Project 1: Diamond Price Against Carat

Alex Bass, Andy Ortiz, Dima Mikhaylov, Seth Harrison

Group 5

Executive Summary

Using a dataset of more than 1,000 diamonds, we analyzed various diamond characteristics that affect diamond price in USD. We also evaluated several claims made by prominent diamond company Blue Nile about those relationships. Generally, we found that the only diamond characteristic that substantially affects diamond price is carats - the higher the carats the higher the price of the diamond.

Methodologically, we analyzed relationships mostly using simple correlations - which checks how related two characteristics are to each other - and created figures examining certain parts of the data more closely. For example, we created a scatterplot showing how certain levels of diamond clarity compared to price (we found no relationship here by the way!). However, to evaluate diamond price and carat, we used a stronger statistical test that can determine if two variables are statistically related and, if so, can yield accurate predictions of price given carat.

In our analysis, we specifically examined the relationship between diamond price and these other characteristics: carat, clarity, cut, and color. As alluded to above, we did not find any strong relationship between price and clarity, cut, or color, but we did find a strong relationship between carat and price. In fact, we found that on average for every 10% increase in carats for a diamond being sold, there is a 19.85% increase in that diamond's price. Our findings contradicted several of Blue Niles claims which suggested a stronger relationship between clarity, cut, and color with price. While our data does not contain an exhaustive list of all diamonds bought and sold, it does give us a good idea of general trends in prices, and we confidently claim to diamond buyers/sellers that they should look to a diamond's carats as the strongest driver of its price.

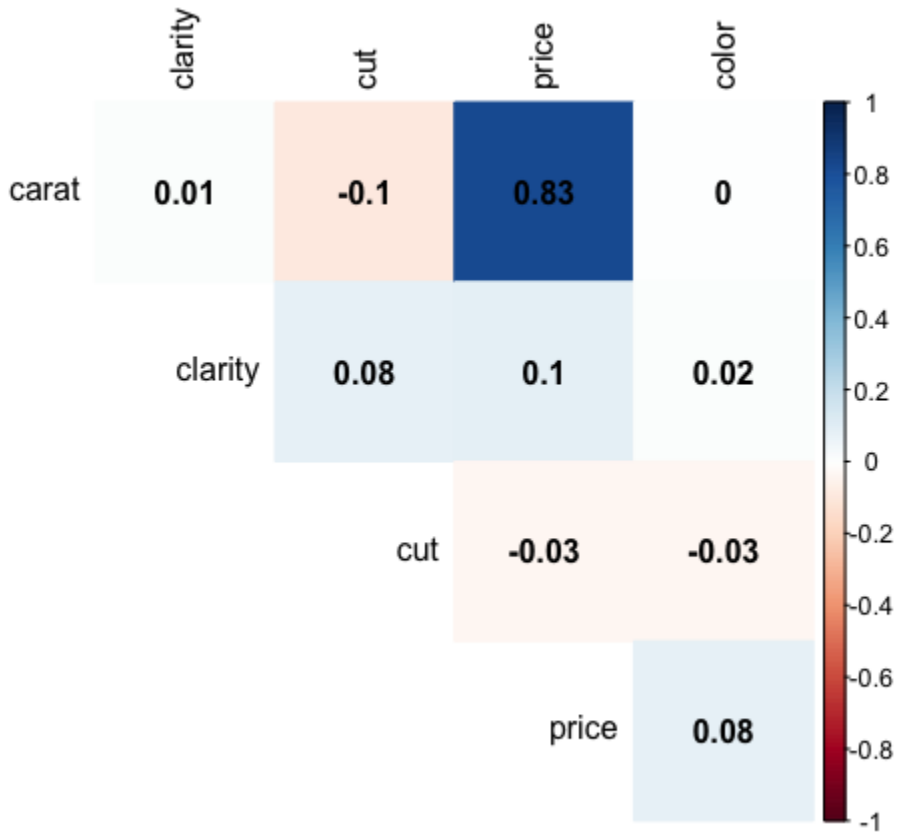
Data Description And General Correlations

For our analysis of diamonds, we used a data set containing various information about more than 1,000 diamonds from bluenile.com and other online sources. This dataset does not contain an exhaustive list from this site, only a subset. The variables in the dataset and their descriptions are below:

- Carat - refers to the weight of the diamond not the size (common misunderstanding) though the weight and size are generally correlated.
- Clarity - this is the purity of a particular diamond. Diamonds with higher clarity (FL, IF) have few blemishes/inclusions that can only be seen in a microscope. Diamonds with lower clarity (VS, VVS) can have several inclusions visible to the naked eye.
- Color - refers to the tone of the diamond. Higher letters (D, E, or F) signify a colorless, cool, and icy look while lower letters (H, I, J) indicate a gradually more yellow tone.
- Cut - refers to how the diamond is expertly cut with higher quality cuts gathering and reflecting more light and lower quality gathering and reflecting less. There are 4 grades from worst to best: good, very good, ideal, and astor ideal.
- Price - the cost of the diamond in USD.

Before diving into some of the analysis and evaluation of Blue Nile's price claims, let's look at an overview of the data and get a feel for some general trends and correlations. Several variables in this dataset are categorical without a continuous scale but all of these variables have a natural ordering. Using this natural ordering, we can

compare the continuous variables (price and carat) to these other ordinal variables via correlation. A correlogram of this dataset was generated below:

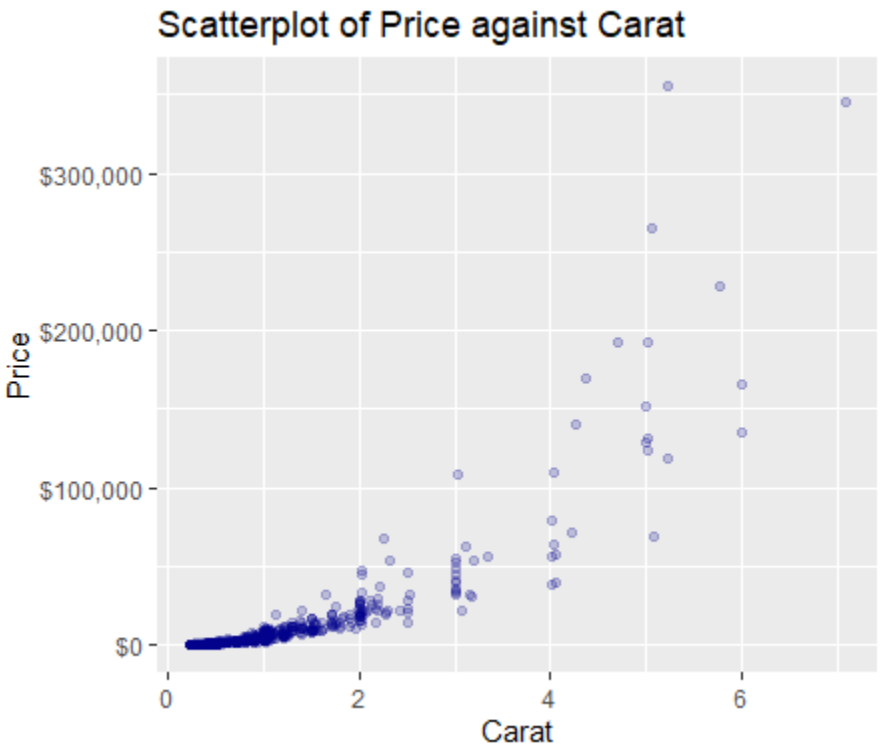


In this chart, lighter colors signify a weaker correlation and darker colors signify a stronger one. As one can see, only two fields in this dataset are correlated: price and carat. This is a strong positive correlation meaning that in general diamonds with higher prices are related to diamonds with higher carats. Interestingly, every other variable shows a weak correlation with the rest of the variables in the dataset. At first glance, this is surprising because it would seem that other variables like the diamond's purity, or color would be more correlated to the price of the diamond (as the Blue Nile website suggests), but this doesn't seem to be the case. A weak correlation between the other variables might suggest that diamonds with various clarities, cuts, and colors are highly

customizable. For example, it may be common that a customer might buy a diamond with high clarity and bad color while another might choose vice versa while the price is unaffected between the two. In the next section, we will evaluate some of the Blue Nile claims using data visualizations from our dataset.

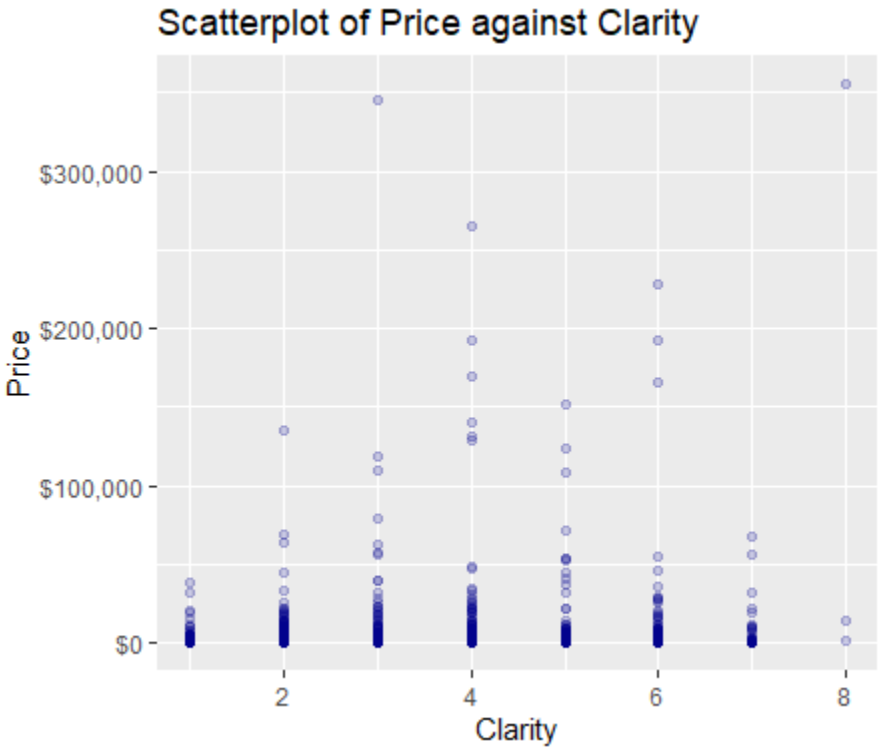
Blue Nile Claims And Visualizations

With respect to a diamond's weight correlating with price, the Blue Nile website clearly describes a reason for the correlation, stating "…higher carat weights are cut from larger rough crystals that are harder to source…" The plot below fits with this reasoning, as the price increases in an exponential fashion with increasing weight.

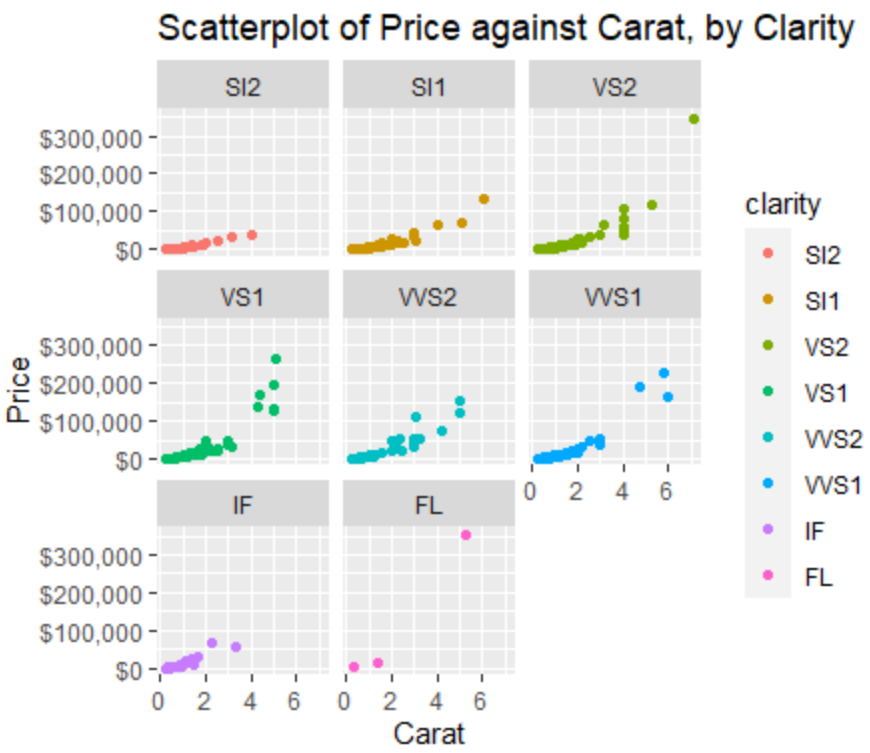


According to the Blue Nile website, "It's also important to note that diamonds with the fewest and smallest inclusions receive the highest clarity grades—and higher price

tags to reflect that." However, as the plot below shows, a higher grade of clarity does not correlate with price (here 1=lowest grade, 8=highest).

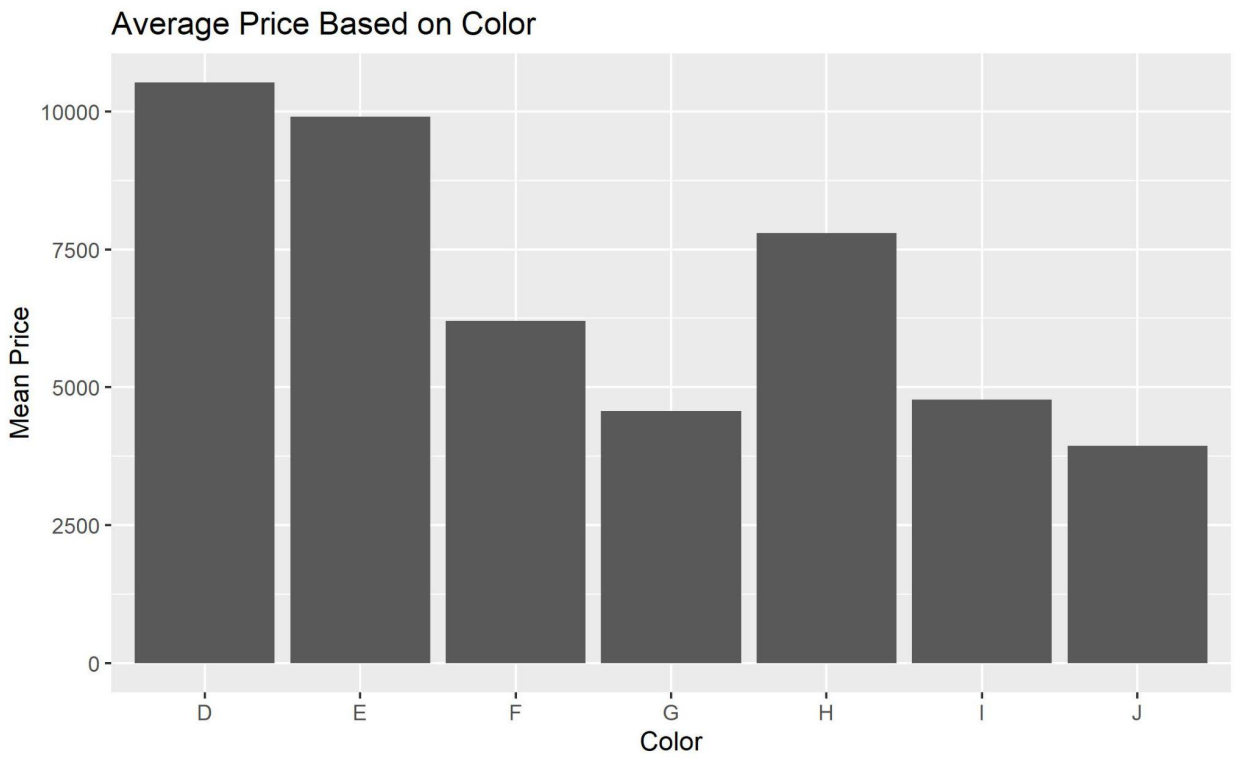


To further demonstrate the dominance of carat weight over clarity, we plot carats vs price for each grade of clarity in a separate facet. The plots are almost indistinguishable:



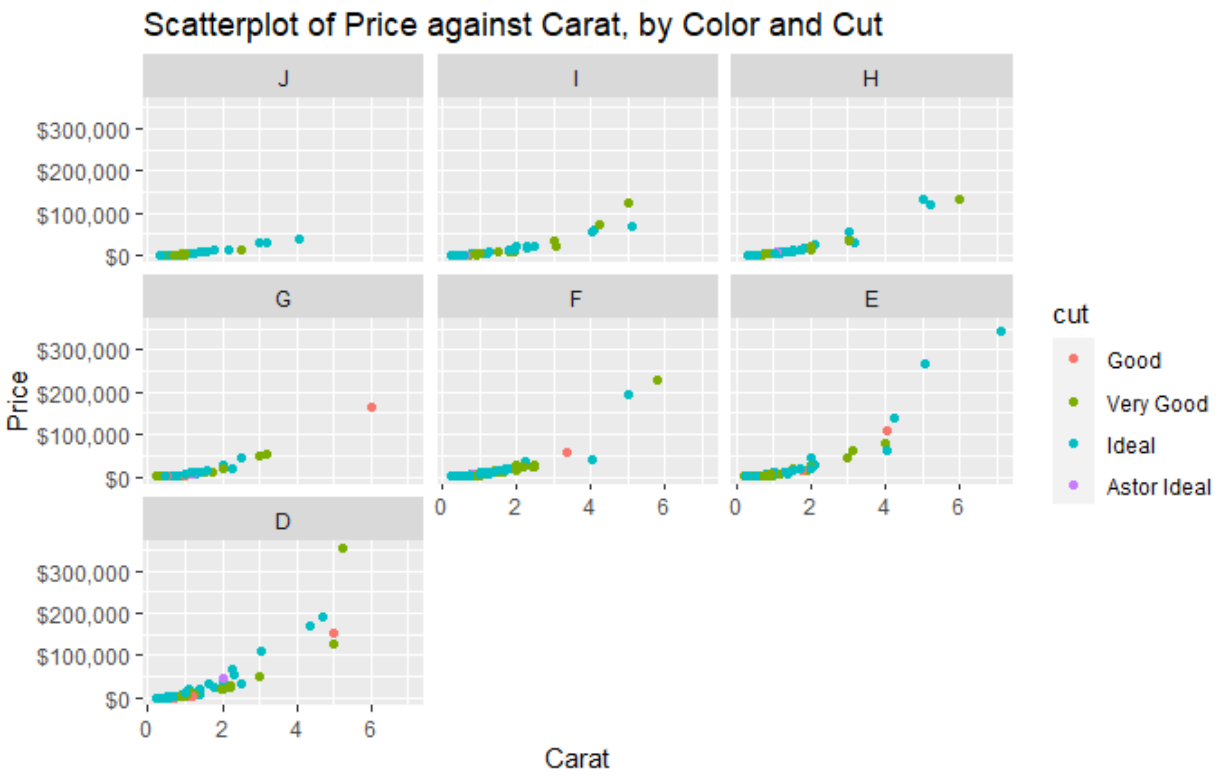
Price Relation to Color and Cut

On Blue Nile's page they make the claim that the two most important factors in a diamond are the cut, followed by the color. Though when looking at the relationship between these two variables it does not appear that either of them are closely correlated with price based on our correlation matrix. Though this may be the case, we can still look at the average price distribution of different colors and cuts.



From looking at the average price of the diamonds based on color, we see a somewhat similar price trend to that which Blue Nile details on their website with H being seemingly more valuable than they claim. This does not necessarily mean that color is a good indicator for predicting the value of a diamond, though it does reflect that some of the claims made by Blue Nile may not be completely unfounded.

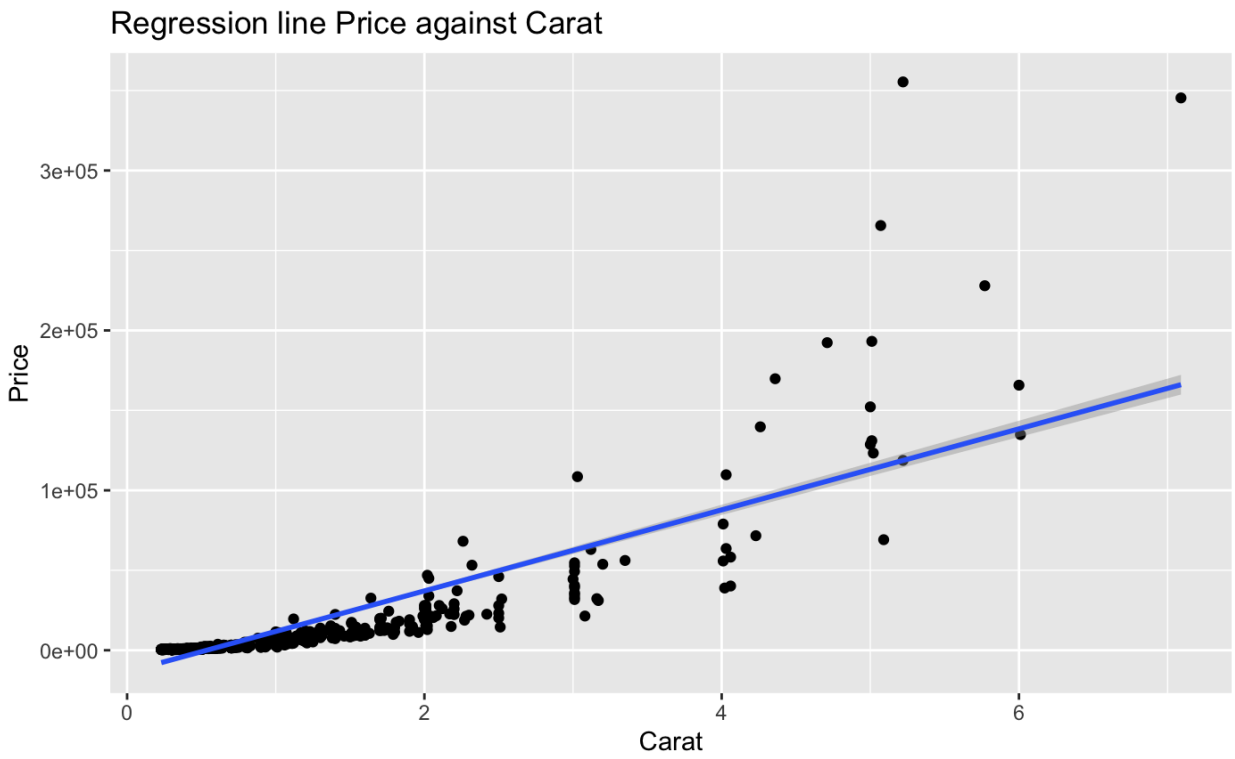
Looking at the average price of different cuts though, there is a somewhat negative relationship between the cut and the price. We see that the Astor ideal cuts have the lowest average price, and the good cuts have the highest. This would lead us to believe that the worse the cut is, the more expensive the diamond will be. It is important to note that there are a few very expensive diamonds in this dataset that could be skewing the price of the lower quality cuts because of factors that are more impactful, such as the carat, which are not factored into this visualization.



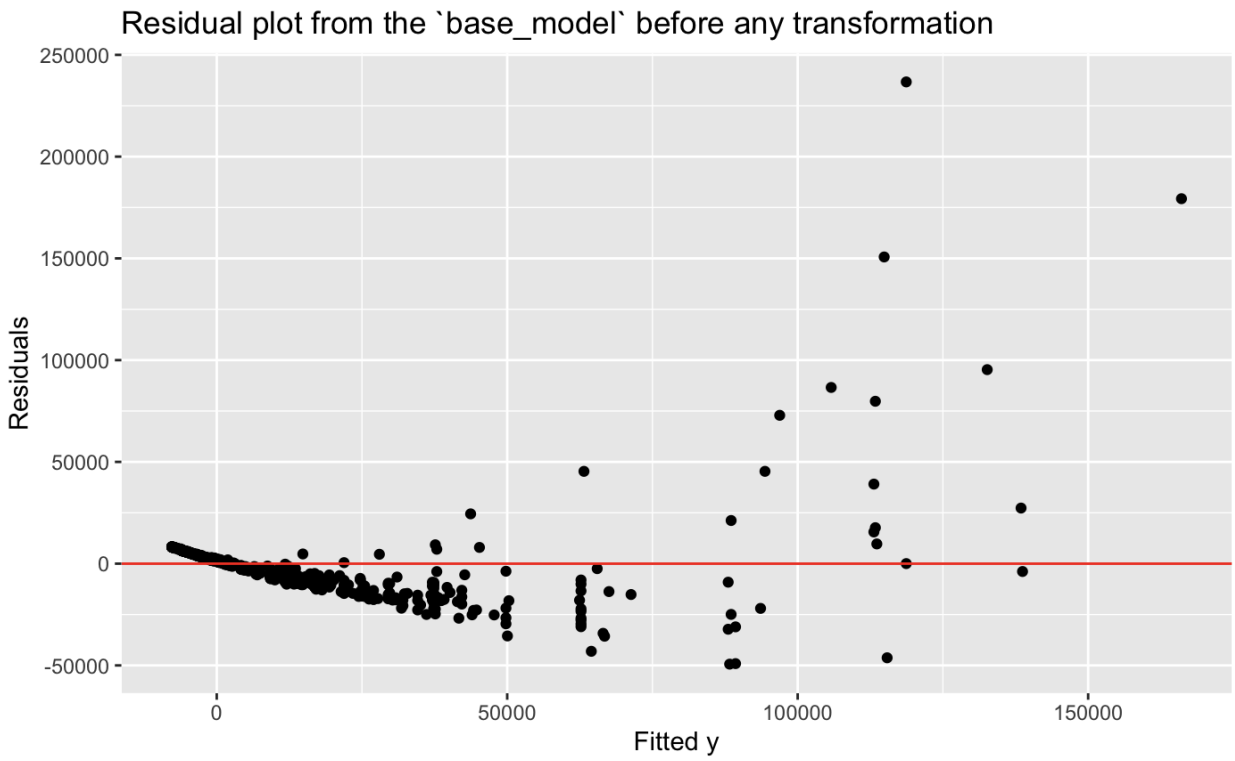
When looking at how the color and the cut of a diamond affect the price, it appears that in general those two factors matter very little compared to the carat. In the visual above we can see that the reason we previously found good cut quality diamonds to be more expensive than the Astral ideal is because many of the good cut diamonds in our data have high carat values. This seems to explain why our correlation between carat and price is so much higher than cut, or color with price.

Fitting Price Against Carat Using SLR

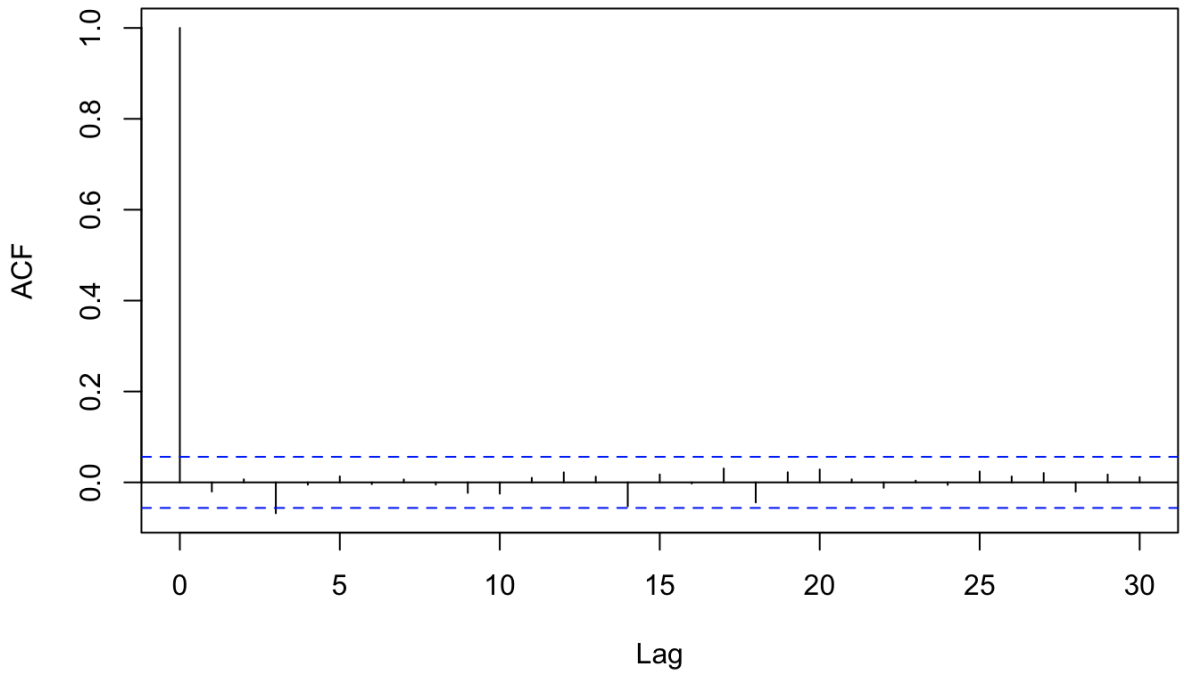
Finally, it is hypothesized that there is a linear relationship between price and carat. Two-dimensional scatterplot was used to visualize the possible relationship between the variables and decide on the regression model parameters. Noteworthy, as shown on the plot below, the relationship is not strictly linear, which is especially critical for the larger values on the x-axis.



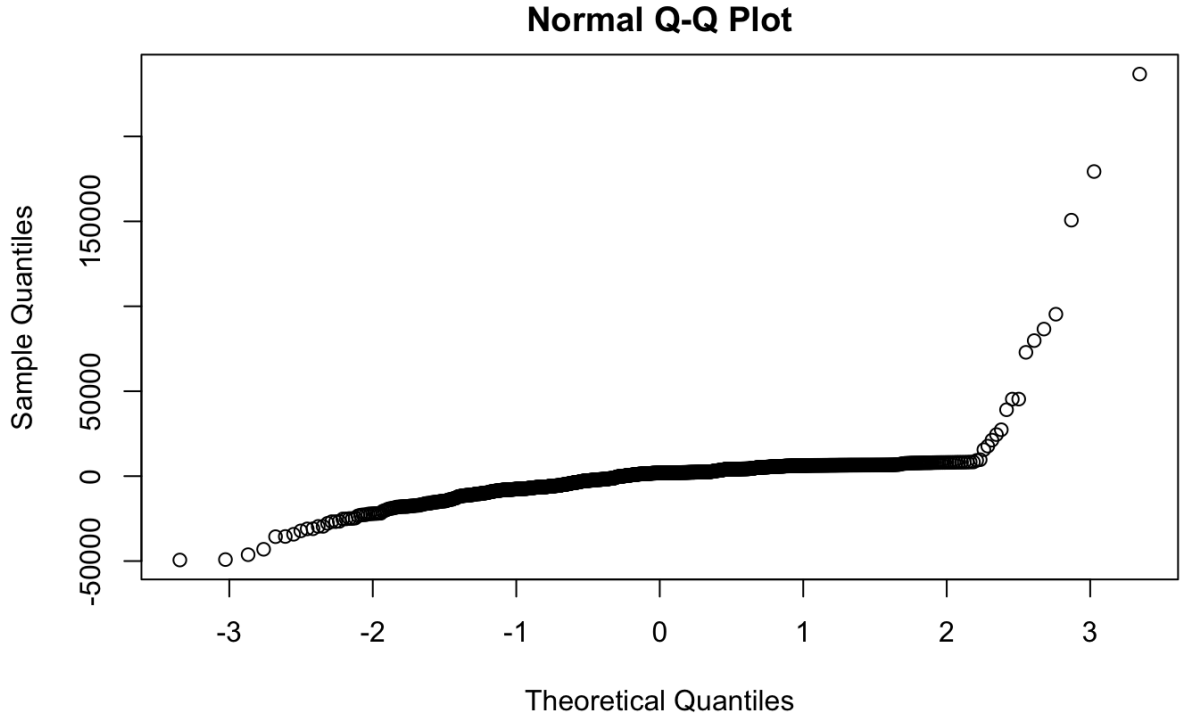
Residual plot from the “base_model” model was used to test the assumptions of a simple linear regression. The plot below shows the predictor's residuals against a fitted dependent variable: overall mean is likely not zero and variance seems to be non-constant.



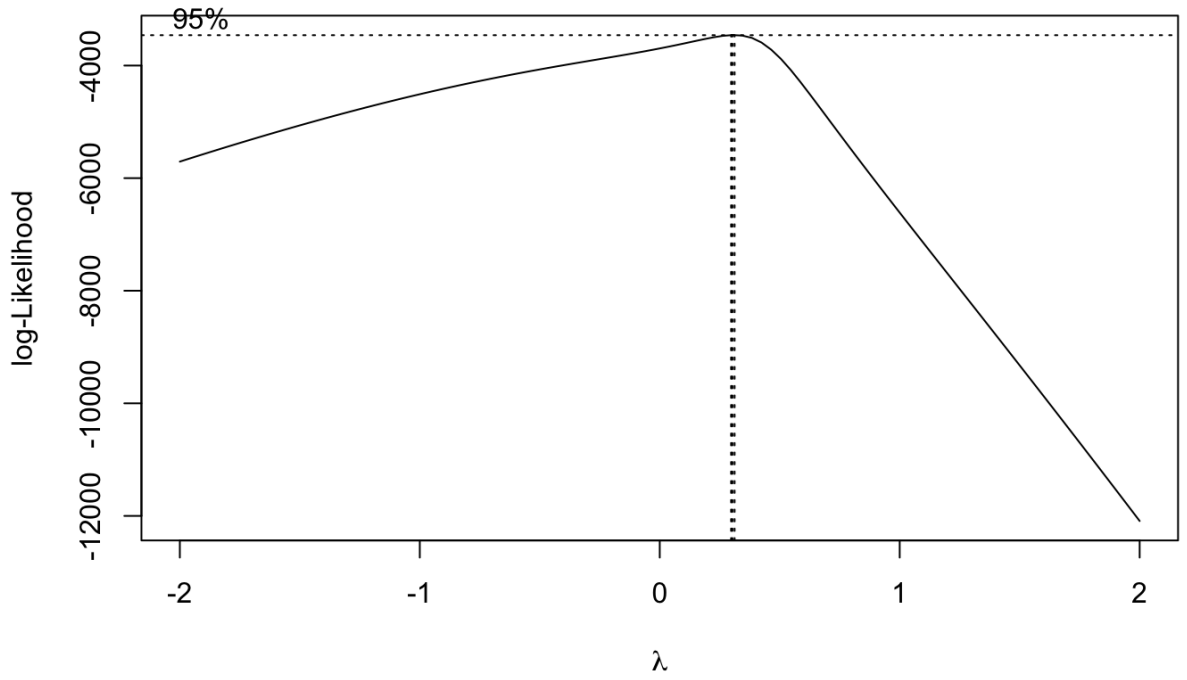
Apart from the obvious heteroskedasticity, residuals seemed to be auto-correlated and non-normally distributed, as shown on the following plots below. First, ACF shows somewhat significant correlation at the 3d lag:



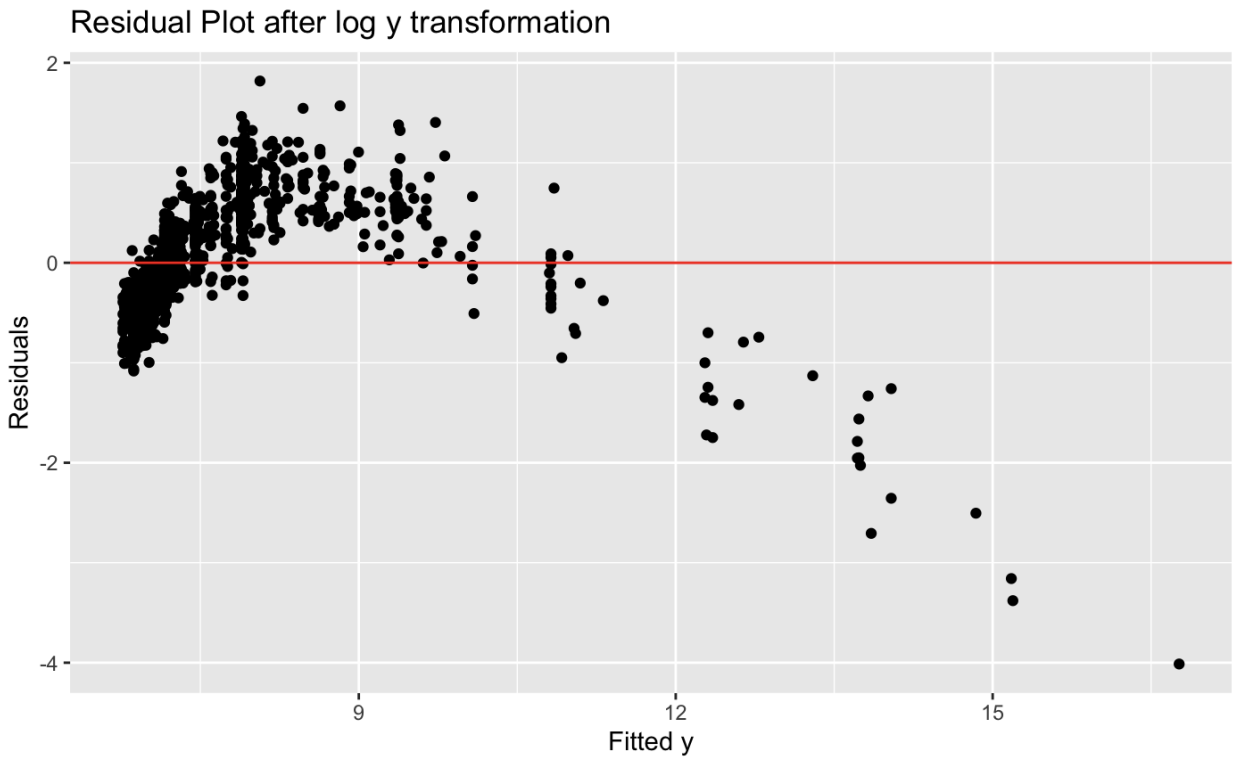
Moreover, sample quantiles deviate from the theoretical expectations:



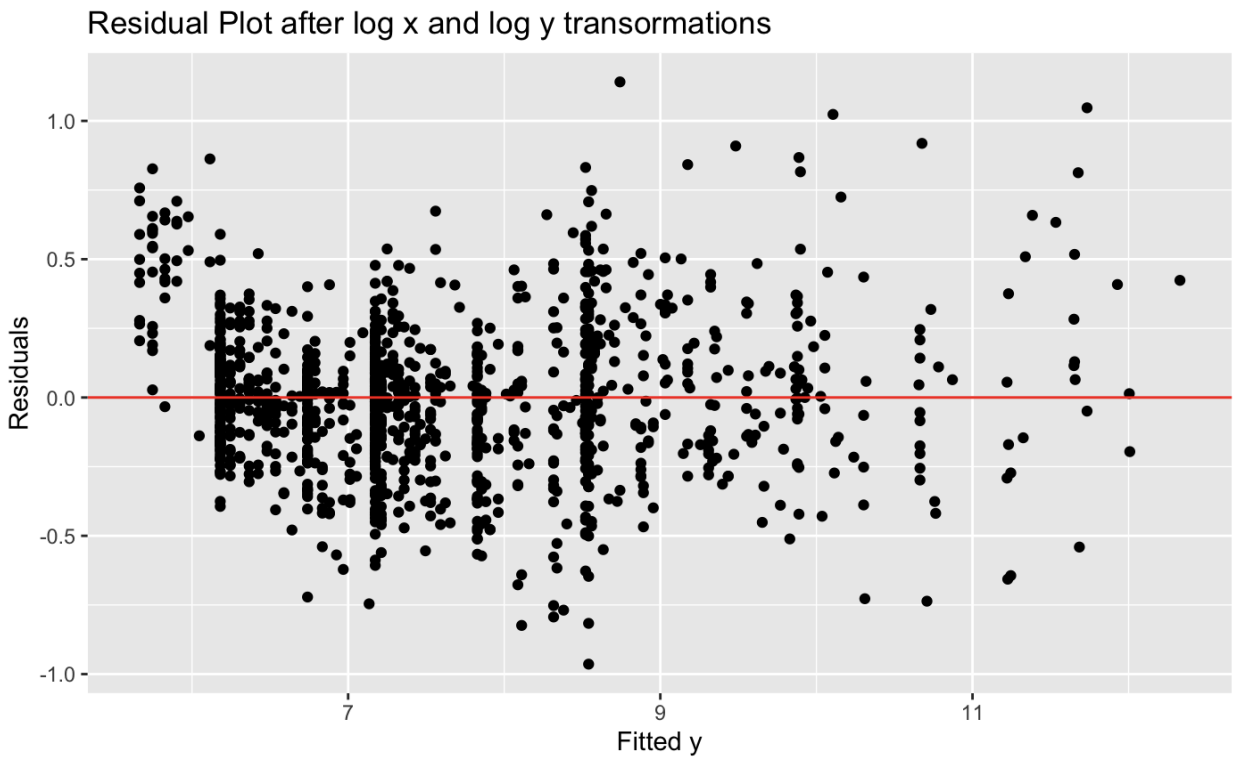
In order to mitigate non-constant variance, log transformation was applied to the dependent variable, Box Cox analysis below:



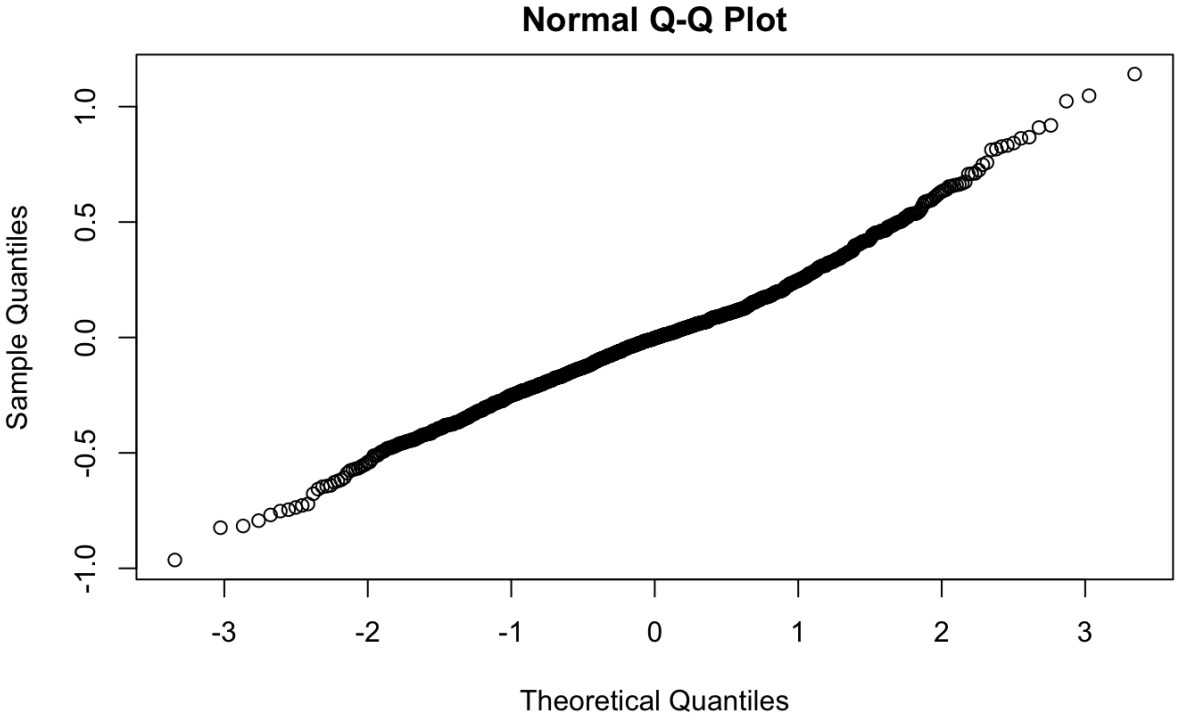
Taking the log of the dependent variable improved variance of the residuals, but did not mitigate the non-zero mean:



In order to improve the overall expectations, i.e. to disperse the cluster of residuals above the horizontal line, another log transformation was performed on the predictor variable. The resulting plot of the residuals is presented below.



The final log-log model met linear regression assumptions reasonably well.



Below is the summary output for the final log-log linear regression model:

```
Call:
lm(formula = price_log ~ carat_log, data = df)

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 ***
carat_log   1.944020   0.012166   159.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2761 on 1212 degrees of freedom
Multiple R-squared:  0.9547,    Adjusted R-squared:  0.9546
F-statistic: 2.553e+04 on 1 and 1212 DF,  p-value: < 2.2e-16
```

The summary output of the log-log model shows good predictive power as measured by R-squared of 95% and good overall significance, measured by very small F-statistics. The resulting intercept is positive and the slop highly statistically significant. The formula for the model can be written as log(Price) = 8.5 + 1.9 * log(Carat). Contextually, this can be interpreted as starting from the constant 8.5%, every additional 1% increase in Carat size leads to approximately 1.9% increase in diamond Price. In other words, the log-log model shows elasticity of diamond price with respect to carat size.