STAT 6021 Project 1

Group 14: Rose Eluvathingal Muttikkal (re4mfy), August Lamb (abl6ywp)
Thomas Burrell (tmb9ccd), Elisabeth Waldron (psa7rm)

2023-06-29

(1) Summary of Findings

According to Blue Nile, one of, if not *the* most important factors in determining the value of a diamond is its cut. The cut of the diamond can affect the perception of both its color and clarity, meaning that a "low quality" diamond in both of those categories, when well-cut, can actually be quite valuable. While Blue Nile claims that this is the quality that most affects price, there does not seem to be strong evidence to back this up. It is possible that the interaction between cut and one or more other variables could significantly sway price, but this is difficult to measure due to the subjectivity in opinion of whether or not the cut sufficiently disguises the less desirable characteristics of other factors. Additionally, in our analysis we discovered that there is a heavy skew towards one level of cut in this dataset, so we are hesitant to draw conclusions without more balanced data.

The second most important factor in the value of diamonds is the color, according to Blue Nile. The color of diamonds refers to how colorless they are and is an important factor when customers decide what they're going to buy. According to Blue Nile, "the absence of color in a diamond is the rarest and therefore, the most expensive." Blue Nile claims that diamonds that are on the more colorless side of the color spectrum should be more expensive, while diamonds with color or slight hues will be less expensive. After analysis of the dataset provided by Blue Nile, this claim did not seem to hold, as the color of the diamond has almost no effect on the price, even when looking at the absolutely colorless 'D' diamonds.

Clarity is another important factor to consider when purchasing diamonds. It is a measure of imperfections within and on diamond surfaces. Blue Nile has made several claims about clarity. They have claimed SI clariy-grade and VS clarity-grade diamonds offer the best value and are the most popular choice. Using the dataset provided by Blue Nile, we explored the frequency of diamond clarity grade, the relationship between clarity and price, and the influence of other factors on the relationship between clarity and price. The analysis supported the claim of SI and VS diamonds being the most popular choice because 70% of the diamonds on sale at Blue Nile are SI and VS diamonds. However, the analysis did not find evidence supporting the claim that SI and VS diamonds offer the best value.

The carat of a diamond refers to the weight of the rock, rather than its overall size. The weight of a carat is seen as being the most important factor in which people believe a diamond's monetary value is set, due to commercialization and capitalism. In reality, there are multiple determining factors such as the color and cut of the diamond that determine a diamond's overall value, according to the Blue Nile. But above all, Blue Nile agrees, if all other variables are in good standing, the carat weight has the largest effect on how the price is outputted.

There are many important factors when customers choose a diamond appropriate for them. For most people however, the price is by far the most important factor. Diamond prices can range from hundreds of dollars to millions of dollars. To investigate what aspects of the diamond affects the price the most, we explored the data through visualization, examined and challenged Blue Nile's claims, and drew conclusions. We found that the carat weight of a diamond is a very strong predictor of the response variable, price. For every one percent increase in the carat weight of the diamond, the predicted price increases by almost two percent. Customers can use this information to gauge how expensive a diamond should be based on its carat weight.

(2a) Description of dataset

Data

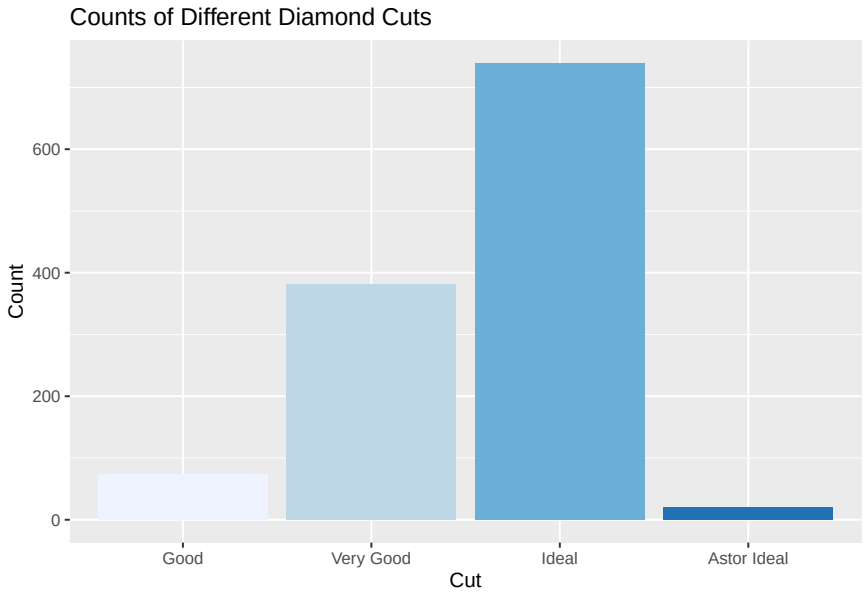
In the Blue Nile diamonds dataset, there are 1214 observations and five variables (**Cut**, **Color**, **Clarity**, **Carat**, **Price**)

Variables

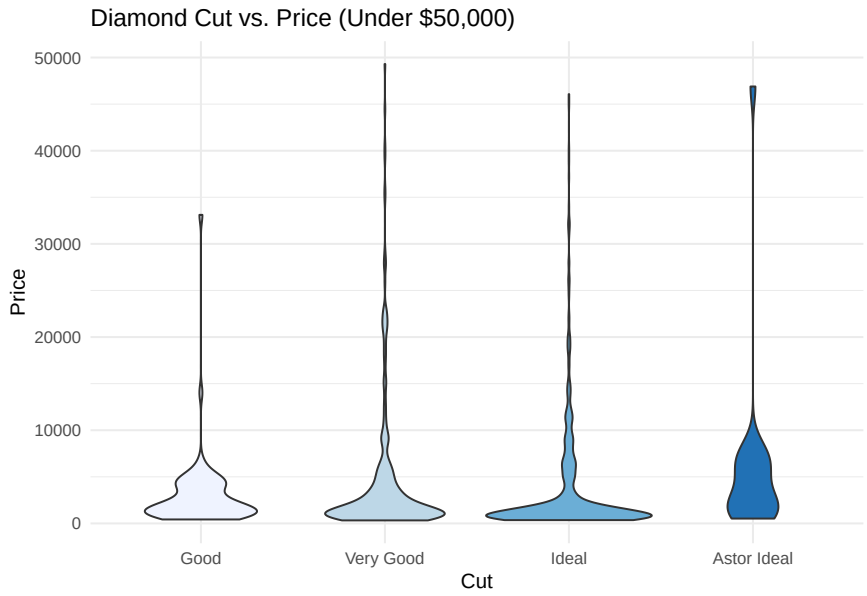
- **Cut**: A categorical variable describing how well-proportioned the dimensions, surfaces, and facets of a diamond are in order to produce the most desirable levels of sparkle and brilliance. From lower quality to higher quality, the levels are `Good`, `Very Good`, `Ideal`, and `Astor Ideal`.
- **Color**: A categorical variable describing how colorless a diamond is along a standardized diamond color chart. From most colorless to least colorless, the levels are `D - Absolutely Colorless`, `E - Colorless`, `F - Colorless`, `G - Near Colorless`, `H - Near Colorless`, `I - Near Colorless`, `J - Near Colorless`.
- **Clarity**: A categorical variable describing degree of imperfections within and on diamond surface. From lower quality to higher quality, the levels are `Slightly Included` (SI), `Very Slightly Included` (VSI), `Very Very Slightly Included` (VVSI), `Internally Flawless` (IF), and `Flawless` (FL). A clarity grading of 1 denotes higher clarity.
- **Carat**: A continuous quantitative variable that refers to a diamond's weight, with one carat being 200 milligrams.
- **Price**: A continuous integer variable. The cost in dollars of each diamond.

(2b) Visualizations

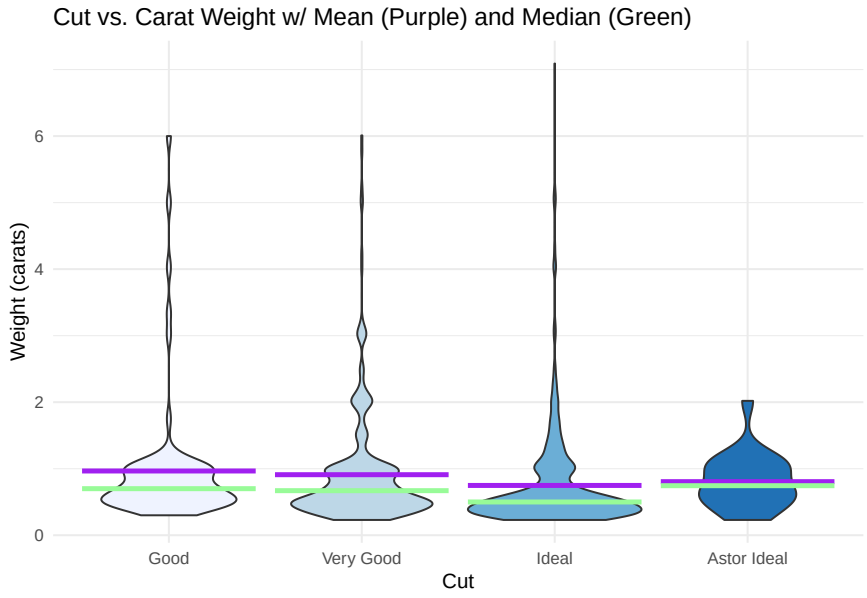
(2b.i) Cut Variable



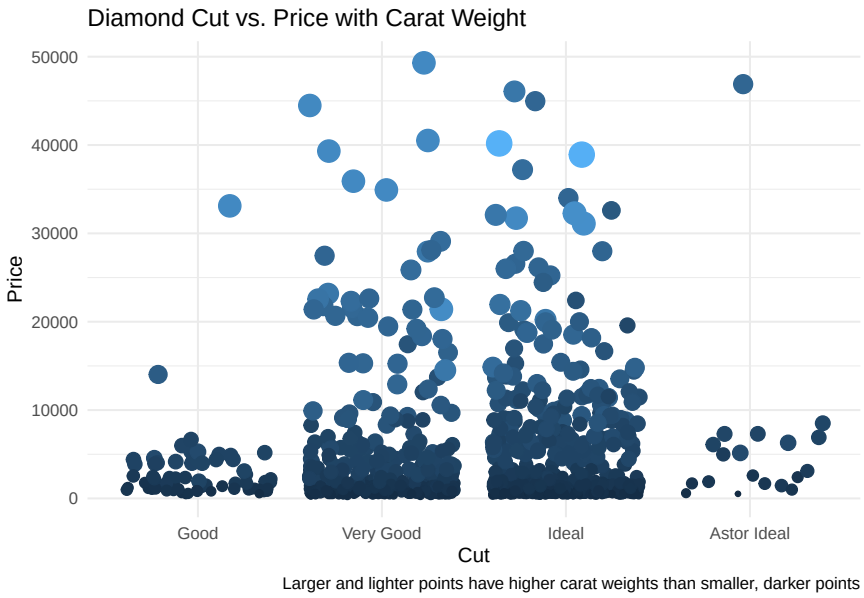
This dataset is not very well balanced by cut. There is a much larger number of `Ideal` cut diamonds compared to all others. There are very low numbers of `Good` and `Astor Ideal` cut diamonds. This is going to make drawing conclusions based on cut alone difficult.



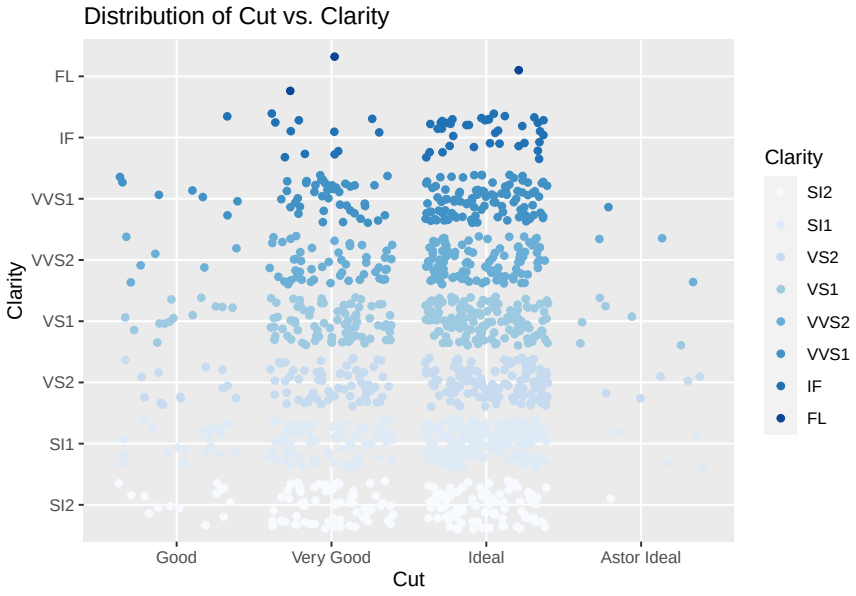
These violin plots show the density of diamonds at certain price levels based on their cut. Because there are a number of outliers in the first graph, I limited the dataset to entries under $50,000 for the second graph in order to improve the visualization. Blue Nile's claim that `Cut` has a significant effect on diamond quality, and thereby price, does not seem to hold here. There are similar densities at certain price points for each cut, but based on their claim one would expect higher densities at higher price points for better cuts. This plot implies that it is just as easy to get a `Good` diamond at $2,000 as it is to get an `Astor Ideal` diamond at $2,000. Additionally, it is important to note that the skewed `Cut` data will have an effect on this visualization, as there are over 700 `Ideal` cut diamonds and only 20 `Astor Ideal`.



There appears to be similar densities of diamonds of a certain weight in each `Cut` category. The one interesting feature I notice here is that `Ideal` cut diamonds have lower average and median weights, possibly from removing more of the diamond by weight to achieve the desired quality. However, because `Ideal` is the largest sample by far, the data is heavily skewed in its favor, so this conclusion may not be correct.



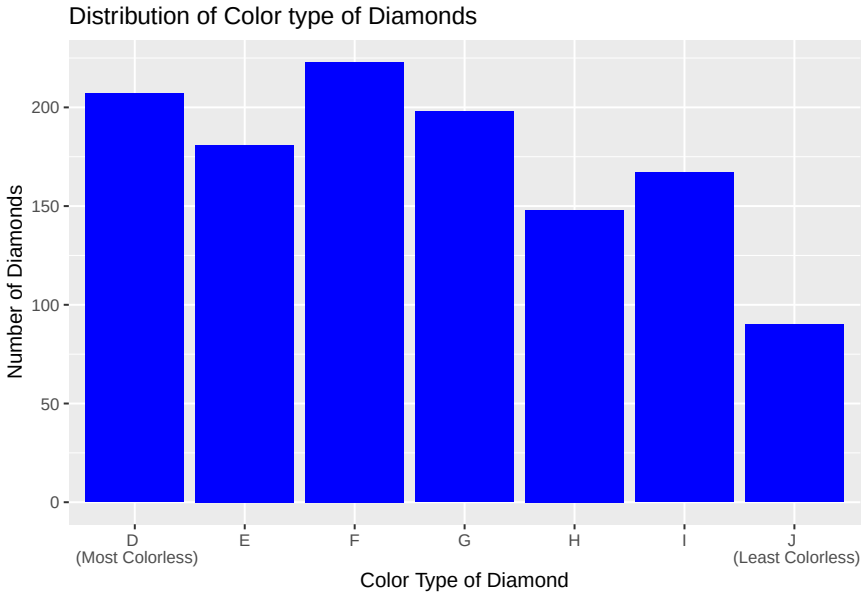
This plot confirms that diamonds of a larger `Carat` weight and higher `Cut` quality have higher prices.



Because there are similar distributions of diamonds with a certain clarity among the different cuts, with a peak in the SI-VS region, there doesn't seem to be a significant relationship between cut and clarity. Blue Nile claims that a well-cut diamond can increase the value of a lower-clarity diamond by disguising the blemishes, however this must be a purely visual improvement, not an actual improvement of clarity.

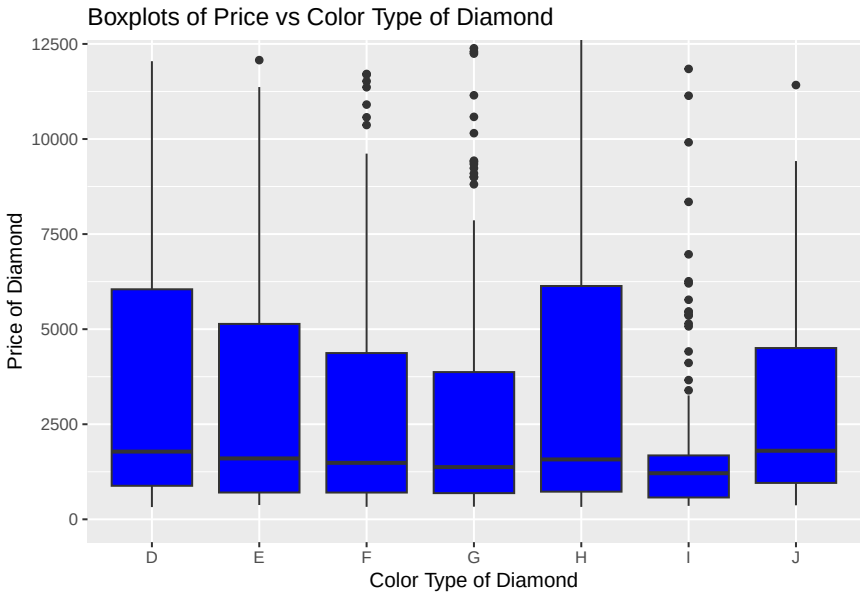
(2b.ii) Color Variable

Summarizing Categorical Variable (Color Type) using a Bar Chart:



In the Blue Nile dataset, there is a pretty uniform distribution of diamond color types, except one color type, with J having the least amount of diamonds. This graph shows that Blue Nile carries a diverse range of colors and in high quantities. It was a bit surprising to see that their most 'rare' and 'expensive' diamond (D) has the second highest quantity among all colors.

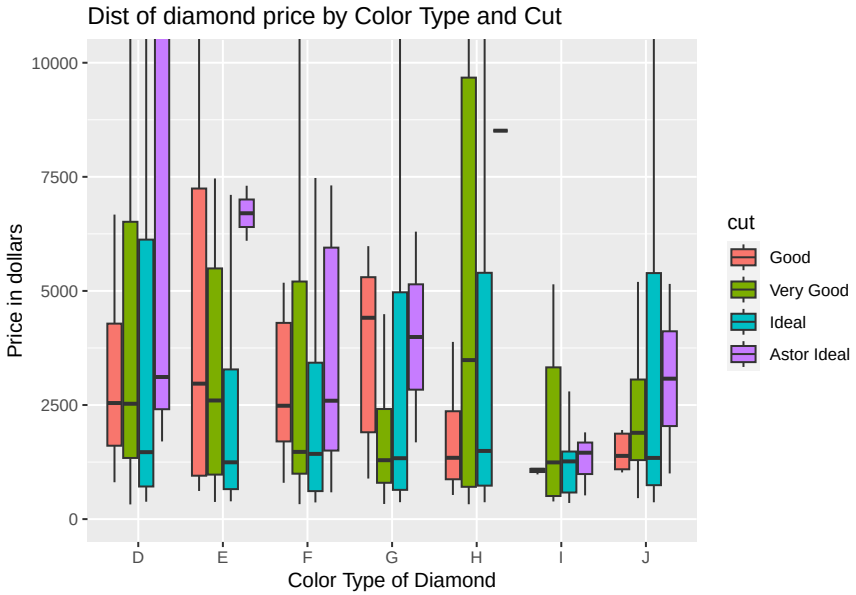
Summarizing Categorical Variable (Color Type) vs. Quantitative Variable (Price) using a Boxplot



The boxplot presented above displays the relationship between the price of a diamond and the color type. We can see that the median price of a diamond does not vary by more than a few hundred dollars even if the

color type is different. The claim from Blue Nile that price is very dependent on the color doesn't seem very substantial based on this plot, but we will continue to investigate through more exploratory data analysis.

Distribution of diamond Price by Color Type and Cut

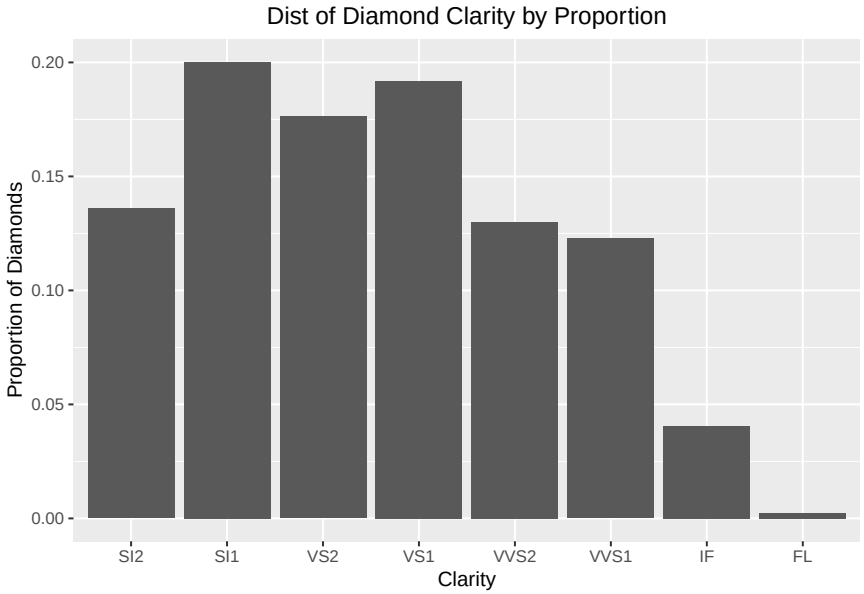


When displaying the price of diamonds by color type and cut, we can see that the cut is very important for the price of the diamond, across all color types. The 'Astor Ideal' cut consistently has the highest median price across all color types (except G).

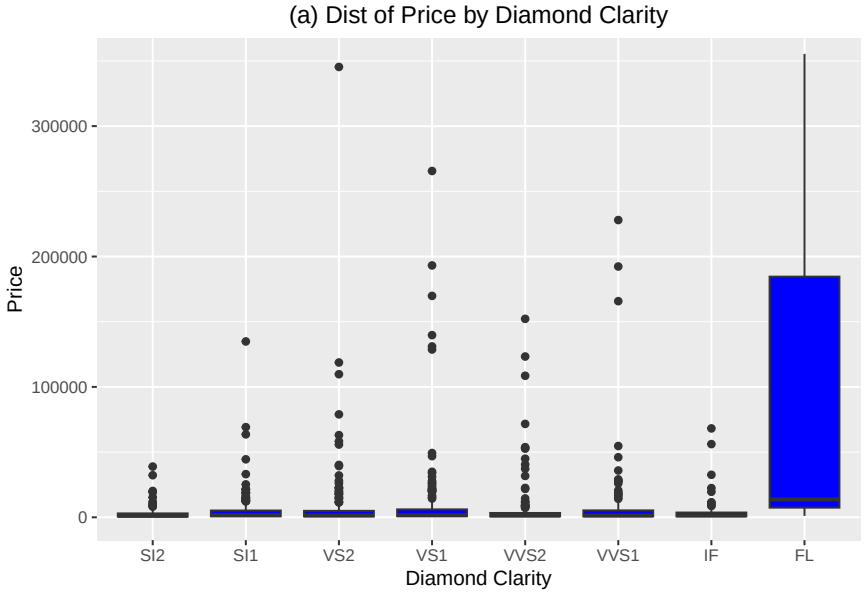
(2b.iii) Clarity Variable

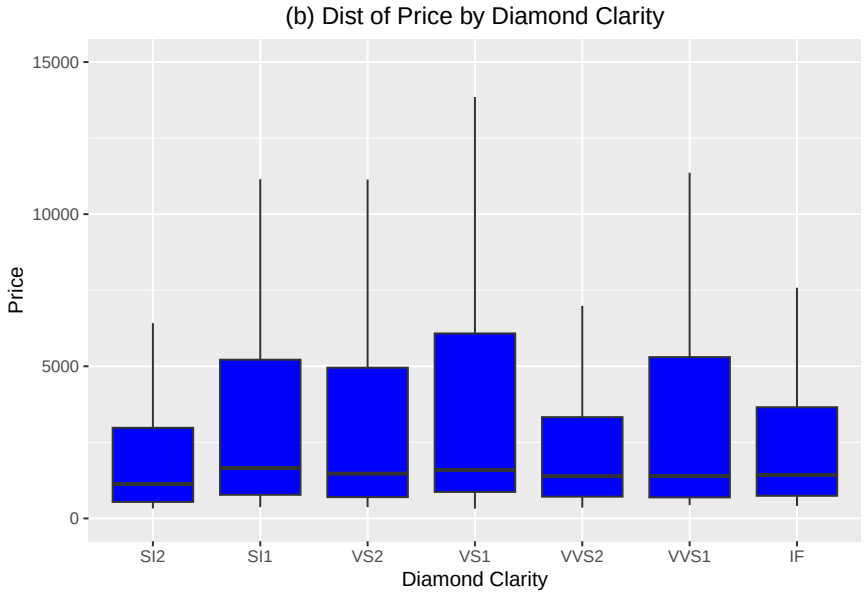
Clarity is used to classify imperfections in diamonds. Surface flaws are blemishes, while internal defects are inclusions. Eye-clean diamonds have small inclusions only visible with magnification. Clarity has 6 categories: Included (I), Slightly Included (SI), Very Slightly Included (VSI), Very Very Slightly Included (VVSI), Internally Flawless (IF), and Flawless (FL) diamonds. A clarity grading of 1 denotes higher clarity. SI diamonds have noticeable inclusions at 10x magnification, VVS diamonds appear eye-clean, IF diamonds are eye-clean, and FL diamonds have no internal or external characteristics. Blue Nile does not sell I clarity diamonds for engagement rings; they were not part of the dataset.

Blue Nile claims: "SI Diamonds And VS Diamonds Are The Best Value", "VS is the most popular choice", "43% of all customers buy VS diamonds," "30% of all customers buy SI diamonds." According to Blue Nile, SI and VS diamonds are not only affordable but also visually comparable to higher clarity grades, making them a compelling choice for customers. Blue Nile emphasizes this option by highlighting that a majority of customers buy SI and VS diamonds. Using the given dataset, we can explore the frequency of diamond clarity grade, how clarity relates to price, and see if any other factors come into play when examining clarity and price.



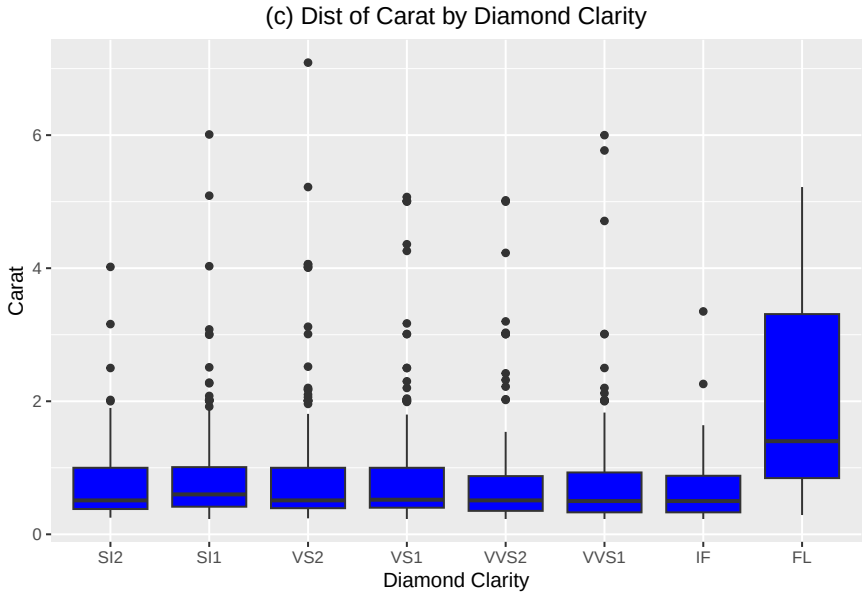
Percentage distribution of diamonds for sale on Blue Nile based on clarity grades, arranged in ascending order. Approximately 70% of the diamonds available for sale are SI or VS diamonds, as depicted in the figure. VS diamonds comprise the largest share of the diamonds for sale. The figure provides compelling evidence to support Blue Nile's claim of customer preference for SI and VS diamonds, as it shows a significant portion of their inventory is comprised of SI and VS diamonds.



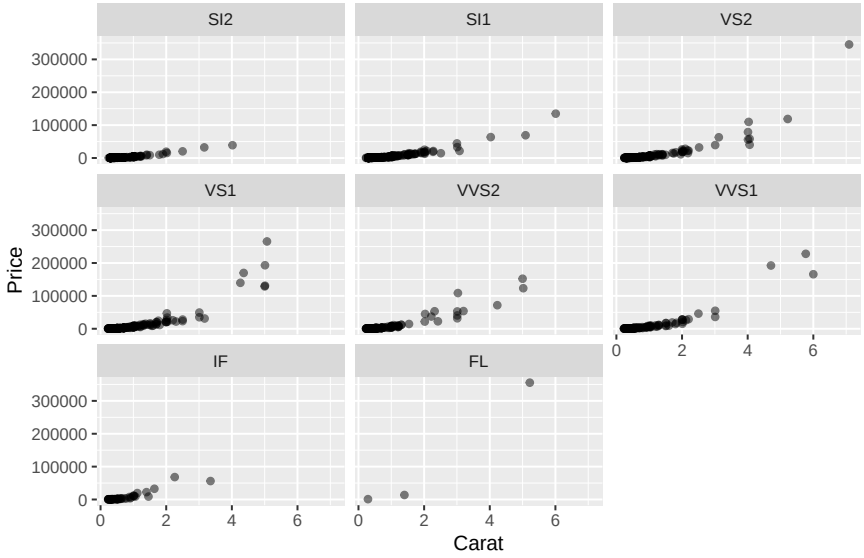


Comparison of price across diamond clarity using side-by-side boxplots. (a) The median and quartile range of price for FL diamonds is higher than all other clarity grades. (b) The median price remains similar across clarity grades SI2-IF. Outlier values and FL diamonds were excluded to provide a closer examination of the other clarity grades. The figure challenges the claim of SI and VS diamonds as having the best value since higher clarity grades, except FL, exhibit similar median prices.

The following data visualizations explore the influence of other factors on the relationship between price and clarity.



(d) Scatterplot of Diamond Price against Carat by Clarity

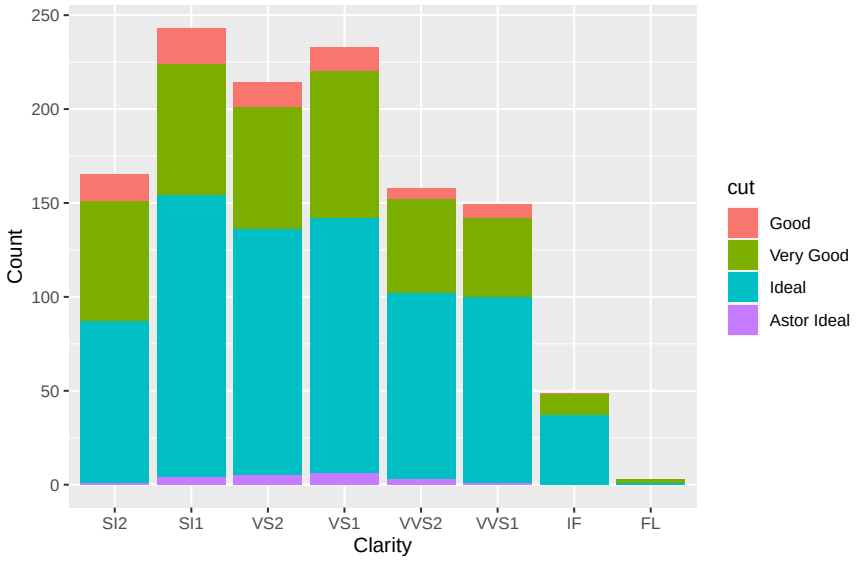


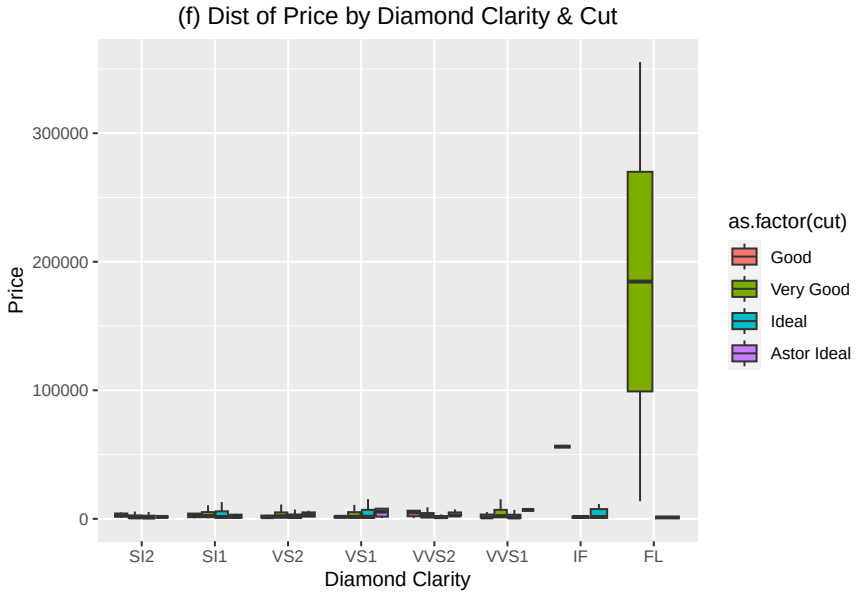
Comparison of price across diamond clarity by carat. (c) The side-by-side boxplot is a comparison of carat across clarity. The median and interquartile range of carat for FL diamonds surpasses all other clarity grades. The distribution of carat across clarity closely mirrors the distribution of price across clarity, with FL diamonds showing a higher median carat, similar to the trend observed in price. (d) The scatterplot depicts the carat on the x-axis and the price on the y-axis, with the data segmented by clarity. Observing SI2-VVS1 diamonds within 2 carats, the price remains similar.

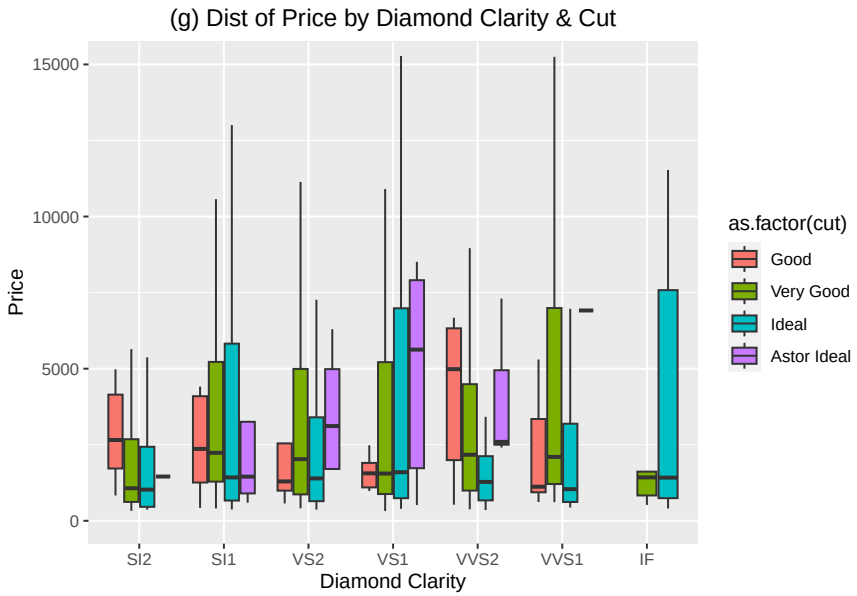
The figure challenges the claim that SI and VS diamonds offer the best value, as higher clarity grades within 2 carats exhibit comparable prices. When customers consider their diamond purchase, this figure indicates that they can opt for a higher clarity grade, up to VVS1, at a similar price to SI and VS diamonds, provided they remain within the 2-carat range. Notably, this size falls well within the average diamond size for engagement rings in the US, which is typically 1 carat (Krueger, 2022).

Krueger, A. (2022, October 20). This is the average carat size for a diamond engagement ring. Brides. https://www.brides.com/average-diamond-size-for-engagement-ring-5179259

(e) Diamond Clarity by Cut



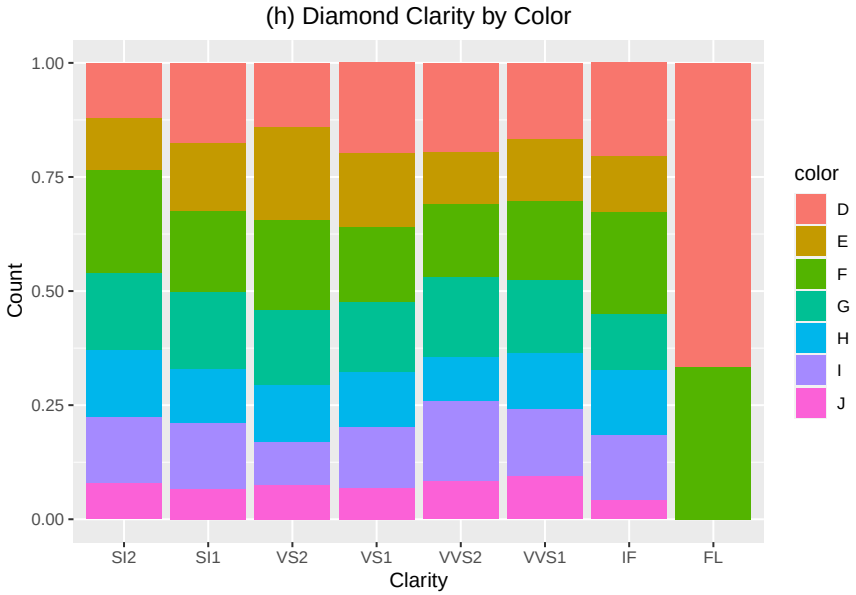


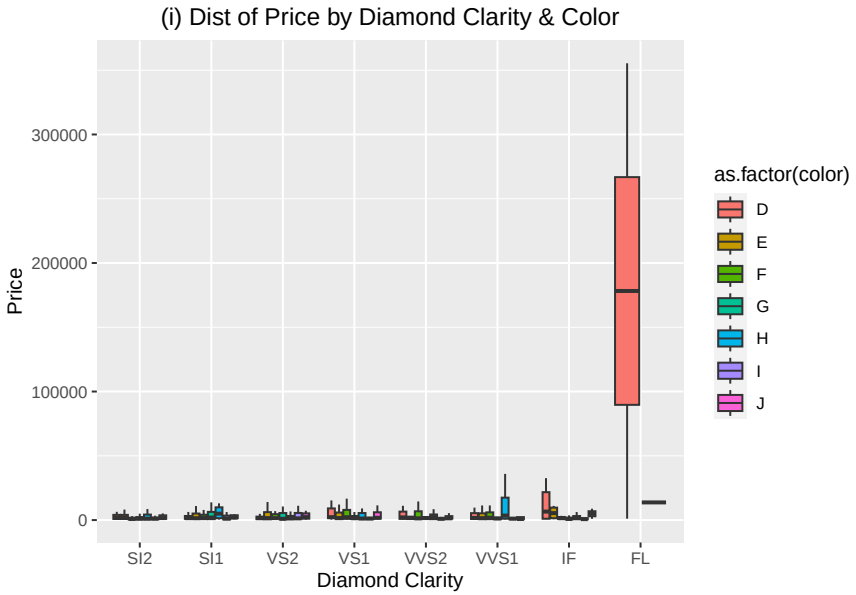


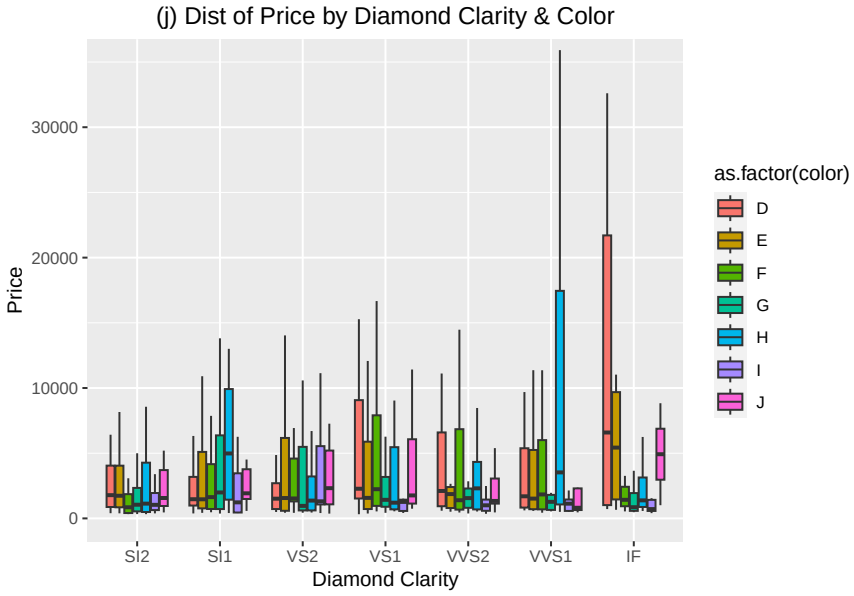
Comparison of price across diamond clarity by cut. (e) The bar chart displays the distribution of different cut types across clarity grades. The Ideal cut dominates across all clarity grades, representing the majority proportion. Notably, the highest quality Astor Ideal cut is limited to the VVS1 diamonds. Furthermore, FL clarity grade is available only in Very Good and Ideal cuts. (f) The side-by-side box plot depicts the clarity on the x-axis and the price on the y-axis, segmented by cut. Outliers were excluded for better visualization. The FL diamond with Very Good cut for exhibits a significantly higher median price compared to other clarity grades, surpassing even the Ideal cut for FL diamonds, which is technically a higher quality cut. (g) To examine the other clarity grades in detail, FL diamonds were excluded. Across all clarity grades, the median price of the Ideal cut is either equal to or lower than the median price of the Very Good cut. Interestingly, a VVS2 diamond with Astor Ideal cut displays a lower median price than VS diamonds with Astor Ideal cut. Moreover, this median price appears equivalent to the 75th percentile price of an SI2 diamond with a Very Good cut, meaning 25% of SI2 diamonds with Very Good cut are more expensive than 50% of VVS2 diamonds with Ideal cuts.

The figure challenges the claim that SI and VS diamonds offer the best value since higher clarity-grade

diamonds with superior cut quality can have lower median prices. Customers prioritizing high-quality cuts may find VVS2 diamonds with Astor Ideal cut to be the best value, as they offer higher clarity grades and superior cut quality with a lower median price than VS diamonds. On the other hand, customers prioritizing high clarity grades may consider IF diamonds with Ideal cut to be the best value, as their median price is comparable to SI1 and VS2 diamonds with the same cut.







Comparison of price across diamond clarity by color. (h) The bar chart illustrates the distribution of color grades across clarity, ranging from D (most colorless) to J (least colorless). Blue Nile claims that D grade diamonds are the most "rare" and "expensive." The majority of FL diamonds are D grade, while other clarity diamonds exhibit a relatively even distribution of color grades. J grade diamonds appear to have the smallest proportion across clarity. (i) The side-by-side box plot depicts the clarity on the x-axis and the price on the y-axis, segmented by color. Outliers were excluded for better visualization. The FL diamond with D color grade exhibits a significantly higher median price compared to other clarity grades, similar to the trend observed with FL diamond with Very Good cut. (j) IF diamonds with D color grade have a much higher median price than other clarity-grade diamonds. However, VVS diamonds with D color grade appears to have a lower median price than VS1 diamond with D color grade. Additionally, The median price of VVS1 with E color grade also appears to be similar or lower than VS1 diamond with the same color grade.

(2b.iiii) Carat Variable

```
##
## 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 <0.0000000000000002 ***
## carat        25333.9      494.4   51.24 <0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```

Linear model equation: y = 25334x - 13551, further indicating a positive correlation and an increasing linear

line due to the the y - intercept being negative when x is 0.

B hat 1: 25333.9-> Increase in price as carat weight increases

B hat 0: -13550.9-> No indicated carat weight predictor price would be $13,550.9, meaning there is a positive correlation indicated.

s: 13560-> statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation. In statistics, a sample mean deviates from the actual mean of a population; this deviation is the standard error of the mean.

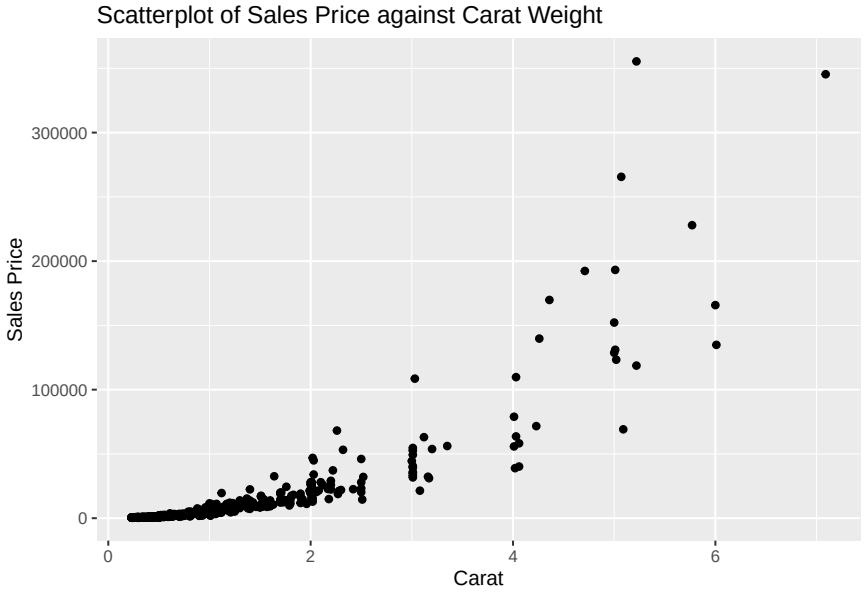
13,560 * 2 = +/-$27,120

F: 2625-> Alternative hypothesis is supported;there is a linear association

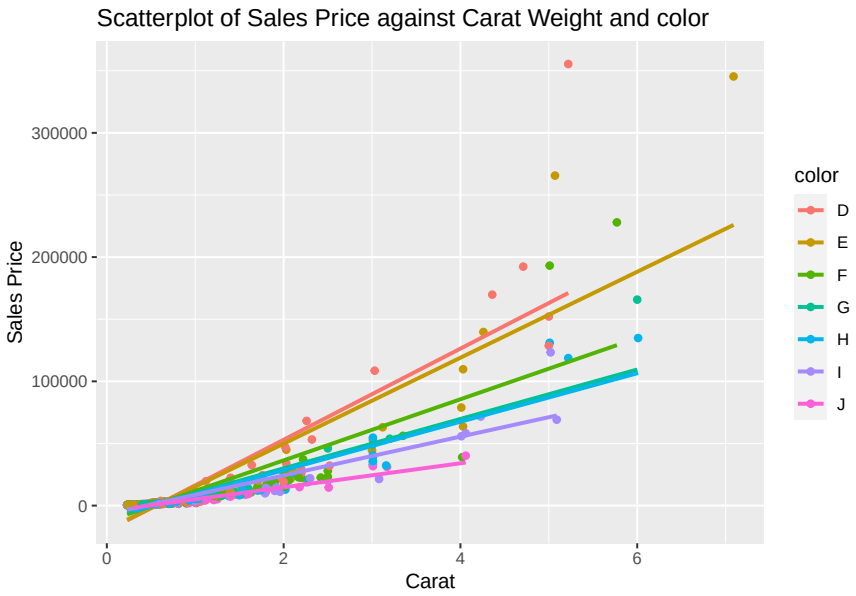
R^2: 0.6839-> 68.39% of the variation in price can be explained by carat weight.

p-value: < 2.2e-16-> reject the null hypothesis

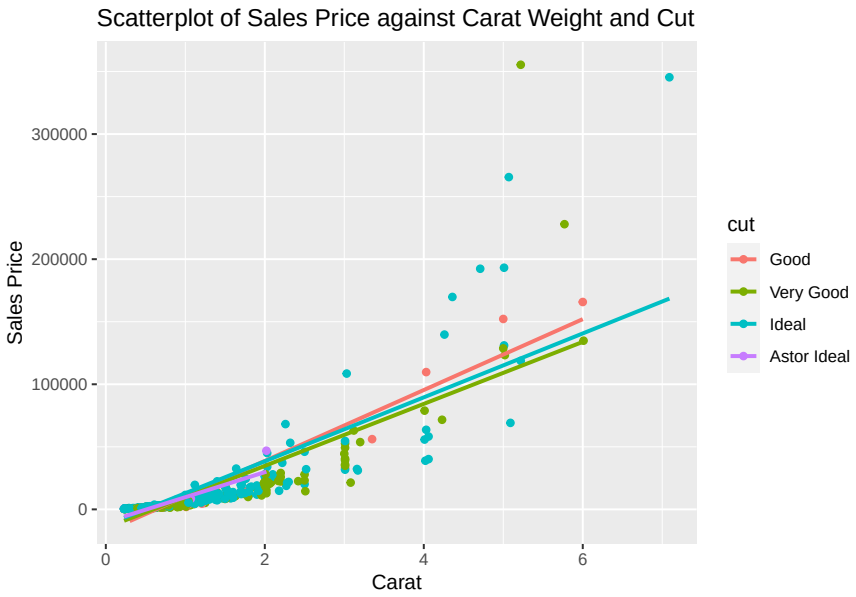
The correlation of Carat with Price is 82.71



There is a strong, positive, exponential, relationship between sales price and carat weight.



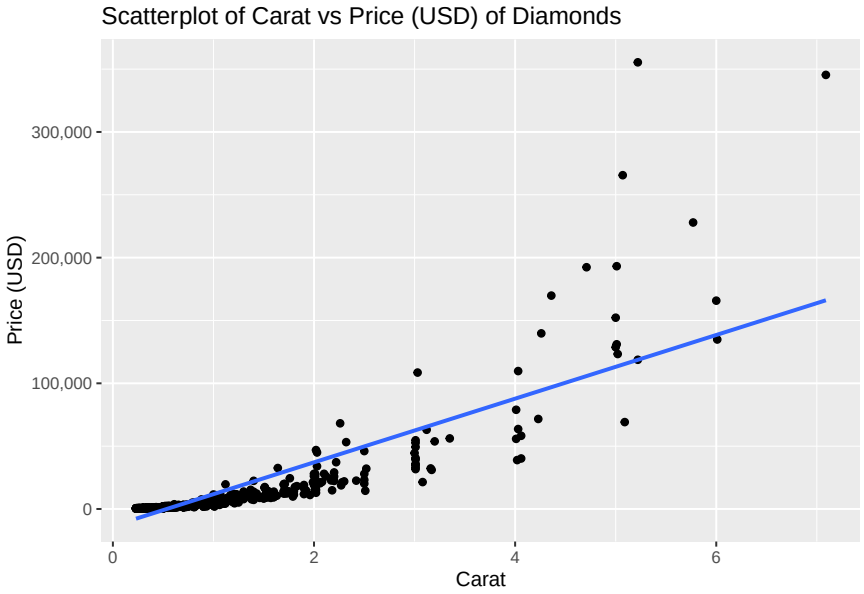
From this scatterplot with regression lines for each color, we can see that the more colorless a diamond is (More toward the D side of the color spectrum), the higher the price is.



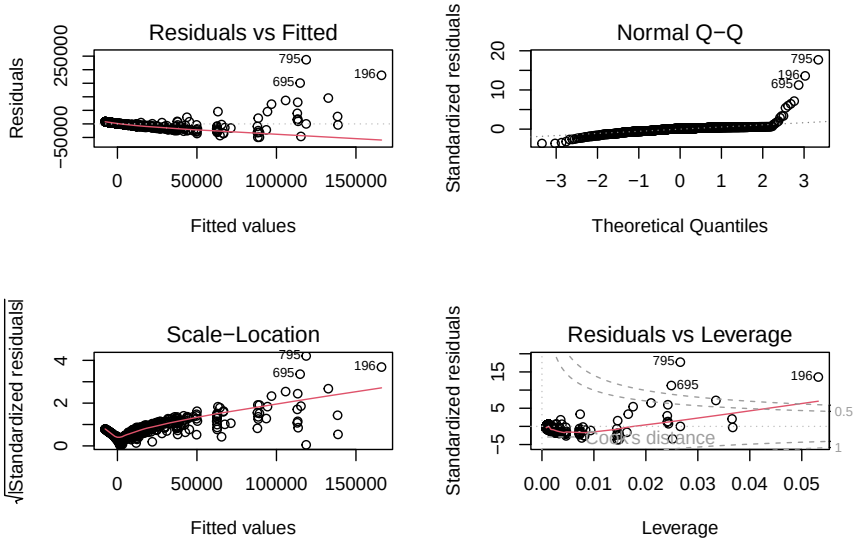
From this scatterplot with regression lines for each cut category, there is not much of a significant difference for each cut category against price and carat weight.

In conclusion, as the weight of the carat increases, the price increases, signifying a positive correlation between both variables. Also, the way in which the weight of the carat is either rounded up or down affects the pricing of the diamond significantly, within the thousands. Lastly, the relation of the qualitative variable, color, also played a significant role in how the price was selected in conjunction with carat weight.

(3) Analysis of Carat Weight on Price

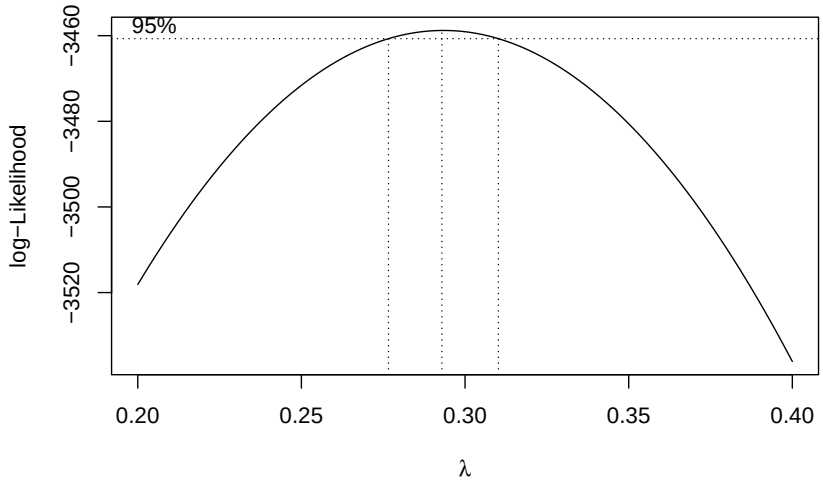


Looking at the scatterplot above, the data appears to follow an exponential trend rather than a linear trend. *Assumption 1* (errors have mean 0) appears to have been violated, meaning we could have biased predictions after fitting the model. To confirm our suspicion that *Assumption 1* has been violated, we will analyze the associated residual plot.



Diagnostic Plots

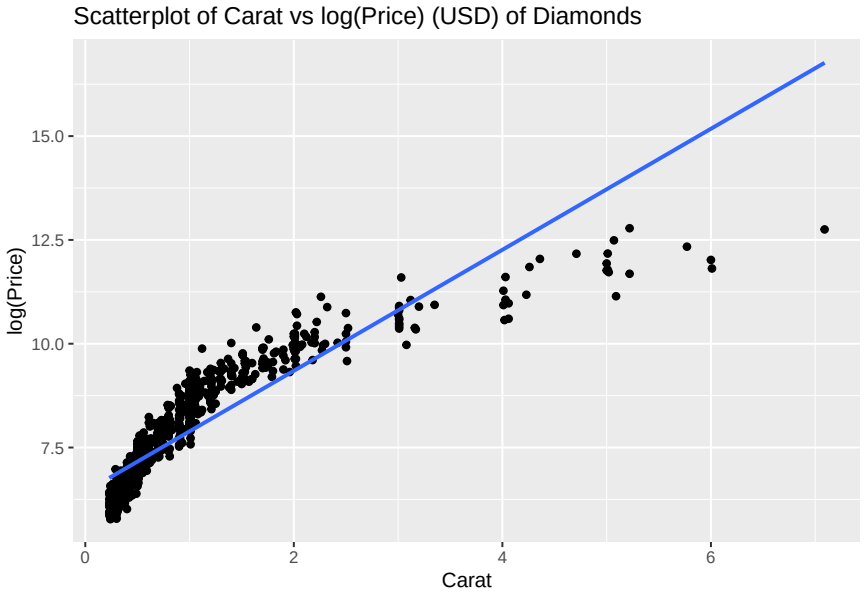
It is clear from the residual plot that both *Assumption 1* and *Assumption 2* (constant vertical variation) have been violated. The values are more centered above the residual line and and fan out into a horn shape. To resolve this, we will transform the response variable (`Price`). A Box-Cox plot will help decide what value of lambda should be chosen.



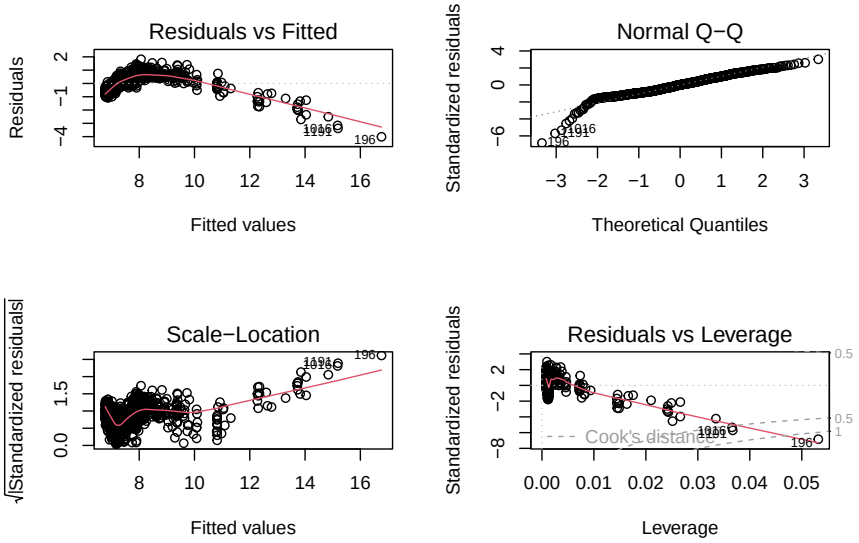
Box-Cox Plot

A Lambda value between 0.27 and 0.31 should be chosen to transform the `Price` variable. However, we should reconsider transforming using this lambda value because it will be difficult to give an interpretation of how `Carat` and `Price` are related. So despite the Box-Cox, we will logarithmically transform `Price`.

1st Transformation: $y \to \log(y)$



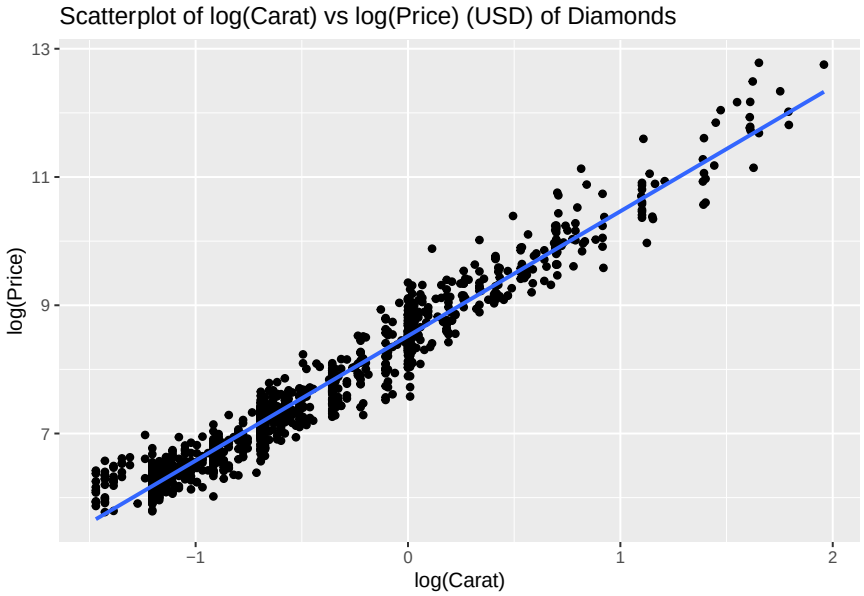
After transforming the response variable, `Price`, both *Assupmtion 1* and *Assumption 2* seem to be violated. We will proceed to analyze the residual plot to check assumptions.



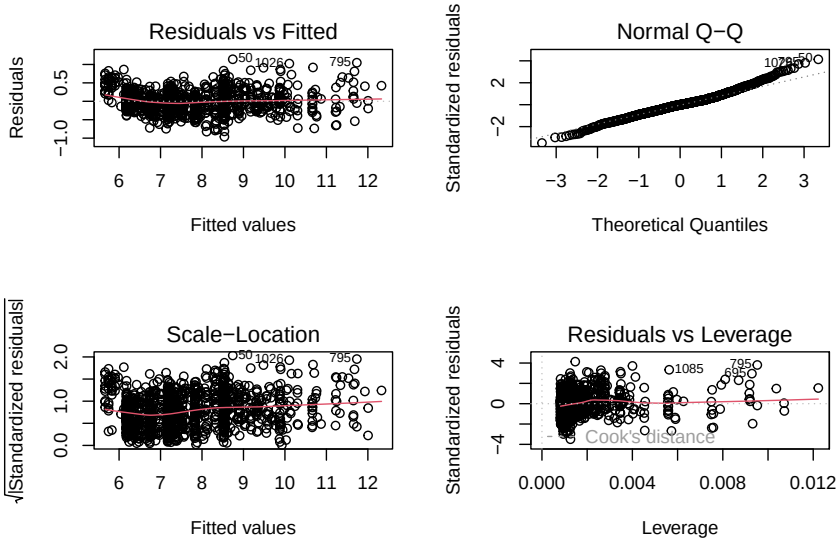
Diagnostic Plots

We can see from the residual plot that the variance is not constant, so we will logarithmically transform the x variable, `Carat`, to account for this. Also there is a clear pattern between the fitted values and the residuals.

2nd Transformation: $x \to \log(x)$



After logarithmically transforming the x (`Carat`) and y (`Price`) variables, *Assumption 1* appears to be satisfied; the data follows a linear pattern. Now, we will check the residual plot to evaluate the other SLR assumptions.



Diagnostic Plots

The log transformations of x (`Carat`) and y (`Price`) have proven to be a good decision, as the residual plot no longer violates the second assumption. There is an even number of values above and below the residual line. There is also no fanning out pattern that we saw in the previous residual plots. The QQ plot shows a fairly normal distribution for the observations.

Simple Linear Regression Output and Equation

```
##
## 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 <0.0000000000000002 ***
## log(carat)  1.944020   0.012166   159.8 <0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```

Equation $y* = 8.5212 + 1.944x*$

With $y*$ being the log transformation of the `Price` variable and $x*$ being the log transformation of the `Carat` variable.

Interpretation of Model For a 1% increase in the carat weight of the diamond, the predicted price increases by approximately 1.944%.