Stat 6021: Project 1

Liam Donoghue, Clay Harris, Hai Liu, and Brian Nolton

2024-10-26

# Section 1: Summary of findings.

We looked at over 1,200 diamonds on the Blue Nile website, each graded on the “4Cs,” or cut, color, clarity, and carat weight. These are what impacts a stone’s beauty and value. We found that despite several Blue Nile claims, the single most important factor about a diamond is the carat weight. Why? Because it’s inarguably the one C that can predict the price of the diamond, regardless of the other Cs.

In fact, for every 1% increase in carat, the price of a diamond increases by roughly 2%.

Blue Nile purports cut is the most important of the 4Cs. That may be true if you’re most interested in a diamond’s more hard-to-measure characteristics like scintillation, fire, brilliance, or balance. But if balancing your checkbook is most important, then start with carat.

Of course, when looking at two diamonds that are the same size, the one with the better cut is typically going to be subjectively “prettier” as well as objectively pricier. (It also happens to be the one that jewelers have the most control over, so there is an incentive for the consumer to think it’s the most important.)

The same can be said for two same-sized diamonds with different color grades or clarities.

Blue Nile claims Color is the second most important C, as more colorless diamonds are rarer. But every diamond Blue Nile carries is either “colorless” or “near-colorless,” and most folks buying diamonds don’t care too much about their near-colorless diamond having “no discernible color,” and the prices reflect this. Whereas the most expensive diamonds Blue Nile carries tend to be graded higher on this color scale, the price of a typical diamond doesn’t change too much when going from the lowest to the highest color.

Blue Nile suggests Clarity is the least important of the 4Cs, and we couldn’t agree more. Most blemishes, even in the lowest clarity diamonds, can’t be seen with the naked eye. They say inclusions become more visible as diamond size increases, and we clearly see this is the case. The best value diamonds also happen down at the lower end of the clarity scale.

One final claim Blue Nile makes is an important one: buy shy. What this means is that deals can be had if you look for a diamond just below a round carat weight number: say 0.9 carat rather than 1, or 1.9 rather than 2. You’ll effectively pay a premium (over what you’d normally pay per carat) for hitting that round number.

# Section 2. Data descriptions, visualizations, and addressing claims.

## 2.1. A description of the data and the variables.

First, we load the data set from the .csv file and have a look at the data structure.

## carat clarity color cut   
## Min. :0.2300 Length:1214 Length:1214 Length:1214   
## 1st Qu.:0.4000 Class :character Class :character Class :character   
## Median :0.5200 Mode :character Mode :character Mode :character   
## Mean :0.8134   
## 3rd Qu.:1.0000   
## Max. :7.0900   
## price   
## Min. : 322.0   
## 1st Qu.: 723.5   
## Median : 1463.5   
## Mean : 7056.7   
## 3rd Qu.: 4640.8   
## Max. :355403.0

## [1] 1214 5

As we can see from the data itself and a quick summary of the data, the data set is composed of 1214 observations and 5 variables with:

* Two variables carat and price that are numeric or quantitative, and
* Three variables clarity, color, and cut that are categorical or qualitative.

Based on the context, the variable price should be considered as the response variable for this analysis. The minimal price of the diamond recorded in this data set is $322 and the maximal price is $355403. Also, 50% of the diamonds in the middle price range are between $723.5 and $4640.8. The mean price of the diamonds is $7056.7.

The second quantitative variable carat measures the weight of each diamond (1 carat = 200 milligrams). The minimal weight of the diamond in carat recorded in this data set is 0.23 and the heaviest diamond is 7.09 carat.

The variable cut measures how well-proportioned a diamond’s dimensions are, including its balance and brilliance. In this data set, we have four levels for the quality of cut, *i.e.*, Good, Very Good, Ideal, and Astor Ideal, from low to high based on the information from the diamond education page.

##   
## Good Very Good Ideal Astor Ideal   
## 73 382 739 20

The variable color refers to how colorless a diamond is. The less color a diamond has, the higher the diamond color grade. In this data set, there are seven levels for the grade of color, *i.e.*, J, I, H, G, F, E, and D, from low to high.

##   
## J I H G F E D   
## 90 167 148 198 223 181 207

Clarity is used to quantify and specify any inclusions naturally occurring during the diamond forming process for both natural and lab diamonds. The variable clarity assesses small imperfections within a diamond. In this data set, there are eight levels for the grade of clarity, *i.e.*, SI2, SI1, VS2”, VS1, VVS2, VVS1, IF, FL, from low to high.

##   
## SI2 SI1 VS2 VS1 VVS2 VVS1 IF FL   
## 165 243 214 233 158 149 49 3

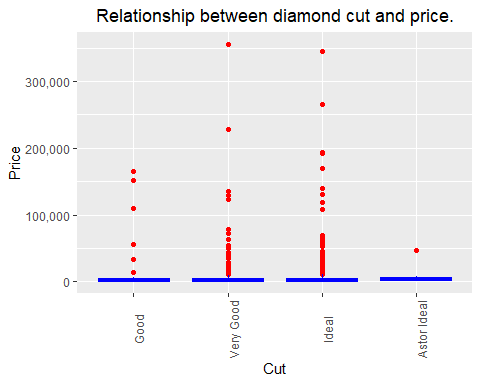
## 2.2. Data wrangling for the visualizations to explore how price is related to the other variables (carat, clarity, color, cut).

To get data well-organized for a good visualization and easy for readers to grasp, we first created a new variable clarity\_less to reduce the levels of clarity according to the information from the diamond education page.

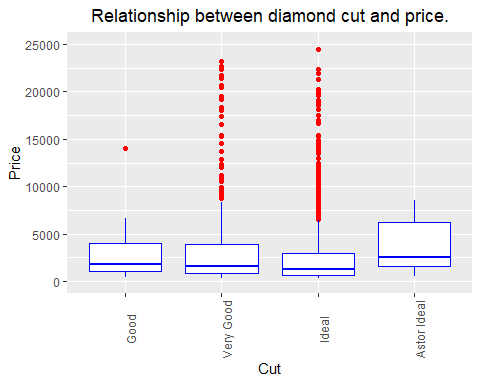
Second, we rearranged the order of the levels for the categorical variables cut, color, and clarity\_less. Here, we rearranged the order from the lowest to the highest in quality based on the information from the education page.

## 2.3. Data visualizations to study the relationship between price and each one of the variables.

We first studied the relationship between price and cut with a box plot.



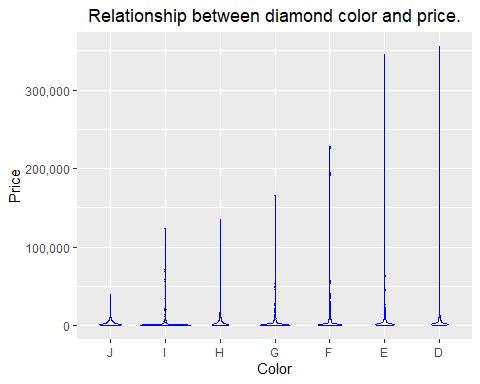
**Figure 2.1A Box plot between diamond cut and price**



**Figure 2.1B Box plot between diamond cut and price under $25k.**

As shown in Figure 2.1A, it is clear that most of the very expensive diamonds (*e.g.*, $25k or more) come with either Very Good or Ideal cut quality. From Figure 2.1B, we can see that the majority of the diamonds are still priced under $10k.

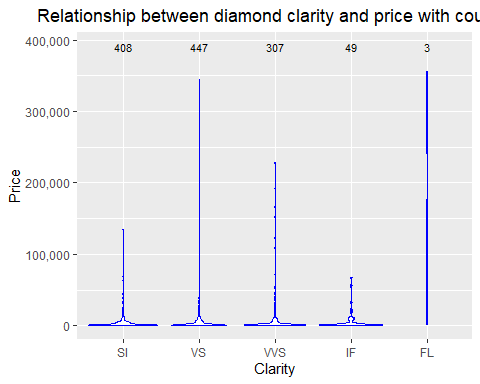
We then looked at the relationship between price and color with a violin plot.



**Figure 2.2 Violin plot between diamond color grade and price.**

From Figure 2.2, we can see that as the color grade goes up, there are more of those very expensive diamonds.

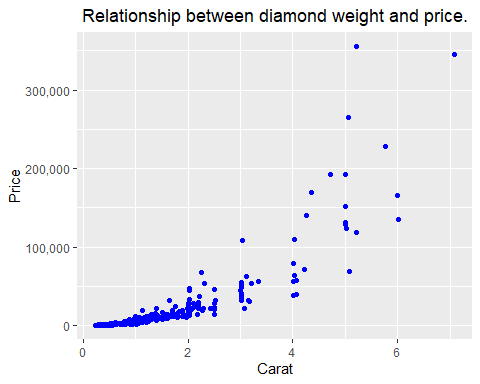
We also looked at the relationship between price and clarity\_less with a violin plot.



**Figure 2.3 Violin plot between diamond clarity grade and price with counts.**

Interestingly, from Figure 2.3 we can see that the low (*e.g.*, SI) and intermediate clarity grades (*e.g.*, VS and VVS) contain the vast majority of the diamonds, while still having a broad range of prices including many more of the most expensive diamonds (*e.g.*, ones that cost $50k and more) than the other clarity grades. Otherwise, the data also reflect that it is very rare to see diamonds with top clarity features such as IF and FL. In this data set, there are only 49 diamonds with IF clarity and 3 with FL clarity.

Finally, we looked at the relationship between price and carat with a scatter plot.

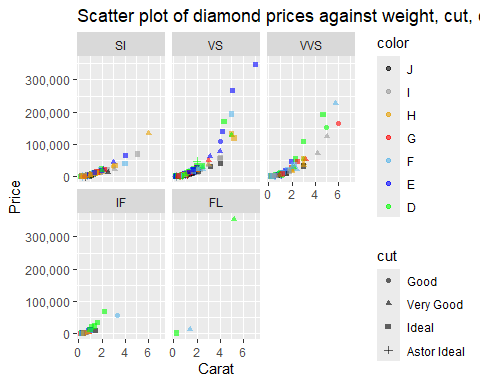


**Figure 2.4 Scatter plot between diamond carat and price.**

As we can see from Figure 2.4, the price exhibits almost an exponential increase as the weight of the diamonds (carat) increases, indicating a strong relationship between diamond price and weight.

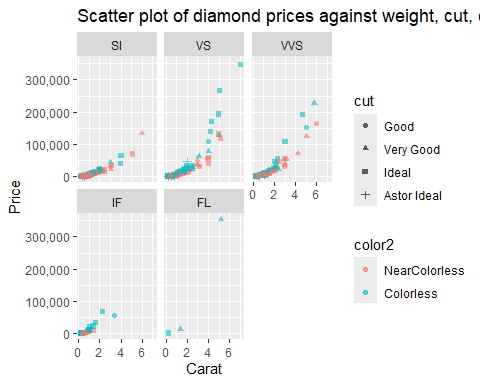
## 2.4. Multivariate visualizations to address how price is related to the other variables at the same time.

We generated scatter plot matrix with each plot displaying the price for diamonds separated by the grade of their clarity.



**Figure 2.5A Scatter plot summarizing diamond price, carat, color, cut, and clarity.**

From Figure 2.5A, we found the variable color started to make the plots difficult to comprehend due to its number of levels (seven), so we created a new variable color2 to reduce the levels of color into two (*i.e.*, Colorless and NearColorless) based on the information from the education page and gave them a correct order.



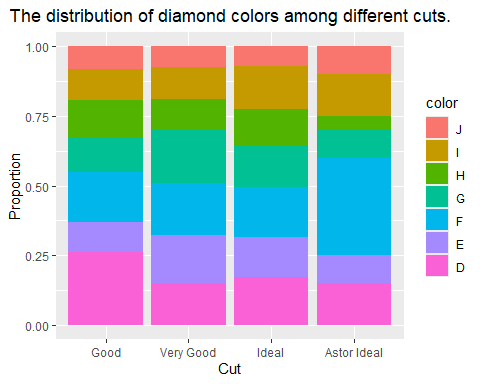
**Figure 2.5B Scatter plot summarizing diamond price, carat, color, cut, and clarity.**

From Figure 2.5B, it’s clear that within each level of clarity, price is positively related to carat. For the diamonds with the same or very similar weight (carat), the ones with higher cut qualities tend to be more expensive than those with lower cut qualities. Also, the ones with higher color grades are likely to be more expensive than the ones with lower color grades.

## 2.5. Pairwise comparison between the variables other than price.

We generated a series of bivariate bar charts to study the relationship between three categorical variables.

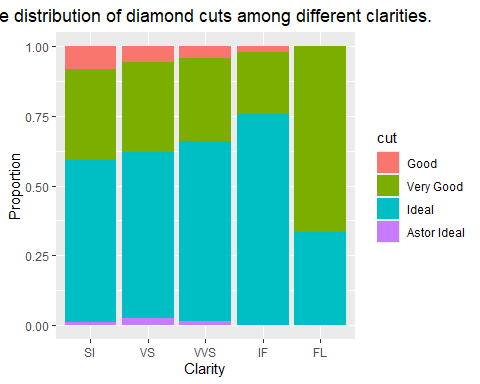
### 2.5.1 Cut vs. Color.



**Figure 2.6 Bar chart summarizing proportions of each color grade among different cut grades of the diamonds.**

Here, we looked at the distribution of the color among different cut grades of the diamonds. We found that the Astor Ideal cut diamonds also have the most top-grade colors (D, E, and F).

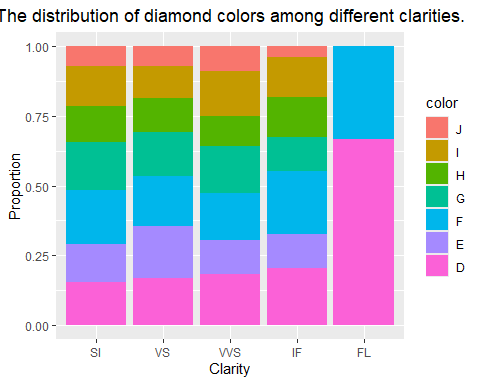
### 2.5.2 Cut vs. Clarity.



**Figure 2.7 Bar chart summarizing proportions of each cut grade among different clarity levels of the diamonds.**

As shown in Figure 2.7, the Ideal cut diamonds take up more percentages as the clarity of the diamonds improves.

### 2.5.3 Color vs. Clarity

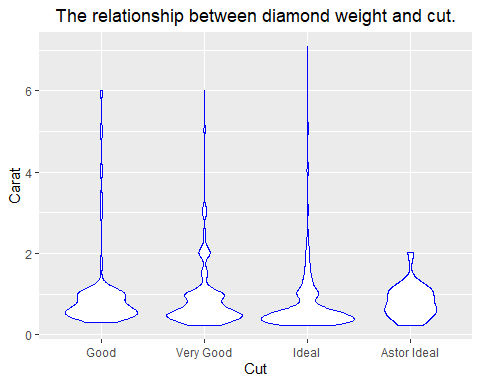


**Figure 2.8 Bar chart summarizing proportions of each color level among different clarity levels of the diamonds.**

From Figure 2.8, we can see that the diamonds with higher grade clarity tend to have a higher proportion of the best color quality (D). For the diamonds with clarity of FL, there are exclusively two top color grades (D and E).

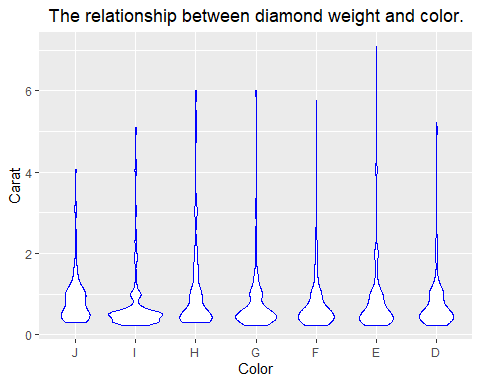
We generated a series of violin plots between carat and each categorical variable.

### 2.5.4 Carat vs. Cut.



**Figure 2.9 Violin plot of diamond carat and cut.**

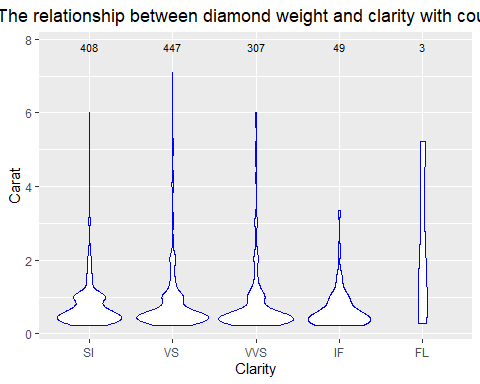
### 2.5.5 Carat vs. Color.



**Figure 2.10 Scatter plot of diamond carat and color.**

From Figure 2.9 and 2.10, we see that the diamonds become rarer and rarer as the weight increases, regardless of their cut and color.

### 2.5.6 Carat vs. Clarity.



**Figure 2.11 Violin plot of diamond carat and clarity.**

We can see from Figure 2.11 that there are fewer diamonds as the carat and clarity both increase, which may reflect the nature of the occurrence rate for diamonds that are both very heavy and very pure.

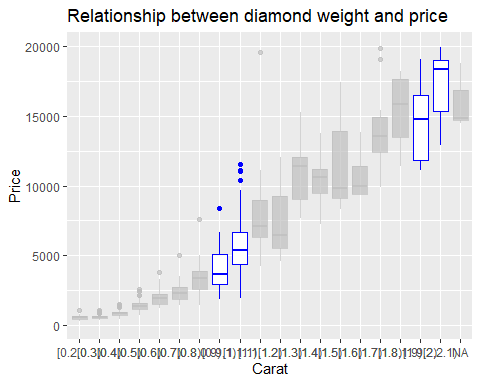
## 2.6 Addressing Blue Nile claims.

On Blue Nile’s website, they make a number of claims, one about each of the 4Cs. We will examine the claims they make about each one of the 4Cs here. We are examining these claims on their individual effect on the response variable of price. We understand that there are other factors in play (like how each of the 4Cs interact with each other and shape), but those are outside the scope of this paper.

### 2.6.1 Claim regarding carat.

**Claim: Buy shy to save money**

This claim suggests that prices “jump” at 0.5 carats, 1 carat, 2, 3, etc., compared to 0.49, 0.9, 1.9, etc. The below chart shows that there is certainly a significant jump in price when going from just under 1 to 1 carat, as well as just under 2 to 2 carats.



### 2.6.2 Claims regarding cut.

**Claim: Cut is the most important of the 4Cs.** **Claim: The more expensive the cut, the more expensive the diamond.**

“Cut measures how well-proportioned a diamond’s dimensions are, including its balance and brilliance. This influences how well a diamond interacts with light, impacting its brilliance, fire and scintillation. A well-cut diamond will look its very best with optimal light performance.” The grade of cut ranges from “Poor” to “Ideal” (the Gem Institute of America uses the term “Excellent” but they are the same thing) for most jewelry stores with Blue Nile adding an additional “Astor Ideal” cut for the best diamonds they can possibly find. Blue Nile says they do not carry the lower cuts of diamonds so their cut range goes from “Good” to “Astor Ideal”. Cut is the one of the 4Cs that jewelers have the most control over. A diamond is whatever color, clarity, and carat it is as it is found in nature (or potentially lab grown). The jewelers may be able to cut around certain imperfections to make it more desirable, but they can only do so much. While the physical dimensions the diamond comes in cannot be changed, they have more control over how they cut the diamond into the grades and shapes (not one of the 4Cs). It is easy to see that they prioritize cut simply by the number of diamonds they carry in each cut grade.

##   
## Good Very Good Ideal Astor Ideal   
## 73 382 739 20

We can see that the “Ideal” cut diamond accounts for the majority of the cuts at Blue Nile, with “Very Good” coming in a distance second. The “Astor Ideal” cut is the least confirming that Blue Nile saves this cut for only the rarest of diamonds that also meet many other strict standards (high marks on the other Cs).

A summary of the numbers of diamonds at different price points against cut is reported in this table:

## cut # > $1000 # > $2000 # > $3000 # > $4000 # > $5000  
## 1 Good 60 35 27 22 11  
## 2 Very Good 272 178 136 108 90  
## 3 Ideal 441 250 209 191 178  
## 4 Astor Ideal 17 12 10 9 8

Here is that same table but listed as a percentage of the diamonds in each price column.

## cut % > $1000 % > $2000 % > $3000 % > $4000 % > $5000  
## 1 Good 8 7 7 7 4  
## 2 Very Good 34 37 36 33 31  
## 3 Ideal 56 53 55 58 62  
## 4 Astor Ideal 2 3 3 3 3

As you can see, the percentage of “Ideal” cut diamonds increases and the percentage of the other cuts stays the same or decreases as a whole as the price of the diamonds increases. This tells us that the better the cut, the higher priced the diamond will be. Interestingly, the means and medians tell a different story. Below is a table showing the medians and means of the prices of the different cuts

## Median Price Mean Price  
## Good 1903 9467  
## Very Good 1744 7758  
## Ideal 1354 6489  
## Astor Ideal 2854 5852

This table seems to indicate that as the cut grade gets better, the median and mean price both decrease.

Evaluating the claim that cut is the most important of the 4Cs is a difficult task. We first want to recognize the bias of what was mentioned earlier about this being the one C that jewelers have the most control over. Thus they want you to think that cut is the most important, thereby validating their work. Second, we must evaluate the claim based on their definition of what the cut of the diamond gets you. According to their website, “The characteristics of a well-cut diamond are superior brilliance, fire, And scintillation.” While these terms are defined on their website, we do not have data on these variables of cut. The only data we have is the effect of cut on price. In the claim about carat, the larger the carat the more expensive the diamond relationship was clearly established, making carat the most important C as it relates to price. So the claim that cut is the most important is not supported by the data, as it relates to price.

The claim that the better the cut the more expensive the diamond seems a little easier to evaluate. Although, we established that the higher the diamond cost, the higher the likelihood that the diamond is an “Ideal” cut (a high grade), the overall measures of center decrease as the cut grade increases. This leads us to believe that there are more factors in play when it comes to price that just the grade of cut. This claim is not supported by the data.

This leads us to believe that Blue Niles’ claim that the better the cut the higher the price is not held up by the data.

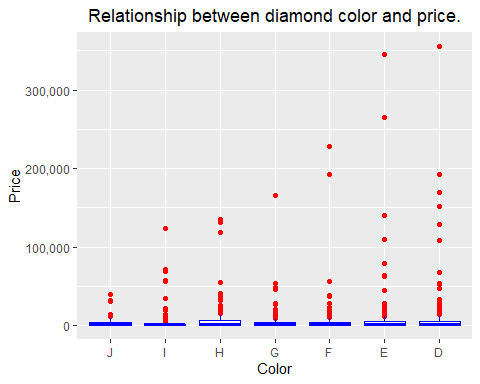
### 2.6.3 Claim regarding color.

**Claim: Color is the second most important of the 4Cs.**

“Colorlessness is a desirable feature in most diamonds and the more colorless diamonds are rarer. The less color a diamond has, the higher the diamond color grade.” Color ranges from “Z” to “D” with a “D” diamond being the most colorless diamond. Blue Nile says they do not carry a diamond with less than a “K” grade making their color scale from “K” to “D”, though our data did not include any “K” grades. We can see the distribution of color here

##   
## J I H G F E D   
## 90 167 148 198 223 181 207

And the impact they have on price here.

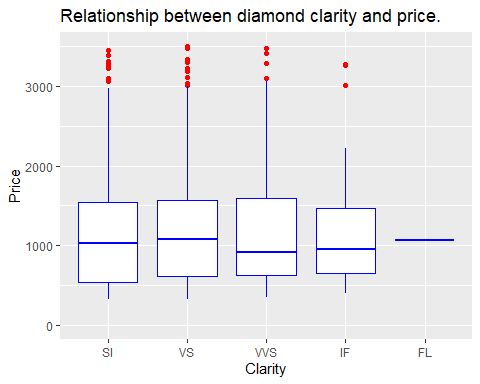


We can clearly see that the higher the color grade (the more colorless the diamond), the higher the price tends to be. The table below shows the median and mean of the price of each of the color grades.

## Median Price Mean Price  
## J 1803 3934  
## I 1212 4779  
## H 1575 7799  
## G 1374 4572  
## F 1485 6205  
## E 1602 9906  
## D 1781 10525

We can see from this table that, in general, the better the color, the higher the price. It is interesting that the “H” color quality has a higher median and mean than the two color grades better than it. The other oddity is that “J” has the highest median, but the lowest mean. While these are interesting notes, they could be due to other circumstances, such as the other Cs, and do not take away from the overall point of the claim. We’ve clearly established that the better the color, the higher the price. Why then, does color take second in importance to cut (as claimed on their website that cut is the most important)? Per Blue Nile’s website, “certain jewelry metals can complement faint color diamonds, so diamond color is a flexible 4C that can vary depending on preferences and jewelry settings.” This means that while color is important, lower grades of color may be hidden or complimented in a way that does not detract from its overall appearance that a decrease in cut grade will. One can argue that the relationship between color and price is better established than cut and price, but the stated flexibility of color (something we have no data to quantify) could account for some of this. So, while important and having a big influence on cost, its flexible nature makes it less important than cut. On the idea that it is second importance, as it relates to price, is a split decision. We will not reject that color is behind cut in importance, but being that we established carat as the biggest importance on price, it would then come third.

### 2.6.4 Claims regarding clarity.

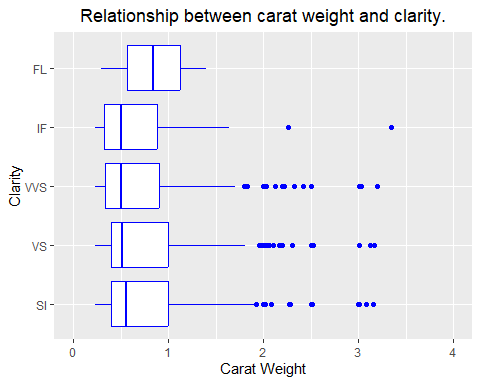


**Claim: When it comes to the four Cs of diamonds, clarity is typically not the most important.**

Per the above box plot, we see similar IQRs and medians for SI and VS. There’s a significant drop when going from “the average” SI or VS diamond to the average VVS or IF diamond, which shows clarity alone does not have much of an impact on a diamond’s price.

**Claim: SI diamonds and VS diamonds are the best value.**

We’ve already shown in Figures 2.5A and 2.5B that for a similar carat weight, a lower cut/color/clarity diamond will typically be less expensive, so if “value” is to mean “more carat weight for your dollar,” then we agree with the claim.

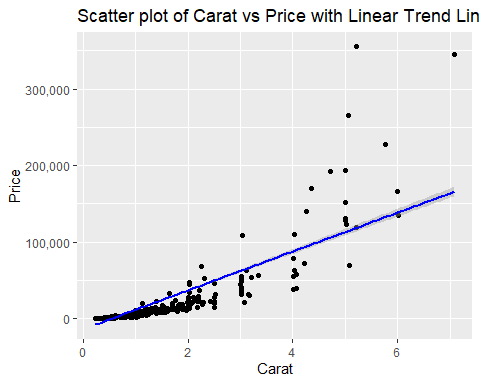


**Claim: Inclusions become more visible as diamond size increases. Conversely: the smaller a diamond, the more difficult it typically is to see any inclusions or blemishes.**

We see higher carat weights correlate with VVS, VS, and especially SI clarities. IQR and total ranges go down the higher you go in grade. Flawless are so rare that we aren’t too concerned with them even though they break the pattern a bit.

# Section 3: Simple linear regression.

## 3.1 Fitting the Regression of Price Against Carat.

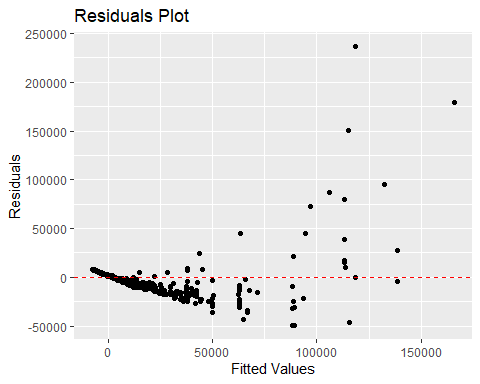


The linear regression model was fitted to explore the relationship between the price of diamonds (response variable) and their carat weight (predictor variable). The scatter plot above displays the data, with carat as the predictor on the x-axis and price as the response on the y-axis. A linear trend line was added to visualize the fitted model.

The scatter plot suggests a positive relationship between carat and price; however, there appears to be some non-linearity, especially for higher carat values where the price increases steeply. Given this pattern, applying a transformation, such as a log transformation to either variable, could potentially improve the linearity and help stabilize the variance of residuals. These transformations will be considered and evaluated later for improving the model fit.

Next steps include evaluating assumptions of linear regression, considering potential transformations, and refining the model accordingly.

## 3.2 Checking the Assumptions of Simple Linear Regression.



### 3.2.1 Mean of Residuals is Zero.

The assumption that the mean of residuals should be zero is not fully satisfied here. The residuals plot shows a pattern where residuals tend to be negative at lower fitted values and positive at higher fitted values. This pattern suggests non-linearity in the relationship between the predictor (carat) and the response (price). To address this issue, a transformation of either the predictor variable, the response variable, or both could help meet the assumption.

### 3.2.2 Constant Variance of Errors.

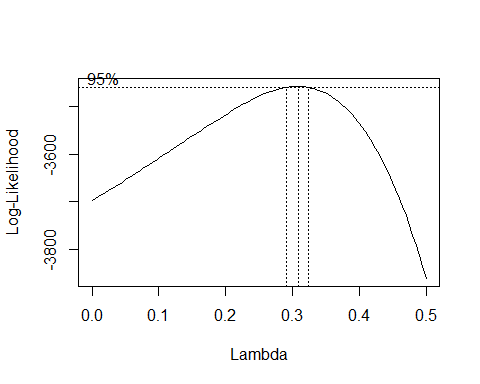
The residuals plot also reveals an increasing spread as the fitted values increase, indicating that the second assumption of constant variance of errors is not met. Transforming the response variable may help stabilize the variance and improve the model’s fit.

### 3.2.3 Addressing the Issues.

When both assumptions are not met, the standard approach is to first transform the response variable to stabilize the variance. If the transformation resolves the constant variance of errors, then we recheck whether the mean of residuals is approximately zero. If Assumption 1 is still violated, then we may consider transforming the predictor variable.

The next step will involve using diagnostic tools such as the Box-Cox plot to guide the selection of an appropriate transformation for the response variable in order to stabilize the variance and improve the model fit.

## 3.3 Transforming the Response Variable.



The Box-Cox plot was used to guide the selection of an appropriate transformation for the response variable to address the non-constant variance in the residuals. The plot showed a peak near a lambda value of 0.3, indicating that a power transformation around this value could help stabilize the variance. We considered two potential transformations:

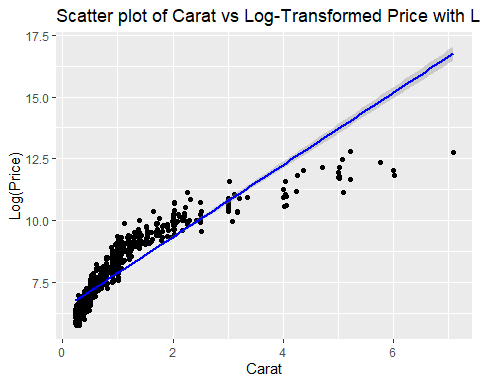
* setting (with lambda = 0.3)
* and (with lambda = 1/3).

Ultimately, we chose a log transformation () for the response variable. This decision was made despite the slight difference from the optimal lambda suggested by the Box-Cox plot, as the practical advantage of interpreting the slope coefficient in the simple linear regression model outweighed the potential error effects. The log transformation allows for a more straightforward interpretation in terms of percentage changes, which is useful in understanding the relationship between carat and price.

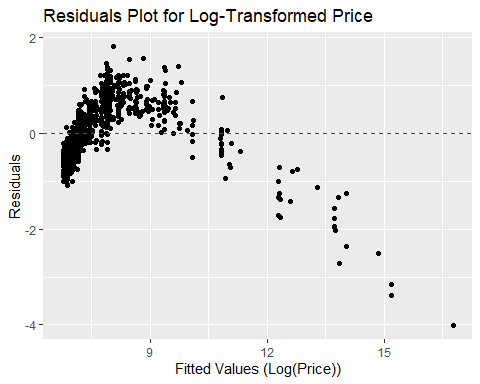
The next step involves fitting the model with the log-transformed response variable and rechecking the assumptions to see if the transformation has adequately addressed the non-constant variance in the residuals. If necessary, transforming the predictor variable (carat) will be considered.

## 3.4 Assessing the Model with Log-Transformed Response Variable.

After applying the log transformation to the response variable (price), we refitted the simple linear regression model to assess if the transformation addressed the issues of non-constant variance in the residuals.



### 3.4.1 Analysis of the Residuals Plot.



The scatter plot of carat vs. log-transformed price, with a fitted linear trend line, shows an improved linear relationship compared to the original fitted model. The residuals plot for the log-transformed model, we observe that the residuals are not evenly scattered across the horizontal axis: at very low and high fitted values, the residuals tend to be negative, and at moderate fitted values, residuals tend to be positive.

Although the vertical variation is fairly consistent in showing a banded appearance in the residuals, there remains more vertical variation at higher fitted values, indicating that some non-constant variance persists. However, we felt confident in this transformation in better addressing assumption 2 than the original fitted model.

### 3.4.2 Addressing Assumption 1.

The next step is to consider a transformation of the predictor variable (carat) to improve the linearity of the relationship and address assumption 1 that the mean of the residuals should be zero.

## 3.5 Transforming the Predictor Variable.

To address the violations of the first assumption, we decided to apply a log transformation to the predictor variable (carat). The choice of transformation was informed by the pattern observed in the scatter plot of the log-transformed response variable () against the original predictor variable (carat). By consulting the scatter plot and comparing it with the reference figure, we identified that the pattern most closely resembled one associated with a log transformation of the predictor.

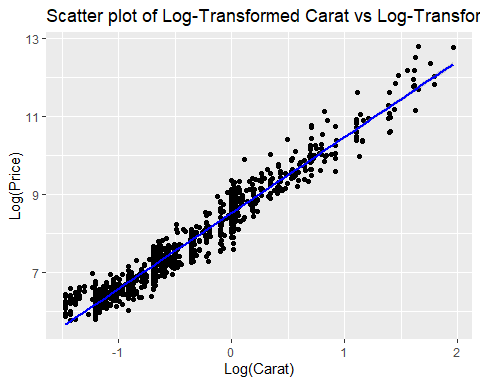
|  |
| --- |
| Transformations for x |

Transformations for x

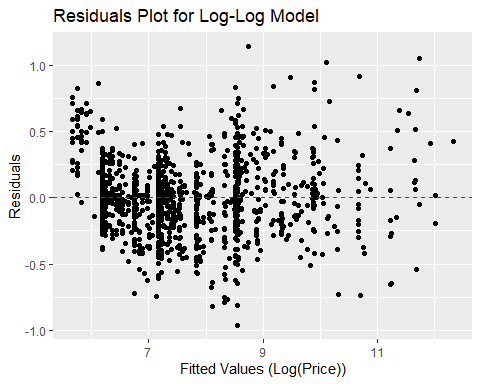
### 3.5.1 Fitting the Model with Log-Transformed Predictor and Response.

After applying the log transformation to both the response variable (price) and the predictor variable (carat), the new model was fitted.

The scatter plot of the log-transformed response against the log-transformed predictor suggests a strong linear relationship, as indicated by the clear linear trend in the plot.



### 3.5.1 Residuals Analysis.



The residuals plot for the model with both variables log-transformed shows no apparent patterns, and the residuals are evenly scattered around zero. This indicates that the assumptions of simple linear regression are now met:

1. **Mean of Residuals is Zero**: The residuals plot confirms that the residuals are centered around zero across all fitted values, meeting this assumption.
2. **Constant Variance of Errors**: The residuals exhibit consistent vertical spread across the range of fitted values, indicating that the variance of the residuals is stable.

With both assumptions satisfied, the log-log model provides a suitable fit for describing the relationship between carat and price.

## 3.6 Log-Log Model.

##   
## 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

### 3.6.1 Model Equation.

The estimated regression equation is:

where:

* represents the predicted log-transformed price.
* represents the log-transformed carat value.

which can be written as:

where:

* The intercept is .
* The slope coefficient for is .

### 3.6.2 Interpretation of the Model Output.

* **Coefficients:**
  + The slope () is highly significant with a -value less than , indicating a very strong relationship between the log-transformed predictor and response variables.
* **and Adjusted** : The model explains approximately of the variability in the log-transformed prices, indicating a strong fit.
* **F-Statistic:** The extremely high F-statistic () with a very small -value confirms that the model is statistically significant and that log(carat) is a significant predictor of log(price).

Overall, the log-log model fits the data well, with a strong linear relationship between the log-transformed variables and both regression assumptions met.

### 3.6.3 Interpretation of the Slope.

The slope coefficient () indicates that for a 1% increase in carat, the predicted price increases by approximately 1.944%.