STAT 6021: Data Analysis for Diamond Prices

Summary of Findings

According to the Blue Nile page, the order-of-importance of variables in influencing diamond price is Cut-Color-Clarity-Carat, with cut being the most important variable. While our visualizations demonstrate that cut is a very crucial factor (much more so than clarity and color), it does not appear to be the most important variable overall. In fact, each diamond is affected by all of its characteristics, which distinctly impacts a diamond's grade.

We were unable to establish a definitive relationship between color and price, nor between clarity and price. When we compared the median prices of each color group against one another – from our analysis, we found no significant differences between the medians amongst the group (i.e. no one color stood out as noticeably more expensive). Similarly, comparing the cuts of diamonds also shows no significant differences between the medians of the groups. An exception to this is the "Astor Ideal" group, which has a larger median than the other groups. Importantly, the sample size of "Astor Ideal" diamonds was only 20 diamonds, which is relatively small compared to the size of the overall sample; this might potentially skew the data. This was the only notable difference, while all other categories of cut showed similar prices. This suggests that, while cut may have a slight influence on the pricing of a diamond, especially if the diamond is an "Astor Ideal" cut, there are many different factors at play and cut is not a determining factor.

Interestingly, despite Blue Nile's claim that the carat weight is the least essential aspect, our analyses actually revealed a high correlation between the carat weight of the diamond and its price. Through statistical analysis, we were able to establish a positive linear relationship between carat weight and diamond price, which means that as carat weight increases, so does diamond price. According to our data, every 1% increase in carat weight resulted in a 1.94% increase in diamond pricing. We conclude that the carat weight is the most influential factor in determining the price of a diamond. However, the price is also a result of a variety of factors, other than carat weight, that affect it uniquely with no clear order of importance.

Data Description

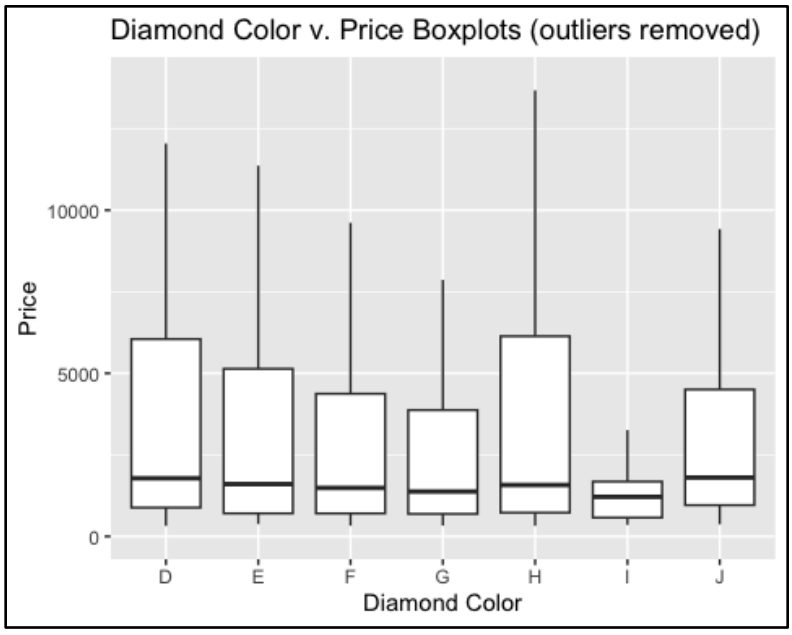
This dataset contains information about diamonds that are sold on BlueNile.com. There are over 1000 diamonds listed in this dataset. Diamonds are characterized by the 4 C's: cut, color, clarity

and carat weight. We will be conducting analysis on how these 4 characteristics affect the price of a diamond.

- Carat: The diamond's carat refers to the diamond's weight. One diamond carat is equivalent to about 200 mg of a diamond.
- Clarity: The clarity of a diamond refers to the inconsistencies on a diamond's surface and within the stone. Clarity comes with the following 6 categories:
  - FL - Flawless (Most shiny)
  - IF - Internally flawless
  - VVS1 / VVS2 - Very Very Slightly Included
  - VS1 / VS2 - Very Slightly Included
  - SI1 / SI2 - Slightly included
  - I1 /I2 /I3 - Included
- Color: The color of a diamond refers to how colorless a diamond is. Typically, the less of a warm hue there is, the higher rated the diamond is.
  - Colorless diamonds: The rarest and highest quality with a pure icy look:
    - D / E / F
  - Near-colorless diamonds: No discernible color; great value for the quality.
    - G / H / I / J
  - Faint color diamonds: Budget-friendly pick; pairs beautifully with yellow gold.
    - K
- Cut: The cut of a diamond measures how well-proportioned and positioned the dimensions of a diamond are. This factor looks at the ratio of a diamond's diameter relative to its depth. "Astor Ideal " is the best cut, then "Ideal ", "Very good", and "Good" respectively.
- Price: The price of a diamond refers to how much the diamond is being sold for.

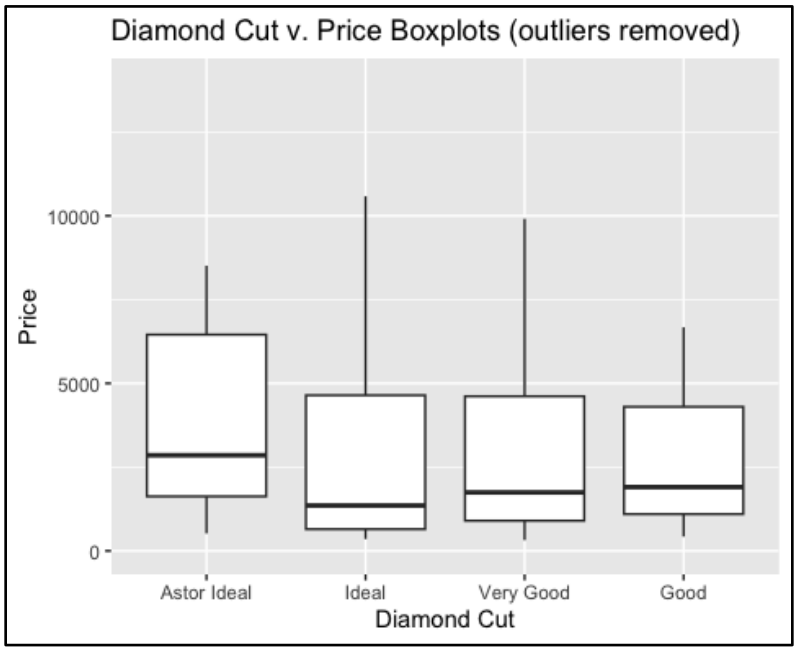
Data Visualizations

- Box Plots



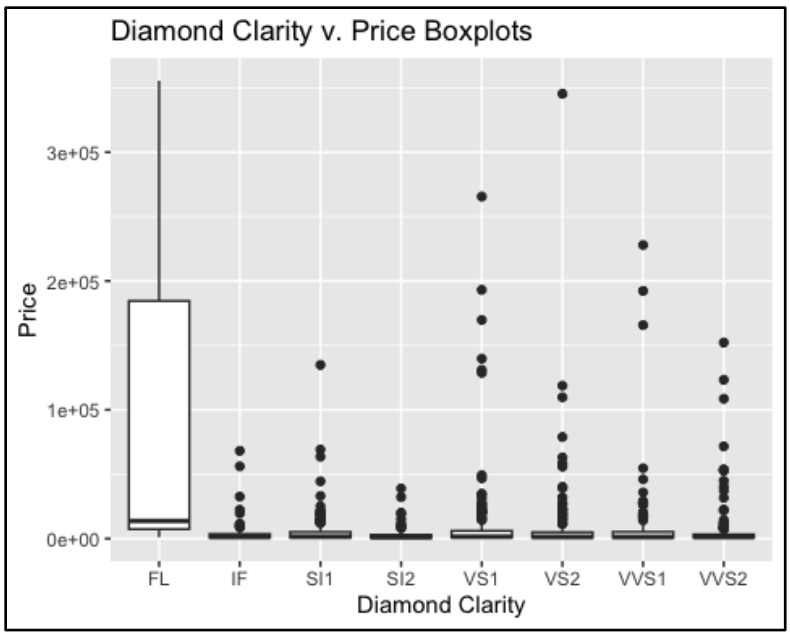
To analyze pricing for diamonds according to their colors, we made a box plot which shows that all diamond colors have approximately similar median pricing. We can assume that the color variable does not have a strong influence on the price of a diamond.

Figure 1. Effect of diamond color on price



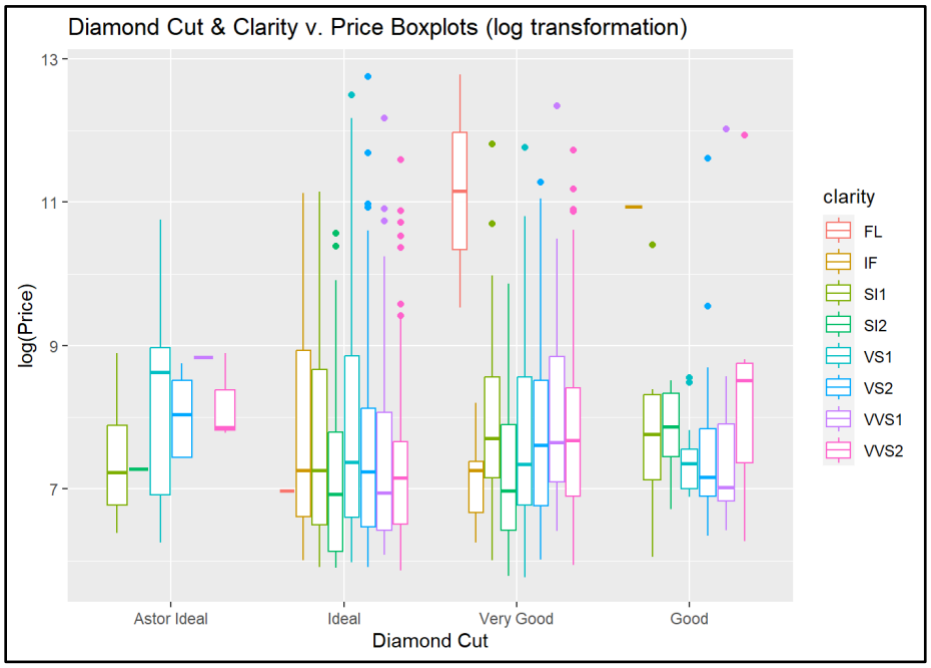
In the figure to the left, we chose to use a boxplot to show the spread of the values for the diamond cut. We found that the price of a diamond is nearly the same for all diamond cuts, with the exception of the Astor Ideal. The best diamond cut, "Astor Ideal," has the highest median price and a significantly wider upper quartile range than the other cuts.

Figure 2. Effect of diamond cut on price



For the price and clarity boxplot, it is important to note that the "FL" category of diamond clarity only has three instances of diamonds associated with it, so the data is potentially skewed because of the small sample size.

Figure 3. Effect of diamond clarity on price



When compared to other clarity categories and cuts, the "FL" category of clarity with the "Very good" cut has the highest median price, as can be seen in the box plot on the left.

Figure 4. Effect of diamond cut on log price by clarity

➔ Barplots

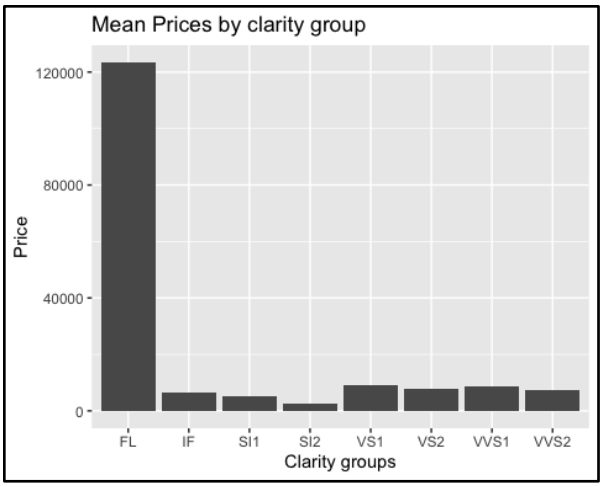


Figure 5a. Effect of diamond clarity on mean price

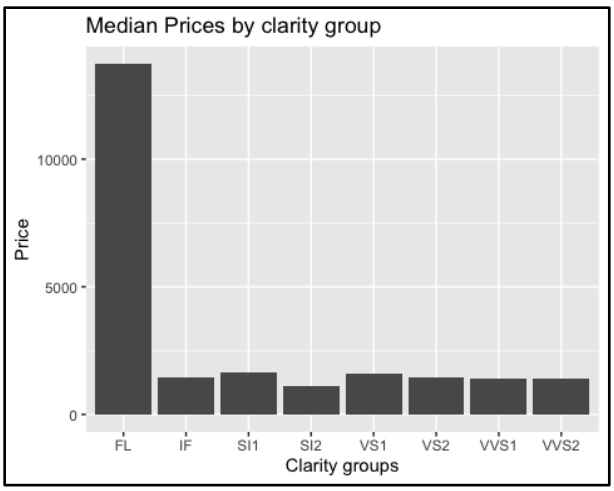
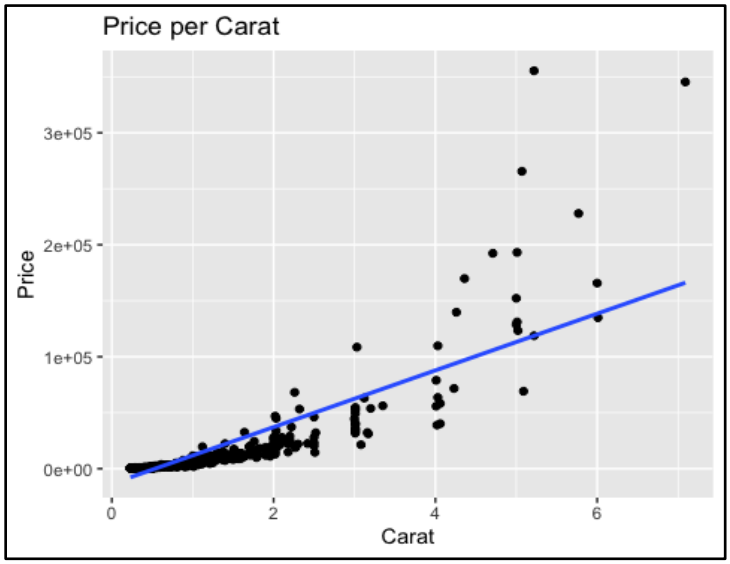


Figure 5b. Effect of diamond clarity on median price

To visualize the data in terms of price against clarity, we calculated the mean prices for each clarity group and plotted a bar graph. From the bar plot of mean prices against clarity categories we can see that clarity "FL" has the highest price over other clarity categories by far (it's important to note that the "FL" group has a sample size of 3, which is relatively small compared to the rest of the data). However, all other clarity groups have similar average pricing. The bar chart that displays the median prices against clarity also shows that "FL" typically has significantly higher pricing over other clarity groups.

➔ Scatterplots



From the simple linear regression plot, we can see that the variables "Carat Weight" and "Price" are highly correlated to each other. There appears to be a clear relationship between the carat weight and the price of the diamond, as larger carat weights correlate with larger prices.

Figure 6. Effect of diamond carat on price

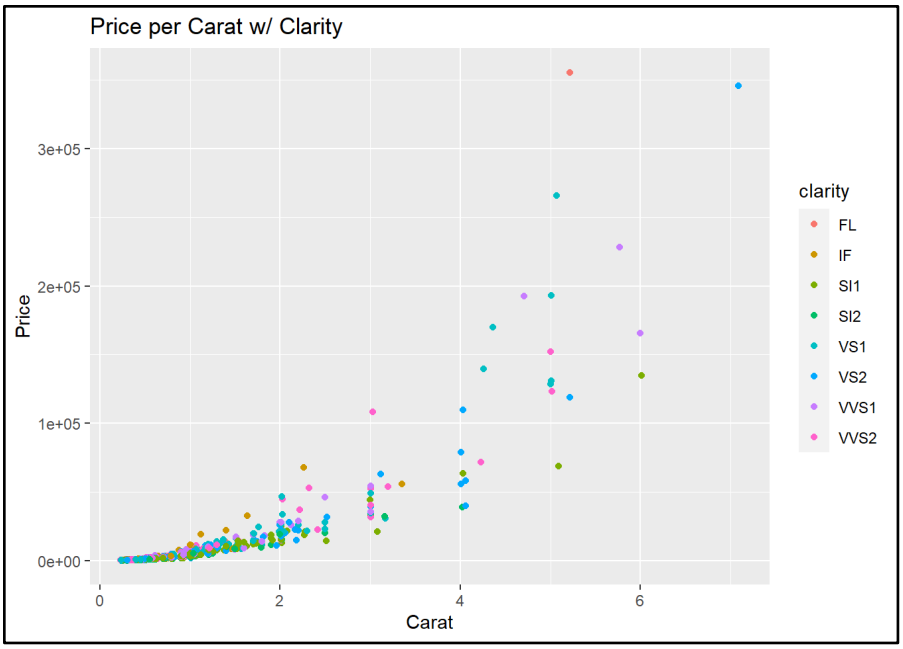


Figure 7. Effect of diamond carat with defined clarity on price

From Figure 3, we did not find any significant relationship between diamond clarity and price. Hence, we plotted the clarity variable on diamond carat v/s price scatterplot (Figure 7). Here we found no significantly definitive relationship between the clarity and the price when plotted with the carat variable as diamonds with differential clarities are randomly scattered across the plot.

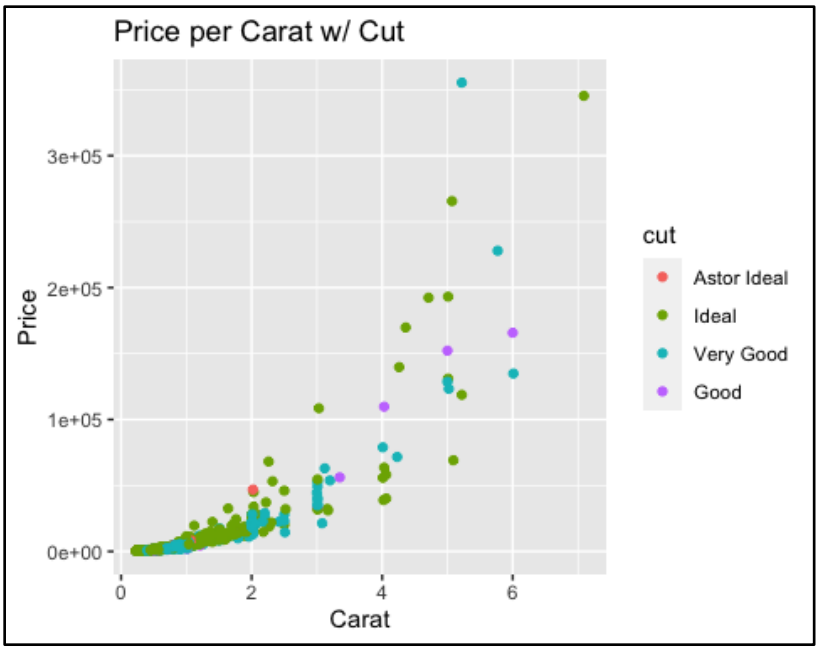


Figure 8. Effect of diamond carat with defined cut on price

From Figure 2, we did not find any significant relationship between diamond cut and price. So next, we plotted the cut variable on diamond carat v/s price scatterplot (Figure 8). Here we found no significantly definitive relationship between the cut and the price when plotted with the carat variable as diamonds with differential cuts are randomly scattered across the plot. Although "Astor Ideal" diamonds are meant to be the highest grade diamonds, there seem to be many "Good" and "Very Good" diamonds that are more expensive than those. From this information, carat weight seems to be a more important variable than cut.

➔ Conclusion

Looking at all the data visualization for price against different variables, what stands out most is that diamonds with the “Astor Ideal” cut have a higher price than other diamonds. However, it is also important to note that the “Astor Ideal” diamonds have a small sample size of n=20 out of a total 1214 diamond samples, so the data may be skewed. We can determine that diamond pricing is affected by a combination of variables, with no single variable completely outweighing the other. In the scatterplot comparing carat weight to price (Figure 8), despite some diamond cuts being considered “Very Good”, the price of the diamond was still high due to its large carat weight. The scatterplot of carat weight against price has a relatively high correlation and clear positive relationship, meaning that carat weight may have a slightly more significant impact on the pricing of a diamond. This demonstrates that pricing is a complicated variable that is affected by numerous factors.

Linear Regression

➔ No Transformations

We performed a simple linear regression on this dataset to test whether there is a linear relationship between the carat and price variables. To begin, we created a scatterplot to view the relationship between these two variables without any transformations. From this plot, we saw a slight positive relationship between a diamond's carat and price with a few outliers.

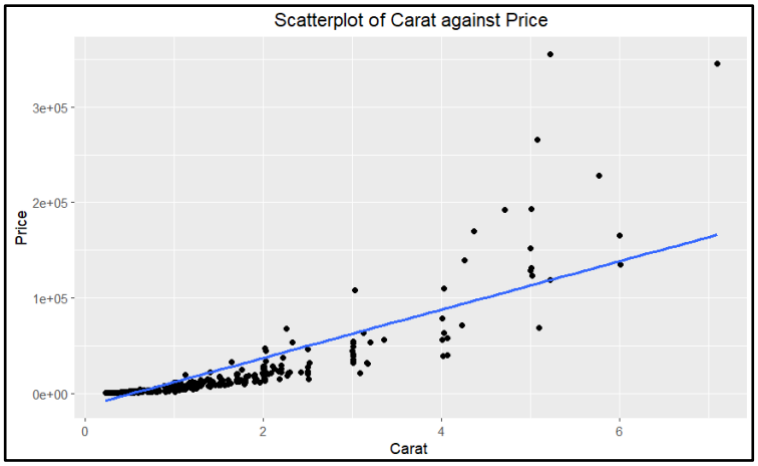


Figure 9. Scatterplot of carat against price

We also created residual plots (below) and checked them against the assumptions of simple linear regression. We found that the residuals were not evenly spread along the horizontal axis and that the vertical variation was not constant across the plot, meaning that the data did not have a mean of 0 or constant variance. Furthermore, there were some outliers that did not fall along the guide line of the Q-Q plot, meaning the normality assumptions were not met. Because both of assumptions 1 and 2 were violated, we decided to transform the y-variable first.

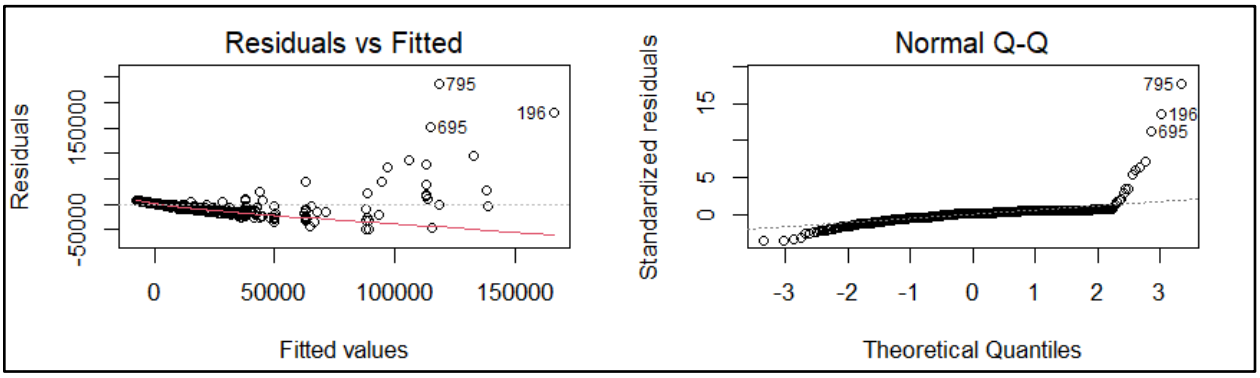


Figure 10. Residual plots to check the assumptions of simple linear regression

We created a Box-Cox plot to help us determine which transformation we should perform on the price variable. Below, the figure shows us the log-likelihood function against lambda. It also gives us the 95% confidence interval of the lambda that maximizes this function, which is the lambda we should choose. However, because the 95% confidence interval does not include

whole numbers, we were not inclined to transform the variable by putting it to the power of the lambda. Additionally, the log transformation of the y-variable would allow us to still easily interpret the regression coefficients. Thus, we chose to log-transform the price variable.

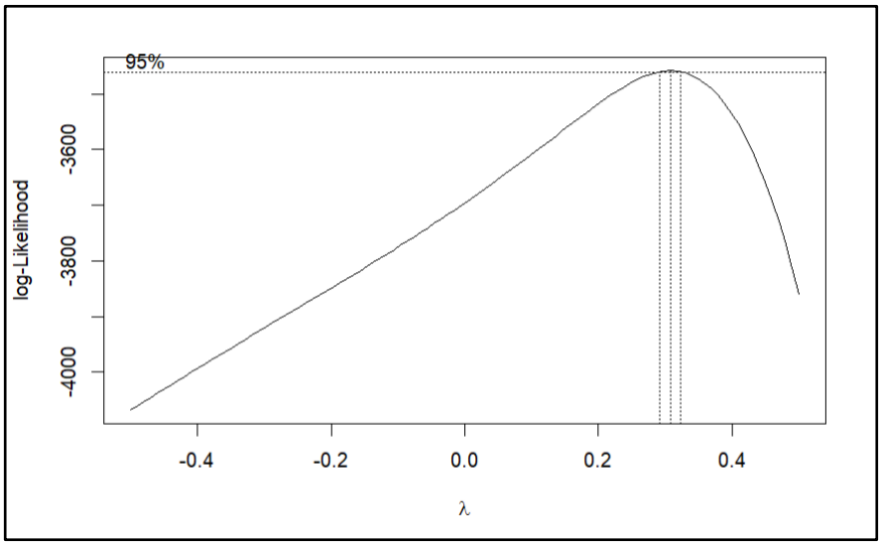


Figure 11. Box-Cox plot

➔ Log-transformed Response Variable

Our new linear regression tests the relationship between the log(Price) and Carat weight. We obtained the scatterplot below after graphing our linear regression model onto the transformed diamonds data set:

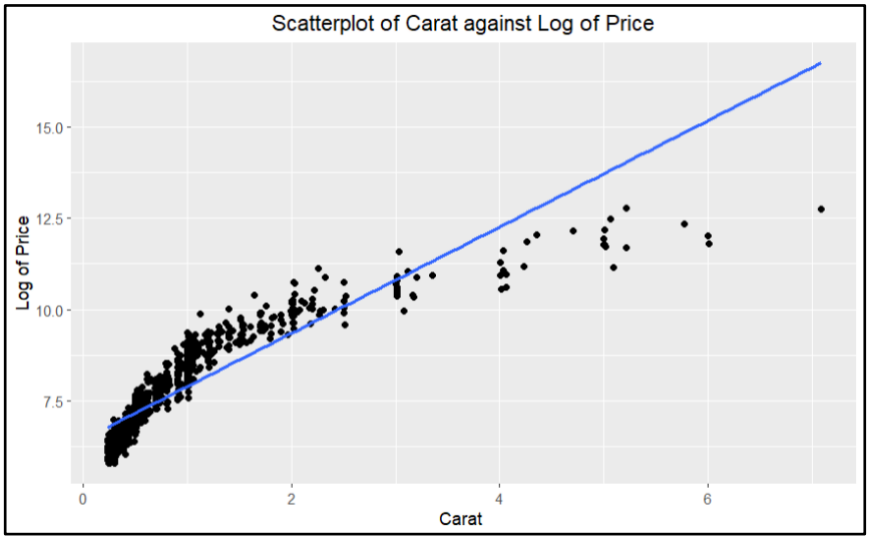


Figure 12. Scatterplot of carat against log price

We see that the data points aren't evenly scattered, and the vertical spread of the data points isn't constant around the regression line. Thus both assumptions 1 and 2 aren't met.

We confirm this with the residual plots below:

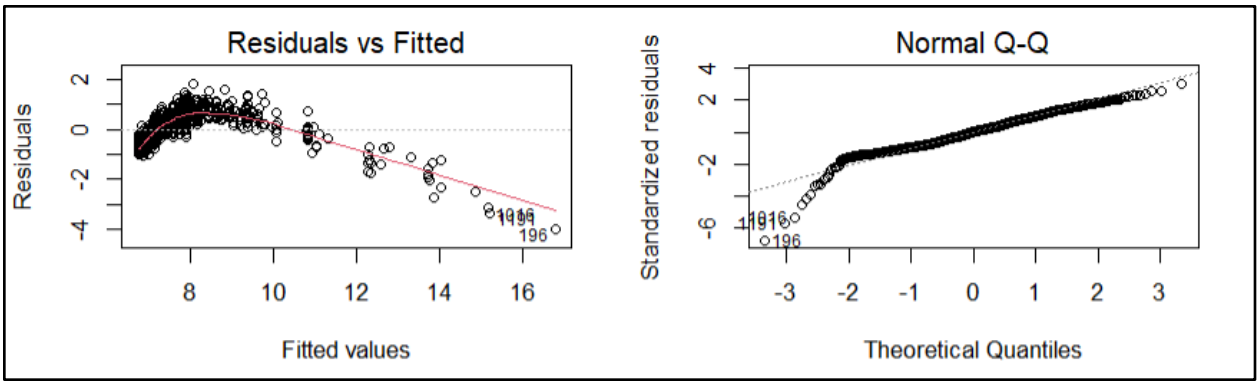


Figure 13. Residual plots for transformed dataset

The residuals are not constant or evenly scattered on the x-axis in the first plot; therefore we wanted to see if another transformation could create a better model to fit our data set. We seek to perform linearization transformation to make our predictor variable, carat, be linear with our transformed log(Price) response variable. Since our scatter plot follows a slight positive curve reminiscent of a log function, we transformed our carat variable to become log(carat).

➔ Log(y)-log(x) Model

We model our log(Price) and log(Carat) in the scatterplot below:

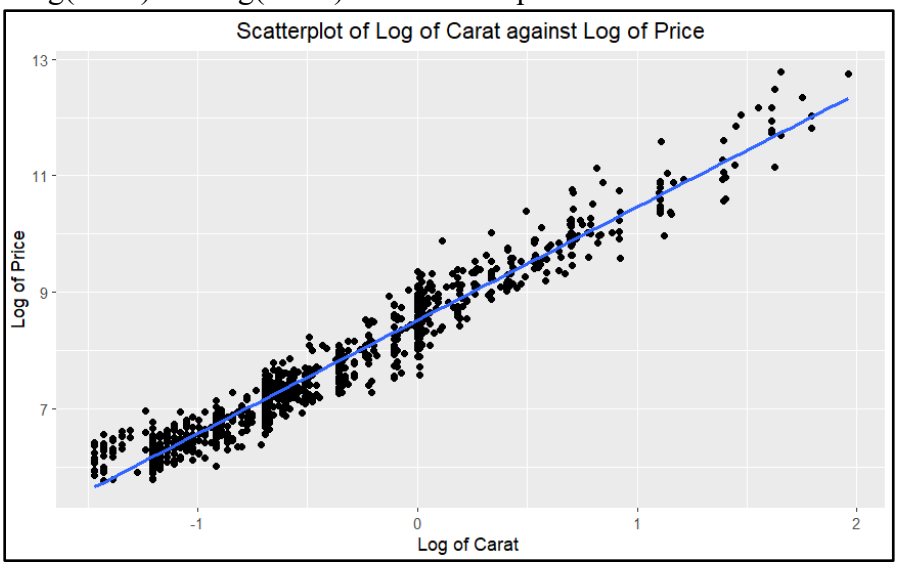


Figure 14. Scatterplot of log of carat against log of price

After the log transformations of both the carat and price variables, we once again tested our transformed data against the assumptions. We created residual plots (below) and found that the spread of the points across the horizontal axis and the vertical variation across the plot were much more even and constant. Both assumptions 1 and 2 were now met. Even the points on the Q-Q plot fell along the guideline, meeting the normality assumption.

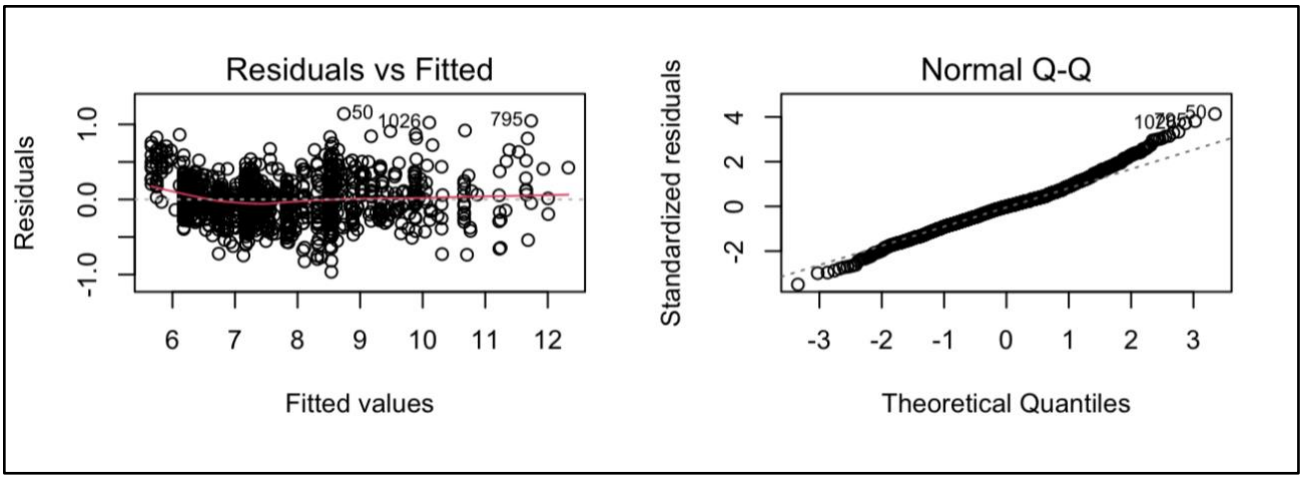


Figure 15. Residual plots for the log transformed carat and price values

➔ Conclusion

```
Call:
lm(formula = log(price) ~ log(carat), data = diamonds)

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 ***
log(carat)  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 16. Summary of linear regression analysis

Because our transformed data now passes the assumptions, we get the following simple linear regression equation:

$$log(Price) = 8.521208 + 1.944020 * log(Carat)$$

We can conclude that for **a 1% increase in carat weight, the predicted price increases by approximately 1.944020%.** We also get a relatively high R^2 value of 0.9547, meaning that our model provides a good fit for our data.

To confirm whether there is a linear relationship between carat weight and price, we conducted a hypothesis test:

$$H_0 = \ There\ is\ no\ linear\ relationship\ between\ log(Price)\ and\ log(Carat)$$
$$H_a = \ There\ is\ a\ linear\ relationship\ between\ log(Price)\ and\ log(Carat)$$

We found the following critical value for our dataset with 1212 degrees of freedom: 3.849143
Since our F-statistic of 25530 > 3.849143, we have evidence to reject the null hypothesis and confirm that there is a linear relationship between log(Price) and log(Carat).