Diamond Price Report

Table of Contents

1 Executive Summary

This study was conducted to explore how price is related to various diamond properties. For this reason, we used a dataset from the Blue Nile website containing the properties of more than 1,200 diamonds as well as their prices. The properties in this study are carat, clarity, color, and cut (referred to collectively as the 4Cs). The study also addresses several claims that are made by Blue Nile or held by the public. This summary will address the study's findings on the relationships between the 4Cs and price, discuss the methods and approaches used in the study, and include a more detailed section on the relationship between a diamond's carat and price.

To conduct this study, we obtained a dataset of over 1000 diamonds that are for sale through Blue Nile and other online vendors, then used R to analyze the relationships between variables both visually and quantitatively. Of the 4Cs, carat (or weight) is the only quantitative variable. The remaining three variables are categorical. Below are the key findings for relationships found between these variables and price.

A commonly held belief is that a higher carat value corresponds to a higher price. The study shows this belief to be generally true. The relationship is not one-to-one, or linear; rather, the price of a diamond often goes up much faster than the carat weight. However, there is much more variation in a diamond's price if the carat weight is large.

Blue Nile claims that higher clarity diamonds are also higher price. This is somewhat supported by the study, although there is not as clear a trend as a consumer may expect. Rather than a smooth upward relationship between clarity and price, most diamonds (from the slightly included to the internally flawless) have similar pricing. Flawless diamonds fetch a much higher price than other clarity diamonds, though it may be worthwhile to perform a follow-on study to determine if there are other factors affecting or obscuring this relationship.

Like clarity, Blue Nile claims that diamonds with a better color grade are more expensive. The study did not find a strong correlation to support this claim, though there is a weak trend for higher quality diamonds to be generally more expensive.

Finally, Blue Nile claims that cut can have the biggest impact on a diamond's price. This study did not find this claim to be true. However, Astor Ideal diamonds, which are exceptionally well-cut diamonds, are more expensive than the other categories. It is possible that a diamond's carat masks the effect of its cut on price; a follow-on study is recommended to study this relationship further.

Relationships between the 4C diamond properties revealed no clear trends or patterns, though the study does note a few interesting findings. First, flawless clarity diamonds are generally higher quality in color and have a higher median carat weight than the other clarity diamonds. Next, jewelers may tend to attempt an ideal cut for most diamonds, though for lower-quality diamonds, they may prioritize preservation of a higher carat weight instead.

To further explore the relationship between carat and price, we developed a mathematical model that predicts the price of a diamond based on its carat weight. This model can help determine what an acceptable price might be for a consumer looking to purchase a diamond of a specific carat, though it does not consider other variables such as the clarity, color, or cut.

Overall, this study found that most claims from Blue Nile were either supported or somewhat supported by data analysis, and that carat is a reliable predictor of a diamond's price. For future studies, we recommend analyzing the relationships between the other diamond properties and price creating a mathematical model with multiple predictor variables.

2 Relationships between Variables

2.1 Data and Variables

2.1.1 Data

The dataset used for this report includes 1214 recordings of diamonds including carat, clarity, color, cut, and price. The dataset is available on the Blue Nile website[1]. We ran a simple validation test to check for quality and confirmed that there are no signs of null values or invalid characters that could potentially provide misleading results over the course of the analysis.

2.1.2 Carat

In this dataset, the Carat variable ranges from 0.23 to 7.09 carats, which indicates a diamond's weight — not size — from lowest to highest. Because higher carat diamonds require larger rough stones, they are rarer and generally more expensive than lower carat diamonds. According to Blue Nile, a higher-carat diamond is not necessarily a better diamond; however, carat weight is often more desirable and can have the biggest effect on price.

2.1.3 Clarity

In this dataset, the Clarity variable has eight possible values: FL, IF, VVS1, VVS2, VS1, VS2, SI1, and SI2, standing for Flawless, Internally Flawless, Very Very Slightly Included 1-2, Very Slightly Included 1-2, and Slightly Included 1-2. These values describe the internal or external flaws of a diamond, which may or may not be visible to either the naked eye or with 10x magnification. According to Blue Nile, SI and VS diamonds are the best value, since they are visually similar but less expensive than the IF and FL diamonds. The dataset does not include diamonds that are Included, or I1-3, which have clearly visible inclusions.

2.1.4 Color

In this dataset, the Color variable has seven possible values: D, E, F, G, H, I, and J. These values describe a white diamond's color, with D–F graded as colorless and G–H and I–J graded as near-colorless. According to Blue Nile, "diamond prices decline or increase in alphabetical order," with a G diamond being less expensive than a D diamond. The dataset does not contain diamonds with K–Z color grades, which have faint (K) or noticeable (L–Z) color.

2.1.5 Cut

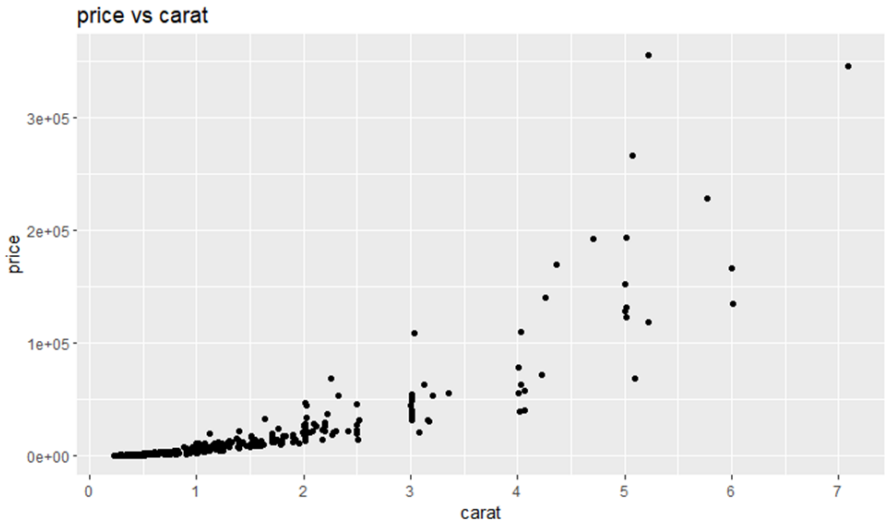
In this dataset, the Cut variable has four possible values: Astor Ideal, Ideal, Very Good, and Good. These values describe a diamond's "light performance" based on proportions, symmetry, and polish. According to Blue Nile, "the cut… can be the biggest factor in the price tag" rather than the carat. This statement is likely indicating that the difference in price is most pronounced when

[1] https://www.bluenile.com/

comparing otherwise similar diamonds by cut. The dataset does not contain diamonds with Fair or Poor cuts, which have lower light performance than Good-cut diamonds.

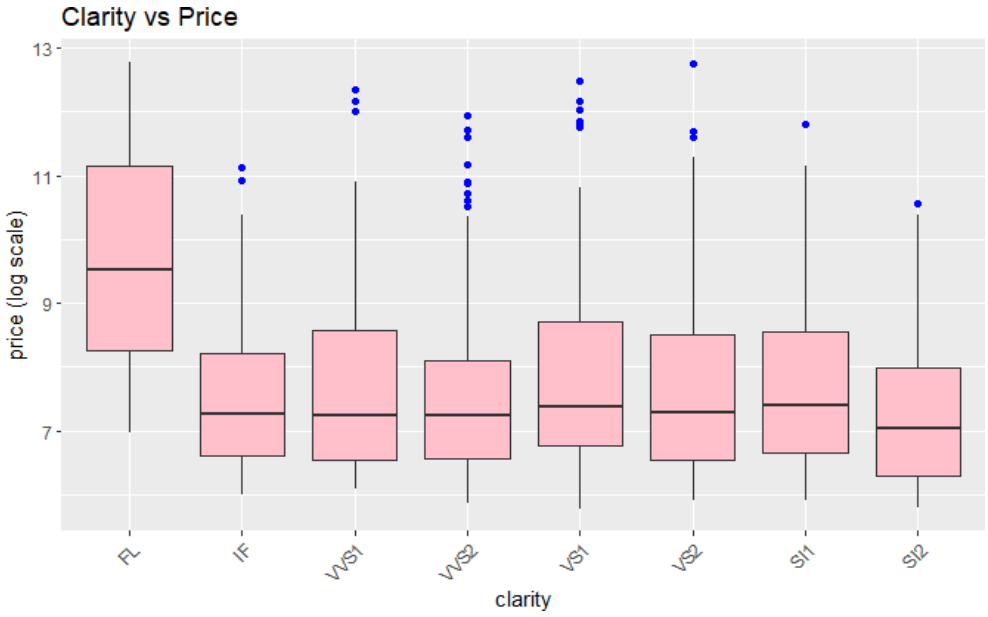
2.2 Relationships with Price

2.2.1 Carat and Price



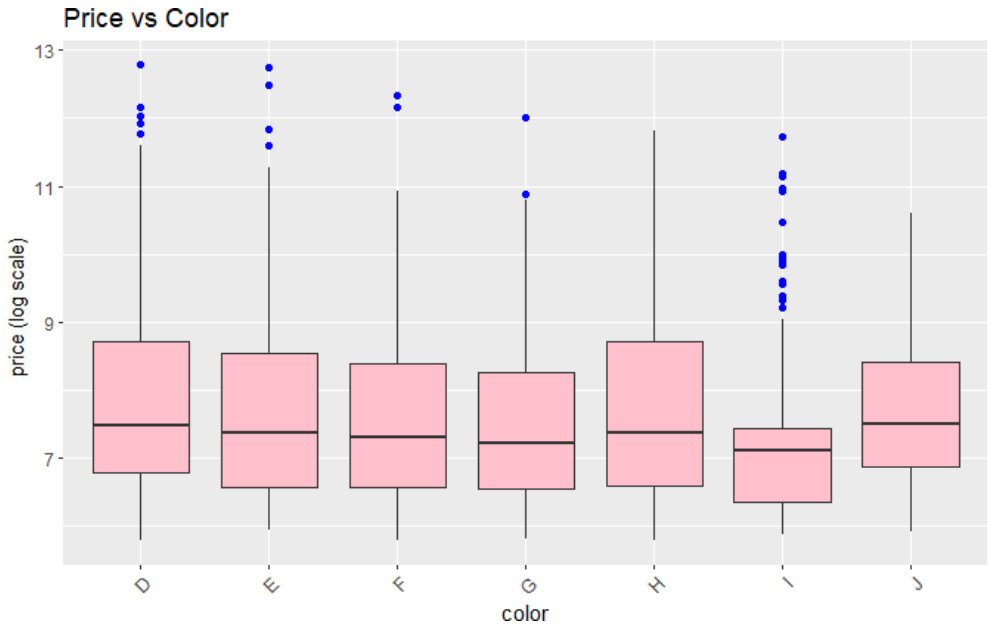
This scatterplot visualizes the relationship between the carat and price of a diamond. A common claim is that carat is the major driver in determining the price of a diamond and that a higher carat diamond will have a higher price — demonstrating a positive relationship. According to the plot, we observe a general upward trend in the points, with the price growing slowly for diamonds up to 1.5 carats and then less slowly for larger diamonds. The variance also increases around 1.5 carats, indicating that the relationship is not necessarily linear. This relationship is contextually logical, as a larger diamond is rarer and should generally be more expensive.

2.2.2 Clarity and Price



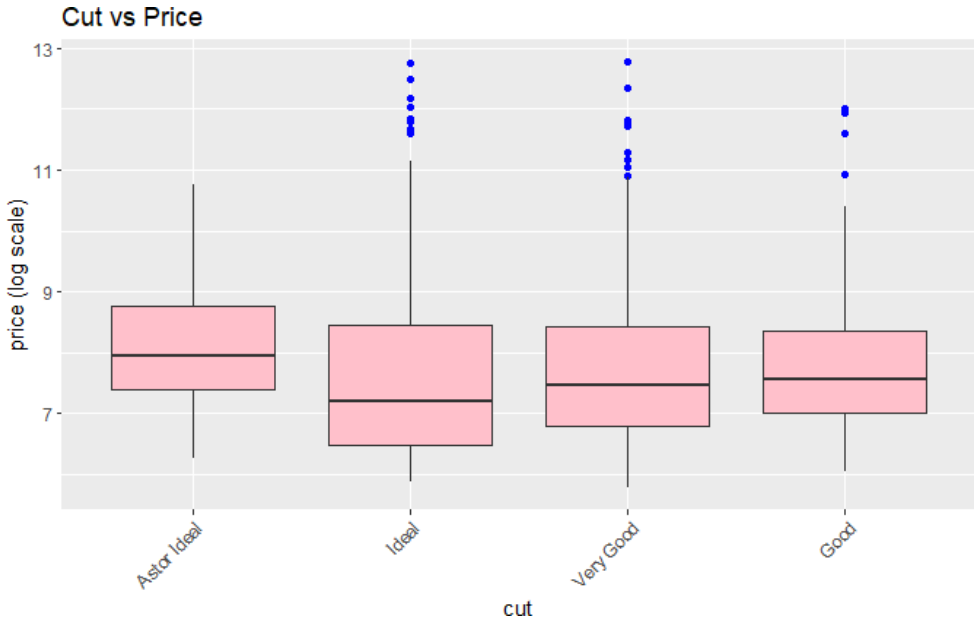
This boxplot shows the relationship between the clarity and price of the diamonds in the dataset. Generally, each category of clarity is right-skewed with outliers at the upper end of price. Blue Nile claims that the VS and SI categories are good values compared to the FL and IF categories. Common belief is that the higher the clarity, the higher the price. According to the boxplot, nearly all categories have similar median prices, and there is no immediately apparent trend for categories other than FL, contradicting the common belief. The FL category has higher first quartile, median, and third quartile values than any other category, indicating that FL diamonds are indeed much more expensive than the other categories of diamonds. However, this does not directly support Blue Nile's claim that FL and IF diamonds are comparable in price against VS and SI diamonds. An interesting note here is that the price range is smaller for FL and IF diamonds than for the other categories. Other variables may be affecting this — for example, a heavier diamond's larger facets may contribute to a lower clarity rating.

2.2.3 Color and Price



The boxplot above represents the relationship between color and price. All color categories follow either normal distribution or slightly right-skewed distribution, with color I containing many outliers. Although the D color grade represents the most expensive color category, there does not seem to be a clear demarcation or trend between the color grades when determining the diamond price. Blue Nile's claim that diamonds are more expensive with a color grade earlier in the alphabet is not immediately apparent, though categories D through G show a somewhat downward trend in the first quartile, median, and third quartile values.

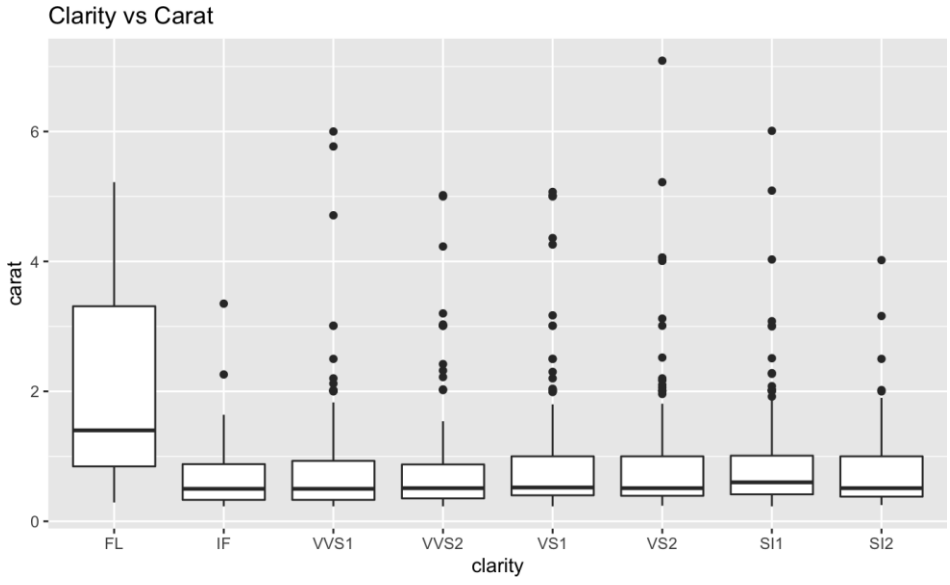
2.2.4 Cut and Price



This boxplot shows the relationship between the cut of a diamond and its price. Blue Nile's claim is that the cut makes a significant difference in price, but the boxplot does not support this claim. If the dataset included diamonds with cuts graded below Good, this claim may be more apparent. According to the boxplot, however, Astor Ideal diamonds do have higher first quartile, median, and third quartile values than the other three categories. These diamonds are only crafted from high-quality diamonds, indicating that based on other properties such as carat and clarity, Astor Ideal diamonds are already likely to be a higher price than diamonds in other cut grades.

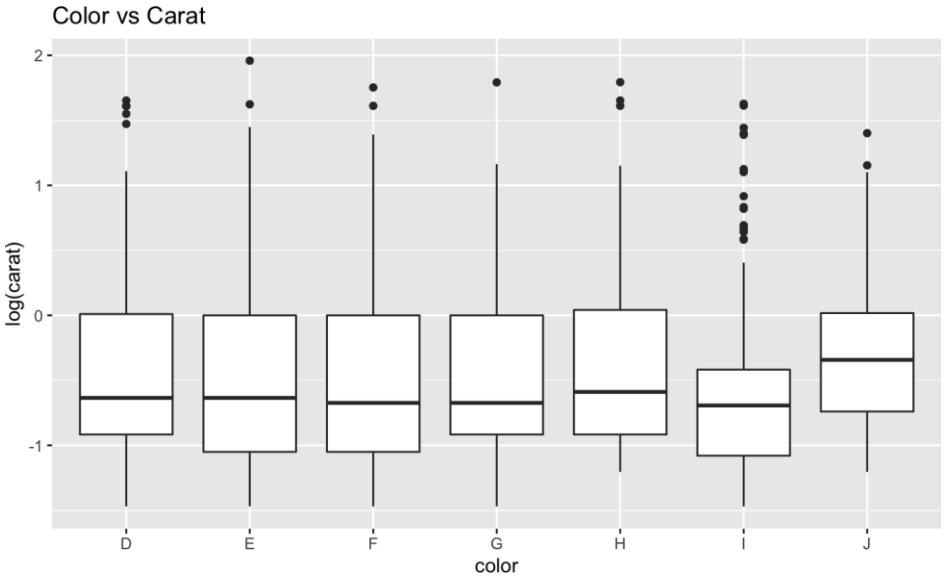
2.3 Relationships Between Other Variables

2.3.1 Carat and Clarity



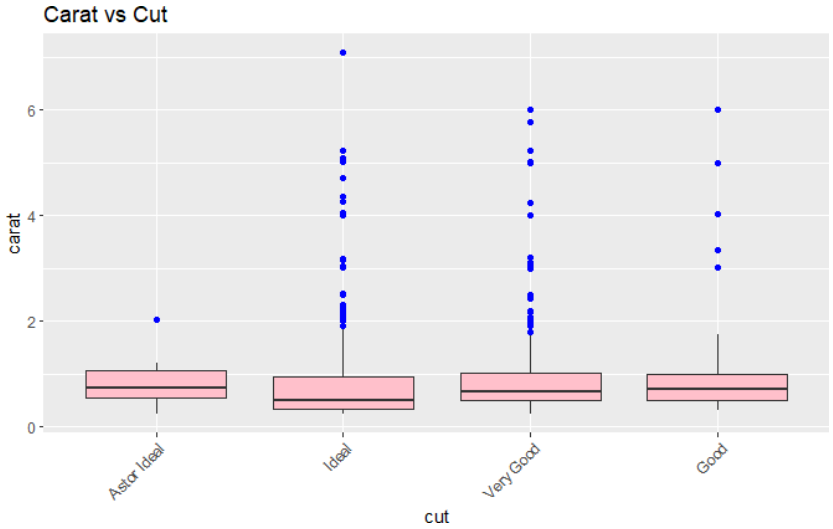
The boxplot above represents the relationship between clarity and carat. In general, the boxplots follow a right-skewed distribution with lots of outliers, aside from FL Clarity diamonds. FL diamonds have relatively higher carat values compared to other clarity diamonds, but the median values stay relatively similar. The boxplot does not show any clear trends, but it is interesting to note that the highest level of clarity is easily distinguishable from the other levels. This may be because few diamonds are of flawless clarity.

2.3.2 Carat and Color



The boxplot above represents the relationship between color and carat — noting here that the y-axis represents the log of carat for more visible boxplots. There is no immediately apparent trend or pattern between the two properties, though it is interesting to note that color I has many more outliers than the other categories.

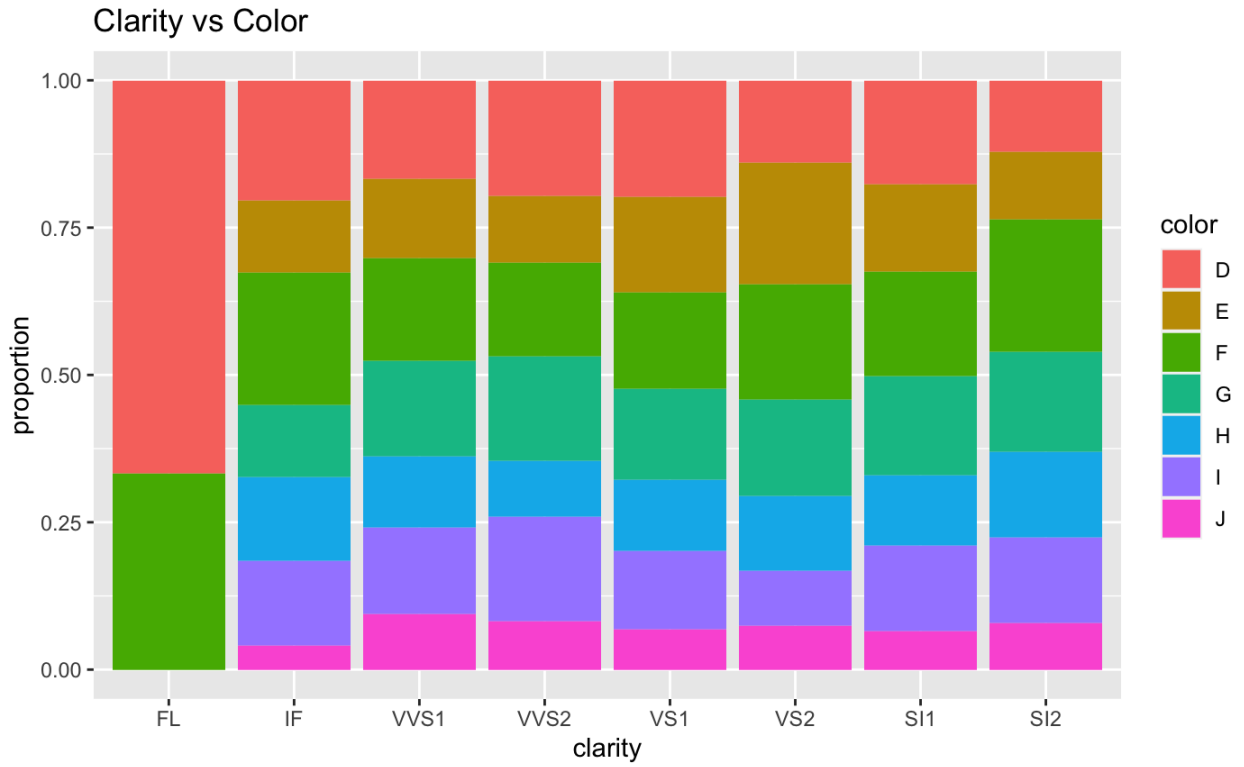
2.3.3 Carat and Cut



The boxplot above shows the relationship between carat and cut. There are many outliers for the Ideal, Very Good, and Good categories, and the median values for these categories seem to show

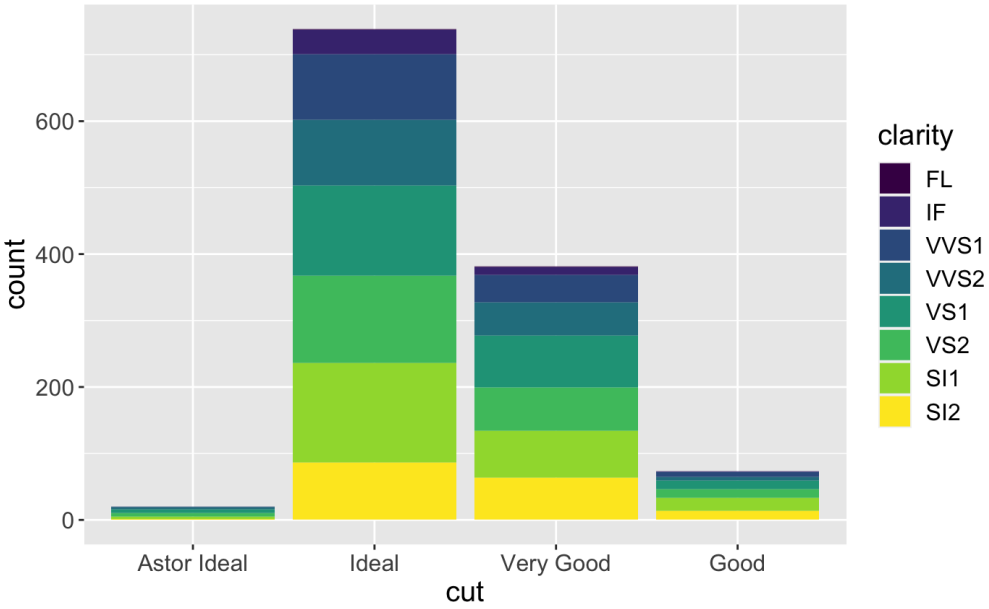
a weak upward trend. This may indicate a market desire for a larger diamond at the expense of a lower quality cut. For Astor Ideal diamonds, the boxplot shows that the range of carat values is much smaller, indicating that the larger diamonds are not always cut at the highest quality — likely due to other variables (such as color or clarity) making the diamond not worth the extra workmanship, even though a high-carat, well-cut diamond would be extremely profitable. In general, this relationship has no trend, but it reveals that high-carat diamonds are not necessarily always cut well.

2.3.4 Clarity and Color

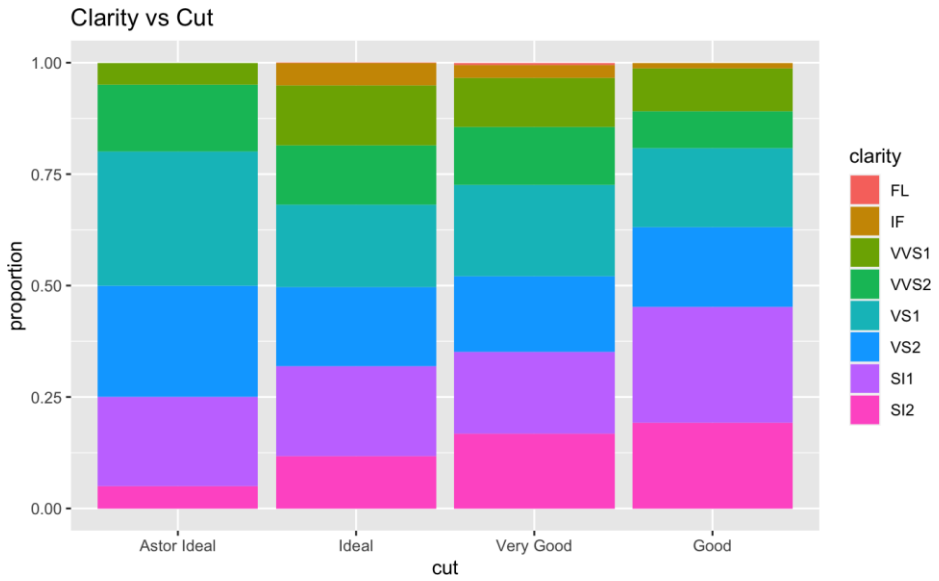


The stacked chart above shows the proportions and relationship between clarity and color. For most of the clarity categories, color is relatively evenly distributed, with no immediately apparent trends. However, the FL clarity only has D- and F-colored diamonds. If this were due to some external factor, such as the conditions each diamond was created in, we might expect a trend. Therefore, the FL category may be different from the others because it has relatively few diamonds. Consumers may believe, however, that a flawless diamond is more often a higher-quality color diamond as well because high quality in one property can be associated with high quality in another.

2.3.5 Clarity and Cut

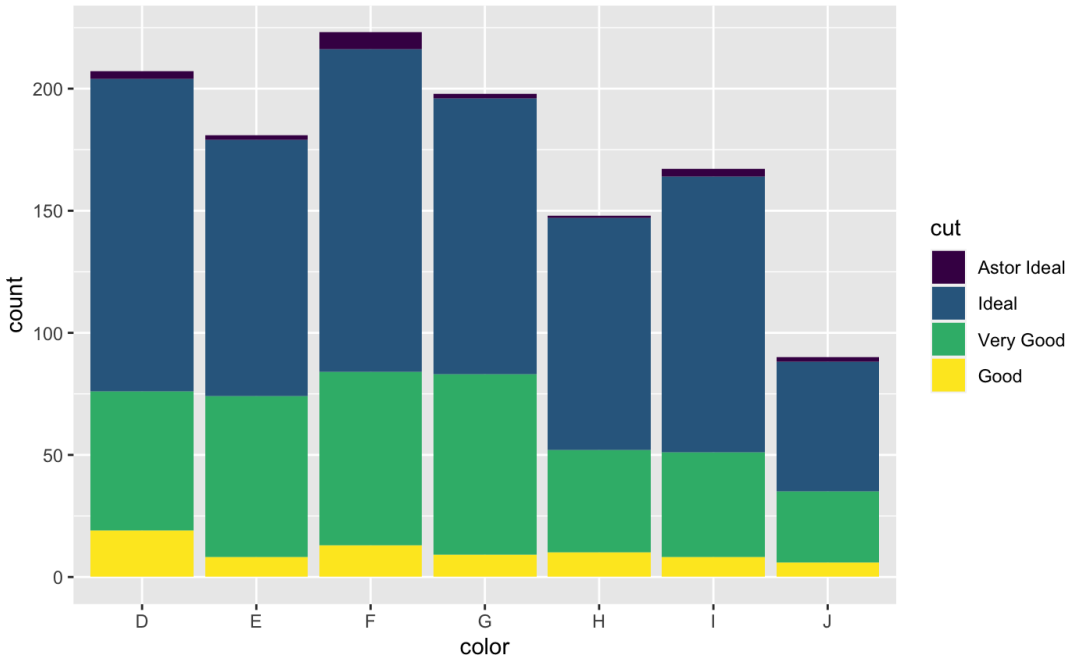


This figure indicates that the majority of the diamonds in the dataset fall under the Ideal or Very Good cut, with most being in the Ideal category. An interesting note is that of the properties, a jeweler only has control over cut, so this figure also indicates that an Ideal cut may be the goal. This makes sense in context, as a well-cut diamond can be sold for a higher price. Astor Ideal diamonds are exceptionally high quality, and their rarity may contribute to the high price.



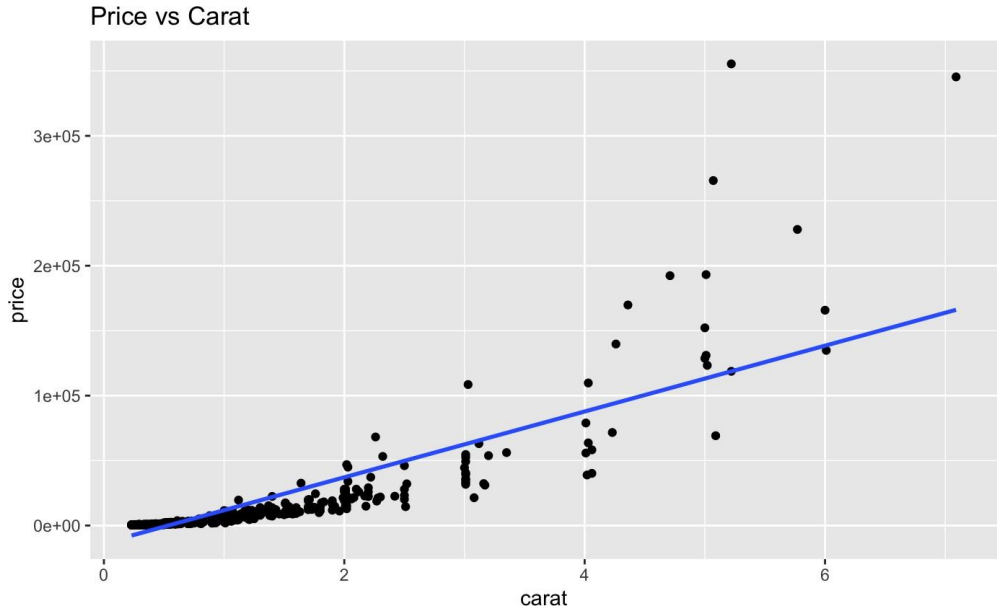
This graph represents that almost all clarities are almost evenly distributed throughout the different cuts of diamonds. However, we do notice that lower-clarity diamonds make up larger proportions of lower-quality cuts — an increasing number of diamonds are slightly included (SI1 and SI2) as the cut decreases in quality. This supports the idea that jewelers may not attempt a higher-quality cut, choosing instead to preserve the carat weight of the diamond at the expense of a better cut.

2.3.6 Color and Cut



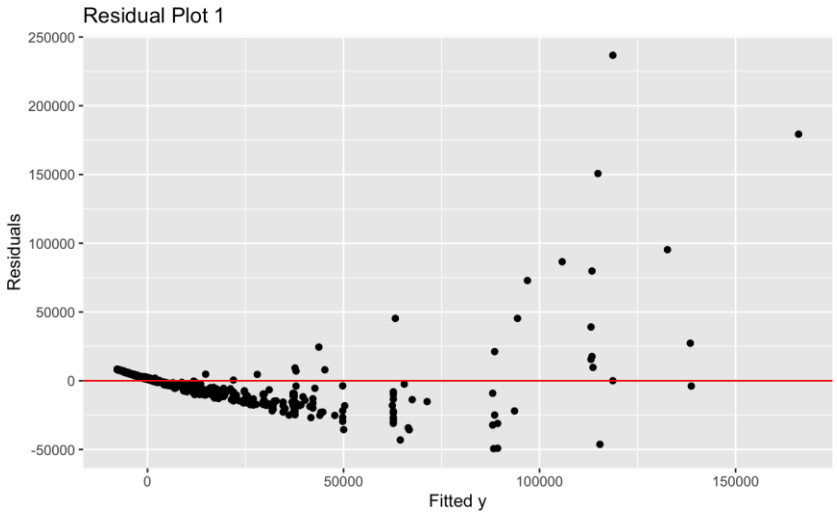
The chart above shows the relationship between color and cut. Similar to a finding of the previous section, the majority of diamonds in each color are of Ideal cut or Very Good cut, with very few diamonds of Good or Astor Ideal cut. In general, a trend or pattern between cut and color was not observed.

3 Simple Linear Regression for Carat vs. Price

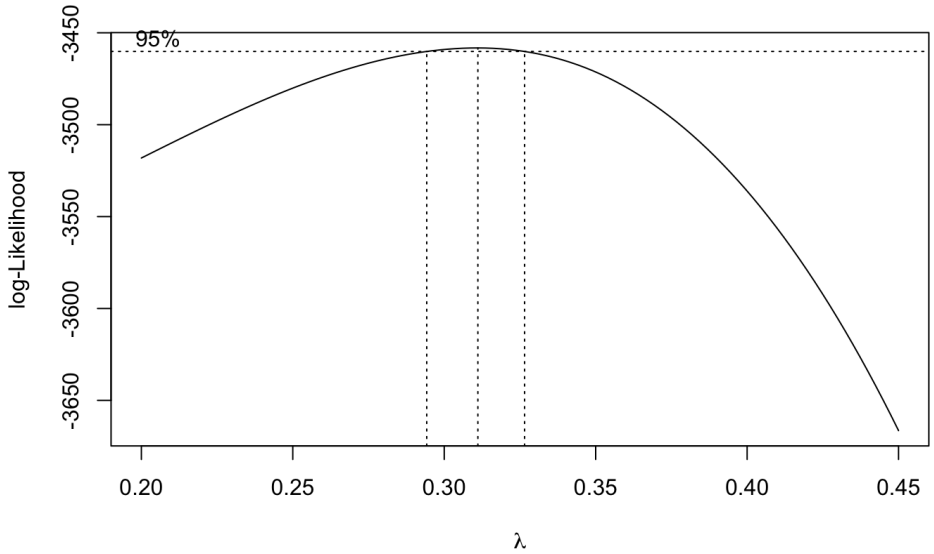


From this scatterplot, we notice a positive correlation between a diamond's price and its carat size, with an increase in the price of the diamond in response to an increase in the carat value. However, in this initial plot, the data points are not equally distributed on both sides of the regression line,

and the relationship between the variables does not appear linear. The vertical spread of the data point is not constant and the assumption that the relationship is linear may not be met. The assumptions that variance for the error term is constant and that the error mean is 0 are also not met. We proceed with plotting residuals to investigate how well the regression model fits the data and to check the linear regression assumptions more closely.

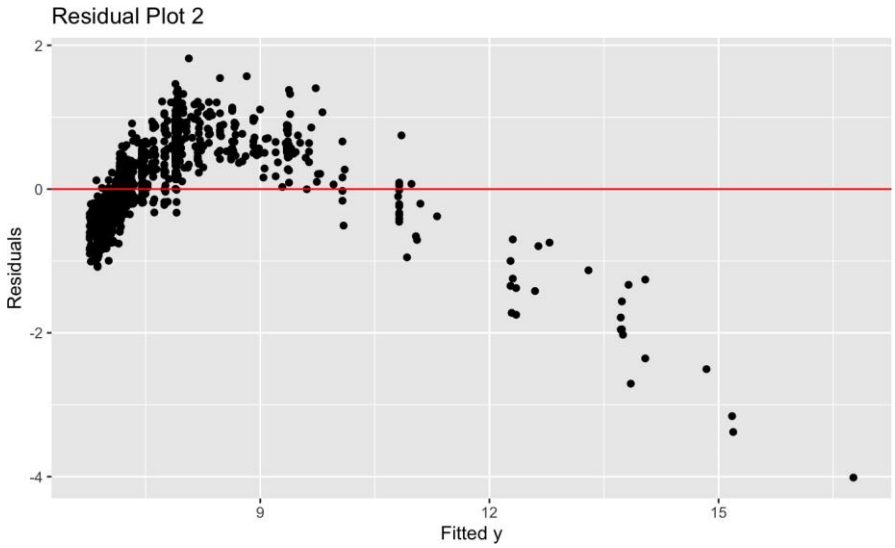


The residual plot shows a curved pattern with non-constant variance, and the variance increases for higher-fitted y values. Just as the first scatterplot shows, the residuals are not evenly scattered on both sides of the horizontal axis. Most of the points are located underneath the regression line. The vertical spread, or variance for the error term, is not constant. In this instance, we would consider a transformation of the response variable. To determine the type of transformation that better suits the model, we will run a Box-Cox analysis.

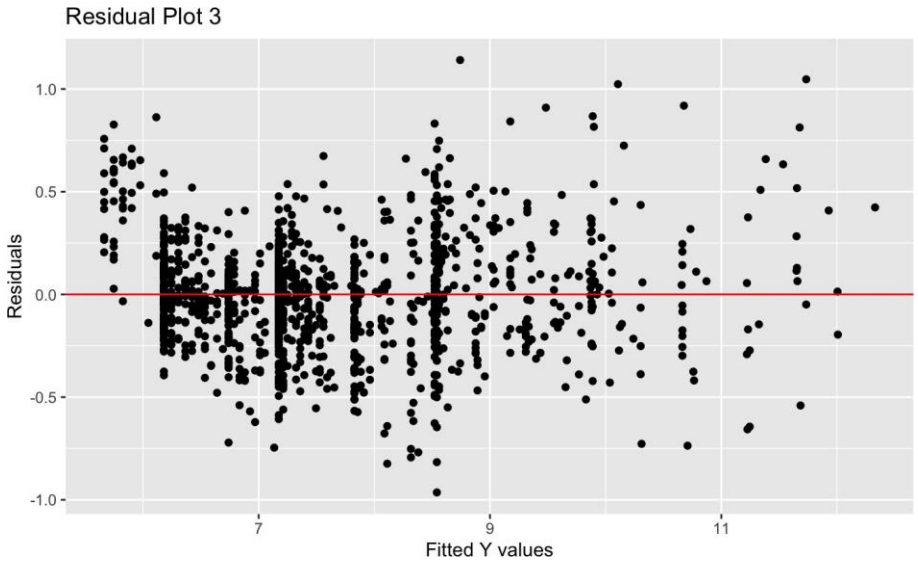


The Box-Cox plot helps us to determine the range of values of lambda we will use in our transformation, as well as the type of transformation that should be applied. A λ value of 1 is equivalent to using the original data, so if the confidence interval for the optimal λ includes 1, transformation is not required. The 95% confidence interval for λ (0.295 to 0.326) does not include

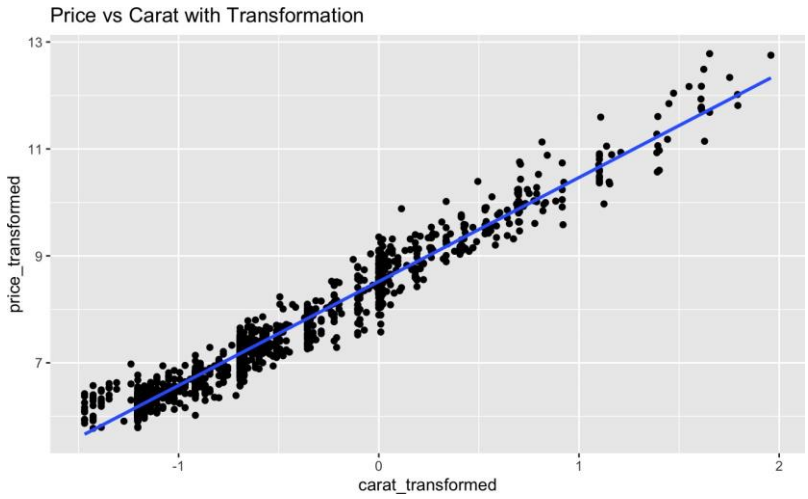
1, so a transformation is appropriate. The estimated value for the optimal λ is 0.31. We know that for λ = 0 we will apply a logarithmic transformation on the y-axis.



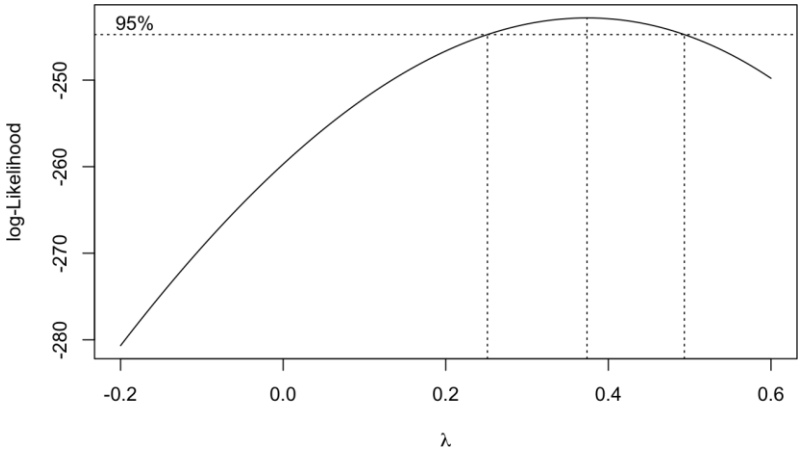
After applying the logarithmic transformation on the response variable, we notice an improvement in the vertical spread of the residuals. However, the residuals are unevenly spread horizontally and show a clear curved pattern; we will perform additional log transformations for the predictor variable.



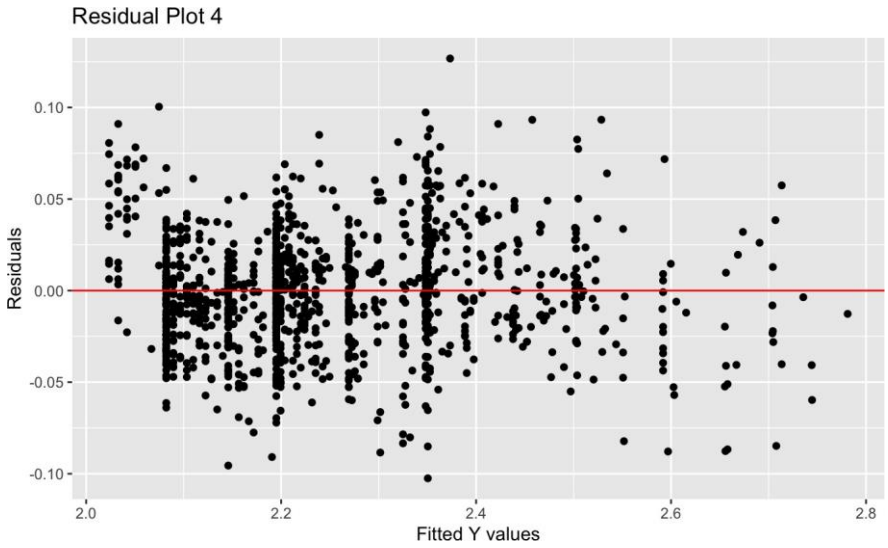
The third residual plot shows no apparent pattern and has constant variance. The points are evenly scattered across the horizontal axis, suggesting that a linear relationship now holds.



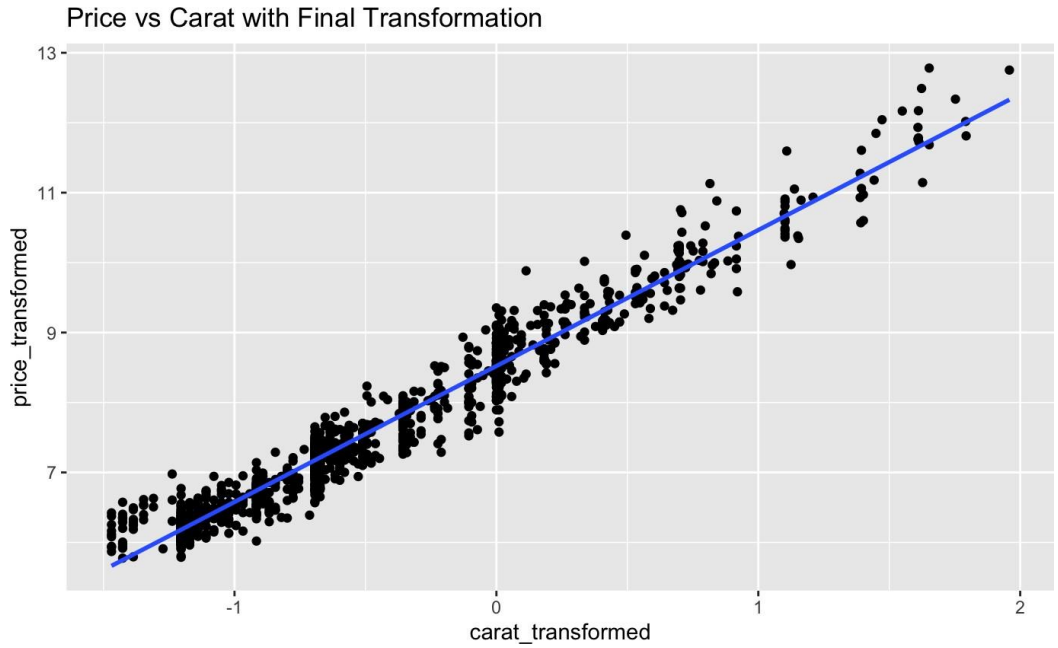
This graph of the transformed predictor and response variables shows a linear relationship, indicating that the transformations were successful.



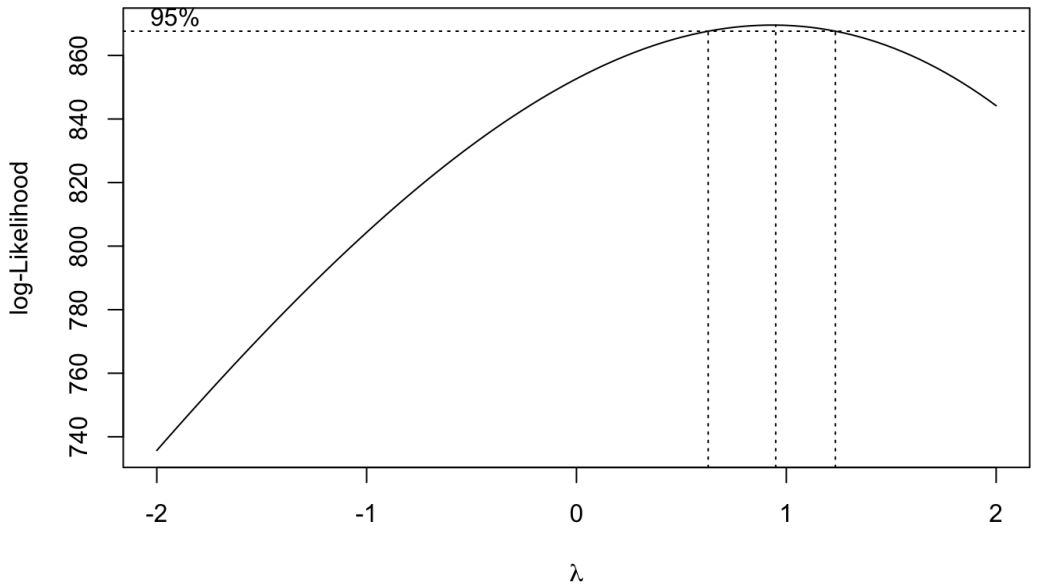
The Box-Cox plot suggests a λ value between 0.2 and 0.5, with 0.38 as the maximum.



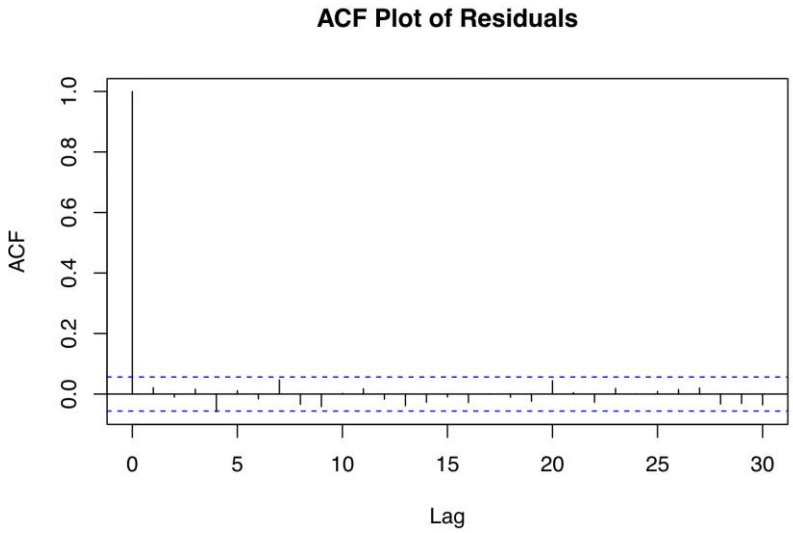
This fourth residual plot shows more constant variance, particularly on higher fitted y-values.



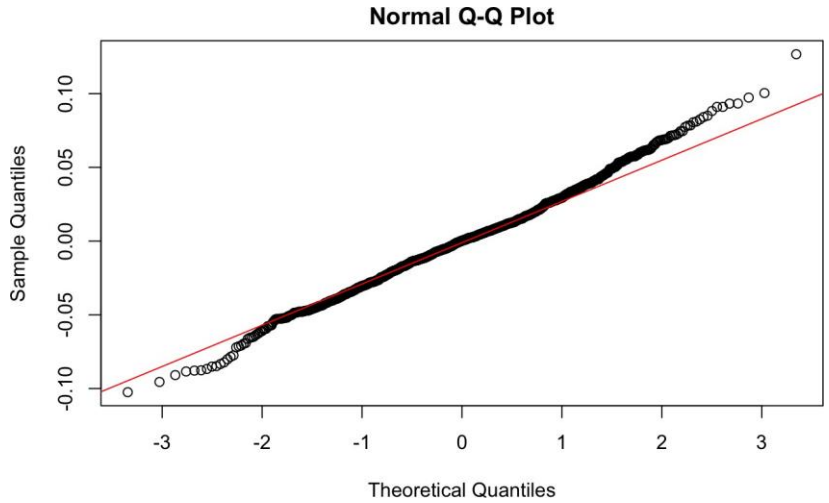
This scatterplot shows a linear relationship, so we will check the assumptions again using a Box-Cox plot, an ACF plot, a QQ plot, and a normality histogram.

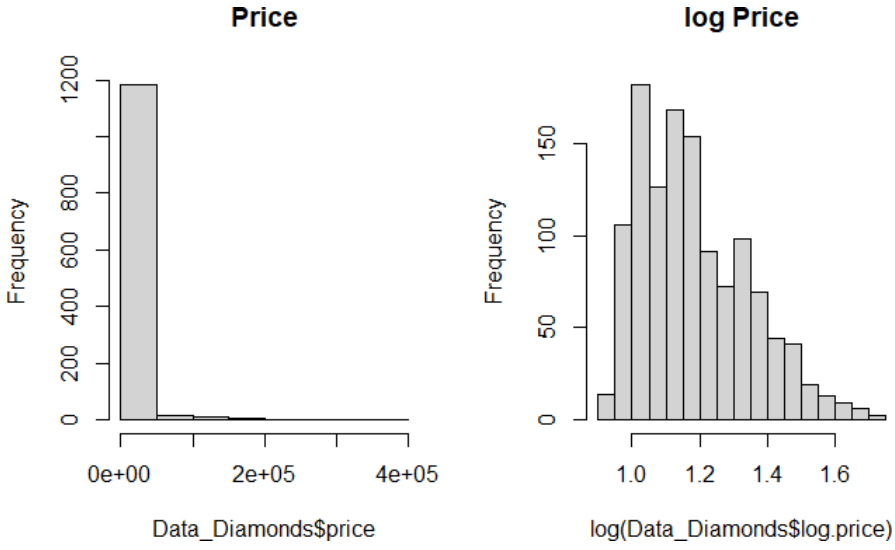


By running Box-Cox plot again with the resulting transformation, we get $\lambda = 1$. Since λ is within the 95% CI, we do not need to continue transforming the variables.



The ACF plot above shows that residuals are not related, indicating that errors are independent.





The QQ plot and histogram above show that price shows a better normality after log transformation. The QQ plot indicates that most plots are following the line, suggesting that the residuals follow the normal distribution. This is also confirmed by the normality histograms, which show the distribution of the price before and after the logarithmic transformation.

The final regression equation is $y^* = 2.348 + 0.221x^*$ where $y^* = (\log(y))^{\wedge}0.4$ and $x^* = \log(x)$.

We then ran a hypothesis test using $H_0: \beta 1 = 0$, $H_a: \beta 1 \neq 0$.

Since the p-value is less than 0.05, we reject the null hypothesis. There is enough evidence to support the statement that there is a linear relationship between the price of diamond and diamond's carat. Our multiple and adjusted R-squared are both 0.9545, which indicates a strong linear correlation. This simple linear regression model is considered a reliable model that can predict diamond prices based on carat values.

3.1 Method and Approach

To fit a simple linear regression for carat and price, we used the R built in function for fitting regression models, lm(). The predictor or explanatory variable for our regression is the Carat variable and the response variable is the Price variable. The lm() function returns a list with a number of different components. The coefficient components give the value of the model parameters, namely the intercept (b_0) and slope (b_1).

Calling summary on a fit model provides more additional outputs to find out if the model is good at predicting the variation in price as a response to a variation in the carat size. The components of most importance are the p-value, the R squared, the standard error and the t-value. The p-value in our hypothesis testing allows us to determine how to interpret the hypothesis testing.

Our hypothesis is $H0 : \beta 1 = 0$, $Ha : \beta 1 \neq 0$. The null hypothesis H0 states that the variable carat has no effect on the response variable price. The alternative hypothesis states that the coefficient influences the response variable. The p-value is the parameter that will tell us whether we can reject the null hypothesis. If the p-value is smaller than 0.05, we can reject the null hypothesis. The p-value for our model is <2e-16, which is smaller than the critical value of 0.05. We can reject the null hypothesis and confirm that a change in the value of the Carat variable is influencing the response variable Price.

After fitting the regression line, we want to check the assumptions of a linear regression are met:

- **Linear relationship:** there exists a linear relationship between the independent variable, x, and the dependent variable, y;
- **Independence:** the residuals or error term ε are independent and have zero mean. In particular, there is no correlation between consecutive residuals in time series data;
- **Homoscedasticity:** the residuals or error term ε have constant variance at every level of x;
- **Normality:** The residuals of the model or errors are normally distributed.

From the residuals plot, the linear model with carat as the only predictor does not exhibit homogeneity of variances. The errors or residuals for the model are linked to the y-variable and they represent the difference between an observed value of the response variable and the value of the response variable predicted from the regression line. Since the variance is related to the

residuals, we will have to transform the response variable to work on the constant variance assumption. Based on this we proceeded with y-axis transformations. After applying a logarithmic transformation on the response variable, we notice an improvement in the vertical spread of the residuals although the horizontal spread can still be improved so we performed a logarithmic transformation of the x axis. The improvement is furthermore confirmed by the Box-Cox result in which we observe that the value of 1 is within the confidence interval.

Once we make sure to transform the variables to meet the linear regressions assumption, we proceed with testing the second assumption, no autocorrelation in the residuals. We do that by generating an ACF plot of residuals and conclude that the assumption of non-autocorrelation in the residuals is met since all ACFs are insignificant for all lags. We ultimately tested the normality of the residuals by generating a QQ plot. We determined that the normality assumption is met since the residuals match their values under normality.

3.2 Conclusions

3.2.1 Findings and Contextual Commentary

As we observed from the data visualizations above, carat is a good predictor and has a strong association with diamond prices. The objective of fitting a linear regression is to measure the relationship between the response (denoted by y) and predictor variables (denoted by x) by fitting the linear equation $y = b_0 + b_1x$, where b_0 is the intercept and b_1 is the slope of the line. This equation helps to understand how a variable can affect another variable and allows us to predict an outcome based on a specific predictor value.

The regression equation is:

$$y^* = 2.348 + 0.221x^*$$

Where $y^* = (\log(y))^{0.4}$ and $x^* = \log(x)$.

The value of the intercept represents the value of y when x is zero. So, in this context, the price of diamond when the carat value is zero is:

$$\exp(2.348)^{0.4} = 2.558$$

Slope tells us how y is affected by x. The value that we got is 0.221, which means that the value of a diamond increases by $\exp(0.221)^{0.4} = 1.092$ for every 1 unit increase of carat.

By substituting these values onto our equation (where x is not mean-centered), we will get $y^* = 2.348 + 0.221x^*$ where $y^* = (\log(y))^{0.4}$ and $x^* = \log(x)$. With this equation, we can predict the estimated price of a diamond given a specific carat size.

The equation may fail if we attempt to extrapolate beyond the carat range in the training data, and a more accurate equation may be possible using multiple diamond properties, such as cut, color, and clarity. However, the high R-squared values indicate that the model will perform well with new testing data that is similar to the training data.