Project 1

Group 1: Alex Lilly, Alex Kendrick, Chi Do, Kristen Rose

2022-10-13

Section I. Executive Summary

Today's diamond market began in the late 19th century in South Africa. Prior to the turn of the 20th century, annual diamond production was under 1 million carats, and by the turn of the 21st century, diamond production increased to over 100 million carats per year. In addition to improvements in mining and sourcing which bolstered the supply of diamonds, the diamond market has changed drastically compared to its status even a decade ago. Numerous online diamond retailers exist today and allow consumers to hand pick diamonds and craft jewelry to fit their desires. As a result, the factors driving the price of diamond jewelry are brought to the forefront of the consumer's attention as they browse the web pages of drop down menus and sliders searching for the perfect diamond.

One of several online diamond retailers is Blue Nile. Founded in 1999, Blue Nile is primarily an online diamond and jewelry retailer, but they also have physical showrooms scattered across the United States. Blue Nile was one of the first companies to leverage the modern online retail format. They provide consumers with a wealth of expert advice on their website when it comes to picking the right diamond. This analysis serves to interrogate that expert advice.

Data was analyzed from 1,214 diamonds sold by Blue Nile of varying color, clarity, cut type, and carat weight, to discern the driving factors behind diamond price. In addition to price drivers, the inter-relationship between diamond qualities was explored. The results of this analysis suggest the following:

1. For a 1% increase in carat weight, there is a 1.9% increase in price. For example, if a 1 carat diamond is approximately $5,000.00, a 2 carat diamond (100% increase in carat weight) is approximately $19,000.00 (190% increase in price) on average.
2. The median price for higher quality cuts of diamond is lower than those of lesser quality. The analysis revealed the median price of Ideal/Astor Ideal diamonds is actually lower than the median price for Good/Very Good diamonds, despite cut having "the greatest influence on a diamond's beauty and sparkle" according to Blue Nile.
3. Behind carat weight, color had the largest impact on price when controlling for the effect of cut type. "Colorless" diamonds on average have a higher price than those which are classified as "Near Colorless".
4. Blue Nile states that most inclusions do not affect beauty because they are too small to see with the naked eye. Despite that, diamonds with the highest clarity grades have prices that reflect that. Our analysis indicated that the assertion of clarity having an effect on price was not necessarily true.

Section II. Data Description and Visualizations

Description of Data

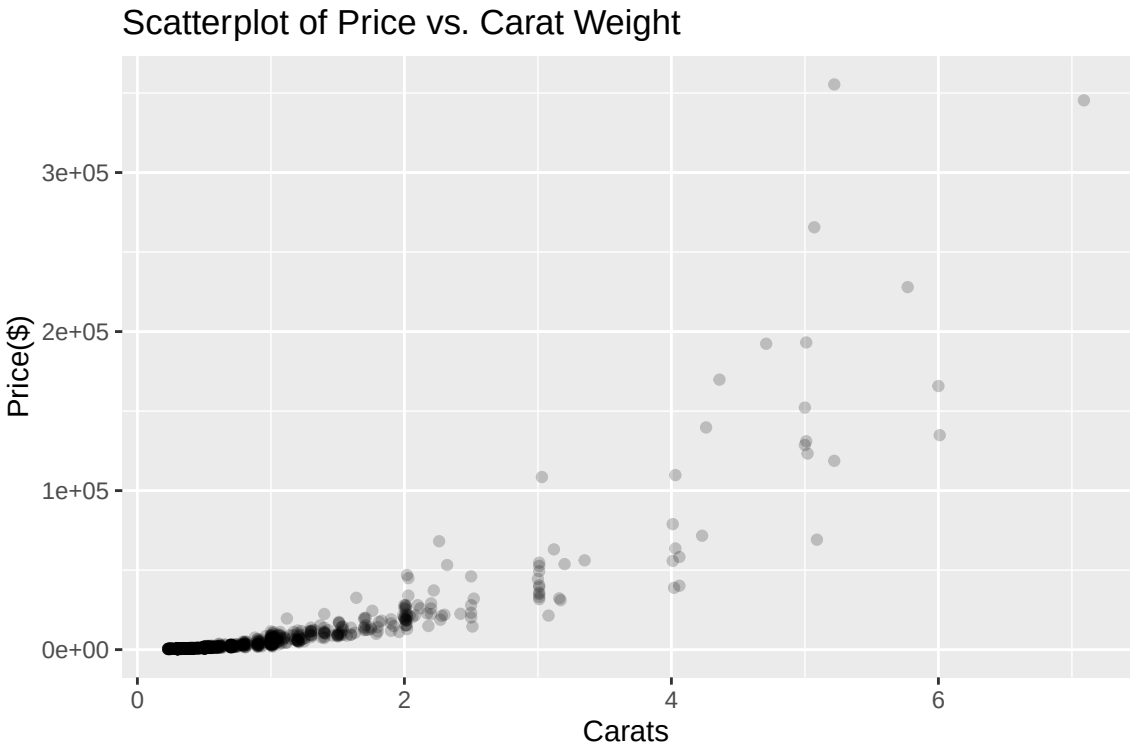
For each diamond in the sample five data elements were collected: price, carat, color, clarity, and cut.

- **Price** is used to measure the value of the diamond in US dollars.
- **Carat** describes the weight of the diamond in the unit carats. Since 1907 one carat is defined as equivalent to 200 milligrams.
- **Cut** is an assessment of how well-proportioned the dimensions of the diamond are and how well shaped and positioned its facets are. The diamonds in our sample range contain cuts graded Good, Very Good, Ideal and Astor Ideal. Ideal cut is a term used by Blue Nile and is equivalent to the grade Excellent used by the Gemological Institute of America.
- **Color** is a measure of how colorless white diamonds are. The less color the higher the grade. The scale used to measure color ranges from Z at the low end to D at the high end. The subset of diamonds in our sample has a minimum rated color of J. Colored diamonds do not use this scale and some intensely colored diamonds are in high demand. For some analysis we bucketed the color grades into near-colorless (G-J) and colorless (D-F) based on the definitions from Blue Nile.
- **Clarity** is used to categorize imperfections on the surface and within a diamond. Surface imperfections are blemishes and internal imperfections are inclusions. The categorizations are determined by how obvious and numerous the diamonds imperfections are. The range of diamonds in our sample from low to high is Slightly Included (SI2, SI1), Very Slightly Included (VS2, VS1), Very, Very Slightly Included (VVS2, VVS1), Internally Flawless (IF), and Flawless (FL). For some analysis we collapsed the number differences in categories (i.e. SI2 and SI1 are bucketed as SI) and combined IF and FL.

Visualizations

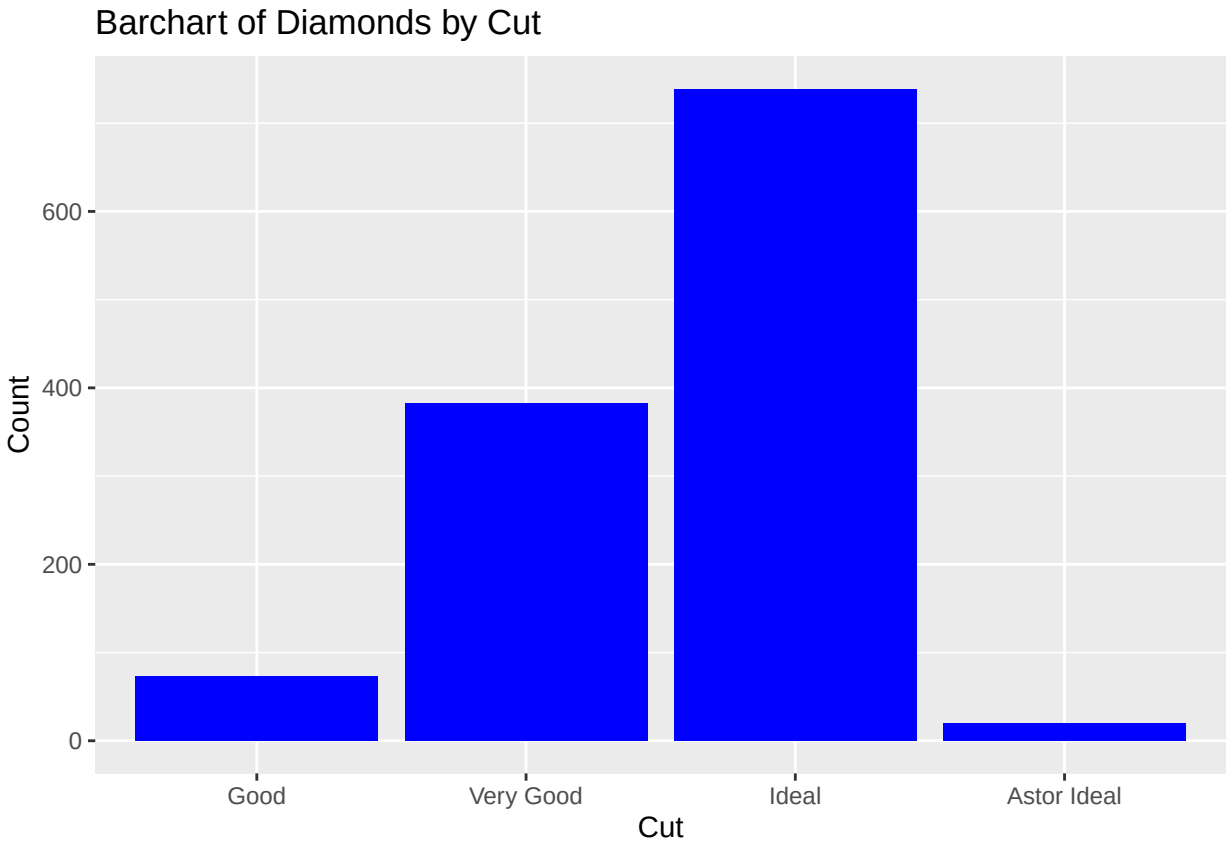
Carat Weight

According to Blue Nile Carat weight has the biggest effect on price. They indicate that this is due to the carat size of the diamond becoming a status symbol. From the graph below we can see that as carat weight increases the price of the diamond also increases aggressively and non-linearly. The variance of the price also increases as the carat weight increases. Further analysis of the Carat weight and its effect on price is explored in the linear model in section 3.

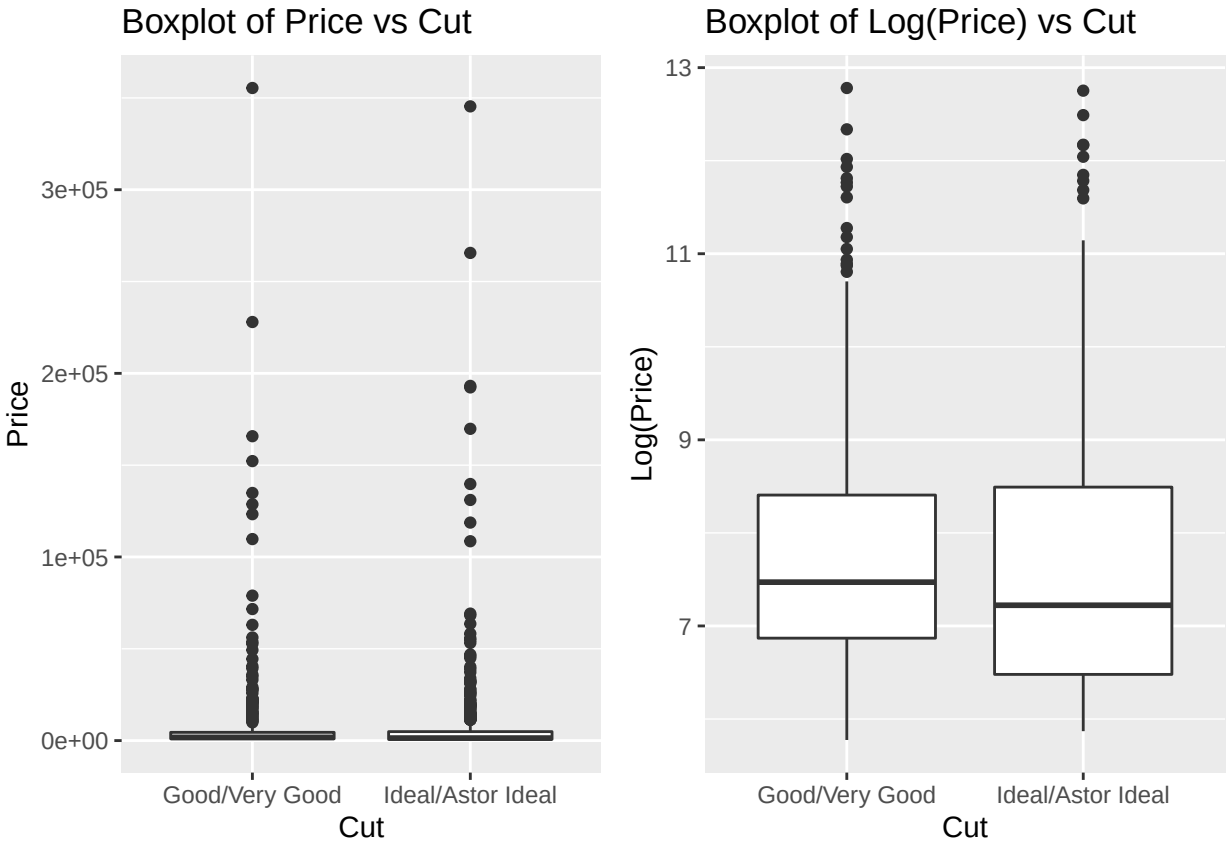


Cut

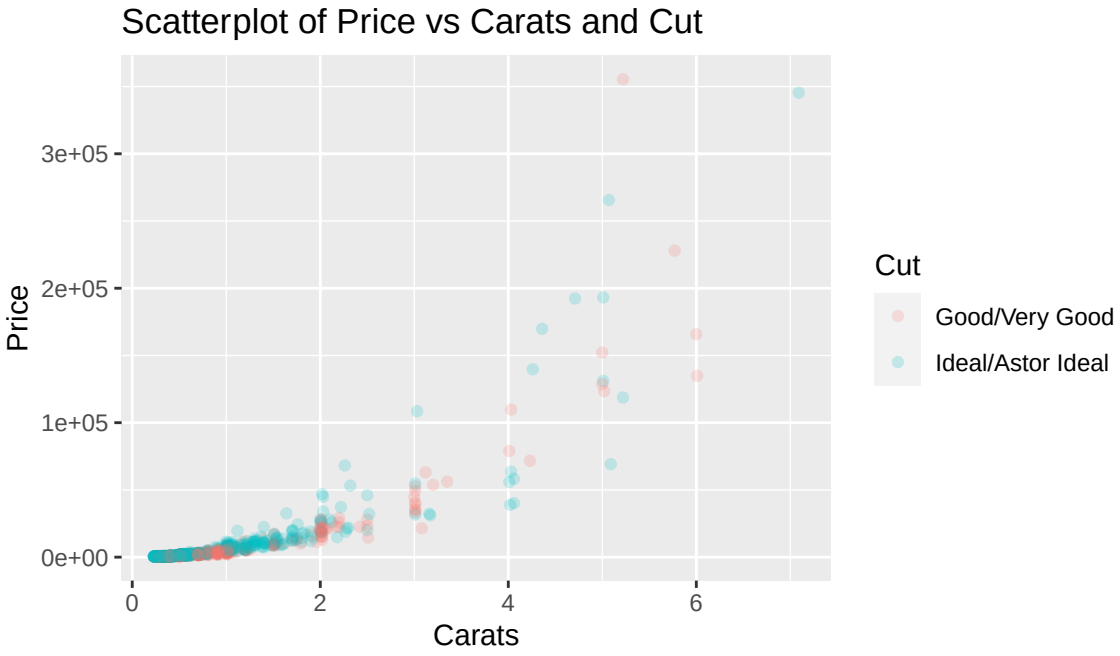
The proportion of diamonds with a Good or Astor Ideal cut in the sample was low, as shown in the graph below. To help with the analysis of the Cut variable the categories we condensed into Good/Very Good and Ideal/Astor Ideal.



Blue Nile considers the cut of a diamond the most important of the 4Cs when determining quality. However from our analysis of cut an improved quality of cut does not always correspond with an increased price of the diamond. The majority of diamonds sold have Very Good and Ideal quality cuts. It could be argued that this is because consumers view the quality of these cuts as good value for the price. When analyzing the price across the different cut qualities across the entire sample the price was not impacted significantly by the cut as shown below.



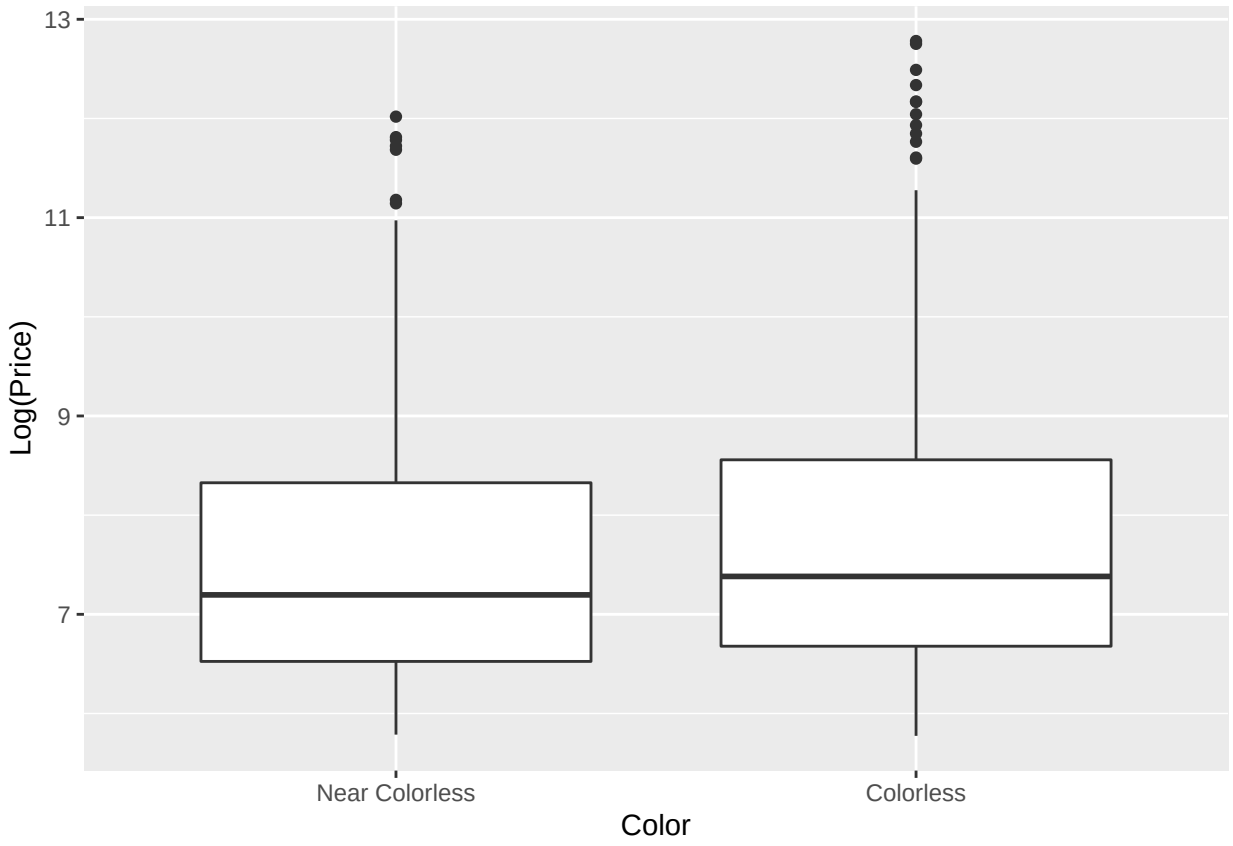
When viewing the boxplots with price on the y axis the large spread of price with a majority of the diamonds clustered at the low end of the range made the boxplots difficult to analyze. To increase the readability of the plot, the log transform was used. Diamond cuts have been collapsed into two categories described earlier. A counterintuitive finding from this plot, based on Blue Niles assertion of the importance of cut, is that the median price for Ideal or Astor Ideal diamonds is slightly lower than the median price for Good or Very Good diamonds. The Ideal or Astor Ideal diamonds, however, tend to range in price more than the Good or Very Good diamonds.

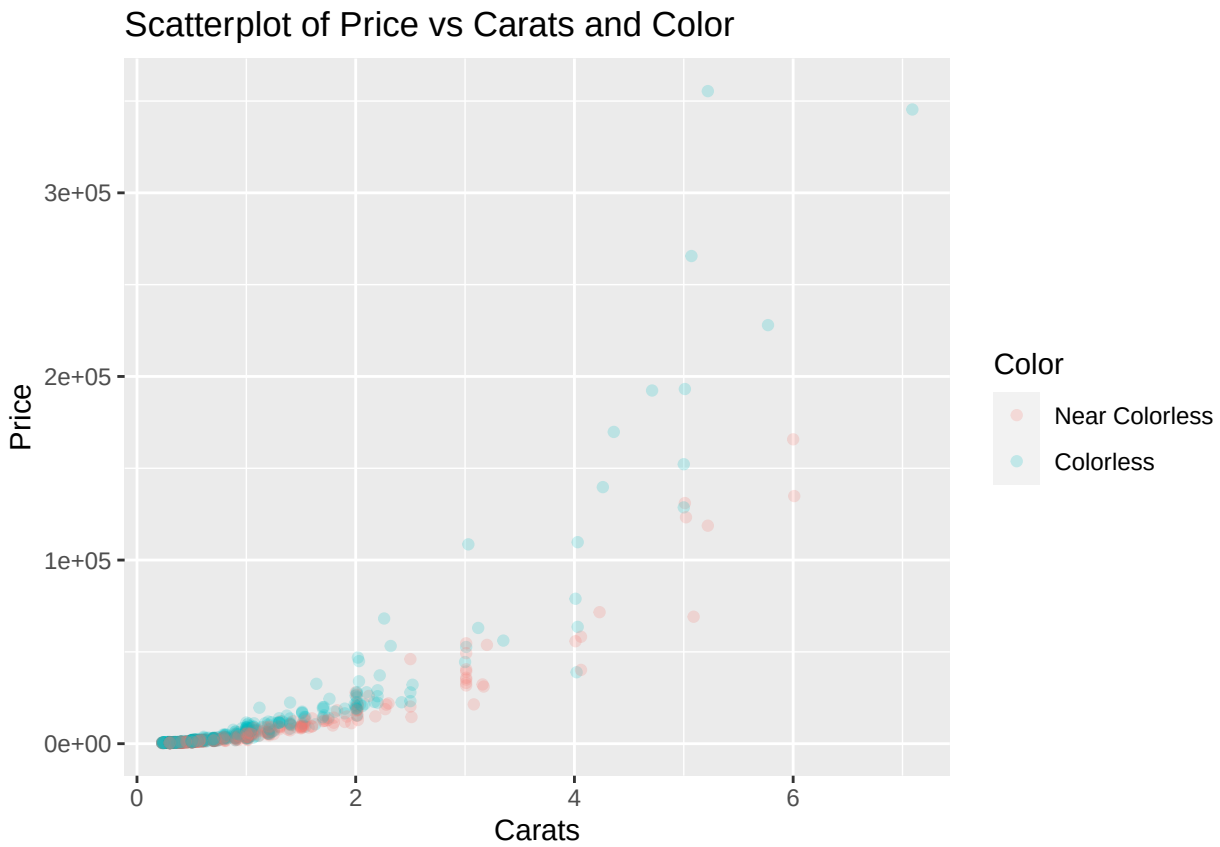


Knowing the importance of Carat weight and its effect on price, we created a scatterplot to explore the effect of both Carat weight and Cut on price. The results from this plot are also surprising. For diamonds with similar carat weights, a better cut does not always have a higher price. This makes it difficult to conclude that the Cut of a diamond has a significant effect on price, despite Blue Nile's assertion that it is the most important of the 4Cs when determing diamond quality.

Color

Similar to price and clarity, it is hard to visualize the relationship between diamond price and color in the left visualization due to high-priced outliers and the majority of diamonds having a lower price. The right visualization shows the relationship between diamond color and the natural log of prices. Diamond color grades have been collapsed into two categories for an easier comparison: colorless (grades D-F) and near-colorless (grades G-J) based on Blue Nile's definitions. The box plot shows that the median price for near-colorless diamonds is slightly lower than the median price for colorless diamonds.

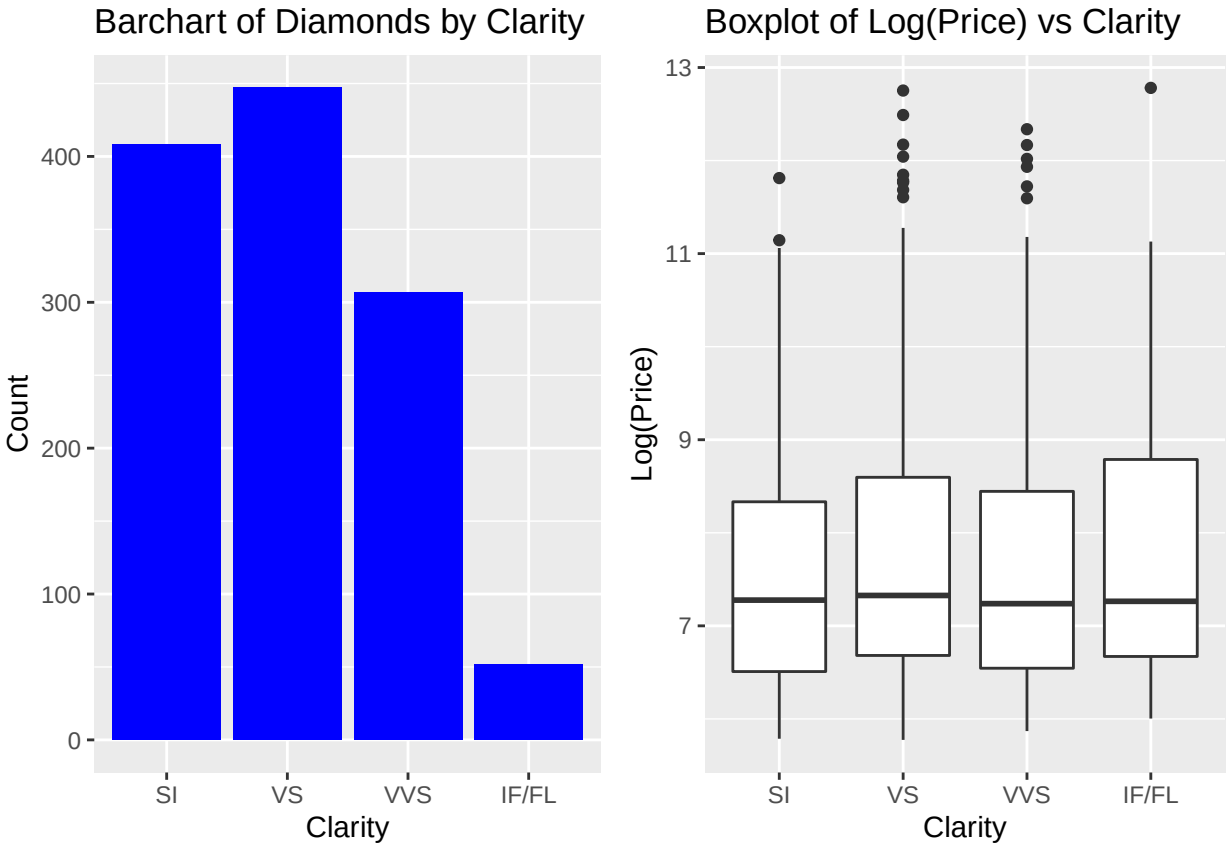

13
11
9
7
Log(Price)
Near Colorless
Colorless
Color


Scatterplot of Price vs Carats and Color
3e+05
2e+05
1e+05
0e+00
Price
0
2
4
6
Carats
Color
Near Colorless
Colorless

Like before with Cut, knowing that Carat weight has a large impact on price we used a scatter plot to analyze the impact of both Carat wieght together with Color on price. From the plot we can see a clear pattern that at similar Carat weights the colorless diamonds tend to have a higher price then the Near Colorless diamonds. This evidence points towards the Color of a diamond having a real effect on the price.

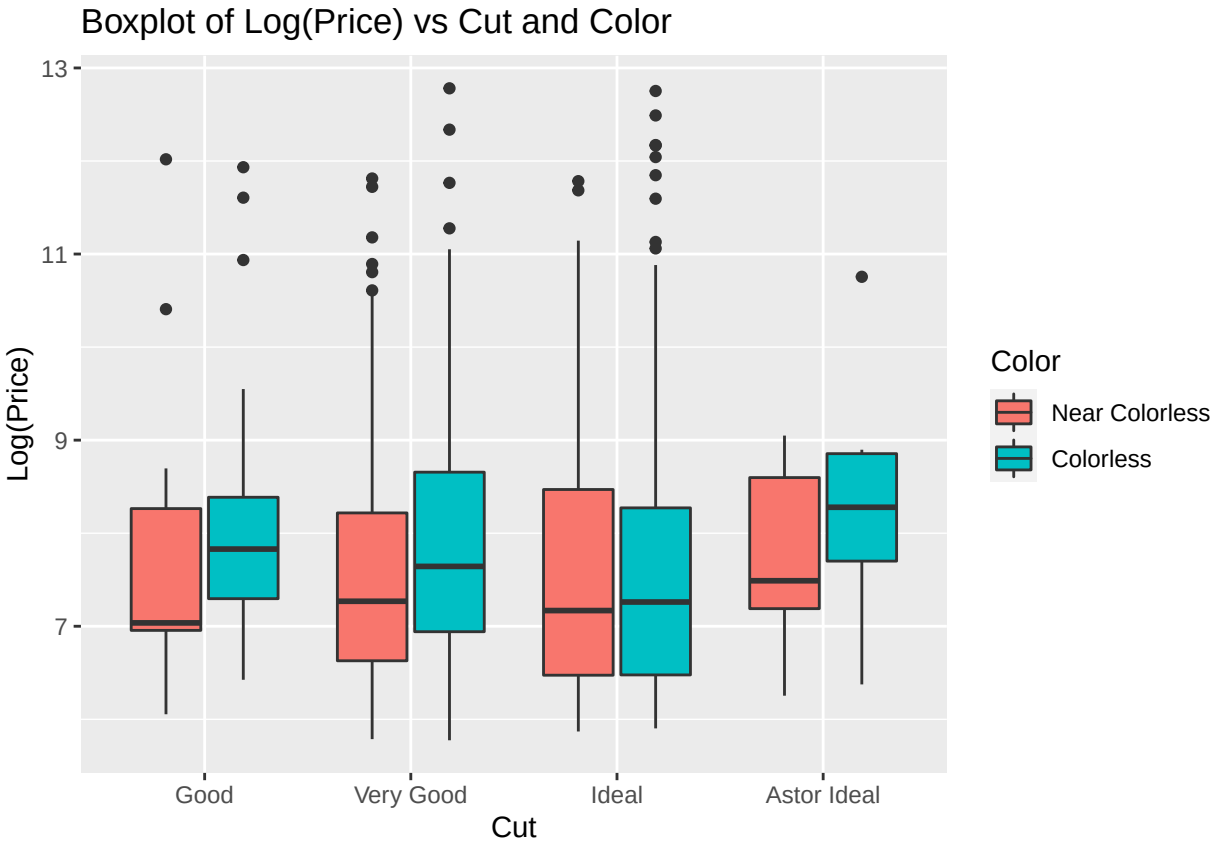
Clarity

There are several values used to quantify the clarity of the diamond based on the visibility and size and location of imperfections in a diamond. We've condensed the clarity levels to help with the analysis by combining the different numbered subcategories (i.e. combining SI2 and SI1) and combining the internally flawless and flawless categories since they make up a low proportion of the sample. This can be seen from the barchart of the different clarities. The chart also illustrates that over half of the diamonds fall into the SI or VS clarity ratings which supports Blue Nile's claim that those are the most popular clarities.

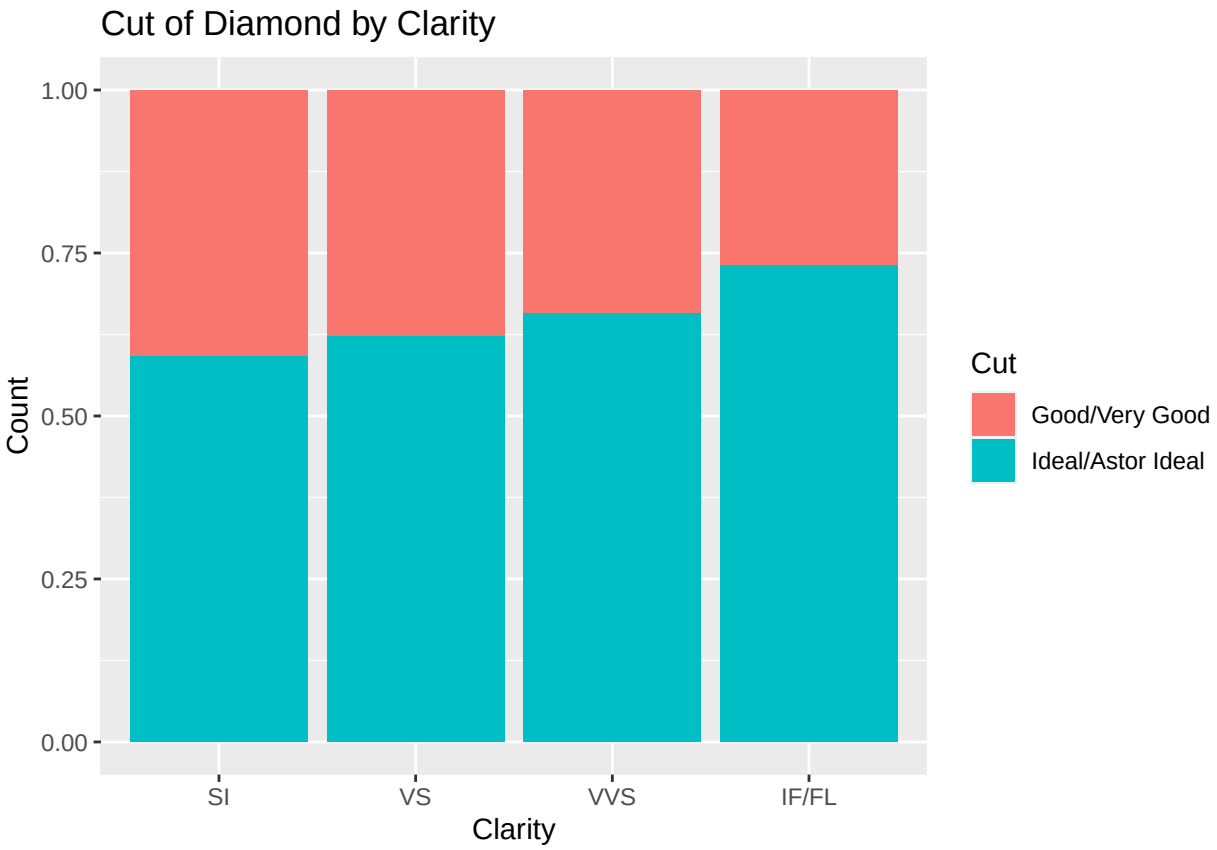


According to the plot, there does not appear to be a significant difference in median price between the clarity grades. This does not follow Blue Nile's assertion that higher clarity ratings correspond to higher prices. It does support the idea that many consumers may not notice clarity differences because they are not visible with the naked eye.

Multivariate Interactions



Blue Nile explains that diamonds cut with "Good" quality represent the top 25% of diamond cut quality, with "Very Good" making up the top 15%, and "Ideal" making up just the top 3%. They go on to describe the trademarked "Astor by Blue Nile" as those which "exhibit outstanding brilliance, fire, and scintillation." Naturally, one would expect there to be a discernible variance in price as you improve the cut type. From the boxplot above, it is evident that the diamond price does not vary significantly with cut type. However, within each cut type category, the median diamond price is higher for diamond colors which are classified as "Colorless" compared to those which are classified as "Near Colorless". This suggests consumers value diamond color to a higher degree than cut type. In order to fully understand this relationship between price, cut type, and diamond color, a multiple linear regression would have to be implemented, as it is very likely the impact of carat weight and clarity are confounding the trends displayed here.

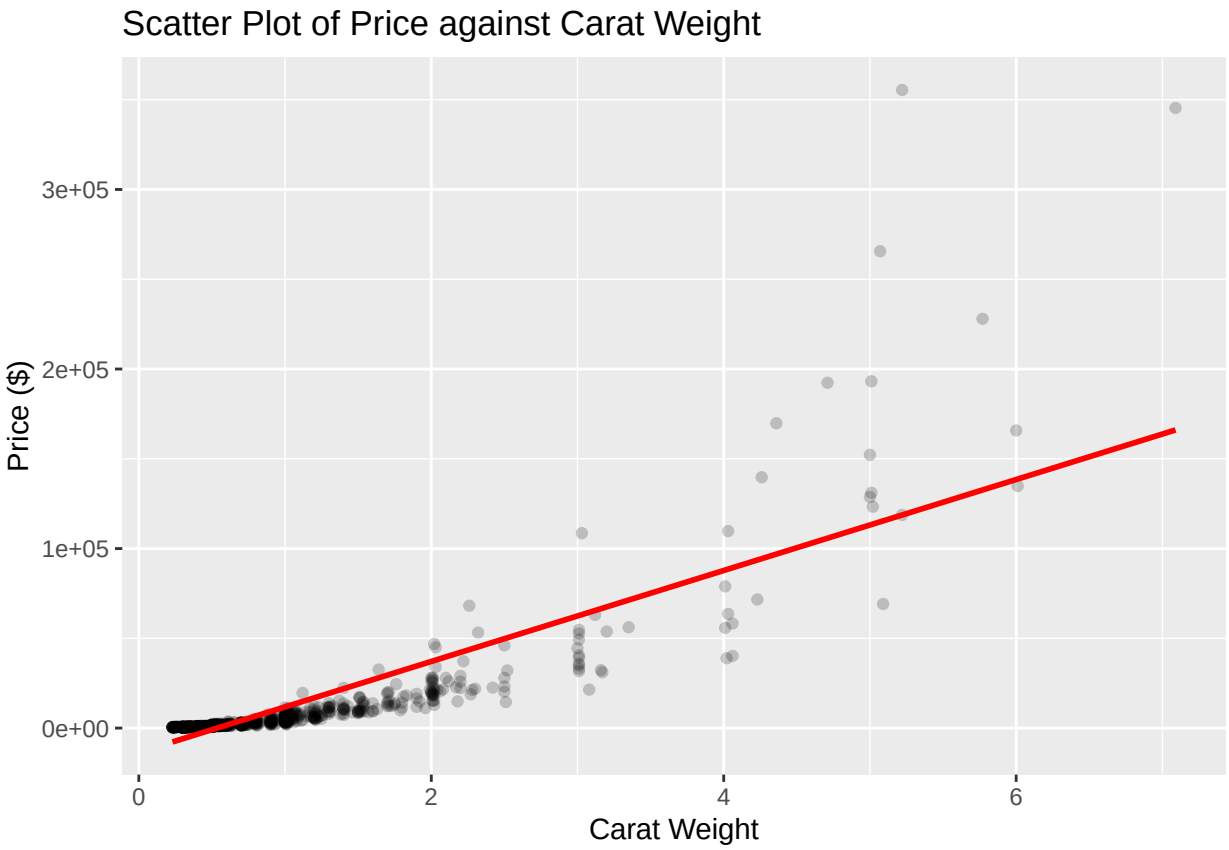


It is also interesting to note the relationship between variables without looking at price. For instance, Blue Nile emphasizes that "a diamond with a flawless clarity grade... can look glassy or dull if the cut is too shallow or deep." Therefore, it is most important to first choose a well-cut diamond before focusing on clarity.

The bar chart above visualizes the sample's relationship between diamond cut and clarity. It is interesting to note that as clarity grade improves, a higher proportion of diamonds tend to have an Ideal or Astor Ideal cut. This is likely because buyers who can afford the cost of an Ideal or Astor Ideal diamond are more likely to have a budget that also can afford a higher clarity grade.

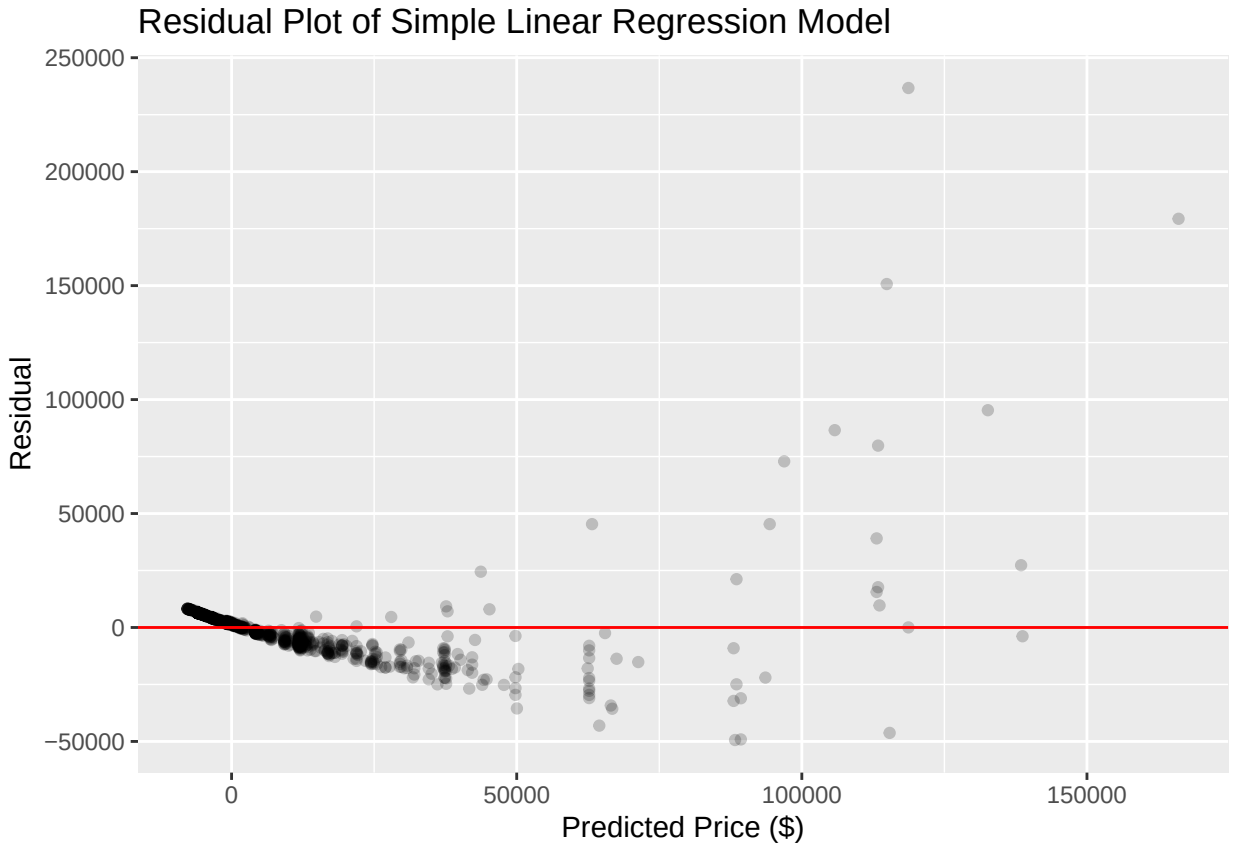
Section III. Simple Linear Regression of Price Agaisnt Carat Weight

A scatter plot of diamond price against carat weight was generated in order to begin exploring the relationship between those two variables.



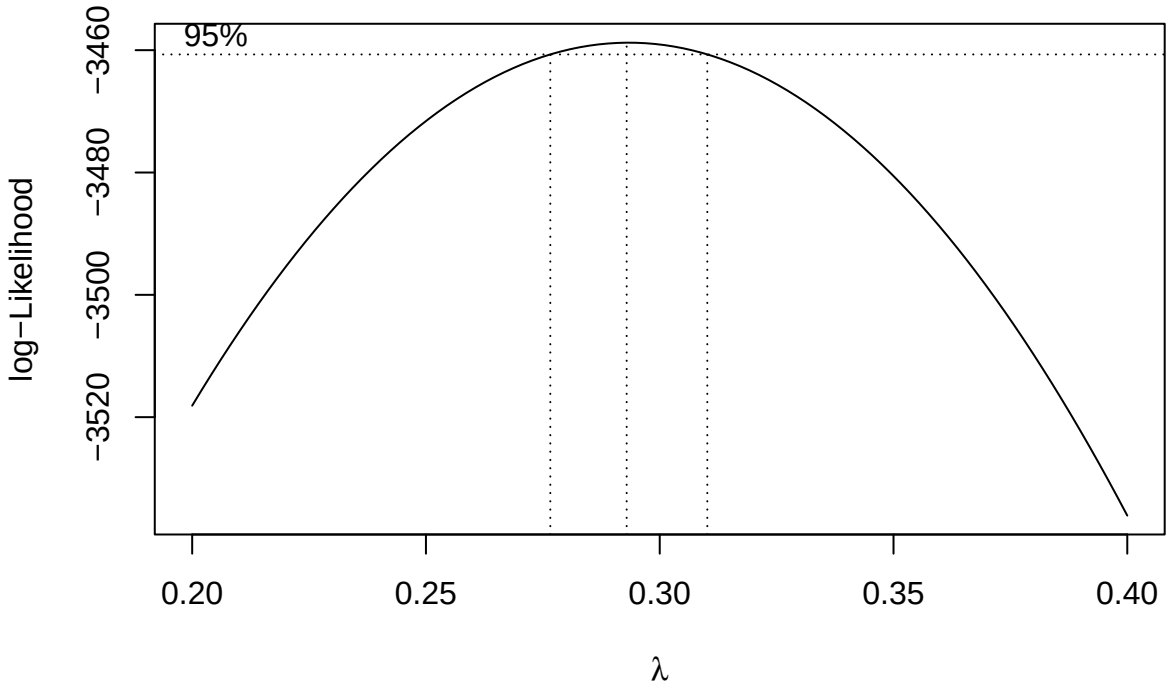
Evidently, as carat weight increases, the price of the diamond also increases non-linearly. There also appears to be a scarcity of diamonds with higher carat weight because the density of data decreases as carat weight increases.

Fitting a simple linear regression model and calculating the residuals yields the following residual plot.

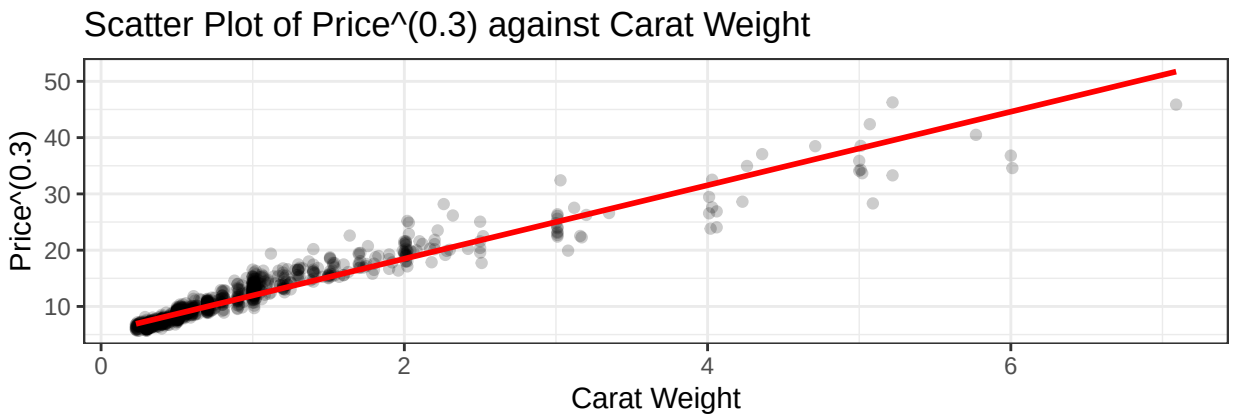


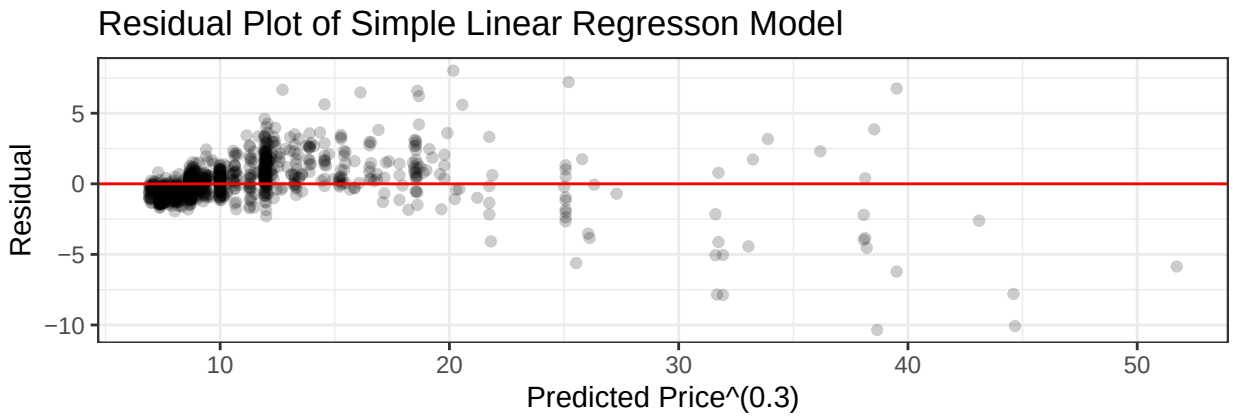
There are regions of the residual plot where the error does not have mean zero, and the variance appears to be increasing as the predicted price increases.

Next, the Box-Cox method was used to determine the most likely transformation that could be applied to the predicted variable, in this case price, in order to generate a distribution that is normal.



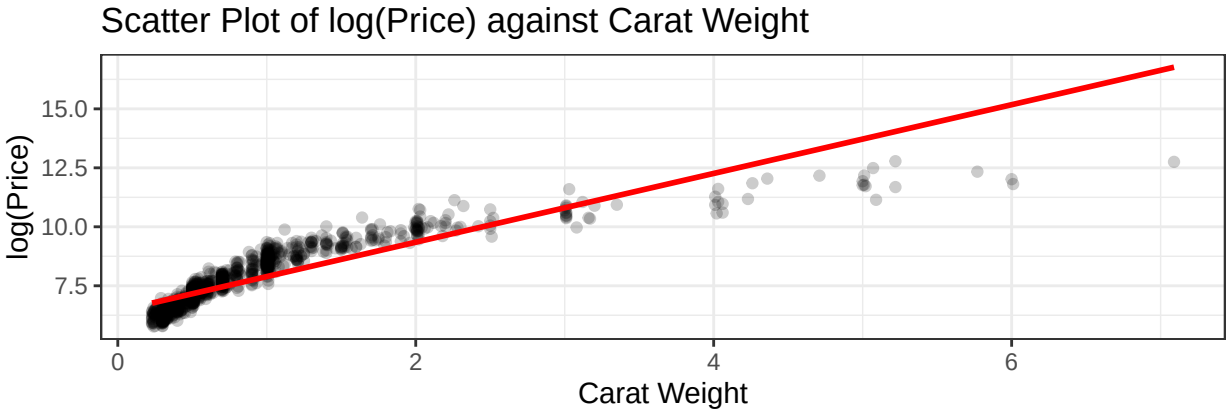
The Box-Cox plot suggests a transform of y^{λ} where $\lambda = 0.3$ would yield a distribution of prices that is most likely to align with a normal distribution. Below are scatter plots of transformed price vs. carat weight and the residual plot of the resulting simple linear regression that was applied.

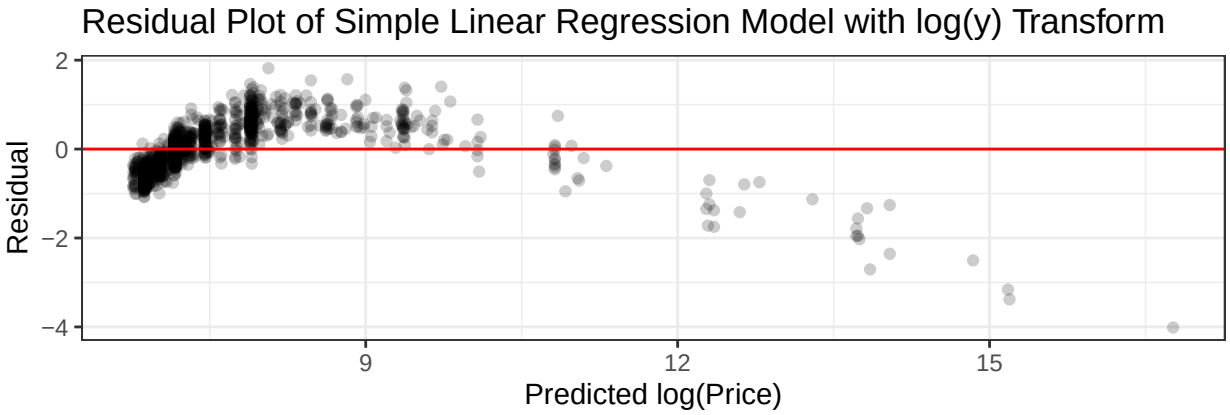




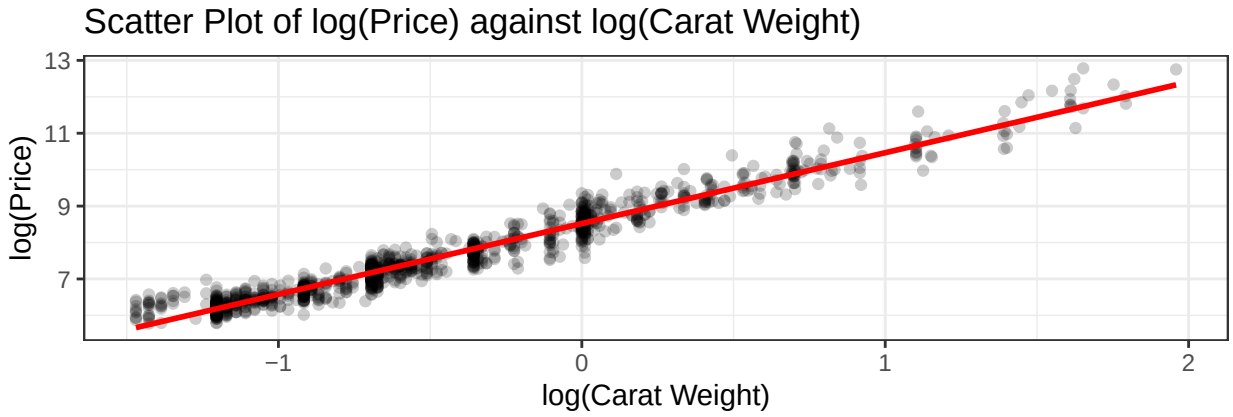
From the scatter plot alone it is evident that the relationship between price and carat weight was linearized partially. In the residual plot, the issue of non-zero error means in some regions persists. Again, the issue of non-constant variance is present. This model can be improved.

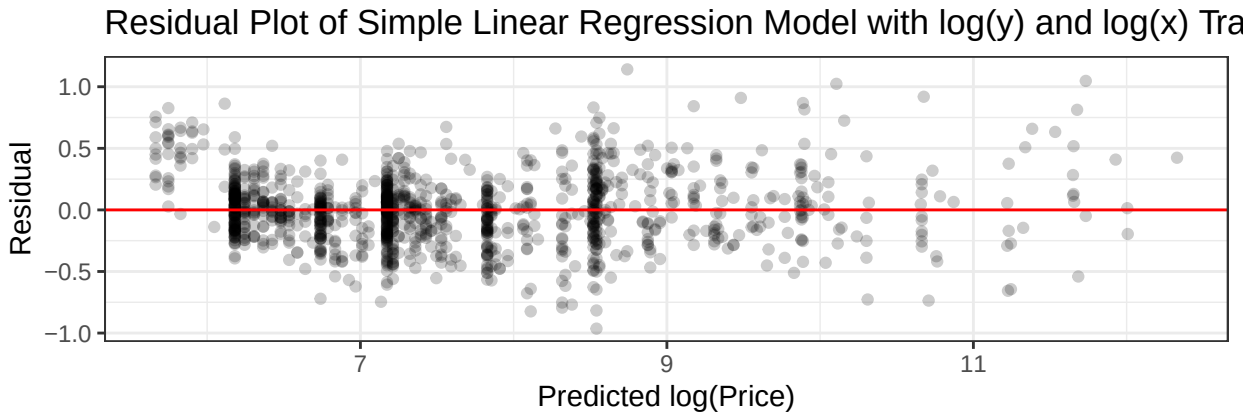
Next, a log transform of price was attempted. Since values of $\lambda < 1$, including logarithmic transforms, all have the same directional effect, it is likely that a logarithmic transform of price will have a similar effect as the lambda $= 0.3$ transformation, just to a higher degree. A logarithmic transform also has the benefit of being more easily interpretable than an exponential transform, which will enable investigators to more easily make inferences about the relationship between price and carat weight. Below are the scatter plots of log-transformed price vs. carat weight and the residual plot of the simple linear regression that was applied.





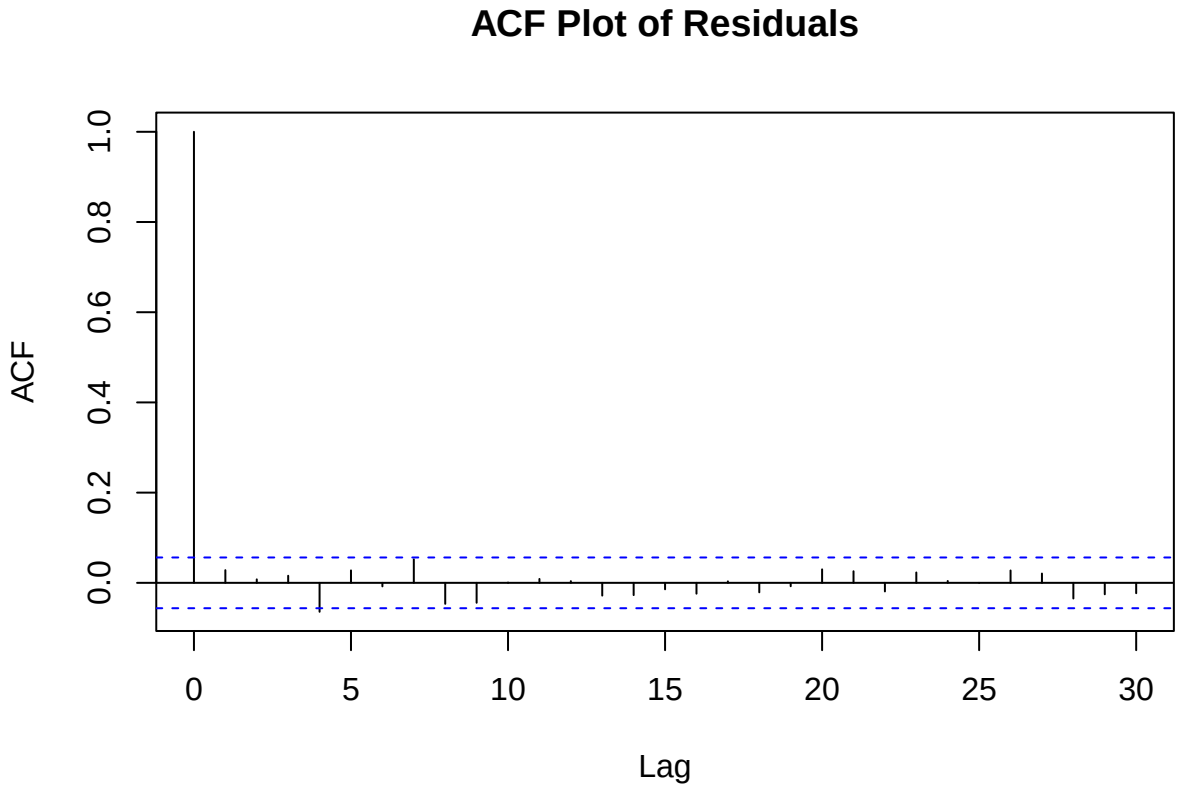
With the natural logarithm transformation of price, the issue of non-constant variance was addressed much better than it was with the $\lambda = 0.3$ transform. However the issue of not having residuals with mean equal to zero persists. To address this, a logarithmic transformation of carat weight was conducted in order to bring the carat weight distribution closer to normal. Below are the scatter plots of log-transformed price vs. log-transformed carat weight and the residual plot of the simple linear regression that was applied.



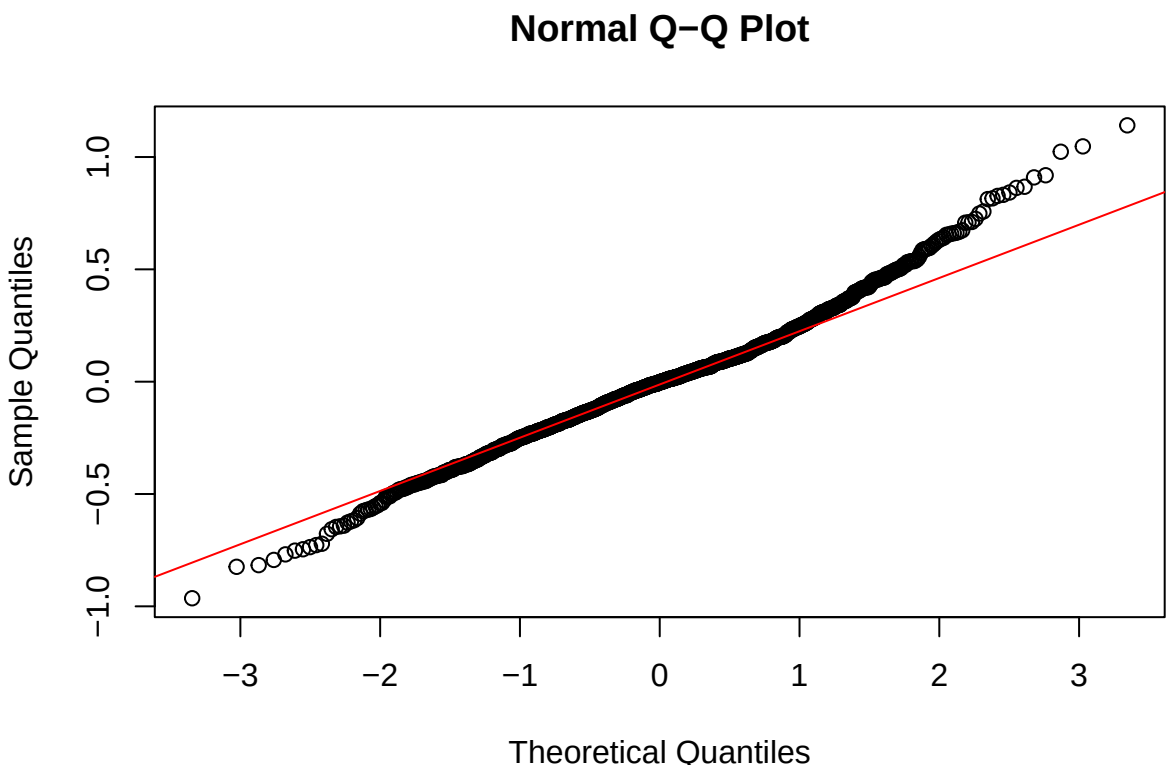


This simple linear regression model satisfied both of the following assumptions:
1. The errors, for each fixed value of x, have mean 0.
2. The errors, for each fixed value of x, have constant variance.

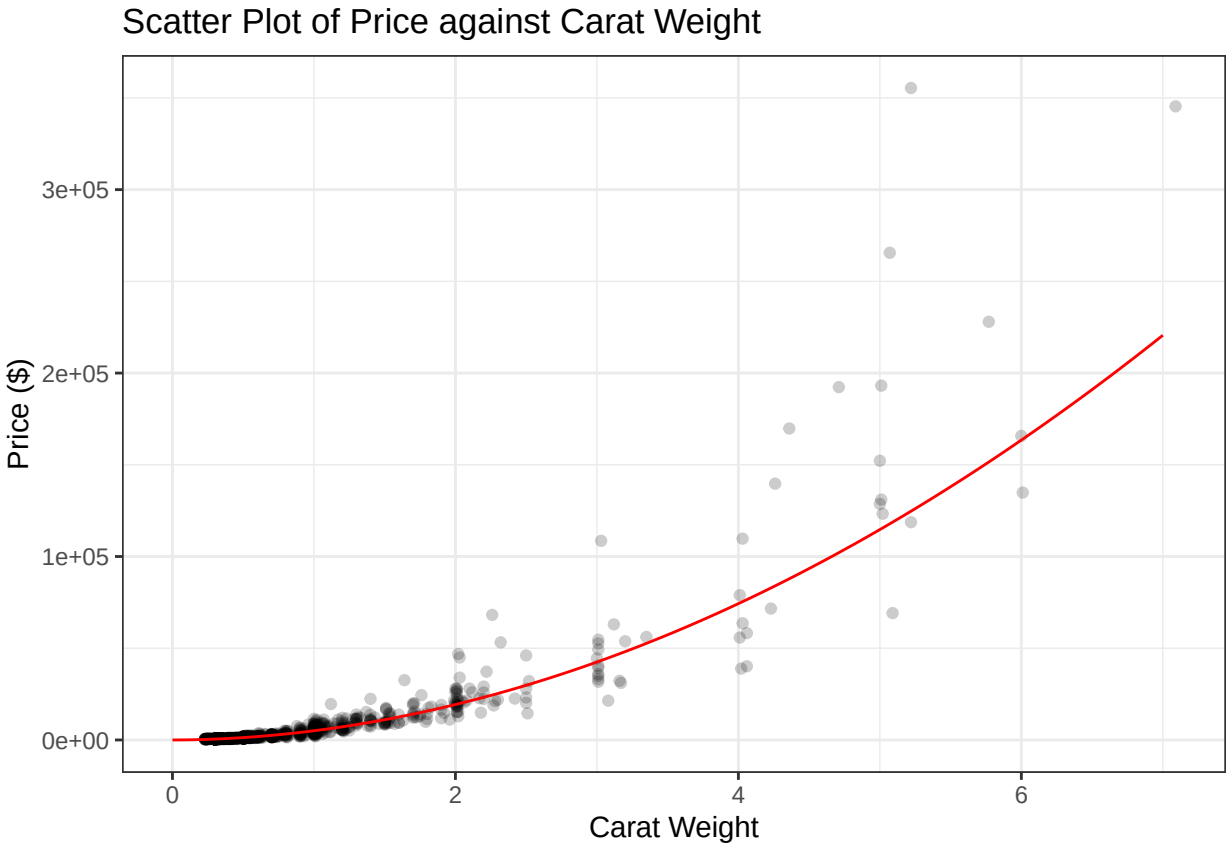
To determine if this model satisfies the assumptions regarding error independence and whether the error terms follows a normal distribution, an auto-correlation function plot and QQ plot were generated.



The auto-correlation function calculates the correlation between error terms of varying window sizes from lag = 0 to lag = 10*log_{10}(N/m) where N is the number of elements in m number of series. In this case, N = 1,214 diamonds, and there is only 1 series of residuals. There are two instances, namely when lag = 4 and 7, where the ACF barely exceeds the threshold for statistical significance at the 95% confidence level. All other lag values are within the 95% confidence interval, indicating that the residuals are mostly independent. This satisfies the assumption that the errors are independent.



There are deviations from normality near the extremes of theoretical quantiles, but otherwise the residuals closely align with the normal distribution. This satisfies the fourth and final assumption for simple linear regression: the errors, for each fixed value of x, follow a normal distribution.



The final regression expression is:

log(Price) = 1.944 * log(Carat Weight) + 8.521208.

To confirm that the model is useful in analyzing the effect of carat weight on price an ANOVA F test was conducted. $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$

```
## Analysis of Variance Table
##
## Response: logprice
##             Df Sum Sq Mean Sq F value    Pr(>F)
## logcarat     1 1946.8 1946.77   25535 < 2.2e-16 ***
## Residuals 1212   92.4    0.08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

| From the ANOVA F test we can reject the null hypothesis and conclude that the coefficient for carat weight is not equal to 0 at the 95% confidence level. So we can conclude that a linear relationship exists between the two variables. | This model has an $R^2 = 0.9547$, indicating that the 95.47% of the variance in the price of diamonds can be explained by this regression equation. As mentioned previously, the benefit of using a log-log regression function over an exponential transformation lies in interpretability. From this regression function, it shows that for a 1% change in carat weight, there is a 1.9% change in price on average.