Arti Patel

Elizabeth Driskill

Mariah Hurt

Sudeepti Surapaneni

**Project 1 Report: STAT 6021**

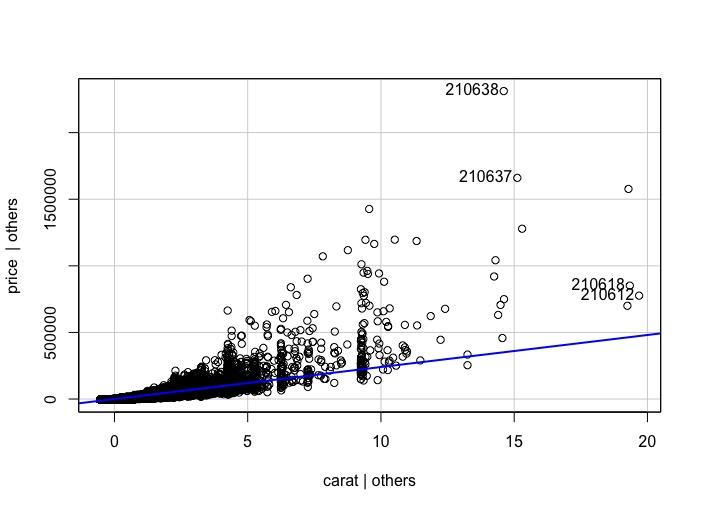
**Introduction:**

The dataset comes from the diamonds available for sale on BlueNile and consists of price, color, cut, clarity, and carat. The four c’s are typically used to assess diamond quality and therefore price. This assessment could be useful to someone shopping for a diamond within a particular budget or someone who has selected a diamond and wants to know if they are paying a fair price. This information could also be useful to someone trying to sell a diamond to see what range they should set their price based on the four c’s. Since we are using cut, color, clarity and carat as determining factors for the quality, we will use these as the independent variables and the price will be the dependent variable. Our high-level goals and objectives for this project included prediction and forecasting of the response variable, the price of the diamond.

**Initial Variable Exploration:**

To start assessing the data we researched each of the four c’s to find out what they mean in context and used the unique() function in R to view a list of unique values for each variable. Three of the four independent variables are categorical variables, meaning they are indicated by characteristics and are not quantitative. According to the Blue Nile website, the color scale ranges from D for the colorless diamonds to J for diamonds with a color. Although some colored diamonds are rare and sought after, colorless D is generally considered the best. The cut scale is similar in that it is categorical, but the categories are ranked from best to poor/fair, ranging from “Astor Ideal” to “Good” in our dataset. Clarity was another categorical variable within our dataset. Again, we visited the Blue Nile website to assess the categories for clarity. Clarity ranges from FL, indicating a flawless diamond, to SI2, indicating a slightly included diamond. Evaluating the structure of the data in R, we saw that each of the categorical variables were already entered in as factors. Since these are categorical, we created dummy variables to represent indicators for each unique value of cut, color, and clarity. The price category simply lists the price of the diamond and the carat category lists the weight of the diamond measured in the unit of carats. Once we completed our initial research on the data provided, we proceeded to developing a model that would satisfy normality assumptions.

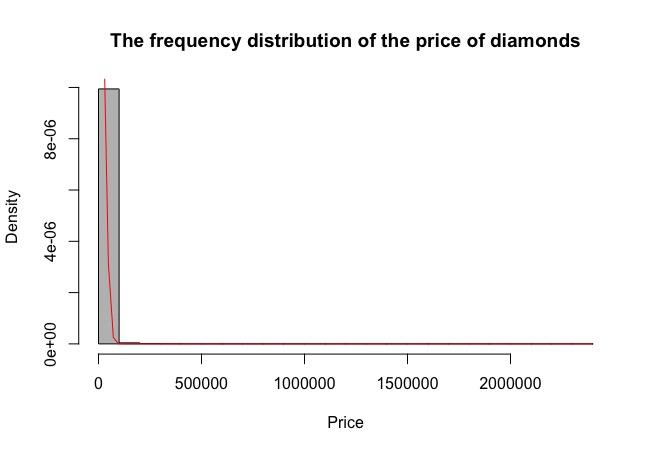
To start examining our quantitative variables, we made an AV Plot of carat to assess the importance of carat as a predictor. The non-zero slope of the AV indicated that carat is an important regressor in the model. We also examined the summary statistics for both price and carat.

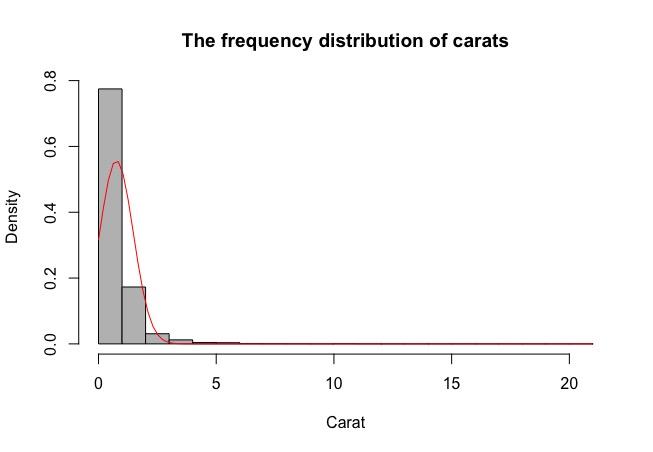


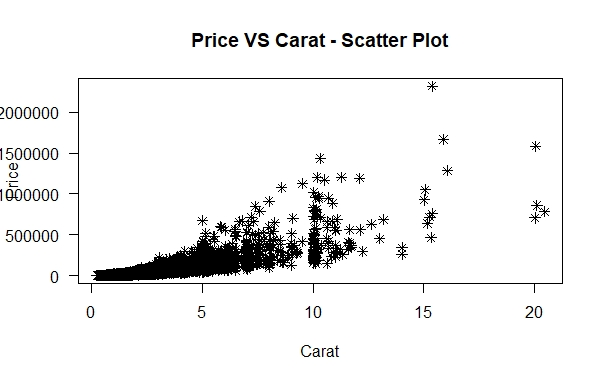
|  |  |  |  |
| --- | --- | --- | --- |
| **Carat** | | **Price** | |
| Min. | 0.2300 | Min. | 229 |
| 1st Qu. | 0.3700 | 1st Qu. | 698 |
| Median | 0.5100 | Median | 1432 |
| Mean | 0.7621 | Mean | 5540 |
| 3rd Qu. | 1.0000 | 3rd Qu. | 4235 |
| Max. | 20.4500 | Max. | 2317596 |

**Model Fitting:**

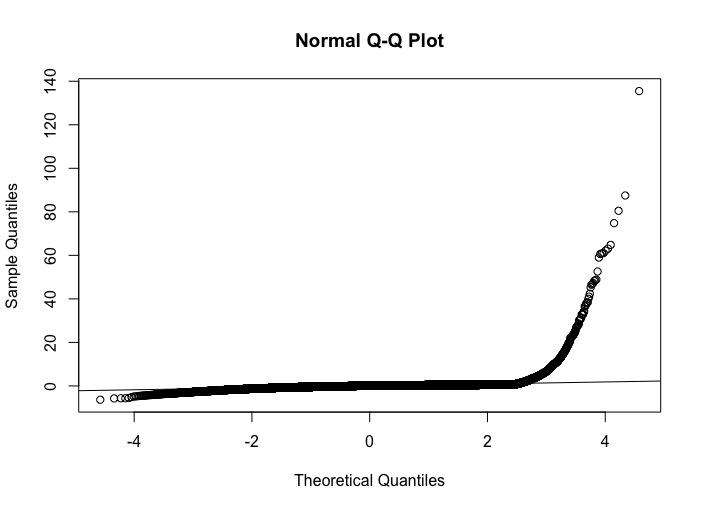
Our initial regression model was : price ~ carat + cut + color + clarity. We also wanted to check if carat interacts with the other categorical variables that would result in multiplied effects in the price, since most buyers who are in the market for a heavy carat may also be looking for the most ideal cut, color and clarity. We decided to use every variable in the data set from the start and then later finalize the model accordingly after data transformations, hypothesis tests to test for significance, and multicollinearity analysis. To determine whether or not we needed to transform the data, we further examined the two quantitative variables, price and carat. Looking at the following scatterplot and histograms of the price and carat, we observed that the data did not follow a linear pattern or a normal distribution.

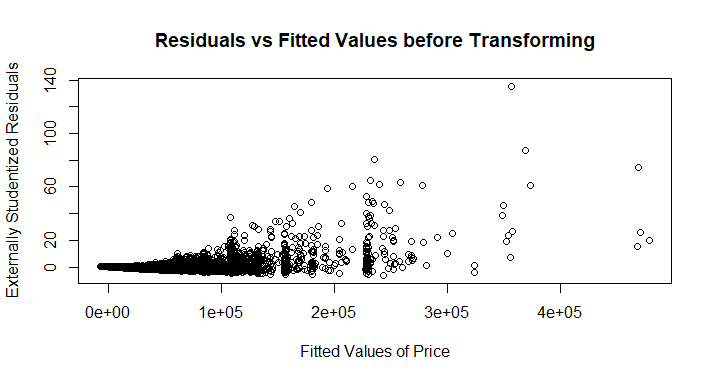






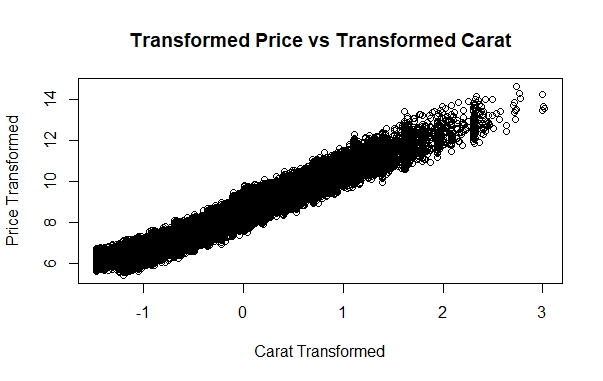
The histograms for price and carat have similar skewed distributions, which made us hypothesize that similar transformations would be needed for these variables. A QQ plot of the single linear regression model with price and carat as well as a plot of the externally studentized residuals vs. the fitted values can be seen below.





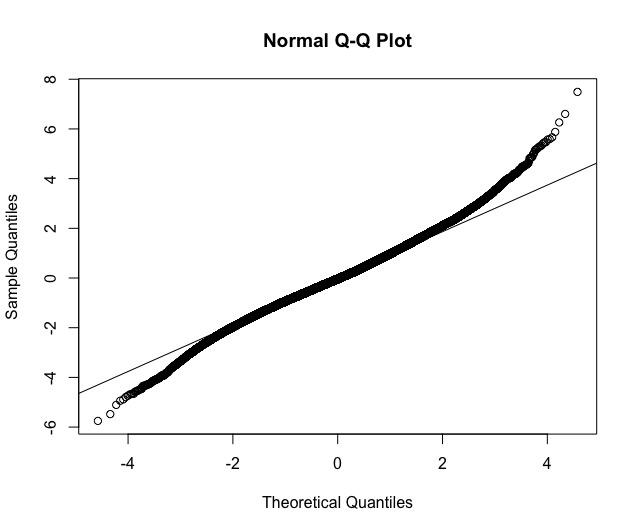
**Data Transformation:**

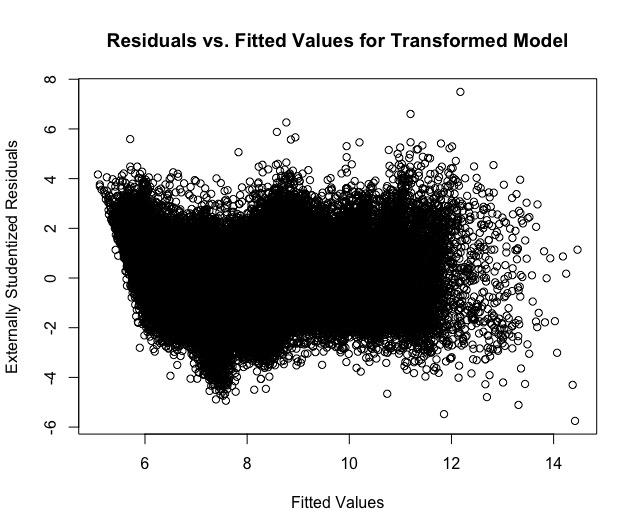
There were significant deviations from the straight line in the QQ plot, and the residual plot did not exhibit constant variance. Because we could not assume normality or constant variance, we could not use our initial regression model and decided to transform the data in order to produce a model that satisfied the normality assumptions. Since the residuals versus fitted values plot showed a pattern proportionate to the square of the expectation of price, we decided to try a natural log transformation of price. In addition, the Blue Nile website also says that the price of the diamond increases exponentially as the carat weight goes up, which supports our theory for transformation. Due to the similarity in distributions of price and carat, we applied the same transformation to carat. The scatter plot of transformed diamonds vs transformed carat now looks linear:



In order to verify our model selection, we applied a box cox procedure to determine a lambda which indicates a transformation parameter. Our “best lambda” was -0.02. Applying (y^lambda - 1)/lambda gives a similar transformation to the natural log. Given this, we chose to keep our natural log model since it seemed to be the most reasonable for our data and seemed to align with the results of our Box-Cox transformation.

The QQ plot as well as a plot of the externally studentized residuals vs. the fitted values can be seen below. Since there were no obvious deviations from the straight line of the QQ plot, we can assume a normal distribution. The plot of residuals versus fitted values looked randomly scattered, so this exhibited nearly constant variance.





After transforming both the carat and price data, we arrived at the model: ln(price) ~ ln(carat) + cut + color + clarity.

**Checking for Model Adequacy:**

To evaluate our data for signs of multicollinearity, we solved for the variance inflation factors in R and received the following output:

carat\_t cut color clarity

1.059753 1.057626 1.037056 1.041533

Because all of these values were small (< 5), we concluded that the data did not exhibit significant multicollinearity. We also produced the correlation matrix in R, and this further confirmed our conclusion that the data did not show multicollinearity because none of the correlation coefficients between variables were significant. The largest absolute value of the correlation coefficients was 0.3398, but this value was still not significant enough to assume that the two variables are linearly dependent. Given all of this information about the relationships between the variables, we decided to include every regressor in our linear model.

We decided to assess if there were any interaction terms that could contribute to the model. The interactions we looked at were carat with color, carat with clarity, and carat with cut. We assessed interaction terms in relation to carat because a very high clarity or desirable color in combination with a large carat might have a heightened effect on price, since it is likely that someone shopping for a heavy carat diamond would also be in the market for an ideal cut, a clear color, and flawless clarity. In order to assess the contribution of each regressor as well as each interaction term to the model, we ran a series of partial F-tests. For our big model, we included all regressors and interaction terms, and for each little model, we removed one term at a time. All of the partial F-tests yielded significant p-values. This indicates that we can reject the null, and there is only a very small chance that the null hypothesis is true. For this reason, we decided to keep all four regressors and all three interaction terms. We decided to use the bigger model as opposed to the smaller model because we are trying to predict values for y (price). If we were trying to estimate parameters for the slope, we would have considered using a smaller model instead. The reason that we did not investigate more interaction terms is because in practice as a buyer, carat is the most dominant predictor of price. Our final multiple linear regression model including interaction terms is:

ln(price) ~ ln(carat) + cut + color + clarity + ln(carat):cut + ln(carat):color + ln(carat):clarity.

**Conclusion:**

Using our final model, we created prediction intervals for price based on sample inputs. Our sample inputs were values from carat taken from each quartile (first quartile, median, third quartile, and maximum). For the categorical variables, the sample inputs were paired with the carat quartiles based on the given spectrums from the Blue Nile website. For example, color “D” is the most preferred, along with cut “Astor Ideal,” and clarity “FL”, so these values were used in the prediction interval for the maximum carat value. Of note, because the data was transformed logarithmically, the R outputs for the prediction intervals had to be exponentiated in order to produce an understandable prediction for price. Our four prediction intervals can be seen below (alpha = 0.05):

Carat = 0.3700, Carat\_t = -0.9943, Cut = "Good", Color = "J", Clarity = "SI2"

fit: 399.6743

lower: 291.3878

upper: 548.2026

Carat = 0.5100, Carat\_t = -0.6733, Cut = "Very Good", Color = "H", Clarity = "VS2"

fit: 1145.11

lower: 834.8892

upper: 1570.602

Carat = 1.0000, Carat\_t = 0.0000, Cut = "Ideal", Color = "F", Clarity = "VVS1"

fit: 6949.417

lower: 5066.73

upper: 9531.677

Carat = 20.4500, Carat\_t = 3.0180, Cut = "Astor Ideal", Color = "D", Clarity = "FL"

fit: 3,973,071

lower: 2,882,965

upper: 5,475,368

This demonstrates how our model can be applied to predict price ranges of diamonds, given a set of other preferred qualities. We only computed the prediction intervals for four sets, but if another user or company wanted to specify other combinations of the variables, the model could still be applied. If you are a buyer with your eye on a particular diamond, you could use this model to determine if the diamond is appropriately priced. Additionally, if a user like Blue Nile wanted to develop an application that accepts user input and predicts price, our model could be used.