Blue Nile Diamond Analysis

Alexandra Ferentinos, Mohammad Farooq, Peter Landers, Shelby Laychak

2023-10-22

Introduction:

Engagements, anniversaries, and birthday celebrations are times of joy and traditionally entail the gift of a precious ring. Rings are often adorned with precious gems or intricately designed bands of precious metals. However, the most popular and valuable gem to adorn these pieces of jewelry are diamonds. Diamonds are precious stones that are sought after for their dazzling brilliance, inherent beauty, and their trend in holding monetary value. These stones garner their worth from the 4 C's: (as coined by diamond distributors as a criteria of diamond quality grading, i.e Blue Nile) Cut, Color, Clarity, and Carat. This combination of characteristics influences the purchase price of these precious gems we well as the popularity amongst gems purchased of a certain variety of the 4 C's.

A dataset from the Blue Nile Company, as well as other internet diamond distributors, containing over 1,000 diamonds for sale listing their selling price along side their graded 4C characteristics. These variables yield interesting statistical nuances when compared to how the 4C variables are related to one another, as well as selling price. After creating data visualizations of the data, it is clear to see what those relationships are and to explore the claims made by Blue Nile Company. Firstly, 'Astor Ideal' cut diamonds neither fetch the highest prices on average, nor do they represent the most expensive diamonds in Blue Nile's collection. This claim is not supported by the data as the 'Astor Ideal' fetches the highest.

Secondly, the data support that diamond clarity is the least important factor in the 4C's influence on pricing. When comparing the different clarity gradings to the diamonds price there is not outstanding clarity type that influences the price, excluding a Flawless clarity as it is rare and flawless clarity are valued at exorbitant prices for the way light refracts in the stone. Additionally, in terms of diamond color D and E are claimed to be the most expensive of the color varieties. From assessing the color to price of the average price by color D and E are in fact the most expensive. This is understandable as D and E color of diamonds affect the most brilliant dazzle when in contact with natural light, making them quite popular and driving the price up. From the conclusion made from clarity is the least import factors in the 4C's the data additionally supports that VS and SI clarity diamonds are the best value. These clarity types are most purchased for diamond pricing, asserting they are the most popular of the clarity grades.

Finally, fitting a linear regression regarding variables price against carat involved manipulation of the x and y variables. Logarithmic transformations were performed on the variables as it was most comprehensible for the coefficients. After these transformations were completed, improvement was seen regarding the diagnostic plots and therefore asserting that the required assumptions were met for this set of data transformations. The overarching conclusion from price against carat analysis is that there is a linear relationship between price and carat after after logarithmic transformations, such that as carat increases price does as well.

Variable Descriptions:

The price of a diamond is based of the 4Cs - carat size, clarity, color and cut. The 4Cs are evaluated based on the standardized grading scale GIA, AGSL, and GemEx. The dataset being used has 1214 observations having five features (price, carat, clarity, color, and cut).

One of the five features includes the carat variable refers to the measure of a diamond's weight. Diamonds with higher carat weight are cut from larger crystals which means that it makes them harder to source than a smaller crystal. When referring to its relationship with price, carat price depends on how rare the diamond is. In our data, carat weight can range from 0.23 to 7.09 carat.

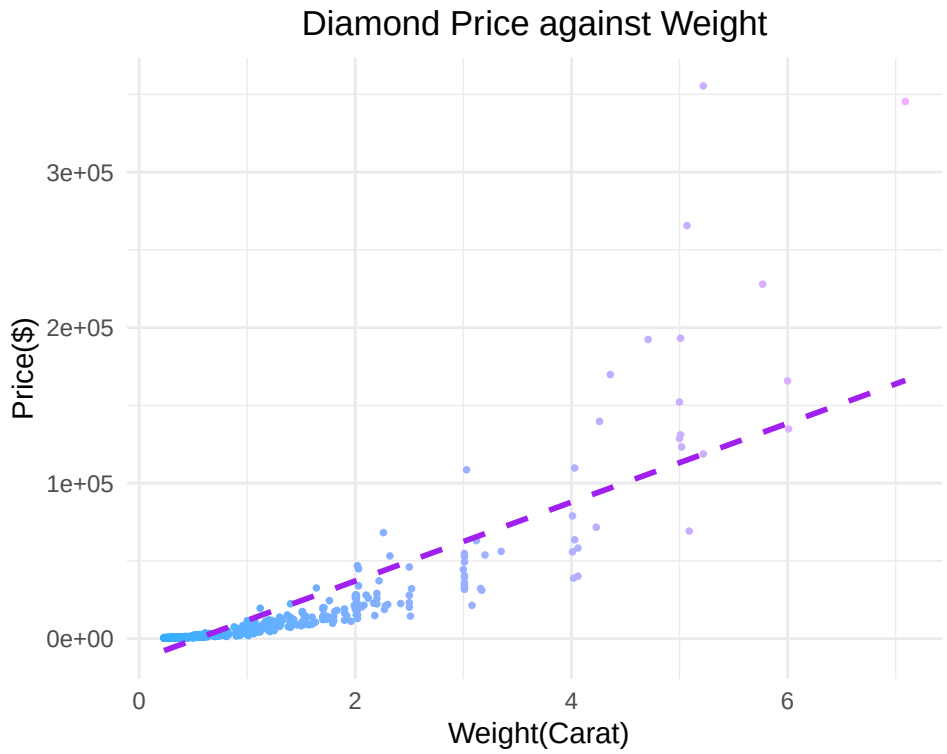
The variable clarity describes the small imperfections within a diamond. Examples include surface flaws or internal defects called inclusions. These inclusions can occur naturally during the diamond forming process. There are five clarity factors: size, number, position, nature, color and relief. Clarity's grading scale includes SI2, SI1, VS2, VS1, VVS2, VVS1, IF, FL in ascending order. SI1, SI2 refers to the group Slightly Included diamonds which means that the inclusions are noticeable by eye. VS1, VS2 refers to Very Slightly included diamonds which means that there are minor inclusions. VVS1, VVS2 refers to Very, Very Slightly included diamonds which means that the inclusions are difficult to see for trained eyes under 10x magnification. They are rare and need an eye clean appearance. IF refers to internally flawless diamonds which means that there might be some surface blemish, but have no inclusions within the diamond. FL refers to flawless diamonds which means that there are no internal or external characteristics. These diamonds are rare to find, less than 1% of diamonds are flawless clarity.

As for a diamond color, it refers to how colorless a diamond is. The less color, the higher the grade and the higher the price. The color scale in this dataset is J (lowest), I, H, G, F, E, D (highest). D, E, F diamonds are colorless which are the rarest and highest quality. Near-colorless diamonds are G, H, I, J with no discernible color.

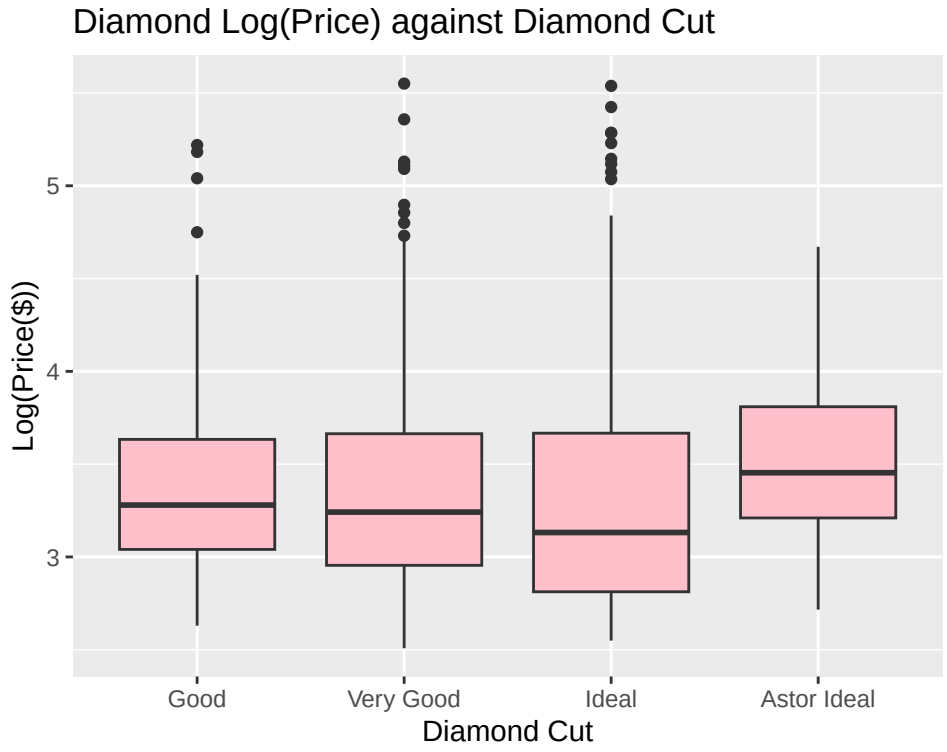
The last C of the 4Cs is cut which measures how well-proportioned a diamond's dimensions are. Cut's grading scale is 'Good' (lowest), 'Very Good', 'Ideal', 'Astor Ideal' (highest). 'Astor Ideal' refers to reflecting the most light as possible. Ideal reflects most of the light. Very Good refers to less than ideal but still reflects light while Good refers to reflecting less than Very Good but still some light.

Variables that were added to the dataset include averages and median values for each predictor variable. This was done to address the claims made by the Blue Nile Company. In order to have improved visualizations, the price variable was adjusted by taking a log of the data values. This allows us to better visualize our box plots seen below:

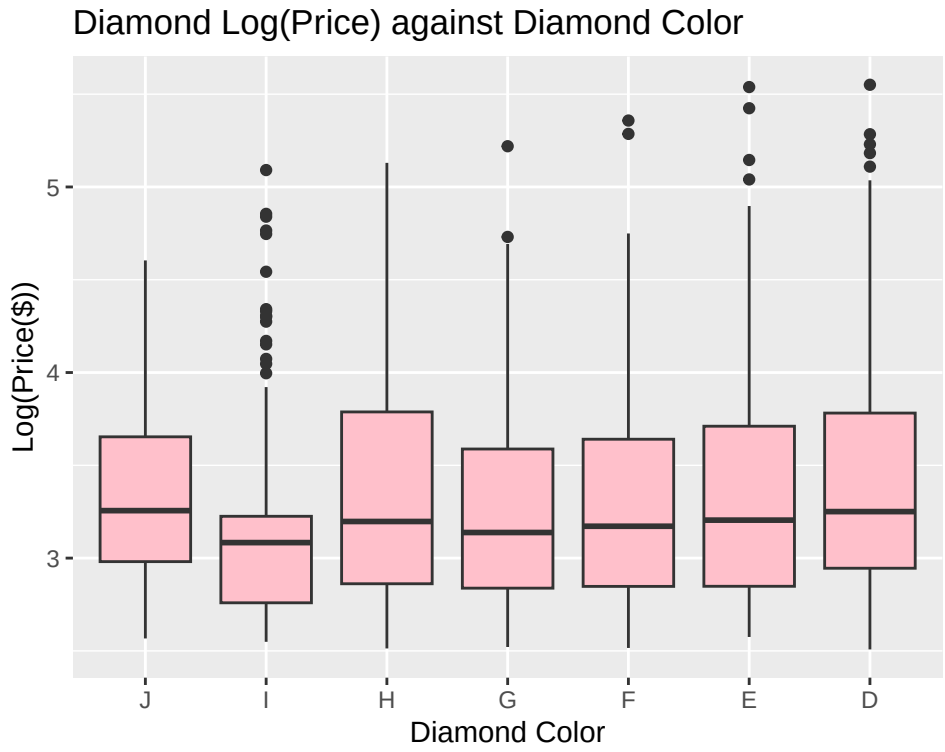
The following scatterplot shows price against weight. This figure shows that an increase in weight has a positive, exponential relationship with price with the most popular diamond weights are those less than 2 carats.



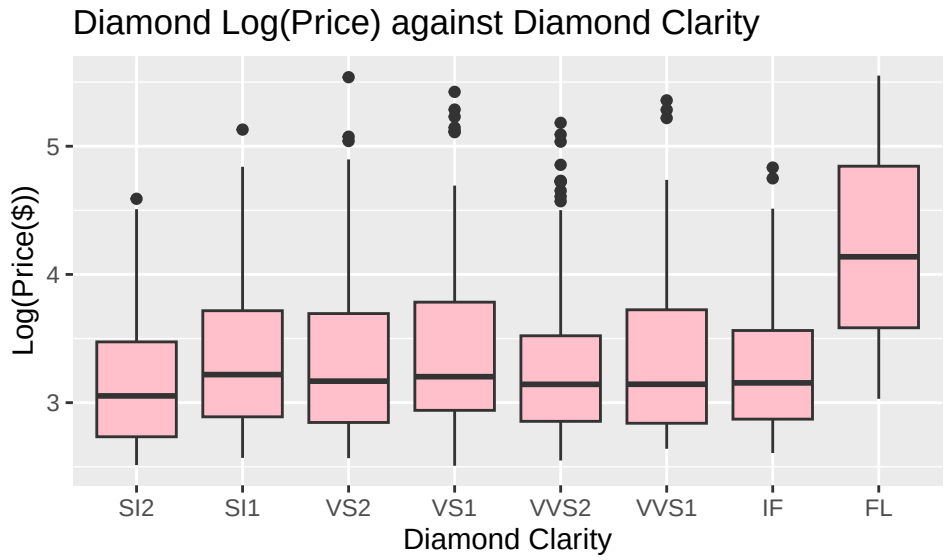
The following figure is a boxplot of price against diamond cut. It shows that 'Ideal' cuts have the highest prices compared to the other categories.



The following figure is a boxplot of price against diamond color. It shows that D and E have the highest prices compared to the other colors.



The following figure is a boxplot of price against diamond clarity. It shows that flawless (FL) diamond clarity has the highest prices.



Claims made by Blue Nile Company

Blue Nile makes several claims on the education pages regarding the "4C's" (carat, cut, clarity, and color) of diamonds on their website including:

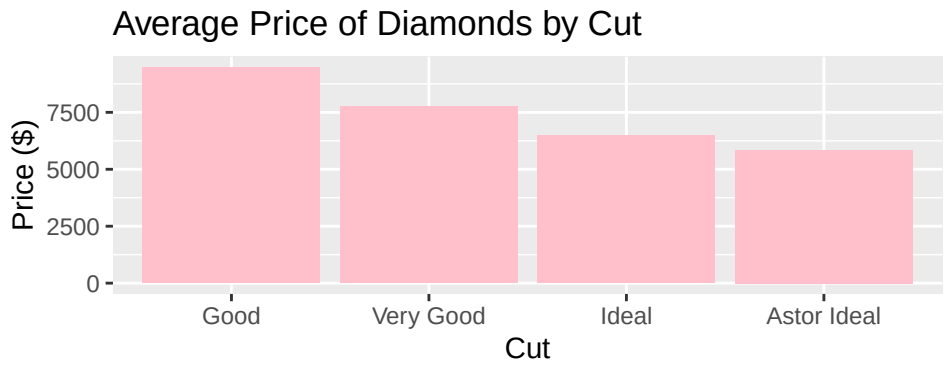
1. 'Astor Ideal' cut diamonds represent the highest quality and therefore fetch the highest prices.
2. Clarity is the least important factor of the 4C's.
3. Diamonds of the colors E & D are the most expensive.
4. 'SI' and 'VS' diamonds are the best value.

Further analysis supports claims 2 and 4 while opposing claims 1 and 3.

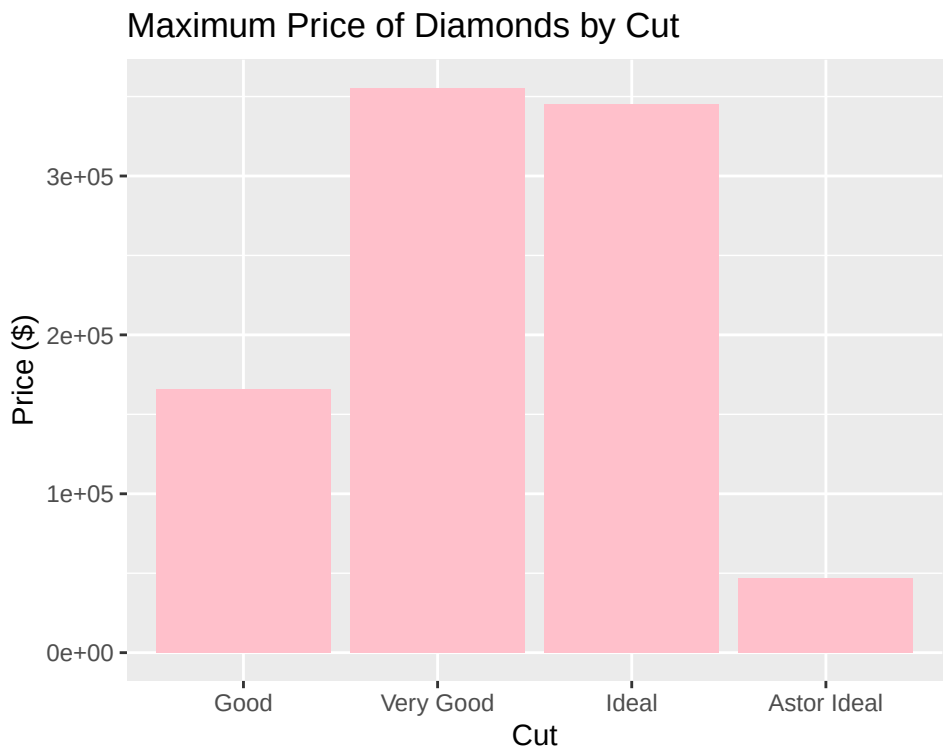
Claim 1:

Cut is a measure of the quality of a diamond's facets and how they support its ability to reflect light. Blue Nile offers four qualities of cut: 'Good' reflecting most light entering (top 25%), 'Very Good' reflecting as almost as much light as 'Ideal' for less cost (top 15%), 'Ideal' (top 3%), and 'Astor Ideal' reflecting the most light (Highest Quality).

First, Blue Nile claims that their self-branded 'Astor Ideal' cut of diamonds represent the highest quality of diamond and are therefore priced to fetch the highest prices. To address this claim, one must review the impact of cut on the price of their diamonds. Logically, one would assume that as cut quality increases the average price of the diamonds should increase as well if Blue Nile's claim is true.



The preceding histogram shows the average price of a Blue Nile diamond against its cut quality. As quality increases, the average price of diamonds decrease. This suggests that Blue Nile's claim is false, however, one could argue that Blue Nile's intention is that the most expensive diamonds in their collection are 'Astor Ideal' and many lesser diamonds that happen to be 'Astor Ideal' exist as well.

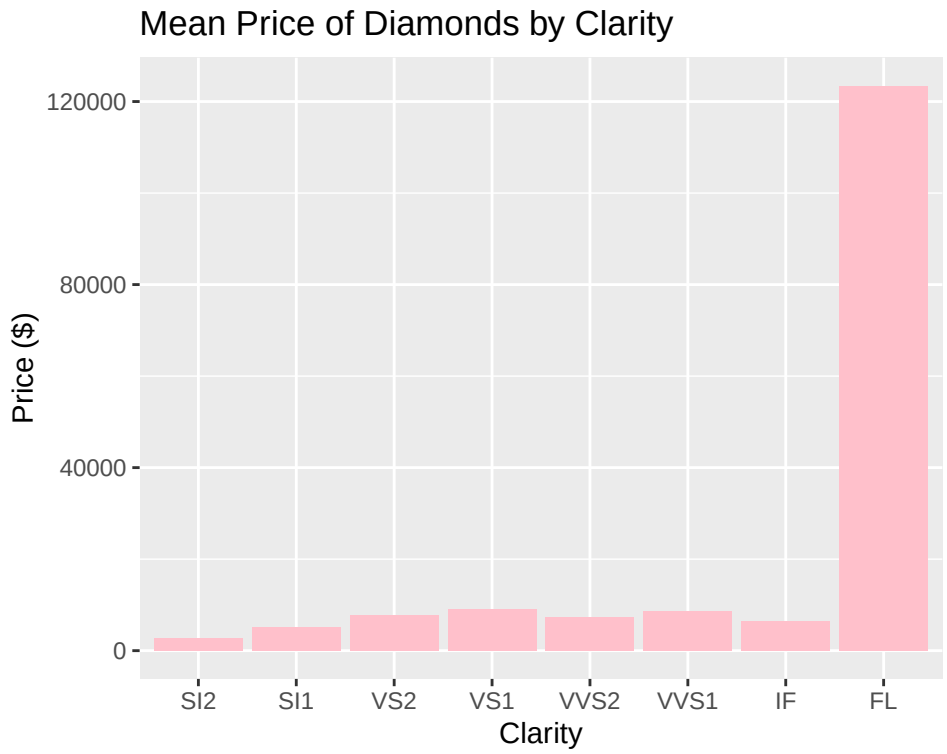


The graph above shows the maximum price of Blue Nile diamonds against their cut quality.'Very Good' and 'Ideal' cut diamonds tower over the maximum price of 'Astor Ideal' diamonds by about 60x. Blue Nile's claim that 'Astor Ideal' diamonds fetch the highest prices must be false.

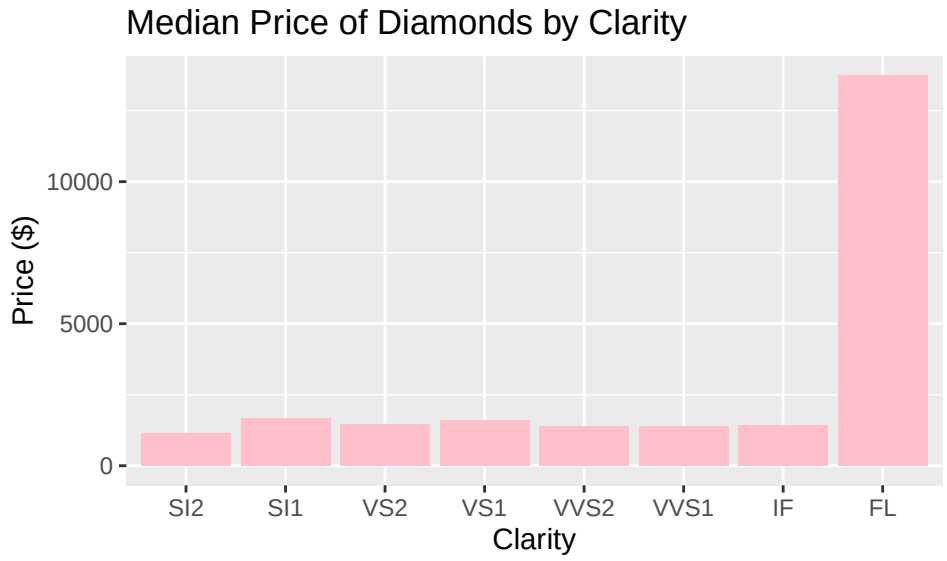
Claim 2:

Clarity is an assessment of flaws or inclusions within a diamond. Blue Nile's collection ranges from Slightly Included (SI2,SI1), Very Slightly Included (VS2,VS1), Very Very Slightly Included(VVS2, VVS1), Internally Flawless (IF), to Flawless (FL). SI diamonds have inclusions visible at 10x magnification, VS and VVS diamond inclusions are more difficult to see at 10x magnification, IF diamonds may have slight surface blemishes at 10x, and FL diamonds have no visible internal or external inclusions. FL diamonds make up less than 1% of all diamonds.

The Blue Nile Company claims that clarity is the least important of the 4C's in terms of price. This translates to the expectation of a marginal change in average price as quality of clarity increases or decreases.



The figure above shows the average price of a Blue Nile diamond against its clarity. Price tends to increase from SI2 to VS1 then oscillates until FL where it skyrockets. Due to the rarity of FL diamonds (<1% of all diamonds), it can be assumed their pricing is an outlier. On average (bar 'FL'), clarity appears to only result in a marginal change of price supporting Blue Nile's claim.

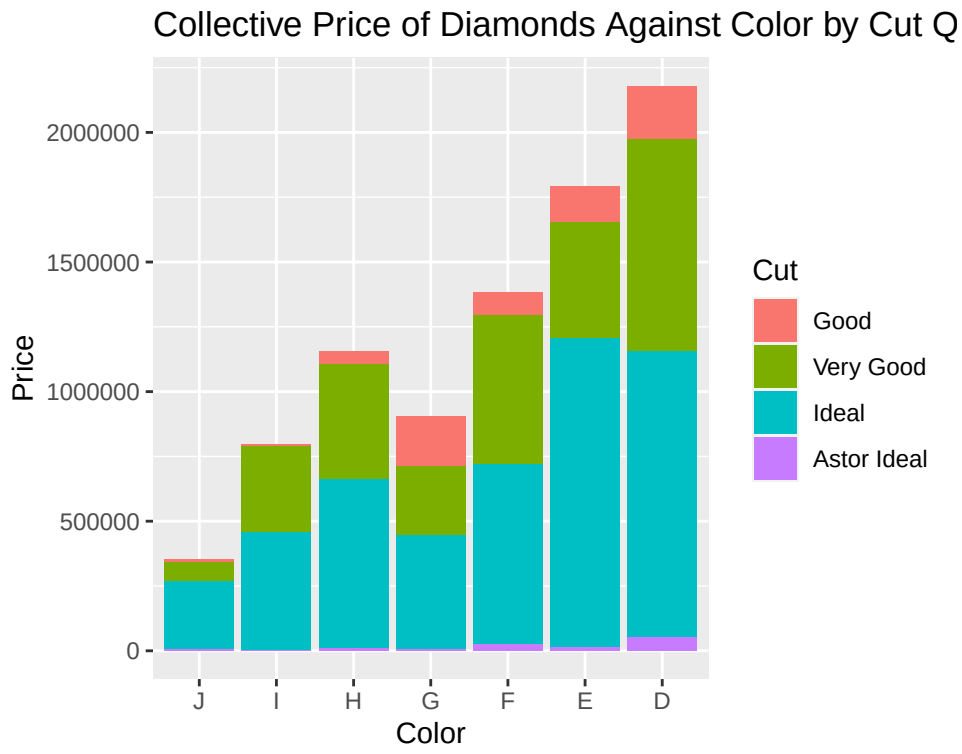


Assessing the median of diamonds by clarity shown in graph above, it is seen that even the median prices of these diamonds is even flatter (bar FL). Median cost diamonds between SI1 and IF seem to have almost the same price across the board, further suggesting that Blue Nile's claim is true.

Claim 3:

Next, the Blue Nile Company claims that "In terms of colors, D & E is the most expensive." Firstly, assessing the figure below, the axis's labels display the diamonds color typing on the x-axis and the Price in units of dollars that the diamond is valued at. At first glance, there is a positive trend of price increase from the color scale D to J. Specifically, diamond colors E and D are the most expensive of the color varieties. Reading from the Blue Nile Company website gives some rationale behind this price to color relationship.

A diamond's color grading is a process that involves assessing where the diamond's color falls on the Gemological Institute of America's (GIA) color scale. This scale begins with the color grading D and concludes with color grading Z. It would be interpreted as odd to start the scale at the letter D instead of A. The justification for this is that D graded diamonds are the first phenotypic feature that consumers tend to purchase, making them standardized starting point of the color scale established by the GIA. The scaling is more defined as there is a schism between D, E, and F color diamonds and G, H, I, and J diamonds. The former grouping is classified as "colorless diamonds." This grouping is rare and not adulterated by color shading or tonality distortion, so they appear more "glass-like," to the naked eye, yielding higher popularity. The latter grouping is classified "near colorless," meaning there is slight clouding in the gem, and it does not meet the standard for the higher color graded diamonds.



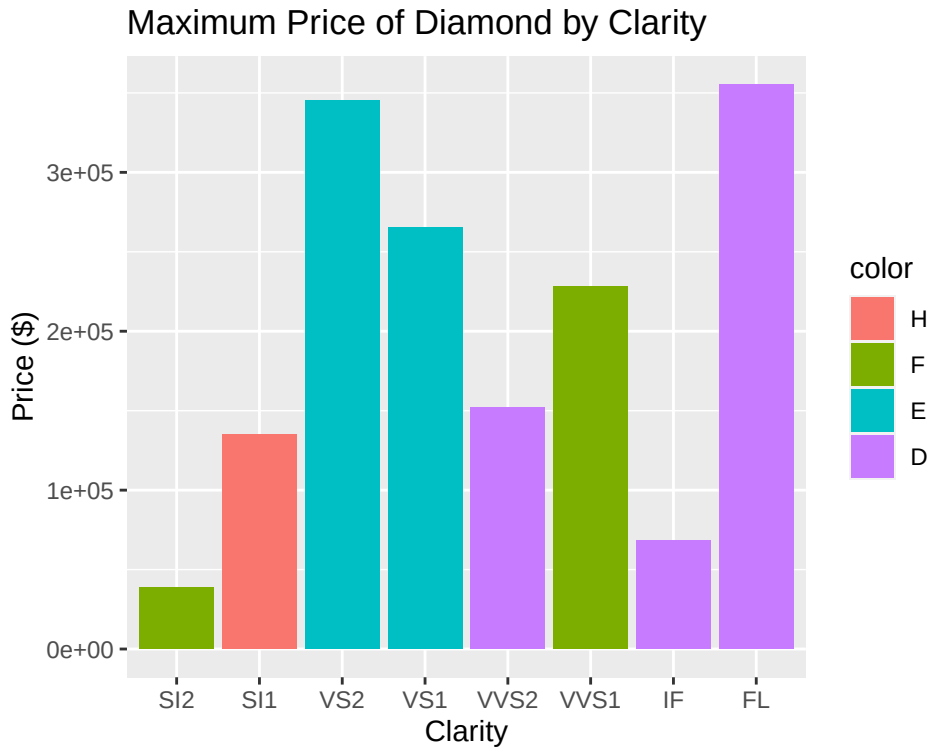
With that nuance understood, referencing the graph shown, it is clear that the D and E diamonds are valued at a higher price than those of the other color gradings. Therefore, the data does support this claim made by the Blue Nile Company.

Claim 4:

Lastly, the Blue Nile Company claims that SI and VS diamonds tend to be the best value or best quality for the price. Blue Nile supports this statement by saying that SI and VS inclusions are difficult to discern

to an untrained eye under a microscope and typically run at a lower price point. So the overall change in quality is negligible to an average consumer, but the price point is significantly lower.

Figures shown in claim 2 describe that the median and average price of each clarity quality is nearly the same between SI1 and IF diamonds which suggests that Blue Nile may be wrong. Furthermore, we can visualize the maximum price among the clarity qualities:



The plot above shows the highest price of a single diamond sold by Blue Nile in each clarity quality and their particular color. Clearly, some SI and VS2 diamonds can be more expensive than IF diamonds vying for a higher cost for lesser quality therein worse value. Blue Nile's claim that SI and VS diamonds are of the best value must be false, and best value should be assessed with a combination of other traits.

While diamonds of various cuts, clarities, and colors all have widely ranging costs, further analysis shows that carat has a direct, positive impact on the cost.

Simple Linear Regression of Price by Carat

According to the Blue Nile website, carat weight has the biggest effect on price out of all of the four C's of diamonds (cut, clarity, color, and carat weight). It should first be noted that a carat refers to the diamond's weight, not its size. Also, according to Blue Nile, a common misconception is that a larger carat weight is always better than a smaller carat weight. They mention on their website that having a well-crafted cut is more important as a high carat weight diamond with a poor cut can look smaller than a low carat weight diamond with a precise cut.

Diamond's have become a status symbol around the world, and as with many other societal misconceptions, the bigger the diamond, the larger the status symbol. Since high carat weight diamonds are seen as more of a status symbol, they in turn fetch higher prices on the diamond sale market. According to Blue Nile, "the relationship between carat weight and price depends on the rarity or availability of a rough crystal"("Diamond

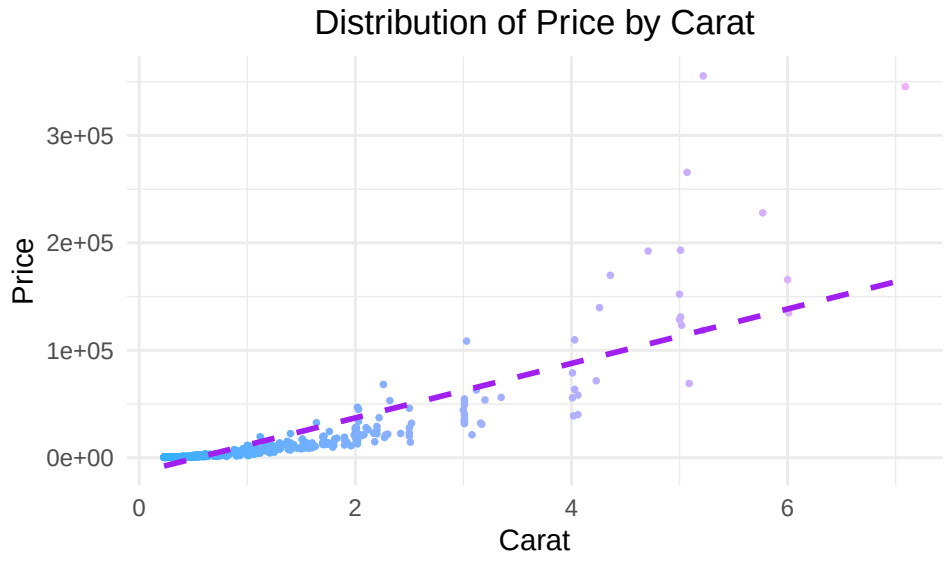
Education," 2023). They also mention that carat price is a function of crystals that have desirable colors and characteristics that can lead to beautiful clarity after a cut.

On their website, Blue Nile's direct statement is that, "prices increase exponentially as carat weight goes up"("Diamond Education," 2023). To test the relationship between price and carat weight, we must employ a simple linear regression and analyze our results to find a more quantifiable relationship.

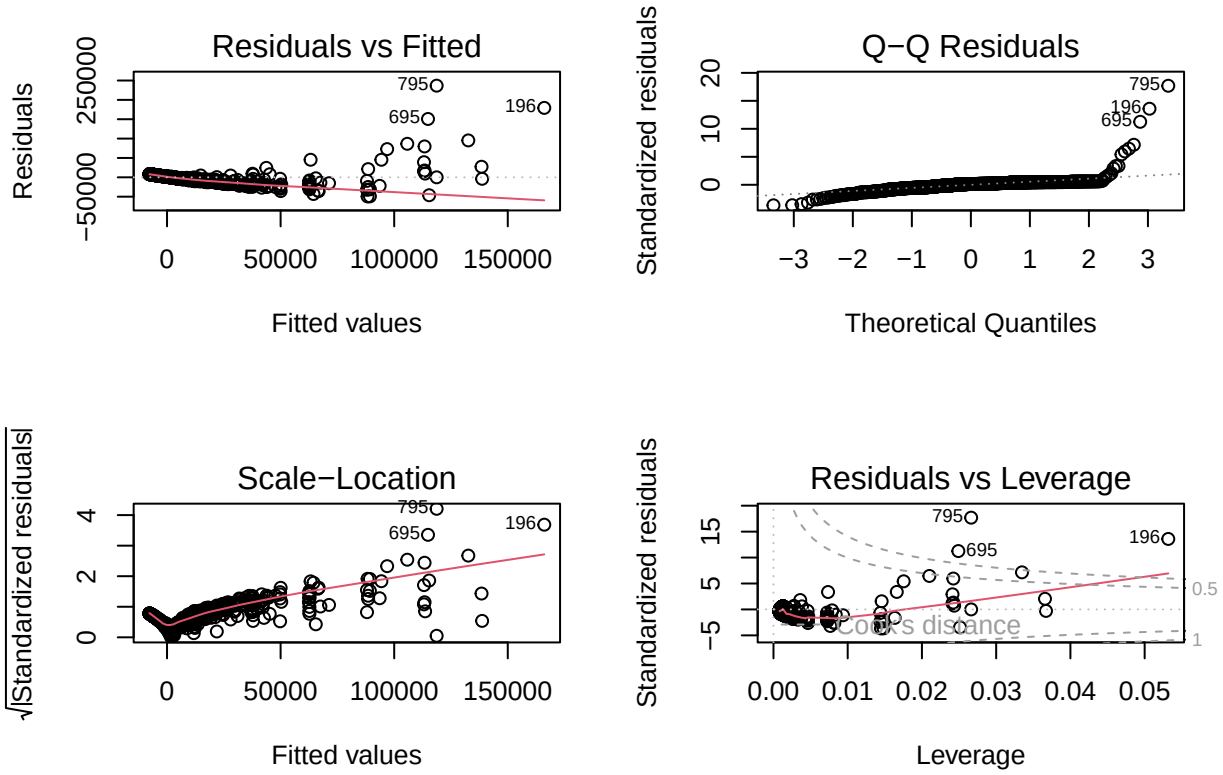
To begin our linear regression analysis we must first plot the diamond data provided into a scatter plot to analyze the initial relationship and to make sure the data meets the basic assumptions needed to perform a simple linear regression (SLR). These assumptions are listed below:

1. The errors have mean 0.
2. The errors have constant variance denoted by σ^2.
3. The errors are independent.
4. The errors are normally distributed.

We must evaluate each of these assumptions before moving forward with our regression analysis. Based on the scatter plot of the distribution of price by carat for the 1214 diamonds seen below, we see that Blue Nile's claim that price increases exponentially as carat weight increases is true based on the data provided. Due to the exponential and non-linear nature of the graph, we can see that some variable transformations will be needed in order to properly perform a regression analysis.



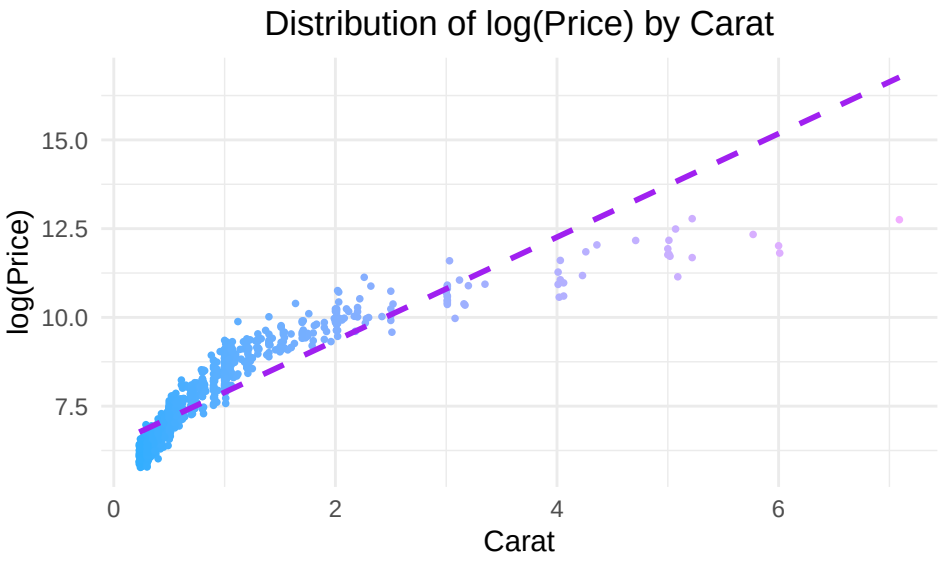
In order to find which assumptions are not being met and to determine which variable transformations to complete, we must look at the residual plots for the distribution of price by carat. These plots are displayed below:



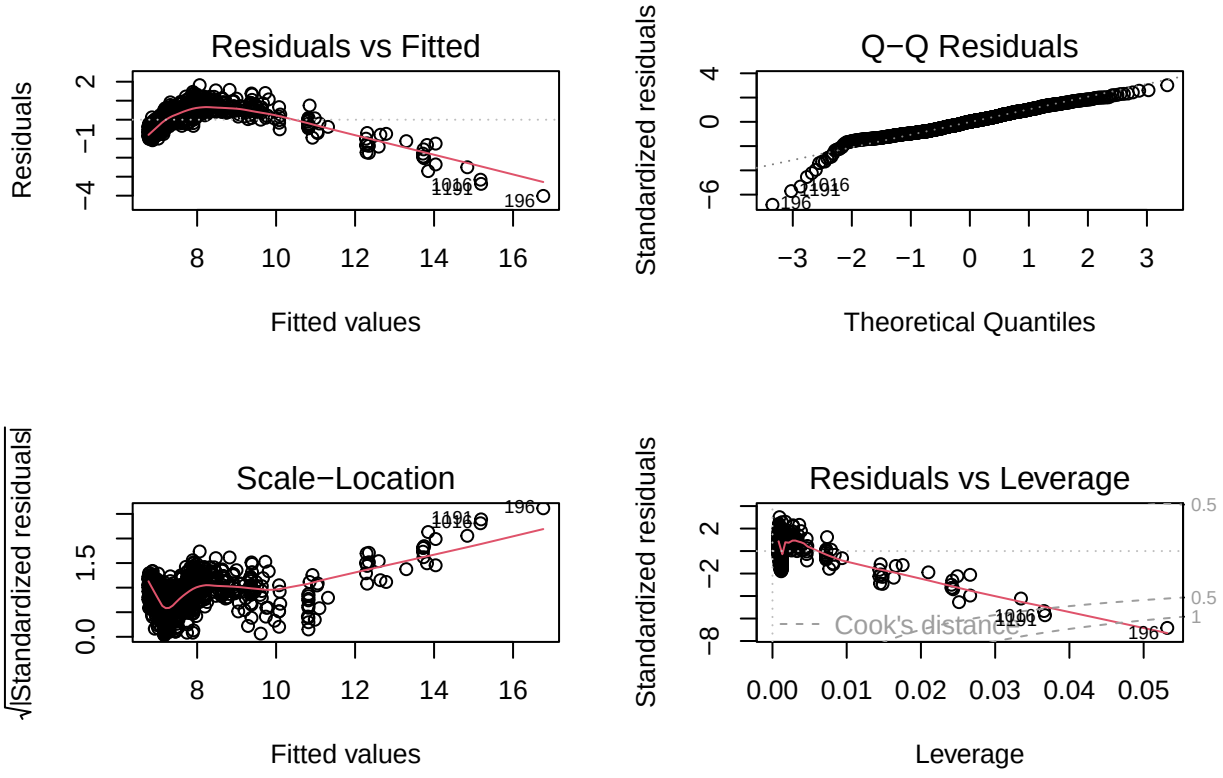
Out of these charts, we must specifically look at the Residuals vs. Fitted plot in the top left of the plot matrix. This chart shows the residuals on the y-axis and the fitted values on the x-axis. The red-line overlayed on this chart represents the average value of the residuals for differing values along the x-axis. If assumption 1 and 2 are being met, this line should be along the x-axis without any apparent curvature and no variation in vertical spread. In our plot, we see a clear descending deviation from the x-axis, informing us that assumption 1 is not being met. Also, to test assumption 2 we must look to see if the vertical spread of the residuals is fairly constant as we move from left to right on the plot. We do not see this in our residual plot, instead the vertical spread increases as we move from left to right.

In order to meet our first two assumptions, we need to perform variable transformations. We start by transforming the response variable so that we can stabilize the variance. As mentioned before, the variance of the residuals increases as we move from left to right. Based on this, we can assume that we need to transform the response variable using $y^* = y^\lambda$ with $\lambda < 1$. We first must try a log transformation on the response variable. Log transformations are preferred over other transformations as the regression coefficients can still be interpreted post-transformation.

After transforming the response variable, price, to log(price), we get our new response variable, y^*. Plotting y^* against carat we get the scatter plot displayed below:



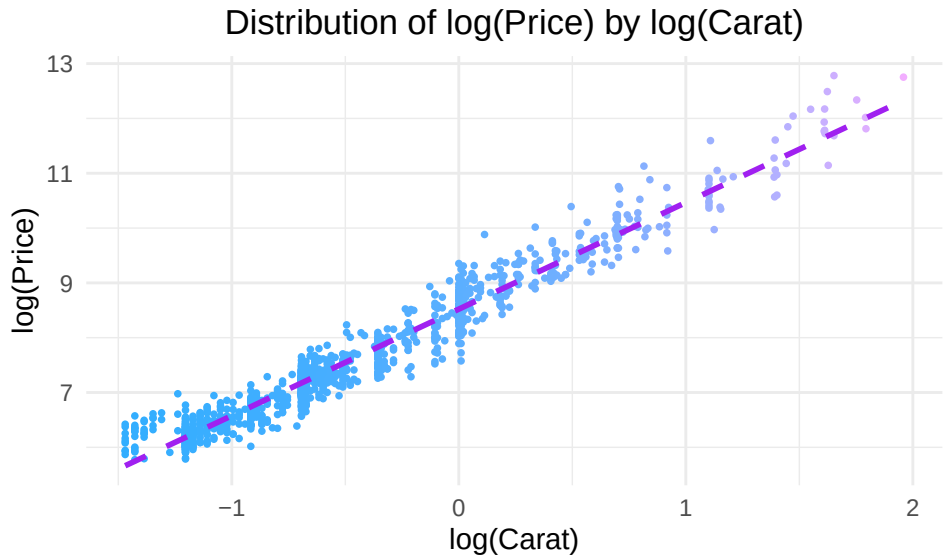
As we can see from the scatter plot, there is now a clear logarithmic relationship between carat and our new y^*. We can clearly see that assumption 1 is not being met due to the non-linear nature of the chart; however, we must use residual plots to check assumption 2.



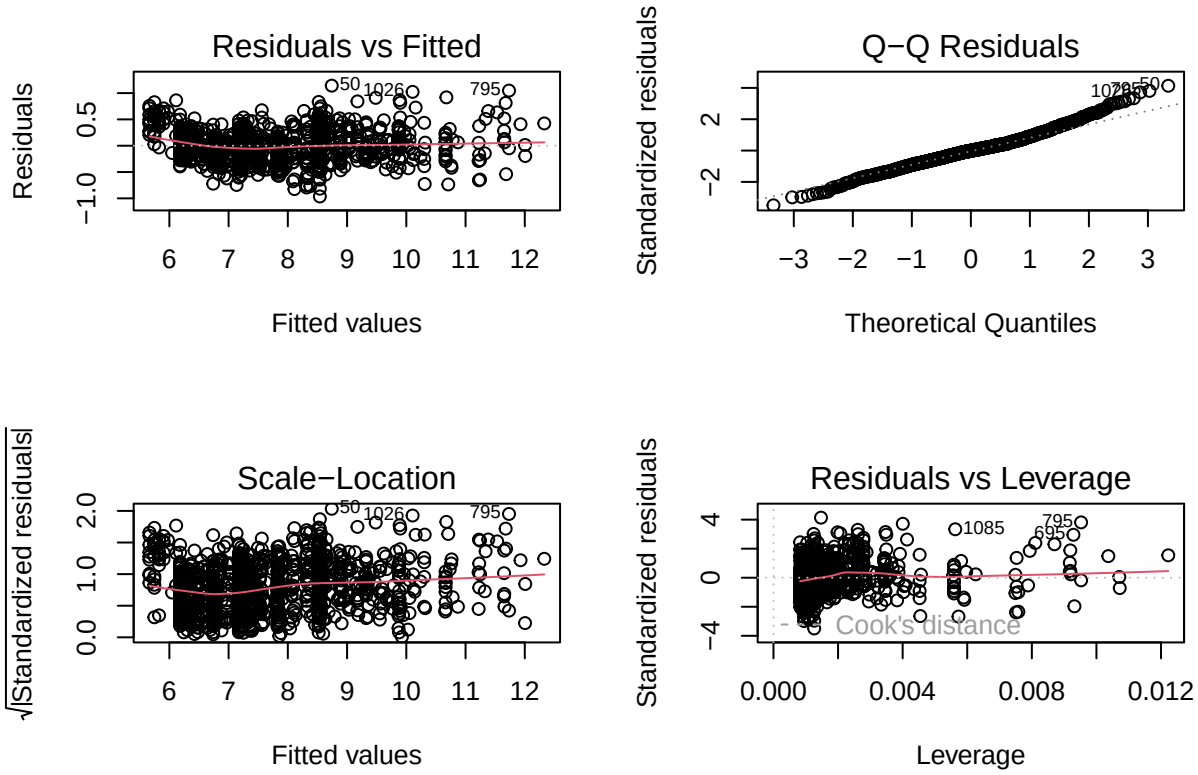
The residual plots shown above confirm that assumption 1 is still not being met as there is a clear curvature in the residuals vs. fitted plot; however, assumption 2 of SLR is now met. There is a fairly constant vertical variance across the red-line moving from left to right showing that our log transformation was successful. Now with our variance stabilized, we must move to transforming the predictor variable in order have

assumption 1 met as well.

In order to transform our predictor variable, we must first try a log transformation as our scatter plot seems to resemble the plot of log(x). We transform our carat data into log(carat) and produce our new predictor variable x^*. Creating a scatter plot of the log of price against the log of carat (y^* against x^*) we get the scatter plot displayed below:



This scatter plot shows us that there is a linear relationship between both new variables. As the number of log(carat) go up, the log(carat) goes up as well in a linear manner. To confirm this and to test our assumptions, we must look at the residual plots of our two new variables. These plots are shown below:



We can see in the residuals vs. fitted plot that the vertical variance is stabilized moving from left to right, so assumption 2 is still met. Also, the residuals are now more evenly scattered across the horizontal axis in the residual plot so assumption 1 is now met as well. To test assumption assumption 4 we must look at the normal probability plot of residuals (QQ plot) in the top right corner of the plot matrix. Since the plots fall closely to the QQ line, we have evidence that the observations follow a normal distribution and assumption 4 is met as well. Lastly, to make sure assumption 3 is met, we must create an ACF plot. This ACF plot is displayed below:

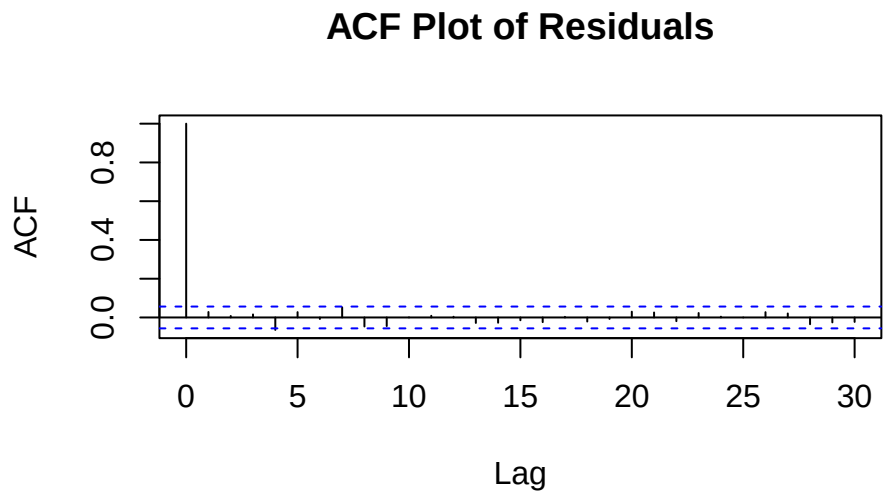


Figure 1: ACF Plot of Residuals

Since none of the ACFs beyond lag 0 are significant, we don't have evidence that the observations are dependent on each other. This shows that assumption 3 is also met. Since all assumptions are now met, we can accurately complete a simple linear regression of the two variables and find the linear relationship.

Regression Equation

```
##
## Call:
## lm(formula = ystar ~ xstar, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96394 -0.17231 -0.00252  0.14742  1.14095
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.521208   0.009734   875.4   <2e-16 ***
## xstar       1.944020   0.012166   159.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2761 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
```

Fitting a linear regression to our two new variables, we can use R to determine our slope($\hat{\beta}_1$) and y-intercept($\hat{\beta}_0$) to determine our regression equation. The slope is 1.944 and the y-intercept is 8.521. The regression equation based on our SLR is as follows:

$$\hat{y}^* = 8.521 + 1.944x^*$$

Since we performed log transformations, we can interpret the slope of the regression. Since both the predictor and response variable were log transformed we can assume that that in general for an a% increase in the predictor, the predicted response is multiplied by $(1 + \frac{a}{100})^{\hat{\beta}_1}$. The following interpretations can be made based on our regression equation:

1. For a 1% increase in the carat of diamonds, the predicted price is multiplied by a factor of $1.01^{1.9444} = 1.0195$.
2. For a 1% increase in the carat of diamonds, the predicted price increases by approximately 1.944%.

Hypothesis Test

To test the hypothesis that there is a linear relationship between our two new variables of log(price) and log(carat), we can use an ANOVA F test. We can set our two hypothesis up as follows:

$$H_0 : \hat{\beta}_1 = 0$$

$$H_\alpha : \hat{\beta}_1 ! = 0$$

Using R to get a summary of our regression analysis results, shown above, we can determine that a regression of two new variables gives us an F-statistic of 2.553e+04 and a p-value of p < 2.2e-16. Since we have such a small p-value, we can reject the null hypothesis. The data supports a claim that there is a linear association between log(price) and log(carat).

Conclusion:

In the complete analysis of the four claims that were stated by the Blue Nile Company alongside the overview of the Simple Linear Regression Model several conclusions were made. Firstly, the claim that 'Astor Ideal' diamonds represent the highest quality and fetch the highest price, was refuted by the data visualizations "Median Price of Diamond by cut" and "Average Price of Diamond by Cut" which was concluded to be false. Secondly, Clarity is the least important factor of the 4C's this claim was supported by the visualization "Mean Price of Diamonds by Clarity" as the graph demonstrates an oscillating with regard to the average price across different cuts. Additionally, the flat behavior of the median price graph indicates that clarity has marginal affect on diamond pricing. Furthermore, the claim that "In Terms of colors, D & E is the most expensive," is supported by the data visualization, "Collective Price of Diamonds Against Color by Cut Quality," which demonstrates the popularity and higher value of D and E colored diamonds. Interestingly the next claim, "SI and VS diamonds are the best value" was confirmed by the previously mentioned two data visualizations. The final claim that, "that SI and VS diamonds tend to be the best value or best quality for the price," was concluded to be inconclusive as other diamond trait combinations must be considered.

The linear regression analysis determined that as the "Distribution of Price by Carat," scatterplot relationship was an exponential one and nonlinear more variable transformations needed to be completed to perform the regression analysis. Transforming the response variable to stabilize the variance then transforming the predictor variable. From these transformations and assessment of the transformed variable scatterplots the linear assumptions were met with the assistance of the corresponding residual plots. Therefore, the conclusion can be made that there is a linear relationship between Log(Carat) and Log(Price) from the dataset. Additionally, that for 1% increase in the carat of diamonds, the predicted price increases by approximately 1.944%.

Appendices:

Diamond Education. (2023). Blue Nile. Retrieved from https://www.bluenile.com/education/diamonds