Group 5: Greg Miller, Michael Puchalaski, Sree Bandakavi, Sean Hersee

Section 1: Results of Analysis

The price of diamonds is a topic from which many controversies arise. However, these outside influences are not the focus of this report. Instead, the focus lies on the combination of what the diamond industry calls the 4Cs of diamonds, something specifically defined by Blue Nile, the origin of the 1214 observation working dataset, as cut, carat, clarity, and color. As described by Blue Nile, these are considered the four main influences, outside of shape, on the overall price of a diamond. The primary influence on a diamond's price is the carat.

Carat, formerly measured as the number of carob seeds needed to balance a scale when weighing a diamond, refers to the weight of a diamond. Currently, 1 carat equates to 200 milligrams (mg), with the ratio remaining constant across any increase or decrease, meaning a 0.5-carat diamond weighs 100mg while a 2.25-carat diamond weighs 450mg. With carat being the most consistent predictor of price when controlling for other contributing factors, our analysis of the Blue Nile data set finds that every 1% increase in weight results in a price increase of 1.944%. In other terms, for every 2mg increase in weight, the price will increase by 1.944%. This means that if a 1-carat diamond is priced at $5037.60, it would be fair to expect its price to increase by around $97.50 for a diamond weighing 2mg more when accounting for carat alone.

Along with carat, Blue Nile has several claims, all with varying degrees of accuracy, with some being more recommendations for consumers than true predictions for how they relate to price changes, all relating to the 4Cs.

Starting with carat, our study finds, alongside its strength as a predictor of price, that Blue Nile's encouragement for consumers to "buy shy," meaning to buy a diamond just shy of a half and full carat, such as a 0.90-carat or a 1.40-carat diamond, for a better value purchase is valid, especially at the lower carat weights. However, we also find that purchasing just above the half and full-carat marks can see similar price decreases from those benchmark points.

While highly rare, only three out of all 1214 in the data set, flawless diamonds are substantially more costly than an internally flawless diamond; our study finds that this stark increase in average price is not the standard when it comes to cut and color. Specifically, the study finds that a colorless diamond, although more expensive on average, has a minimal increase in average price compared to the following near-colorless average price. Next, quality does not always come at a high cost, with many Astor-Ideal diamonds coming in the form of lower-carat diamonds, resulting in a lower average price than lower quality, cut-wise, diamonds.

Overall, the price of a diamond has more to do with the combination of these factors, where a diamond that rates as flawless, has an ideal cut, is colorless, and is a high carat will price in the hundreds of thousands while a diamond with slightly less extremes can still be reasonably priced.

Section 2: Data and Variables & How the Four C's Relate to Price

Description of the Data and Variables

The dataset describes over 1000 diamonds for sale through the Blue Nile website: http://www.bluenile.com/.

The 1214 observations in the dataset are complete with five variables: The 4Cs of diamonds, carat, clarity, color, cut, and price. Below is a detailed description of each variable as described by Blue Nile:

Carat - Carat measures a diamond's weight and refers to an exact weight of 200 milligrams. Typically, heavier high-quality diamonds are rarer than ones with lower carat weights, and diamond prices can reflect this.

- **Carat Group -** This is an additional variable created by the team. It is a grouping of diamonds by carat for easier comparison. There are six distinct groups:
    - Diamonds < 1.00 carat
    - Diamonds between 1.00 - 1.99 carats
    - Diamonds between 2.00 - 2.99 carats
    - Diamonds between 3.00 - 3.99 carats
    - Diamonds between 4.00 - 4.99 carats
    - Diamonds between > 5.00 carats

Clarity - Clarity assesses minor imperfections within a diamond. Clarity is used to qualify and describe the nature of any inclusions that occur during the diamond-forming process. Diamonds are rated:

- I1, I2, I3: These diamonds have apparent inclusions and are not available through Blue Nile; thus are present in our data
- SI1, SI2: Slightly Included Diamonds: These diamonds have slight inclusions that are only visible when viewed from the side
- VS1, VS2: Very Slightly Included Diamonds: These have very minor inclusions, which can be perceived only at 10x magnification
- VVS1, VVS2: These diamonds have little to no inclusions and are considered very rare
- IF: Internally Flawless diamonds have minor surface blemishes but are visually eye-clean
- FL: Flawless diamonds encompass less than 1% of all diamonds and are nearly impossible to find.
    - **Clarity Group -** This is an additional variable created by the team. It is a grouping of clarity ratings where the SI1 and SI2 are grouped as SI with a similar grouping made for VSI, while IF and FL are left on their own.

Color - Diamond color refers to how colorless a diamond is. Colorlessness is desirable in most diamonds; the more colorless diamonds are rarer. The less color a diamond has, the higher the diamond color grade. Blue Nile claims that this is the second most important of the 4Cs. Diamond color rating starts at D and continues alphabetically till Z for the worst. Diamond price tends to decrease in alphabetical order as well.

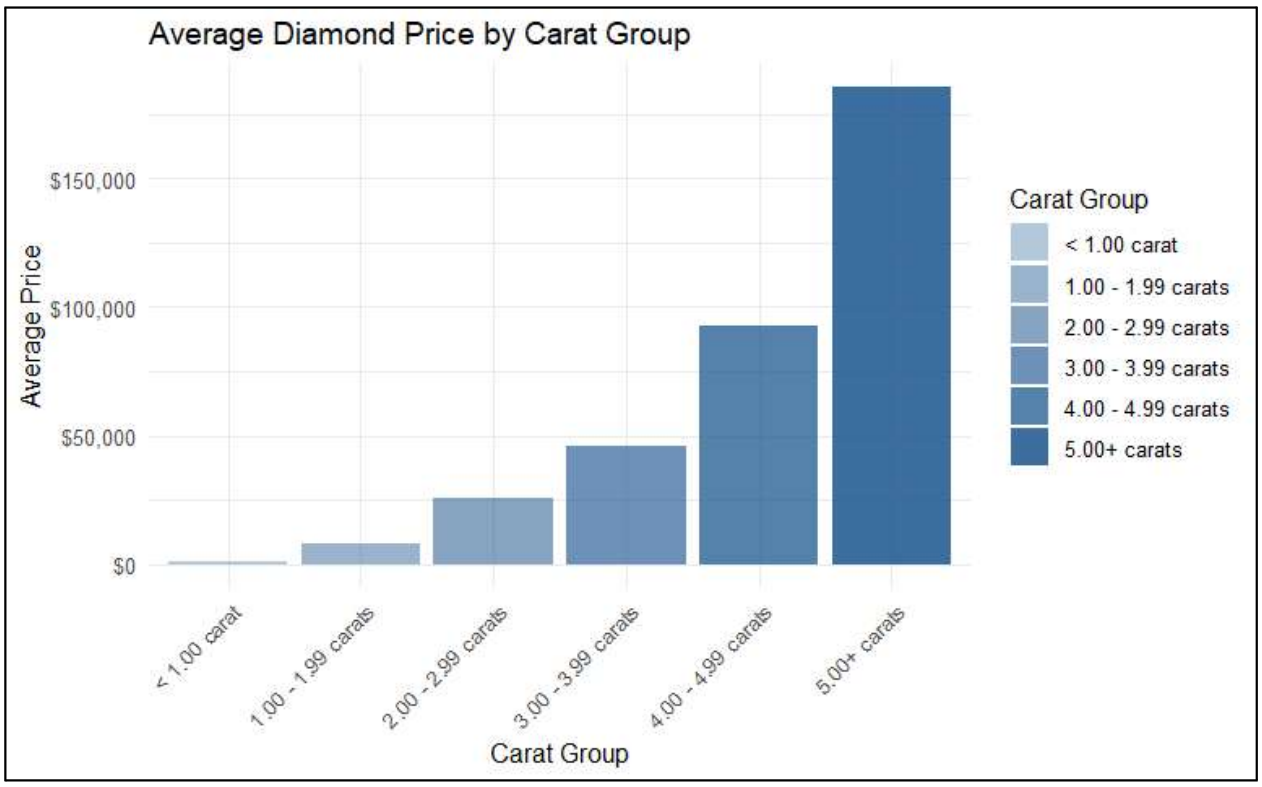
- D-F: The rarest and offer a pure icy look
- G-J: Near colorless and generally considered excellent value for the quality
- K: Budget-friendly, but has a faint color.
- L-Z: A noticeable warm yellow color; Blue Nile does not sell these.
- **Color Group -** This is an additional variable created by the team. It is a grouping of color rating D-F into “colorless” and G-J into “near-colorless”

Cut - Cut measures how well-proportioned a diamond’s dimensions are, including its balance and brilliance. Diamond cut is considered the most important of the four Cs. Cut is scored Poor, Fair, Good, Very Good, and Astor Ideal. In this study, the lowest score cut quality is Good, while the highest score is Astor Ideal.

Price - Price is the dollar amount Blue Nile lists for each diamond. This is arguably the critical factor that most prospective buyers are looking at. In this report, we plan to understand the relationship between the 4Cs and the price.

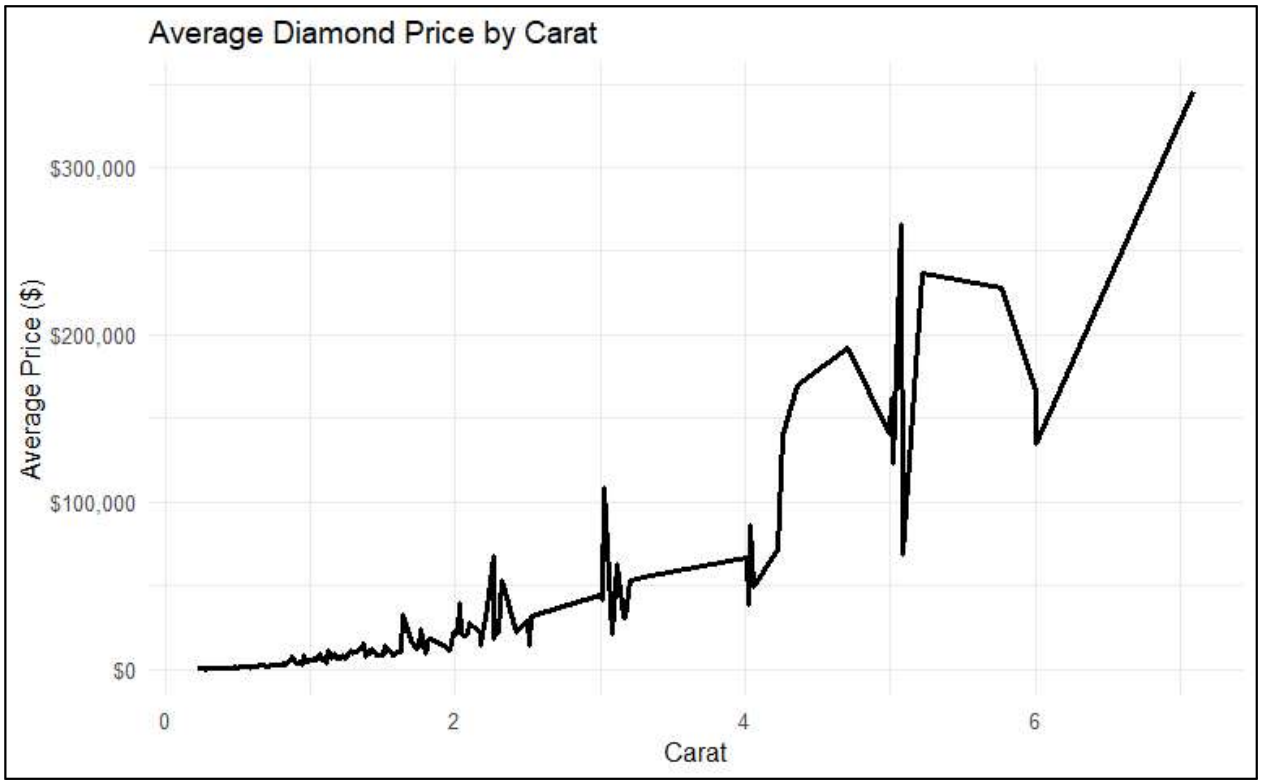
How Price is Related to Carat Weight

On their diamond education page, Blue Nile claims that "Carat has the biggest effect on price." Looking at the average price per carat group, there is a pronounced increase as carat weight increases. Compared with clarity, color, and cut (which are expanded on later in the report), the carat increases the price significantly.

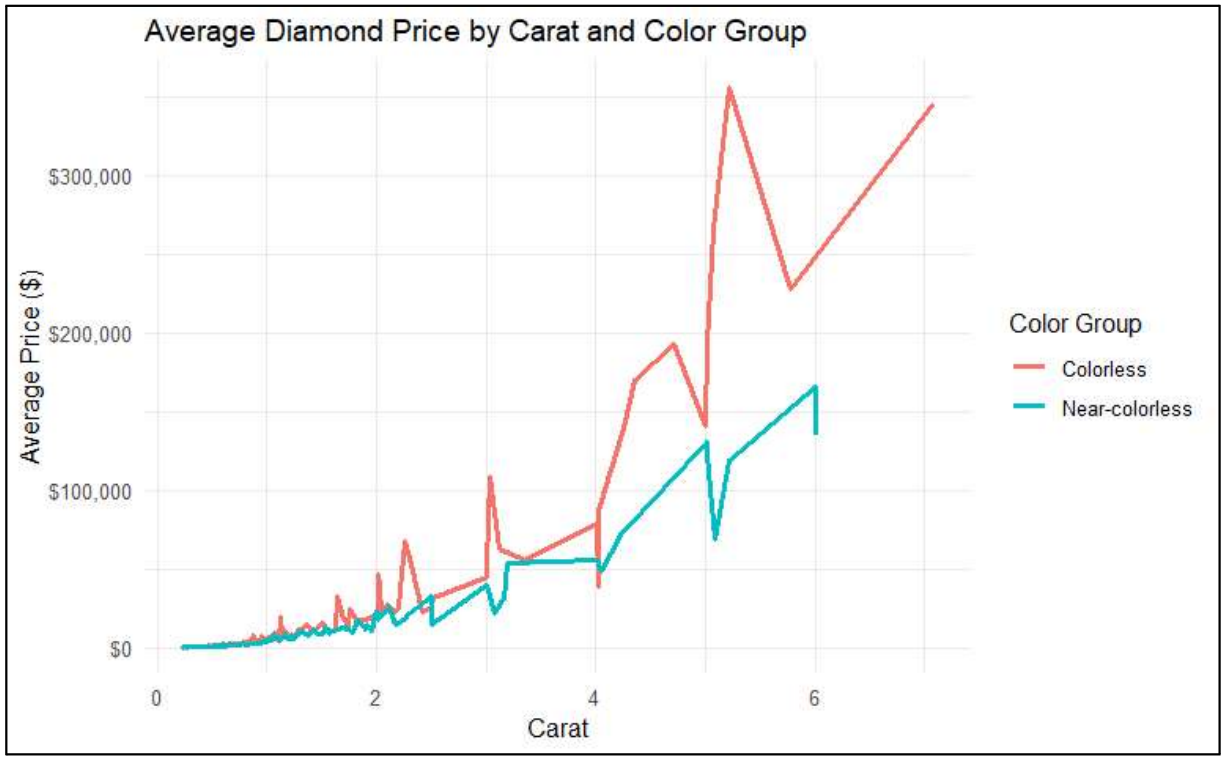


A second claim made by Blue Nile about carat weight is that money can be saved by buying slightly below whole and half-carat marks, which is cleverly described as "buying shy" to save money. Select a carat weight slightly below the whole and half-carat marks. For example, instead of a 2.00-carat diamond, consider buying a 1.90-carat weight. This saves money, and the slight difference is not substantial.

This claim is more difficult to prove. The graph below shows the average diamond price by carat, and while we do see price spikes just beyond the 3, 4, and 5-carat marks, they are all followed by decreases.



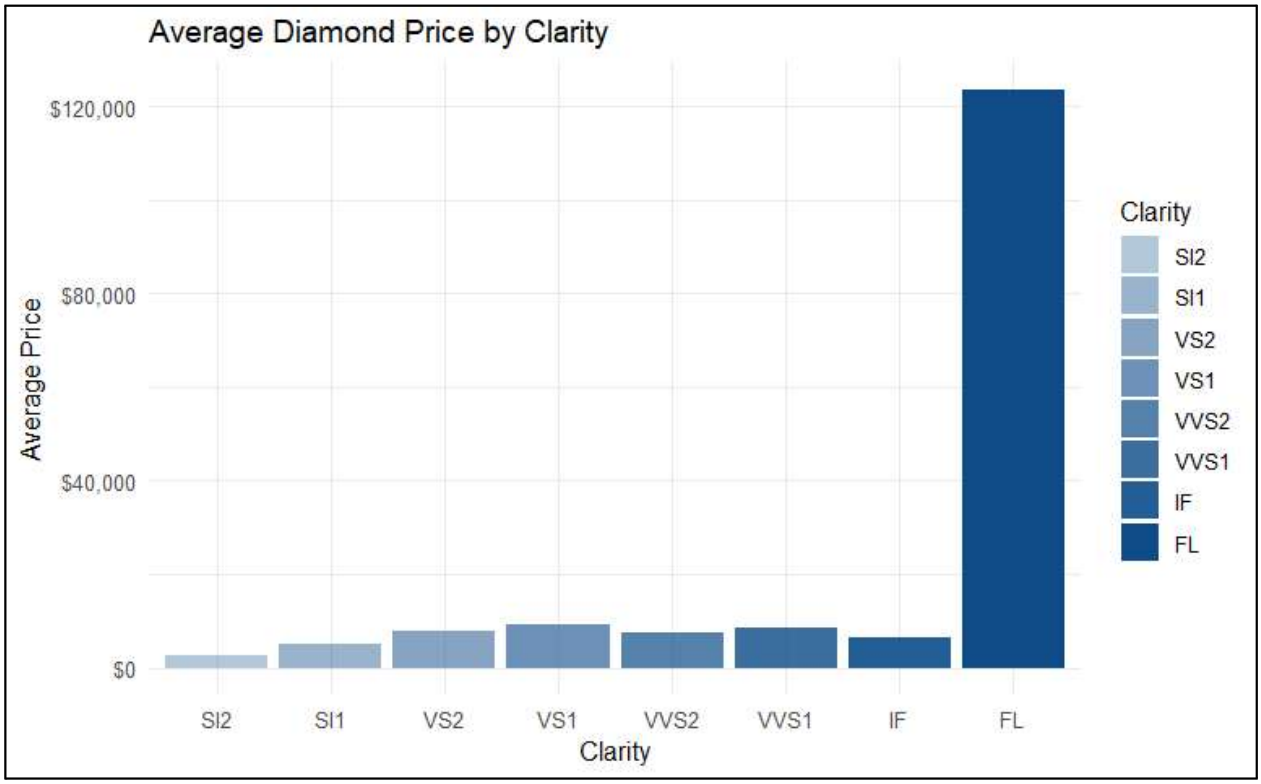
Even while trying to account for other variables like color in the figure below, the same volatility in price is seen. I would amend Blue Nile's claim about buying shy to save money and add that you can also save money by buying just above the full carat marks.



How Price is Related to Diamond Clarity

Clarity has little effect on diamond price until you get to the highest grade of flawless (FL). At that grade, the price increases by around a factor of 10. According to Blue Nile, "For the best value, select a diamond with inclusions that can't be seen through the crown without magnification (also known as eye clean diamonds), like a diamond with a slight inclusion (SI) or very slight inclusion (VS) clarity grade."

The data supports this claim. While flawless diamonds are the highest quality, you can save considerable money by buying a diamond of slightly lower clarity.



Along with this claim, they also report that clarity is also affected by a diamond's shape and size. While we are not given the shape with our data set, we are given data for cut, from which the quality of the shaping of the diamond can be extrapolated. However, this claim is more conflicting. Instead, it seems that the carat has little to no true influence over the clarity quality, with clarity being visibly well distributed in relation to the overall rarity of each clarity level, from highest quality (FL) to lowest quality (SI) at Blue Nile. Similar statements can also be made about the distribution of cut quality across these clarity groups as well. This can be plainly seen in the chart below, as well as the corresponding tables for each of the facet-wrapped visualizations shown.

Group 5: Greg Miller, Michael Puchalaski, Sree Bandakavi, Sean Hersee

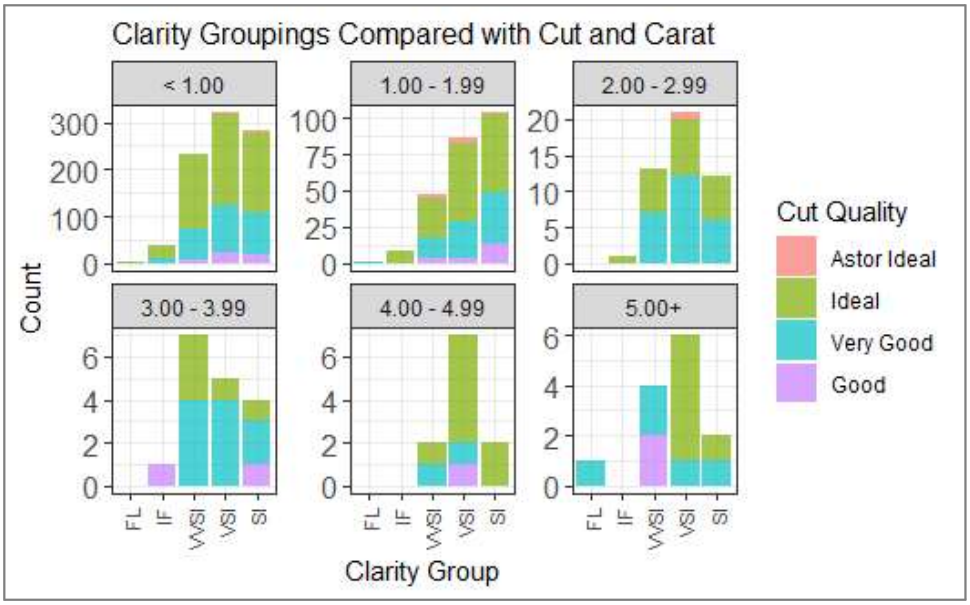


Table for Carat Group: less than 1.00

		Summary	
Clarity	Cut	Count	Percentage
FL	Ideal	1	0.11
IF	Ideal	28	3.18
IF	Very Good	11	1.25
SI	Astor Ideal	4	0.45
SI	Good	19	2.16
SI	Ideal	172	19.55
SI	Very Good	89	10.11
VSI	Astor Ideal	6	0.68
VSI	Good	21	2.39
VSI	Ideal	194	22.05
VSI	Very Good	101	11.48
VVSI	Astor Ideal	2	0.23
VVSI	Good	7	0.80
VVSI	Ideal	160	18.18
VVSI	Very Good	65	7.39

Table for Carat Group: 1.00 - 1.99

		Summary	
Clarity	Cut	Count	Percentage
FL	Very Good	1	0.41
IF	Ideal	8	3.25
SI	Astor Ideal	1	0.41
SI	Good	13	5.28
SI	Ideal	54	21.95
SI	Very Good	36	14.63
VSI	Astor Ideal	4	1.63
VSI	Good	4	1.63
VSI	Ideal	54	21.95
VSI	Very Good	24	9.76
VVSI	Astor Ideal	2	0.81
VVSI	Good	4	1.63
VVSI	Ideal	28	11.38
VVSI	Very Good	13	5.28

Table for Carat Group: 2.00 - 2.99

		Summary	
Clarity	Cut	Count	Percentage
IF	Ideal	1	2.13
SI	Ideal	6	12.77
SI	Very Good	6	12.77
VSI	Astor Ideal	1	2.13
VSI	Ideal	8	17.02
VSI	Very Good	12	25.53
VVSI	Ideal	6	12.77
VVSI	Very Good	7	14.89

Table for Carat Group: 3.00 - 3.99

		Summary	
Clarity	Cut	Count	Percentage
IF	Good	1	5.88
SI	Good	1	5.88
SI	Ideal	1	5.88
SI	Very Good	2	11.76
VSI	Ideal	1	5.88
VSI	Very Good	4	23.53
VVSI	Ideal	3	17.65
VVSI	Very Good	4	23.53

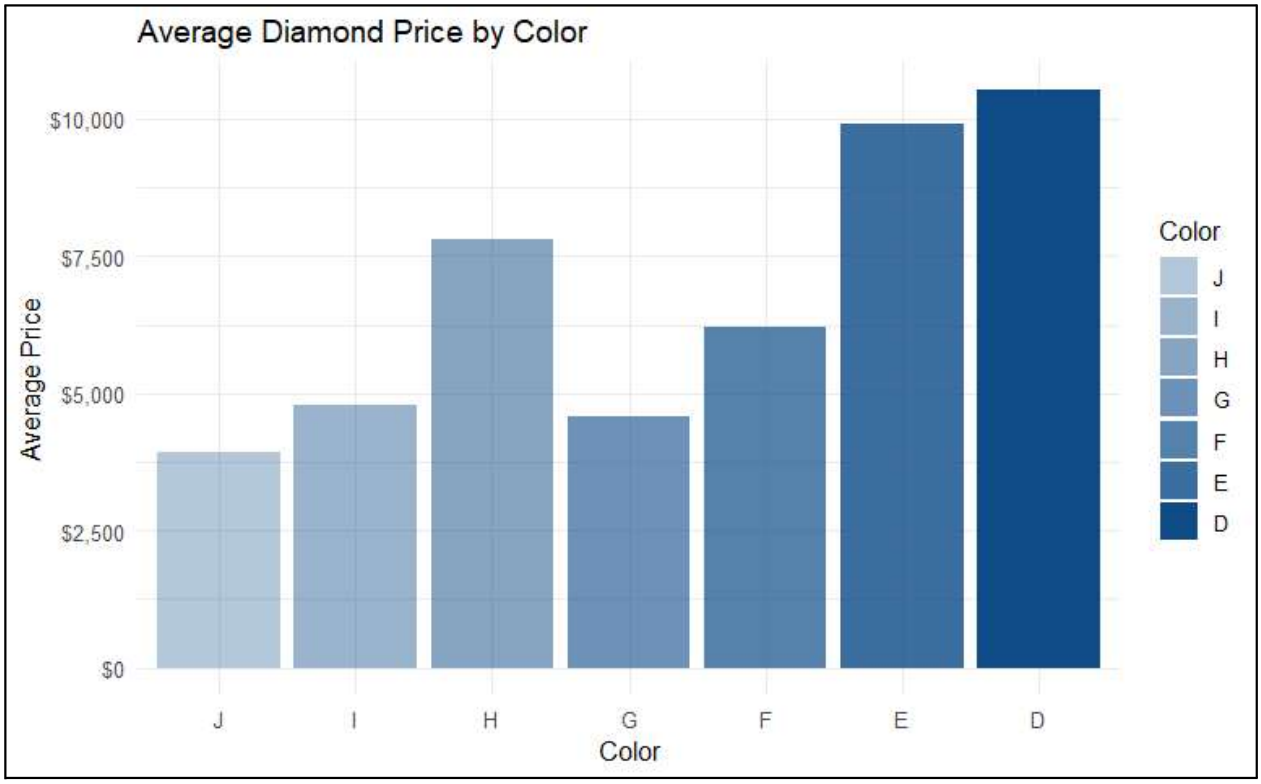
Table for Carat Group: 5.00+

		Summary	
Clarity	Cut	Count	Percentage
FL	Very Good	1	7.69
SI	Ideal	1	7.69
SI	Very Good	1	7.69
VSI	Ideal	5	38.46
VSI	Very Good	1	7.69
VVSI	Good	2	15.38
VVSI	Very Good	2	15.38

How Price is Related to Diamond Color

As colorlessness increases in a diamond, so does the price. On the diamond education page, Blue Nile makes the following claim about color:
"Diamond prices decline or increase in alphabetical order. For example, a diamond with a G color grade is less expensive than a diamond with a D color grade."

The data generally supported this; while a color rating of H has a slightly higher average price than G or F, the general trend holds true.



	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*	*Significance*
Color D (Intercept)	10525.2	1672.2	6.294	4.32E-10	***
Color E	-619.3	2448.3	-0.253	0.8004	
Color F	-4320.6	2322.1	-1.861	0.063	.
Color G	-5953.5	2391.6	-2.489	0.0129	*
Color H	-2726.6	2589.9	-1.053	0.2926	
Color I	-5745.8	2502.5	-2.296	0.0218	*
Color J	-6591.1	3037.8	-2.17	0.0302	*

Residual standard error: 24060 on 1207 degrees of freedom

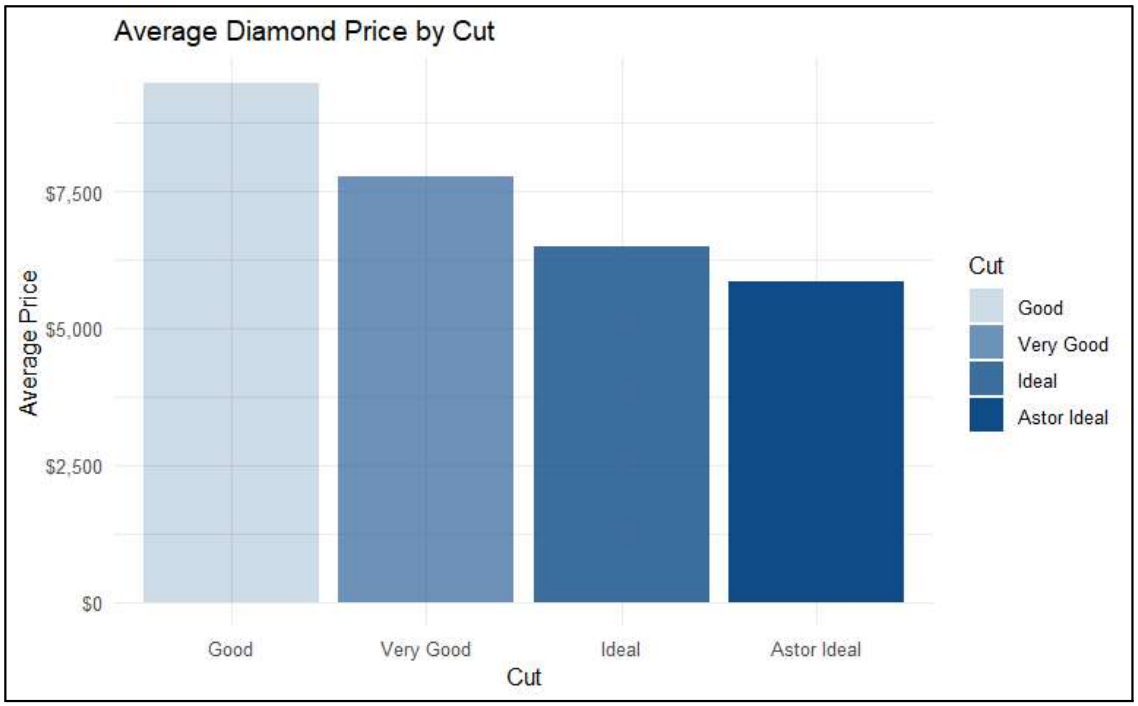
Multiple R-squared: 0.01016, Adjusted R-squared: 0.005237

F-statistic: 2.064 on 6 and 1207 DF, **p-value: 0.05474**

The output from a linear regression model that examines the relationship between color and price is shown above. The Intercept is the baseline color, D, and represents the estimated cost of a diamond and how it changes compared to the change in color. Looking at the estimate, the price tends to be lower than color D; however, it was found that Colors G, I, and J were the only colors that showed significant price decreases. Color as a whole does not have a significant effect on price, which is confirmed with an R^2 value of 1%, suggesting that color causes minimal variation in the price. Color has a small but statistically weak effect on price.

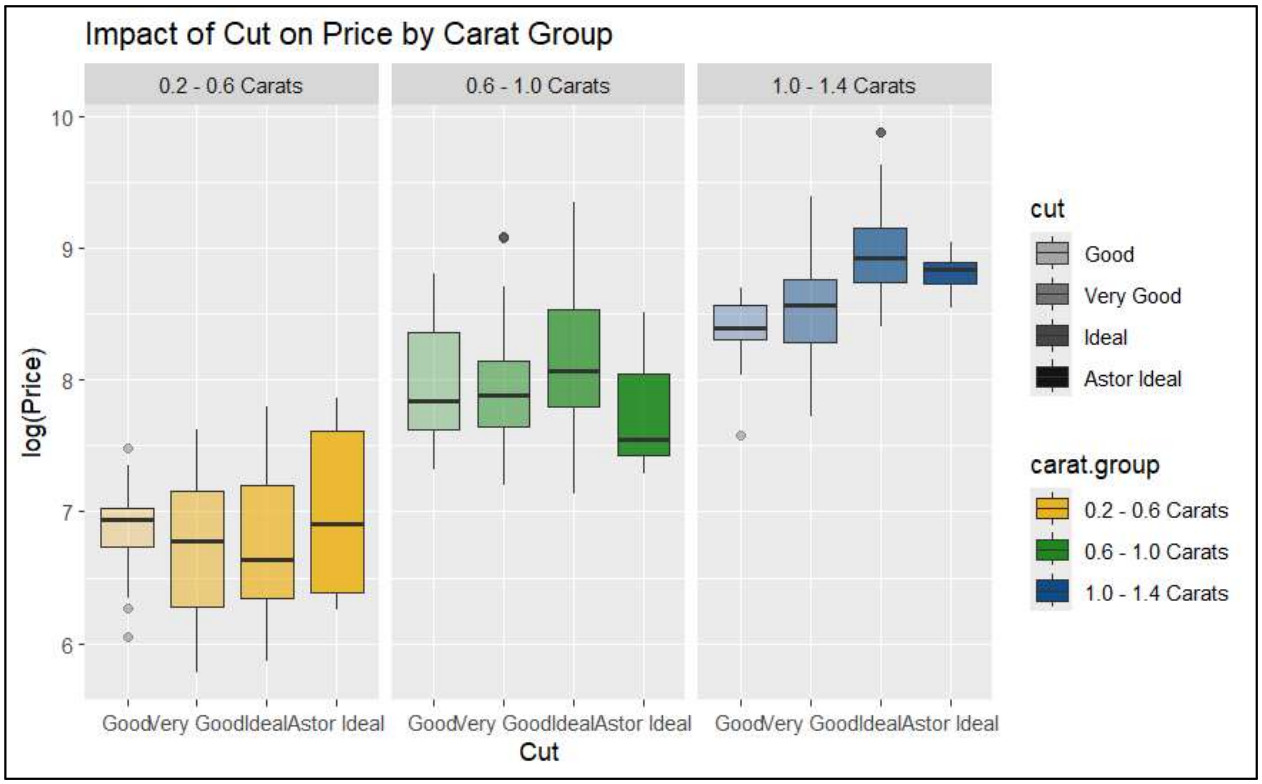
How Price is Related to Diamond Cut

One would expect the price of a diamond to increase with cut quality; after all Blue Nile claims that cut is the most important of the four C's when it comes to selecting a diamond. The chart below shows that diamond price decreases when the cut quality increases.



However, when we consider carat, which has the largest effect on the price of a diamond, there is not an obvious negative correlation between cut and price. The boxplot below shows that price remains fairly constant across cuts when accounting for carat groups.

Group 5: Greg Miller, Michael Puchalaski, Sree Bandakavi, Sean Hersee



Section 3: Regression of Price against Carat Weight

Performing a regression of Price against Carat Weight on the Blue Nile diamond dataset required transformations of both the response (Price) and predictor (Carat Weight) variables before the results could reliably be used to interpret the data. The steps involved were:

1) Checking the data for a linear relationship

2) Creating a linear regression and checking the assumed assumptions

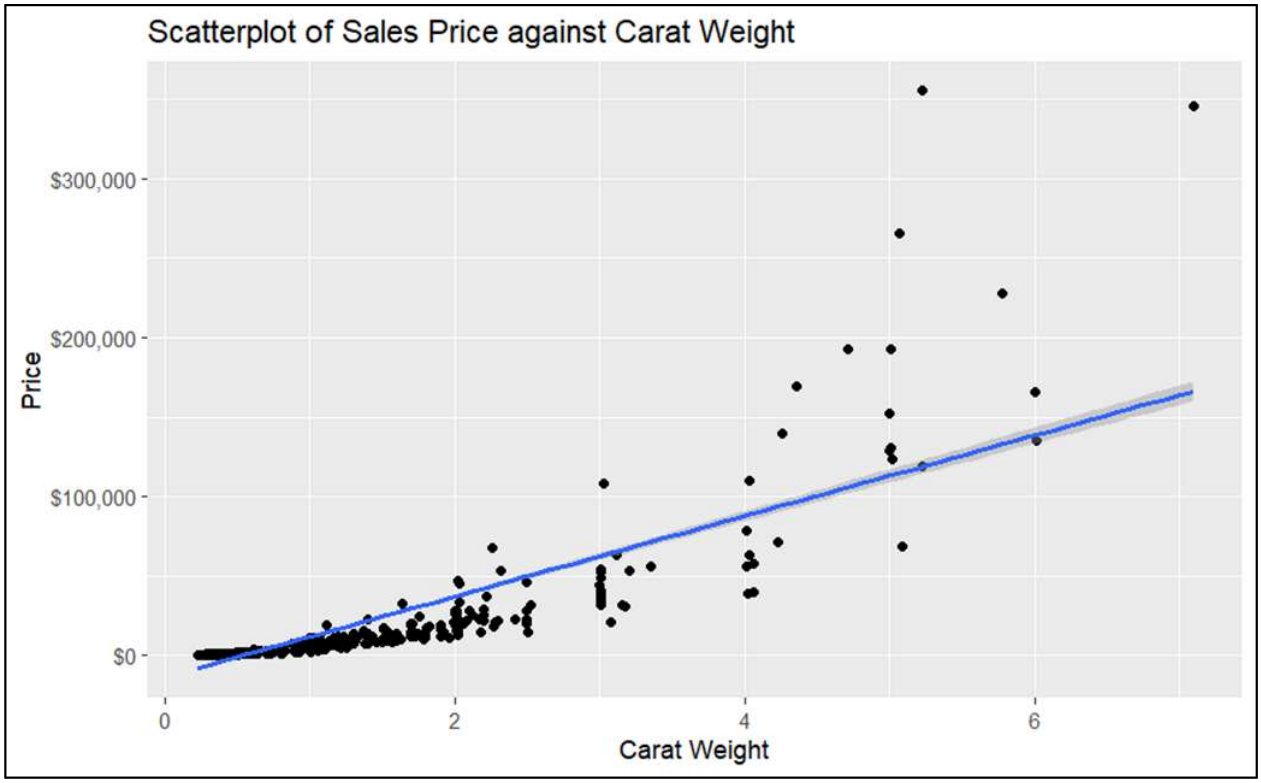
3) Determining the transformation of the response variable

4) Determining the transformation of the predictor variable

5) Final confirmation of linear regression assumptions and interpretation of the model

Step 1: Confirm Linear Relationship

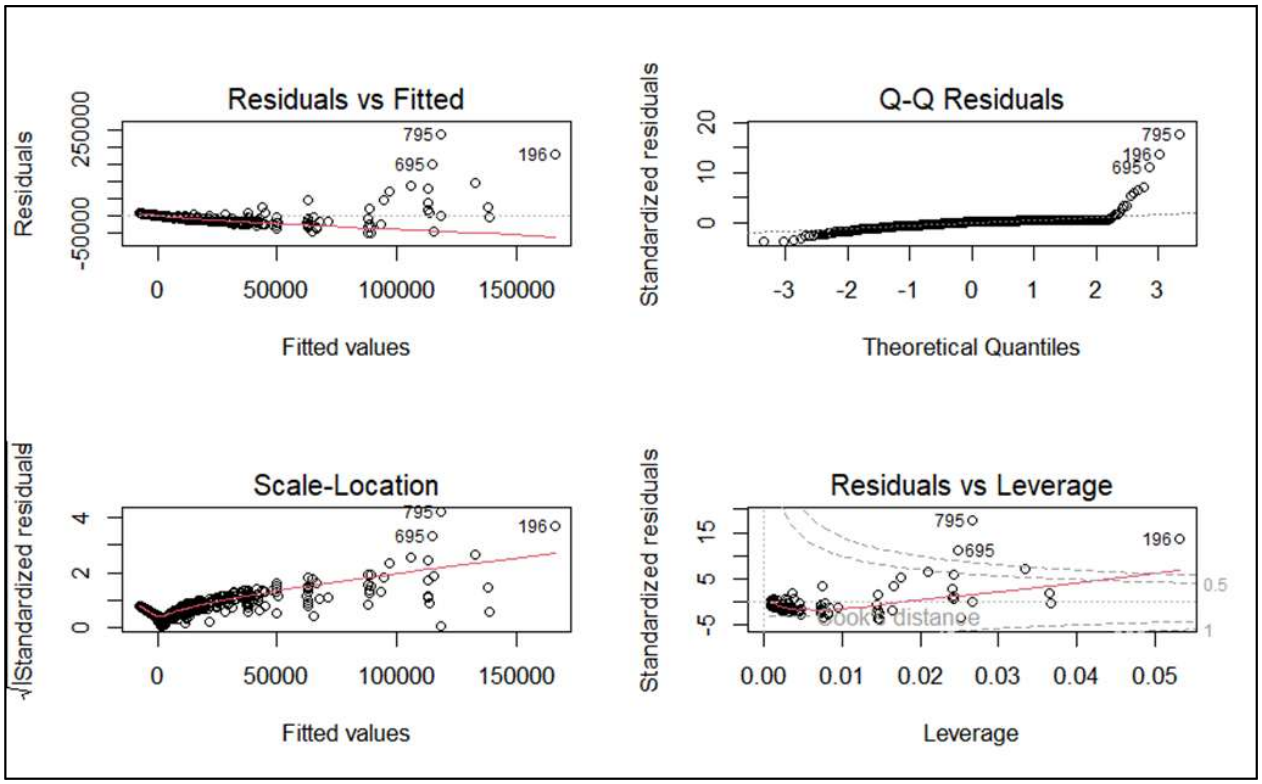
A scatterplot of price against carat weight was produced to assess the potential for a linear relationship between the two variables. As seen in the chart below, a clear positive linear relationship exists.



We also note that the scatterplot indicates there are likely to be linear regression assumptions that are not met. Specifically, we note that the data points are not evenly distributed on either side of the regression line, suggesting a failure to meet linear assumption 2 (errors have constant variance), and the magnitude from the regression line varies across values of the x-axis suggesting a failure to meet linear assumption 1 (errors have a mean of 0).

Step 2: Creating a Linear Model and Checking Linear Assumptions

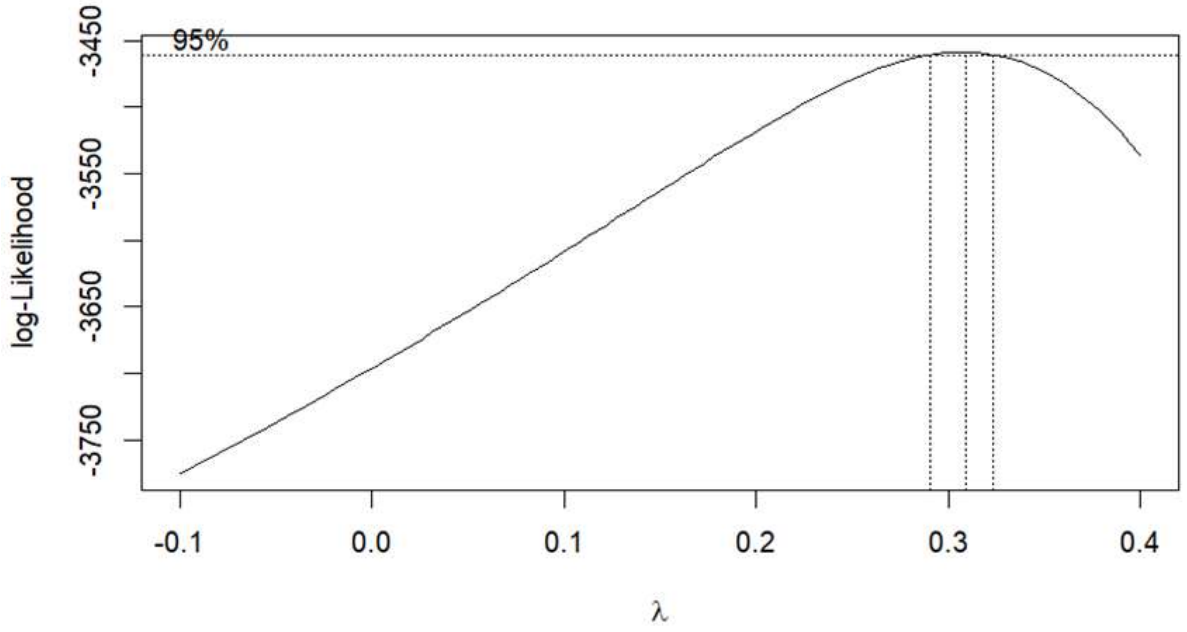
We used information from the linear model to produce residual plots. These plots allow us to validate the hypothesis that the model in its current form fails to meet the linear assumptions 1 and 2. As for the other assumptions, it is reasonable to assume each data point is independent of other data points due to the nature of the data. Assumption 4 can be assessed later after adjusting for 1 and 2.



Because we have determined that neither linear assumptions 1 or 2 are met, this must be compensated for. The first step in compensating for these assumptions is transforming the response variable to correct for assumption 2: errors have a constant variance. The residuals against fitted values plot is used to confirm that the variance of the data points from the x-axis, the error mean, does vary as we move horizontally. This further confirms that this transformation is needed and helps guide us in selecting our likely lambda value range for our box-cox plot.

Step 3: Determining the transformation of the response variable

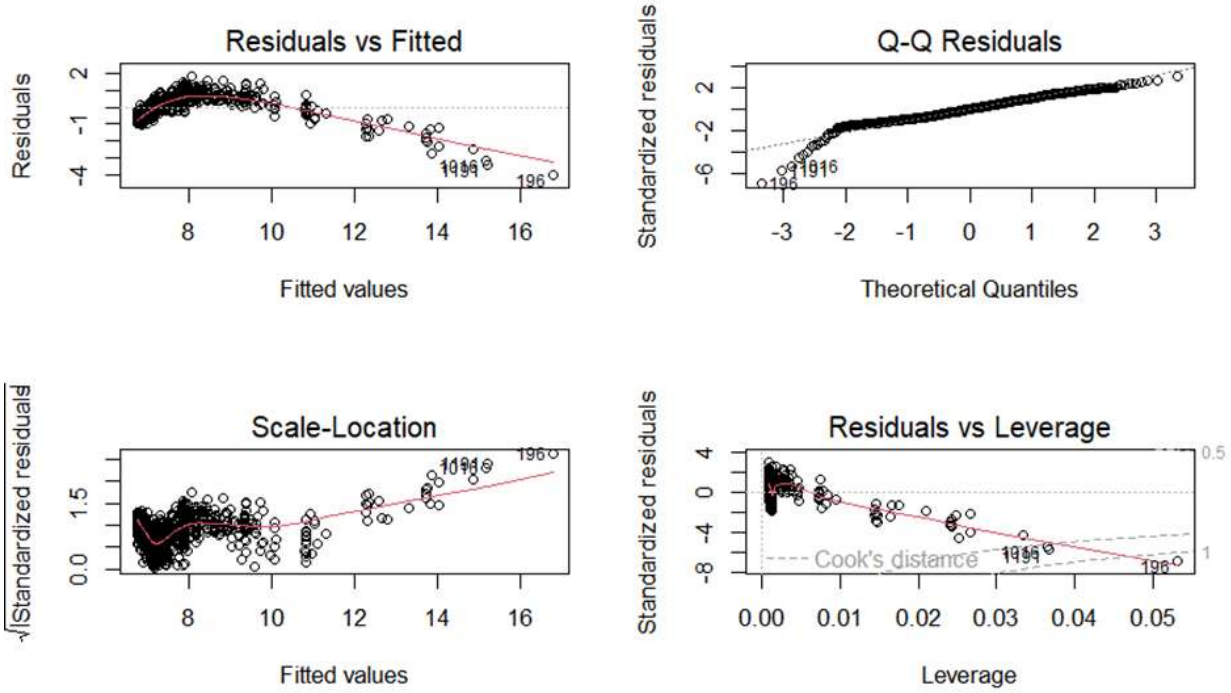
We use the box-cox plot to determine the transformation required to correct the response variable for assumption 2.



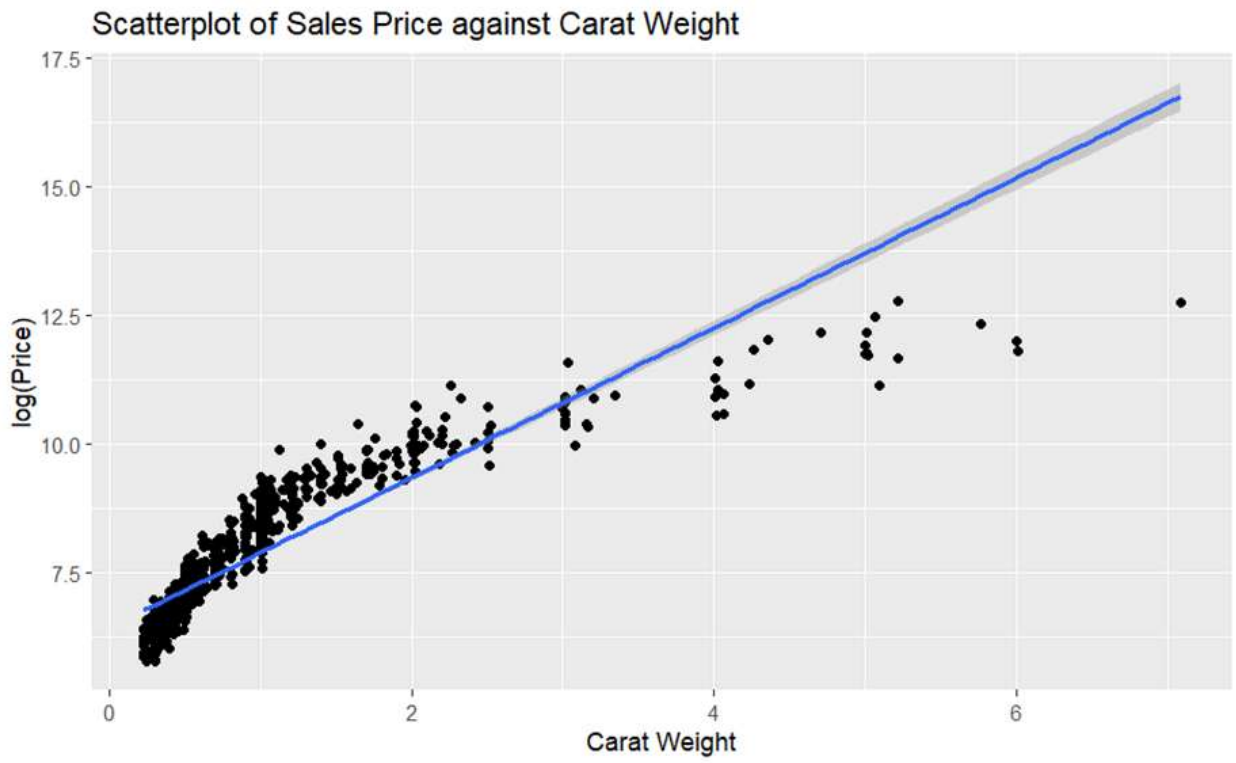
Based on the box-cox plot, we determined that applying a natural log transformation of the response variable is appropriate. Although zero is not within the confidence interval, we found it appropriate to round down to 0 to apply this transformation, a standard practice we confirmed via literature research. This decision's primary goal was to preserve the ability to interpret the regression coefficients meaningfully after the transformation.

Step 4: Determining the transformation of the predictor variable

We replot the residuals and assess the theory that assumption 1 is still unmet.



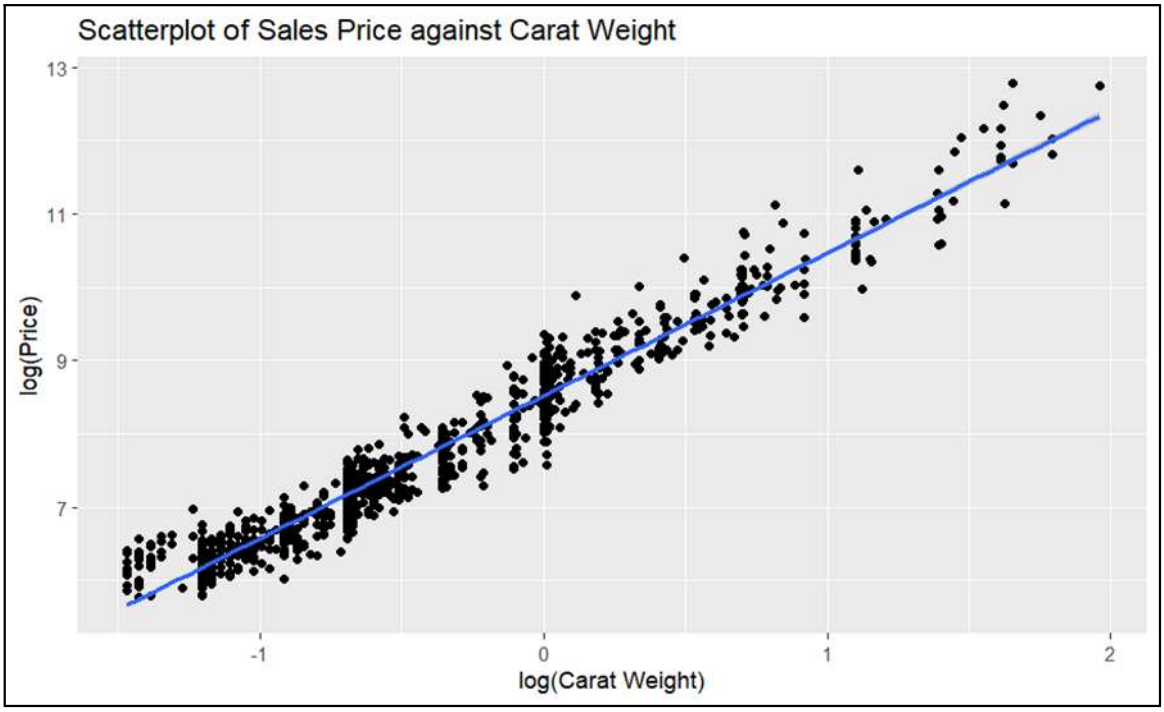
We assert that the assumption of errors having a mean equivalent to zero is not met. This assertion is made on the grounds of an uneven distribution of residuals on either side of the horizontal axis. To assess the next step, a new scatterplot is generated of the transformed response of price against carat weight. The shape of this latest plot will allow us to take an educated approach to moving forward with the transformation of the predictor.

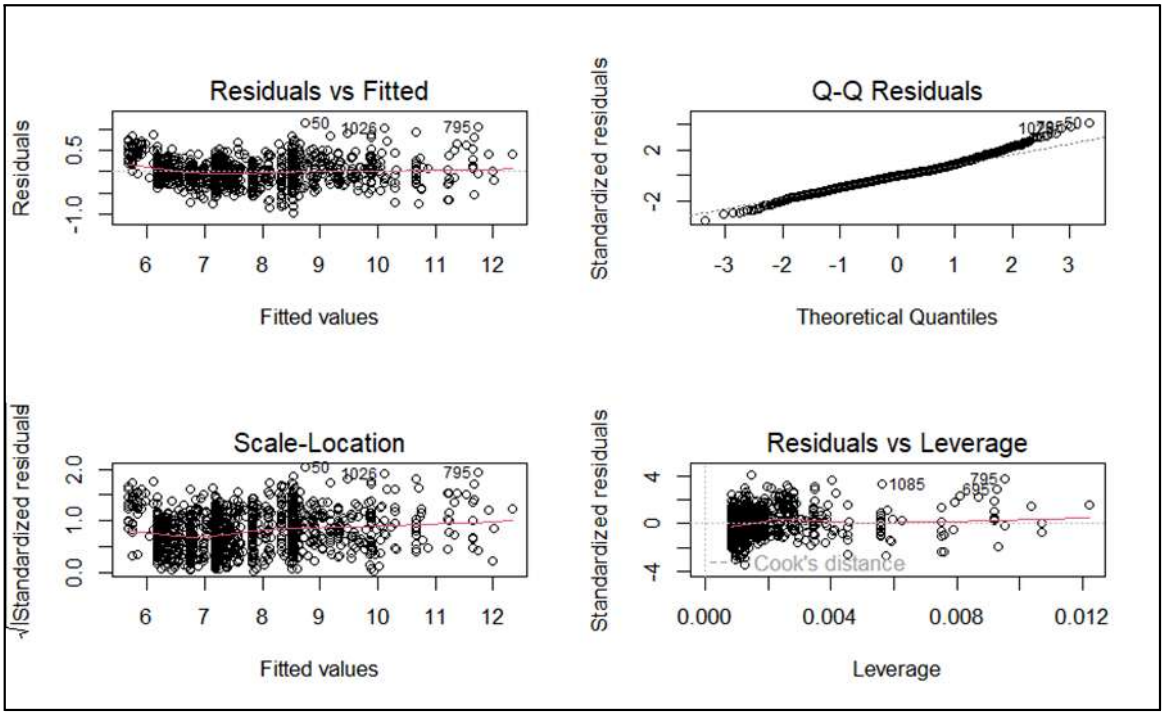


Using the shape of the scatterplot, we determine that a natural log transformation of the predictor variable (Carat Weight) is appropriate.

Step 5: Transformation of Carat Weight and Final Confirmation of Assumptions

We plot the results of the log transformations to both variables to confirm that all linear regression assumptions are now met.





Using the scatter plot and residual plots for the transformed variables, we assert that the assumptions mentioned previously are now met. Along with these, summary data of the adjusted model yields a p-value of 2.2e-16, suggesting a highly significant model. We are now highly confident in using this linear model to make inferences from the modified data set when controlling for all other contributing variables that can influence price.

Conclusion:

$$y^* = 8.521 + 1.944x^*$$

There is a strong positive linear relationship between the price of the diamond and its carat weight. The regression equation showing this relationship is:

where: y* = ln(y) and x* = ln(x) such that the predicted price is e^y*

Interpreting this log-log transformation of both variables indicates that for every 1%, 2mg, increase in the carat weight of the diamond, the price increases by 1.944% from the previous price point. This concept is expanded upon in section 1 of this report: “if a 1-carat diamond is priced at $5037.60, it would be fair to expect its price to increase by around $97.50 for a diamond weighing 2mg more when accounting for carat alone.”