Diamond Pricing: Assessing the Impact of the 4C's on Blue Nile Diamonds

Group 8: Jacob Kuchta, Emmanuel Leonce, Bardia Nikpour, Victor Ontiveros

2024-03-24

Contents

Report Summary

High-Level Overview

Our study aims to uncover what makes diamonds valuable, focusing on four main factors known as the "4C's": cut, color, clarity, and carat. Imagine each diamond as a unique blend of these qualities, like a fingerprint.

Firstly, the cut of a diamond refers to how well it's shaped and polished to reflect light, giving it that beautiful sparkle we all love. Next, color is about how clear or colorless a diamond appears, with less color often making it more valuable.

Then there's clarity, which is all about the tiny imperfections inside and outside the diamond. A higher clarity grade means fewer imperfections, making the diamond more valuable. And lastly, carat measures the weight of the diamond, with heavier diamonds usually costing more.

To explore how these factors affect the price of a diamond, we're analyzing data from Blue Nile, a popular jewelry retailer. We're looking at over 1,214 diamonds they have for sale, examining how their cut, color, clarity, and carat relate to their prices.

Our Findings

Our research confirms that a diamond's cut quality significantly influences its price. Diamonds with superior cuts, such as Astor Ideal cuts, tend to be priced higher due to their enhanced brilliance and sparkle. When purchasing a diamond, consider the cut quality to ensure maximum beauty.

Contrary to common belief, the color of a diamond by itself doesn't greatly impact its price. This means you have more flexibility in choosing a diamond with a hint of color, which can be equally stunning and more affordable.

While flawless diamonds command high prices, you don't need to prioritize the highest clarity grade. Diamonds with slightly lower clarity ratings can still be beautiful and offer better value for your money.

Finally, the weight of a diamond, measured in carats, is a significant determinant of its price. Even a small increase in carat weight can lead to a noticeable price hike. Our study found that for a 1% increase in carat weight, the price of a diamond increases by approximately 2%. So, when shopping for diamonds, consider your budget and priorities in selecting the right carat weight.

Using easy-to-understand summaries, charts, and graphs, we'll break down the data to see if the traditional ideas about diamond value hold true. Our goal is to help everyone, from diamond enthusiasts to industry experts, gain a clearer understanding of what makes each diamond unique and valuable.

Exploring the Data Set and Variables

The Data Set

We are utilizing a data set of 1,214 diamonds that are for sale at http://bluenile.com. Our data provides information on the 4C's of a diamond as well as the price in which each diamond is being sold for. The 4C's of a diamond may be used as indicators or predictors of a diamond's price, and we will be deep diving into the relationships that these characteristics have with a diamond's price within the remainder of this report.

Variable Descriptions

The 4C's represent:

1) Cut: A diamond's cut is a characteristic representing how well-proportioned the dimensions of a diamond are, and how these surfaces, or facets, are positioned to create sparkle and brilliance. The scale of a diamond cut quality ascends from Good to Very Good to Ideal to Astor Ideal. It is important to understand that the cut of the diamond is not synonymous with the shape of the diamond. Rather, the cut assesses light performance of the diamond, whereas shape is related to the outline of the diamond. Cut is said to be the most influential characteristic on the long-term value of a diamond.

2) Color: A diamond's color characteristic refers to how colorless a diamond is. This color, or lack thereof, is always more visible in bigger diamonds. There are techniques in which shape the shape of a diamond can hide color better than others, but relatively speaking, the scale of diamond color quality ascends from Faint Color diamonds (K color) to Near-colorless diamonds (G, H, I, J colors) to Colorless diamonds (D, E, F colors).

3) Clarity: A diamond's clarity is a characteristic representing the assessment of small imperfections on the surface and within the stone. These surface flaws are referred to as blemishes, while the internal defects referred to as inclusions. Because of the fact that a diamond's clarity is nearly impossible to detect with the naked eye, some consumers may view a diamond's beauty as not being affected by a lower clarity grade. The five factors that influence a diamond's clarity are: Size - If there is a large or very noticeable blemish or inclusion, the diamond's clarity grade will suffer. Number - This is the number of easily seen blemishes or inclusions. Simply put, having the least number of these will lead to

a higher clarity grade. Position - Interestingly, the location of a blemish or inclusion within a diamond's anatomy will affect its clarity rating. If it is closer to the more visible parts of a diamond (under the table or close to a pavilion), this position will turn its inclusions into reflectors, further amplifying the affect of the inclusion. Nature - This characteristic refers to the nature of the inclusion and how that affects the diamond's durability. Color and relief - The color and relief rating refers to how easily a blemish or inclusion is seen. It is a measure of the contrast between the characteristic and surrounding diamond. The scale of a diamond's clarity ascends from Included Diamonds (I1, I2, I3), which exhibit obvious flaws, to Slightly Induced (SI) diamonds (SI1, SI2) to Very Slightly Induced (VS) diamonds (VS1, VS2) to Very, Very Slightly Induced (VVS) diamonds (VVS1, VVS2) to Internally Flawless (IF) diamonds to Flawless (FL) diamonds, which represent less than 1% of diamonds and exhbiit no flaws whatsoever.

4) Carat: Potentially the most commonly known characteristic of a diamond, the carat is an objective measurement of the diamond's weight. Carat weight is often mistakenly associated as solely the size of a diamond. But in fact, a high carat weight diamond with a poor cut may appear a smaller diamond than one with a smaller carat weight and a very good cut. Typically, due to the nature of how the average consumer heavily weights the carat weight of a diamond when thinking of quality assessment, the carat weight can be the most influential of the 4C's.

```
##      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
```

This summary table helps us to identify that:

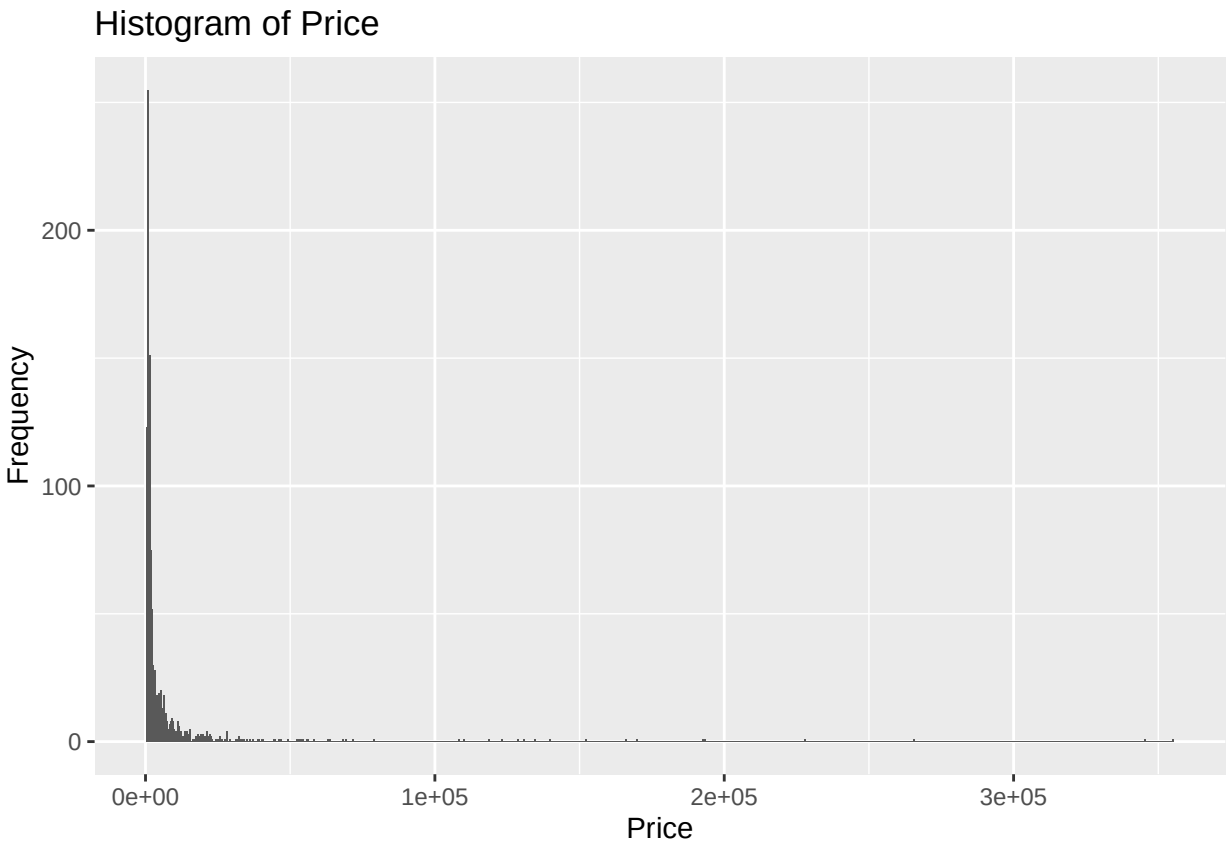
1) The minimum carat weight of a diamond in the data set is 0.23, while the maximum is 7.09. The median carat weight is 0.52, indicating that half of the diamonds have a carat weight below this value and half have a weight above it.

2) Clarity, Color, and Cut are categorical variables.

3) The minimum price of a diamond in the data set is $322, while the maximum is $355,403. The median price is $1,463.50, indicating that half of the diamonds have a price below. this value and half have a price above it.

The mean price is approximately $7,056.70, which is higher than the median, indicating a right-skewed distribution with some high-priced diamonds pulling the mean upwards.

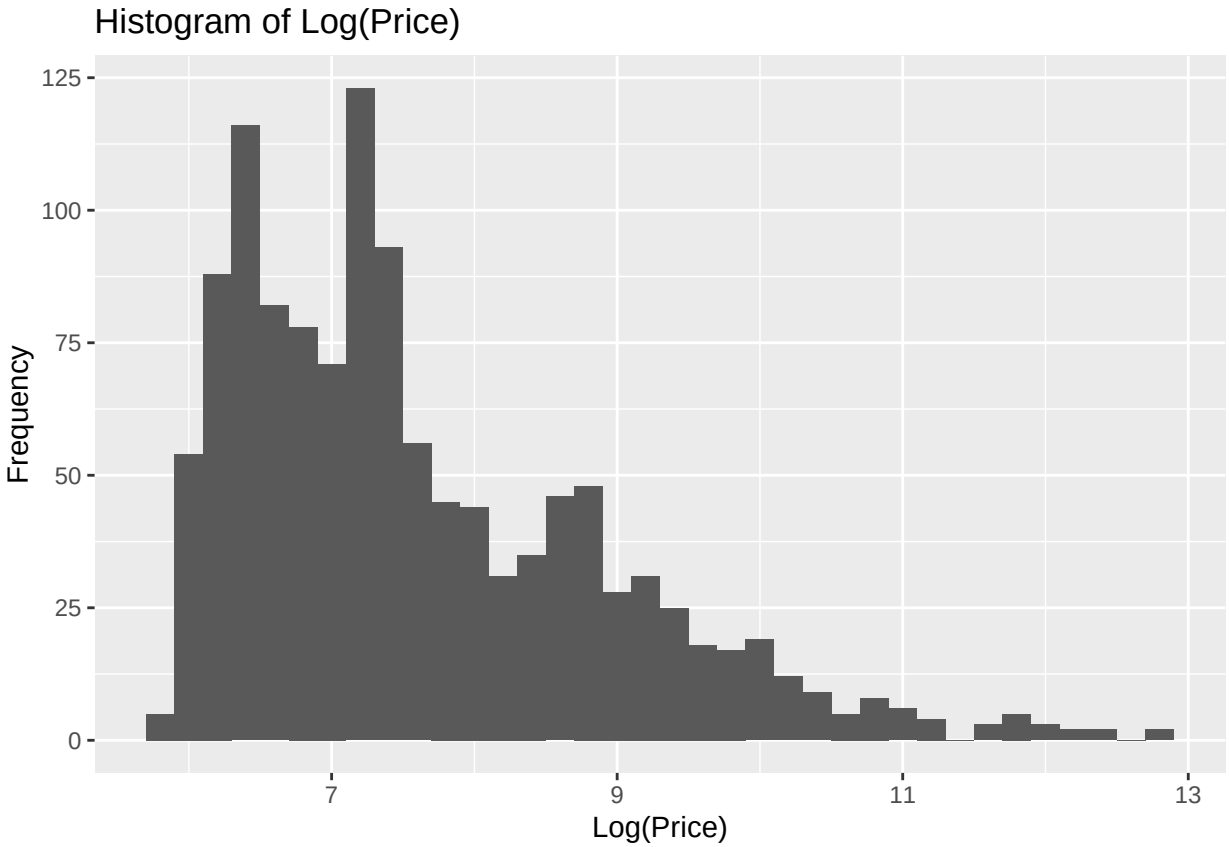
Price



The price histogram shows asymmetrical distribution (right-skewed).

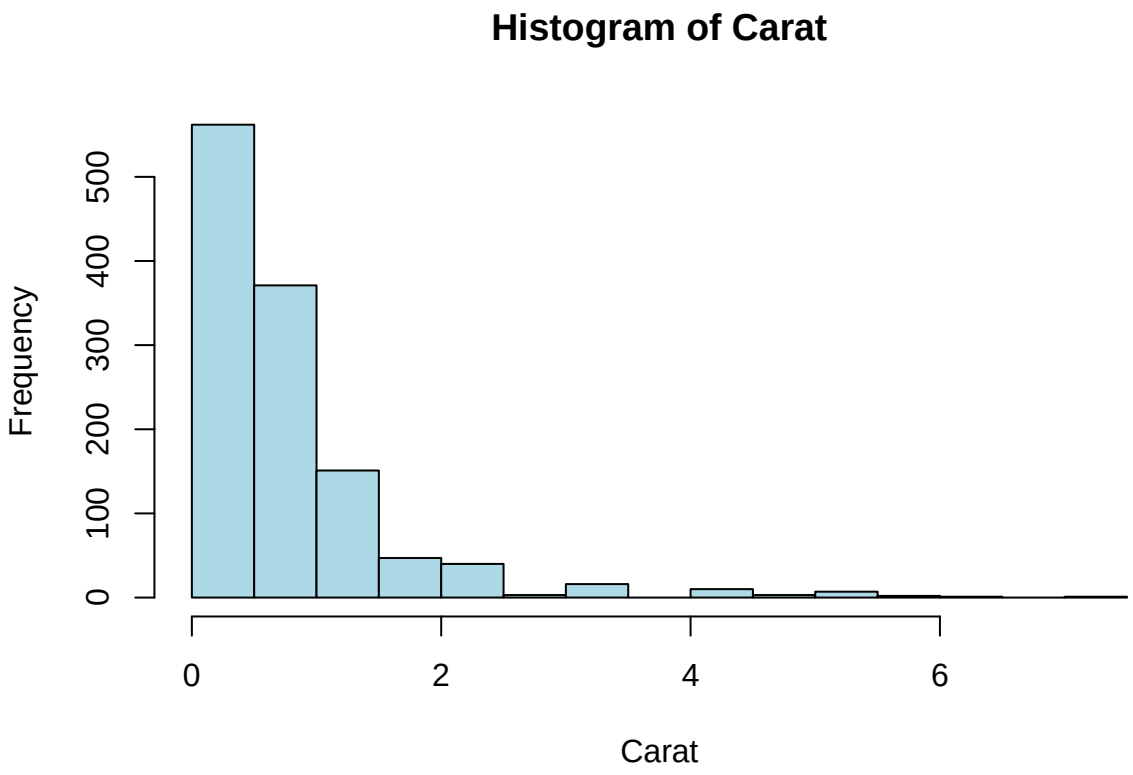
```
## [1] "Skewness of Price: 8.76328219723659"
```

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. In simpler terms, it quantifies the degree to which the data distribution deviates from symmetry. A skewness value of 8.76 indicates that the distribution of the price variable is highly skewed. In this case, it's typically recommended to apply a transformation to make the distribution more symmetric. We will perform a log transformation to do so.



We observe that our price variable now displays a more symmetrical distribution, yet it is still right-skewed. This means indicates that the majority of diamonds in our data set are less expensive, and we have fewer expensive ones.

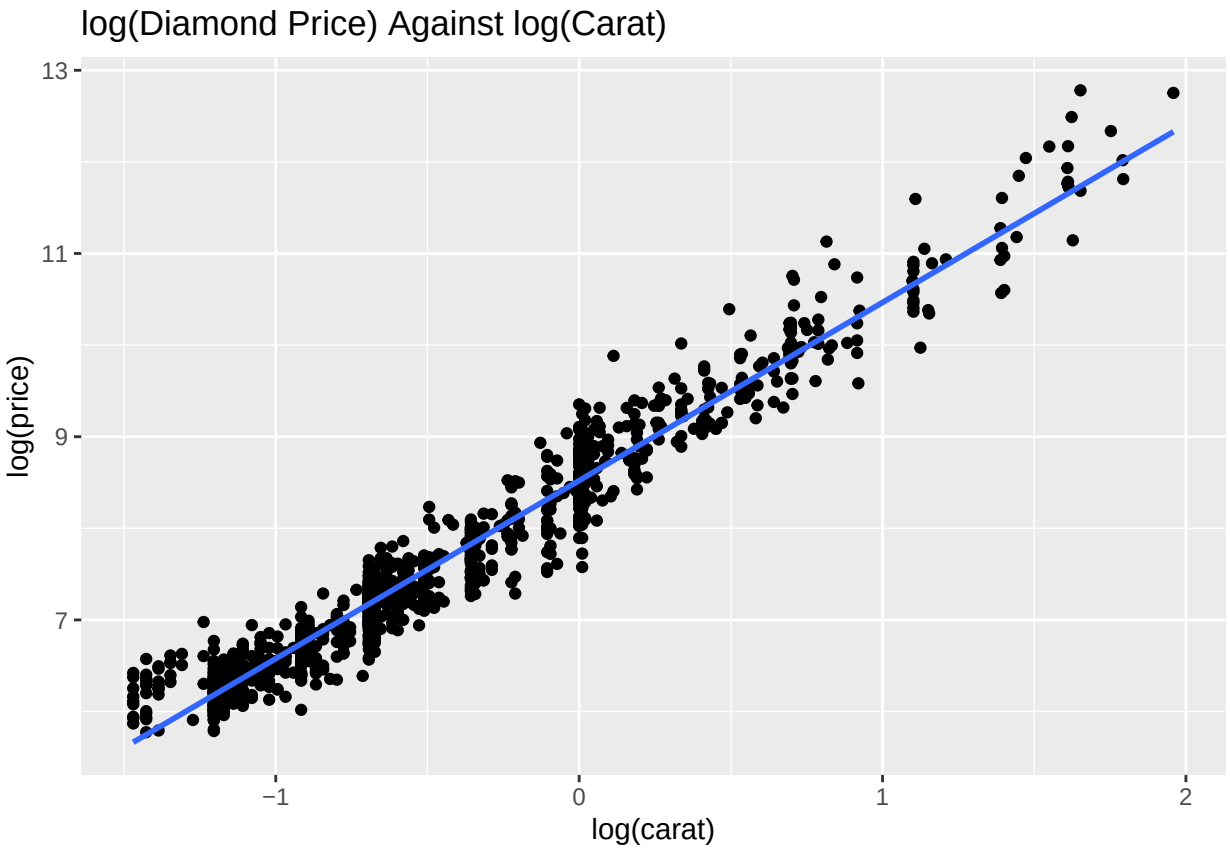
Carat



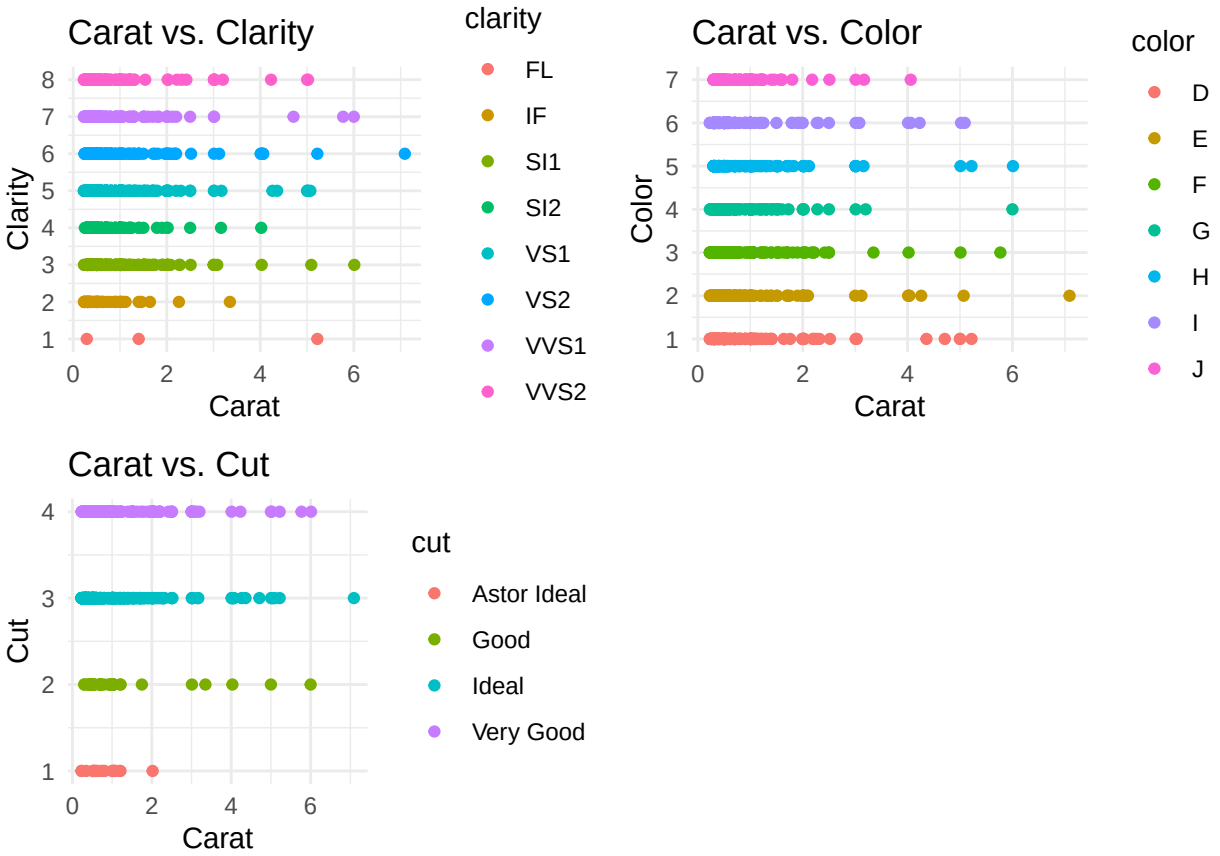
We observe that the distribution of carat weights of diamonds within our data set is right-skewed. This indicates that the majority of diamonds in our data set have lower carat weights, and there are fewer diamonds with higher carat weights. Notice, this distribution has a fairly similar shape to that of price, therefore, we suspect there may be a strong relationship between the carat and price variables.

To best assess this relationship of the quantitative variable, carat, on price, we will perform a log transformation of both the carat and price variables. This is further explained in the Simple Linear Regression of Price Against Carat section of this report.

```
## `geom_smooth()` using formula = 'y ~ x'
```



We see that there is a very strong positive linear relationship between carat and price. This indicates that as the carat weight of a diamond increases, so does the price of the diamond. This supports the claim that movies, mass media, and advertising have created a scenario in which people may put an undue emphasis on carat in regard to a diamond's quality.

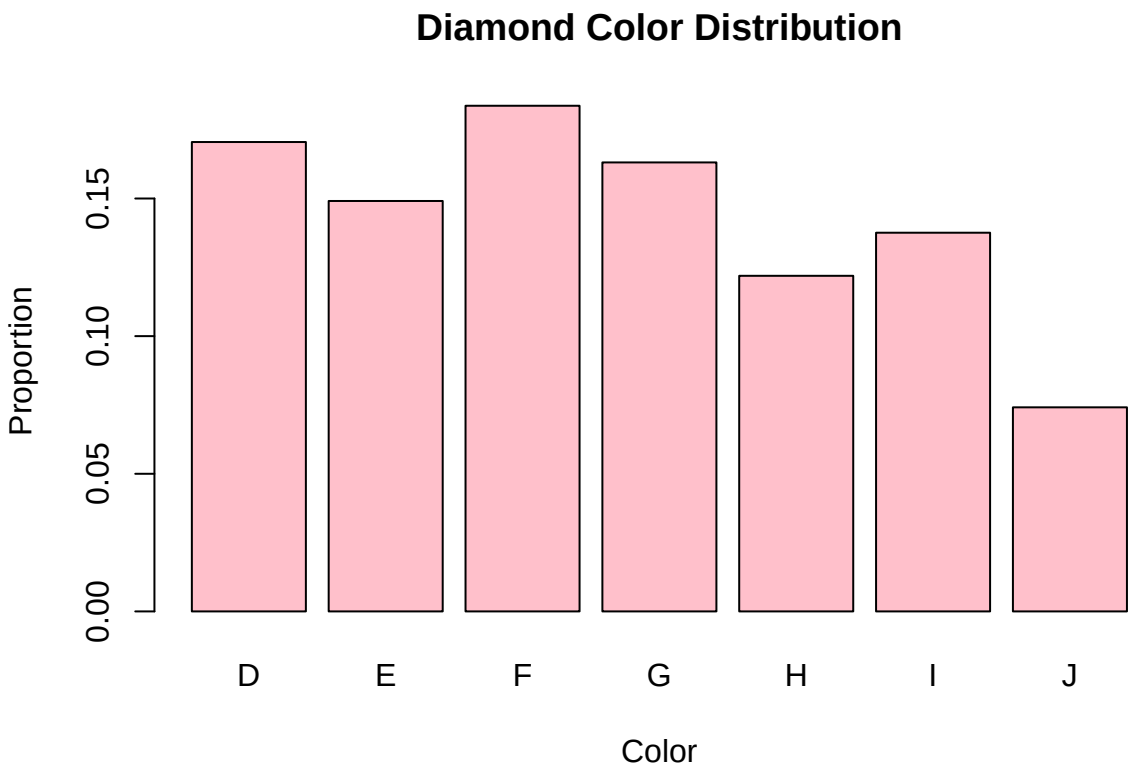


We now inspect the relationship that carat has with the other variables that comprise the 4C's. When analyzing carat againt clarity, we see that as the size of the diamond increases, there are less FL and IF diamonds. This falls in line with reason in that while FL and IF diamonds are hardest to find in general, even rarer still are FL and IF diamonds of high carat weight.

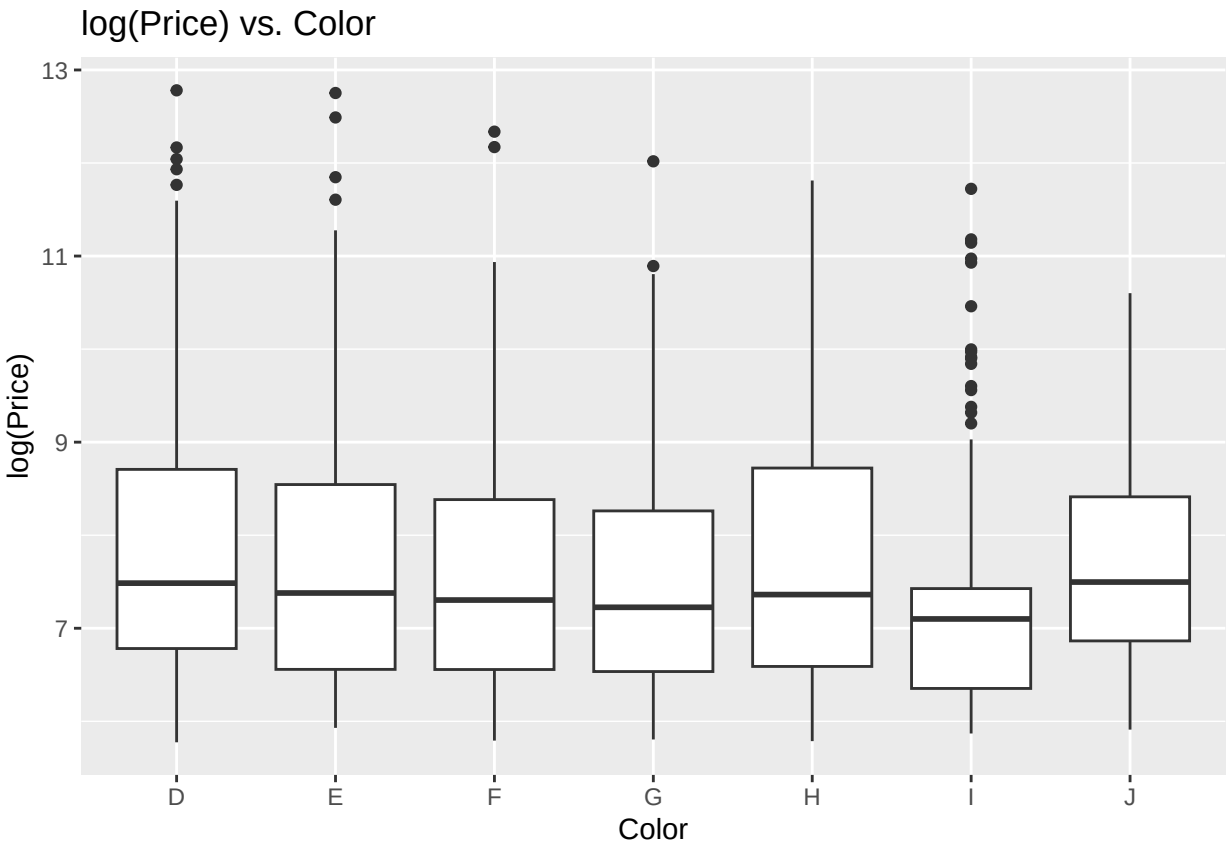
When looking at the relationship between carat and color, we see that for the diamonds of high carat weight, the colors D, E, F, I, and H are most prevalent. As a diamond size increases, the more likely it is to show color. Therefore, consumers should tend to purchase colorless diamond grades (such as D, E, F) when purchasing high carat diamonds. Logically, it makes sense then that a diamond retailer might have more stock of colorless high carat diamonds rather than near colorless or faint color high carat diamonds due to the respective demand of each.

Inspecting the relationship between carat and cut, we see that the selection of Astor Ideal cut diamonds in our data set exhibits the lowest carat weights. It is worth noting the claim that cut may be more influential on a diamond's value than size, as Astor Ideal diamonds best optimize light performance and create the most impressive sparkle and therefore may not necessitate high carat diamonds to be very valuable.

Color

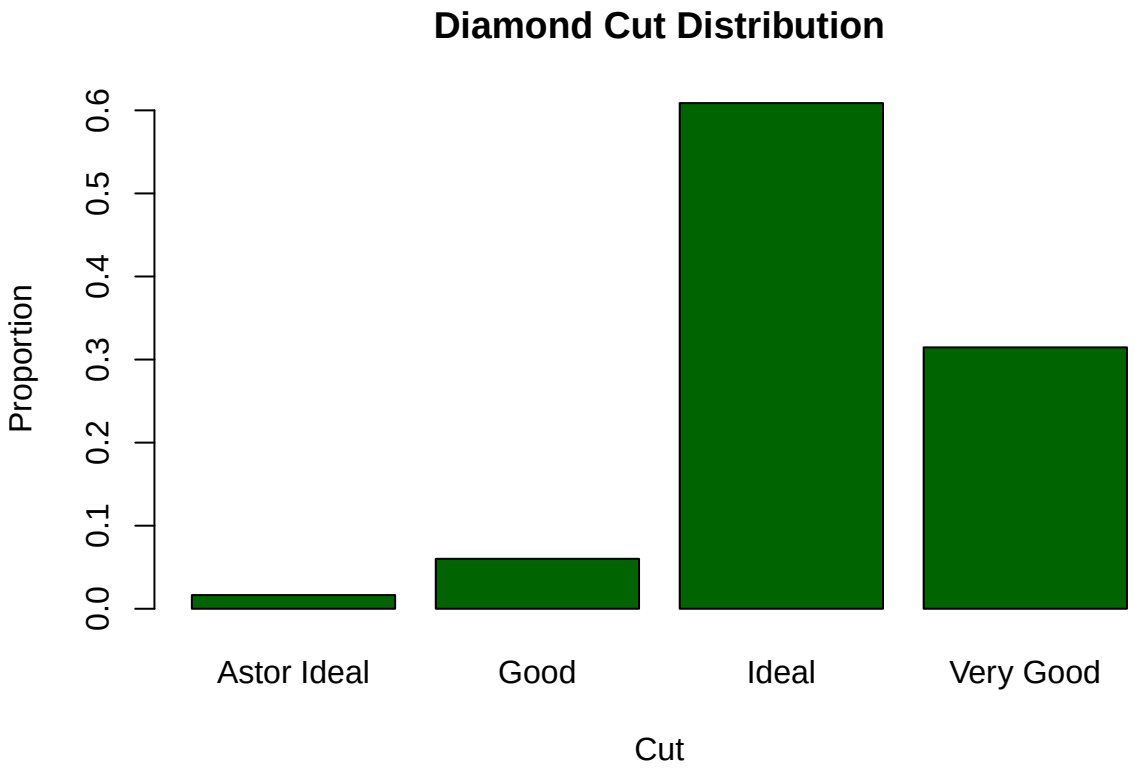


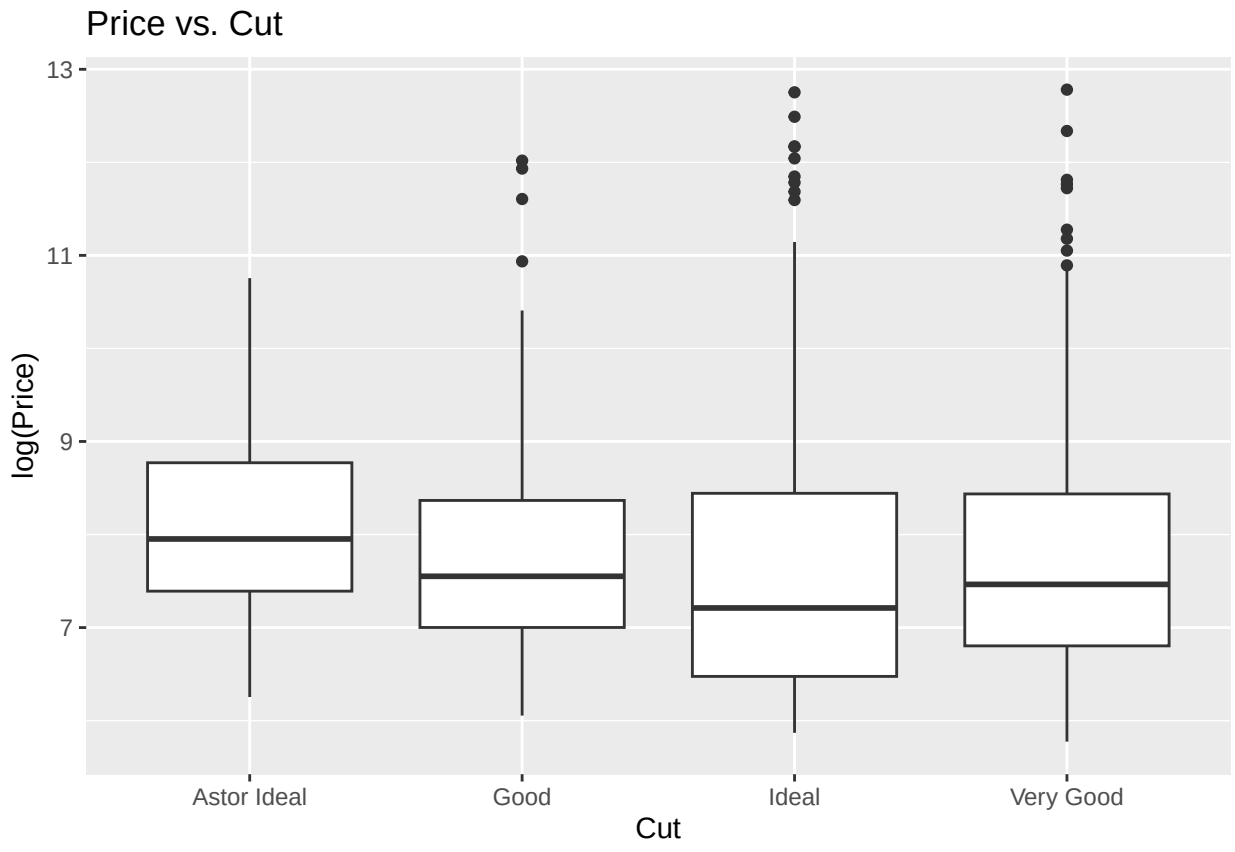
We assess the distribution of diamond colors within our data set using the above barplot. This barplot shows a fairly even distribution of diamond colors, with each bar representing the percentage of diamonds with a specific color grade.



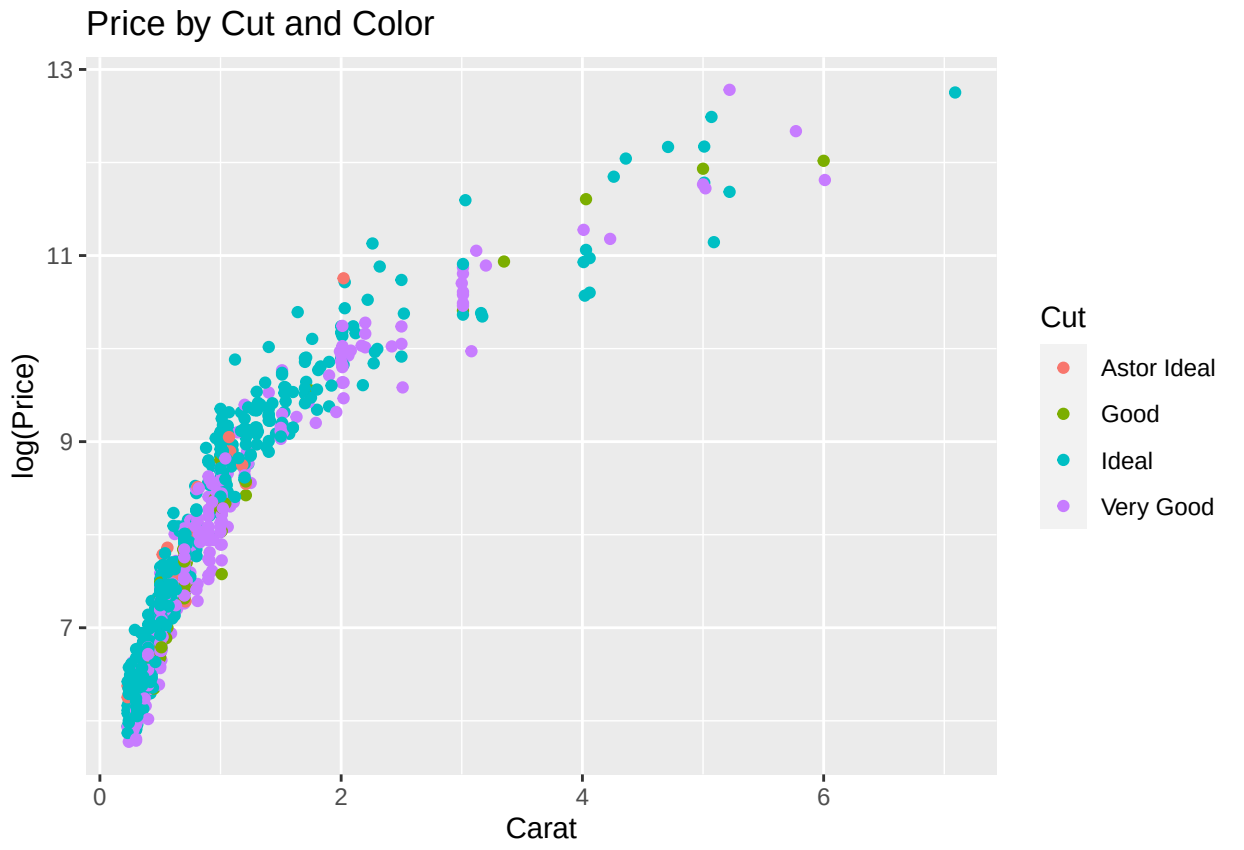
We see by inspecting the bar plot that the distribution of price within each color grade does not differ greatly. Inspecting the medians in particular, we see that the median price is nearly identical across all color grades. With this information, we glean that color grade does not have a significant impact on the price of a diamond in our data set. This differs from the belief that the most colorless diamonds (grades D, E, F) shall be significantly more expensive.

Cut


Diamond Cut Distribution
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Astor Ideal
Good
Ideal
Very Good
Cut

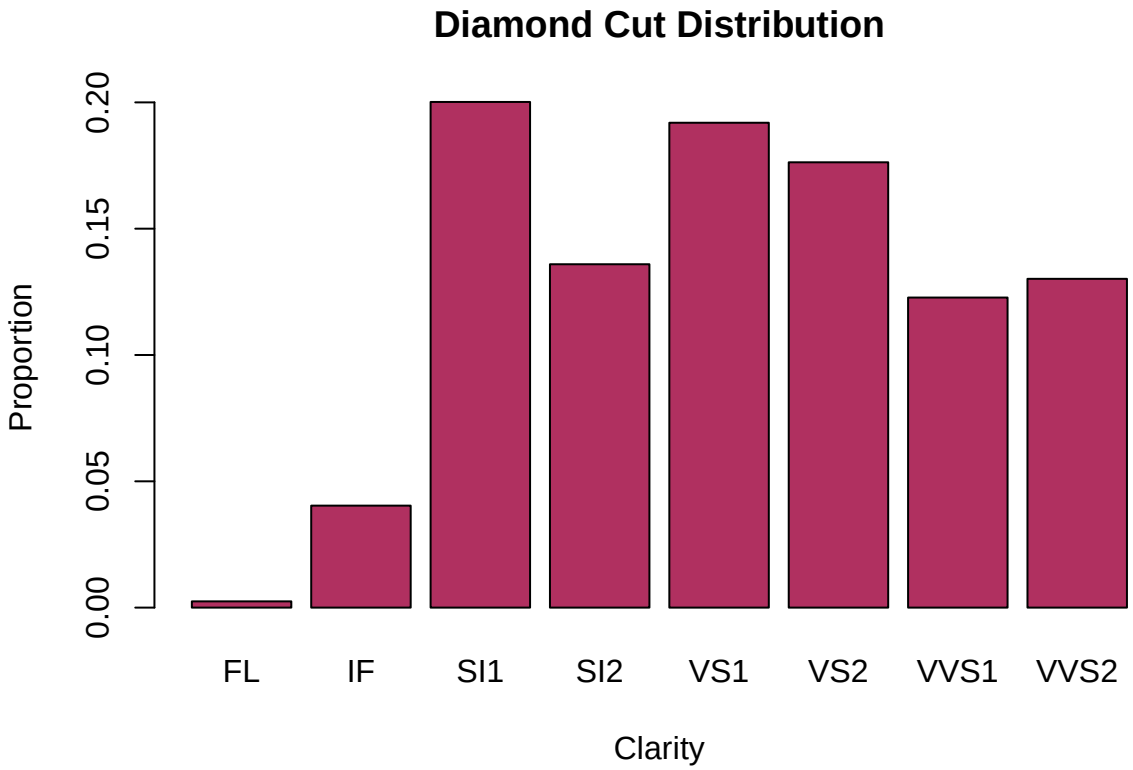

Price vs. Cut
log(Price)
13
11
9
7
Astor Ideal
Good
Ideal
Very Good
Cut

We see that diamonds of cut Astor Ideal quality do stand out as more expensive in both the median price and the interquartile range of price. We do also see that the price distributions across the other quality grades of cut do vary. This provides evidence to support the claim that while most people focus on carat, the cut of a diamond is influential on a diamond's price as well. Interestingly however, we see that the median price of Ideal diamonds is actually less than those of Very Good and Good cuts. This means that in the context of our diamond sample, we have evidence that conflates with the accepted thinking that Ideal diamonds are of higher diamond cut quality than Very Good and Good cut diamonds and therefore should be more expensive.

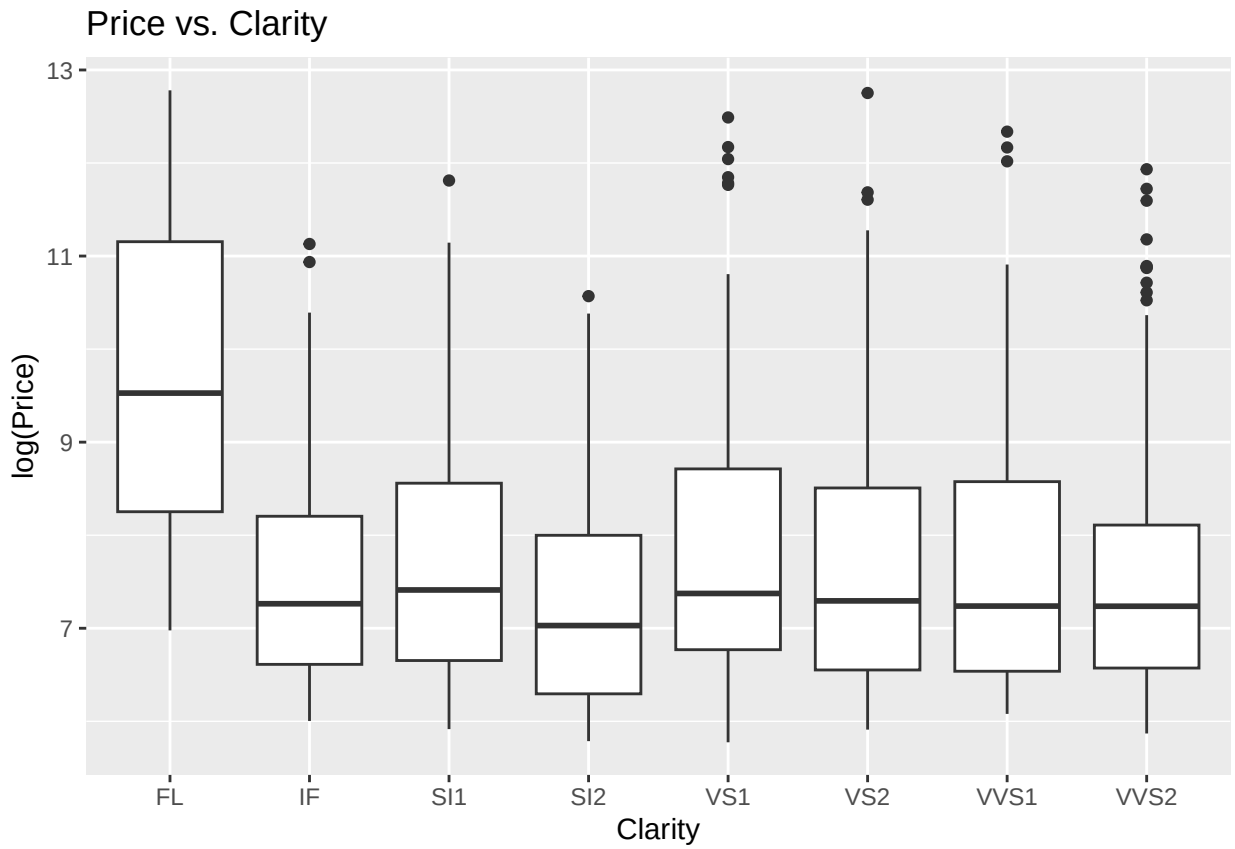


We can see on this graph the strong relationship between price, carat, and cut. Again, the Ideal and Very Good cuts represent the majority of diamonds within our data set.

Clarity



In looking at the distribution of diamond clarity within our data set, we observe that Flawless (FL) diamonds and Internally Flawless (IF) diamonds are the most rare within our data set. This makes sense as these two clarity grades represent diamonds that are free of inclusions or that in which small surfaces blemishes are only visible under a microscope.

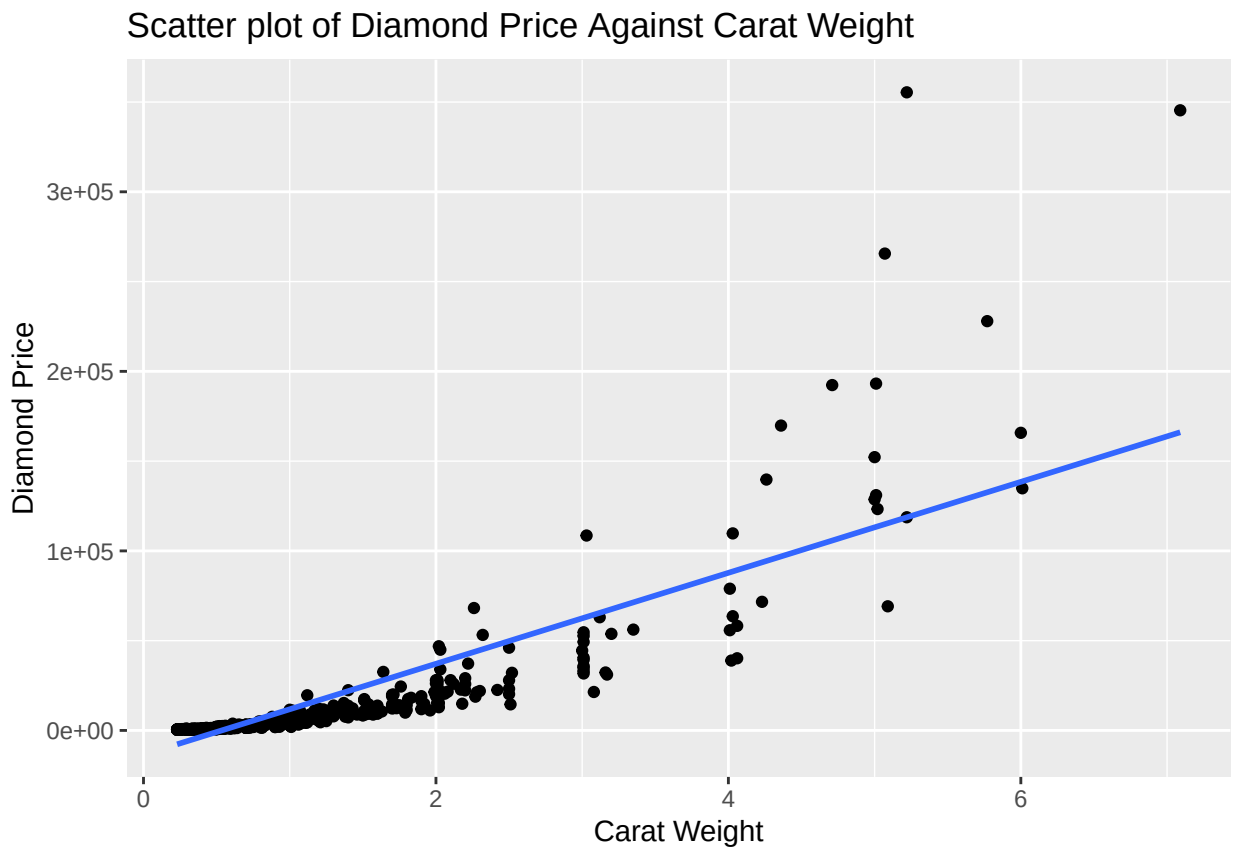


Based on the barplot of diamond price against clarity above, we observe a significantly higher median price associated with FL diamonds than any other clarity grade of diamond. This observation, along with the rarity of FL diamonds as specified in the previous commentary, supports the claim that FL diamonds are practically impossible to find and are therefore the most expensive. Inspecting the price distributions of the other clarity grades, we see that SI2 diamonds have the lowest median price. This relates to their definition of being the worst clarity quality of the diamonds found in our data set. Interestingly, we see that both the median price and interquartile range of price associated with IF diamonds is lower than SI1 and VS1 diamonds, which is unexpected given that diamonds of IF grade represent the second-highest level of quality.

Simple Linear Regression of Price Against Carat

First, we inspect the relationship between carat and price using a scatter plot.

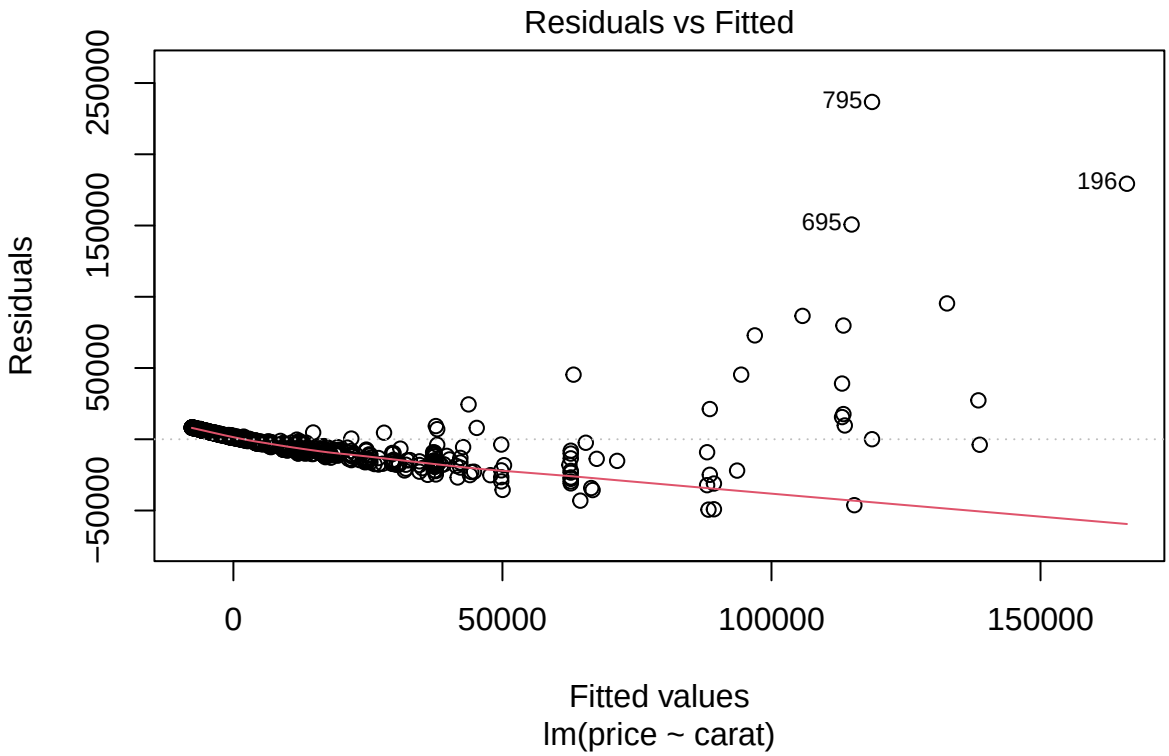
```
## `geom_smooth()` using formula = 'y ~ x'
```



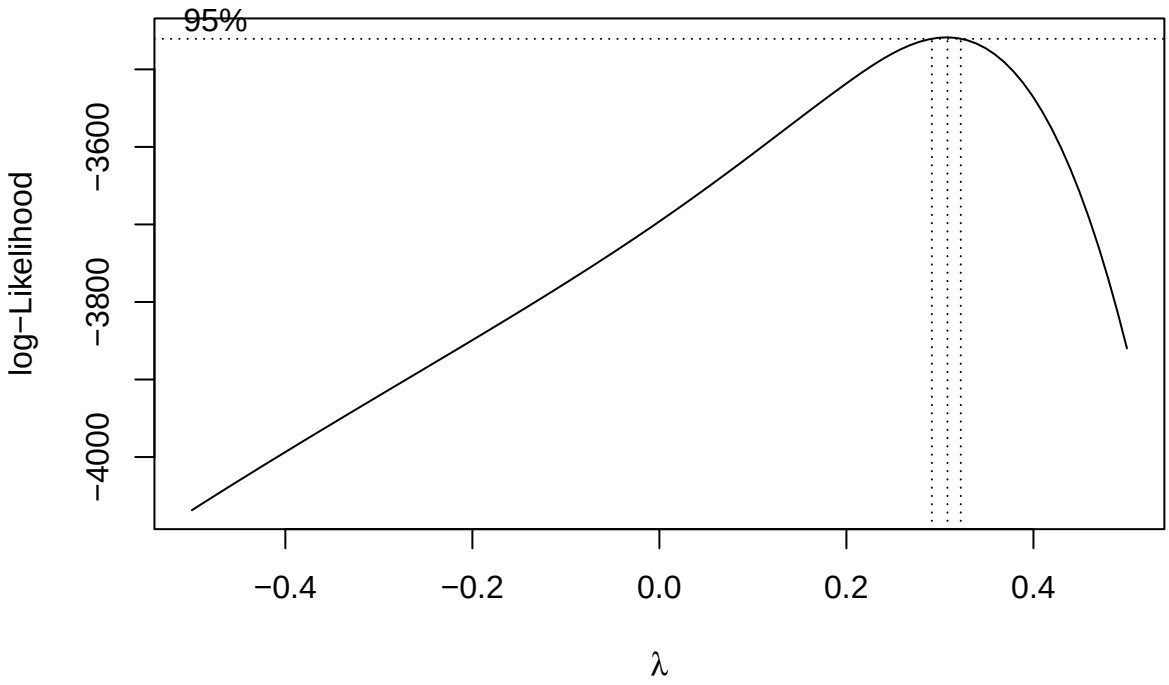
We see that there may be a linear relationship between carat and price, yet our regression assumptions appear to be violated. Let's inspect further.

```
##
## Call:
## lm(formula = price ~ carat, data = Data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49375  -5048   1867   4965 236711
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13550.9      559.7  -24.21   <2e-16 ***
## carat        25333.9      494.4   51.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13560 on 1212 degrees of freedom
## Multiple R-squared:  0.6842, Adjusted R-squared:  0.6839
## F-statistic:  2625 on 1 and 1212 DF,  p-value: < 2.2e-16
```

In our first regression model, we use carat to predict price. We observe a strong p-value associated with the ANOVA F test of this model. Interpreting coefficients, we see a price increase by $25,333.90 for each one-unit increase in carat. But, let's check the assumptions.



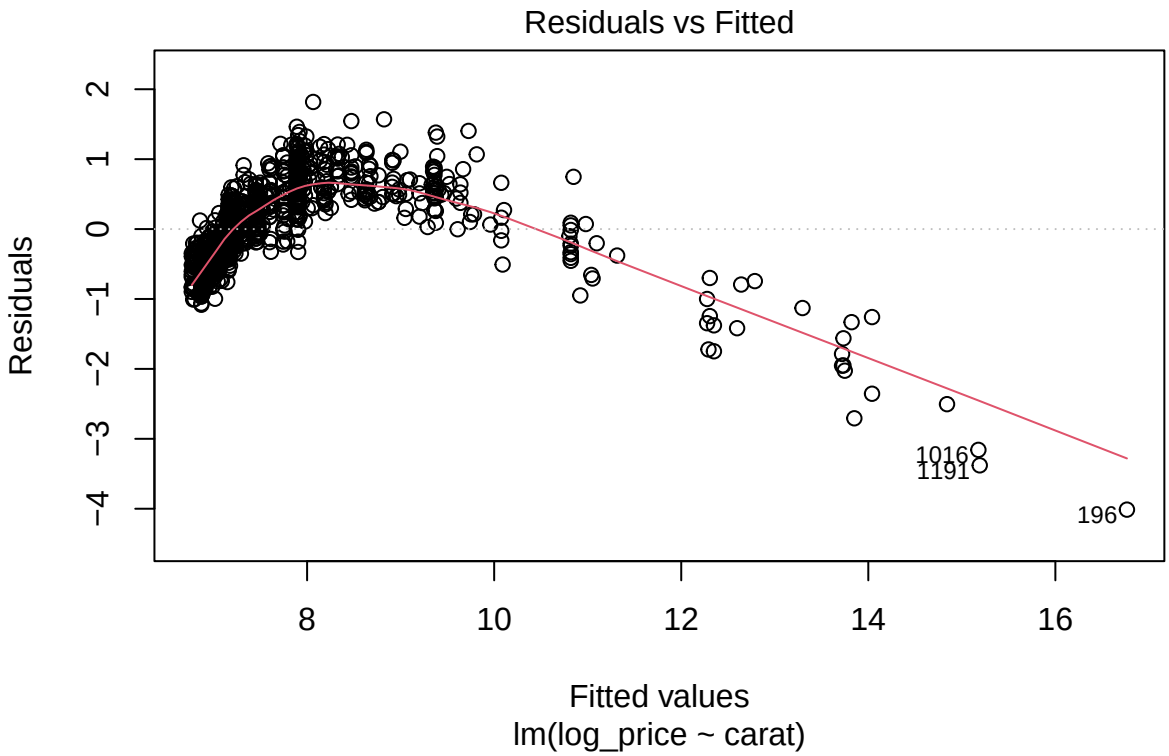
We see that while we have created a significant prediction model, the regression assumptions are violated. In referencing the Residuals vs. Fitted plot from the figure above, we see that residual values are primarily negative as the value of carat is small, and then are primarily positive as the carat grows. So, the mean of the errors is not equal to 0 within each range segment of predictor values. Furthermore, we observe residual variance growing wider into a cone shape as the value of carat grows. Therefore, residual variance is not constant and we must now look to transform the response variable, as we see both of our main regression assumptions are violated.



Our Boxcox plot suggests transforming the response variable to the power of 0.3 (lambda = 0.3). Yet, because it is so close to zero, we will try to raise the response to the power of 0 (log transformation) in order to interpret the coefficients of our resulting regression equation in a way that we otherwise would not be able to do if we were to transform the response variable to the power of 0.3.

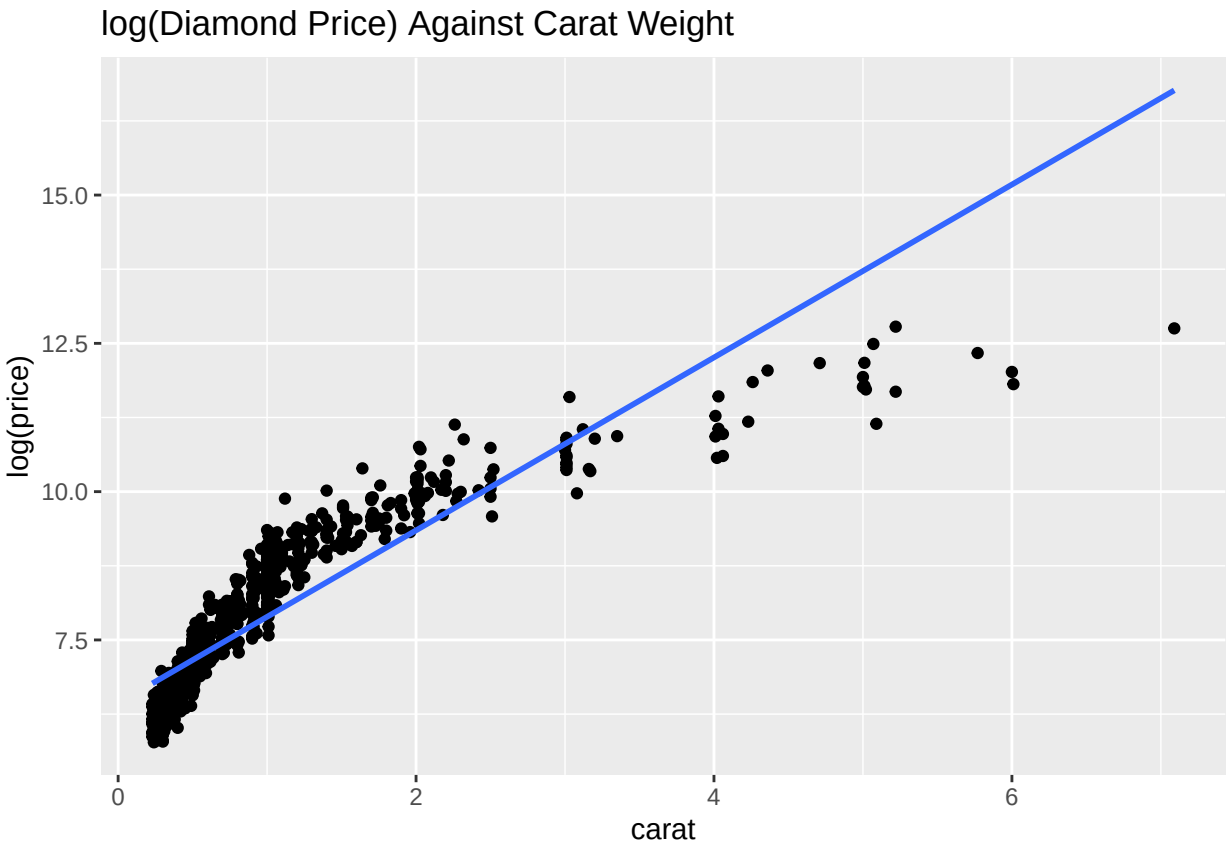
```
##
## Call:
## lm(formula = log_price ~ carat, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0128 -0.4350  0.0067  0.4139  1.8178
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.43235    0.02490  258.37   <2e-16 ***
## carat        1.45739    0.02199   66.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6032 on 1212 degrees of freedom
## Multiple R-squared:  0.7837, Adjusted R-squared:  0.7835
## F-statistic:  4392 on 1 and 1212 DF,  p-value: < 2.2e-16
```

Upon performing a log transformation of the response variable, we again observe a strong p-value associated with the ANOVA F test of the model. But, we must again check that the regression assumptions are not violated.



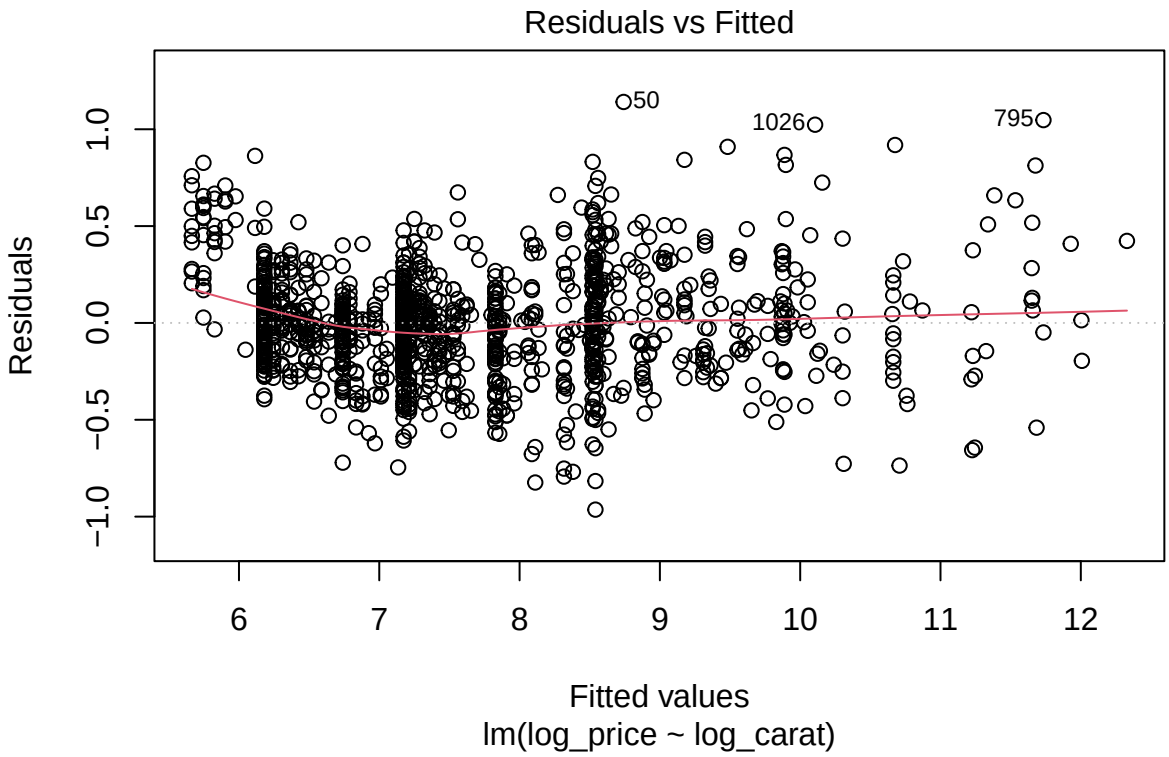
We see that we have handled the assumption in which residual variance must be constant, as shown on the Residuals vs. Fitted graph above. However, we still see that the errors do not have a mean of 0, as the residuals have negative values at the beginning and end of the range of our predictor, with positive values in the middle. Seeing this violation still present, we must look to transform the predictor variable.

```
## `geom_smooth()` using formula = 'y ~ x'
```

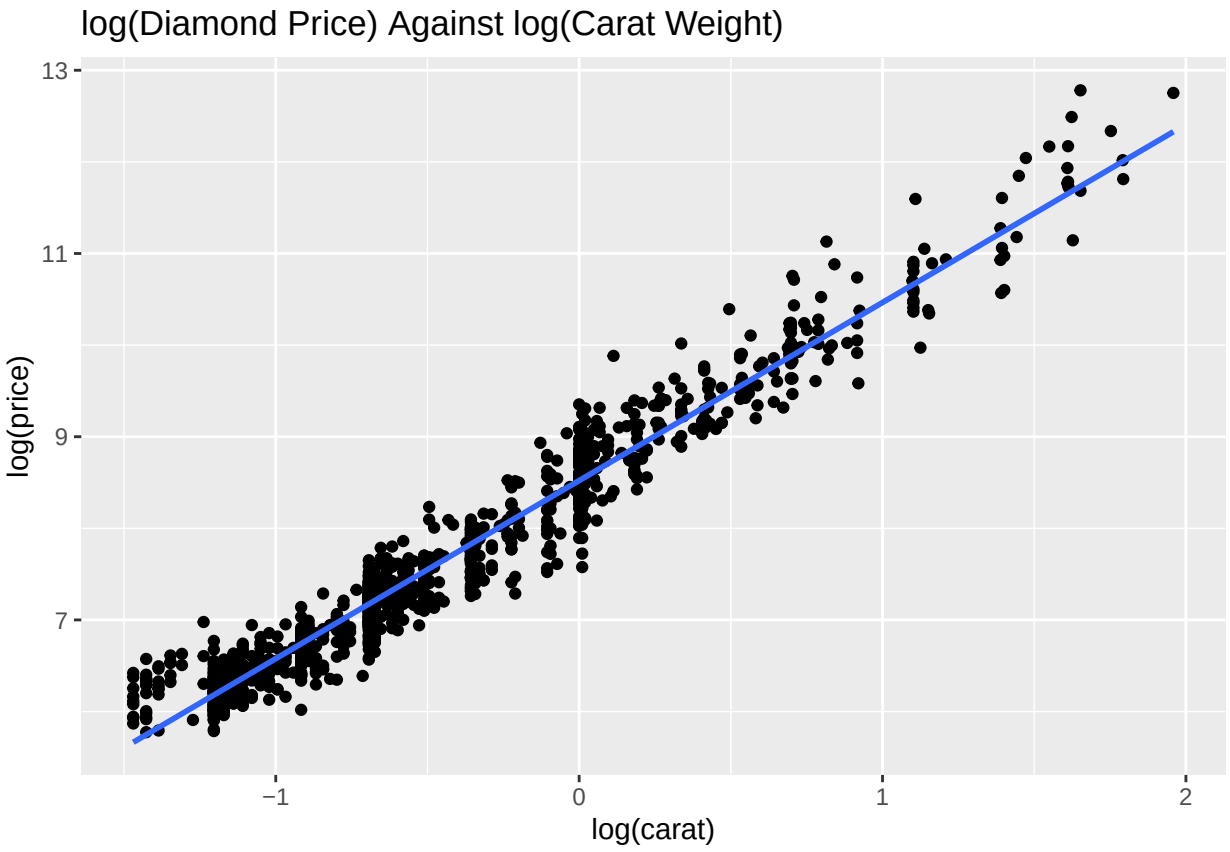


We see from the scatter plot of log_price against carat embodies a logarithmic function's curve, so we will try a log transform of the predictor.

```
##
## Call:
## lm(formula = log_price ~ log_carat, 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 ***
## 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
```




```
## `geom_smooth()` using formula = 'y ~ x'
```



Upon performing a log transformation of the response variable, our model now utilizes log(carat) to predict log(price). We again observe a strong p-value associated with the ANOVA F test of the model. We also now see from the Residuals vs. Fitted graph that the residuals form a horizontal band around 0. The residuals are now evenly scattered positively and negatively, and our residual variance is constant. We can also see from a scatter plot of log(price) against log(carat) that our transformations have now effectively created a strong linear relationship. Given that we have a strong p-value and have satisfied the regression assumptions, we can interpret our regression model. The regression model is: $log(price) = \beta_0 + \beta_1 log(carat) + \epsilon$. Therefore, our regression equation is: $\hat{y} = 8.521208 + 1.944020(x)$. Interpreting this equation contextually, for a 1% increase in carat weight, the price of a diamond increases by approximately 1.94402%.