Group 9 Project Report

Alec Pixton, Darreion Bailey, David Hernandez, Venkat Viswanathan

1. Summary of Findings

Blue Nile is a diamond and engagement ring business focused on making it possible for consumers to shop for high-quality diamonds at great value. Blue Nile specializes in providing consumers with diamond buying tips and education guides, including the 4Cs of Diamonds, which is an educational guide to understanding the unique characteristics of stones as they relate to their Cut, Color, Clarity, and Carat Weight, to help consumers choose the perfect wedding band.

Blue Nile asserts multiple claims on their website and provides an analysis of diamond characteristics (Cut, Clarity, Color, and Carat Weight) in comparison to price to educate its consumers and assist them with selecting the best wedding band. Our goal was to test the accuracy of each of Blue Nile's claims and provide supporting evidence using Blue Nile's inventory of diamonds and their recorded characteristics.

In researching the claims for the Cut characteristic of diamonds, Blue Nile asserts that a diamond's cut is the most important factor to consider compared to Color, Clarity, and Carat. Therefore, consumers should spend most of their money on selecting the best cut grade, the Astor Ideal cut. Our analysis supports Blue Nile's claim that Ideal cuts, in general, dominate the upper price range of their inventory of diamonds and confirms that cut is a significant factor in price influence, as illustrated in *Figure 1: Cut vs. Price*.

According to the Gemological Institute of America (GIA) color scale, the industry standard for diamond color grades ranges from D (i.e.: the most colorless diamond) to Z (i.e.: a light yellow or brown diamond). While Blue Nile does not offer the L-Z color grades, they assert that diamond prices will decrease in alphabetical order. Our analysis supports Blue Nile's claim, as colors D, E, F, G, H, I, and J are listed on the website and in the dataset as the most to least expensive in alphabetical order (see *Figure 3: Color versus Price*).

A diamond's clarity is the assessment of non-visible and visible (i.e.: "eye-clean") imperfections on the surface and within a diamond. Blue Nile claims that Clarity is the least important characteristic since imperfections are not usually seen with the naked eye. However, Blue Nile acknowledges that fewer and more minor scratches on a diamond will receive the highest clarity grades, such as VS, VVS, IF, and FL, and are more expensive. Blue Nile recommends that a consumer should select a clarity grade that is not too expensive, such as the FL and IF diamond clarities, and should select a diamond with inclusions that cannot be seen with the naked eye, such as the VS (Very Slightly Included) and SI (Slightly Included) diamonds. In our analysis of Clarity versus Price, Blue Nile's claim that diamond imperfections are the least important is supported in *Figure 5: Frequency of Diamond Clarity Grade* because consumers prefer the VS and SI clarity grades over the more expensive imperfection-free clarity grades IF and FL. Additionally, *Figure 2: Clarity versus Price* supports the idea that the higher clarity grades are generally more expensive.

According to Blue Nile, consumers should buy below half and whole-carat values to save significant money, as other people will never notice a diamond's slight difference in weight. Our analysis of Blue Nile's claim is supported because there is a higher frequency of diamond rings sold for the Ideal cut, suggesting its popularity among consumers and that the cut of a diamond versus its weight is more important, as illustrated in *Figure 4: Cut added for more context*.

As shown in Figure 4, we categorized a diamond's carat weight by creating a low and high variable for weight. Blue Nile claims that the price per carat of diamonds is a better deal if you buy half and below whole carat values; for example, buying 1.9 carats instead of 2 carats is a better value. However, the data does not support

their claim. Our analysis concluded that comparing the low diamond weight with price is the same as comparing the high carat weight with price. In other words, low and high-carat weights are highly correlated with the price of a diamond. The supporting evidence of our analysis is determined by the R-squared value, which indicates how well the data fits our model. In simple linear regression, a statistical method for understanding the relationship between two variables, such as price and carat, R-squared denotes the proportion of variance in the response variable, price, explained by the predictor, carat. In other words, R-squared values closer to 1 indicate a strong relationship, while values closer to 0 indicate a poor relationship between price and carat. Our analysis revealed an R-square of 0.9547, suggesting a strong relationship between price and carat.

2. Visualizations

2.1 Data & Variables

This dataset provided by Blue Nile describes more than 1200 different diamonds that are for sale on their website, and is a reflection in some ways of their standards as a diamond merchant. The dataset is in a .csv format. The following are the variables which comprise the dataset: Carat, cut, clarity, color, and price. An additional variable (carat_cat) was also created to address a claim in the next section.

Carat refers to a diamond's weight, which is a quantitative variable. Cut refers to the quality of the dimensions of a diamond (i.e. how the surfaces are positioned to create sparkles) and consists of 4 grade categories: Astor Ideal, Ideal, Very good, and Good. Clarity refers to the number of imperfections internally and externally about or within the stone itself, and consists of 11 grades, 8 of which Blue Nile sells: SI1, SI2, VS1, VS2, VVS1, VVS2, IF, and FL diamonds, with FL diamonds being the most rare and expensive. Color refers to the purity of the color, and Blue Nile's business consists of 3 categories across 8 grades: Colorless diamonds (D, E, and F grades), Near-colorless diamonds (G, H, I, and J), and Faint color diamonds (K).

Additionally, we have trimmed each visualization to be $0 < x < 25000$ which excludes outliers and makes the patterns identified below more apparent.

2.2 Price vs. 4Cs & Addressing Claims

Cut

Using *Figure 1* (*Cut vs. Price*), we can see that Astor Ideal cuts have a higher average price than any other cut, which is to be expected given that it is the highest cut grade. We can also see that Ideal and Very good are almost identical, with Ideal having a slightly higher average price than Very Good. From this, we gather that a higher cut grade does correlate to a higher price.

Using *Figure 4 (Carat vs. Price)* - where cut is added for context - we can see that the Ideal cut dominates the upper half of the distribution. This fact, coupled with the tightness of the distribution and the steep gradient (density) seen in *Figure 1 (Cut vs. Price)*, confirms the claim on the Blue Nile website that the cut of the diamond **can be** the biggest factor on price.

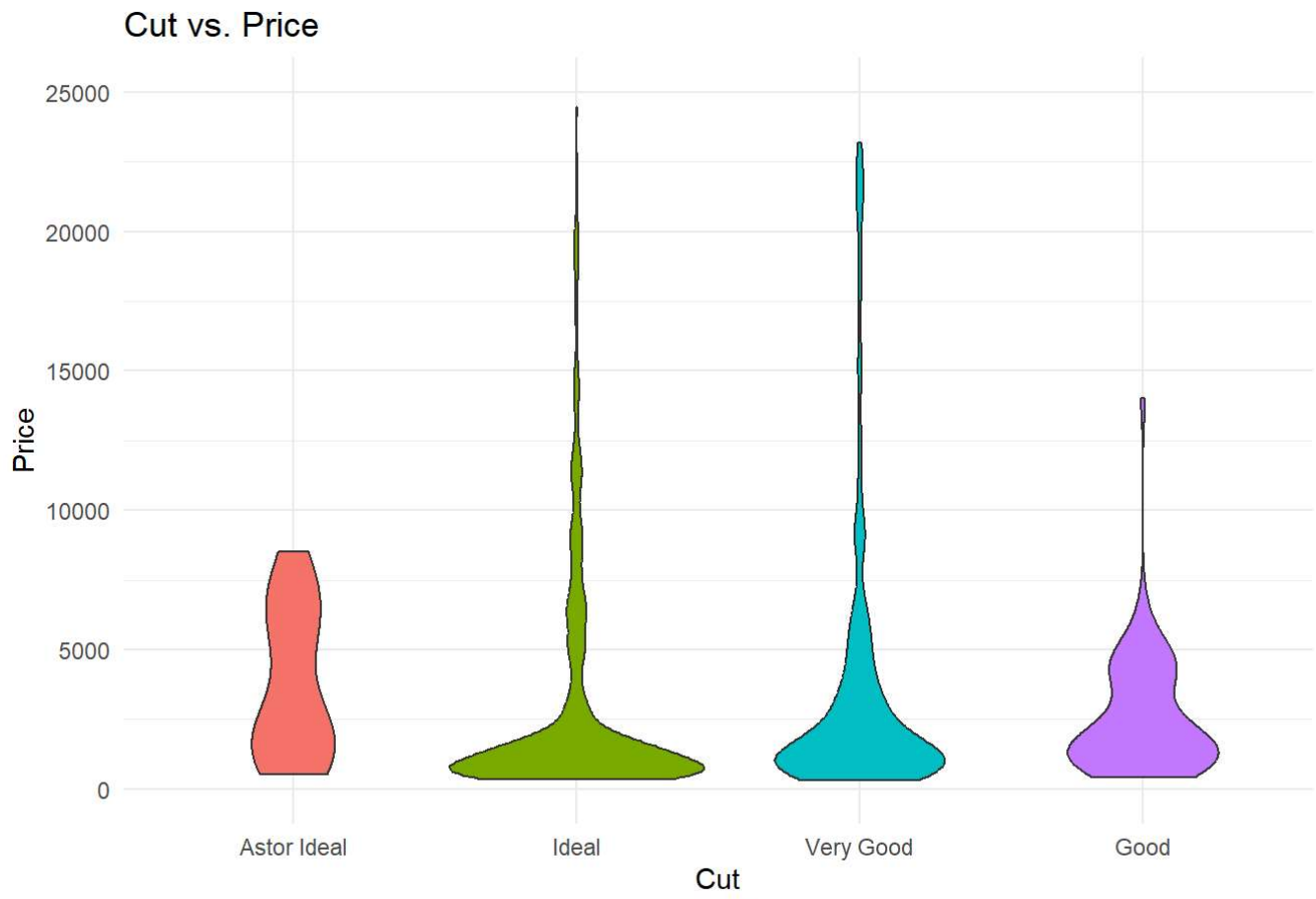


Figure 1

Figure 4: 'Cut' added for more context

Clarity

Using *Figure 2* (*Clarity vs. Price*), we can see that Flawless (FL) diamond clarities are the most expensive. Given that there are so few FL clarity diamonds in this dataset, this would also give credence to the claim made by Blue Nile that less than 1% of all diamonds are FL clarity, though this is hardly provable given our sample size.

Additionally, using *Figure 5 (Frequency of Diamond Clarity Grades)*, we can see that the distribution of diamond clarities are such that VS1 and VS2 are the most popular clarity among Blue Nile, which confirms their claim as such.

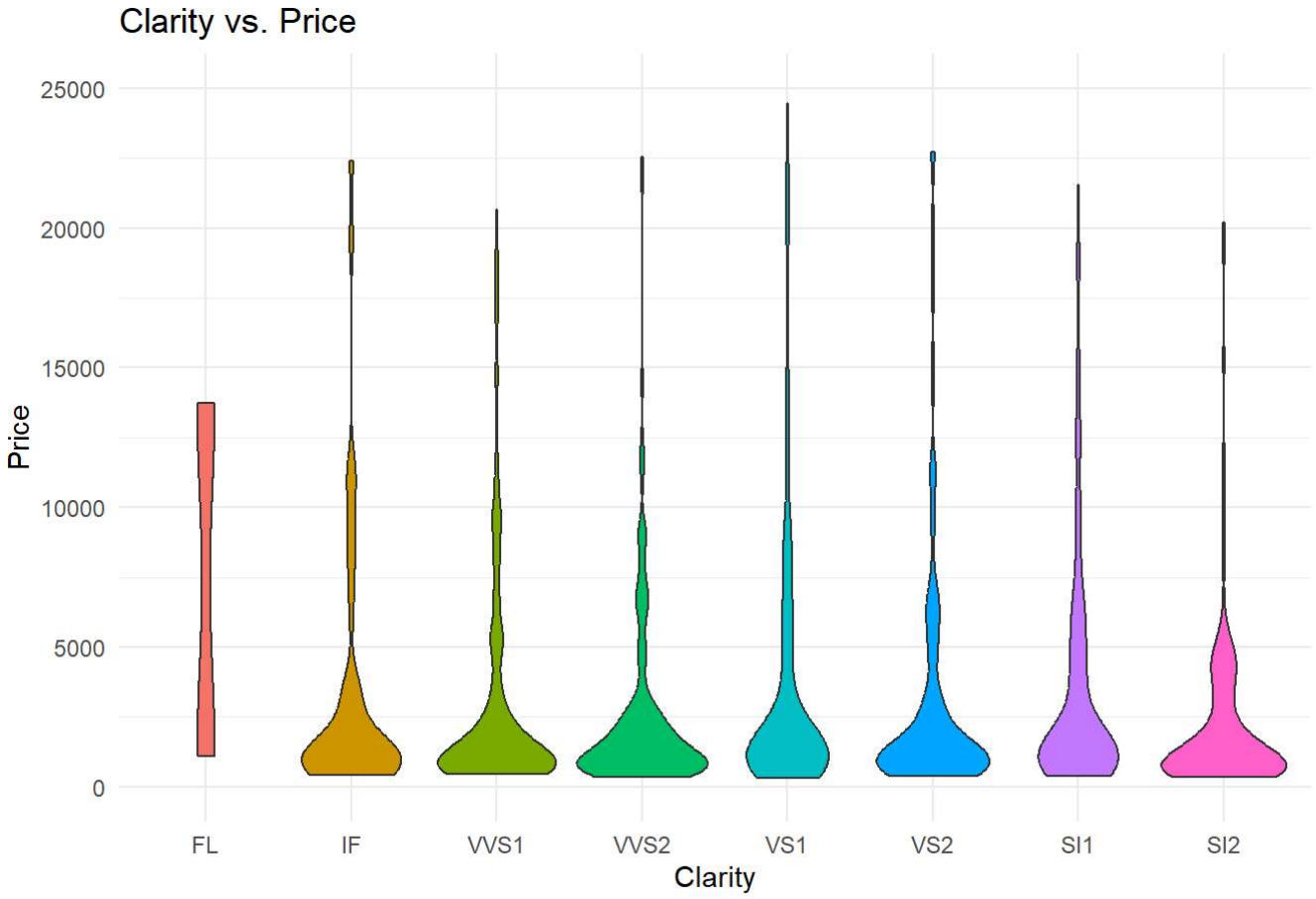


Figure 2

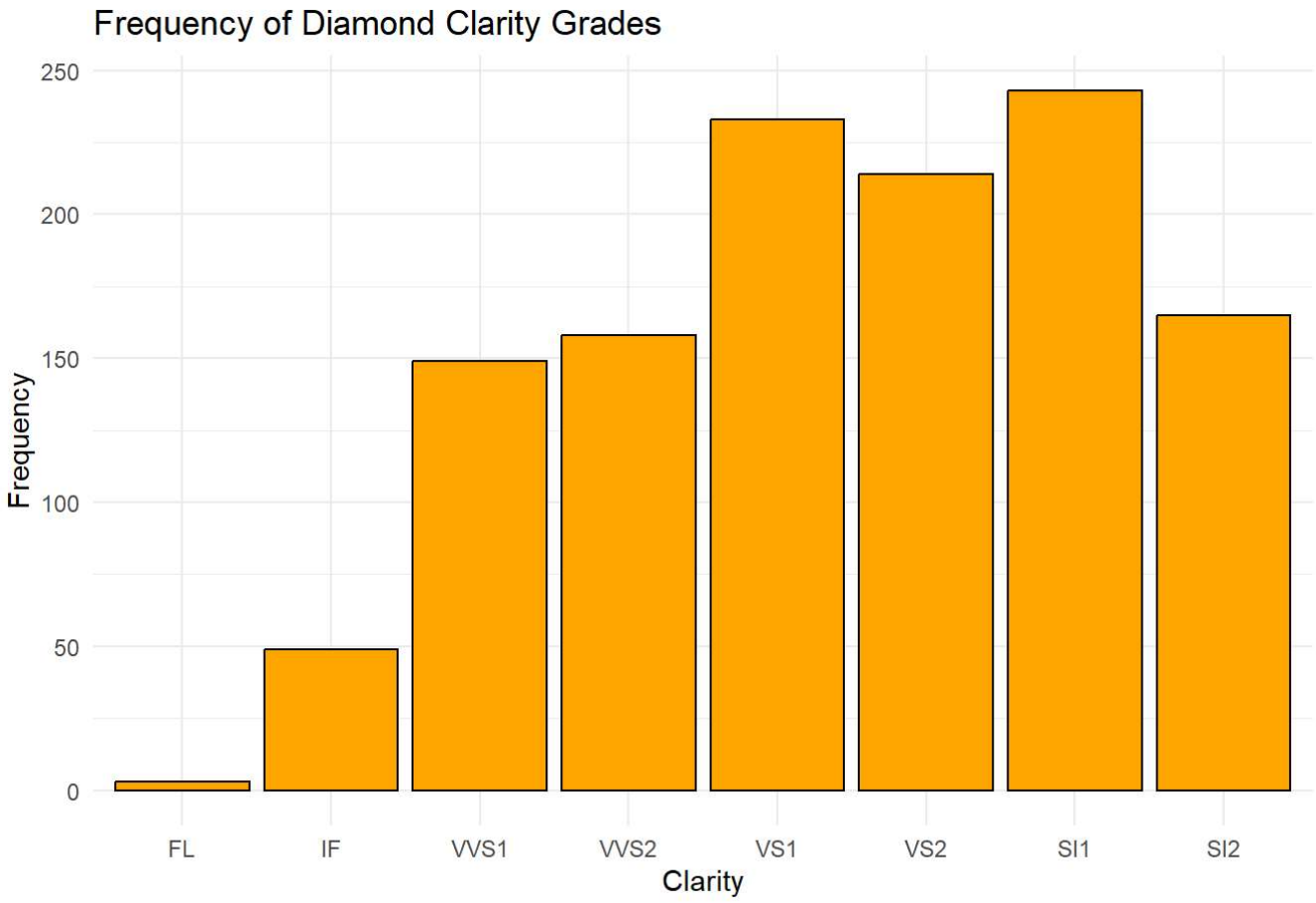


Figure 5

Color

Using *Figure 3 (Color vs. Price)*, we can see that D, E, and F color grades are by far the most expensive, which is to be expected given that they are the highest quality color grades according to Blue Nile. Additionally, we see that G, H, and I price values hover around the midpoint, which confirms their claim that these color grades are a great value for their quality.

However, as we can see in *Figure 6 (Frequency of Diamond Color Grades)*, Blue Nile appears to sell the least amount of J color grade diamonds, which fails to confirm their claim that D, E, and F color grades are the rarest among all color grades. Although, this dataset is not a representation of all diamonds harvested, and this data could be a reflection of Blue Nile standards and marketing influence.

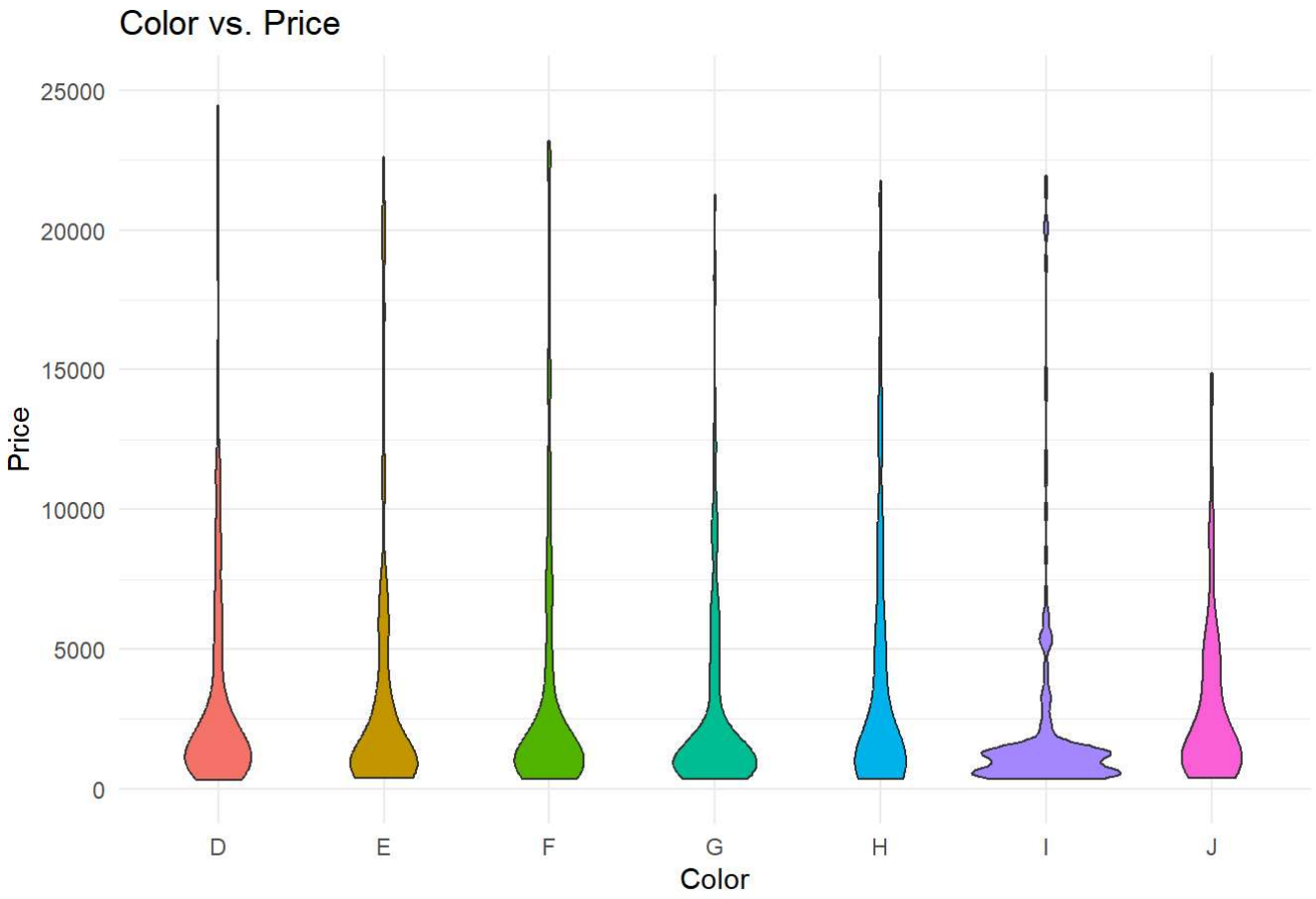


Figure 3

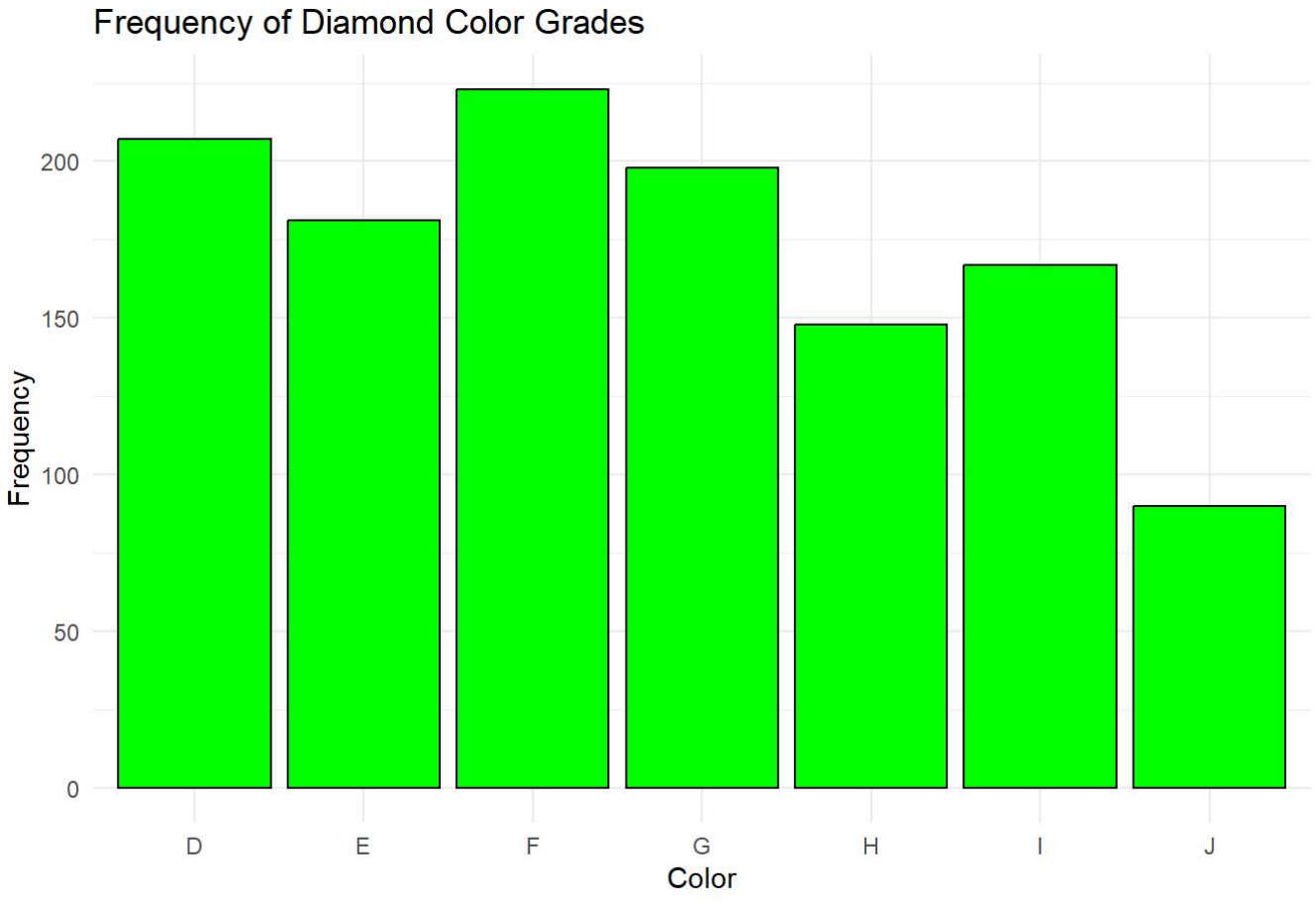


Figure 6

Carat

Using *Figure 4 (Carat vs. Price)* - where cut is added for context - we see a very tight distribution and a steep gradient. This indicates that carat has a very strong relationship with price. In comparison to the other of the 4Cs, the distributions are not quite as tight, and the gradient is not quite as steep, confirming their claim that carat weight is the biggest factor on price.

Using *Figure 7 (Carat vs. Price by Carat_cat)*, we can infer based on the visualization that given the density of data points around half and full carat increments that the claim on the Blue Nile website which suggests that buying slightly below half-carat and whole-carat values will save significant money for little compromise is not confirmed. This point will be further addressed in detail in SLR section of this report.

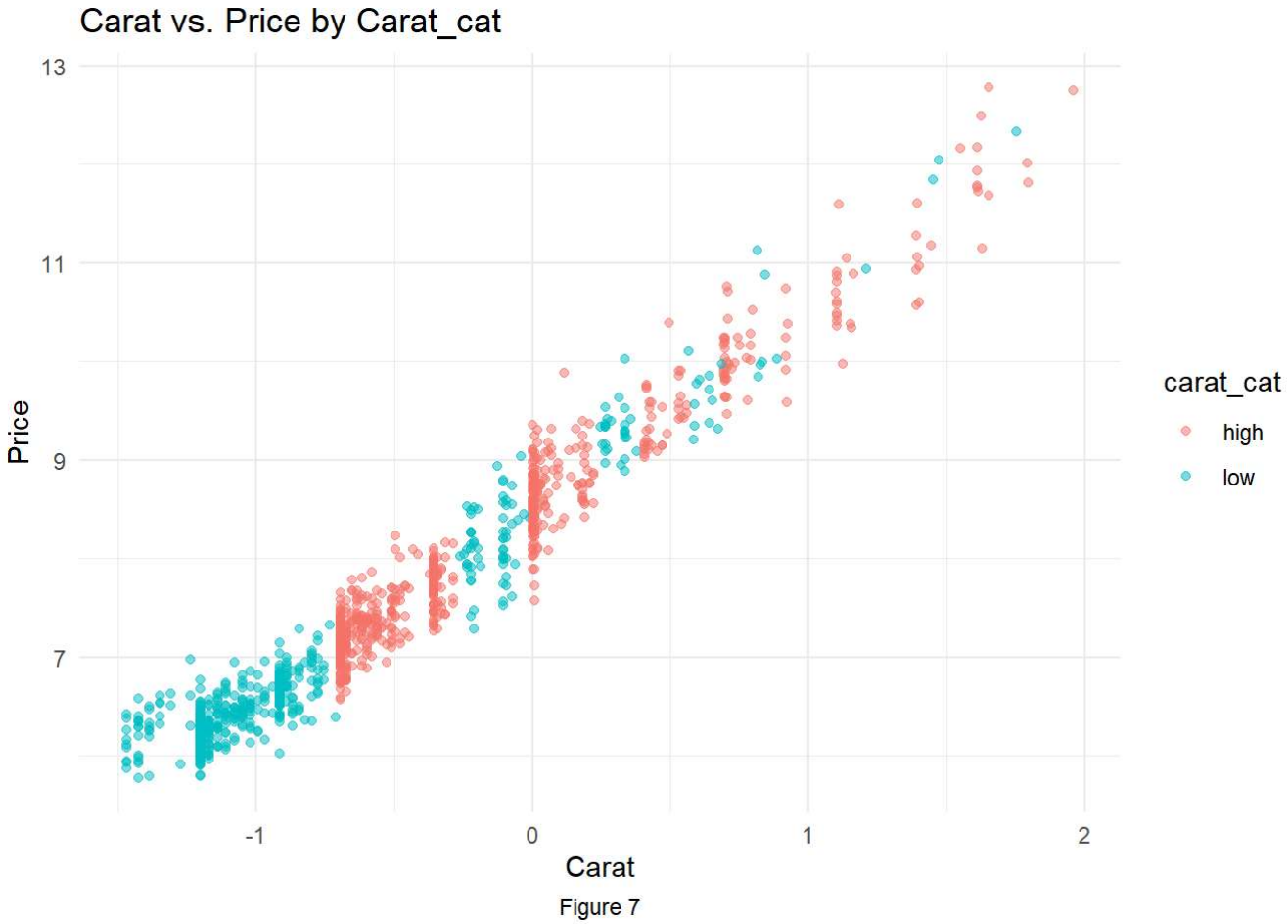
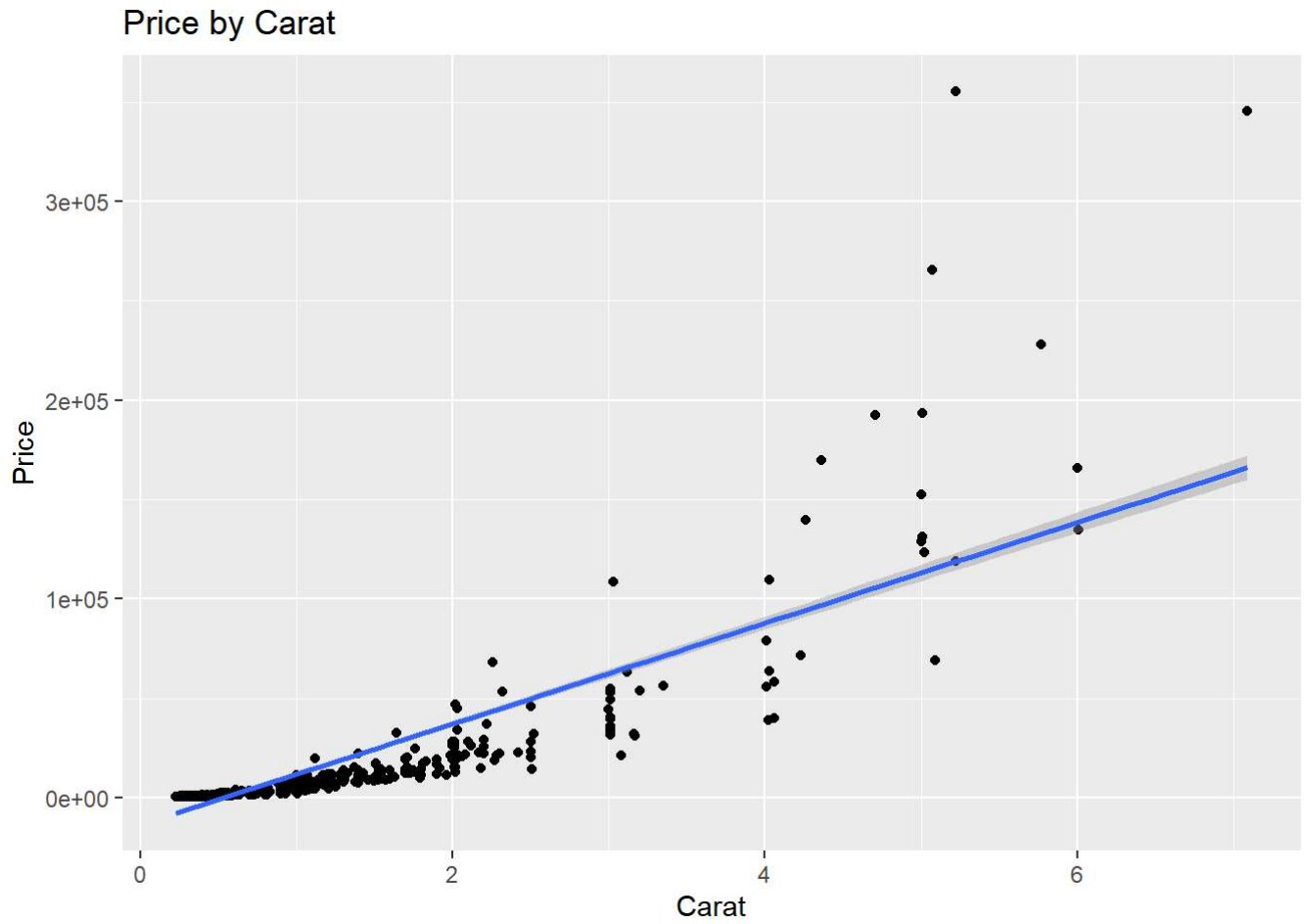
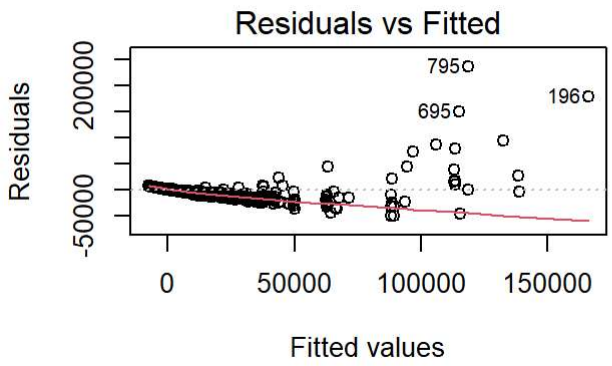
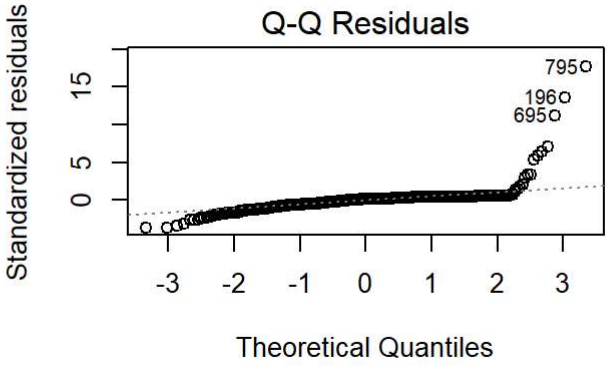


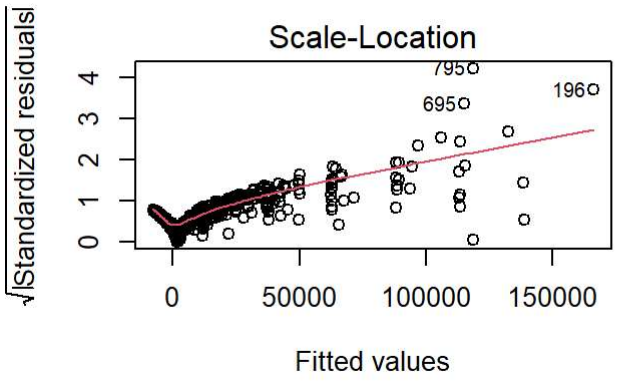
Figure 7

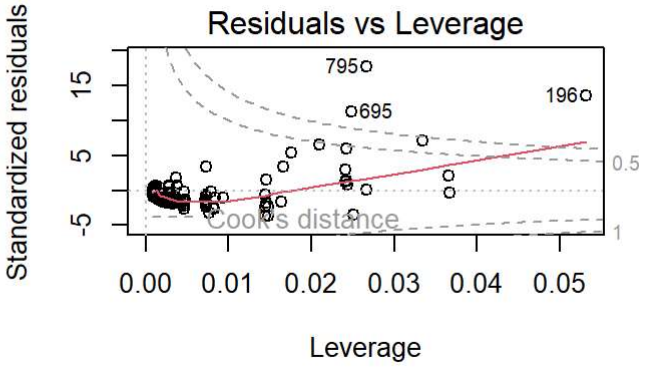
3. Simple Linear Regression Model


Price by Carat
Price
3e+05
2e+05
1e+05
0e+00
0
2
4
6
Carat

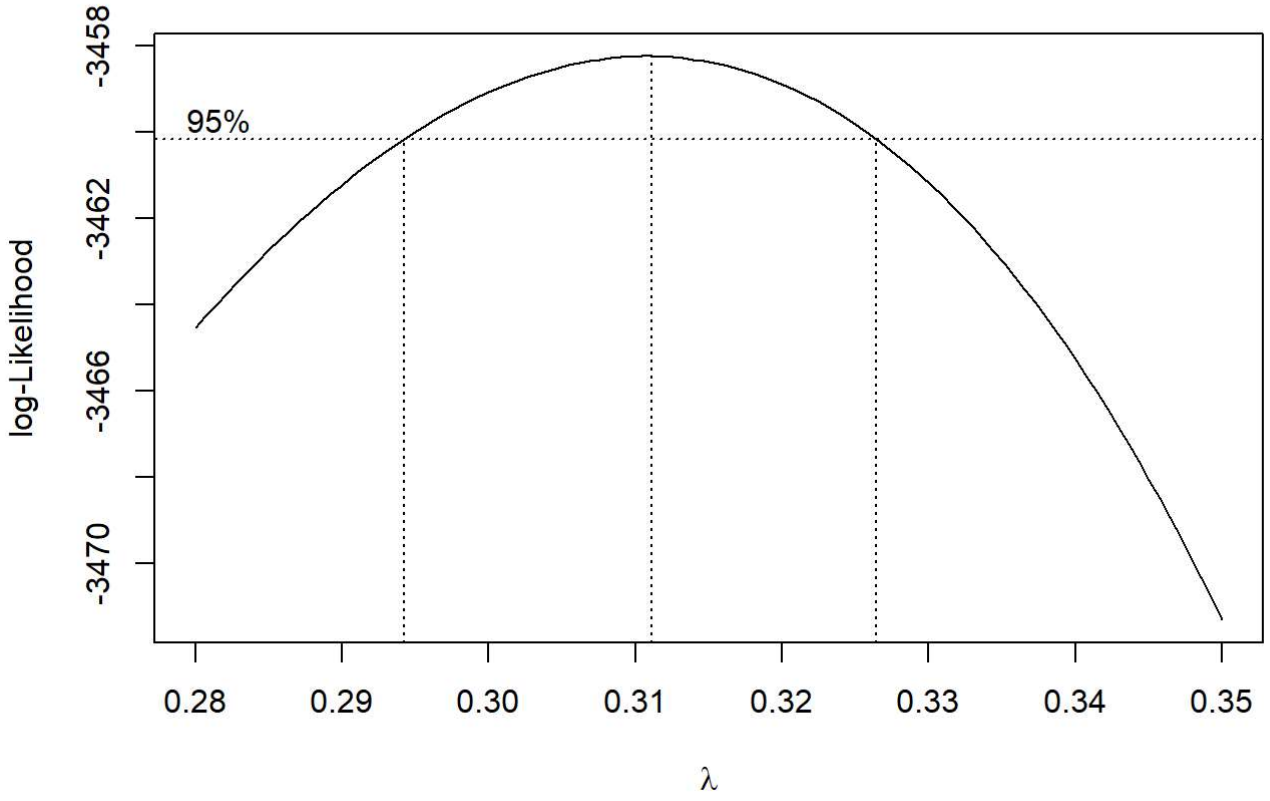

Residuals vs Fitted
Residuals
200000
-50000
795
196
695
0
50000
100000
150000
Fitted values


Q-Q Residuals
Standardized residuals
15
5
0
795
196
695
-3
-2
-1
0
1
2
3
Theoretical Quantiles

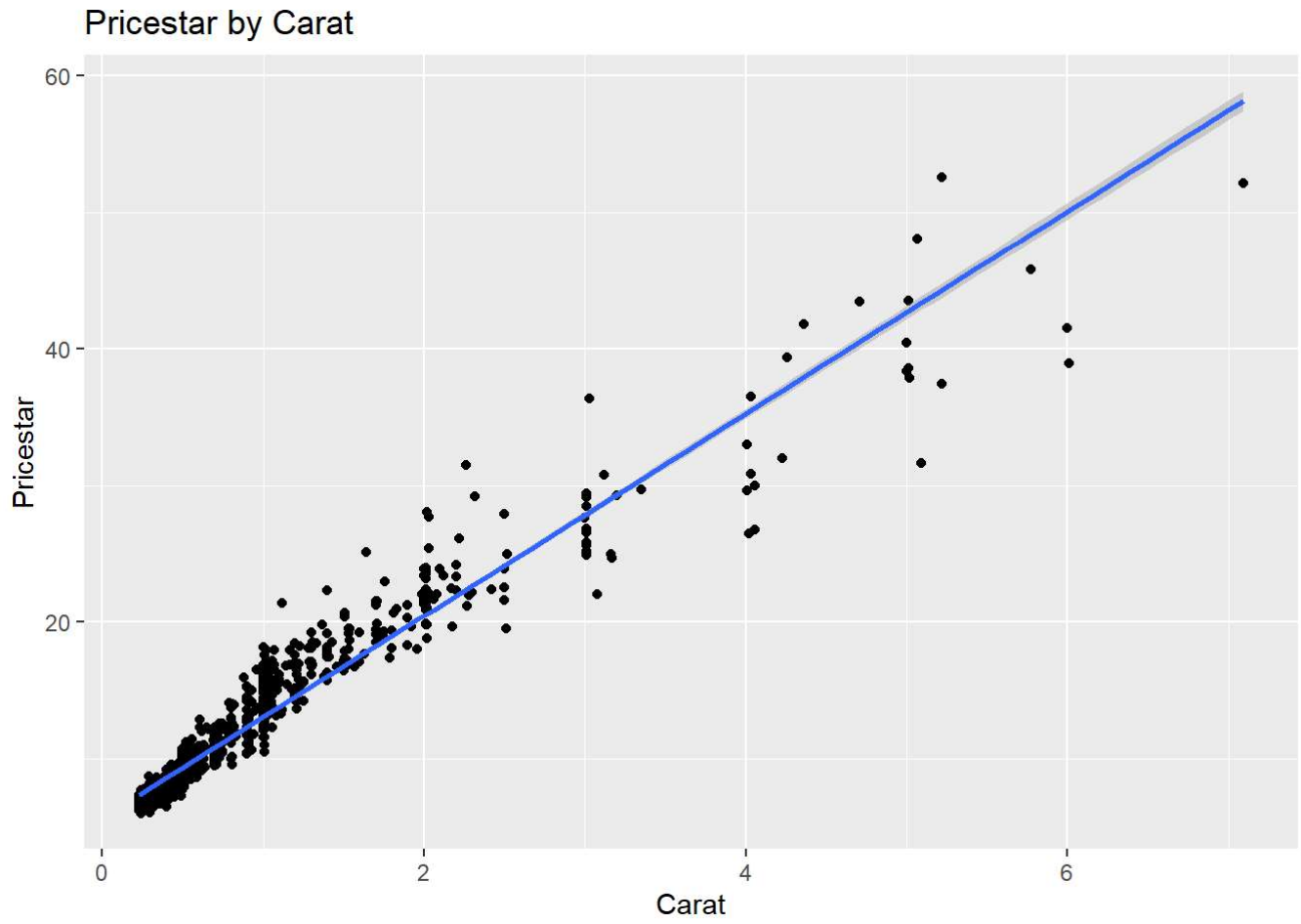

Scale-Location
√|Standardized residuals|
4
3
2
1
0
795
196
695
0
50000
100000
150000
Fitted values

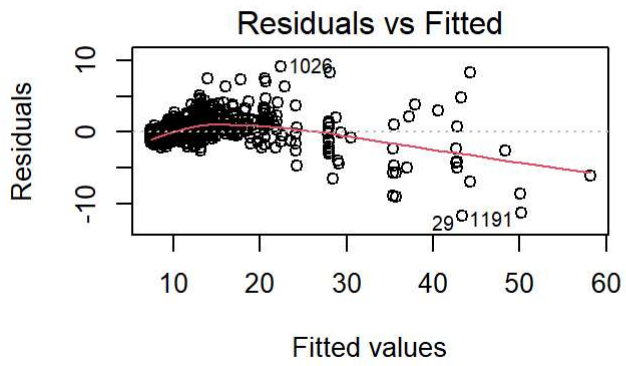

Residuals vs Leverage
Standardized residuals
15
5
-5
795
695
196
0.5
1
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
Leverage

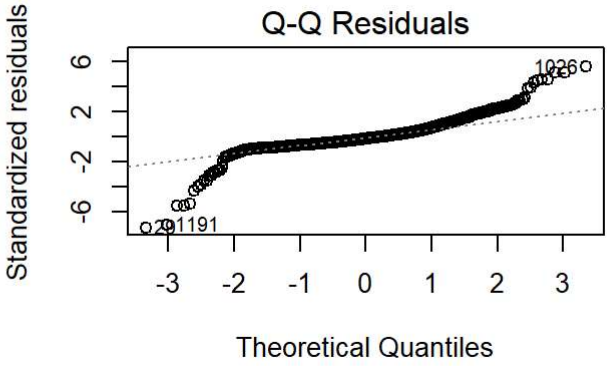
The scatterplot shows a nonlinear pattern and data points are fitted unevenly around fitted line. Vertical variation of the price increases as carat increases. The residuals plot also shows high vertical variation and uneven fitting around the line. Since the assumptions required for simple linear regression are not met we will attempt a transformation to linearize the data. We will attempt a transformation of the response variable first to address the vertical variation.

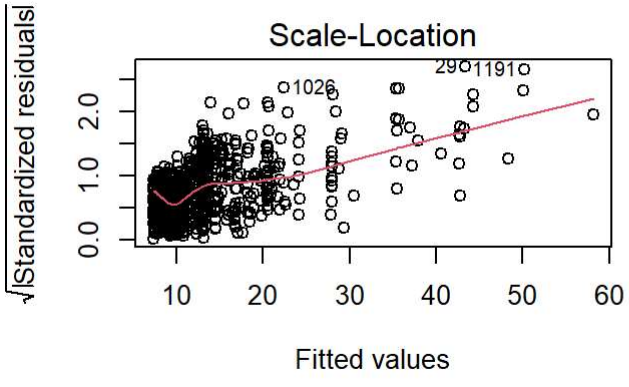


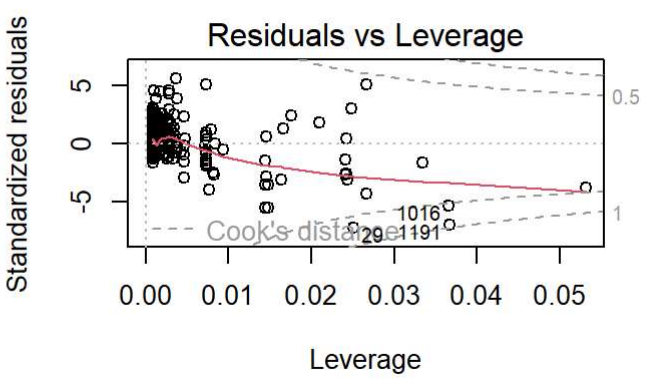
The boxcox suggests a lambda value of $(0.31) * pricestar = price^{0.31}$


Pricestar by Carat
Pricestar
Carat
60
40
20
0
2
4
6

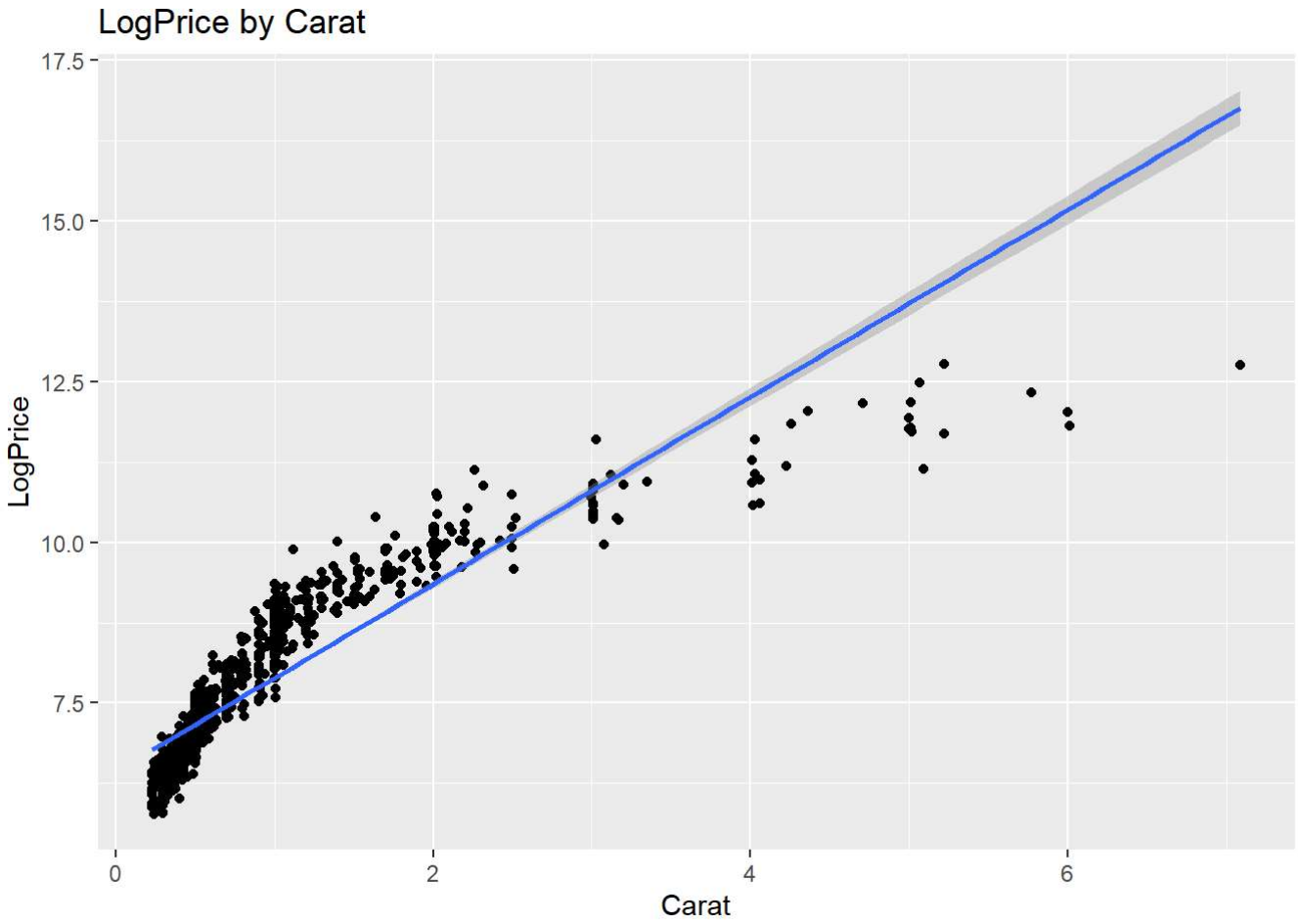

Residuals vs Fitted
Residuals
Fitted values
1026
29
1191
10
0
-10
10
20
30
40
50
60

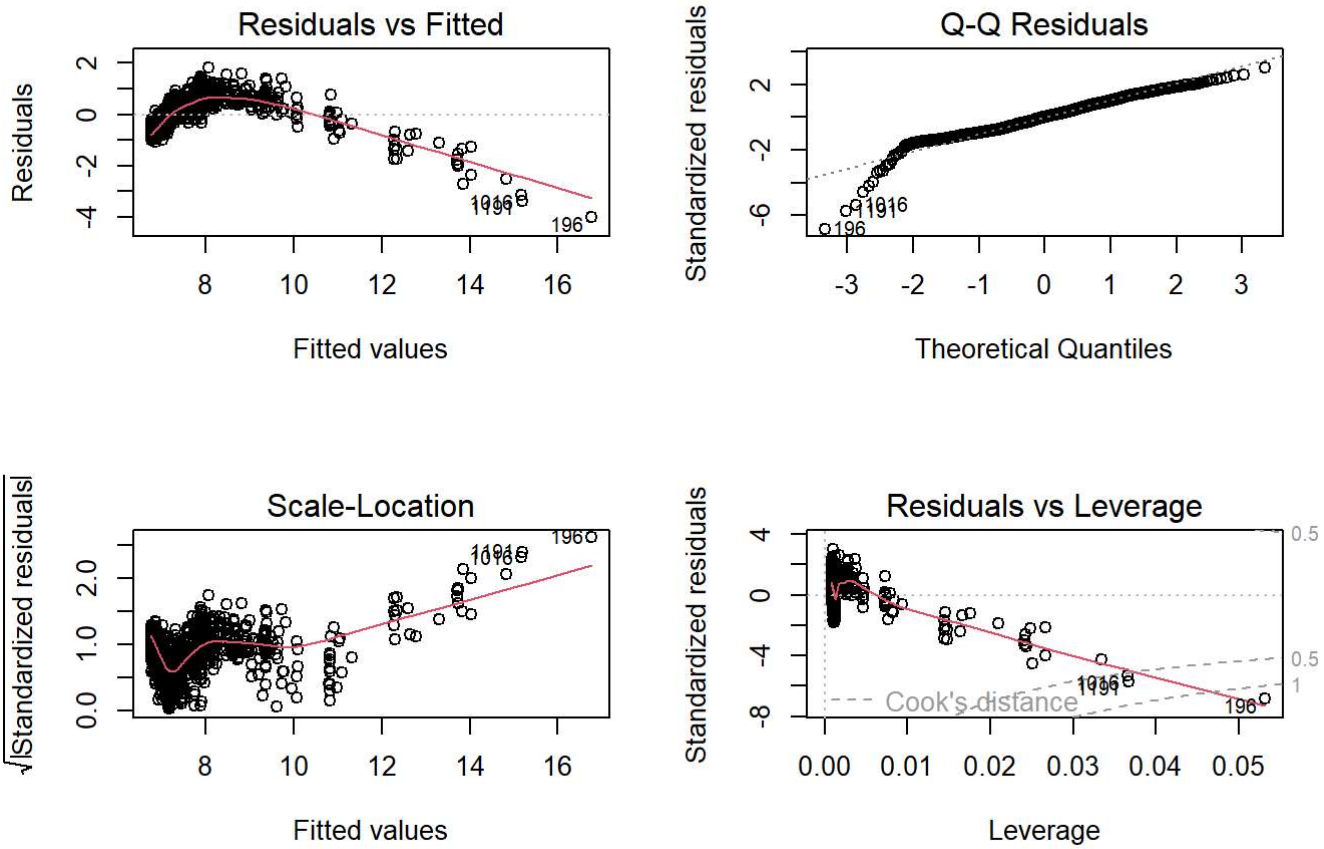

Q-Q Residuals
Standardized residuals
Theoretical Quantiles
1026
29
1191
6
2
-2
-6
-3
-2
-1
0
1
2
3


Scale-Location
√|Standardized residuals|
Fitted values
1026
29
1191
2.0
1.0
0.0
10
20
30
40
50
60

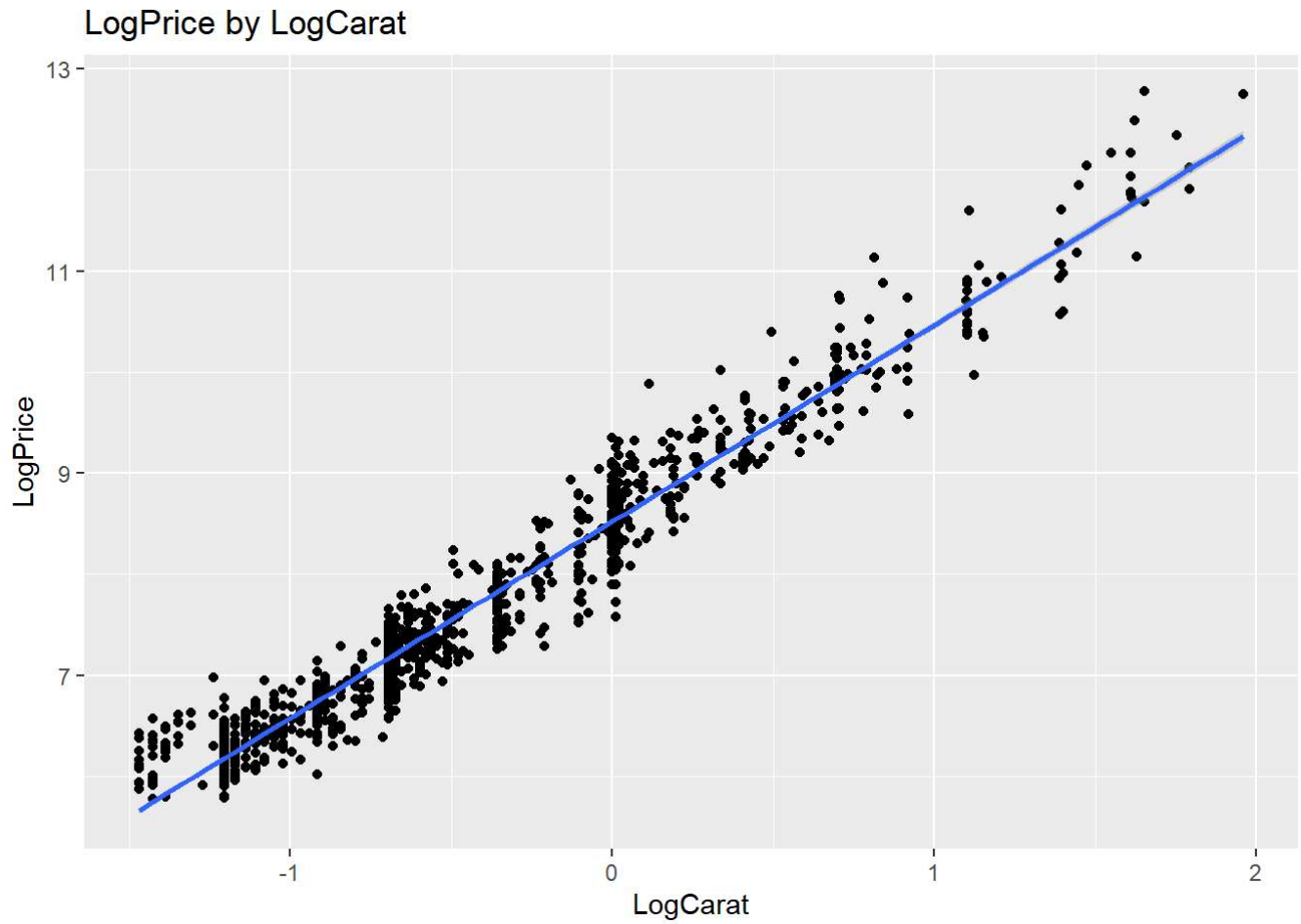

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
1016
29
1191
5
0
-5
0.5
1
0.00
0.01
0.02
0.03
0.04
0.05

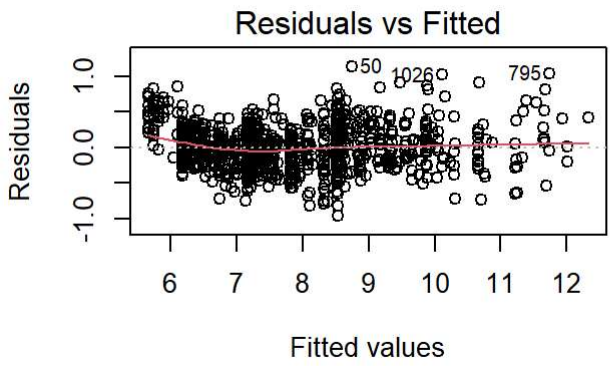
The pricestar scatterplot and residuals plot show improvement of the vertical variation, however the assumption is still not satisfied. To further dampen the variation we chose a smaller lambda value: $\lambda = 0$, which is a log transformation.

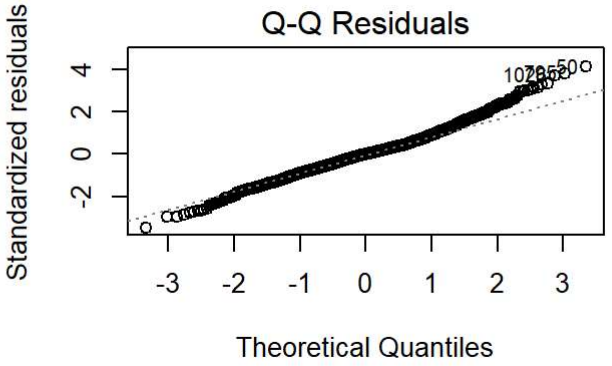


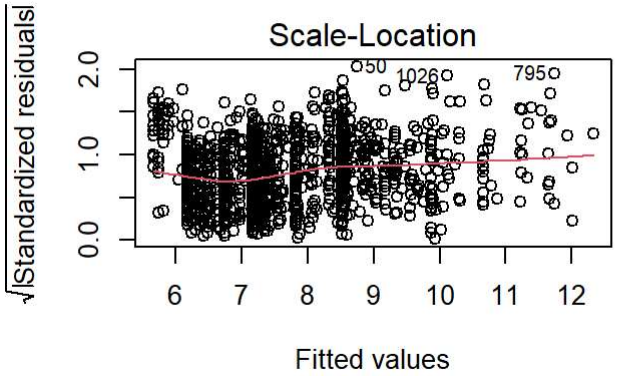


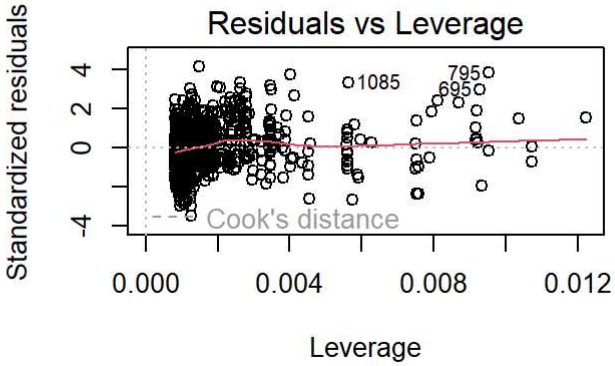
The scatterplot and residuals plot show that the vertical variation is constant across the x-axis. Since this assumption is met we can now address the nonlinearity of the data. To do so, we will perform a log transformation on the predictor carat.


LogPrice by LogCarat
LogPrice
LogCarat
13
11
9
7
-1
0
1
2


Residuals vs Fitted
Residuals
Fitted values
50
1026
795


Q-Q Residuals
Standardized residuals
Theoretical Quantiles


Scale-Location
√|Standardized residuals|
Fitted values
50
1026
795


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
1085
795
695

Both the scatter plot and residuals plot show linear data spread evenly across the regression line. The data has constant vertical variation along the x-axis. Because the basic assumptions for simple linear regression are met we can evaluate the model.

```
##
## 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
```

From the summary table we can create the regression equation. $logprice = 8.521208 + 1.944020 * logcarat$

$logprice = log(price) logcarat = log(carat)$

To test the validity of our model we evaluate an F test.

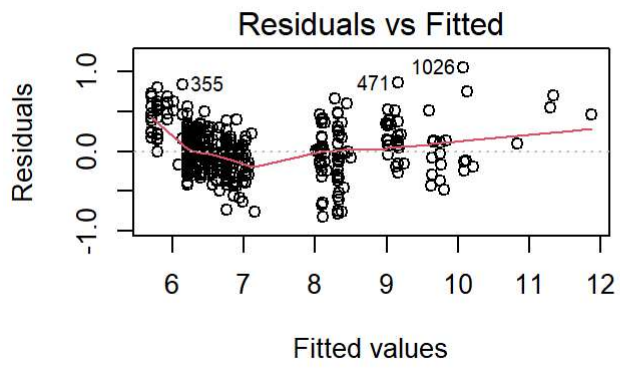
The null hypothesis is $H_0 : \beta_1 = 0$. There is no linear relationship between logcarat and logprice.

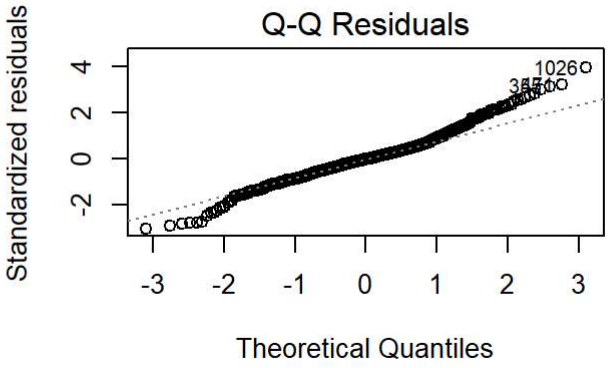
The alternative hypothesis is $H_a : \beta_1 \neq 0$. There is evidence of a linear relationship between logcarat and logprice.

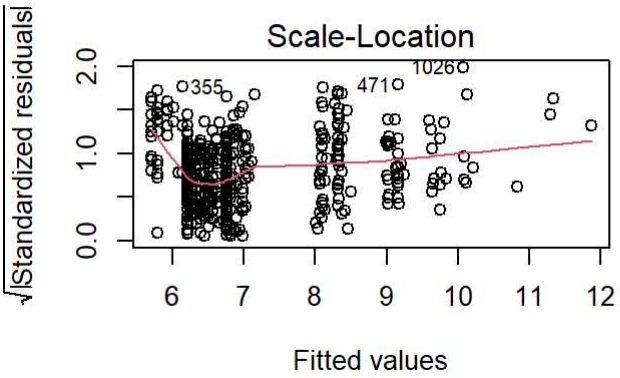
The p-value of the F test from the summary is $< 2.2 * 10^{-16}$ This p-value is much smaller than 0.05 so we reject the null hypothesis. There is evidence to support the linear relationship between logcarat and logprice.

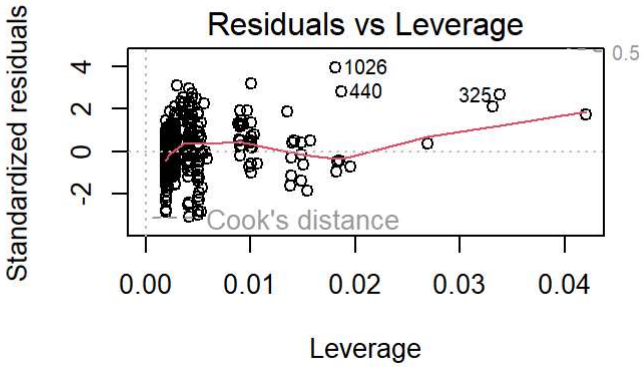
Finally, the adjusted R^2 value is 0.9546, this is a high number indicating a high linear correlation.

Figure 7 shows diamonds split into 2 categories: `high` and `low`. Blue Nile claims that diamonds bought just below half-dollar and whole-dollar values are a better deal than diamonds just above half and whole-dollar values. High diamonds are those with carat values from x.00 to x.25 and x.50 to x.75, where the term “high” refers to the price of the diamond against those that are slightly higher or lower in carat weight. Low diamonds have carat values from x.25 to x.50 and x.75 to x.00. For example: given that a 2.61 carat diamond is essentially a 2.5 carat diamond but you are paying for the weight of the diamond, a 2.61 carat diamond would be overvalued or on the `high` end price wise compared to the 2.5 carat diamond, conversely for Low diamonds in their respective range. To evaluate the Blue Nile’s claim we compare the mean price/carat of the low category against the mean price/carat of the entire set of diamonds.










```
##
## Call:
## lm(formula = logprice ~ logcarat, data = Data, subset = carat_cat ==
##     "low")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82321 -0.15829 -0.00852  0.12859  1.05010
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.51512    0.02084  408.69   <2e-16 ***
## logcarat     1.91888    0.02092   91.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2689 on 520 degrees of freedom
## Multiple R-squared:  0.9418, Adjusted R-squared:  0.9417
## F-statistic:  8411 on 1 and 520 DF,  p-value: < 2.2e-16
```

We perform a t-test comparing β_1 from low_carat and β_1 of the entire sample.

$H_0 : \beta_1 = 1.944020$ The null hypothesis is that the price/carat for low diamonds is 1.944020.

$H_a : \beta_1 \neq 1.944020$ The alternative hypothesis is that the price/carat for low diamonds is not 1.944020

```
## [1] -1.702471
```

The t-statistic is 1.702471

```
## [1] 1.964467
```

The critical value is 1.964467

Because the t-statistic is less than the critical value, we fail to reject the null hypothesis. The data supports that the low diamonds price/carat is not different from the entire sample. We conclude that Blue Nile's claim is inaccurate.