Project 1: Examining Blue Nile’s Claims

Addison Nitto, Arda Unal, Christopher Lee, Sae’von Palmer

2025-03-23

Section 1

Blue Nile Inc., an online jewelry retailer, is known for offering a broad selection of diamonds. On their website, they include a guide designed to help consumers make informed purchases, with detailed information about how a diamond’s price depends upon the four Cs: Cut (quality in which the diamond interacts with light), Color (how clear the diamond is), Clarity (measure of the diamond’s flaws or imperfections), and Carat (weight of the diamond). To find out if these claims actually hold true in practice, researchers Addison Nitto, Arda Unal, Christopher Lee, and Sae’von Palmer analyzed over a thousand diamonds.

In their study, the research team found several discrepancies between Blue Nile’s claims and the actual data. While Blue Nile states that “SI diamonds will have a lower price than VVS diamonds,” Palmer notes that “diamonds of varying clarity grades often have similar median prices with substantial overlap in their price ranges.” Similarly, Blue Nile’s claim that SI or VS clarity grades are “much less expensive” than higher grades isn’t fully supported by the findings. Regarding color, Blue Nile claims “the absence of color in a diamond is the rarest and therefore, the most expensive.” Unal confirms this is partially true, stating that “Diamonds with better color grades (D, E, F) tend to show higher prices compared to lower grades (I, J),” but the research shows the differences aren’t as dramatic as Blue Nile suggests.

Perhaps the most significant finding concerns diamond cut. Lee writes that a “diamond cut’s impact on price is not as strong as carat weight; carat is the strongest predictor of price.” Nitto adds that “while better cuts do tend to have higher median prices than lower quality cuts,” the differences aren’t as significant as Blue Nile’s statement that “Diamond cut is considered the most important of the four Cs” would lead consumers to believe. The team discovered that consumers may achieve better value by prioritizing carat weight while making modest compromises on the other Cs, rather than following Blue Nile’s marketing emphasis on cut as the primary consideration. Their analysis revealed that when a diamond’s carat weight increases by just 1%, the price typically jumps by nearly 2% - demonstrating the substantial premium placed on heavier stones.

Understanding these relationships between diamond characteristics and price can help consumers make more informed purchasing decisions, potentially saving thousands of dollars by focusing on the attributes that truly drive value rather than those emphasized in marketing materials. By knowing that carat weight is the dominant price factor, buyers can make strategic trade-offs among the 4Cs to maximize value while staying within their budget.

Section 2

Variables

The variables in the data set are as follows:

- Carat: Carat is the weight of the diamond, with the measurement being in carats. This is a continuous numeric variable with the Min being 0.23 and the Max being 7.09.
- Price: Price is the cost of the diamond in US dollars. This is also a continuous numeric variable, with the min being 322 and the max being 355403.

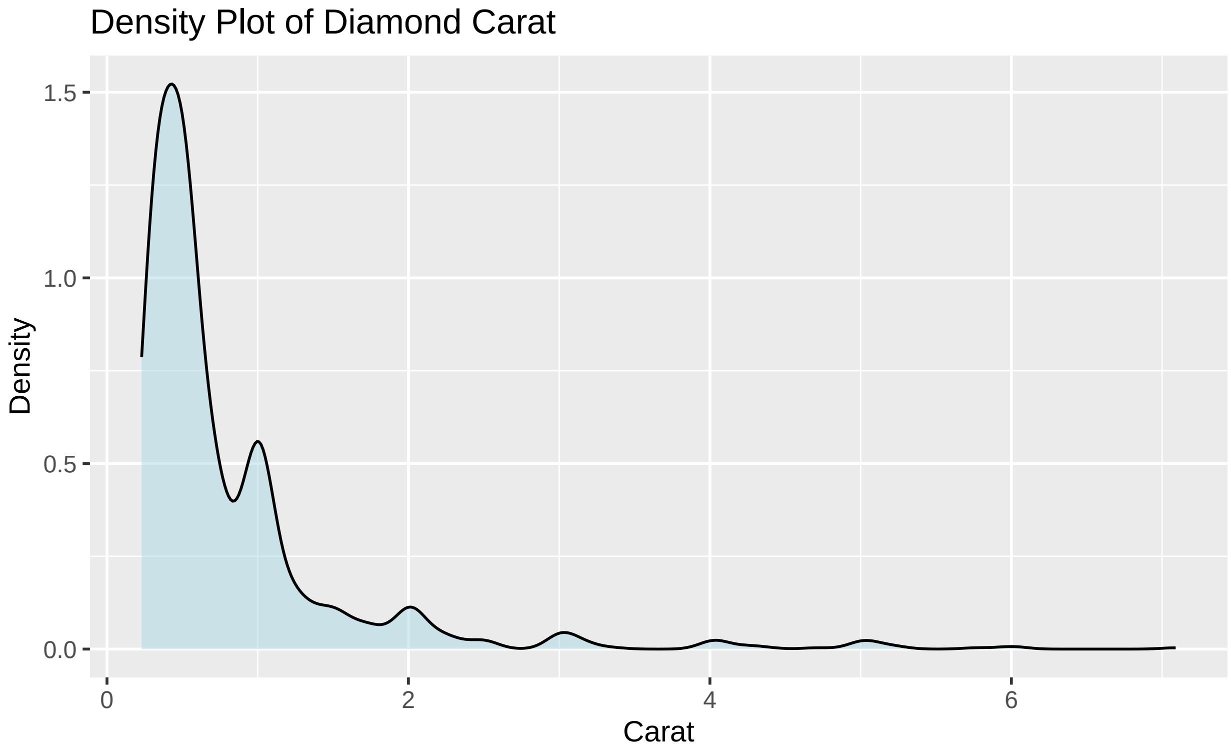
- Cut: Cut is the quality in which the diamond interacts with light. While there are different styles of cuts in this instance, we are measuring the quality of the diamond's cut categorized as Astor Ideal, Ideal, Very Good, and Good. This variable is ordinal.
- Color: Color refers to how clear a diamond is. Color is graded on a scale, ranging from D (colorless) to J (noticeable color). This is an ordinal categorical variable.
- Clarity: Clarity is a measure of the diamond's flaws or imperfections, categorized as FL (Flawless), IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, with FL being the best and SI2/I1 being the worst. This is also an ordinal categorical variable.

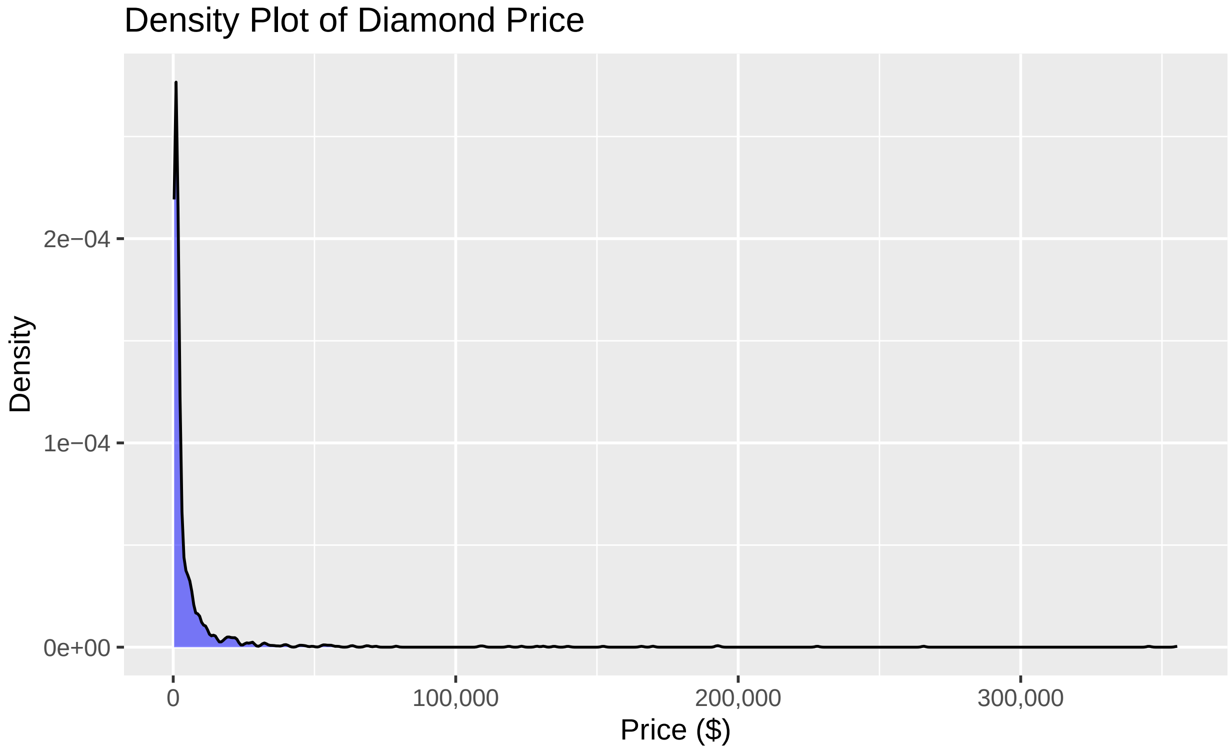
Data Examination

Summary of the data:

```
##      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  
```

Examining the distribution of the variables:





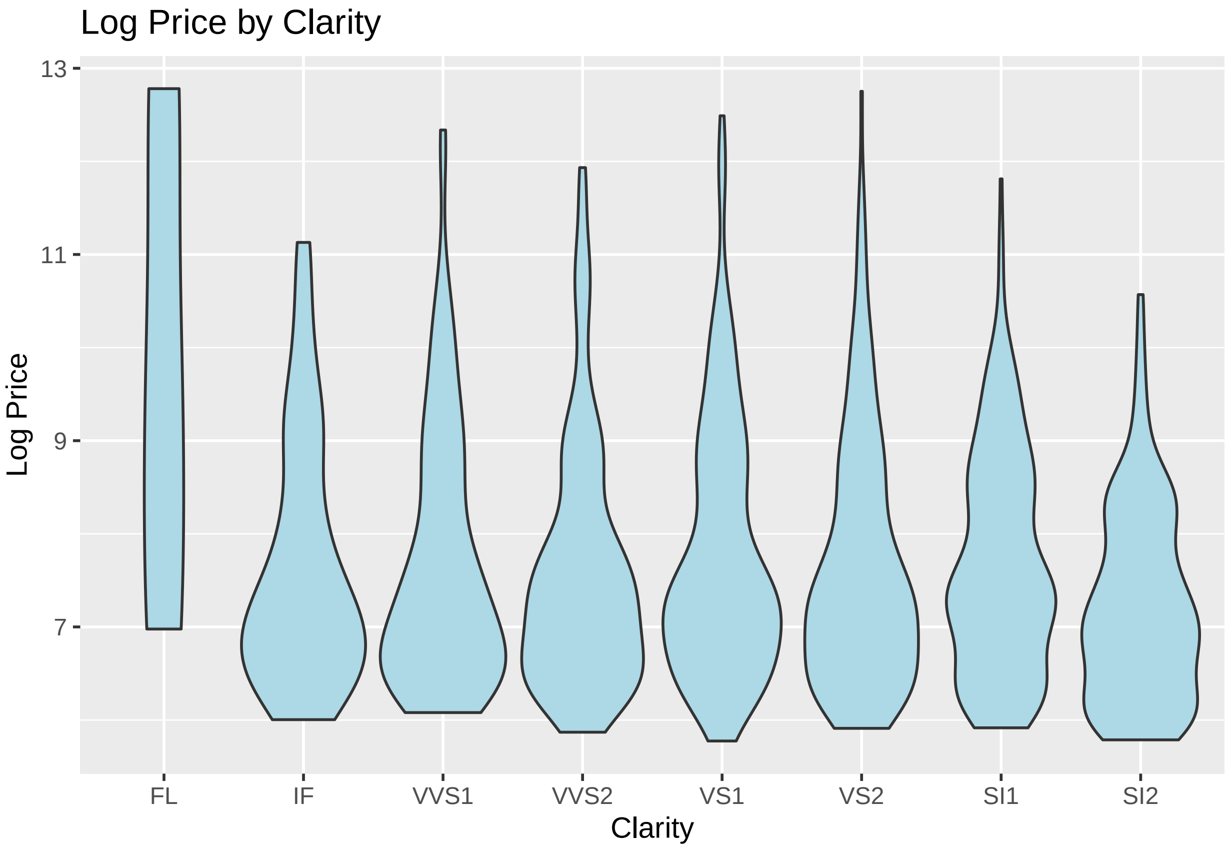
Number of diamonds per category:

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

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

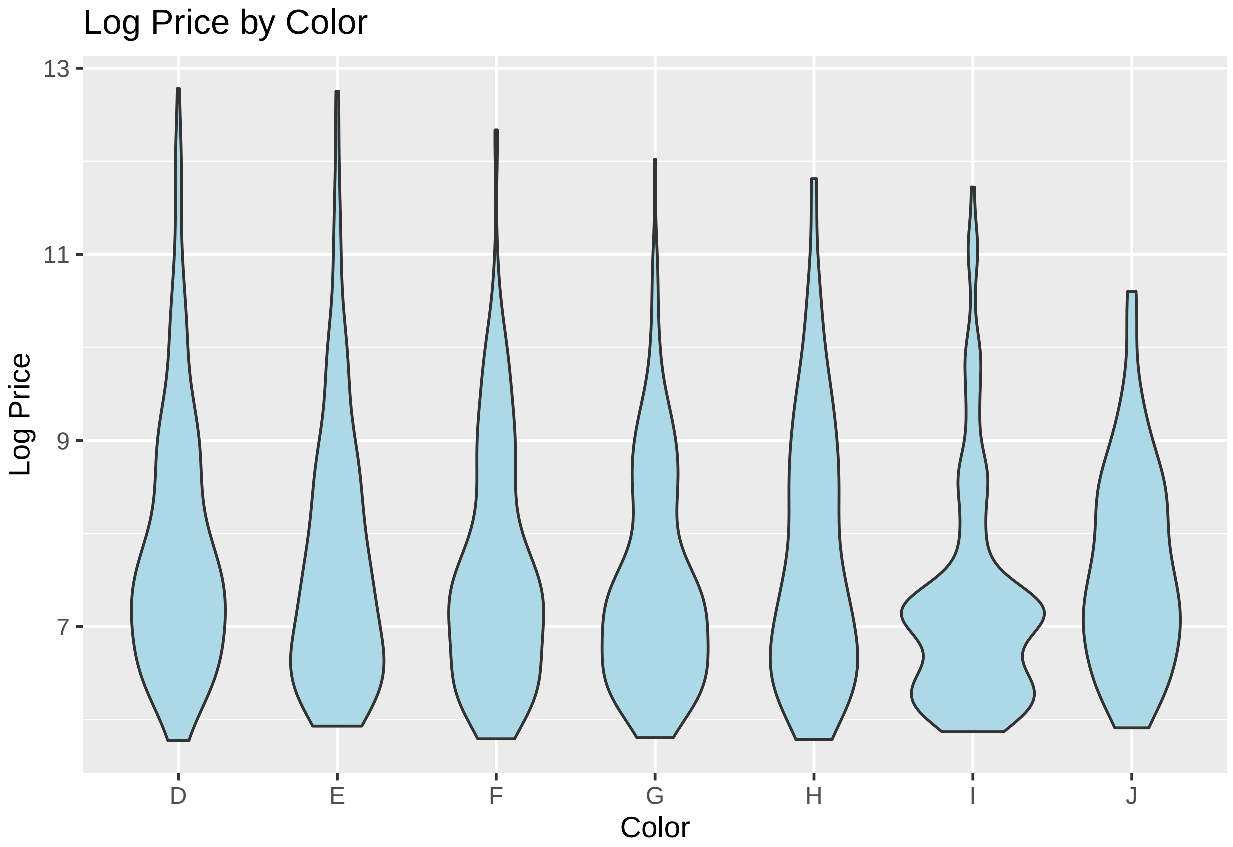
```
##
## Astor Ideal        Good       Ideal   Very Good
##          20          73         739         382
```

Data Visualizations



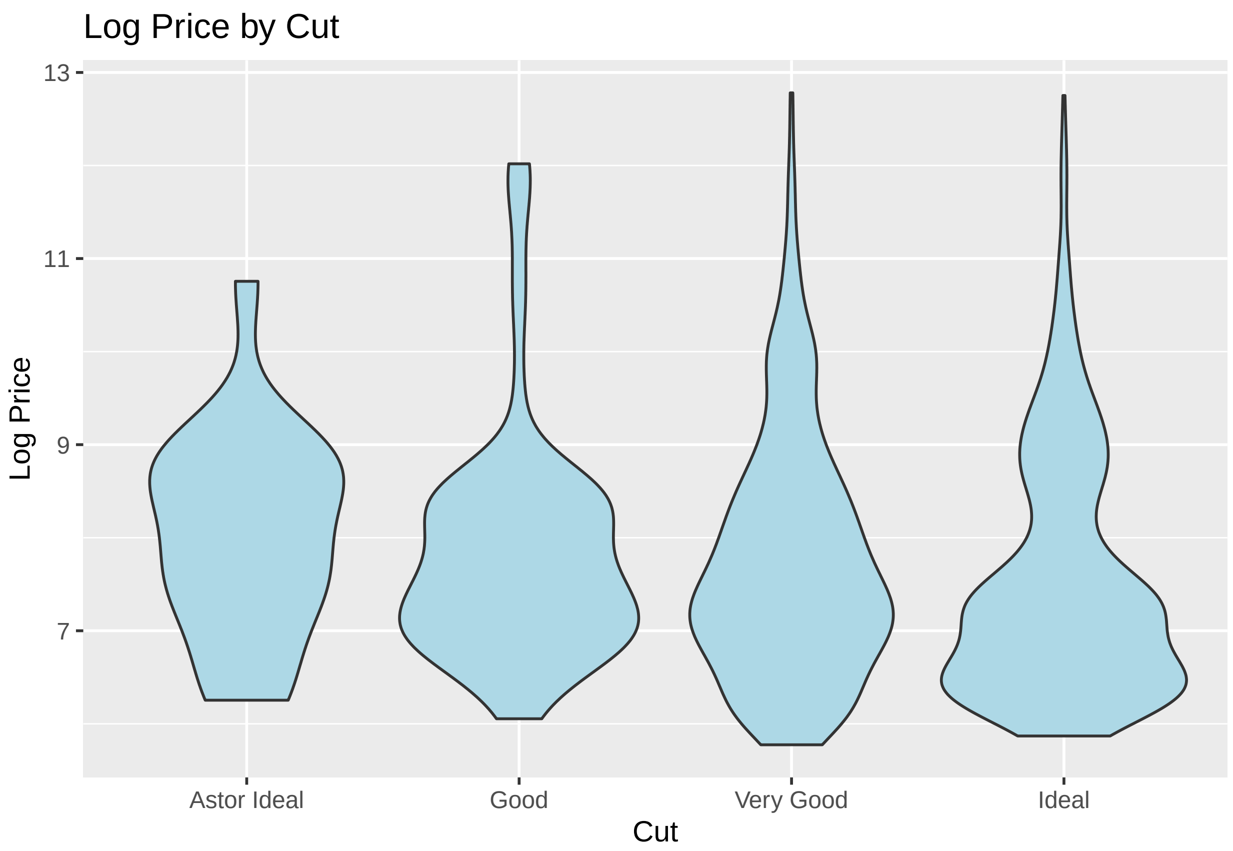
Claims made by Blue Nile:

- "Assuming all the other diamond gradings are the same, a diamond positioned higher on the diamond grading chart will carry higher value than more included diamonds. For example, SI diamonds will have a lower price than VVS diamonds."
  - Response: Aside from FL clarify diamonds, the median price of diamonds based on clarify is roughly the same across clarity types. In this claim, Blue Nile says SI diamonds are lower priced than VVS diamonds, but the violin plot shows the median price of the two categories are roughly the same.
- "For the best value, select a diamond with inclusions that can't be seen through the crown without magnification (also known as eye clean diamonds), like a diamond with an SI or VS clarity grade. These diamonds are much less expensive and look the same as the higher grades, visually."
  - Response: Blue Nile claims SI or VS clarity grades are much less expensive and than higher clarity grades when controlling for the other C's. However, the violin plot shows it's possible for higher clarity diamonds to cost less than SI or VS clarity grade diamonds, thus refuting Blue Nile's claim.



Claim made by Blue Nile:

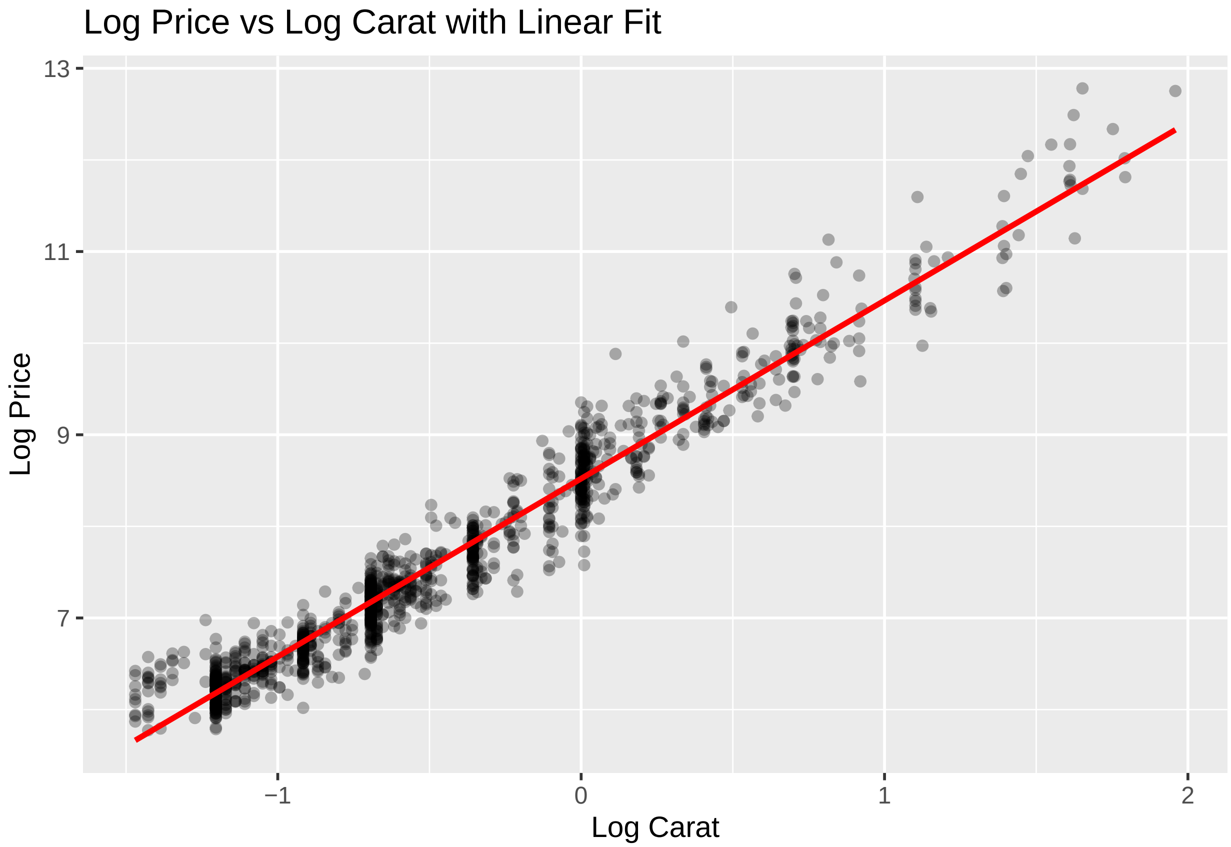
- "The absence of color in a diamond is the rarest and therefore, the most expensive."
  - Response: Diamonds with better color grades (D, E, F) tend to show higher prices compared to lower grades (I, J).



Claims made by Blue Nile:

- "Diamond cut is considered the most important of the four Cs."
- "The Ideal cut diamond, and the super-ideal Astor by Blue Nile™, are the most expensive diamond cuts… "
  - Response: Ideal and Very Good cuts tend to have higher price distributions, while Good cuts have lower median prices. While the violin plot shows there is a not a clear normal distribution, the Blue Nile claim that cut is the most important C is wrong.

Simple Linear Regression of Price Against Carat



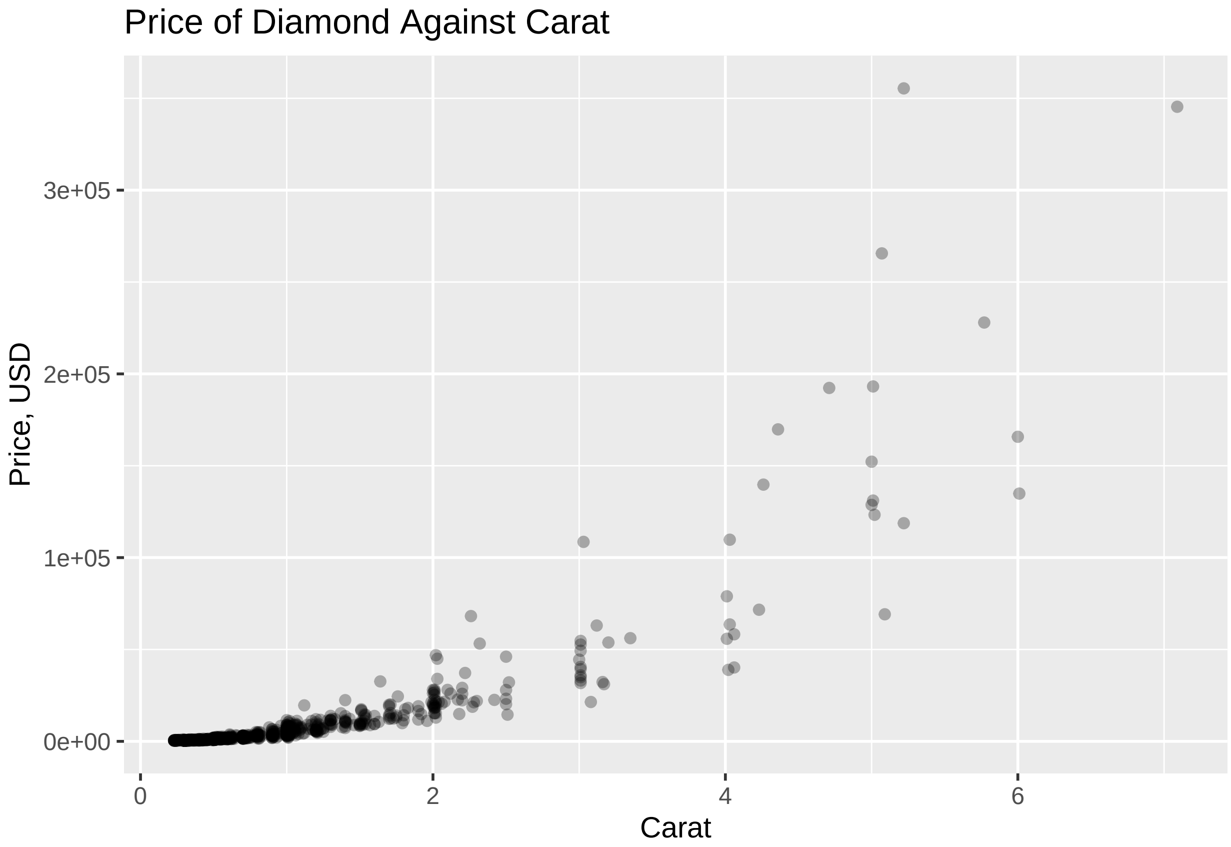
Summary

Blue Nile claims that diamond cut is considered the most important of the four C's. However, diamond cut's impact on price is not as strong as carat weight. Carat is the strongest predictor of price. The violin plot shows overlapping distributions depending on the grade of cuts, especially between "very good" and "good", indicating their impact on price is not as important. In contrast, carat size has a consistent impact on pricing and the scatter plot shows the strong correlation in that. That being said, carat is the most important of the four C's based on the data we have.

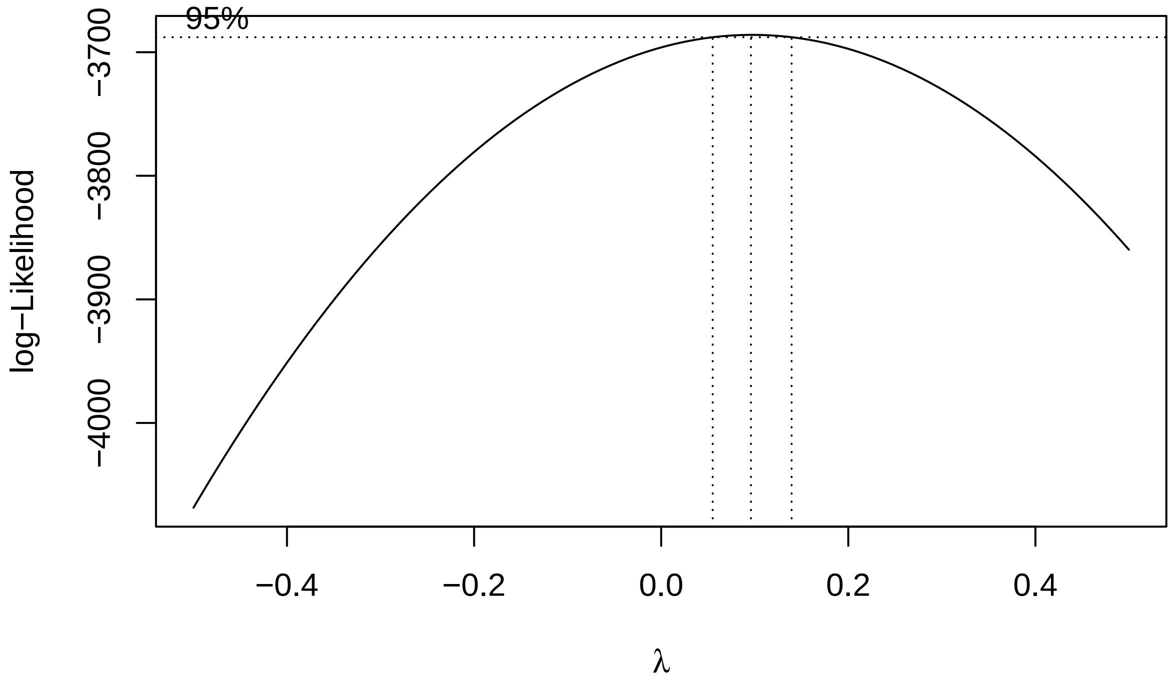
Section 3

Fitting a Linear Regression for Diamond Price Against Carat

First, plot price against carat in a scatter plot to visually check how the data looks.

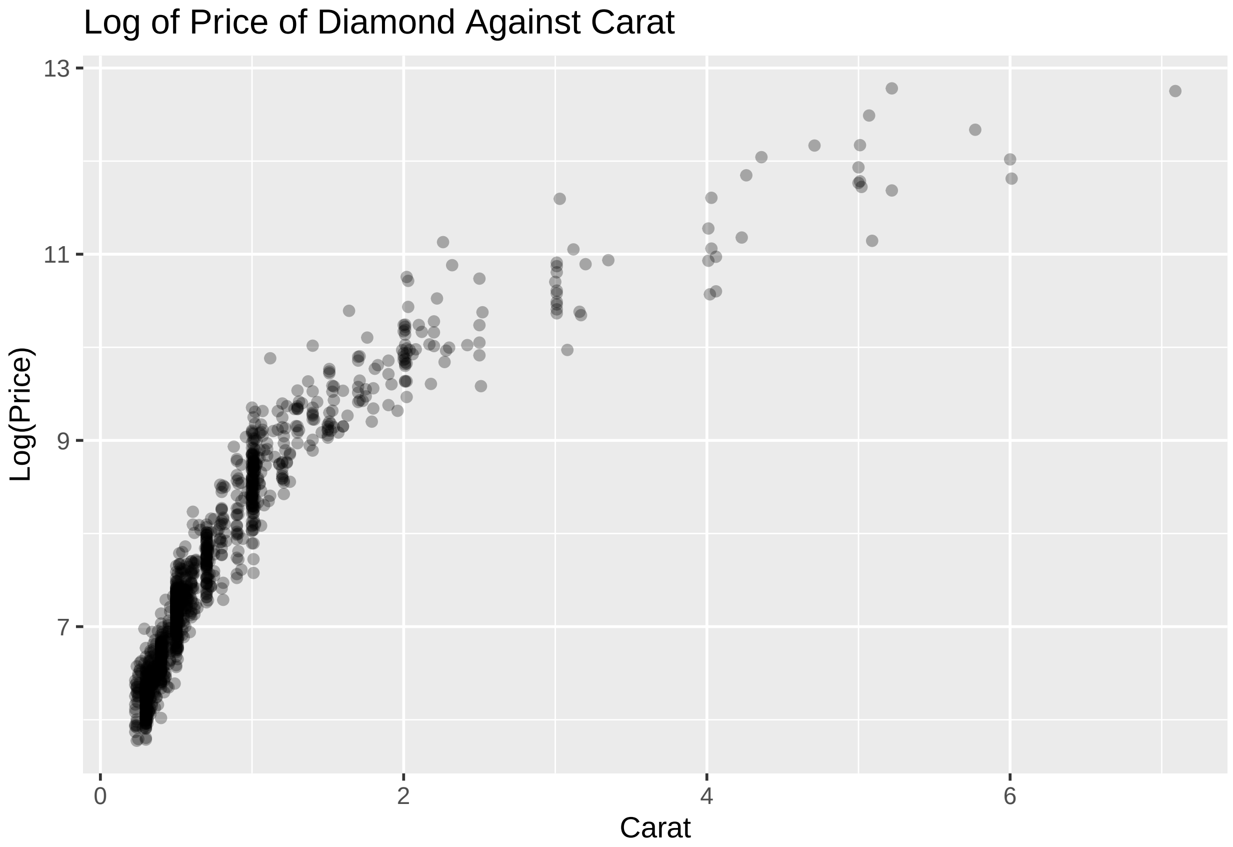


Due to increasing variance, we check the box cox plot to see what λ to select, knowing that λ needs to be < 1.



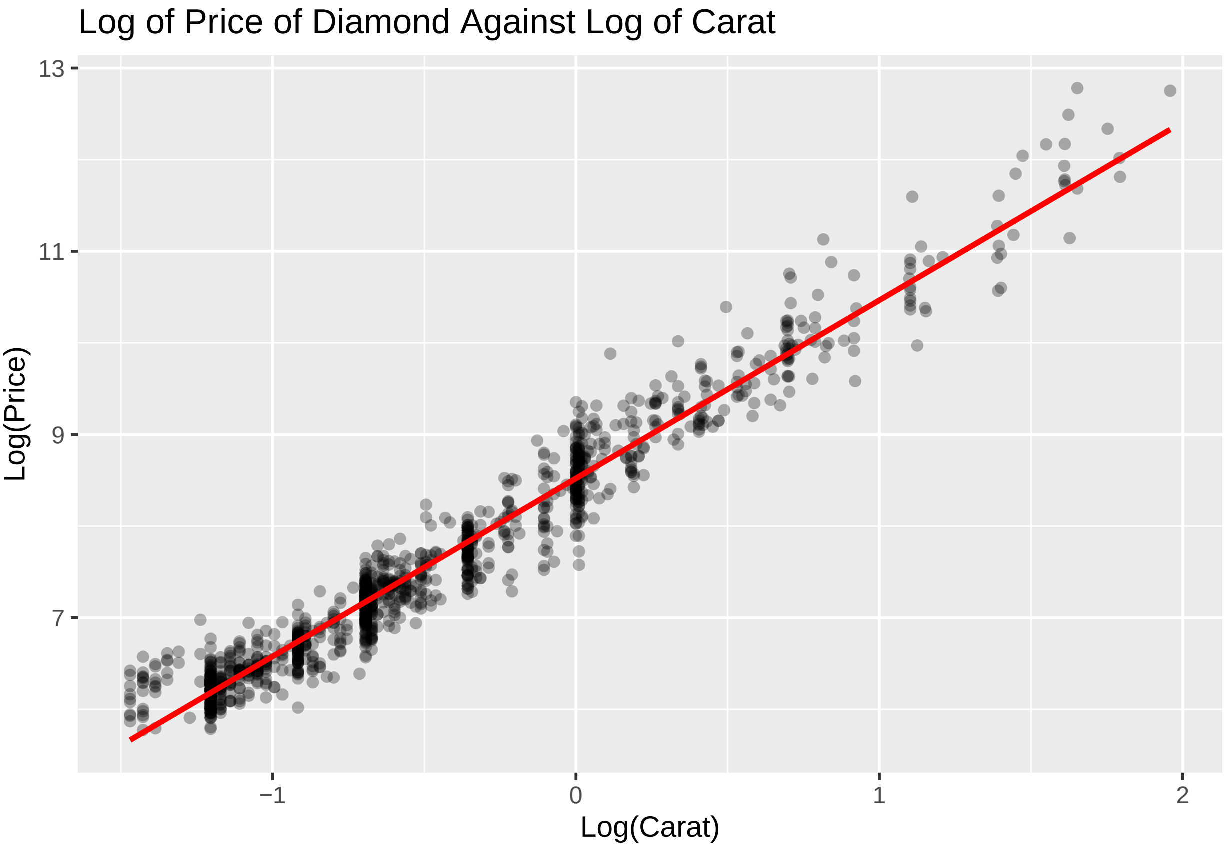
We see that the 95% confidence log-likeliehood for the regression of price against carat is near 0, therefore we can select $\lambda = 0$ and thus use a log transformation on the response variable to meet assumption 2.

We transform the response variable and re-check the regression model via scatter plot.

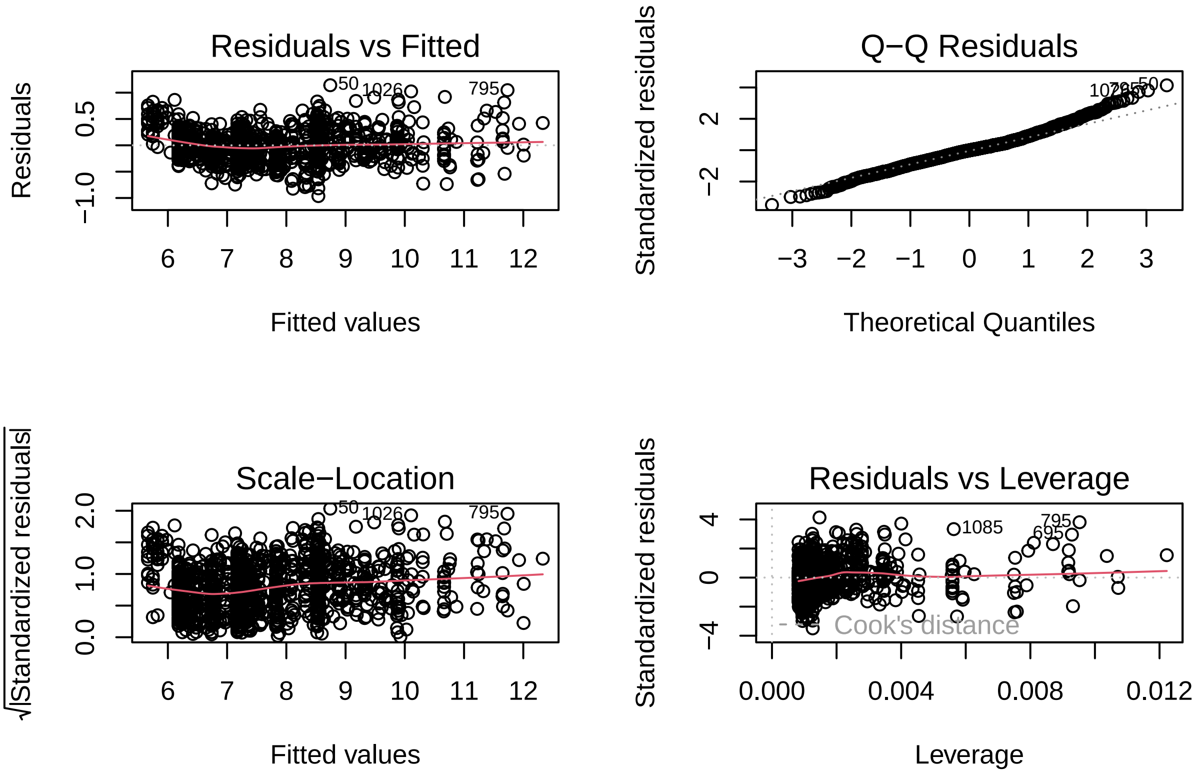


Transforming the response variable with a log transformation fixed the increasing variance the original data showed, satisfying assumption 2.

However, the data does not follow a straight line. Since the data seems to follow a log plot, we select a log transformation for the predictor variable. We then re-plot and re-check the regression assumptions via scatter plot.



We see via the scatter plot that log transforming the predictor variable fixed the curving nature of the previous scatter plot and thus all the regression assumptions are met. We can verify via the residuals plot.



Assessing Simple Linear Regression Assumptions

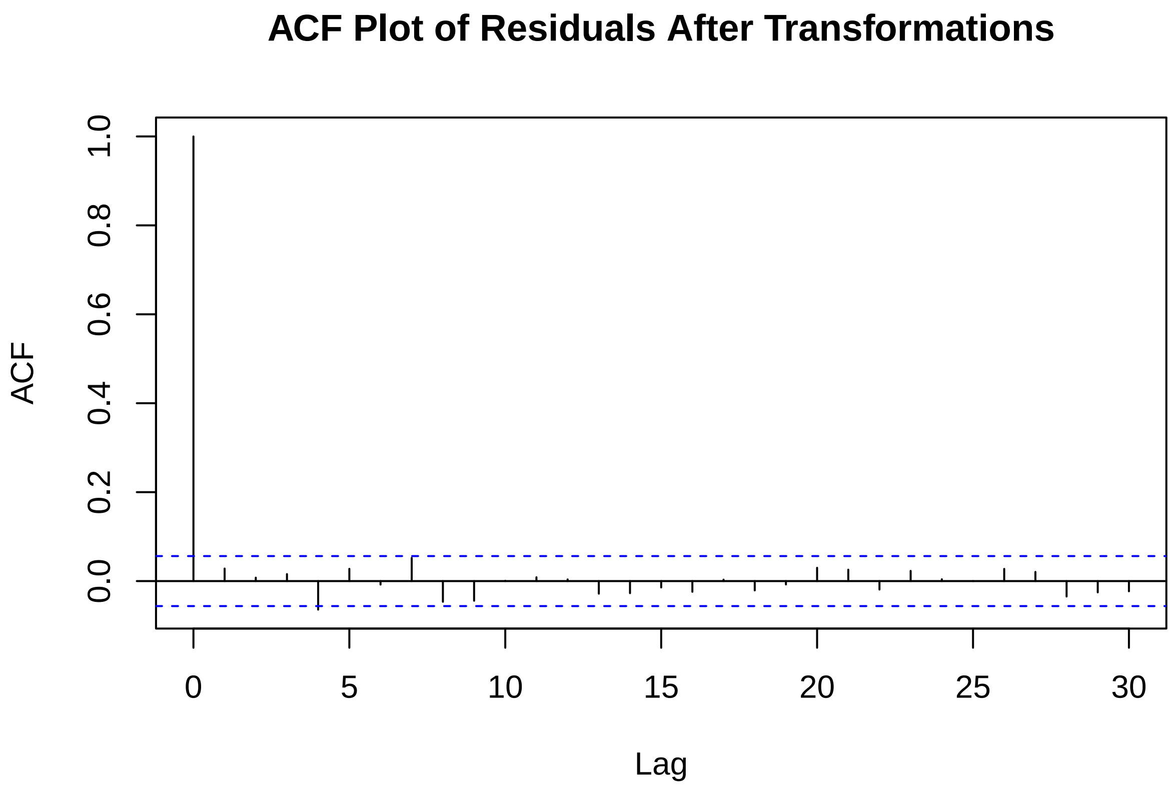
Assumption 1 - Errors have mean 0

- Based on the residuals plot, the mean of the errors (red line) lies fairly in line with the x-axis, indicating that the average mean across the data is equal to 0. The mean of the errors is in a straight line and not in any other shape. Therefore, assumption 1 is met.

Assumption 2 - Errors have constant variance.

- Based on the residuals plot, the variance of the errors from the mean and from the x-axis is fairly even across the plot. There is no increasing or decreasing variance throughout the data. Therefore, assumption 2 is met.

Assumption 3 - Errors are independent



- The data can be checked for independence using an autocorrelation function (ACF) plot, shown above. Generally, the ACF values values are between the critical values (blue dashed lines), indicating no significance in the residuals. Therefore, we can conclude that the data is independent and assumption 3 is met.

Assumption 4 - Errors are independent

- Based on the QQ plot, the residuals generally fall along the diagonal QQ line, indicating the data is generally normally distributed. Therefore, assumption 4 is met.

Regression Analysis

Summary of regression with transformed variables:

```
##
## Call:
## lm(formula = logPrice ~ logCarat, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96394 -0.17231 -0.00252  0.14742  1.14095
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.521208   0.009734   875.4   <2e-16 ***
## logCarat    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
```

The adjusted regression equation is

$$\hat{y^*} = 8.5212 + 1.944x^*$$

where $\hat{y}^* = log(y)$ & $x^* = log(x)$.

Based on the adjusted regression equation, for every 1% increase in carat for a diamond, the price of the diamond is multiplied by $(1.01)^{1.944} = 1.0195$.

Put another way, for every 1% increase in carat for a diamond, the price increases by 1.95%.