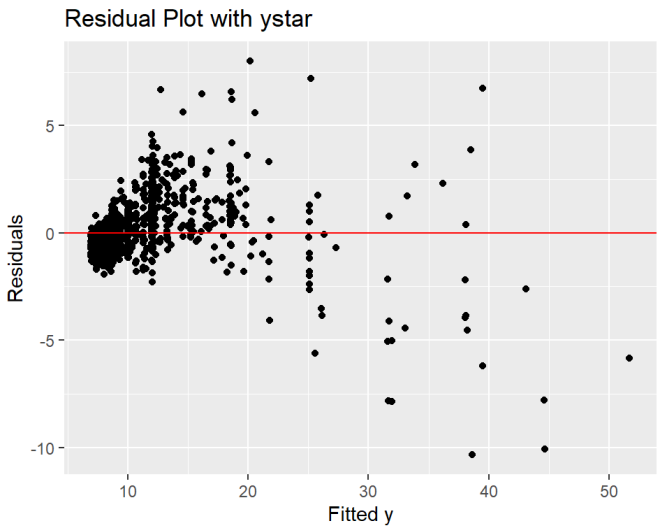
Given the variation of observations having a more curved nature in comparison to the linear model as carat size increases AND the observations fanning as carat size increases, we can say initially that assumptions 1 (mean zero errors) and assumption 2 (constant variance) are not met. As such, given the current state of the data, it would not be appropriate to fit a linear model and develop a regression line as it exists currently.

To examine the direction we should take with transforming the data to better construct observations that we feel confident in fitting a model to our team will construct the box-cox plot and first focus on the constant variance assumption.

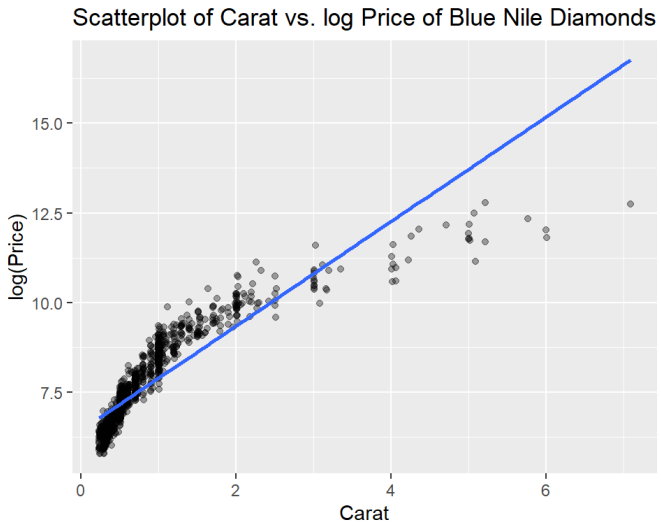


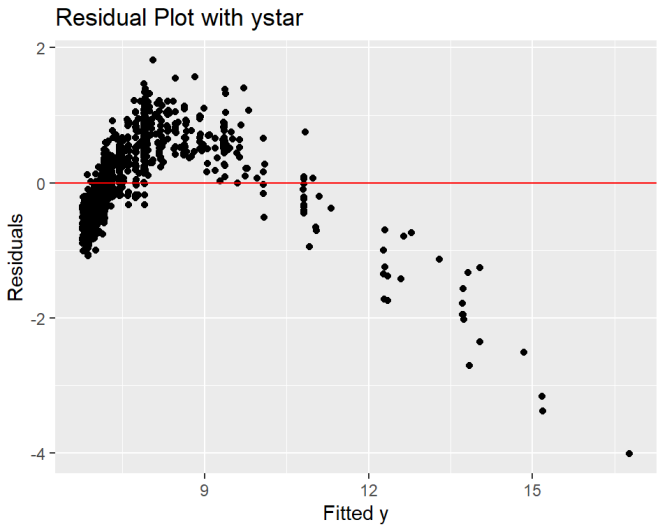
Based on the box-cox plot displayed, we chose to use a lambda of 0.3 to transform y as a starting point. This decision is based on knowledge that the 0.3 falls within the 95% CI for the box-cox plot and for a lambda value of less than 1 but greater than 0, one should choose a value approaching 0 within the CI.





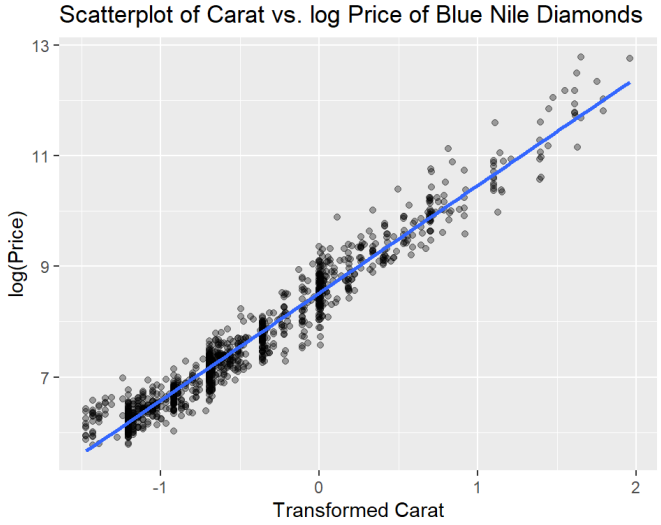
Clearly, the lambda = 0.3 transformation did not reduce the increasing variance enough to satisfy the regression assumptions according to the residual plot. Therefore, our team as a next step chose a lambda further from 1 (closer to) and less than 0.3, increasingly closing in on 0. As such, we will try the log transformation next.

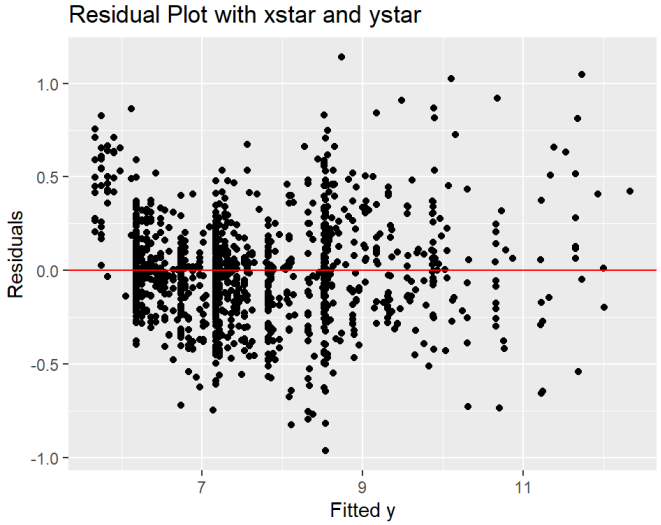




After using a log transformation, the constant variance assumptions appears to hold. This is a positive indication that the log transformation is the appropriate transformation to assist in the failed assumption 1 and 2 observed previously.

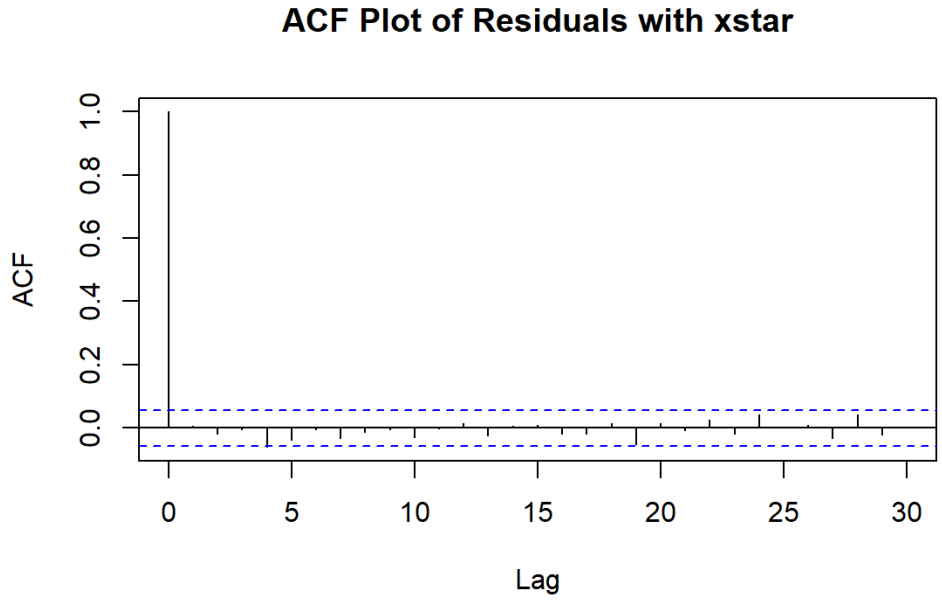
Next, our team examined the transformation of our predictive variable x (carat) in order to satisfy the first regression assumption and flatten the curved pattern in the residual plot for a tighter regression line. If our team were to apply the log transformation to each variable, the Power Law would be in effect (please see further details in next section).

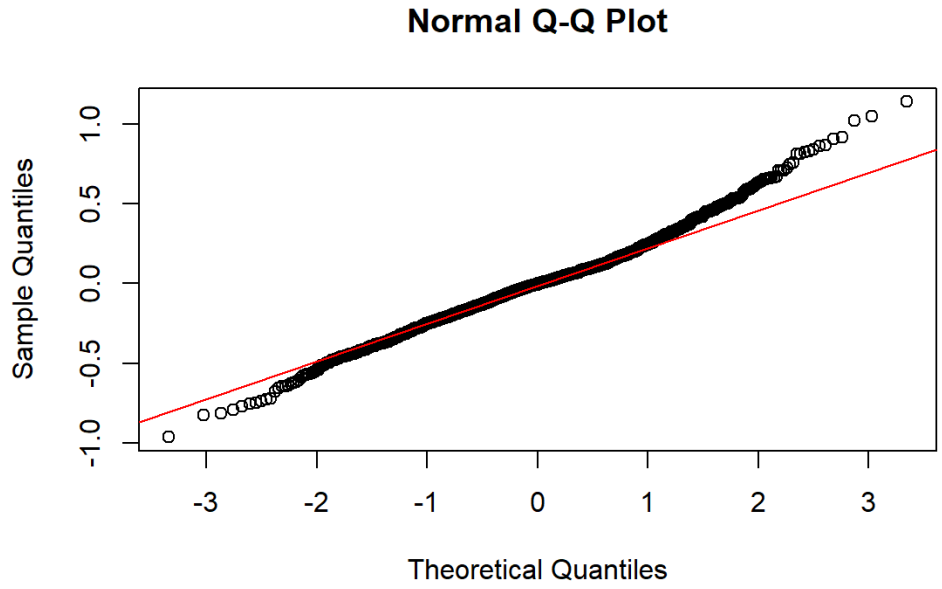




Following the transformation of the predictive and response variables using the log function, we can see that the linear relationship (left chart) is far tighter, removing the prior curve that was prevalent with the higher priced diamonds or those with a greater carat size. Similarly, the variance issue where the distance between observations to the theoretical mean line was inconsistent is far more consistent following the transformation (right chart). Given these transformations we are confident that applying the Power Law to the data has transformed the x and y values to a form appropriate for regression calculations. For awareness, the usage of the Power Law is the functional relationship between the two variables indicating that a relative change in the predictive variables leads to a proportional relative change in the response variable (response variable varies as power of predictive variable).

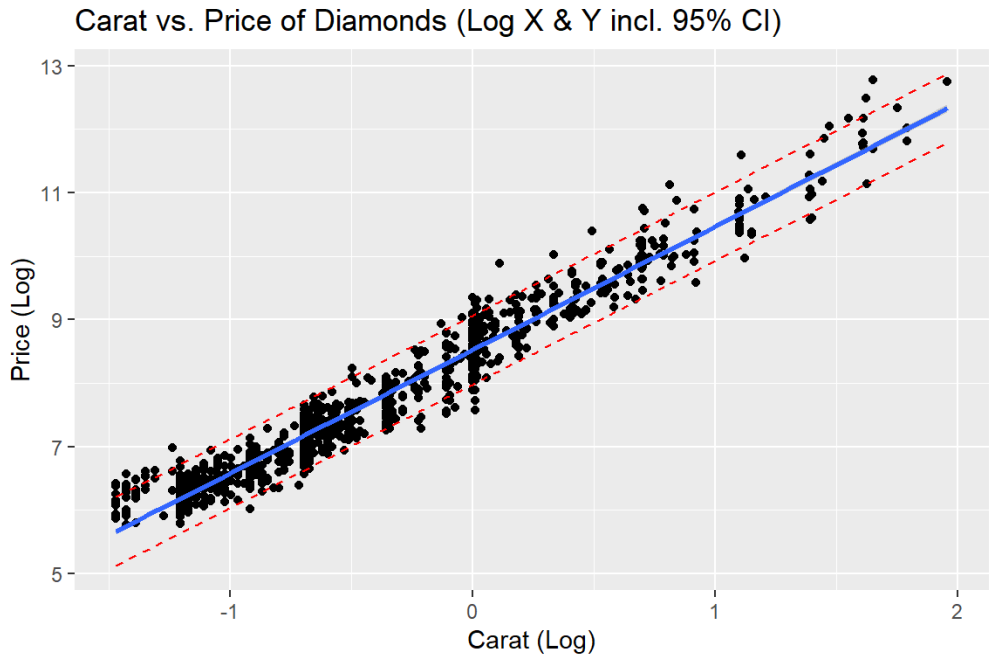
Following the log transformation, our team has confidence in producing the autocorrelation plot to indicate the existence of uncorrelated residuals as well as the Normal Q-Q plot tracing our transformed observations against the theoretical representation of the expected value under normality.





In viewing the ACF plot, although multiple variables touch the critical line around lags 4 and 18, we concluded that this is not a drastic impact to the model and does not prevent us from further progress. Additionally, from viewing the Normal Q-Q plot, one may see that on either end the actual transformed observations trail off from the theoretical line. This again did not trigger any flags given the level of variance remaining from the data.

Finally, to conclude the regression analysis our team mapped the 95% confidence interval against the newly regressed scatterplot of log of price against log of carat. We can be 95% confident that a new Blue Nile diamond, when assessed by carat size, will fall within the below threshold defined by the red dotted lines.



Commentary on Regression Assumptions

- Sample - The sample taken of the original Kaggle Blue Nile data set in order to develop the given data used for the project was produced randomly.
- Independence - Observations within the sample dataset are independent of one another.
- Linearity - The relationship between the predictive variable (x) and the mean of the response variable (y) is linear.
- Homoscedasticity - The variance of the residual is the same for any value of the predictive variable.
- Normality - Observations of X and Y are normally distributed following the use of the Power Law to apply a log function to each of the variables of interest.
- Interaction Effects - Linear regression ignores further interaction effects between price and variables not carat.

Model Output

In order to inform our model interpretation, our team produced a regression summary and an ANOVA table of the initial results prior to transformation as well as with our proposed log transformations. Although we understand the regression assumptions were not met for the initial model (non-linear relation with increasing variance) we have included both versions of our model below for the purpose of comparison only.

Regression output of the *initial* (no-transformation) model:

```
##
## Call:
## lm(formula = price ~ carat, data = Data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49375  -5048   1867   4965 236711
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13550.9      559.7  -24.21   <2e-16 ***
## carat        25333.9      494.4   51.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13560 on 1212 degrees of freedom
## Multiple R-squared:  0.6842, Adjusted R-squared:  0.6839
## F-statistic:  2625 on 1 and 1212 DF,  p-value: < 2.2e-16
```

ANOVA table of the *initial* (no-transformation) model:

```
## Analysis of Variance Table
##
## Response: price
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## carat        1 4.8290e+11 4.8290e+11  2625.3 < 2.2e-16 ***
## Residuals 1212 2.2294e+11 1.8394e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Although we are unable to interpret this model in context due to the fact that the regression assumptions do not hold, we found it interesting to report that the R^2 was 0.68 suggesting that this model (if the assumptions were to hold) would explain 68% of the variation in price using a diamond's carat.

Regression output of the *transformed* model:

```
##
## Call:
## lm(formula = price_log ~ carat_log, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41863 -0.07484 -0.00109  0.06403  0.49551
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.700714   0.004228   875.4   <2e-16 ***
## carat_log   1.944020   0.012166   159.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1199 on 1212 degrees of freedom
## Multiple R-squared:  0.9547, Adjusted R-squared:  0.9546
## F-statistic: 2.553e+04 on 1 and 1212 DF,  p-value: < 2.2e-16
```

ANOVA table of the *transformed* model

```
## Analysis of Variance Table
##
## Response: price_log
##             Df Sum Sq Mean Sq F value    Pr(>F)
## carat_log    1 367.18  367.18   25535 < 2.2e-16 ***
## Residuals 1212  17.43    0.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The regression output above indicates that by introducing the Power Law to transform the response and predictive variables, we enhanced our model from being able to explain only 68% of the variation in the regression to explaining 95% of the variation in the regression using only variables price and carat. As described above, the regression assumptions are met for this version of our model. Therefore, we will move forward with the interpretation of this well-performing model in the next section.

Final Model and Interpretation

The regression equation of the transformed observations can be summarized as follows:

$$price^* = 3.701 + 1.944(carat^*)$$

where:

$$price^* = log(price)$$

$$carat^* = log(carat)$$

In the case of simple linear regression, both an ANOVA F-test and a t-test for the slope parameter give the same result. Both test whether the slope parameter is different from 0 as described in the hypotheses below:

- $H_0 : \beta_1 = 0$
- $H_a : \beta_1 \neq 0$

As expected, F-test and t-test both resulted in a p-value of 2.2e-16. We can therefore, reject the null hypothesis that the slope parameter is equal to zero and conclude there is significant evidence that there exists a linear relationship between $log(carat)$ and $log(price)$ of Blue Nile diamonds. Because we performed a log transformation to both the predictor (carat) and response (price), we can be confident in our interpretation of the regression coefficients below:

- **Intercept** ($\hat{\beta}_0$): Because the predictor is on a log-scale, the intercept tells us the expected response when the predictor is 1 (since $log(1) = 0$). Therefore, from our estimated $\hat{\beta}_0$, the estimated $log(price)$ when a diamond has 1 carat is 3.701.
- **Slope** ($\hat{\beta}_1$): For a 1% increase in carat expected price increases 1.944%.

Futher, our team constructed a 95% confidence interval to assess our confidence in the model's estimate for $\hat{\beta}_1$ as shown below.

```
##    2.5 %   97.5 %
## 1.920152 1.967888
```

From the interval above, we are 95% confident that the true percent price increase associated with a 1% increase in carat weight is between 1.920% and 1.967%.