STAT 6021: Project 1

Taylor Tucker, Mahin Ganesan, Wyatt Priddy

2023-10-20

Section 1: High Level Results of the Analysis

Our report studied the relationship between various factors associated with the grading of diamonds and the price of the diamonds. We found that the carat weight, a carat being roughly equal to 0.2 grams, is the most important factor in determining the price of the diamond. Thus, we analyzed the relationship between the carat weight of the diamond and its price. We found that, if the carat weight of a diamond increases by 1%, then the price of the diamond will increase by between 1.92% and 1.97%.

Section 2: Data Description

Inspection of the data shows 5 variables being utilized within the *diamonds4.csv* data set with 1,214 observations.

carat	clarity	color	cut	price
0.51	SI2	I	Very Good	774
0.93	IF	H	Ideal	6246
0.50	VVS2	D	Very Good	1146
0.30	VS1	F	Ideal	538
0.31	SI1	F	Ideal	502

The information in this section mostly comes from the following source:

Blue Nile. "The 4Cs of Diamonds," n.d.

Dataset Variables

The quality, and consequently cost, of diamonds is measured with the "4Cs of Diamonds": the `cut`, `color`, `clarity`, and `carat` weight, which all appear in the data set, along with the variable `price`.

The `cut` of a diamond measures how "well-proportioned a diamond's dimension are", as well as, once being processed, the "balance and brilliance of its facets". The facets of a diamond are the sides which are formed on the diamonds surface to give it its reflective, glittery appearance. This variable is stored as a character in the data set.

`Distinct value counts of ‘cut‘ variable:`

Cut	Frequency
Astor Ideal	20
Good	73
Ideal	739
Very Good	382

* *descriptions of* `cut` *categorical variable*

The `color` of a diamond refers to the colorlessness of the diamond. Diamonds can often be yellowish, or off-white, depending on impurities in the source material. Therefore, the purest of diamonds are the ones which are completely colorless. This variable is stored as a character in the data set.

`Distinct Value Counts of ‘color‘ variable:`

Color	Frequency
D	207
E	181
F	223
G	198
H	148
I	167
J	90

* *descriptions of* `color` *categorical variable*

The `clarity` of a diamond measures the cloudiness or impurities that can occur during diamond formation, due to impurities or foreign material. This variable is stored as a character in the data set.

`Distinct Value Counts of ‘clarity‘ variable:`

Clarity	Frequency
FL	3
IF	49
SI1	243
SI2	165
VS1	233
VS2	214
VVS1	149
VVS2	158

* *descriptions of* `clarity` *categorical variable*

The `carat` weight refers to the physical weight of the diamond, and is often the most important indicator for showing how large, and subsequently how valuable, a diamond is. One carat is roughly equivalent to 0.2 g, or 0.00705 oz. This variable is stored as a double in the data set.

Finally, the `price` refers to the price ($USD) that the particular diamond sold for. This variable is stored as an integer in the data set.

Created Variables

During the analysis portion, we had to create variables to meet the assumptions of a simple linear regression, namely `xstar` (x^*) and `ystar` (y^*). The variable `xstar` is the log of the `carat` variable, calculated as $x^* = ln(carat)$. The variable `ystar` is the log of the `price` variable, similarly calculated as $y^* = ln(price)$

Visualizations and Descriptions

To better understand how `price` is related to the other variables in the data set visualizations were created to see the relationship between price and the variables `carat`, `clarity`, `color`, and `cut`. These graphs will give us an understanding of how the price changes with increasing measures of quality. The increase in quality, both for numeric and categorical data, can be seen on the following scale provided by BlueNile:

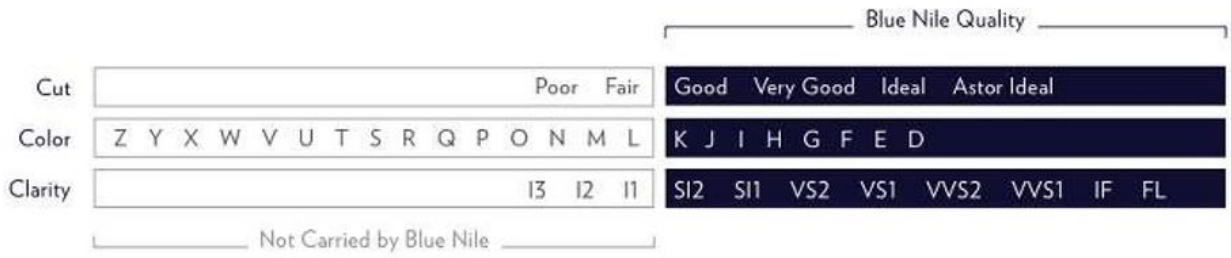
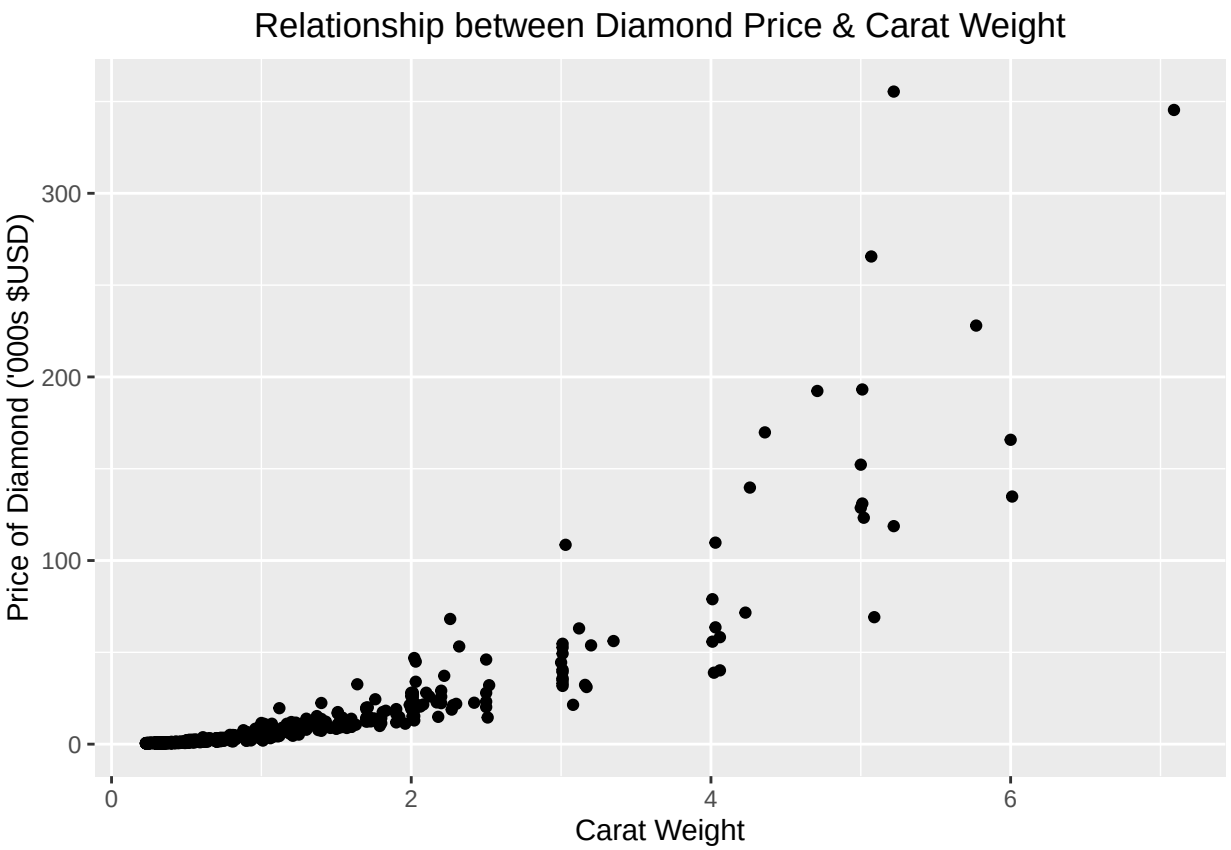


Figure 1: Quality Scale of Diamonds

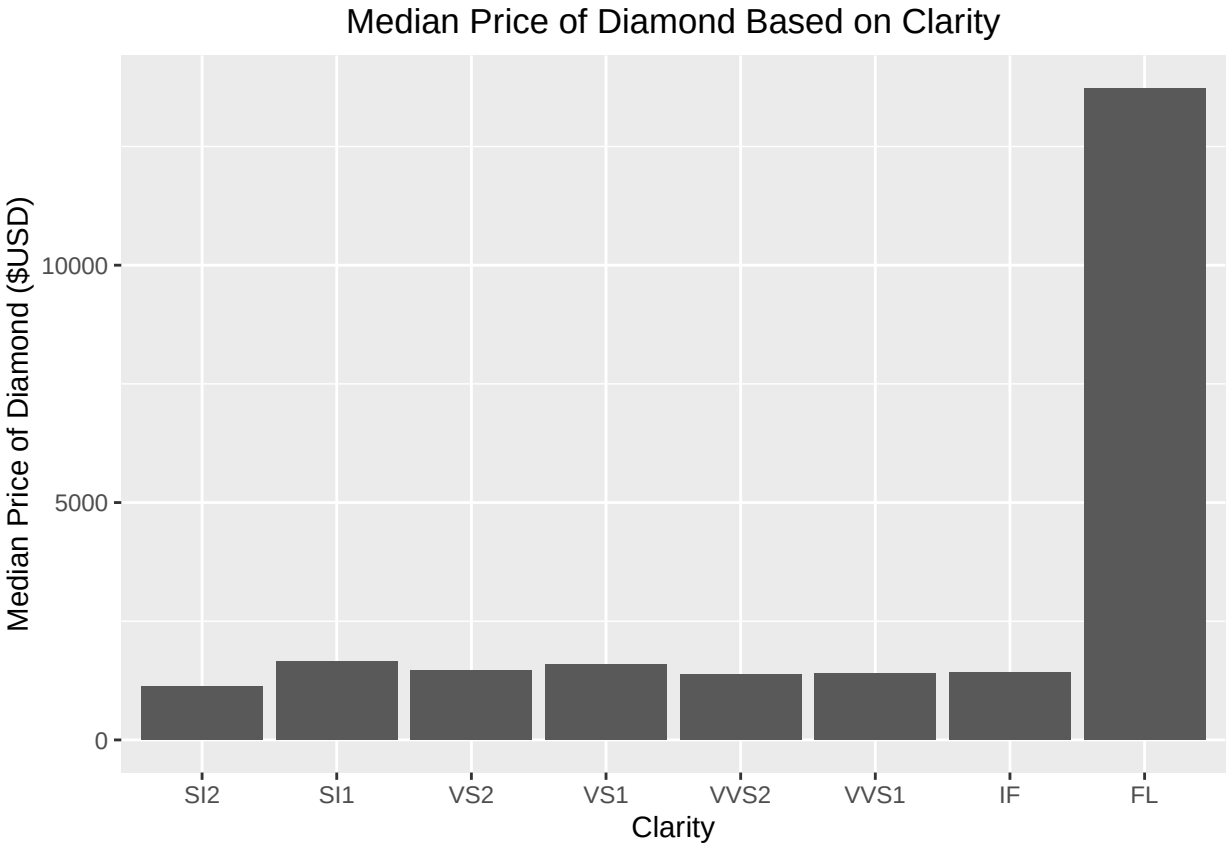
Price vs. Carat Weight

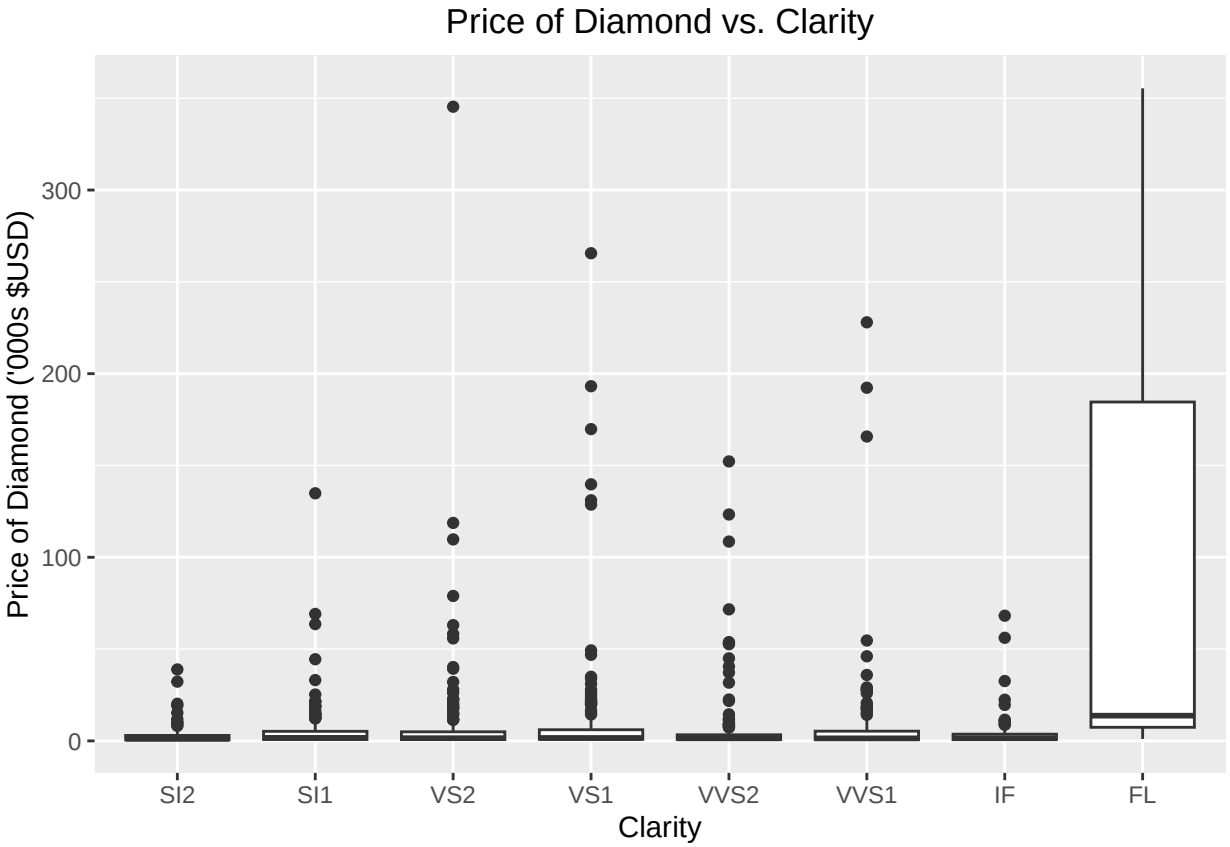


We can see from this scatter plot that the relationship between the price of a diamond and the carat weight appears exponential. As the carat weight increases, the price substantially increases. Carat weight is the only numeric variable in the data set other than price, and this method of using a scatter plot allows us to clearly see an upwards trajectory in the relationship between the carat weight and the price.

Price vs. Clarity

To order the categories for `clarity` to match the given scale, we applied data manipulation to refactor the levels of the variable based on the tiered quality levels provided. We then created a new data frame, where the diamonds were grouped by the clarity factor and the median price was applied. We plotted both boxplots of the `price` vs. the `clarity`, as well as bar plots of the median `price` vs. `clarity`. These plots will give us an idea about how `price` is related to `clarity`. The associated plots are:

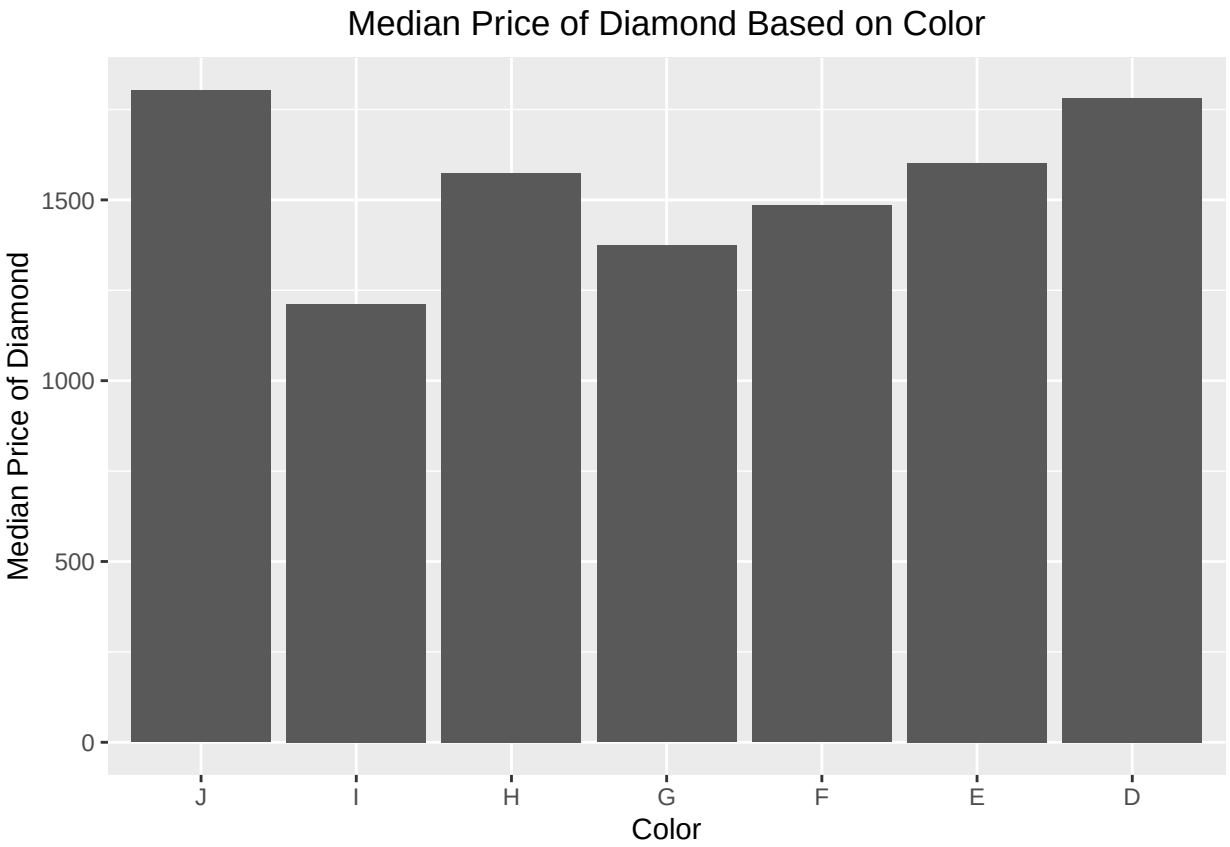


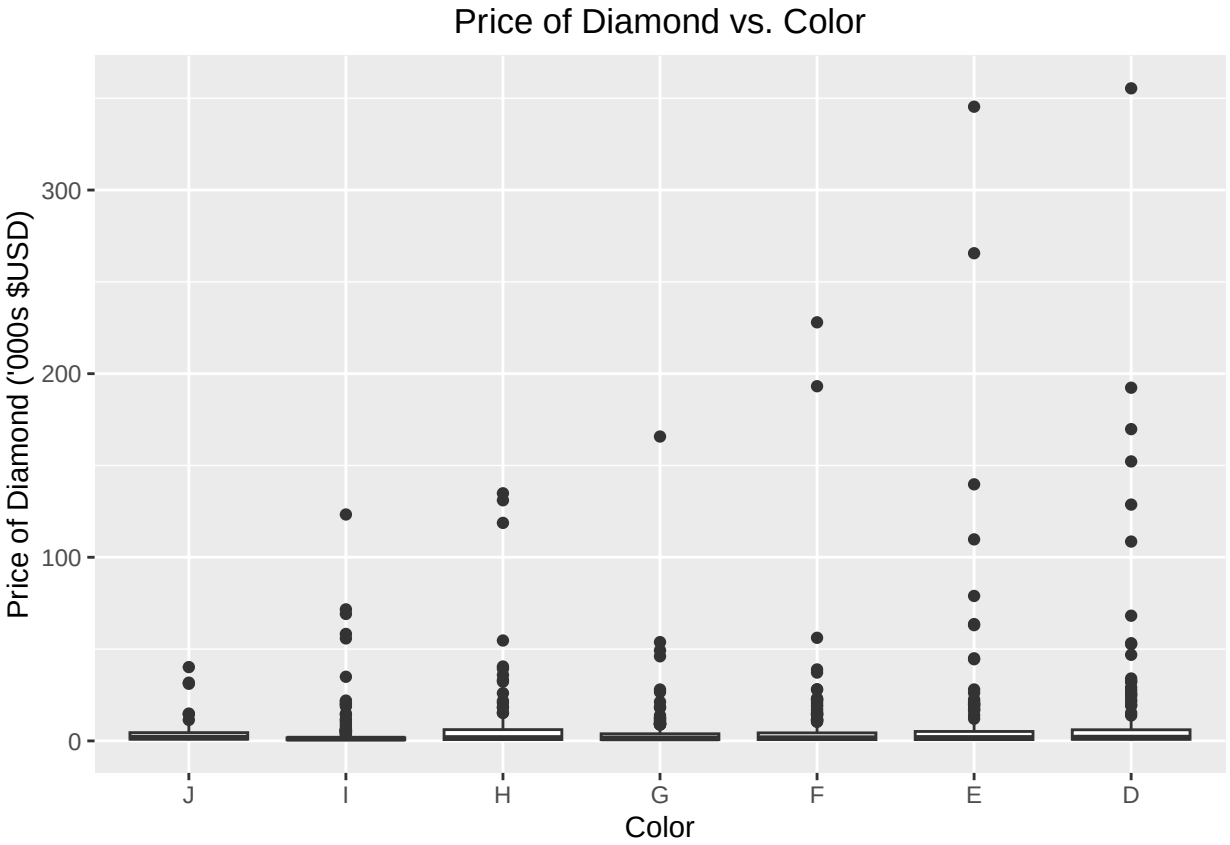


Shown in the bar plot above, the price increases dramatically when the clarity reaches the level of FL, which is flawless. We see that the median price of the diamond is not particularly affected by the clarity. However, based on the box plot, the upper end of the prices for each quality do appear to have an upwards trend.

Price vs. Color

To order the categories for `color` to match the given scale, we applied data manipulation to refactor the levels of the variable. One note regarding the refactoring: when refactoring using all of the potential `color` levels, we received warnings that the color level `K` was not found in the data. Thus, we removed that level from the refactoring. A new data frame was created where the data for diamonds were grouped by `color` and the median of the price was taken for each `color`. We plotted both box plots of the `price` vs. the `color` in addition to bar plots of the median `price` vs. `color`. These plots will give insights with regards to how `price` is related to `color`. The associated plots are:

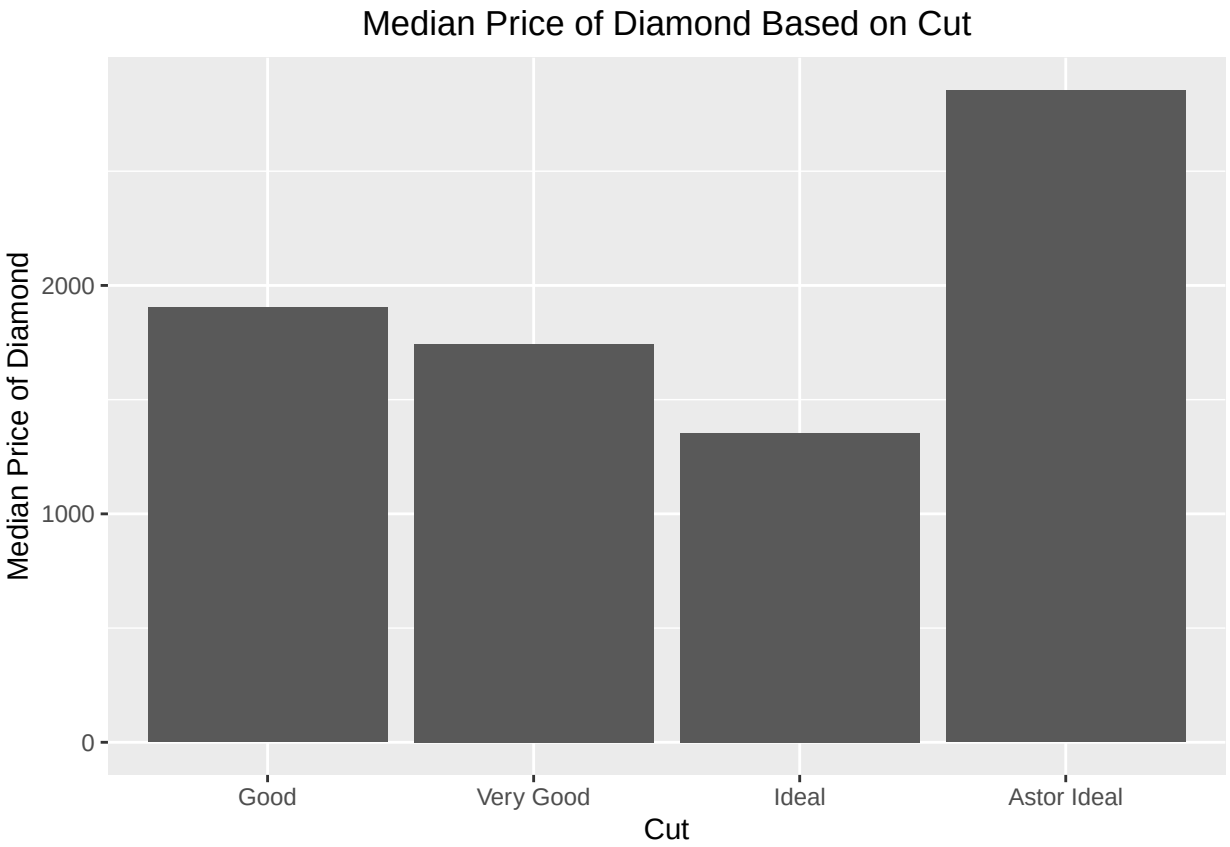

Median Price of Diamond Based on Color
Median Price of Diamond
1500
1000
500
0
J
I
H
G
F
E
D
Color

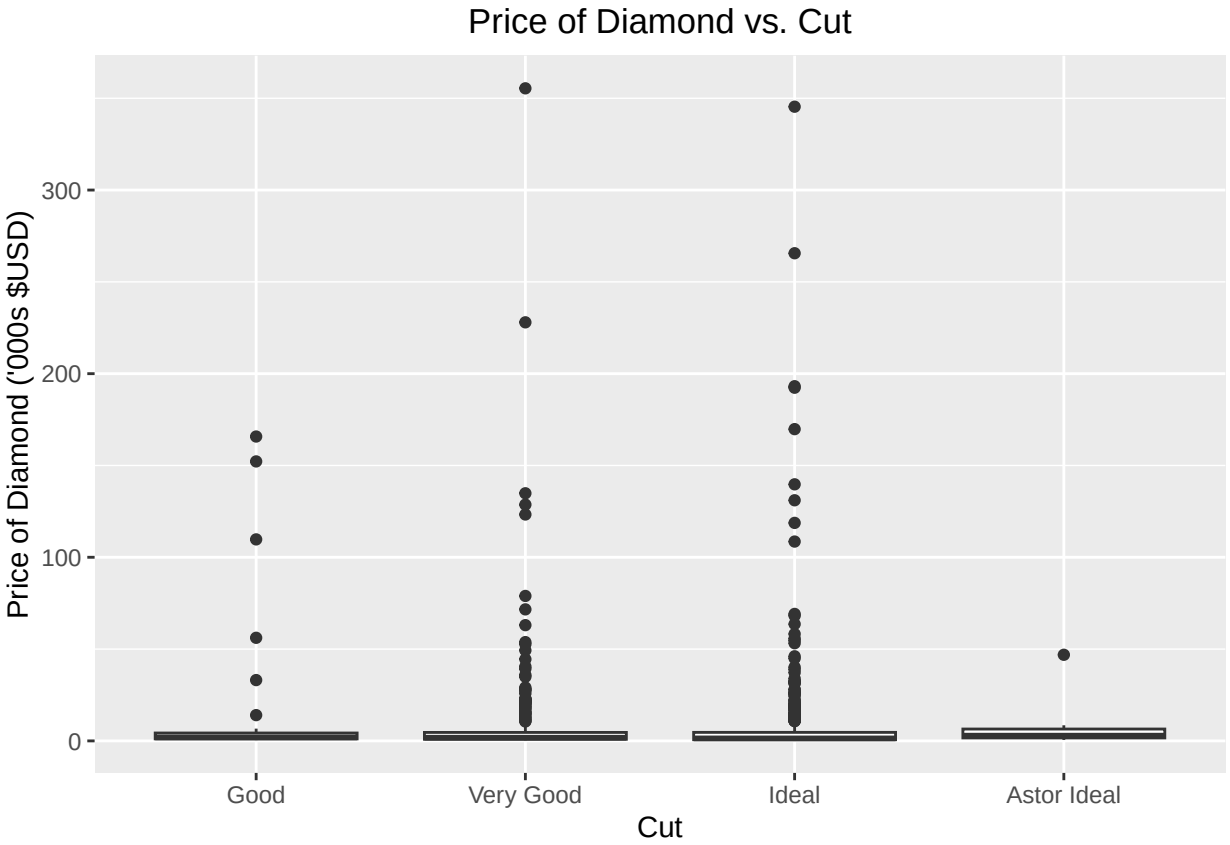


Based on the bar plot, we do not see much of a relationship between color and the price of the median diamond of each color. In fact, the highest median price for a diamond appears for the lowest color quality on the scale: `J`. Ignoring the median price for the color level `J`, there is an upwards trend leading us to believe that the distribution of `J's` prices are skewed to the higher prices. When looking at the box plots, the upper end of the diamonds sold at each color appear to have an upwards trajectory. That being said, the color does not appear to be a major factor in determining the price of a diamond, but there is certainly a loose positive relationship between `color` and `price`.

Price vs. Cut

To order the categories for `cut` to match the given scale, we refactored the levels of the variable. We created a new data frame, where the data for diamonds were grouped by the `cut`, and the median of the price was taken for each group. We plotted both box plots of the `price` vs. the `cut`, as well as bar plots of the median `price` vs. `cut`. These plots will give us an idea about how `price` is related to `cut`. The associated plots are:


Median Price of Diamond Based on Cut
Median Price of Diamond
2000
1000
0
Good
Very Good
Ideal
Astor Ideal
Cut

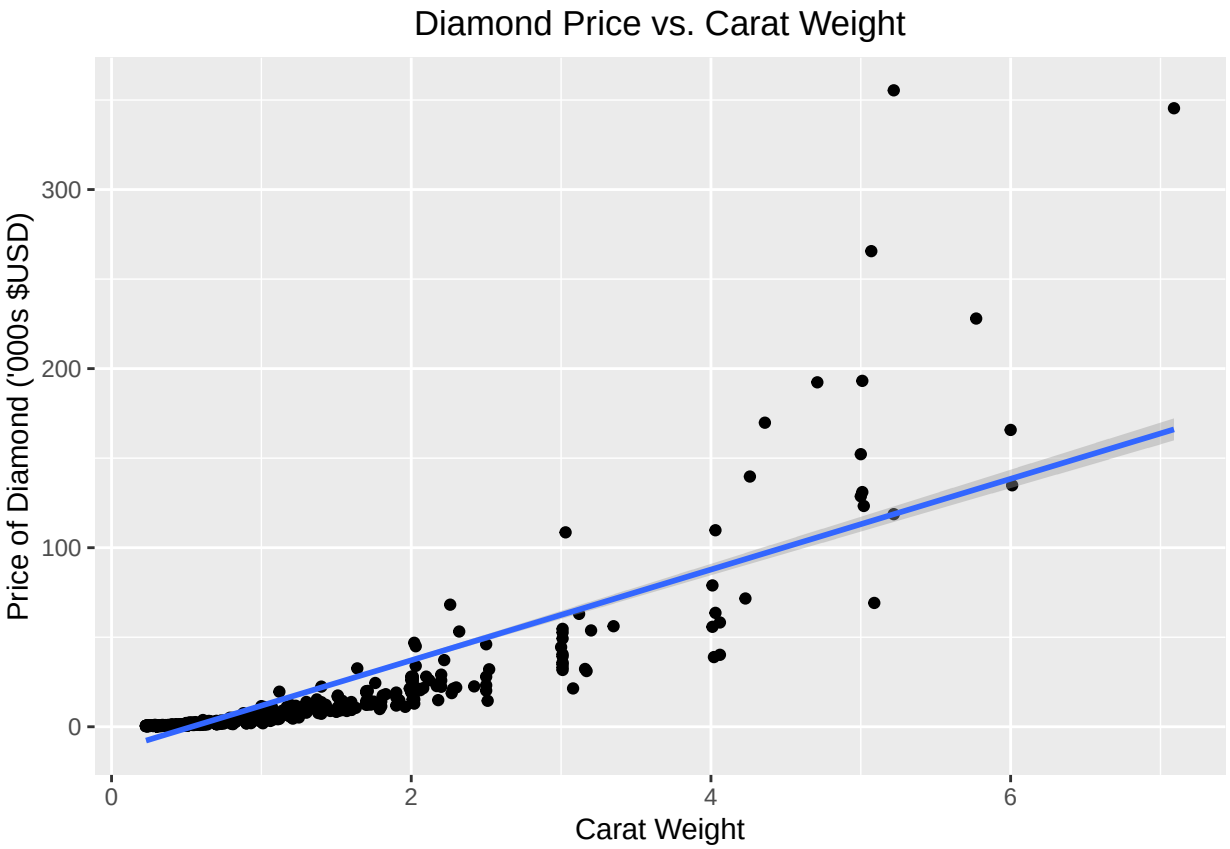


As seen with color and clarity, the bar plot does not show much of an upwards trend in median prices based on the cut of the diamond. In fact, with the exception of the Astor Ideal cut, the median prices decline as the cut quality improves. This trend continues when looking at the box plot. There is no discernible upwards trend in the upper end of prices as the cut of the diamonds improve. Based on this information, it appears that the cut is not a relevant variable when it comes to determining the price of a diamond

Based on the analysis above, the variable that best appears to relate to the price of a diamond is its carat weight, which agrees with the claims made on the BlueNile diamond education page. We will use the carat weight variable as the predictor variable moving forward in our regressions. However, based on the exponential appearance of the graph, some data transformations may be needed.

Section 3: Regression Description

First, we must take a look at the scatter plot of the response variable *price* against the designated predictor variable *carat*.

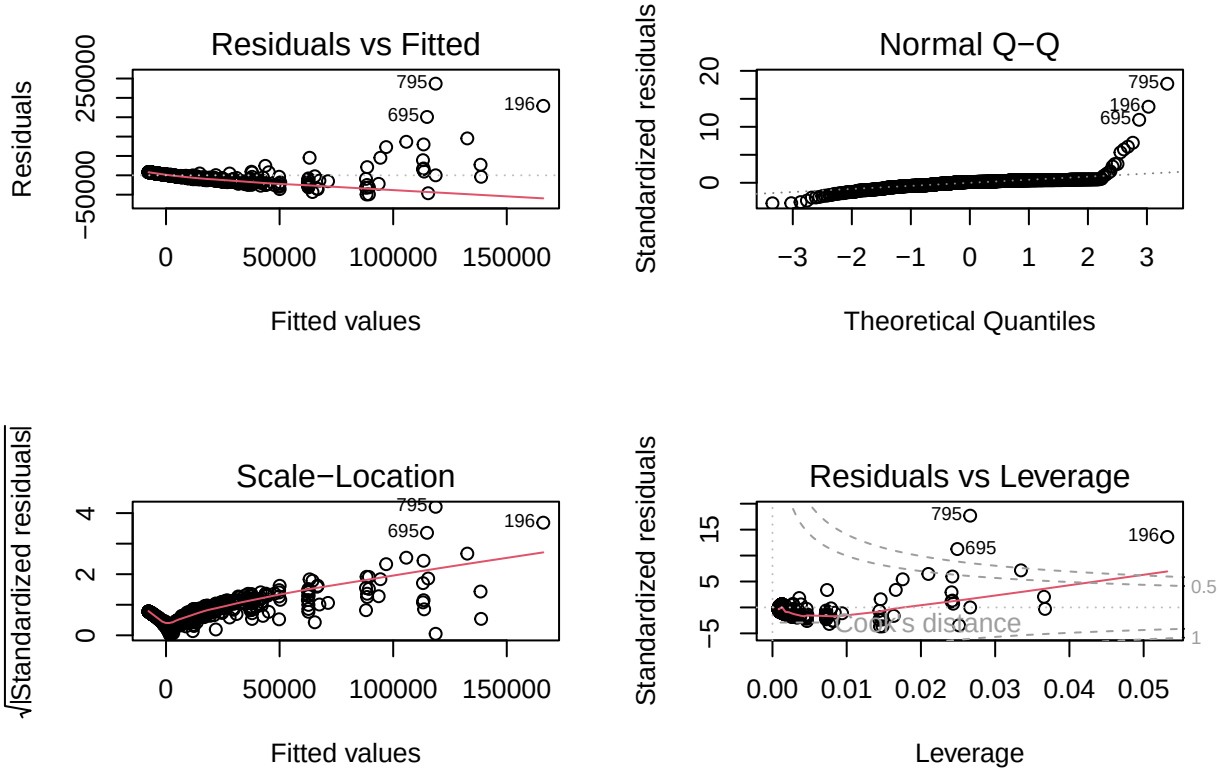


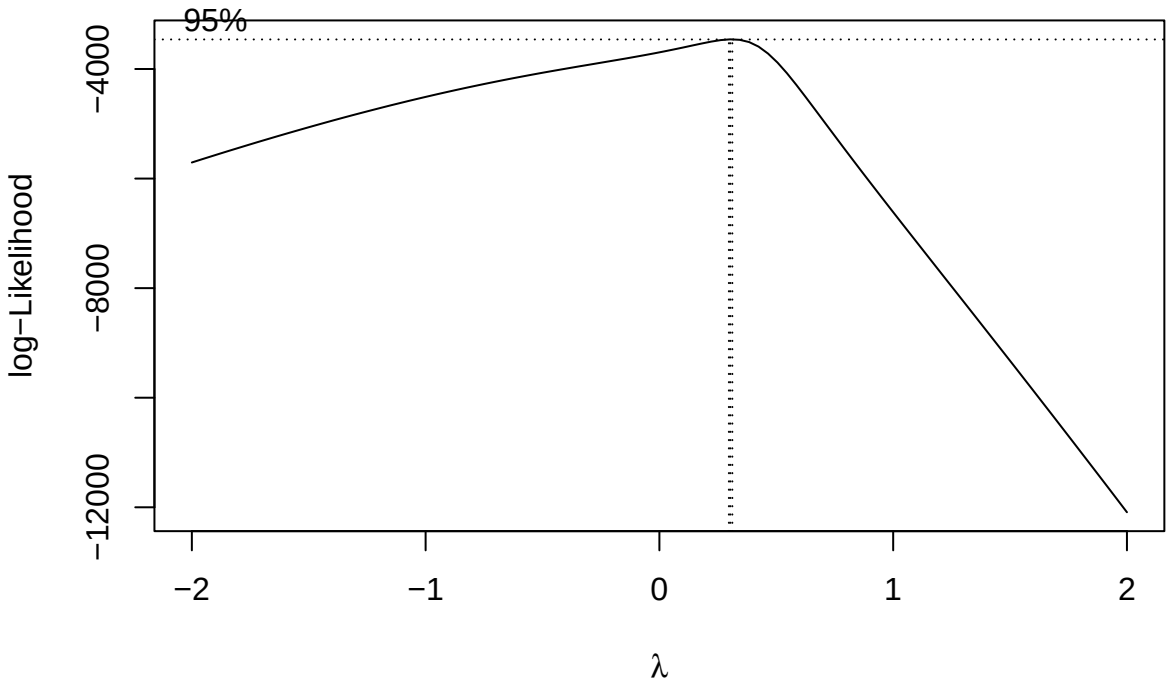
The scatter plot comparing carat and price looks exponential rather than linear. This would indicate that a transformation is needed, likely a log transformation.

The simple linear regression assumptions are as follows:

- The residuals have mean 0.
- The errors have constant variance.
- The errors are independent.
- The errors are normally distributed

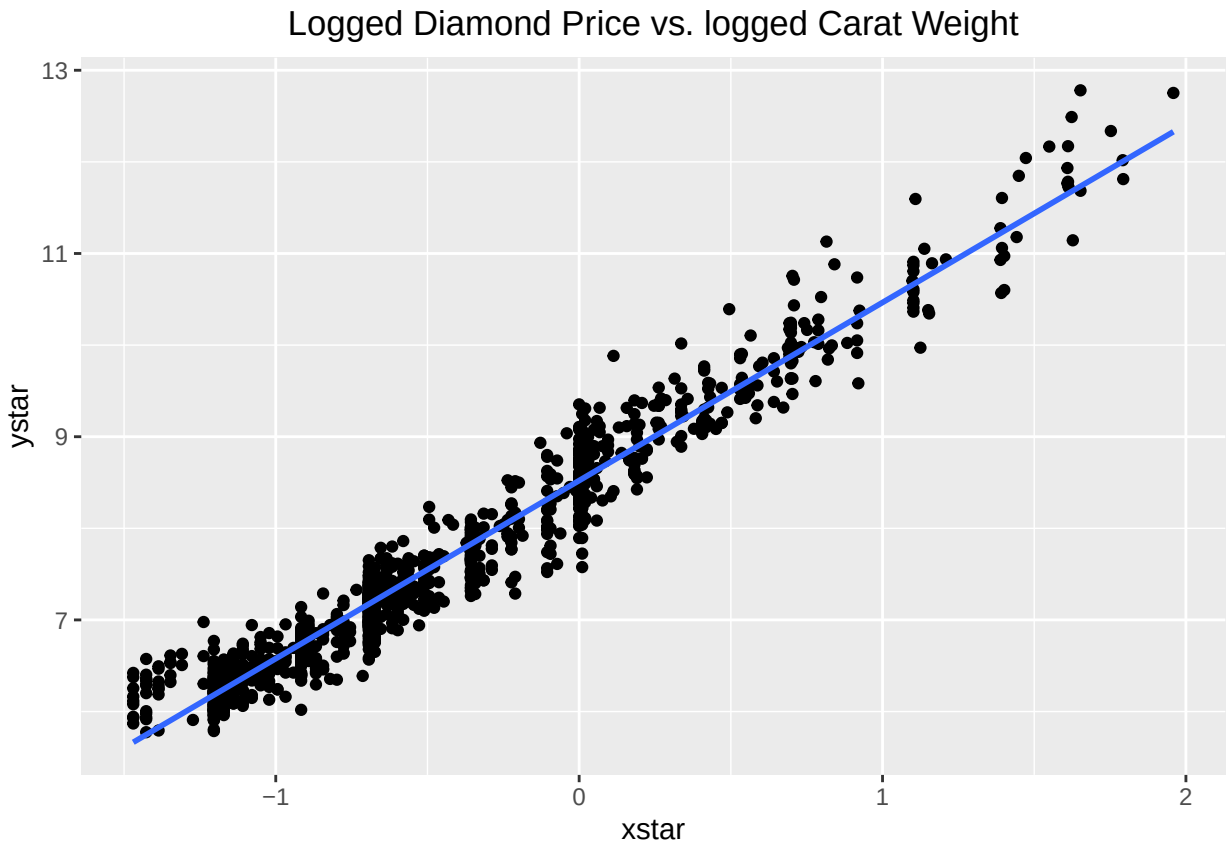
To confirm, the diagnostic plots are produced:


Residuals vs Fitted
Residuals
Fitted values
795
695
196
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
Leverage

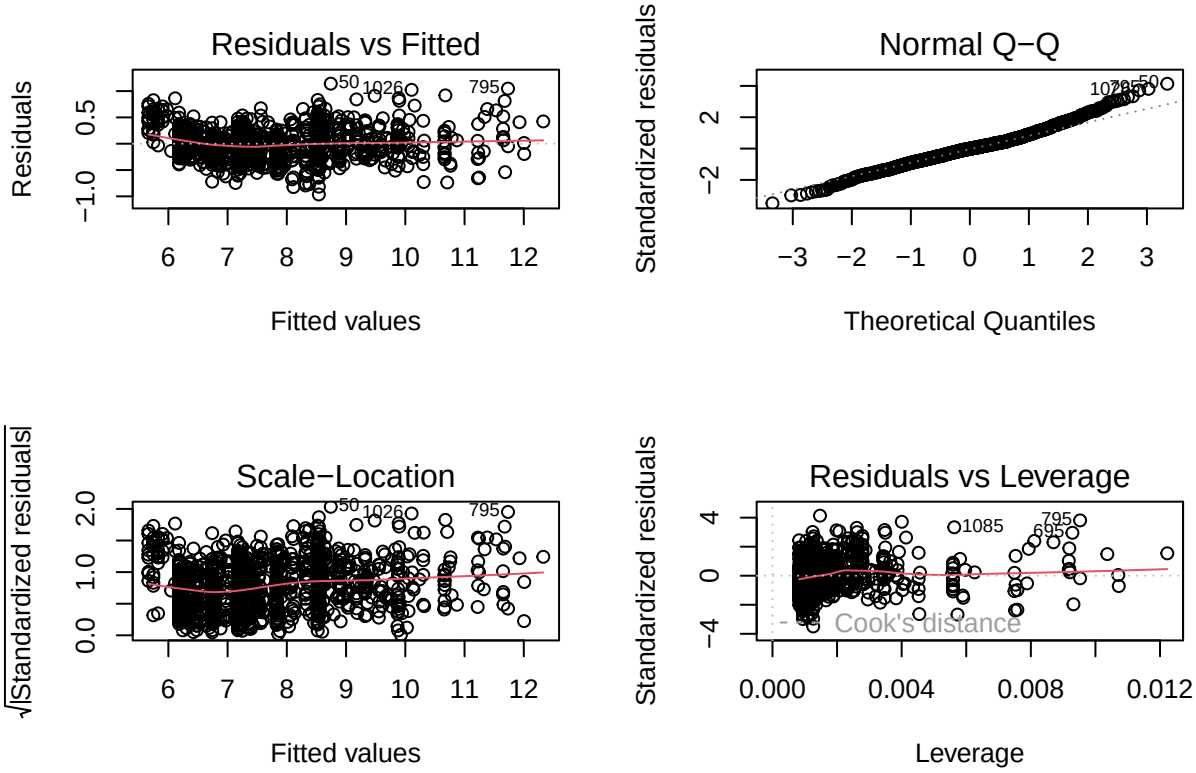


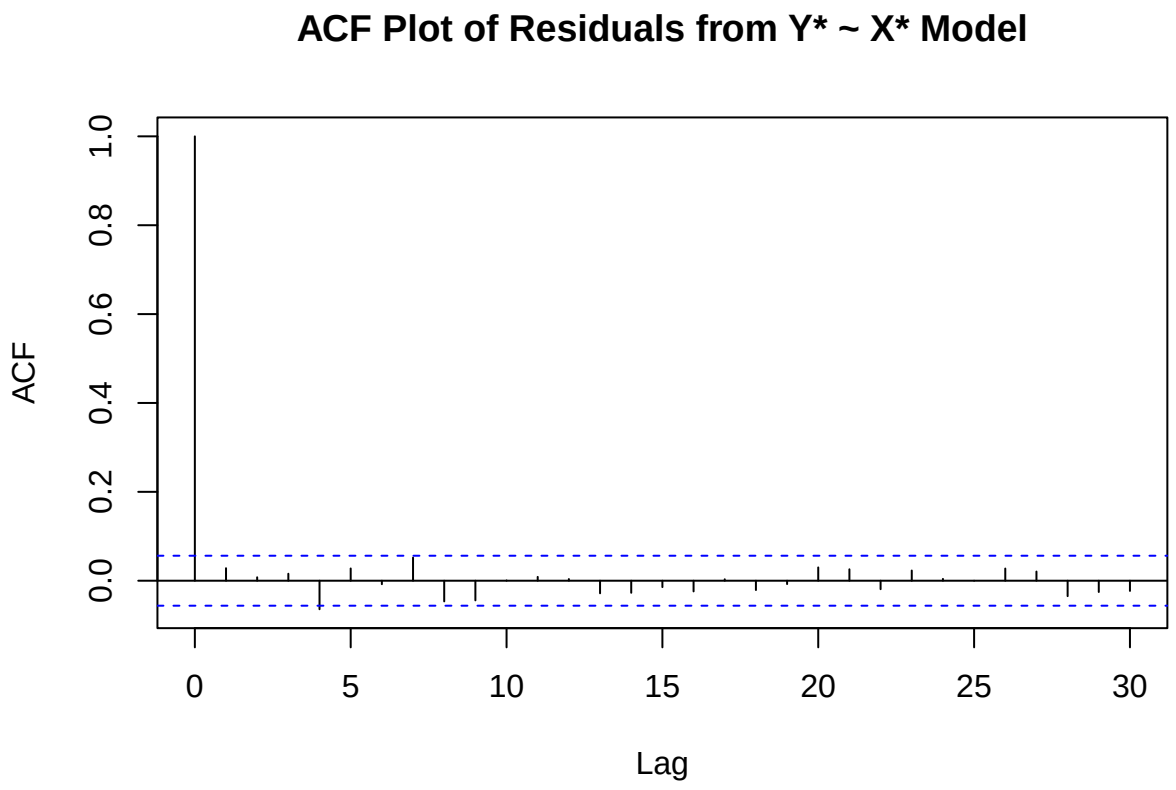
Based on the diagnostic plots, it seems that both regression assumptions 1 and 2 have been violated since the errors do not have mean 0 nor constant variance. The mean of the residuals appears to decrease, and the residuals appear become more variable. If both assumptions are violated, the best solution is to log transform both the x and y variables. The Box-Cox plot agrees with this since the confidence interval does not include 1, indicating the y variable must be transformed.

We will create the variable `xstar`, which is defined as $x^* = ln(carat)$ and the variable `ystar`, defined as $y^* = ln(price)$:



As we can see from the above scatter plot, the relationship appears much more linear than original analysis before. To be sure we have remediated the violations of the regression assumptions, we look at the diagnostic plots once more:


Residuals vs Fitted
Residuals
Fitted values
50
1026
795
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
1085
695



After log transforming both the `x` and `y` variables, the diagnostic plots also pass the assumption, with errors appearing to have both a mean of 0 and constant variance. Further, the Q-Q plot shows that the residuals have a fairly normal distribution. Based on the ACF plot, we can also see that the errors are independent. Thus, all four regression assumptions are met.

So, our hypotheses are:

$H_0 : \beta_1 = 0$

$H_0 : \beta_1 \neq 0$

When we run the regression, we see:

```
##
## Call:
## lm(formula = ystar ~ xstar, 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 ***
## xstar       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
```

Based on the results of the model with both the response and the predictor being log transformed, `xstar`, $x^* = ln(carat)$, is statistically significant at the 99% level with a p-value of 2.2×10^{-16}. This means that the log of carat is statistically significant in predicting the log of the price of the diamond.

The regression equation is $ln(price) = 8.521 + 1.944ln(carat)$. This means that for every 1% increase in the carat weight, the price of the diamond is expected to increase by around 1.944%.

The 95% confidence interval is:

```
##                2.5 %   97.5 %
## (Intercept) 8.502110 8.540306
## xstar       1.920152 1.967888
```

Which shows that we are 95% confident that for each 1% increase in the carat weight of a diamond, its price will increase between 1.92% and 1.97%.