Angela Boakye Danquah, David Chen, Evan Mitchell, Jason Zhang
STAT 6021
Dr. Jeffery Woo

# Predicting Diamond Prices

## Executive Summary

Our project focuses on using a simple linear regression model to identify the strength of the linear relationship between diamonds' carat weight and price. The data originates from bluenile.com, the Blue Nile Diamonds Jewelers organization's website. The first section of our report describes the dataset in detail to highlight the type of data available to us. Next, we listed the changes that we made to the dataset and subsequently generated visualizations to explore our data. These visualizations were used to address the Blue Nile's claims about the diamond features that have the greatest impact on the price; we found that clarity may be a more important factor than the Blue Nile asserts. Next, we transformed the predictor and response variables to use in the model. Two linear regression models were constructed: one predicts prices by colorless diamond carat weights, while the other predicts prices by near-colorless diamond carat weights.

Both models indicated that the price of diamonds tends to increase as the carat weights increase.

# Introduction

The purpose of this report is to examine the features of a diamond that contribute to its price using simple linear regression. This report uses data obtained from bluenile.com, the website for Blue Nile Diamond Jewelers. The featured dataset reports the carat weight, clarity, color, cut, and price for 1,214 diamonds. The variable “carat” is given as the diamond's total weight. “Clarity” reports whether a diamond is an “Included” diamond, a “Slightly Included” diamond, a “Very Slightly Included” diamond, a “Very Very Slightly Included” diamond, an “Internally Flawless” diamond, or a “Flawless” diamond. The variable “color” reports a diamond's color based on the Gemological Institute of America (GIA) color scale. “Cut” is reported as a categorical variable with each diamond being classified as “Good”, “Very Good” “Ideal”, or “Astor Ideal”. The response variable, price, is reported in USD. With this information, we aim to explore the relationship between price and the other listed features as well as the relationships between these features.

## Exploratory Data Analysis

Some minor changes were made to the dataset before analysis was carried out. First, we added a new column to the dataframe called "recolor." "recolor" is based on the "color" column: colors D, E, and F are identified as "colorless", and colors G, H, I, and J are identified as "near-colorless." We added another column to the dataframe called "regrade" which is based on the "clarity" column: clarities SI1 and SI2 are identified as "SI", clarities VS1 and VS2 are identified as "VS", clarities VVS1 and VVS2 are identified as "VVS", and clarities IF and FL were unchanged.

We first used a scatter plot (Figure 1) to show the individual data points of price ($) by carat count. Each point has a color corresponding with either the Colorless or Near-colorless categorization.

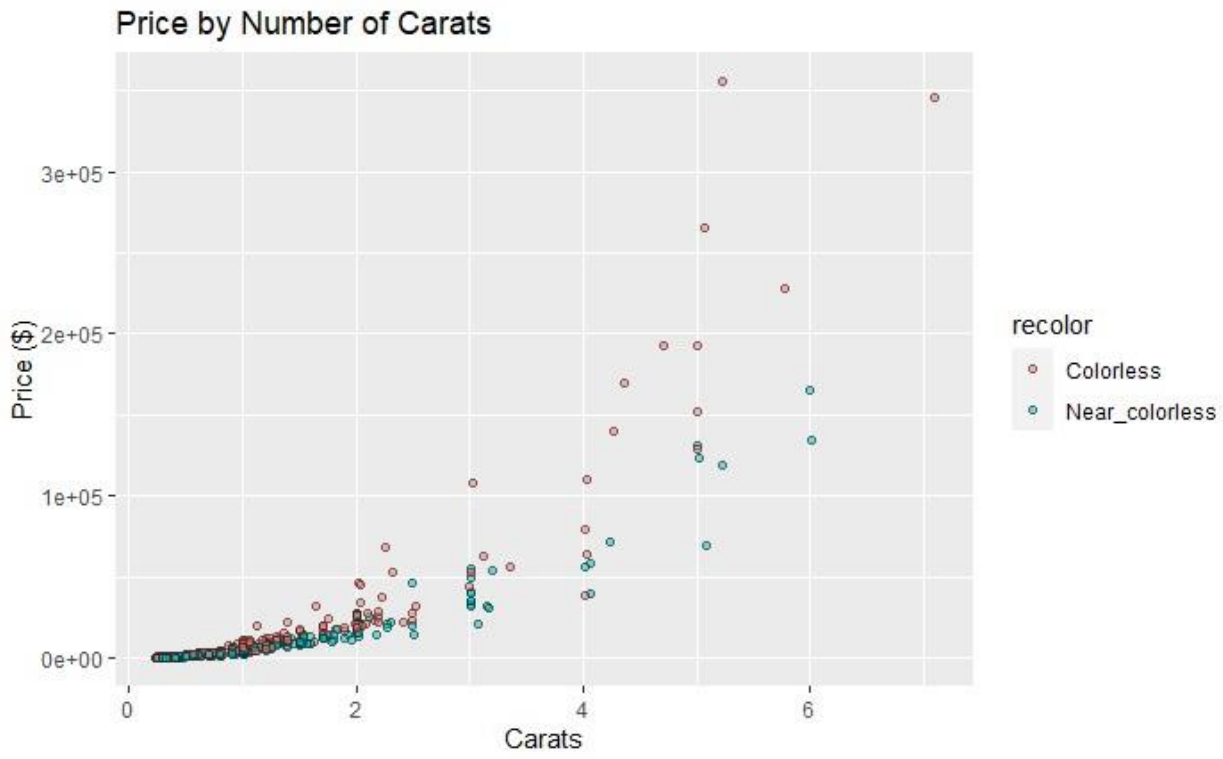


***Figure 1*** *Price by number of carats scatterplot, points colored by "recolor" factor*

The scatterplot shows that the price of a diamond tends to increase as its carat weight increases. We can also clearly see that the relationship between price and carats is not strictly linear.

Next, we used boxplots (Figure 2) to visualize the relationship between color and price.

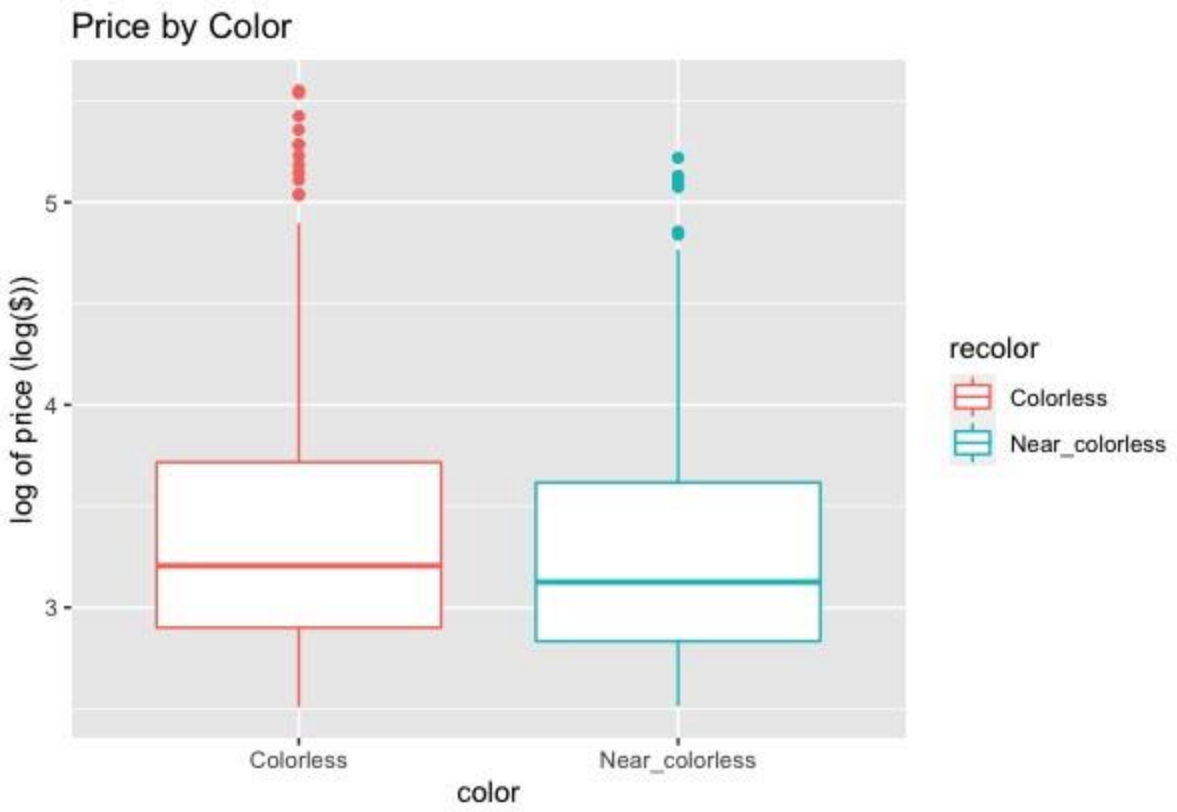


***Figure 2*** *Log (base 10) of price by color, side-by-side boxplots*

The boxplots above show that the median price of colorless diamonds is slightly higher than that of near-colorless diamonds. They also show that colorless diamonds take on the highest prices in the dataset overall.

The following set of boxplots (Figure 3) demonstrate the relationship between clarity and price.

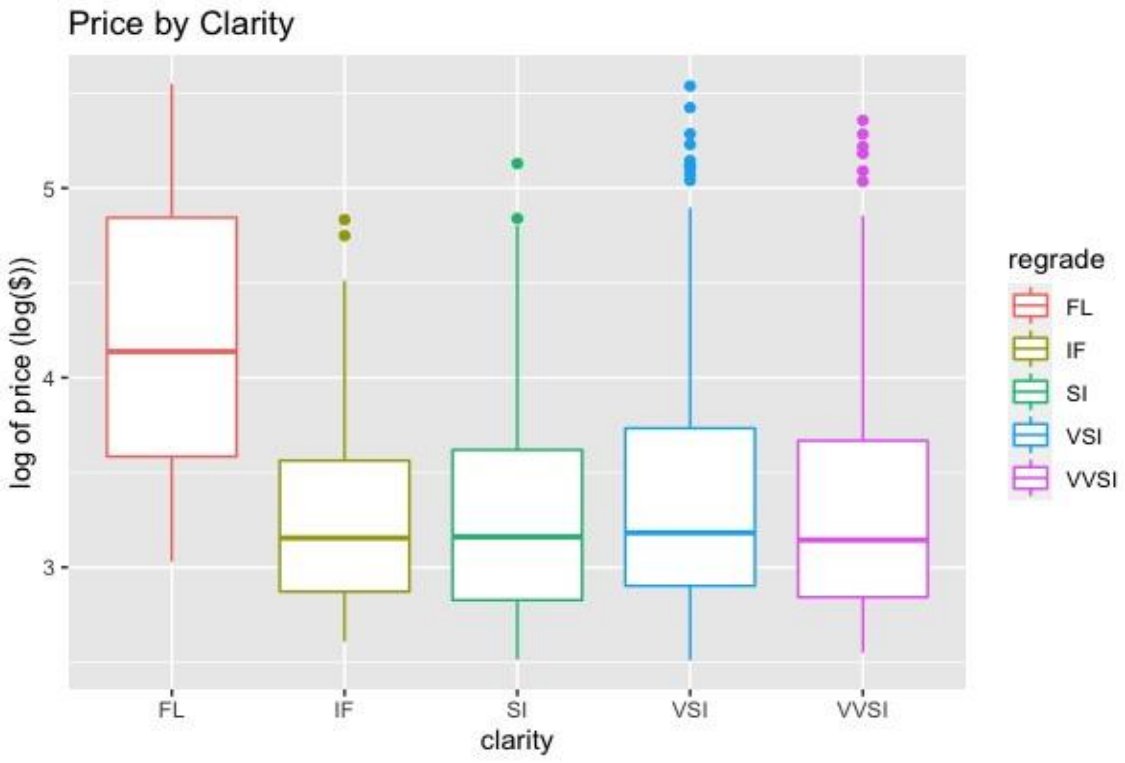


***Figure 3*** *Log (base 10) of price by clarity, side-by-side boxplots*

There are minimal differences in price between the various clarity categories besides diamonds designated as "flawless." These diamonds have a nearly 10 times higher median price than those of the other clarity categories.

Our final set of boxplots summarizes the relationship between cut and price.

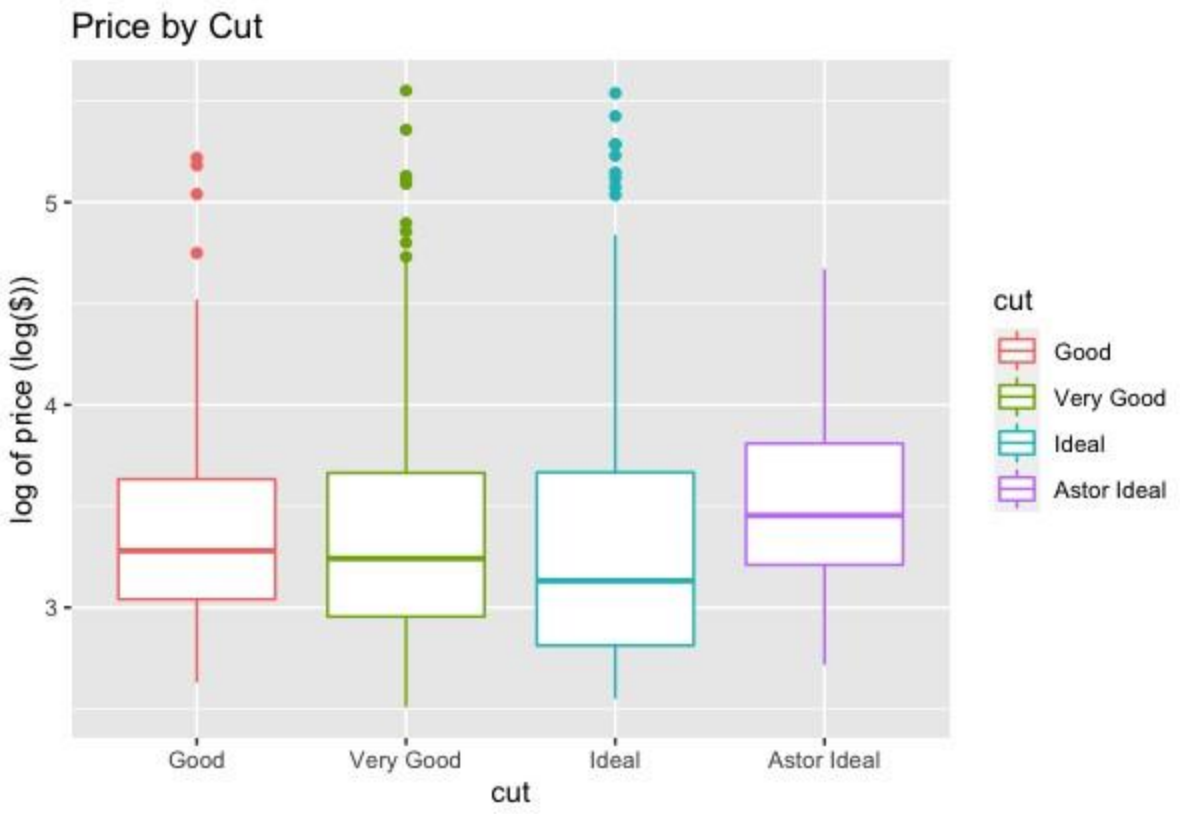


***Figure 4*** *Log (base 10) of price by diamond cut, side-by-side boxplots*

The median prices for diamonds classified as “Good” and “Very Good” are in about the same place; however, diamonds classified as “Very Good” take on higher prices between the two. The median value for diamonds classified as “Ideal” is the lowest overall, but some of these diamonds take on high values similar to those classed as “Very Good”. The median price of diamonds classified as “Astor Ideal” is the highest in the dataset; however, the range of prices for these diamonds is the smallest, most likely because these are the rarest diamonds.

The Blue Nile website claims that the feature of a diamond that has the greatest impact on its price is cut followed by color, carat, and finally clarity. Although our boxplots show that median prices do increase to a certain extent as the cut quality increases, the clearest difference in price

is summarized by the clarity boxplots. The price of diamonds in the 25th percentile of flawless diamonds is higher than or very close to the prices of diamonds in the 75th percentile of the remaining clarity categories. This indicates that clarity may be a more important factor than the Blue Nile website asserts.

## Data Modifications

Before any transformations, our initial scatter plot (Figure 1) revealed that the relationship between price and carats is not linear, resulting in a violation of our first assumption for linear regression. Our residual plot (Figure 5) also revealed that there was not constant variance in the residuals. The variance in the residuals was low at the beginning and grew larger as carats increased.

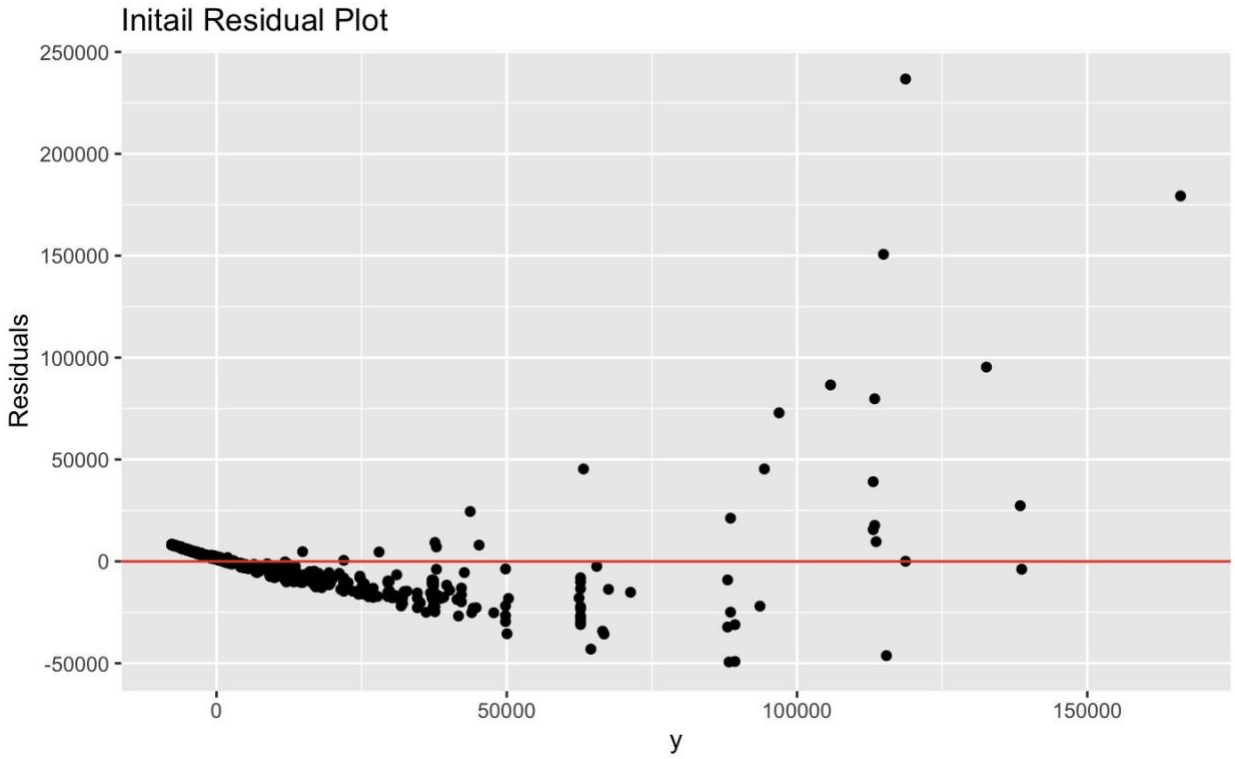


***Figure 5*** *Initial residual plot, no transformations applied*

Because of these assumption violations, we decided we needed to perform a transformation. After exploring different transformations, we decided that the best set of transformations was to transform both the response variable and the predictor using log transformations. Figure 6, the resulting scatter plot, shows a linear relationship between the two transformed variables

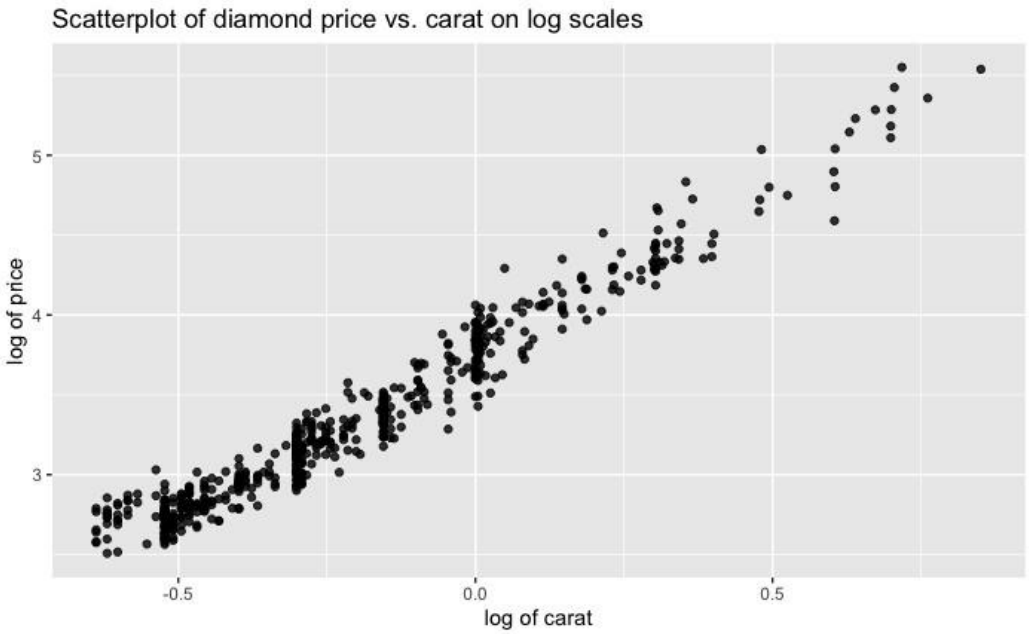


***Figure 6*** *Final scatter plot of log(price) vs log(carat)*

Furthermore, the resulting residual plot supports the assumption that residuals had a constant variance since there is no apparent fanning pattern in the residual plot (Figure 7).

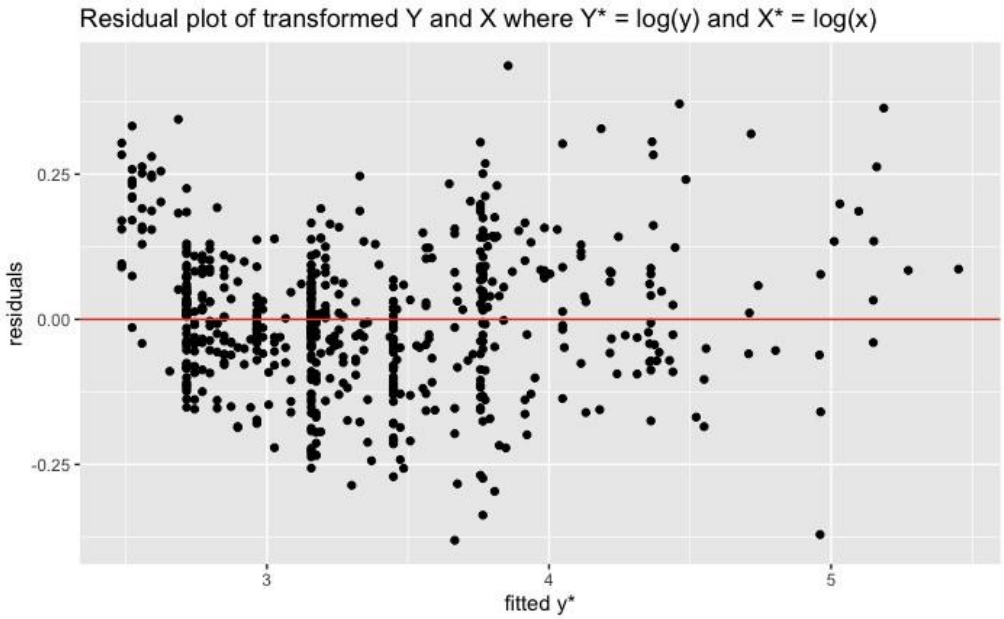


***Figure 7*** *Final residual after log transformations of price and carat*

While there is no way of knowing if the observations are independent, the resulting autocorrelation plot in Figure 8 shows that the residuals are uncorrelated, and the observations are not in any way ordered.

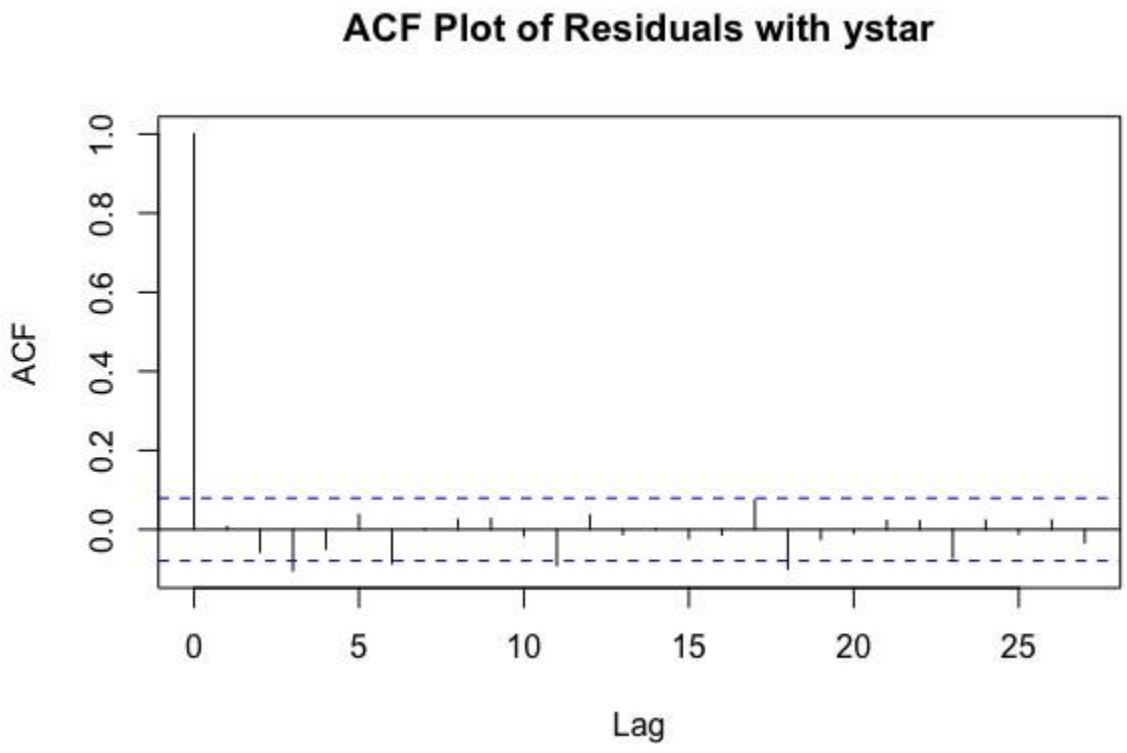


***Figure 8*** *Autocorrelation plot for residuals of transformed data*

Finally, the normal probability plot in Figure 9 reveals that the residuals follow a normal distribution, satisfying the normality assumption for linear regression.

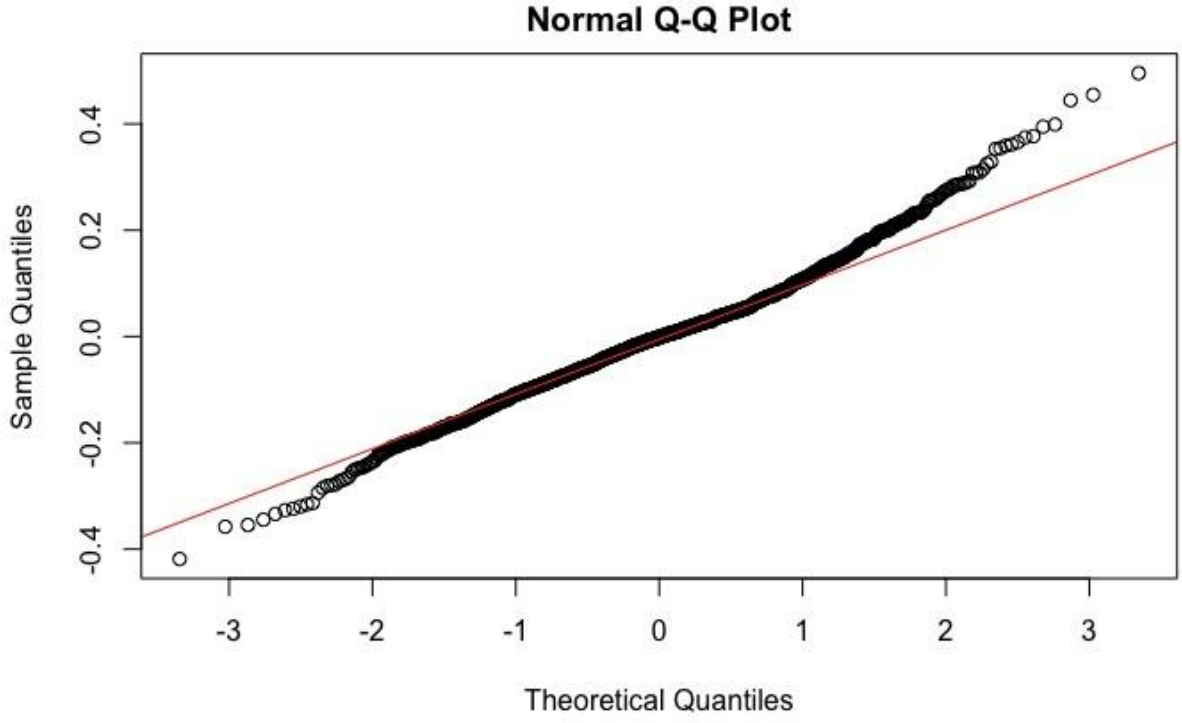


***Figure 9*** *Normal probability plot of residuals*

## Simple Linear Regression

After cleaning and transforming our data, we began by creating a simple linear regression model on the full dataset, using the number of carats to predict the prices of the diamonds. And the linear model for $\log_{10}$(prices) against $\log_{10}$(carats) yields $ß_0$ =3.700 and $ß_1$= 1.944 with $R^2$ = 0.9547. We then performed a hypothesis on the slope to determine whether it is statistically significant. The null hypothesis of our test was $ß_1$=0 and the alternative was $ß_1 \neq 0$. The p-value for slope is 2.2e-16, much smaller than p* of 0.05. Thus, we can reject the null hypothesis at significance level $\alpha$=.05, and conclude that there is a linear relationship between the log of price and the log of carat weight for all diamonds and that the value of our slope is statistically significant.

Regression line for the entire set of diamonds:

$$\log_{10}(y) = ß_0 + ß_1 \log_{10}(x) \log_{10}(y)$$

$$= 3.700 + 1.944 * \log_{10}(x)$$

While examining the visualization, we could see that, from figure 1, different diamond color categories seem to have different slopes as most colorless diamonds are more expensive than those near-colorless ones at the same carat weight.

Thus we decided to perform linear regression independently for the two color-categories of diamonds: colorless group and near-colorless group. Similar to the entire dataset, we transformed prices and carats by taking logarithms of base 10. To further visually confirm our assumptions that these two diamond groups by color have linear relationships after log transformations and their linear regression models have different slopes, we plotted the log-transformed prices and carats by group, see figure below, and fitted a straight line for both groups.

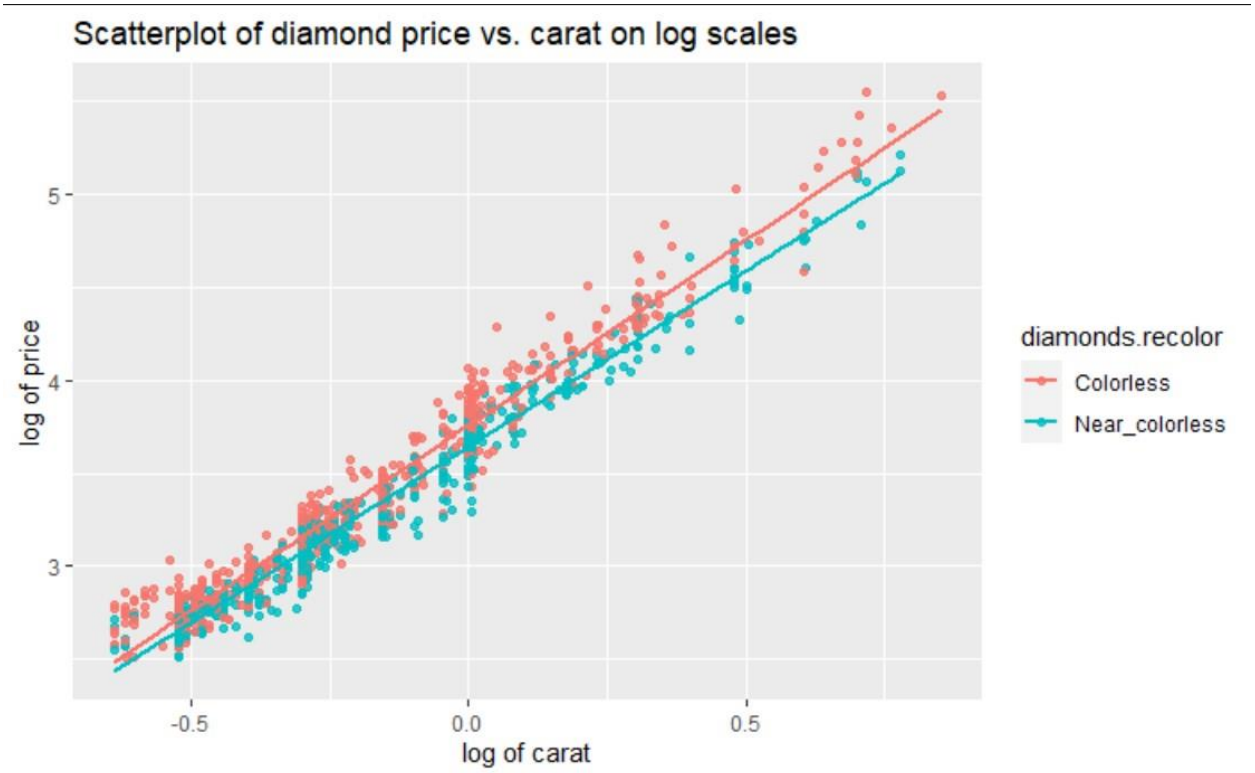


***Figure 10*** *Color-coded scatter plot of log(price) vs log(carat)*

First, we generated the residual plot for the near-colorless group of diamonds to confirm once again our assumptions, see figure below.

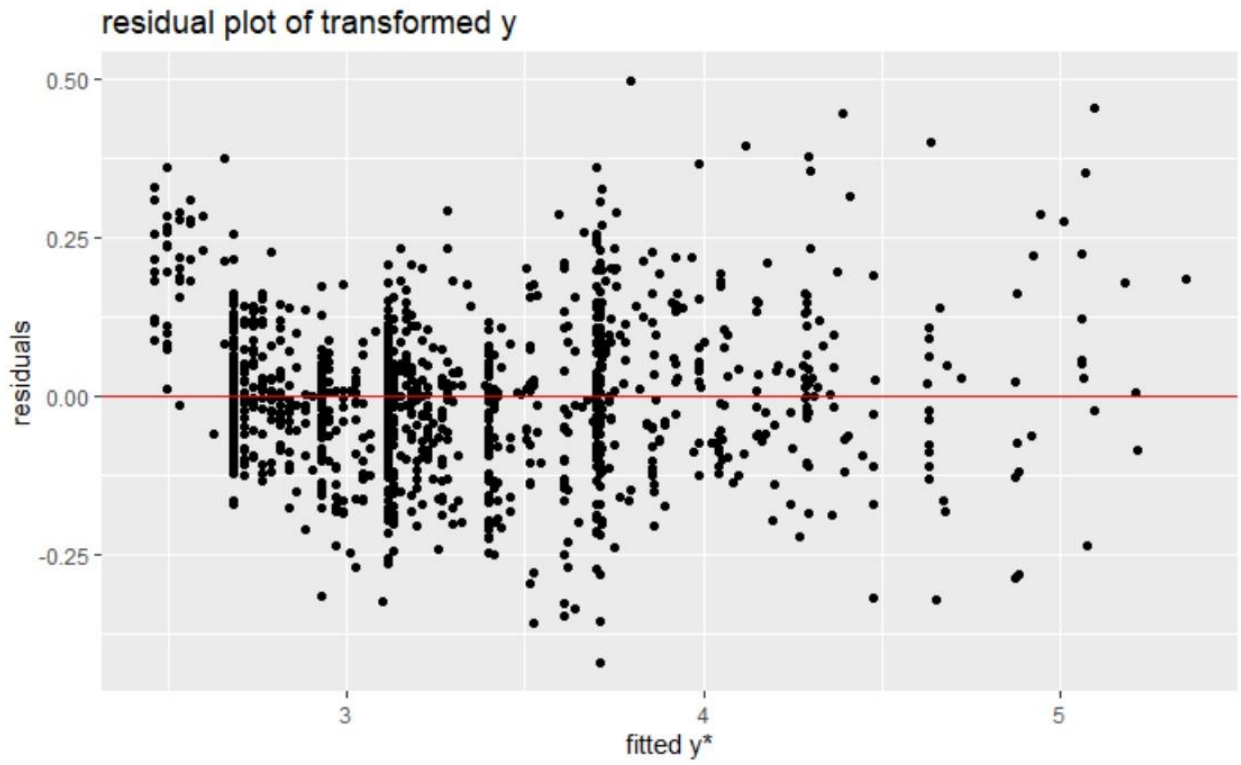


***Figure 11*** *Residuals for near colorless diamonds*

Proceeding to summarize the linear regression model, we obtain the following statistics for the near-colorless group of diamonds: $R^2$ =0.9694

This $R^2$ is greater than the $R^2$ for the overall set, indicating a better fit when we divide the diamonds by color groups.

Regression line for near-colorless group:

$$\log_{10}(y) = ß_0 + ß_1 \log_{10}(x) \log_{10}(y) =$$

$$3.6433 + 1.8925 * \log_{10}(x)$$

Second, we generated the residual plot for the colorless group of diamonds to confirm once again our assumptions, see figure below.

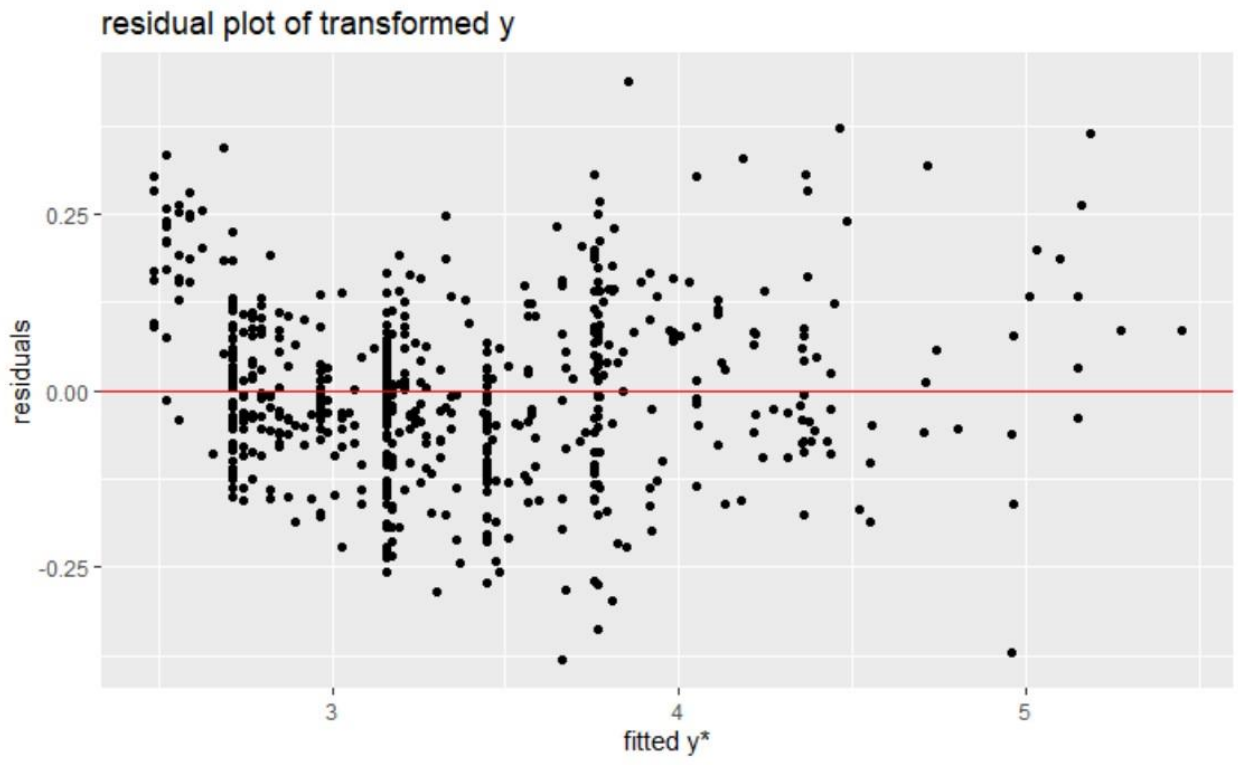


***Figure 12*** *Residuals for colorless diamonds*

Proceeding to summarize the linear regression model, we obtain the following statistics for the colorless group of diamonds: $R^2$ =0.9558

Again, similar to the $R^2$ for the near-colorless group, the $R^2$ for the colorless group is greater than the $R^2$ for the overall set, indicating a better fit when we divide the diamonds by color groups.

Regression line for colorless group:

$$\log_{10}(y) = ß_0 + ß_1 \log_{10}(x) \log_{10}(y)$$

$$= 3.757 + 1.992 * \log_{10}(x)$$

Similar to the p-value and the hypothesis test for the linear regression for the entire diamond dataset, the p-values for both the near-colorless and the colorless groups are in the $10^{-16}$ scale, indicating statistically significant linear relationships at the α=.05 significance level between $\log_{10}$ of prices and $\log_{10}$ of carats.

# Conclusion

Without considering additional categorical predictor variables such as clarity, color, and cut, we could find a linear fit between the log of prices of a diamond and the log of its carat weight at a high degree of confidence.

By simply separating the diamonds into two color categories, near-colorless and colorless, we obtained better $R^2$ values for the two subgroups than the $R^2$ of the entire dataset. This necessitates fitting linear regressions by color group.

By comparing the slope of colorless diamonds ($ß_1$ of 1.992) and near-colorless diamonds ($ß_1$ of 1.8925), we can tell that the group of colorless diamonds has a larger slope than that of the near-colorless group, indicating that, from incremental carat increase, incremental price increase for colorless is larger than that of near-colorless. This observation has practical implications for diamond shoppers. When one buys diamonds, one will need to pay more for a colorless diamond than for a near-colorless diamond at the same incremental gain in carat weight. This can be significant because our linear regression slopes between log of price and log of carat translate to exponential coefficients between actual price and carat. However, this may not be significant to someone making a decision on a matter that's not quantifiable such as love.