Blue Nile Diamond Report

Tatev Gomtsyan, Taryn Trimble, Jacqui Unciano, Lingzhen Zhu

2023-06-26; Group 12

Section 1: Analysis Report

Introduction

This report explores the relationship between the 4C of diamonds (cut, carat, color, and clarity) and the price of the diamond. In addition, a simple linear regression analysis is done in order to determine if carat weight and price have a linear relationship. The data set is provided by the certified diamond retailer company, **Blue Nile**. Their diamonds are analysed and graded by the Gemological Institute of America (GIA) which is one of the most reputable diamond labs.

Findings Between the 4Cs and Price

After analysing the relationship between the 4Cs and the price of the diamonds, there is strong evidence that the carat weight is the largest influencer in determining the price of a diamond. The other three factors will need more in depth analysis in order to determine their statistical significance as slight differences between the cut, clarity, and color grade and the price were noted. **Blue Nile** has explained that while the quality of a diamond is determined by a balance of the carat weight, clarity grade, color grade, and cut grade, the price can mostly be influenced by the carat weight due to media and adverts.

However, they have also stated that the cut can be the biggest factor in the price of a diamond, rather than the carat weight. While the data set provided by **Blue Nile** definitely has some very high quality cuts of diamonds with a more expensive price tag, it is noted that for the most part, the price does not visually seem to be too different between the different grades of diamond cuts, suggesting more in-depth analysis should be done.

In terms of the color and clarity of a diamond, **Blue Nile** has stated that while these variables can influence the price of a diamond, there are other factors that affect the appearance of the variables. For example, the metal used for a setting can impact the appearance of the color. The size and shape of the diamond can mask the inclusions or make the inclusions more visible. This is mostly supported by the data set as it's apparent that for most of the clarity grades, the price range is not too different for most of the diamonds recorded. It is noted that there is still a slight difference between groups of grades. The same can be said for the difference between prices in terms of color grade. Visually, there seems to be an insignificant difference, but it's recommended to analyse these two variables more.

These observations about the relationship between the 4Cs and price is supported by our analysis. Information about the data set and how each visualization was made, along with thet appropiate in-depth analysis can be found in section 2 of this report.

Section 2: Data Description and Visualization

About: Carat

The carat refers to the *weight* of the diamond, not the *size* of the diamond. Larger diamonds are sourced from larger crystals, meaning it's relatively harder to find than smaller crystals. Therefore, the price of a large carat diamond depends on how rare the larger crystals are. This would suggest that the larger the carat is, the more expensive the diamond is. In our data set, the variable name is *carat.*

About: Clarity

For clarity, diamonds are grouped into 11 different clarity grades, ranging from low (Included -I1,I2,I3- Diamonds) to high (Flawless -FL- Diamonds) grade. Clarity refers to how many 'inclusions' are within a diamond as the more inclusions a diamond has, the more negatively it affects the beauty of the diamond. In our data set, the variable name is *clarity.*

About: Colour

Color refers to how colourless a diamond is. There are 8 color grades for diamonds, ranging from low (K) to high (D) grade that are offered at **Blue Nile**. The other lower grades (L to Z) are not offered at **Blue Nile**. Diamonds with higher ratings strive to be as colorless as possible, whereas diamonds with lower ratings tend to exhibit more noticeable hues as we progress along the color diamond scale chart. A colorless diamond appears transparent, while diamonds lower on the rating chart may exhibit a warm tint or hue. In our data set, the variable name is *color.*

About: Cut

The cut of a diamond refers to the light performance of a diamond. It is based on a combination of factors: proportions, symmetry, and polish (the overall surface condition of facets). There are four grades to the GIA scale for cut, ranging from "Poor/Fair" to "Excellent/Ideal". **Blue Nile** has their specialty "Super Ideal" ranking called *Astor.* **Blue Nile** does not sell poor/fair diamonds. In our data set, the variable name is *cut.*

About: Price

The price is the dollar amount of each diamond in USD. For the most part of the visualizations, the price will either be reflected in thousands or log-thousands for easier comprehension/visualization. In our data set, the variable name is *price.*

Additional Information on Dataset Changes

The analysis on the relationship between the predictor variables and response variables were done using R. The packages used were ggplot2, dplyr, and MASS for data wrangling and visualization.

Due to the variables Clarity, Color, and Cut having a grade order, the variables were refactored using the factor() function in order to reflect the grades from lowest to highest when creating the graphs and visualizations.

Graphs and Visualizations

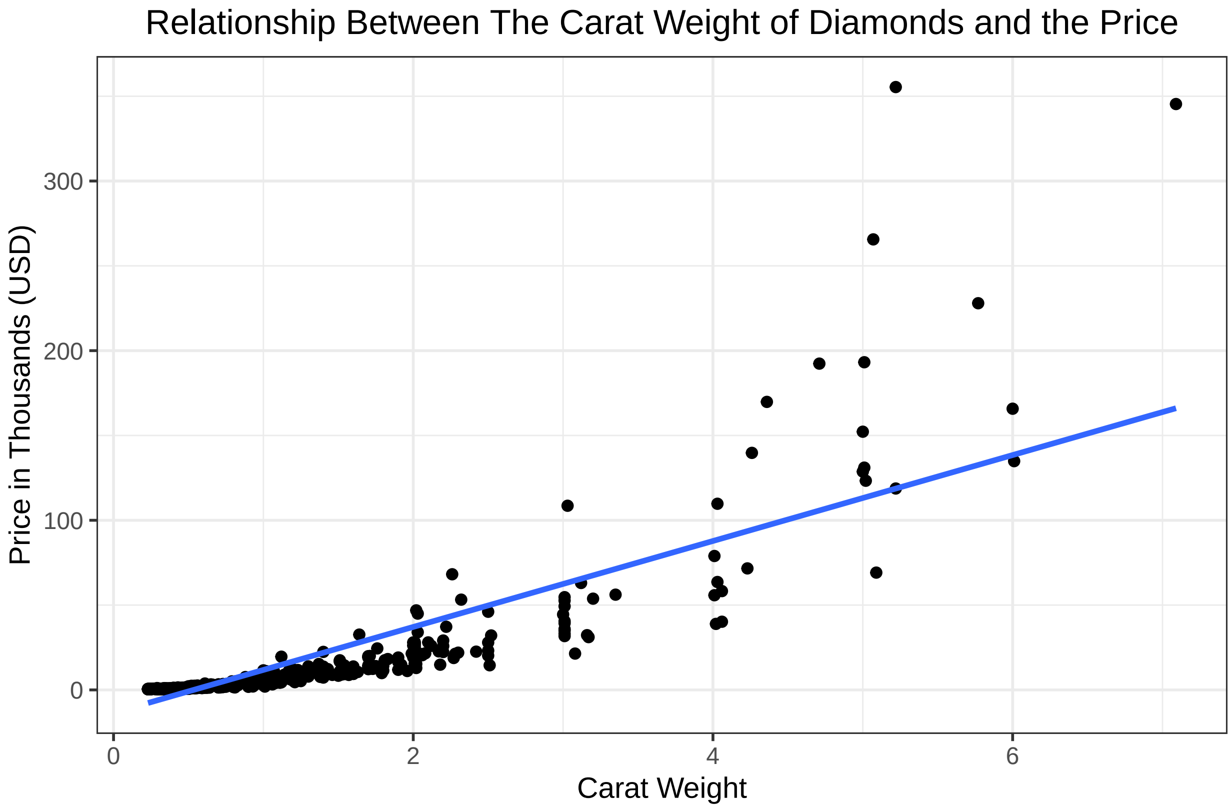


Figure 1

Refer to Figure 1. Generally, the price goes up when the carat goes up. The relationship between the carat weight and the price seem to follow a curvilinear trend, which suggests a transformation for the carat variable. The variance also seems to be increasing, suggesting a transformation for the price variable. This will be further discussed in section 3.

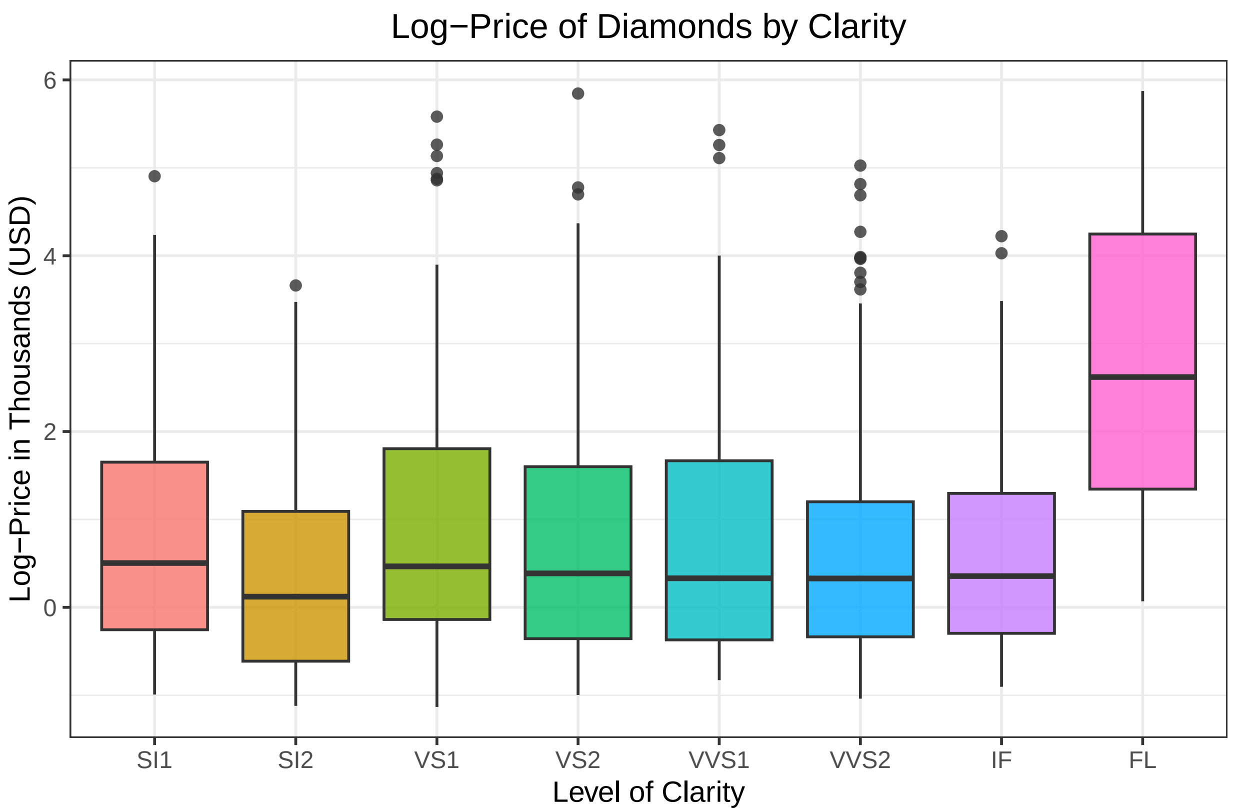


Figure 2.1

Figure 2.1 above clearly displays the relationship between the log-price and clarity of diamonds. Observing the median supports that the grade of clarity does not significantly change the price of the diamonds. But it is important to remember that the sample size of both IF and FL is much smaller than the other clarity categories. So it appears there is no relationship between the two but there is some slight uncertainty.

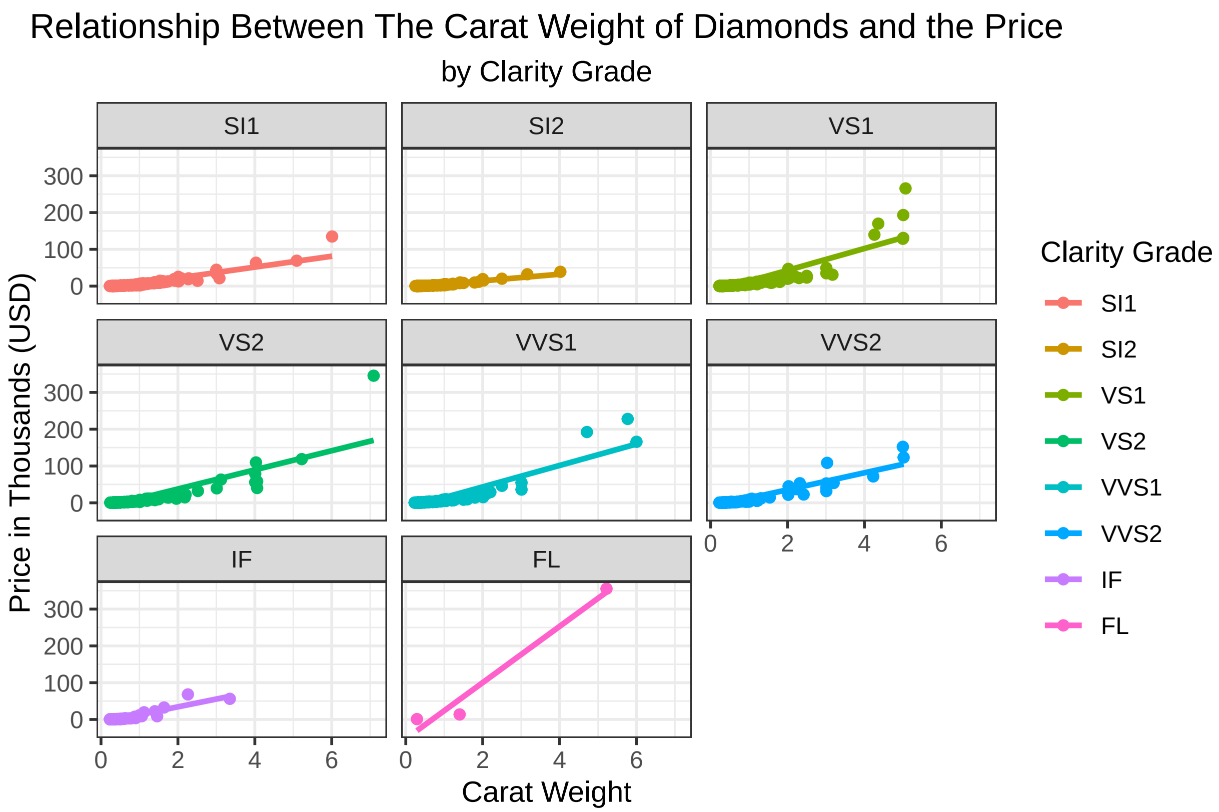


Figure 2.2

Figure 2.2 shows the relationship between carat weight and clarity grade. Mostly, the lower carat weight diamonds are around the same price range, regardless of clarity grade. This graph further supports the observations concluded in Figure 2.1. It is currently unknown if the difference is statistically significant. Most of the different clarity grades in Figure 2.2 also seem to rise in price in a similar fashion as well since most of them seem to have a curvilinear relationship with price. The two Slightly Included grades and the Flawless grade are noted to be seemingly more different to the other 5 categories.

A better visualization is shown below (Figure 2.3).

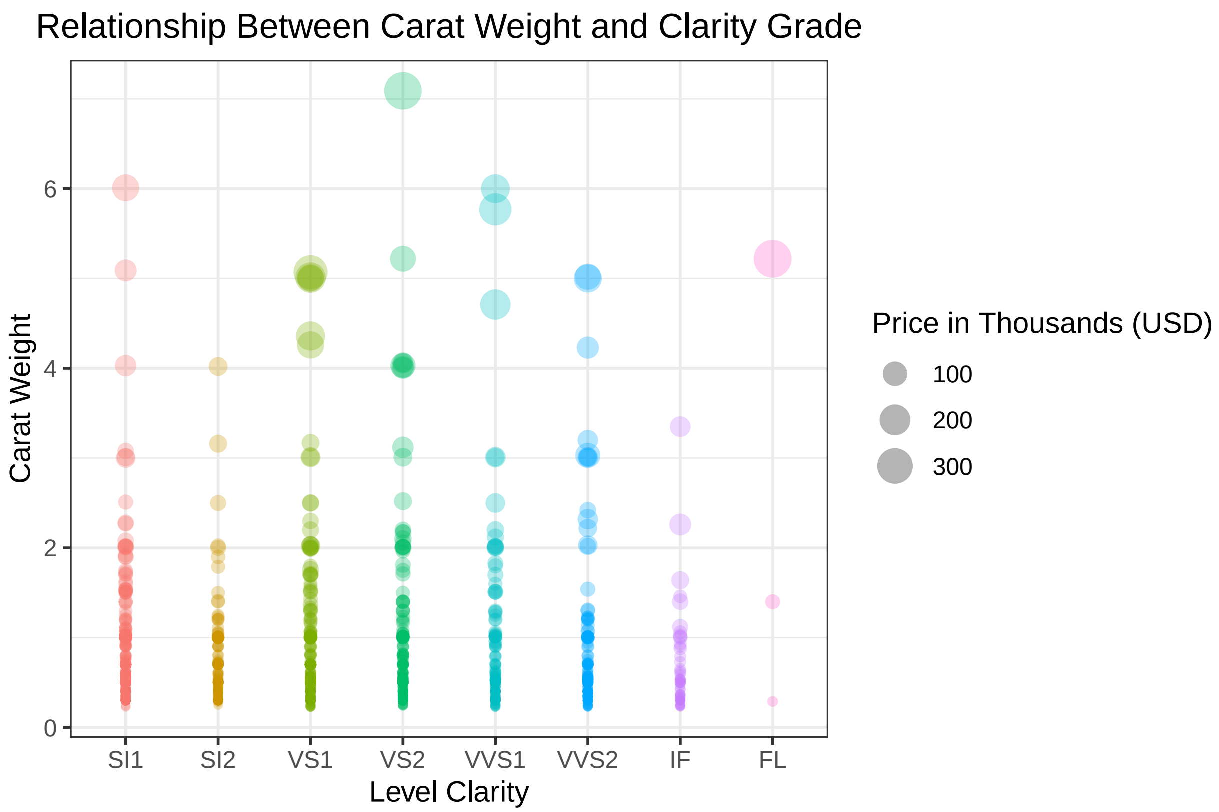


Figure 2.3

Figure 2.3 shows a clearer image of price increase when both the carat and the clarity increase. Although there are some slight outliers, which is likely due to the other variables given by **Blue Nile**. The graph also gives a clearer image of how few data points there are in the IF and FL clarity levels.

Please note that while **Blue Nile** does offer a limited selection of preset jewelry with I1 diamonds, there were no records of Included Diamonds within this data set.

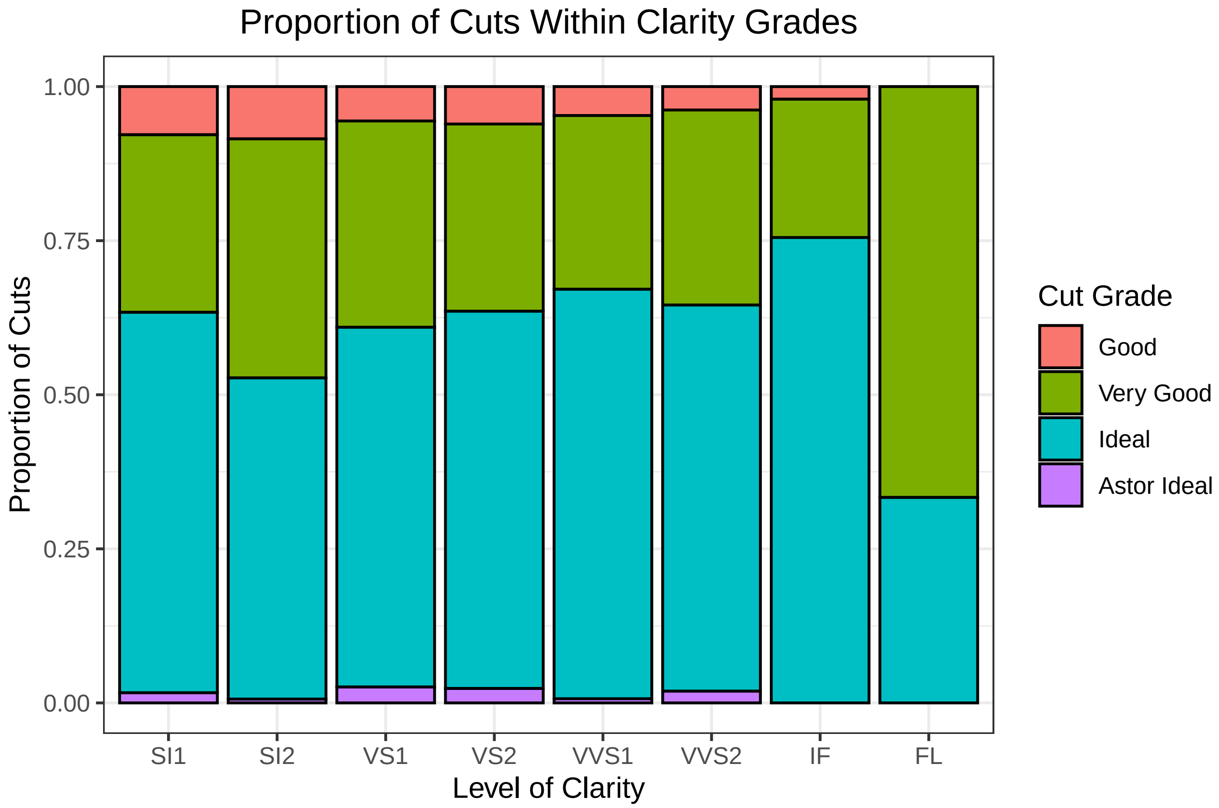


Figure 2.4

Figure 2.4 above displays the relationship between the type of cuts and the levels of clarity. This shows that most of the cuts are ideal regardless of each level of clarity.

It is observed that for the highest two levels of clarity, there appears to be no diamonds with the Astor ideal cut. This is likely due to the rarity of the IF and FL level diamonds. Within the 1214 diamonds we were given data for, 49 were IF clarity and 3 were FL clarity. Regardless of the diamonds not having any Astor Ideal cuts, they still make up the most expensive diamonds in the data set.

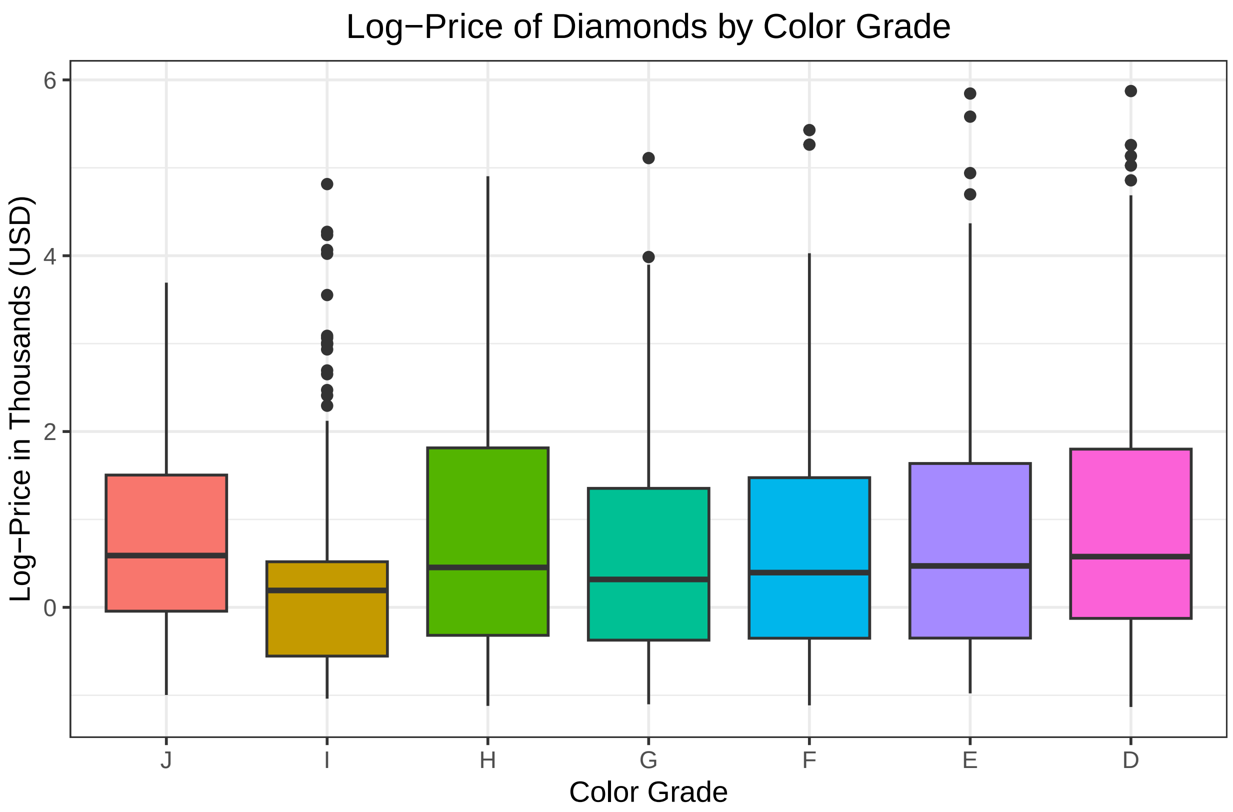


Figure 3.1

A log transformation was done on the price variable for a better view by taking the log of the price variable. The resulting graph, Figure 3.1, suggests that the color of the diamond does not significantly affect the price as the median seems to stay relatively constant throughout the different color grades. An exception may be noted with color grade I, but the difference may or may not be negligible.

Note that we do not have all the type of color in our data set.

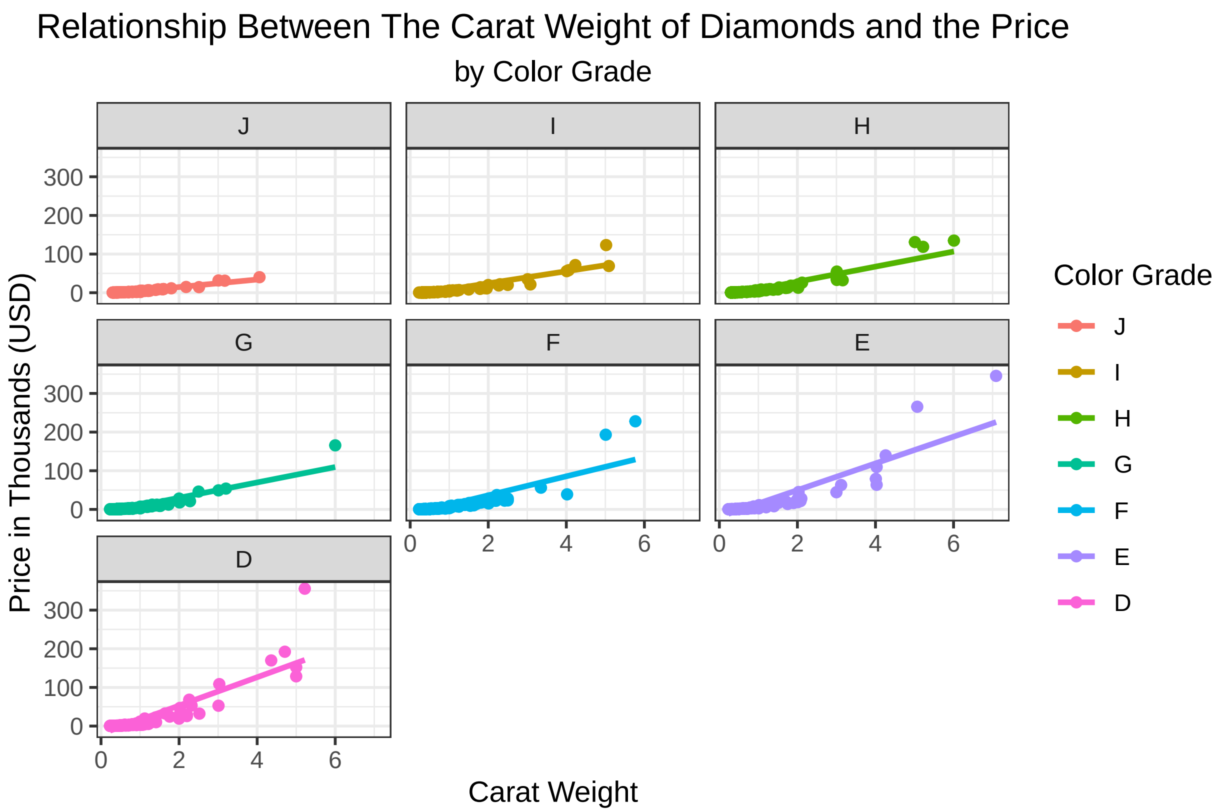


Figure 3.2

Refer to the graph labelled Figure 3.2. As shown by the graph, there seems to be a faster rise in price as color grade increases, but the difference may or may not be statistically significant. It also seems to be that as color grade increases in quality, the curvilinear trend becomes more prominent. A supplementary visual is provided with the same information (Figure 3.3).

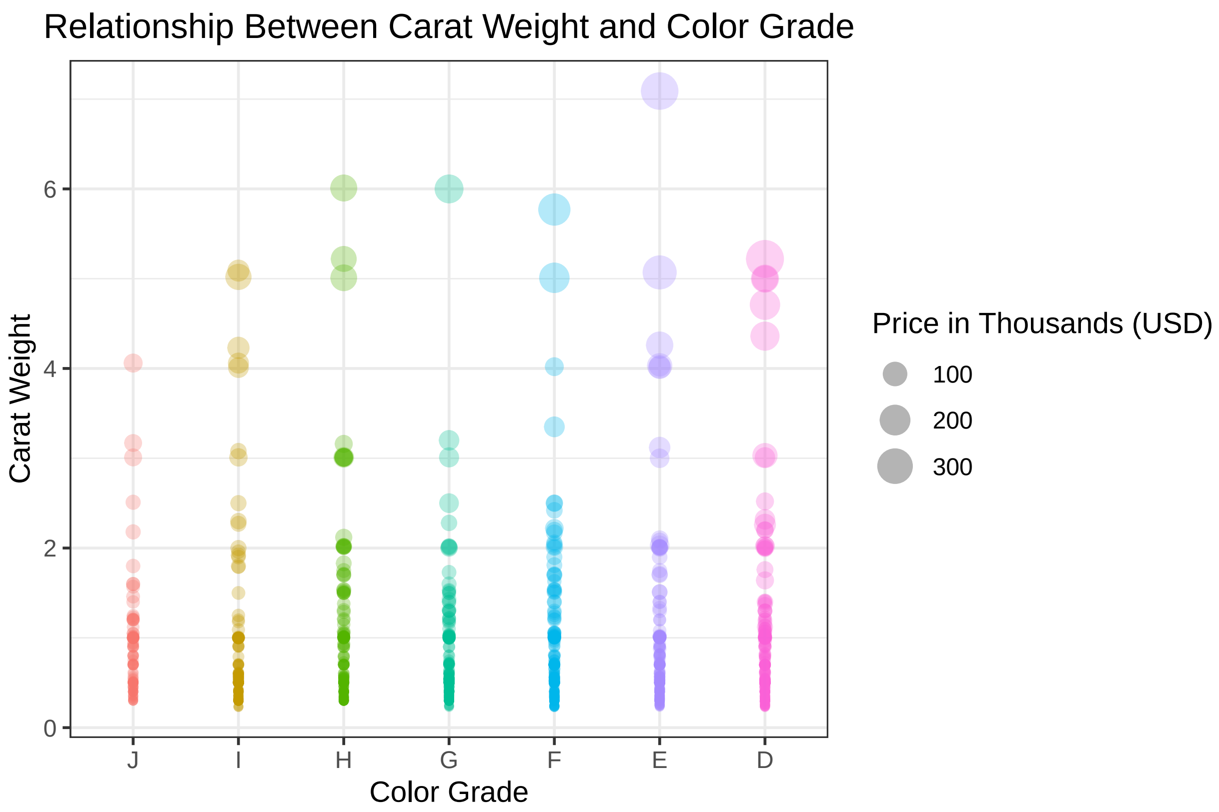

Relationship Between Carat Weight and Color Grade
Carat Weight
0
2
4
6
J
I
H
G
F
E
D
Color Grade
Price in Thousands (USD)
100
200
300

Figure 3.3

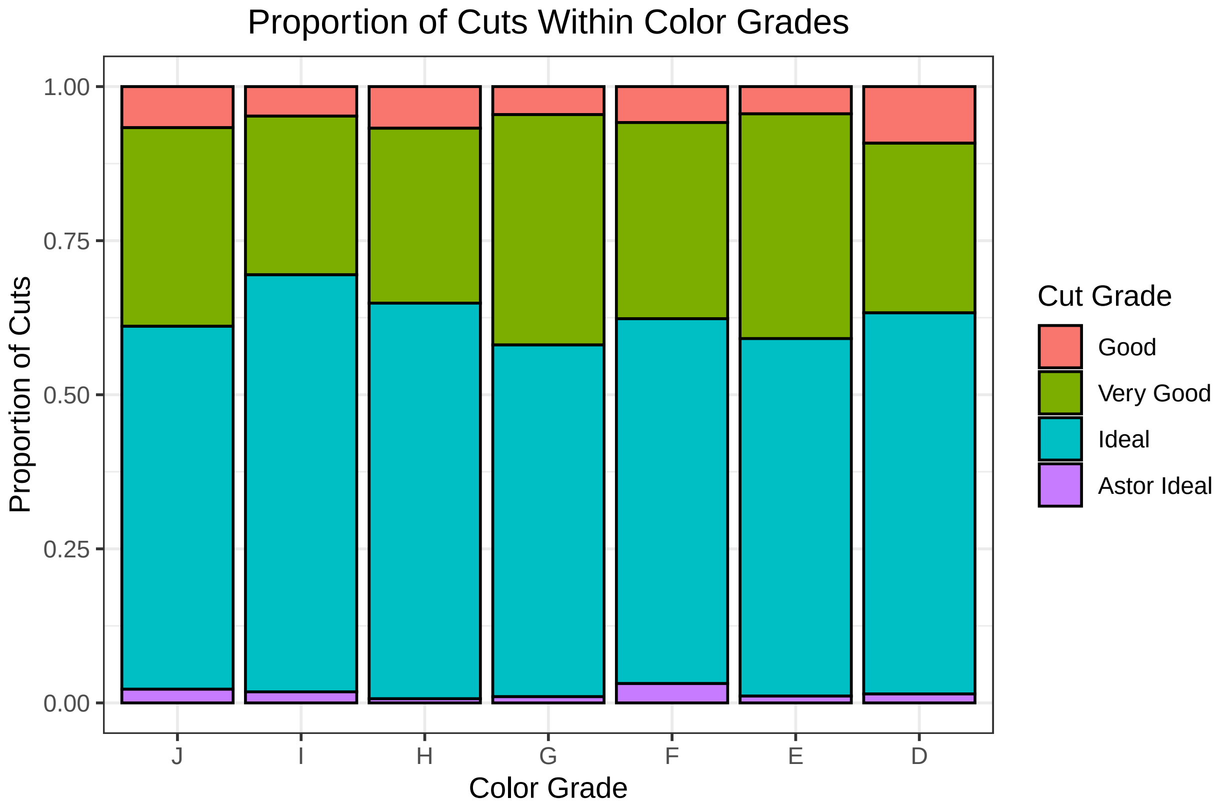


Figure 3.4

The relationship between cut type and colour grade is shown in Figure 3.4. It is noted that there is not a significant difference cut type across color grade. This indicates that a relationship may not be present between these two variables.

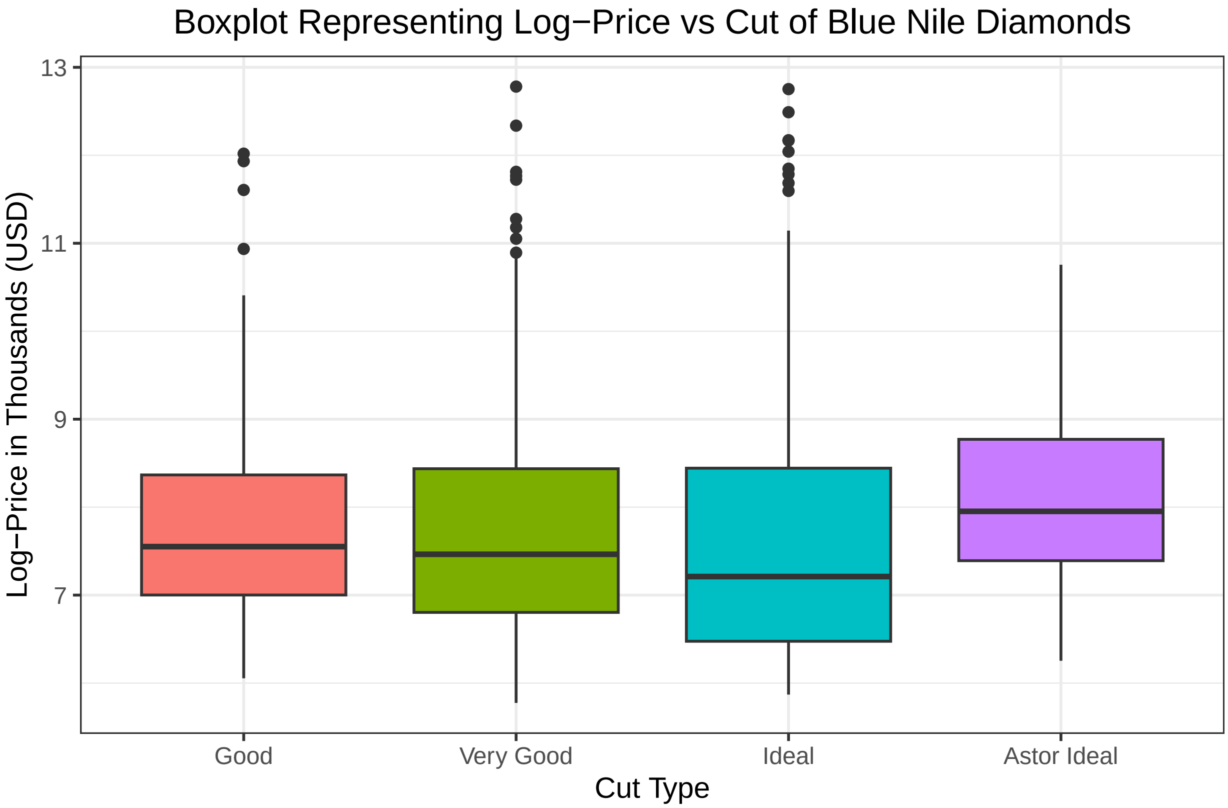


Figure 4.1

Figure 4.1 shows the summary statistics for the log-price to create a more visually effective boxplot. Judging by the figure, we can see that the IQR of prices of all cut types is about 7-9. Astor Ideal has the highest median out of the cut types, but the difference may or may not be negligible. Good and Very Good cuts have almost identical medians, but the Very Good cut has a larger range of prices and a much higher maximum price. Surprisingly, Ideal cuts have the biggest IQR and the first quartile of this box has the lowest starting point.

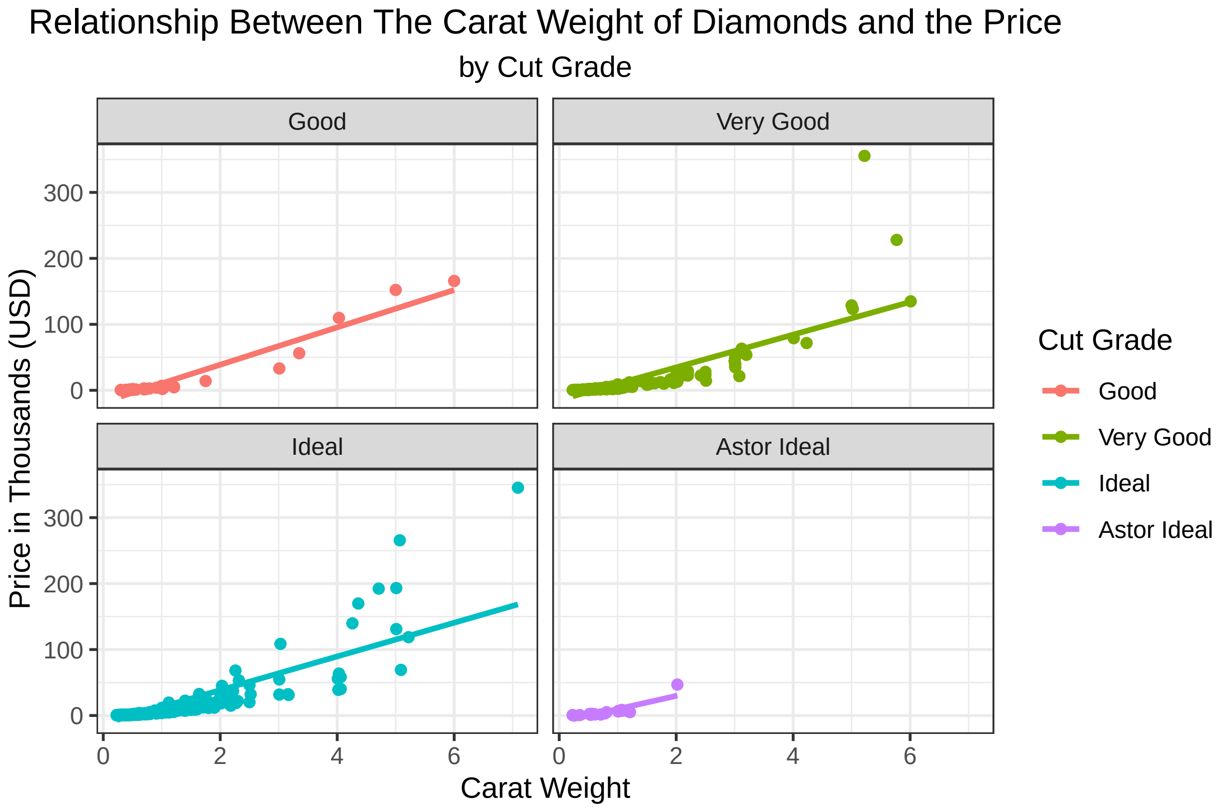


Figure 4.2

Figure 4.2 shows that Very Good and Ideal cuts seem to have a more curvilinear trend. But, similar to the overall conclusion drawn with the clarity and color variables, the difference might not be statistically significant. More testing needs to be done. In addition, As carat increases, price also increases for all diamond cut types. By observation, Good cut has the steepest slope. The highest prices belong to Very Good and Ideal cuts. According to the data set, Astor Ideal cuts usually have smaller carats and lower price, but **Blue Nile** has stated that their super-ideal Astor is the most expensive cut, which contradicts the observation.

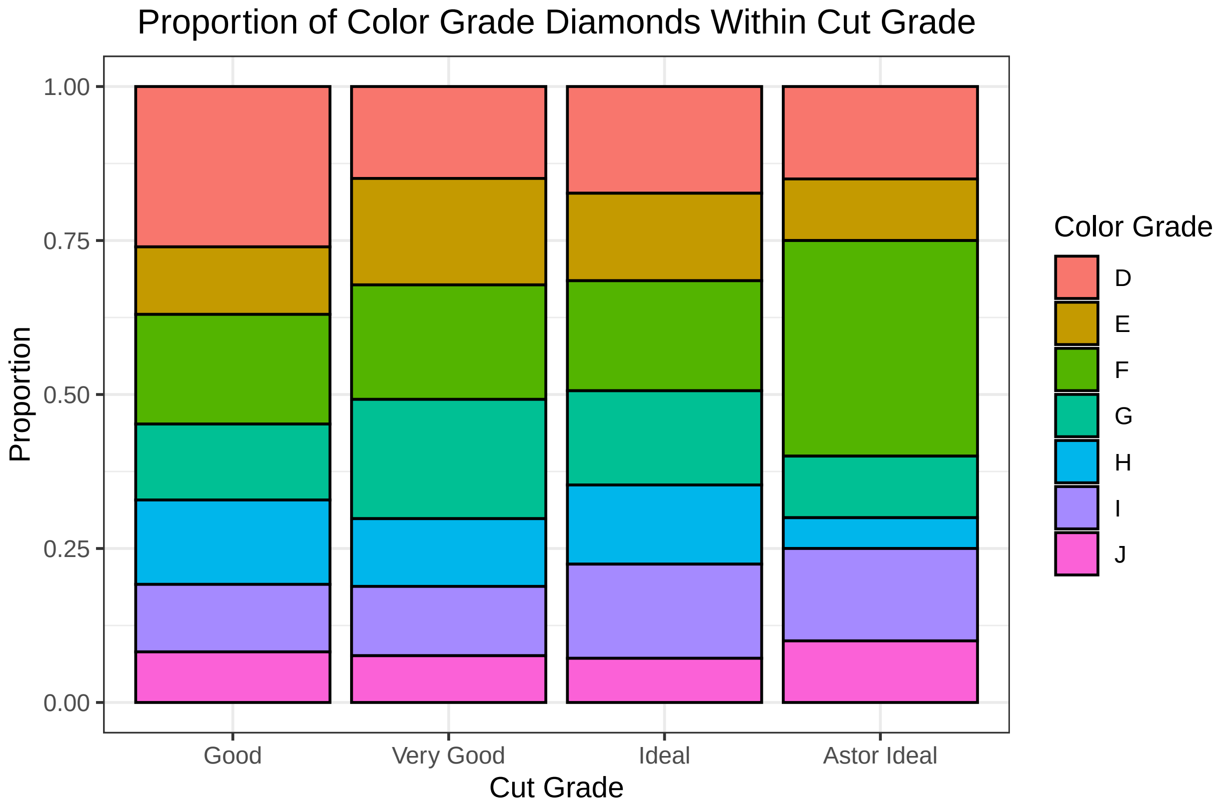
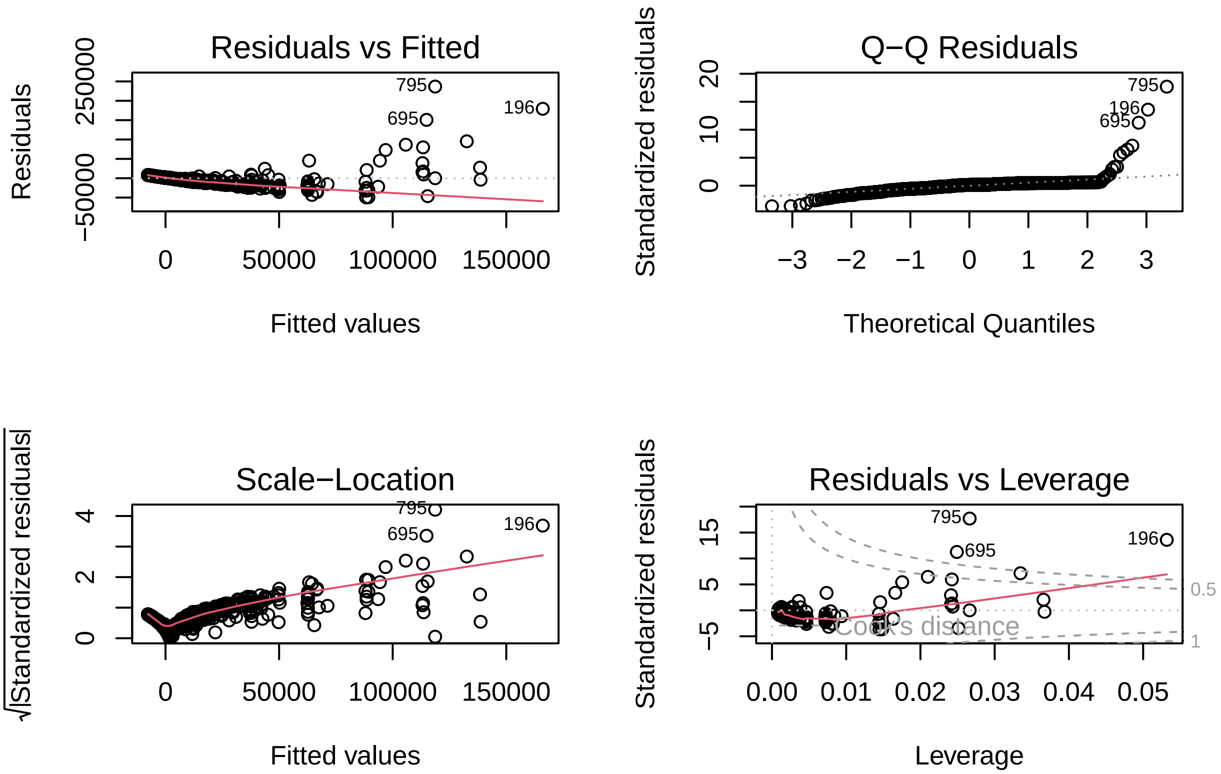


Figure 4.3

The relationship between cut and color was observed using a stacked bar graph (Figure 4.3). The highest quality color is considered 'D', and the lowest, 'J'. Overall, there doesn't seem to be too much difference in the colors of each cut. However, we can make the observation that Good cut has the largest proportion of D color, while Astor Ideal has the largest proportion of F color. When thinking about the price points of these cuts and the quality of colors, we know that Astor Ideal doesn't have a high price point, which is in line with the color quality. On the other hand, we know that Very Good and Ideal cuts have the highest prices, but this is not reflected in the colors that are used for these cuts. All colors seem to be used pretty evenly for both cuts, with no overwhelming use of any color to create an association with price. This suggests that there may not be a relationship between the two variables.

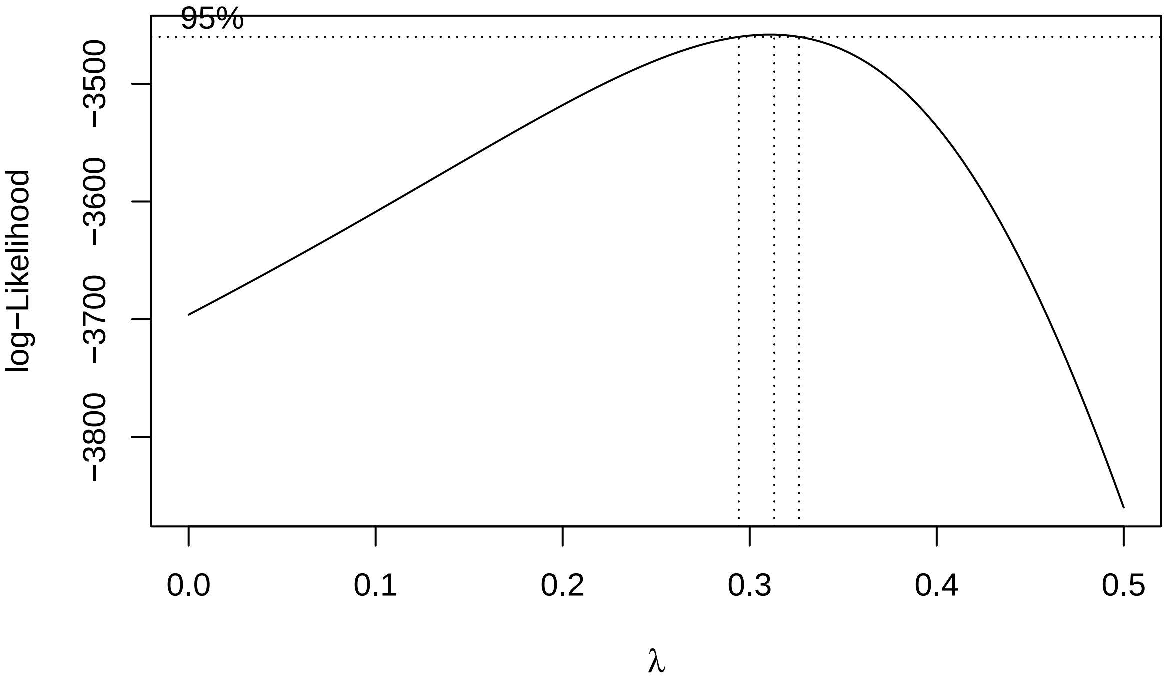
Section 3: Simple Linear Regression for Price Against Carat Weight

Figure 5.1



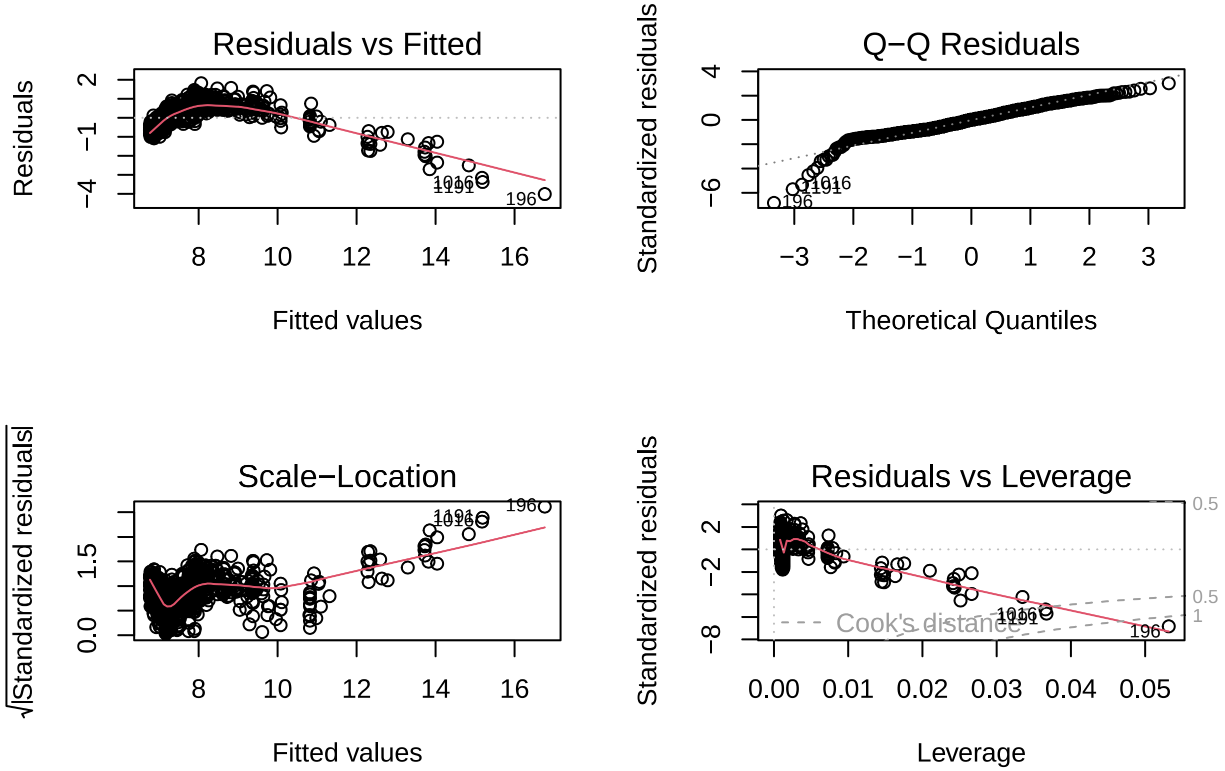
A simple linear regression analysis was done for the variable carat weight. However, as mentioned before and seen in the residual plots (see: Figure 5.1), the variance increases as the price (fitted values) increases and the regression line seems to either under- or over-estimate the price. This indicates that a transformation on the response variable (price) and possibly the predictor variable (carat) may be needed in order to successfully create a linear regression model.

Figure 5.2



The boxcox function from the R package MASS was used to determine an appropriate transformation for the response variable (Figure 5.2).

Figure 5.3


Residuals vs Fitted
Residuals
2
−1
−4
8
10
12
14
16
Fitted values
1016
1191
196
Q−Q Residuals
Standardized residuals
4
0
−6
−3
−2
−1
0
1
2
3
Theoretical Quantiles
1016
1191
196
Scale−Location
√|Standardized residuals|
1.5
0.0
8
10
12
14
16
Fitted values
1191
1016
196
Residuals vs Leverage
Standardized residuals
2
−2
−8
0.00
0.01
0.02
0.03
0.04
0.05
Leverage
Cook's distance
0.5
0.5
1
1016
1191
196

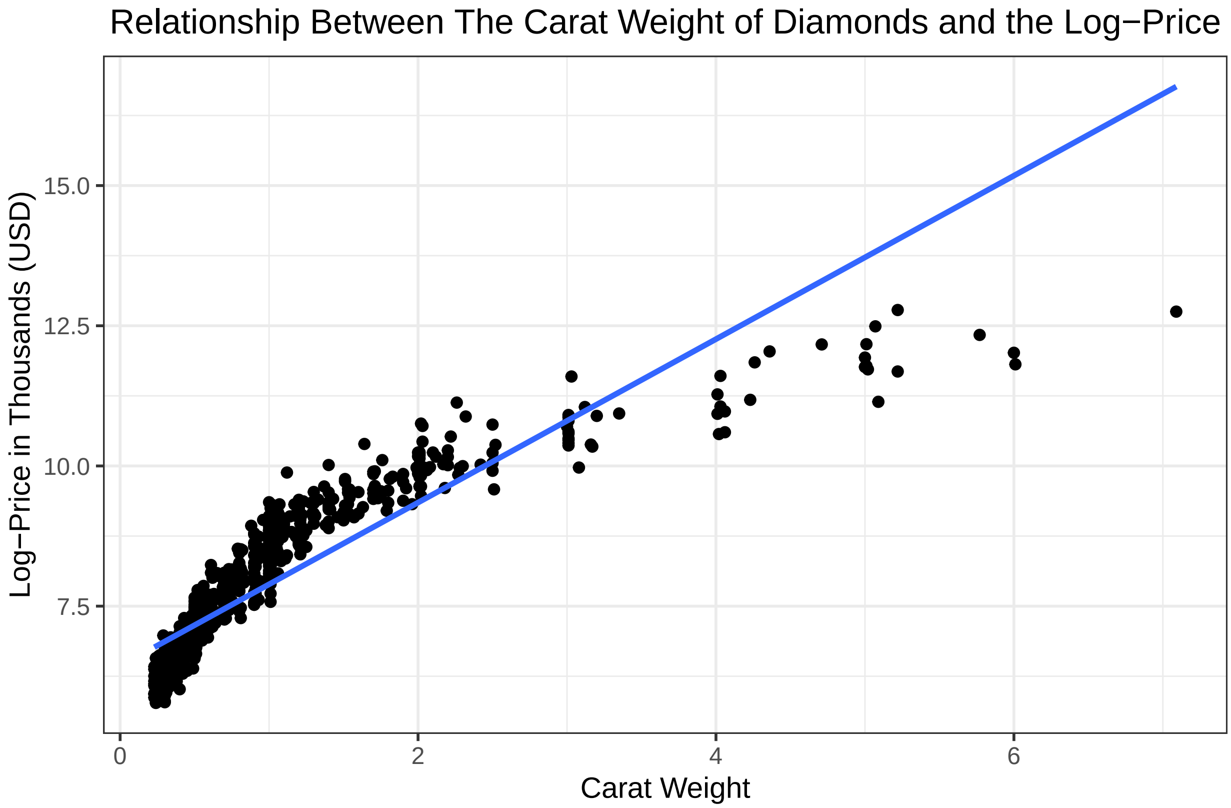


Figure 5.4

Based on the results, a log transformation was done on the response variable and the residual plots were redrawn (Figure 5.3). Figure 5.3 shows that the variance has stablised.

However, the scatter plot (Figure 5.4) still shows the log-price and carat weight relationship to be a curvilinear trend, indicating that a tranformation of the predictor variable (carat weight) is still needed. Based on the trend, a log transformation on the predictor variable was done.

Figure 5.5

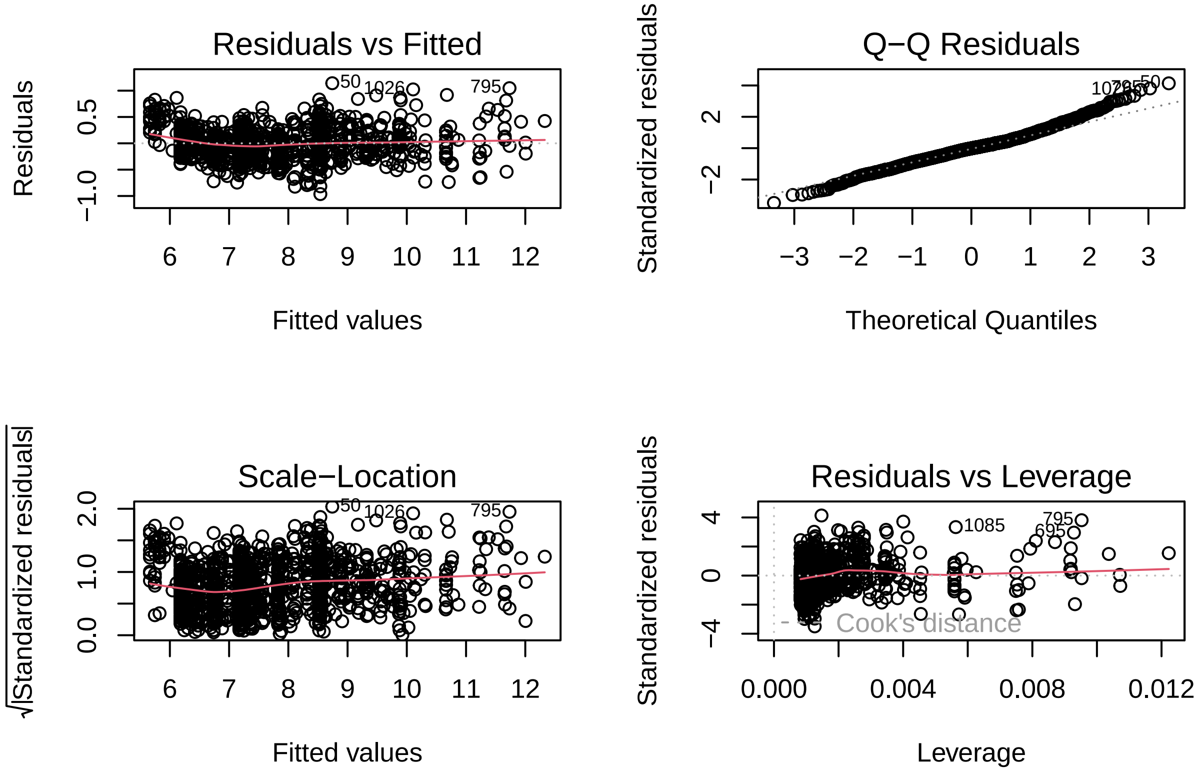


Figure 5.5 ascertains that the assumptions for simple linear regression hold. The data points are evenly distributed over the regression line, meaning the mean of the residuals for each fitted value is equal to zero, and the QQ plot shows the quantiles to be along the 45-degree angle, indicating that the distribution is normal. In addition, the variance over the horizontal band is fairly constant, as seen in the residual plot.

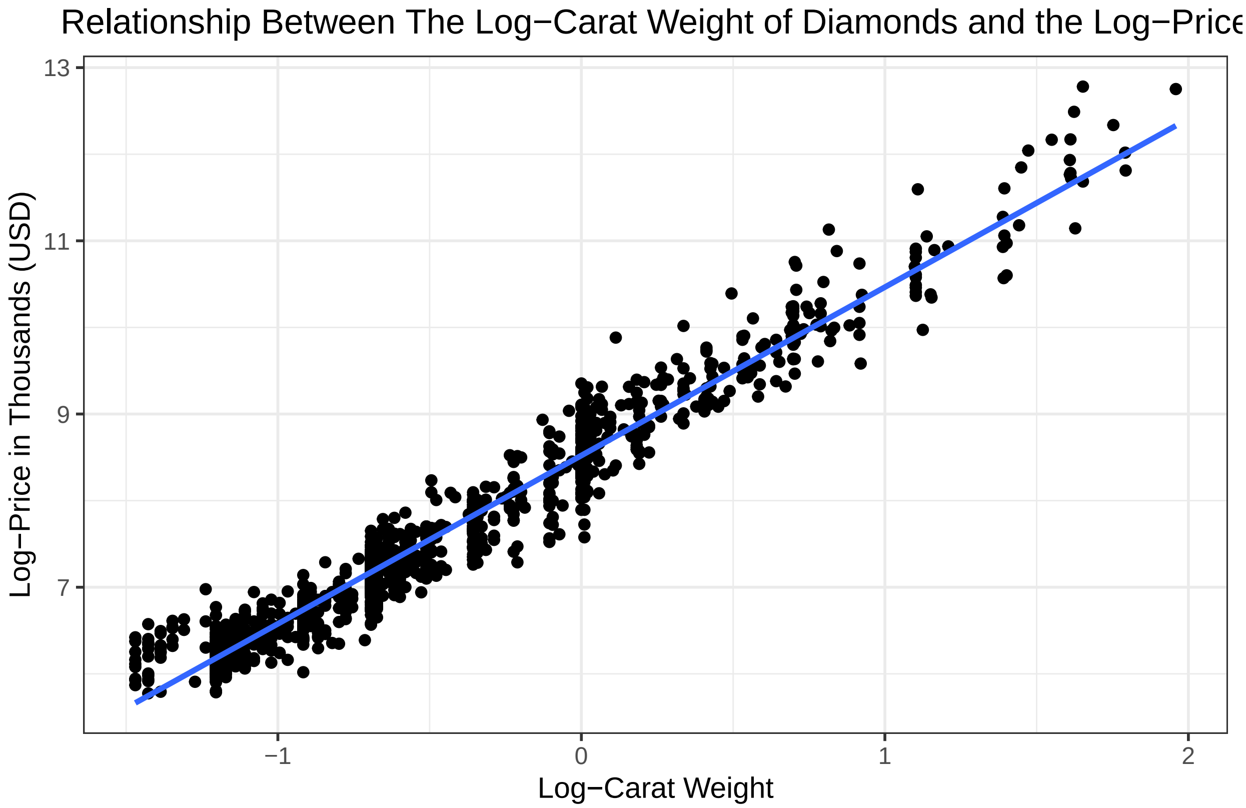


Figure 5.6

The final scatterplot (Figure 5.6) displaying the linear relationship between the carat weight and price of **Blue Nile** diamonds is shown.

With the regression analysis on the relationship between carat weight and diamond price complete, we can confirm that the two variables have a positive, linear relationship. This means that as for 1% increase in carat weight, the predicted price of a diamond by a certain value, provided that other possible influential variables are held constant.