How Do The Four Most Important Characteristics of Diamonds Factor Into Their Sales Price?

Tyler Hinnendael (dcc7qe@virginia.edu)

Shrikant Mishra (stm5ne@virginia.edu)

Matt Scheffel (mcs9ff@virginia.edu)

Rexwell Minnis (rlm4cr@virginia.edu)

University of Virginia School of Data Science

STAT 6021 Linear Models for Data Science

October 16th, 2022

Section 1: Executive Summary

This statistical analysis focuses on the prices of diamonds in a dataset from Blue Nile. The purpose of this analysis is to use various statistical practices and data visualizations to explore how the price of a diamond relates to its other major variables: carat weight, clarity, color, and cut. These variables are commonly known as the "4 C's." The dataset features 1214 observations (diamonds) and their associated 5 variables. Some additional variables were created to reflect any transformations made to the data. In addition to exploring these relationships, the analysis also addresses certain claims made about diamonds on the Blue Nile website. There are four major claims with implications for this analysis:

1) a diamond's cut is the most important characteristic for a diamond relative to price;
2) color is the second most important characteristic for a diamond;
3) clarity is the least important characteristic for a diamond;
4) the higher the quality of a diamond, the higher the price will be.

For the project, three major analyses were carried out: univariate analysis, bivariate analysis, and multivariate analysis. For our calculations, log transformations were applied to the price of the diamond and this new variable was compared to the 4 C's.

The univariate analysis gives very simple observations of the different diamond variables with results stemming from frequency graphs. Results show that a majority of the observations in the dataset fall between 0-2 carats. The most frequent styles of cuts are "Ideal" and "Very Good." ("Astor Ideals", on the other hand, were very infrequent.) The color of the diamonds were more evenly distributed, with "F", "D", and "G" being the most frequent. The clarity of the diamonds tended to be skewed towards the lower quality end, with SI1, VS1, and VS2 being the most frequent. (FL and IF diamonds were very rare in the dataset.)

The bivariate analysis compares the price (with log transformation applied) to the 4 C's. The analysis of price and carat weight confirms that an increase in carat weight leads to an increase in price. The relationship, however, is not exactly linear. As the carat weight increases, price tends to increase at a faster rate. The analysis of price and cut does not show a significant pattern amongst the different styles of cut - indicating the claim that cut is the most important characteristic of a diamond is inconsistent with our data. The analysis of price and color showed a clear trend that as the quality of the diamond's color decreased, so did price. This affirms the website's claim that color is a highly important factor for a diamond's price. The analysis of price and clarity, aside from the "FL" outliers, did not establish a significant trend or pattern. This tends to hold with the website's claim that clarity is the least important of the variables.

The multivariate analysis looked deeper into the relationship between color and the other variables. Using a "price per carat" variable , the analysis found that diamonds with a "D" color and a "Very Good" cut had the most expensive price per carat. This helped solidify the finding that cut was not the most important quality - otherwise the Astor Ideal diamonds would have had the most expensive price per carat. The analysis also found that diamonds with D and FL clarity grades have a significantly higher price per carat, and that colorless + "H" diamonds have the most expensive prices per carat for all clarity grades.

Finally, our simple linear regression analysis for price against carat (after transforming some of the variables) determined the following linear model: **y* = 1.944x* + 8.521,** where y* = log(y) and x* = log(x). In this model, y is the price and x is the carat weight. Essentially, for every 1% increase in carat weight, the estimated price will approx. increase by 1.94%. Through an ANOVA F test, we found that there was a linear relationship between the final transformed variables for carat weight and price.

Section 2: Data and Variables

The "diamonds4" dataset includes 1214 observations (different diamonds) and 5 features which make up the columns of the dataset. The 5 features are a diamond's price, carat weight, cut, color and clarity. The numerical variable "price" describes the price of each diamond measured in dollars.

Carat (Weight)

The numerical variable "carat" describes the weight of the diamond, measured in carats. The metric equivalency of the carat system allows for a level of uniformity throughout the world. Although carat weight is not synonymous with size, sizes tend to be similar for diamonds of similar weight. The value of a diamond tends to rise disproportionately as the weight increases due to the rarity and limited supply of larger diamonds in the world. In this data set, the minimum diamond weight is 0.23 carats, the maximum weight is 7.09 carats, the mean weight is 0.813 carats, and the median weight is 0.52 carats.

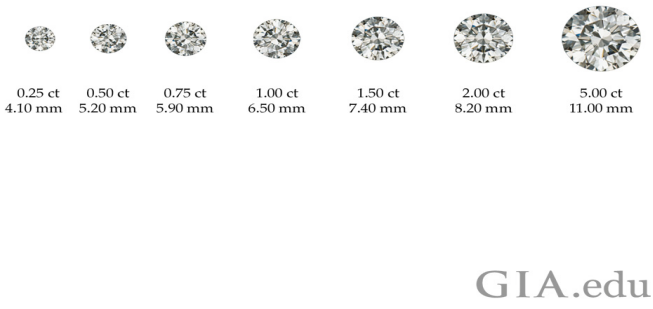

Figure 1

Clarity

The categorical variable "clarity" refers to the amount of blemishes and inclusions present in a diamond and how visible they are at different levels of magnification. The grades or sub-categories for the "clarity" variable, in descending order, are: FL (flawless), IF (internally flawless), VVS1 (very very slightly included), VVS2, VS1 (very slightly included), VS2, SI1 (slightly included), SI2, and I (included). This data set includes observations ranging from FL to SI2. Flawless diamonds - the highest grade on the clarity scale - have no inclusions or blemishes visible under 10x magnification. VVS and VS diamonds have minor inclusions that are difficult to see under 10x magnification. SI1 and SI2 diamonds have inclusions that can sometimes be seen by the naked eye. Included diamonds -

the lowest grade on the clarity scale - have inclusions that are obvious at the 10x magnification level and can sometimes be seen by the naked eye.The most frequent clarity grade in this data set is “SI1” with 20.02% of the observations.

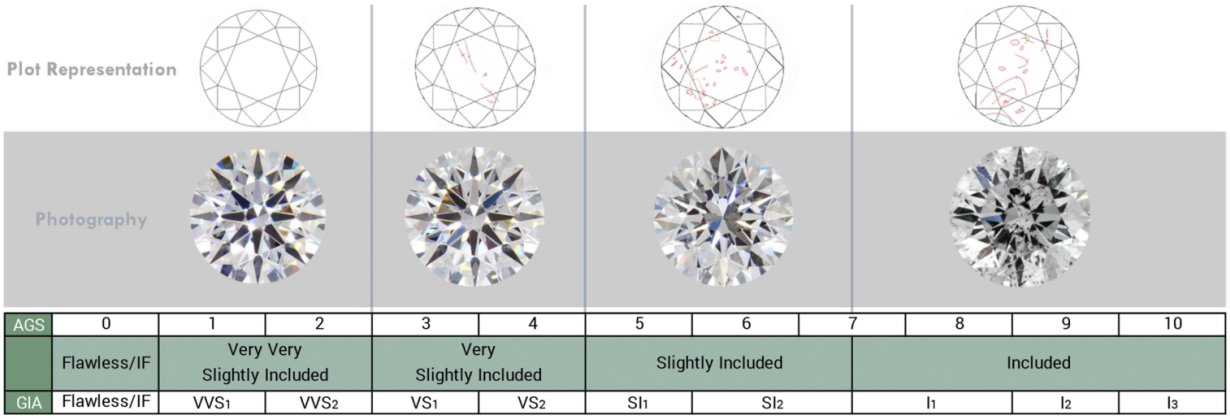

Figure 2

Color

The categorical variable “color” refers to the color grade of the diamond. This diamond color scale ranges from D-Z, with the highest quality diamonds being colorless “D” diamonds. A diamond is considered ‘colorless’ if its color grade is between D-F and it is considered ‘near colorless’ if its color grade is between G-J. Diamonds with a color grade from K-Z tend to have varying noticeable tints of yellow. Colorless stones tend to command the highest prices. All observations in the dataset range for this analysis are either colorless or near colorless. The most frequent color grade in this data set is “F” with 18.37% of the observations.

Figure 3

Cut

The categorical variable "cut" refers to the shape, faceting, symmetry, proportion and finish of a diamond. Diamonds with the highest quality cuts maximize the amount of light that is reflected from the diamond. The grades or sub-categories for the "cut" variable are, in descending order: Astor Ideal, Ideal, Very Good, and Good. The most frequent cut grade in this data set is "Ideal" with 60.87% of the observations.

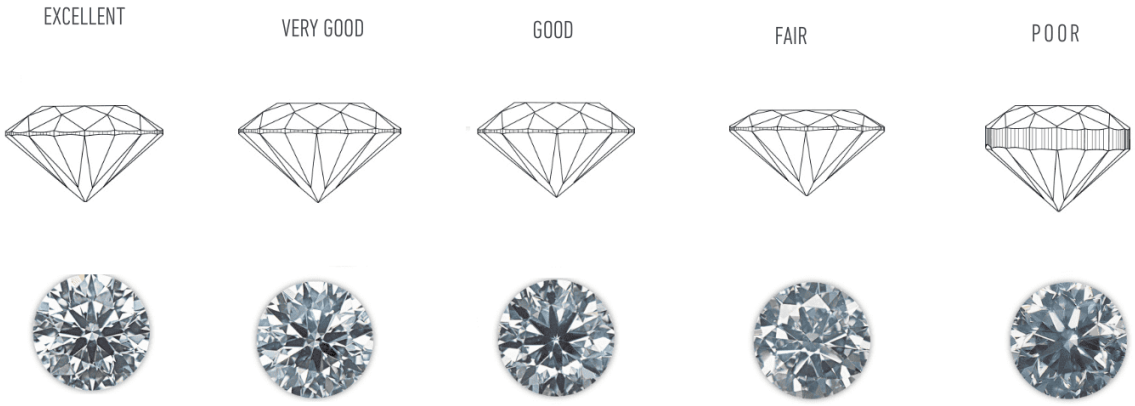

Figure 4

Price

The price is influenced by the other four features. In this data set, the minimum price is $322, the maximum price is $355,403, the mean price is $7,056.74, and the median price is $1,463.50. The disparity between the mean and median price show the presence of outliers in the data set.

<u>Transformations</u>:

log_price

Additional variables were created in the process of data exploration. A variable called "log_price" was created to transform the price values into a condensed scale which better represents the values without such a large influence from outliers. The variable "log_price" was calculated using the log function on each 'price' observation.

price_per_carat

A variable called "price_per_carat" was created to normalize each diamond price in terms of its weight in carats. This variable preserved the context of diamond weight while allowing visualizations of the other three C's. The "price_per_carat" variable was calculated by dividing each 'price' observation by the corresponding 'carat' observation.

Visualizations and Descriptive Statistics:

Univariate Analysis

The frequency plots below show the count of each of the 4 C's sub-categories. The graph in the top-left corner shows the frequency of the diamond "carat" weights. Nearly all of the observations are between 0-2 carats. The graph in the top-right corner shows the frequency of the diamond "cuts". The most frequent cuts are "ideal" and "very good", with very few "Astor Ideal" diamonds. The graph in the bottom-left shows the frequency of the diamond "colors". This data set includes a wide spread of multiple observations of each color grade. The most common colors are "F", "D" and "G", although even the least common grades represent a decent amount of the observations. The bottom-right graph shows the frequency of the diamond "clarity" grades. The most common grades are on the lower-quality end of the scale, being SI1, VS1, and VS2. There are very few FL and IF diamonds in the data set.

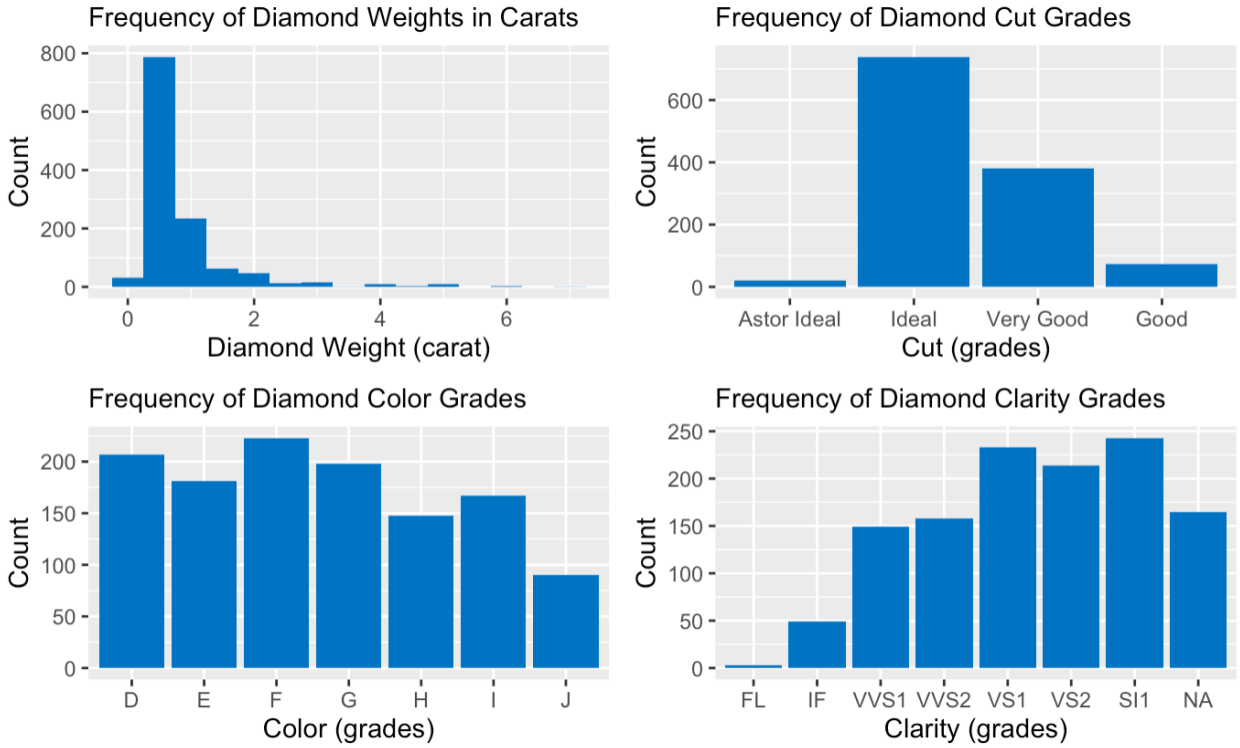

Figure 5

Bivariate Analysis

The scatter plot and bar charts below show how price is related to the other variables: carat weight, cut, color, and clarity. The relationship between the numerical variables - price and carat - is examined using a scatter plot with price in its original form (i.e. before the log transformation). The relationship between price and the other three categorical variables is explored using a bar chart and the mean of the log(price) values for each variable's sub-categories. The calculations were as follows,

using the top-right graph as an example. First, the log(price) was calculated for every observation. Second, each observation was grouped by the type of diamond cut: Astor Ideal, Ideal, Very Good, and Good. Third, the mean(log(price)) was calculated for each group. Fourth, the mean(log(price)) is conveyed on the y-axis of each bar chart to show the impact of each grade on the diamond's price.

The top-left graph shows a scatter plot of price and diamond weight in carats. The plot confirms the description of the impact an increase in carat weight has on price - this is not a linear relationship, because price increases at a faster rate as the weight increases. We can see that most diamonds weigh less than 2.5 carats and cost less than $5,000. This is consistent with the values for the median and mean of both price and carats. We also see the presence of potential outliers with values greater than 4 carats and $10,000.

The top-right graph shows the mean of the log(price) for each type of diamond cut. We see that the "Astor Ideal" cut has the highest mean price and the "Ideal" has the lowest. There is not a significant descending pattern from each grade of diamond cuts. **This is inconsistent with the claim on the Blue Nile website, which states that "cut is the most important characteristic of the 4C's"**. If that were the case, we should expect to see a pattern - the higher quality cuts should have a clearly higher price-point than the lower quality cuts. We do not see this pattern in the bar plot.

The bottom-left graph shows the mean of the log(price) for each type of diamond color. We see that the "D" color has the highest mean price and "I" has the lowest. Although "H" color has a high mean price, there is a clear trend in the data - as the quality of the color grade decreases, the mean price goes down. **This is consistent with the claim on the website that diamond "color" is the second most important of the 4 C's.**

The bottom-right graph shows the mean of the log(price) for each type of diamond clarity grade. We can see that "FL" (flawless) diamonds are significantly more expensive than diamonds of other clarity grades. The change in the mean price between clarity categories is mostly flat. The claim on the website is that clarity is the least important of the 4 C's. If this were false, we would expect to see a clear pattern between increasing clarity grades and increasing prices. Because there is not a clear pattern, we conclude that an increase in clarity does not necessarily increase the price of the diamond, with the exception of FL diamonds. **Thus, our analysis is consistent with the claim from Blue Nile that clarity has little impact on price.**

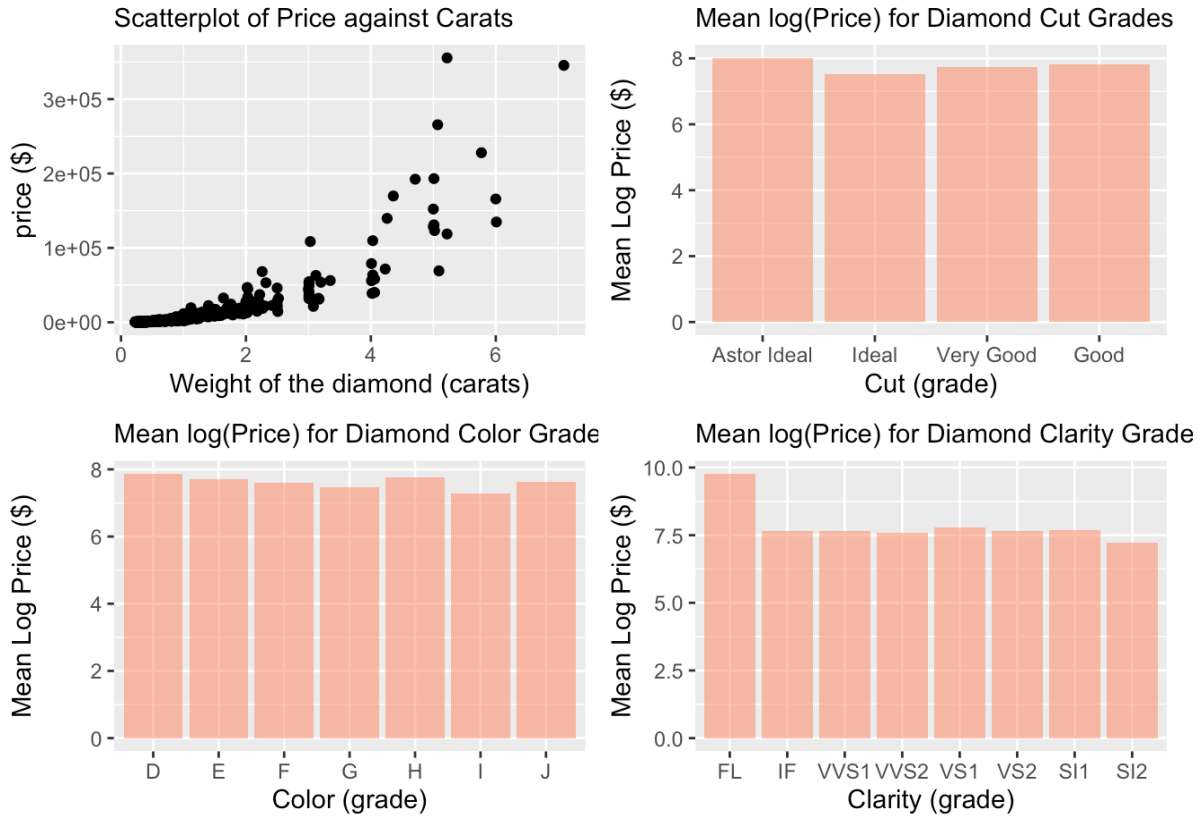

Figure 6

Based on the bivariate analysis, we decided to further explore the relationship between a diamond's color and its other features. **We support the website's claim that a diamond's color is an important feature, however, we do not see evidence that a diamond's cut is the most important feature.**

The violin plot below shows the distribution of prices for each color grade. The y-axis has been transformed to log(price) to show the distribution of the majority of the values with less influence from outliers. The x-axis shows each color grade category in descending order, from colorless to near colorless. We see that there is a general pattern in these distributions. The higher-grade colorless diamonds are generally a higher price than the lower-grade, near-colorless diamonds. **This is consistent with the bar chart above and the claims on the website.**

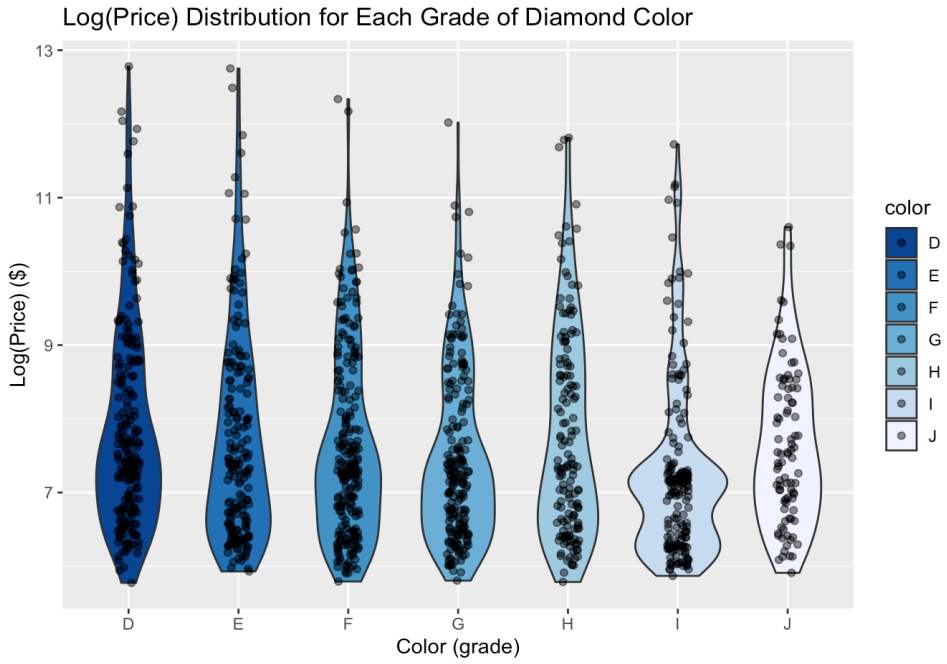

Figure 7

Multivariate Plots

To further explore the relationship between color and the other C's, we created two multivariate graphs. These graphs use the price-per-carat variable (described above) to capture the weight of the diamond in combination with the other variables that impact a diamond's price. The price-per-carat variable also scales any diamonds of disproportionately large or small size, which could have a significant impact on price. On the y-axis of both graphs below is the 'Mean Price-per-Carat.'

In the first graph, on the x-axis are all combinations of diamond color and cut grades. The graph shows that diamonds with "D" color and "Very Good" cut have the most expensive price per carat. **This is inconsistent with the website's claim that cut is the most important quality, which would imply that we should "Astor Ideal" cut diamonds be the most expensive (per carat).** The graph shows that diamonds of "I" color and "Good" cut are least expensive (per carat). We can see that the mean price per carat generally increases as color grade improves, for all types of diamond cuts.

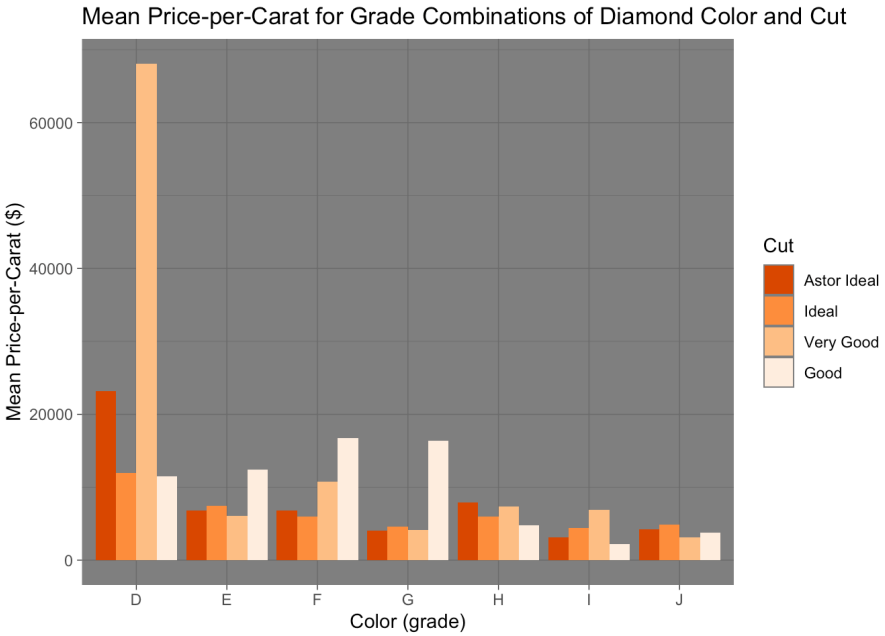

Figure 8

In the second graph, on the x-axis are all combinations of diamond color and clarity grades. The graph shows that diamonds that are “D” and “FL” are significantly more expensive (per carat) than all others. We can see that colorless diamonds, and “H” diamonds, are the most expensive (per carat) for all grades of clarity.

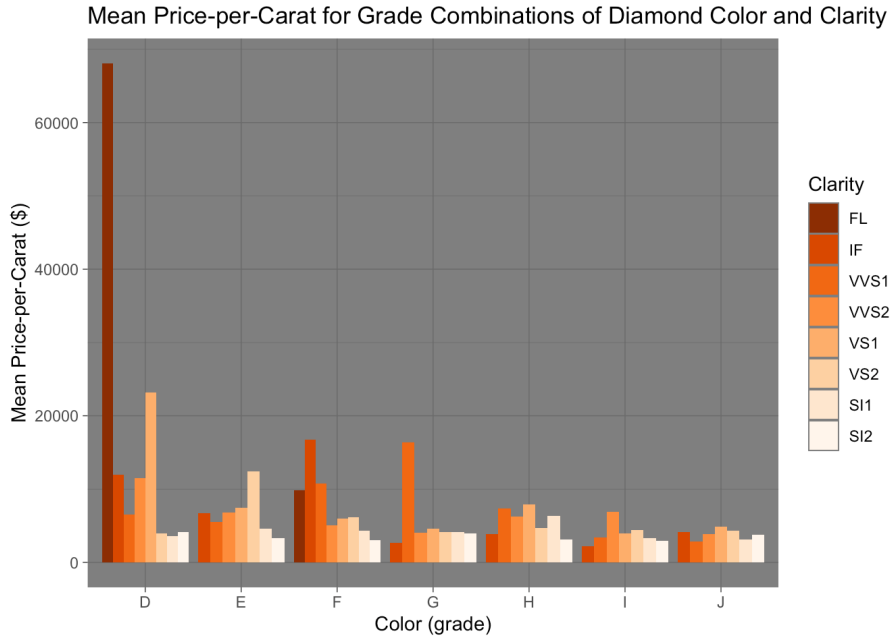

Figure 9

Section 3: Simple Linear Regression for Price against Carat

In our analysis we assume that price is the dependent variable and carat, clarity, color and cut are independent variables.

Price Against Carat Weight:

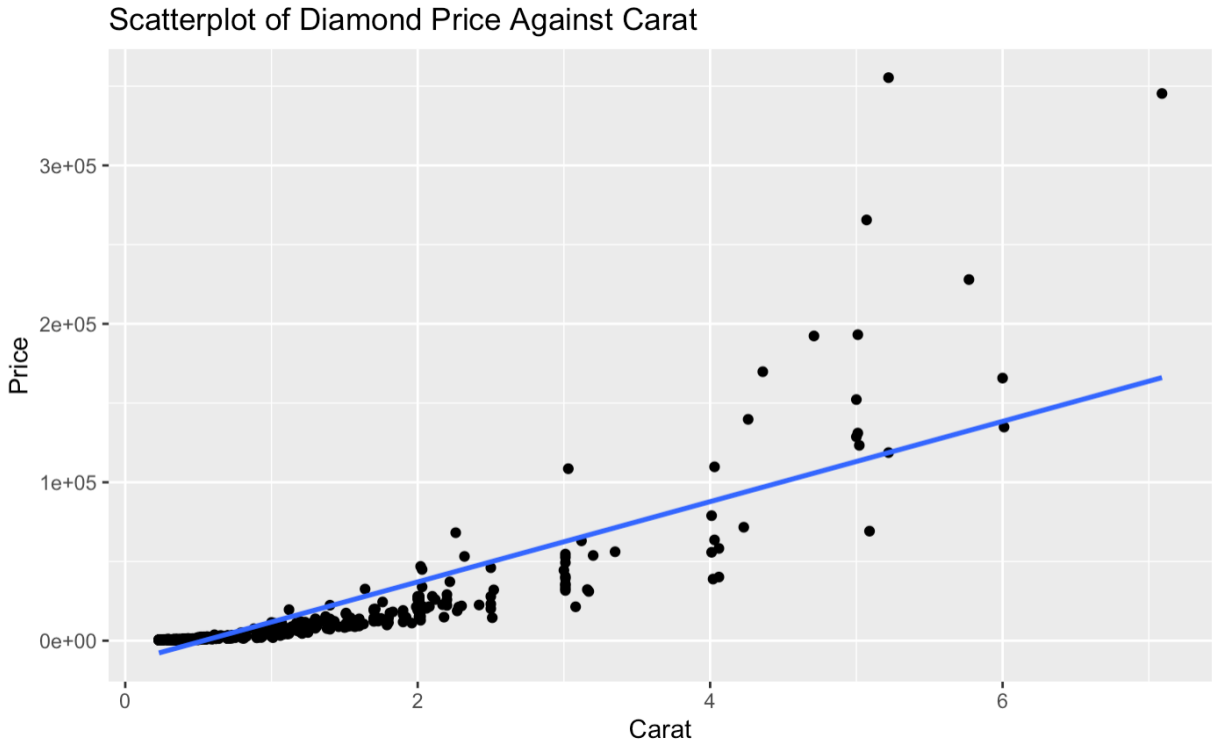

Figure 10

Just from a visual inspection of our plot, we can already tell that certain assumptions are not going to be met. First we will check for the first two SLR assumptions. A residual plot can help us assess these first two assumptions:

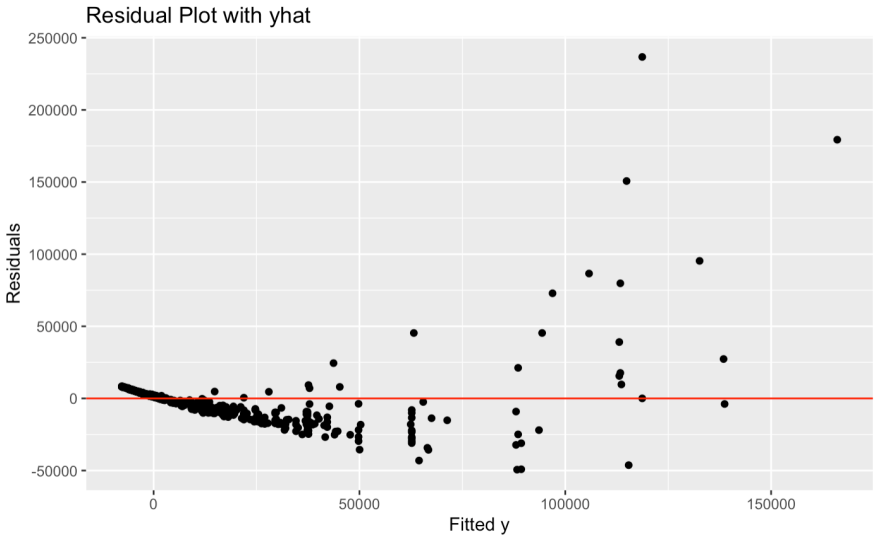

Figure 11

Assumption 1 is about the errors in the residual plot having a mean of 0. Since the points do not seem to be randomly scattered over and under the x axis, and because there seems to be a curvature pattern, we can say that $E(\epsilon)$ does not equal 0, so Assumption 1 is not met.

Assumption 2 is about having a constant variance among the values of y-hat. Since the spread of the points on the plot seems to increase as we go along the horizontal axis, Assumption 2 is not met either.

Since both assumptions are not met, we will first try to transform the response variable, price, to try to meet Assumption 2. For this, we will need to first look at a Box Cox plot of our current linear model.

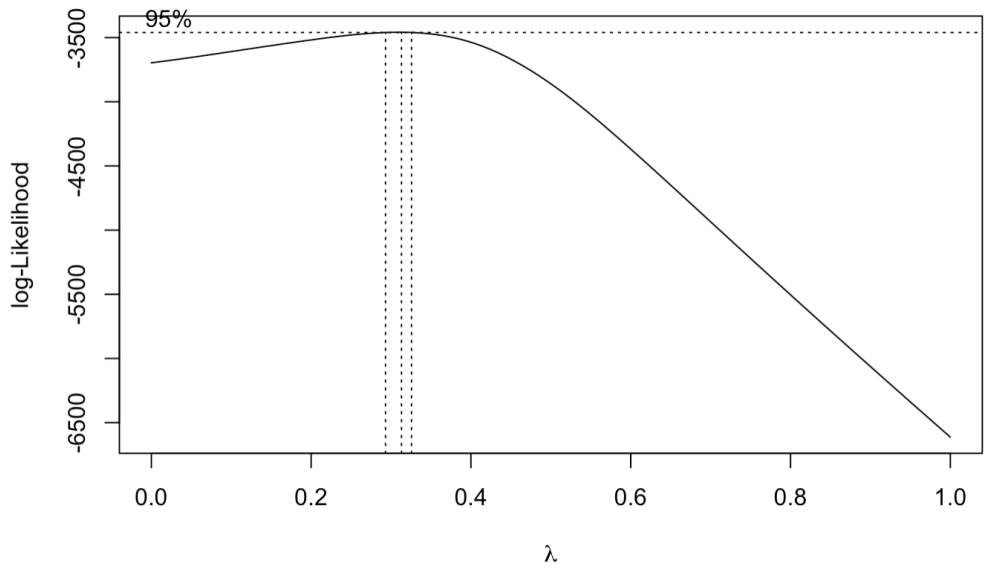

Figure 12

Because the value of the suggested lambda seems to be close enough to 0, and because using a log transformation can lead to more interpretable results, we can just assume $\lambda = 0$. In that case, according to the box cox method, we must transform our response so that $y^* = \log(y)$.

The new residual plot with y*:

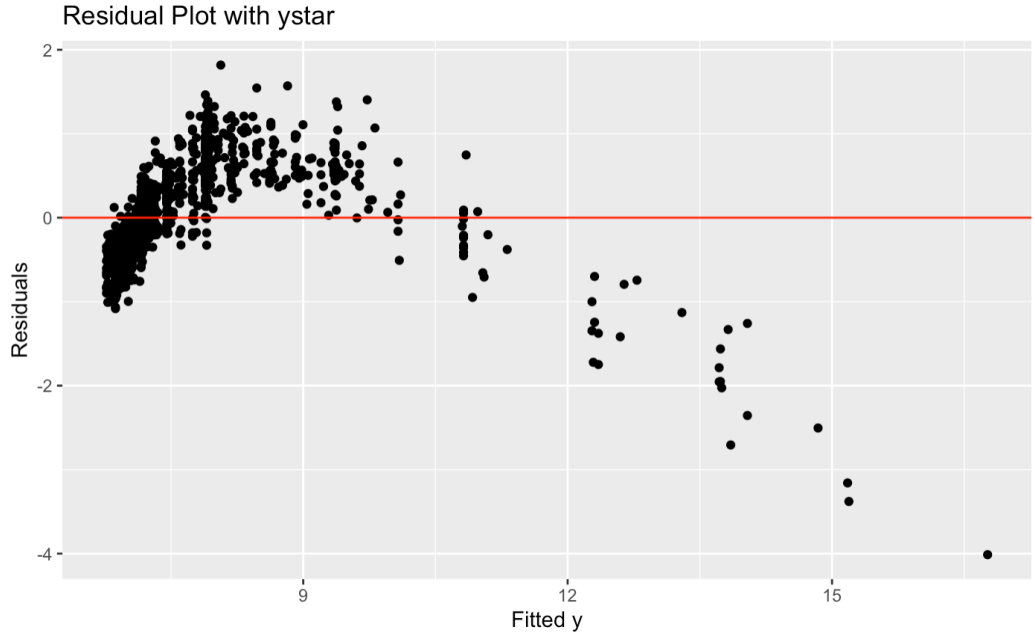

Figure 13

Looking at this new residual plot, we can see that the variance is now more constant as we move along the x axis of the plot. Assumption 2 is now met.

To meet Assumption 1, we must transform the predictor variable, which is carat. Taking a look at the new plot, which now compares y* against carat weight:

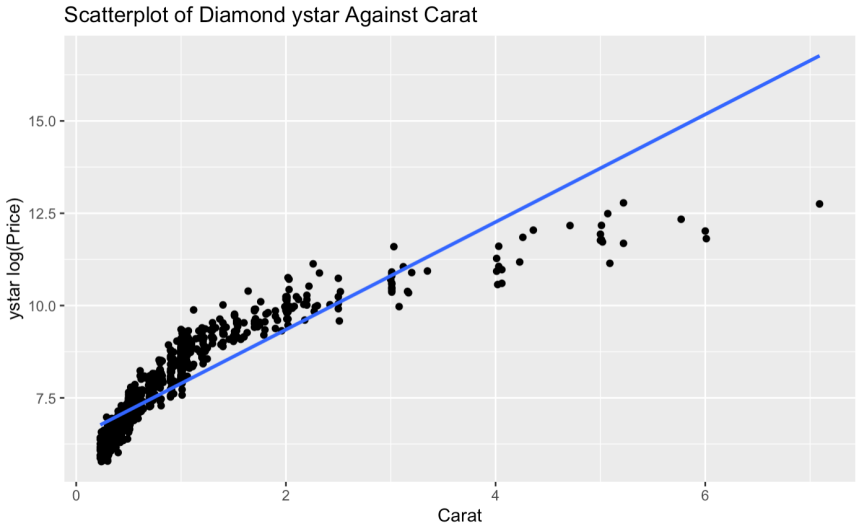

Figure 14

Looking at the plot above, we can see that the points seem to follow a logarithmic pattern. Thus, we can transform the predictor so that x* = log(x).

Taking a look at the new residual plot after transforming the predictor:

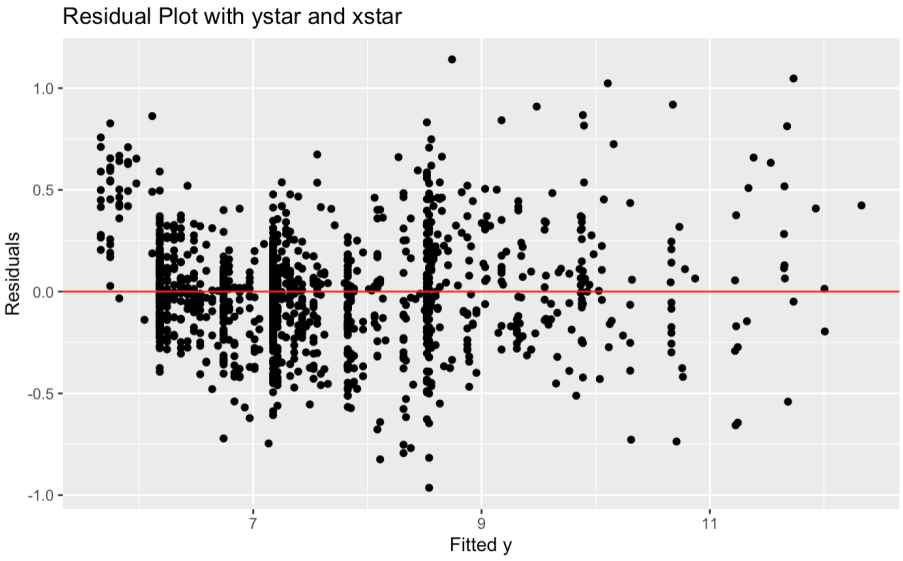

Figure 15

The points seem to be equally spread out above and under the x axis, thus Assumption 1 is met. The spread is still relatively constant as we move along the x axis so Assumption 2 is also still met.

Since both of the first 2 assumptions have been met, we no longer need to transform our predictor variable or our response variable. The next step is to test that our errors are independent, or Assumption 3.

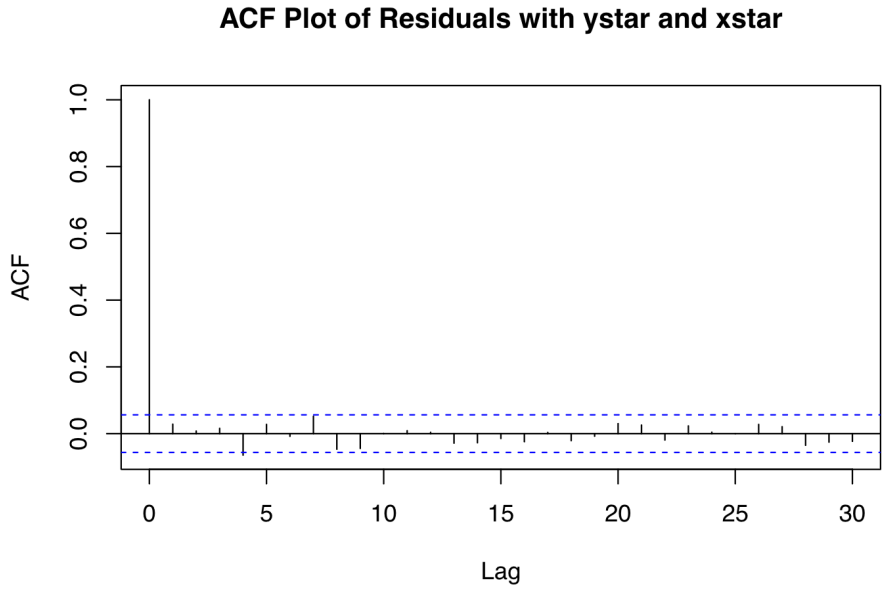

Figure 16

There is one lag value where the line crosses the blue band, but all other lag values seem to be within the blue band. Since we have no information that the Blue Nile diamond dataset was not randomly selected, we can confirm that Assumption 3 is also met.

Finally, to check that the errors follow a normal distribution, we can plot a Q-Q Plot. If the sample quantile points line up close enough to the theoretical line, we can assume that Assumption 4 is met.

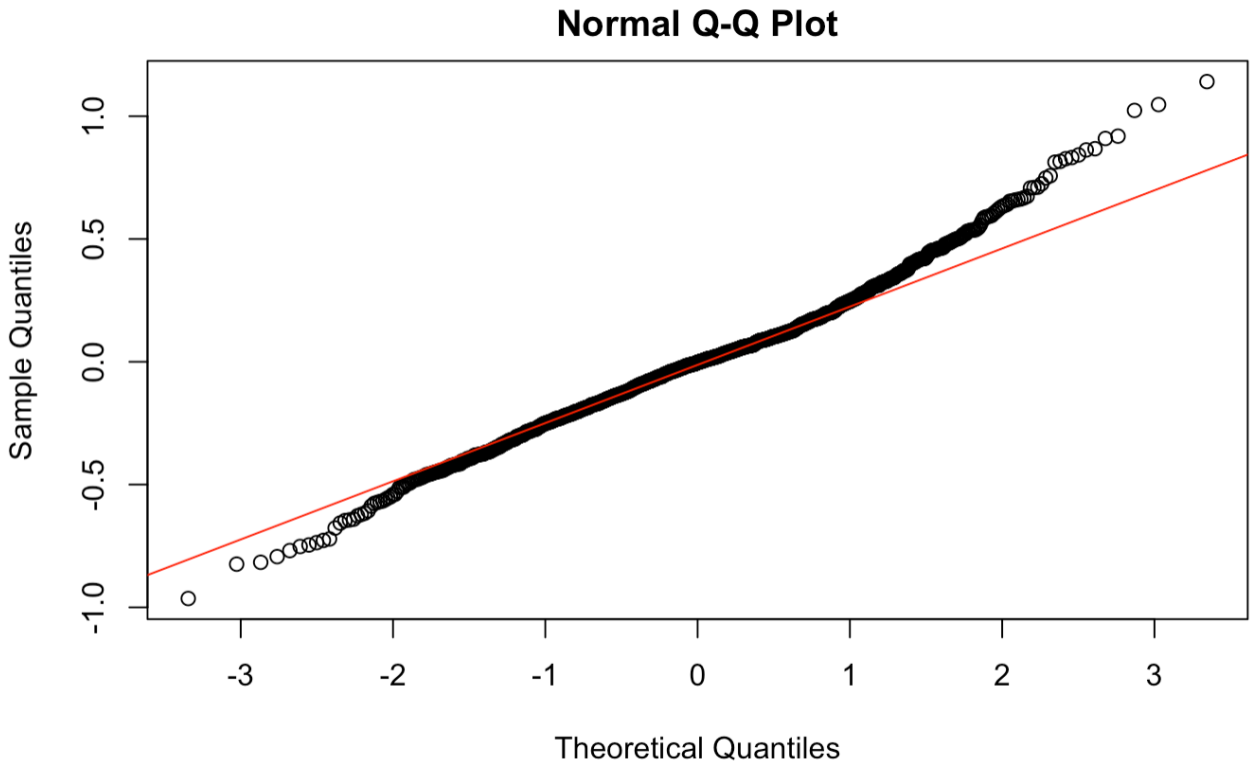

Figure 17

Even though the sample quantile points in the plot differ from the theoretical line towards the ends of the horizontal axis, they seem to match up with the line for most of the values, so we can confirm that Assumption 4 is met.

Now that all of the Simple Linear Regression assumptions have been confirmed as met, here is the plot of our transformed price against transformed carat weight:

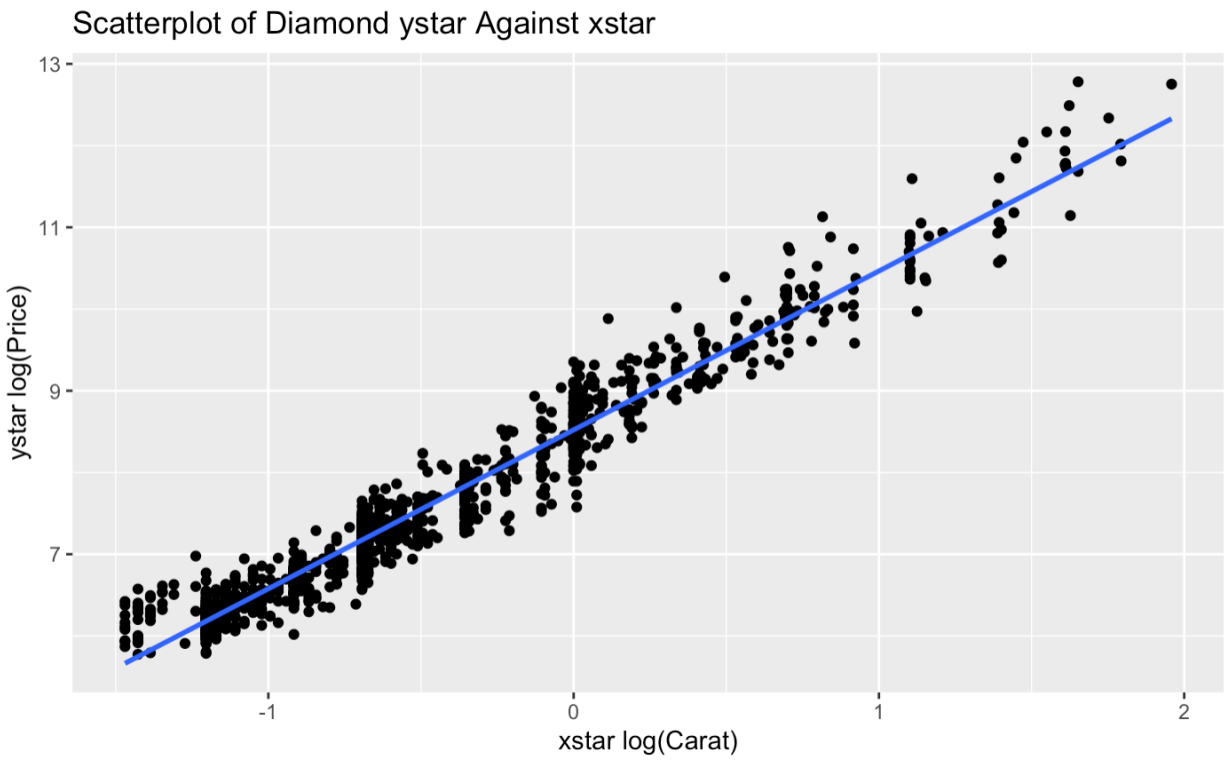

Figure 18

The final linear model that we propose is **y* = 1.944x* + 8.521,** where y* = log(y) and x* = log(x). Here, y is the price and x is the carat weight.

Because we did a log transformation on both our predictor and response, our regression equation is still interpretable. The slope of the regression, which is $\widehat{\beta}_1 = 1.944$, means that for every 1% increase in the carat weight, the estimate of the price will approximately increase by 1.944 %. The intercept of this equation, when x* is equal to 0, is 8.521.When x* = 0, the carat weight is equal to 1. Thus, $\widehat{\beta}_0 = 8.521$ means that when carat weight is 1, the model estimates the price to be $e^{8.521} = \$5{,}019.19$. In this model, predicting the price when the carat weight is 0 is not interpretable.

Conducting a hypothesis test to confirm whether there exists a linear relationship between our transformed variables, we will use the ANOVA F test with $\alpha = 0.05$ and the following hypotheses:

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

The results of the hypothesis test are as follows:

```
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
```

Figure 19

The p-value for the ANOVA F test is 2e-16, which is less than $\alpha = 0.05$, so we reject our null hypothesis. We have sufficient evidence that there is a linear relationship between the transformed variables, x* and y*.

Additionally, from the output above, our model has an R^2 value of 0.9547, which shows that approximately 95% of the variation in the transformed response variable y* is covered by our transformed predictor x*.

Finally, our proposed model also has a correlation value of 0.977. This shows that there is a strong, positive association between the transformed variables.

Conclusion:

For our final conclusion, we return to the four major claims with implications for this analysis from the Blue Nile website:

1) a diamond's cut is the most important characteristic for a diamond relative to price;
2) color is the second most important characteristic for a diamond;
3) clarity is the least important characteristic for a diamond;
4) the higher the quality of a diamond, the higher the price will be.

Our analysis found Claim #1 to be inconsistent with our results. While Blue Nile's website asserted that cut is the most important characteristic with respect to price, our data analysis did not reflect this. No significant descending pattern was found amongst the grade of diamond cuts. If this claim was true, we would expect our data analysis to demonstrate a clear pattern of increasing price points as diamond cut quality improved. After a log transformation to the price, even the highest quality of cut (Astor Ideal) has a barely distinguishable difference in price from the lowest quality of cut. While diamonds with the highest quality cut are certainly on the most expensive end of the spectrum, our data does not support that the quality of the diamond's cut is the leading factor for the higher price. The multivariate analysis looked deeper into the relationship between cut and the other variables.

Using a “price per carat” variable , the analysis found that diamonds with a “D” color and a “Very Good” cut had the most expensive price per carat. This helped solidify the finding that cut was not the most important quality - otherwise the Astor Ideal diamonds would have had the most expensive price per carat.

Our analysis found Claim #2 to be consistent with our results. Blue Nile’s website stated that color was the second most important characteristic, and our analysis reflected the idea that color played an important role in determining the price of the diamond. A clear trend is evident in the data that as the quality of the color grade decreases, the mean price decreases as well. This trend is seen in both the bivariate and multivariate analyses.

Our analysis found Claim #3 to be consistent with our results. Blue Nile’s claim that clarity was the least important factor in the price of the diamond held up with our analysis. While Flawless diamonds carry a significantly higher price than other grades of clarity, the remaining grade levels from our data demonstrate no clear pattern between increasing clarity grades and increasing prices. This lack of a trend can be seen in both the bivariate and multivariate analyses.

Our analysis found Claim #4 to be consistent with our results. While this is a more vague claim than the other 3, the data still lends support to the claim that attaining a higher grade tends to result in a higher price - particularly when the grade is the highest grade that can be attained. For example, the Astor Ideal cut grade stands out significantly from its peers in regards to the price of that diamond. Flawless cut diamonds and D color diamonds also outperform their relative peer diamonds. And, the diamonds with the heaviest carat weight trend towards being the most expensive. While the highest grade categories for each characteristic may result in some outliers, the data surrounding them still suggests that “high quality” will result in “high price.”

References:

https://www.bluenile.com/education/diamonds

Images:

- https://beyond4cs.com/color/
- https://4cs.gia.edu/en-us/blog/4cs-diamond-quality-most-important-c/
- https://www.americangemsociety.org/buying-diamonds-with-confidence/4cs-of-diamonds/understanding-diamond-clarity-the-4cs-of-diamonds/diamond-clarity-scale/
- https://www.gemonediamond.com/diamond-cut/
- https://diamondsonthekey.com/the-4cs-of-diamonds/