STAT 6021: Project 1

Greg Madden, Maxwell Levinson, Andrew Setaro, and Trey Hamilton

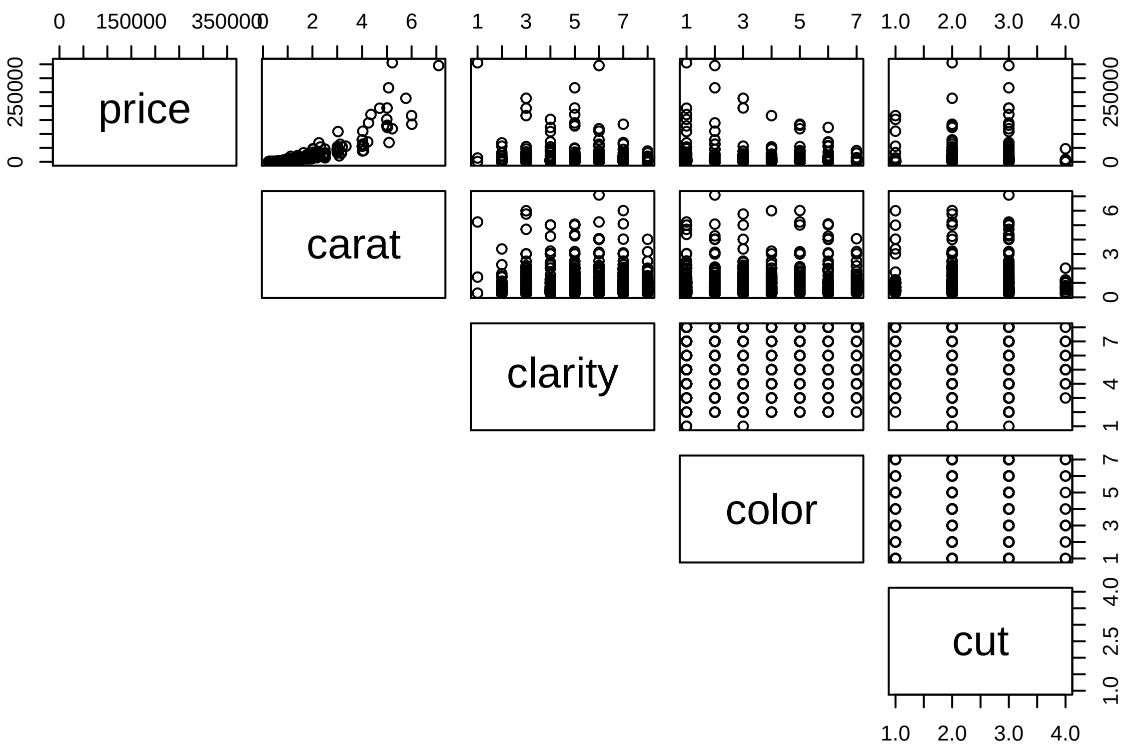
3/3/2022

We have been approached by Blue Nile to perform the following tasks:

1. Use data visualizations to explore how price is related to the other variables (carat, clarity, color, cut), as well as how the other variables may relate to each other. Address the various claims on the diamond education page on Blue Nile.
2. Fit an appropriate simple linear regression for price against carat.

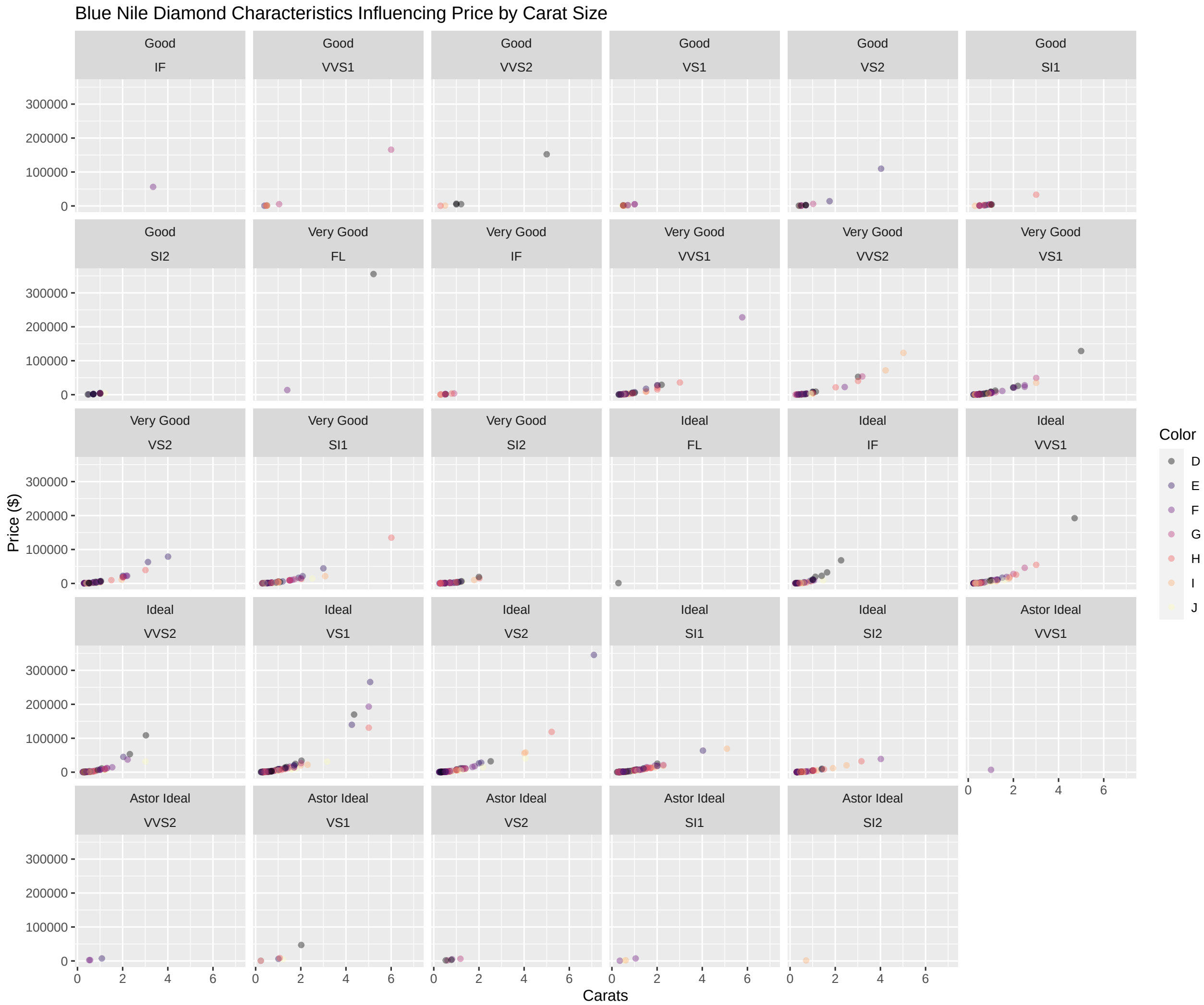
Description of the Data and Variables:

Plotting a pair-wise scatterplots for the data below:



Note that the x-axis above corresponds to increasing desireability of the factored categorical variables: clarity, color, and cut. In the above scatterplot matrix how price appears to have the clearest linear relationship with carats.

Plotting the response variable of interest (price) against carat, faceted by cut and clarity, with color indicating different diamond color characteristics:

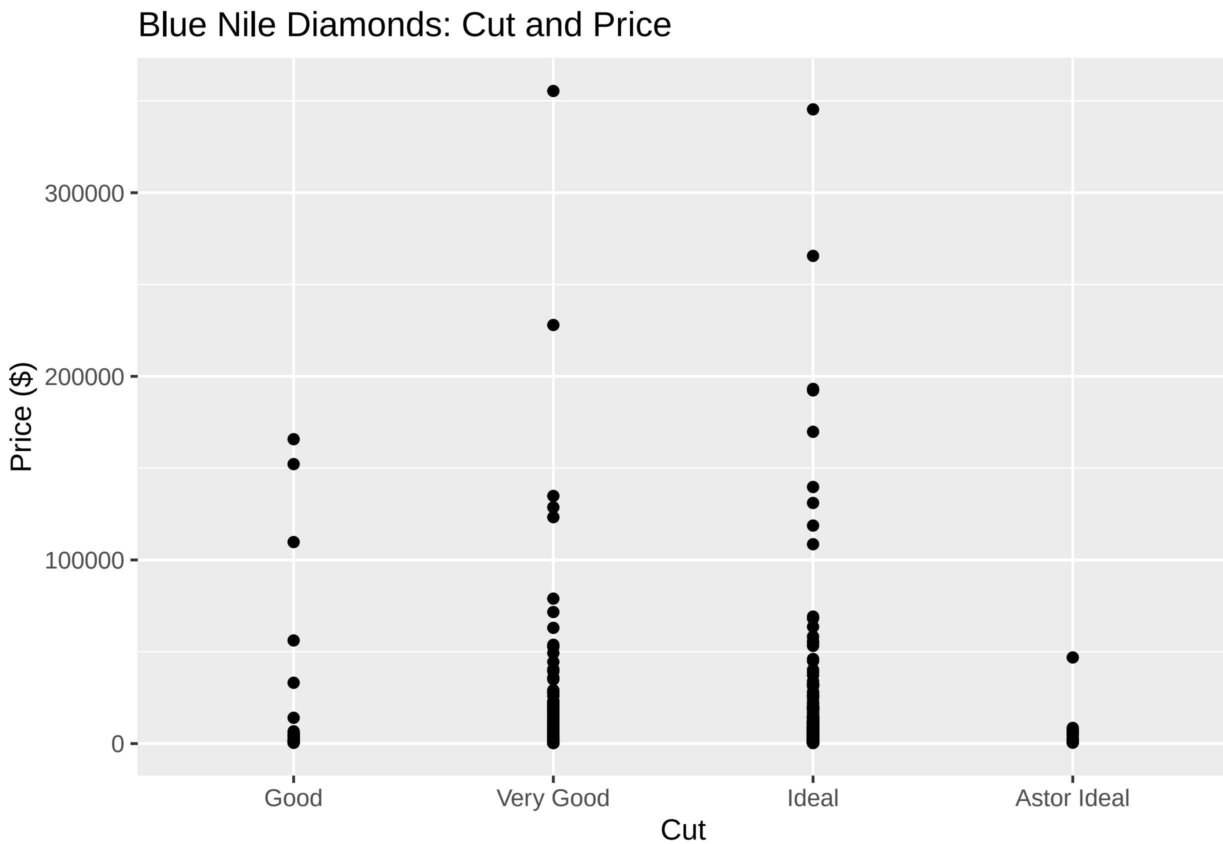


In the above plot, you can see that certain cut/clarity combinations (e.g., Good cut and VVS2 clarity) have different slopes in terms of their price ~ carat relationships. For example, ideal cut with FL (Flawless) clarity appears to have a higher price per additional carat size than the Ideal Cut with SI2 (Slightly Included). In addition, more desireable colors (i.e.D-F) appear to cluster at the lower price ranges.

Addressing various claims on the diamond education page on Blue Nile:

- Cut: https://www.bluenile.com/education/diamonds/cut +"A diamond's cut refers to how well-proportioned the dimensions of a diamond are, and how these surfaces, or facets, are positioned to create sparkle and brilliance. For example, what is the ratio of the diamond's diameter in comparison to its depth? These small, yet essential, factors determine the diamond's beauty and price."

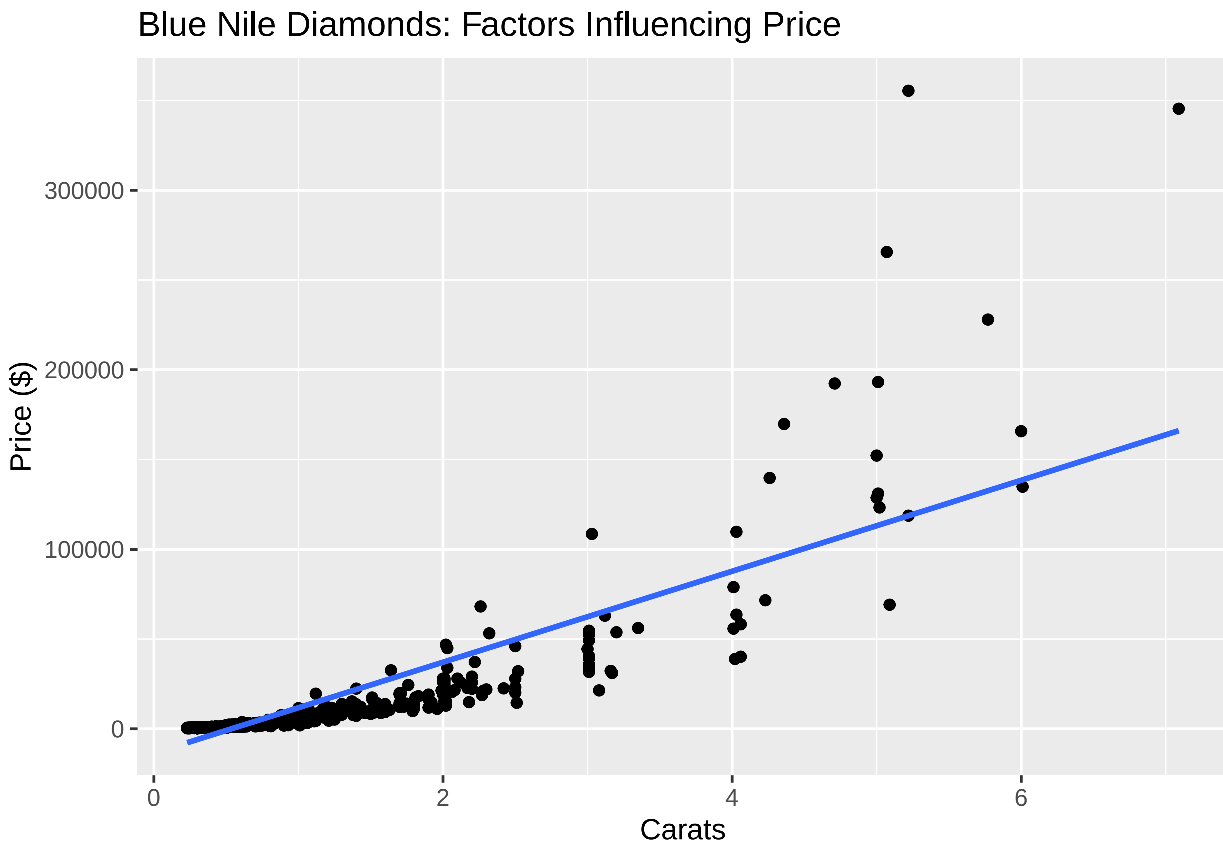
Assertion above is that better cuts correlate with higher price. Let's check the scatterplot to see if that bears out in the data:



As shown above, increasing quality of diamond cut by itself does not seem to have a linear relationship with price, contrary to the Blue Nile's claim.

Description of how we fitted the regression of price against carat:

First checking a scatterplot for Price ~ Carat:



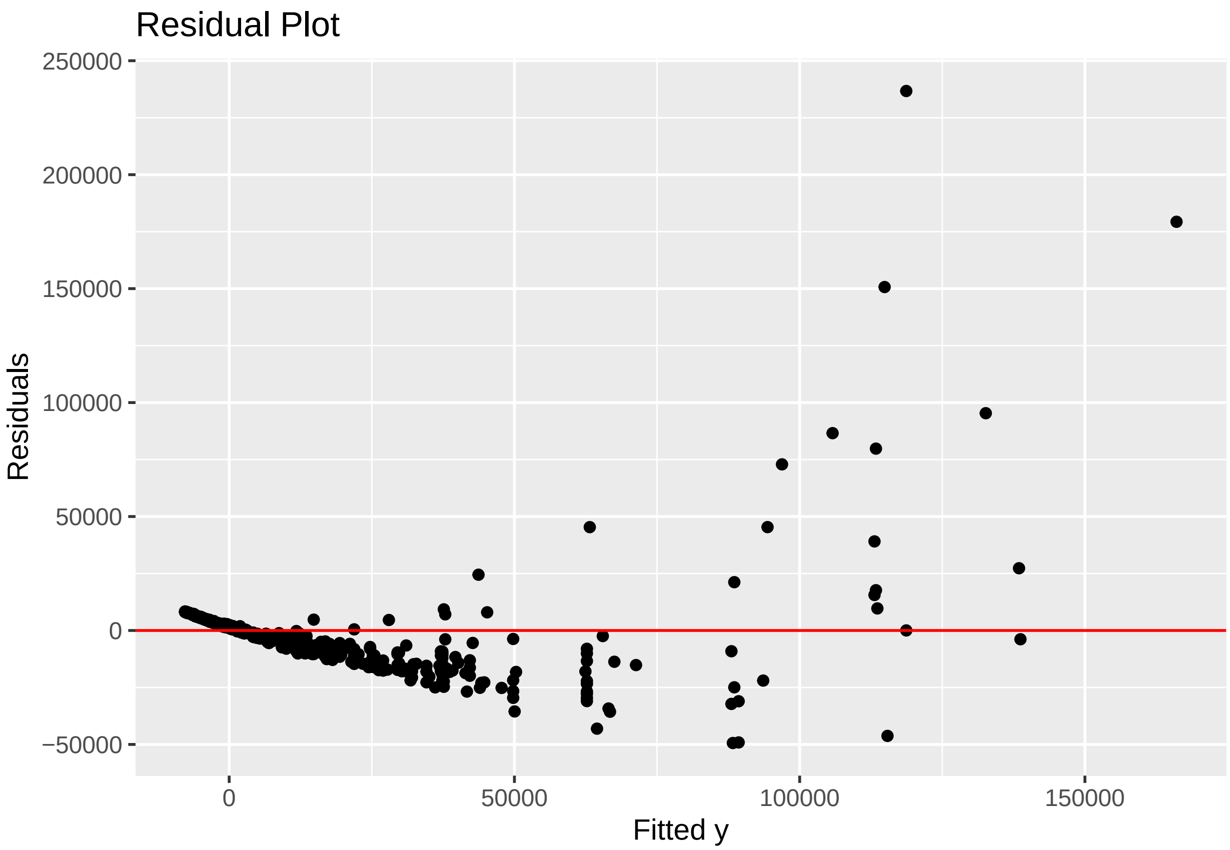
Note that the relationship between Price~carats is roughly linear, however the variance of price over carats does not appear to be constant.

Fitting a preliminary simple linear regression model:

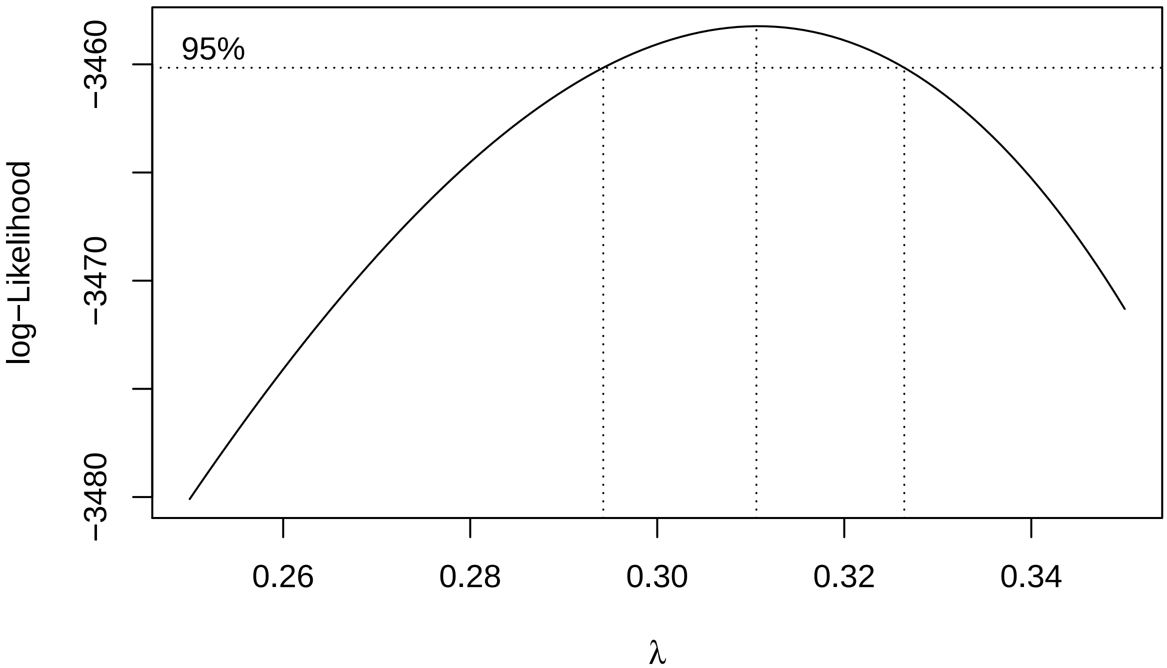
```
##
## Call:
## lm(formula = price ~ carat, data = Data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49375  -5048   1867   4965 236711
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -13550.9      559.7  -24.21 <0.0000000000000002 ***
## 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
```

Plotting the residuals.

Constant variance and mean of error = 0 assumptions do not appear to be met.

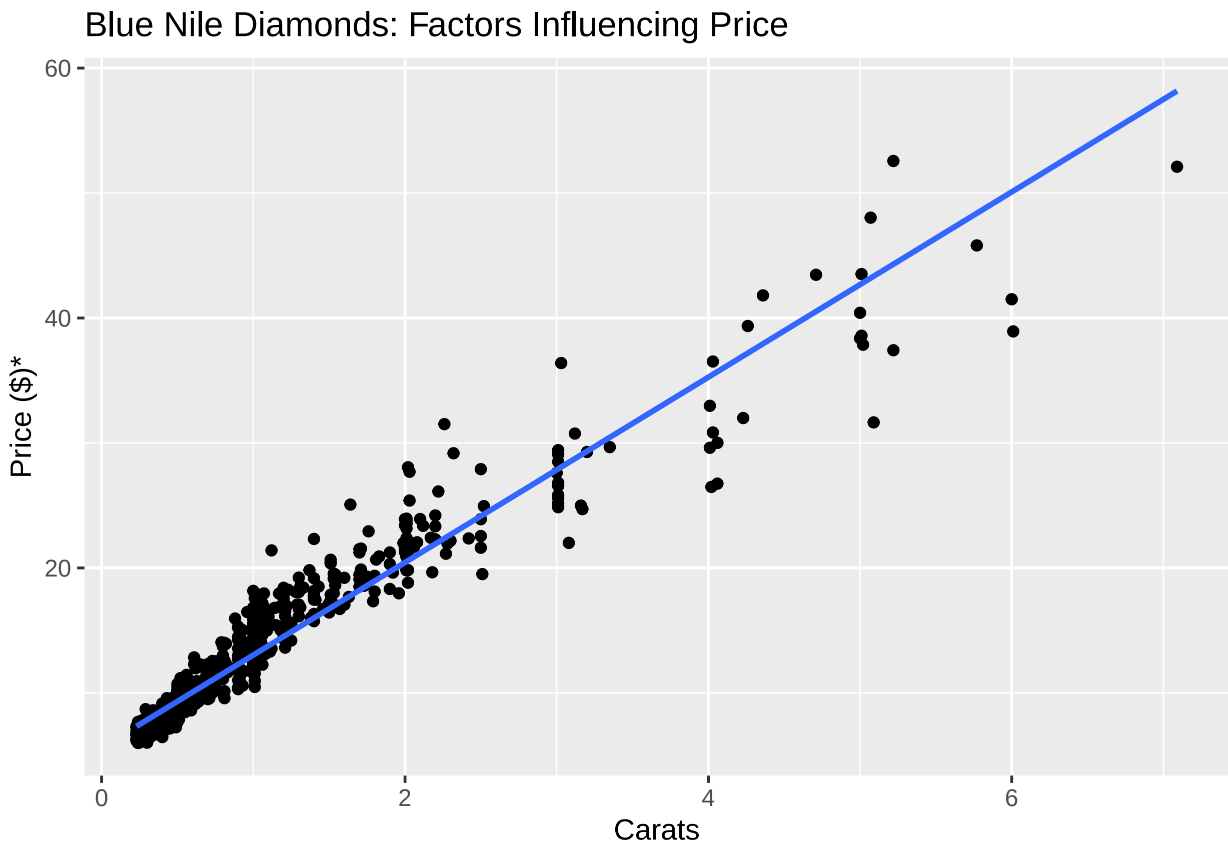


***Variance is not constant so attempting to transform y first. Will start with boxcox plot to see what the optimal lambda may be.

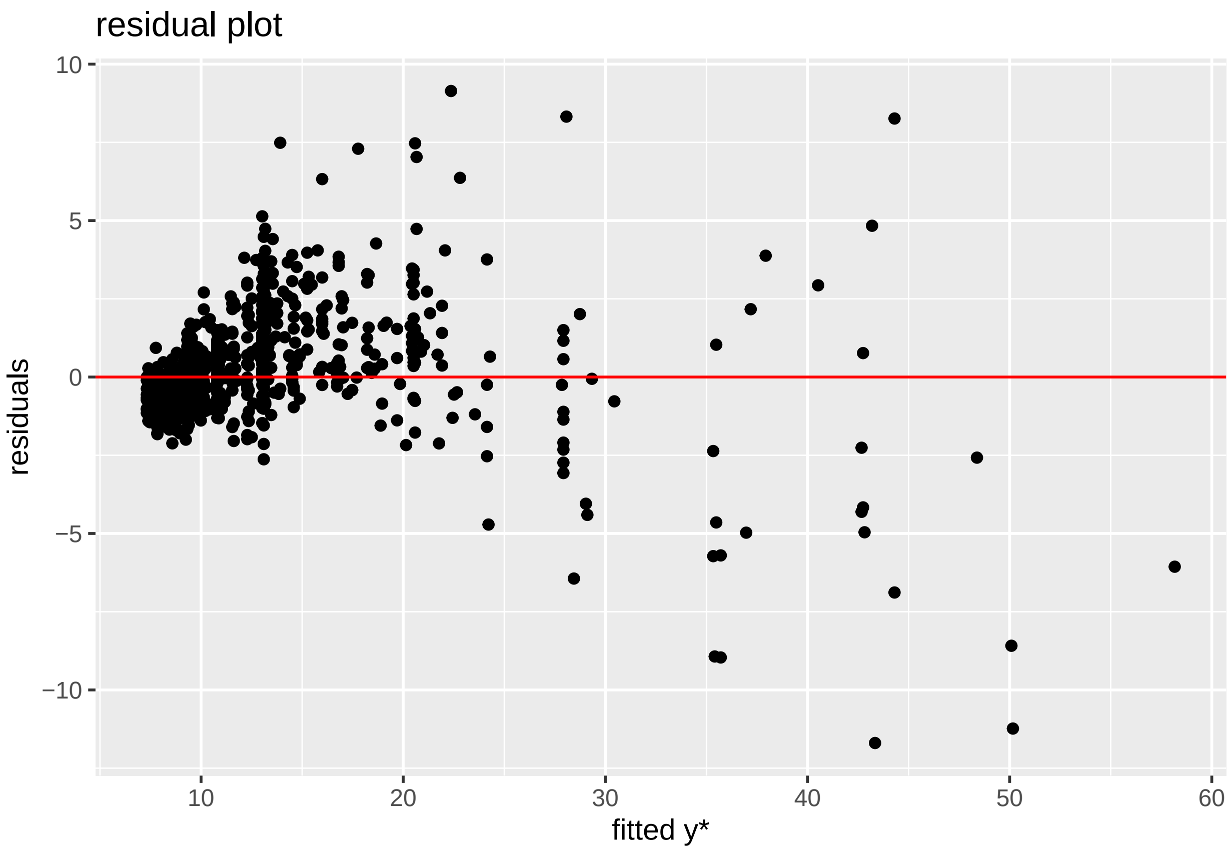


It appears that a lambda of 0.31 would be appropriate.

Replotting the scatter plot with the transformed y variable.



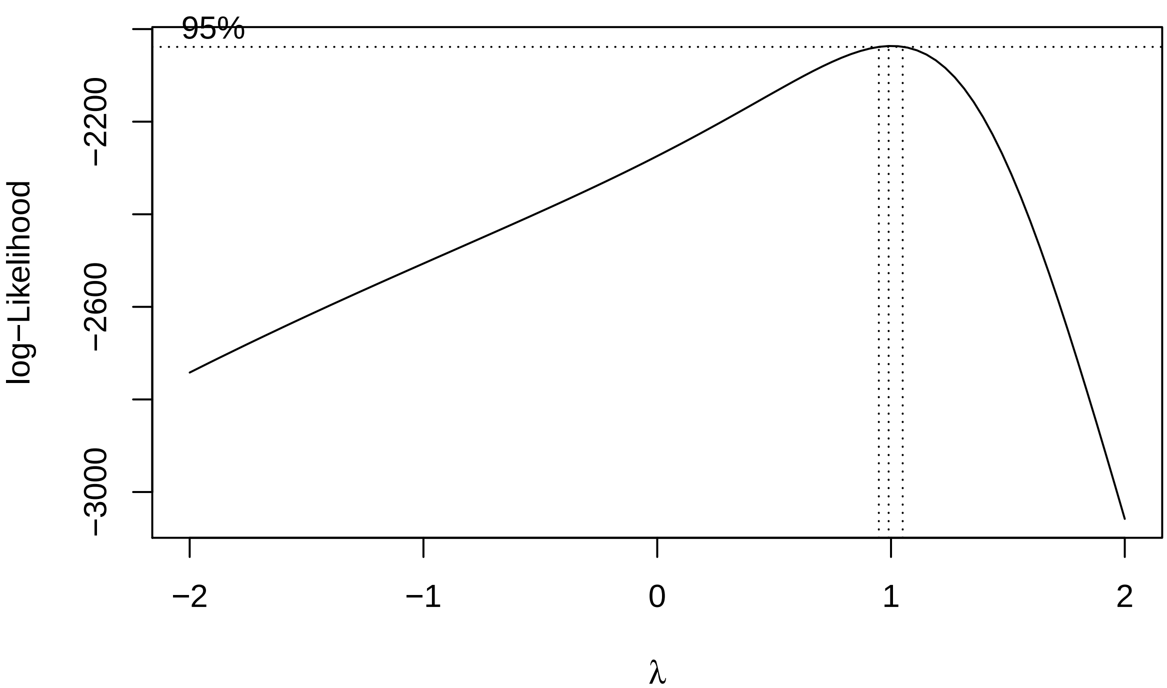
New residual plot:



Note: variance looks better but mean of errors still not equal to zero over x so will attempt to transform the x variable. Given the curved appearance, will try a square root transformation.

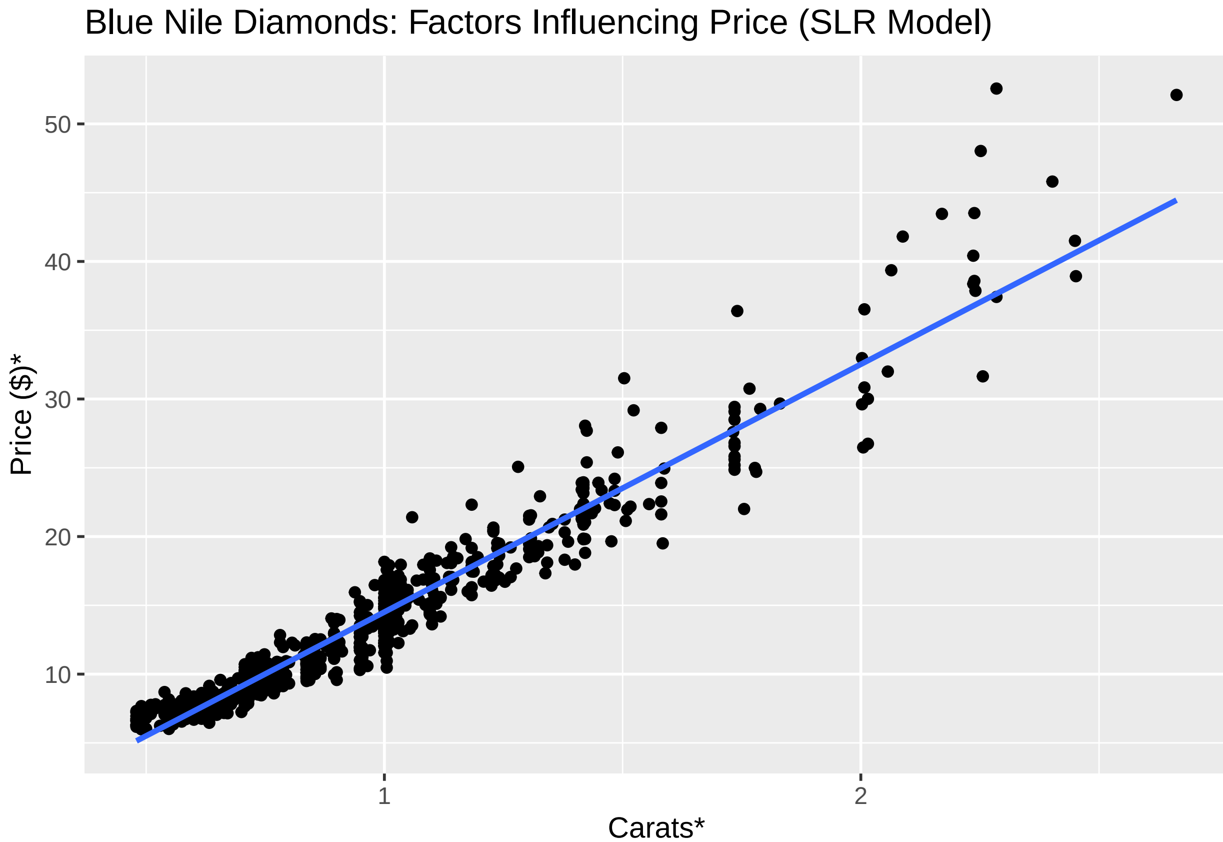
Confirming that the box cox plot looks better.

Boxcox plot:

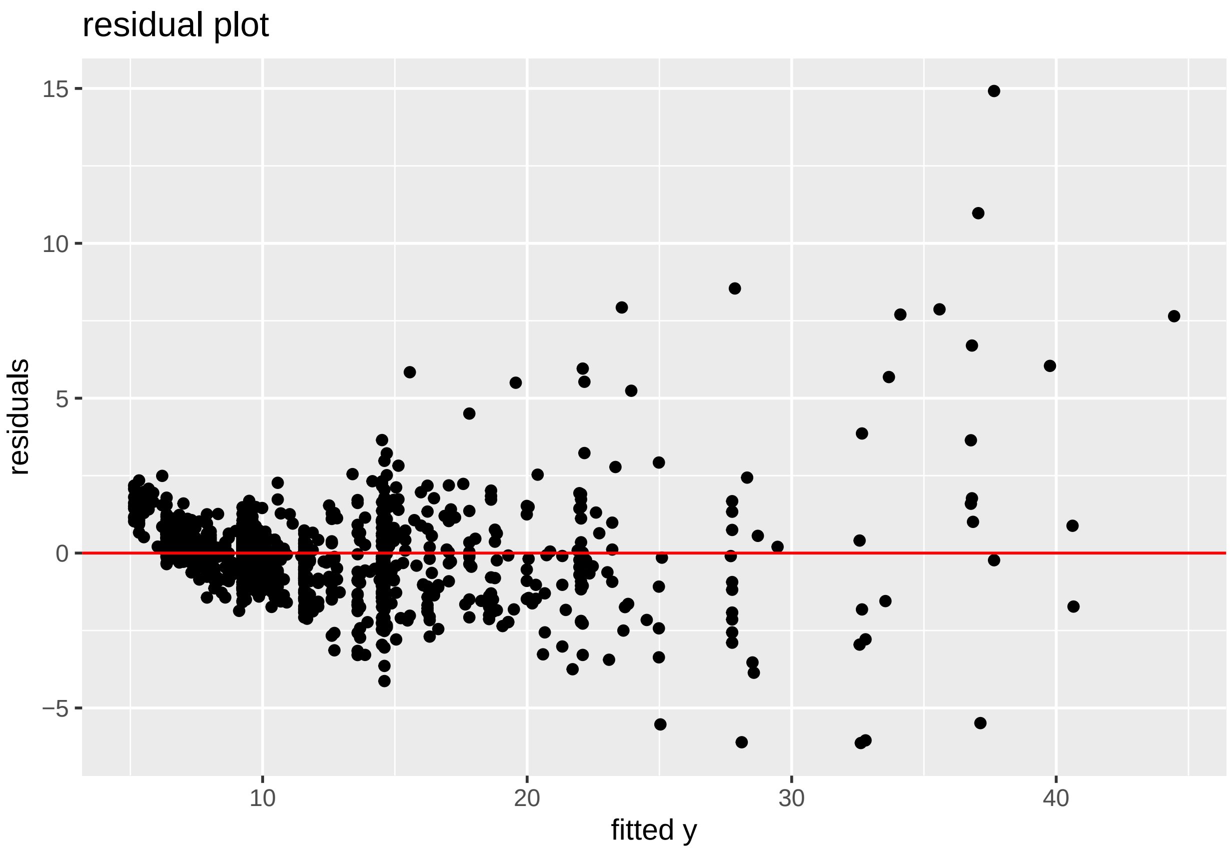


Now I see that confidence interval includes 1. Next step is to consider whether or not to transform x.

Plotting the scatterplot using the transformed x (sqrt(x)) and y (y^0.31) variables.

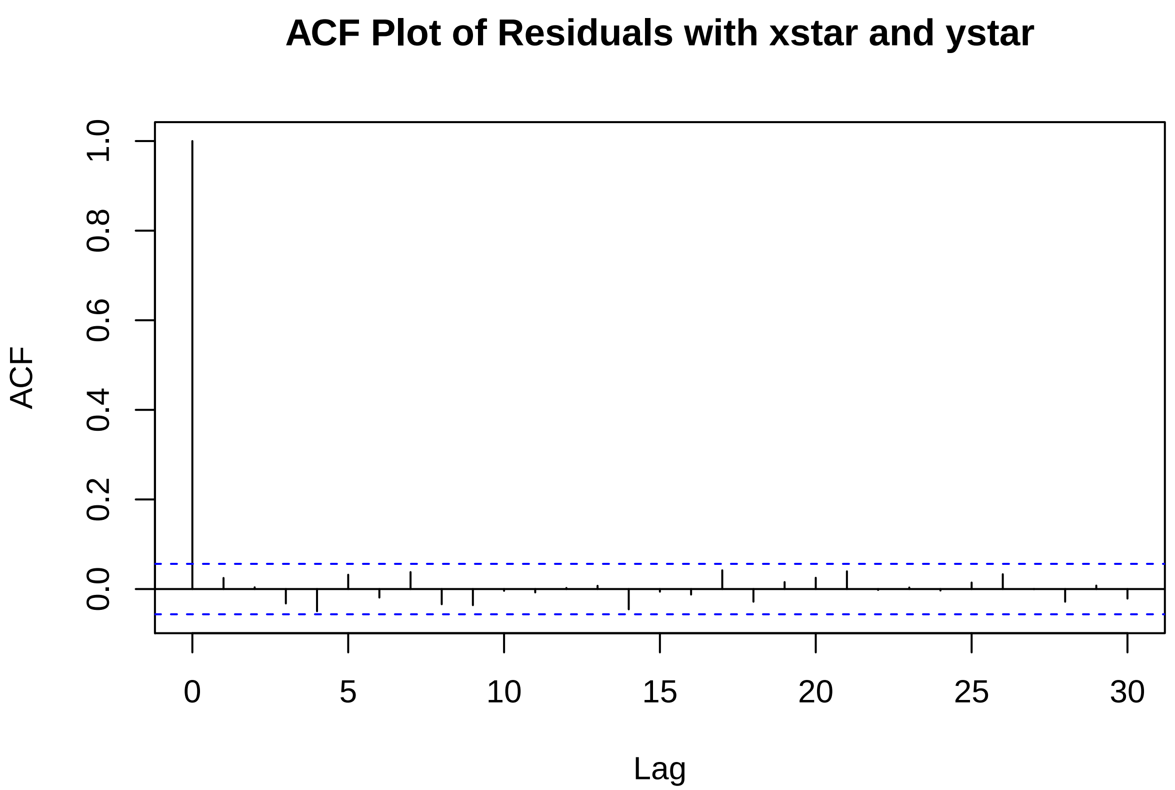


Fitting new model and creating another residual plot:



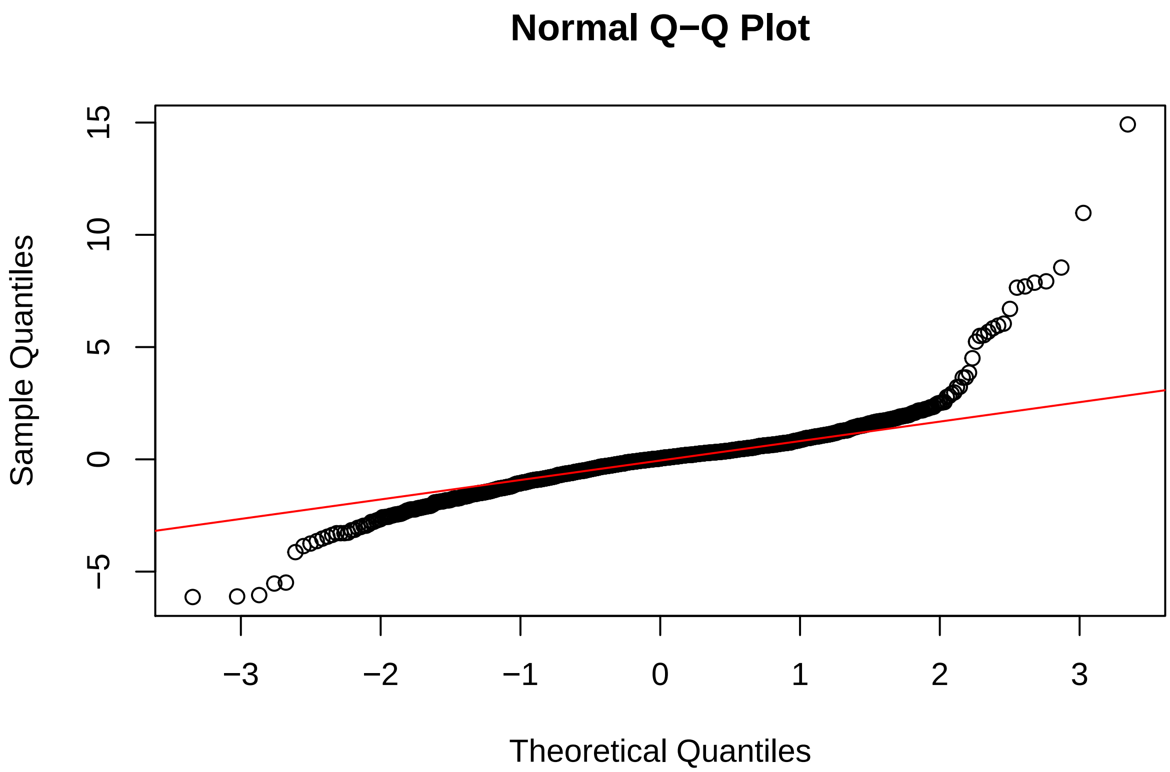
Not a perfect fit but overall improved and adequate for prediction.

ACF Plot:



Errors do not appear correlated to each other.

QQ Plot:



Normality assumption is not met but this may be the least important.

Summarizing the final model below:

```
##
## Call:
## lm(formula = ystar ~ xstar, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1320 -0.6377  0.0373  0.5315 14.9172
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  -3.4936     0.1137  -30.73 <0.0000000000000002 ***
## xstar        18.0085     0.1261  142.87 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.436 on 1212 degrees of freedom
## Multiple R-squared:  0.9439, Adjusted R-squared:  0.9439
## F-statistic: 2.041e+04 on 1 and 1212 DF,  p-value: < 0.00000000000000022
```

Conclusions:

In conclusion, the blue nile data contain various characteristics for individual diamonds including carat (weight of the diamond), clarity, color, and cut; clearer, colorless, ideal cut, and heavier diamonds being more desireable and thus correlating with higher price. Certain claims by the website such as that the quality of a diamond's cut determines the price - do not appear to be supported by the data.

We fitted an appropriate simple linear regression for price against carat. Doing so required transformation of both the x and y variables to approximate the assumptions required for linear regression.

We can summarize our final regression equation for diamond price (y) as the following:

$y^* = 18.0085x^* - 3.4936$

where $x^* = \sqrt{x} \textit{ and } y^* = y^{0.31}$

As you can see in the model summary printed above, the t test for slope P-value for xstar (<0.0000000000000002) indicates that we can reject the null hypothesis that the slope is equal to zero; in other words, carat is a significant predictor of diamond price.

Executive Summary: