UVA Stat 6021 Project 1, April 5, 2021

Thomas McIntyre (gem5cm), Tulsi Ratnam (tr9sq), Manpreet Dhindsa (mkd8bb), and Jonathan Shakes (hqe7rd)

Executive Summary

Diamond pricing can be complex, taking into consideration many different factors including the "4 C's": carat weight, cut, color and clarity. This project uses 1,212 observations of diamonds from the online diamond retailer, Blue Nile, to build a statistically based model that, when provided with the 4C's, can estimate diamond prices. The project aims to lay the groundwork for a tool that would help consumers be better informed diamond shoppers.

Though a simpler model using fewer inputs would have been sufficient, we want to create a model where customers can toggle the predictive variables to reach a desired price point. The below equation shows that when there is an increase of x% in the carat weight, the predicted price is increased by a factor of $(1 + x/100)^{2.232326}$while the categorical variables are held constant. Our model also shows promise for helping consumers who purchase from other diamond retailers.

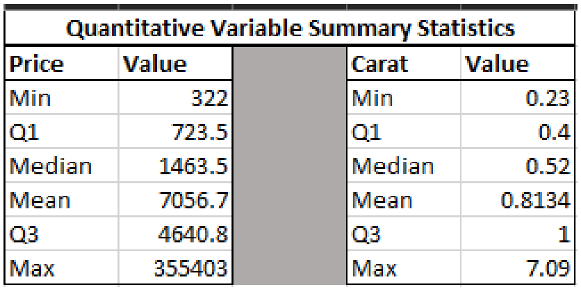
$$log(Price) = log(Carat + 0.1) + Clarity + Color + Cut.$$

The following paper will first assess relationships between the 4C predictor variables and whether they affect the response variable, price. Next, we outline the steps taken to transform the predictors and build a multiple linear regression model. Finally, we will discuss whether the model is useful enough to estimate diamond pricing from other retailers such as Costco and Daniel William.

Exploratory Data Analysis

Our goal is to choose a regression model that predicts diamond prices using the “4 C’s.” Carat is the only quantitative variable of the four, while Color, Clarity and Cut are all categorical variables.

We investigate the counts of unique levels of each categorical variable and see that two of the levels have relatively few observations, which may impact our model. Specifically, Clarity type “FL” has 3 observations, less than 1/20th the number of the next smallest Clarity type. The Cut type called “Astor Ideal” has 20 observations, which should be large enough to make statistical observations but is still small compared to the next-smallest Cut level, “Good,” with its 73 observations. External background information[1] clarifies that the word “Astor” in “Astor Ideal” is a brand label, not an independently verifiable distinction in Cut, which may suggest the potential to combine “Astor Ideal” and “Ideal” into a single category. Below is the summary statistics table of the two quantitative variables.

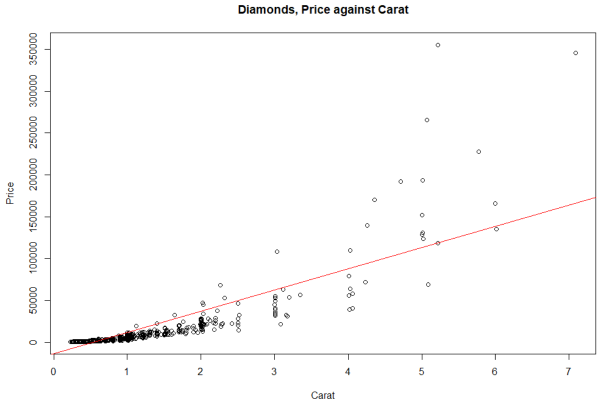


Quantitative Variable Summary Statistics				
Price	Value		Carat	Value
Min	322		Min	0.23
Q1	723.5		Q1	0.4
Median	1463.5		Median	0.52
Mean	7056.7		Mean	0.8134
Q3	4640.8		Q3	1
Max	355403		Max	7.09

The data’s Price variable displays a positive skew, with the mean Price ($7056) greatly exceeding the median Price ($1463). The Carat variable also displays a positive skew.
The Q3 Price value is 6.4 times larger than the Q1 Price value, whereas the Q3 Carat value is 2.5 times larger than the Q1 Carat value. If there were a pure linear model between the two, then such a large discrepancy would be unlikely.

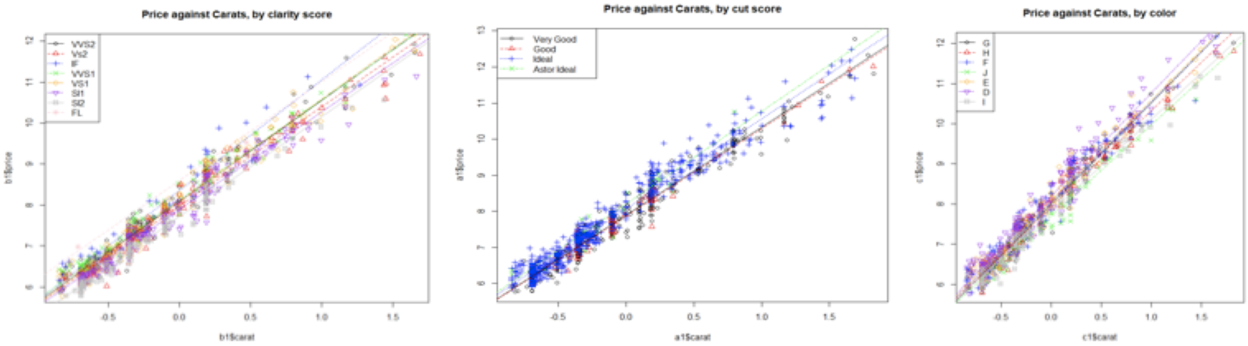
[1] https://www.reddit.com/r/Diamond/comments/7ugtpk/blue_nile_ideal_vs_astor_ideal/

To better understand the relationship between the only two quantitative variables, we can produce a scatterplot showing Price against Carat.



The scatter plot confirms our suspicion, from the summary stats, that Price does not increase linearly with Carat.

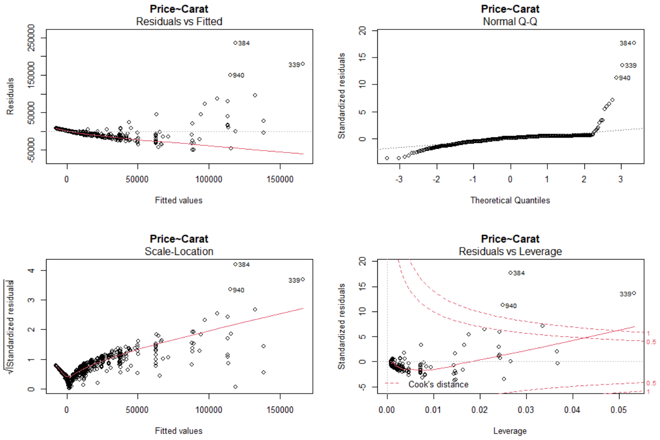
To see if the Price and Carat data are linear, differentiated mainly based on categorical variables, we can produce scatter plots that separate levels of different categorical variables. These plots can be seen below in the order of Clarity, Cut and Color. Our most significant observation from the plots above, prior to model selection, is the lack of a clear linear relationship in the data.



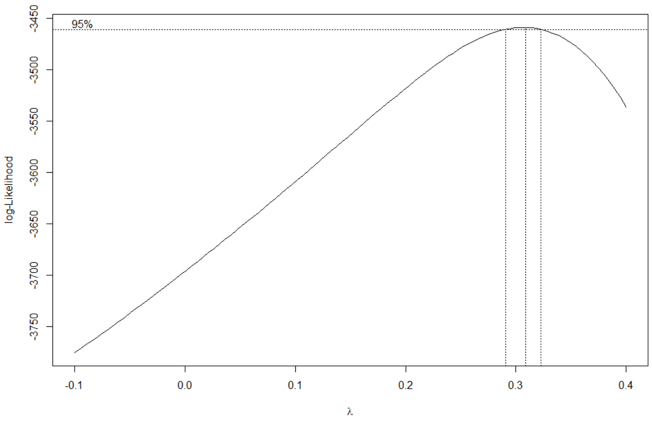
Breaking out the data by Clarity does not eliminate the need to transform the quantitative variables. The black circles representing “VVS2” Clarity, for example, still display a non-linear, exponentially increasing slope that starts below the black-solid VVS2 line and then curves above it at high Carat values. However, it does show a variety of different slopes for each Clarity score, suggesting that we cannot easily eliminate Clarity as a predictor of variation in Price. Breaking the levels of Cut into separate regression lines of Price by Carat does not solve the distribution issues of the data. The blue crosses of the "Ideal" Cut, for example, still suggest an exponential upward curve. With the original data, the slopes for the different levels of cut do not appear to be all equal. However, the need for transformation is still apparent from this plot. Each level of Color has a different slope, but the linear models do not match the data observations. We can also see that the values of Color that are commonly considered most valuable (D and E Colors) have the steepest slopes, and the least valuable J and I Color slopes have flatter, but still positive, slopes.

Model Building Process

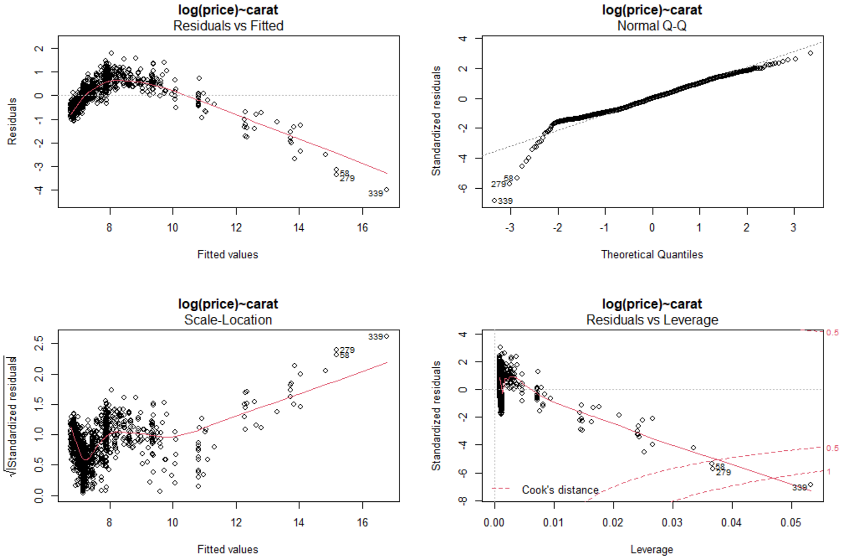
We start the process of model selection by addressing the non-linear pattern of Price against Carat. To discover an appropriate transformation of Price and/or Carat, we view assumption plots for a simple, non-transformed model: Price~Carat.



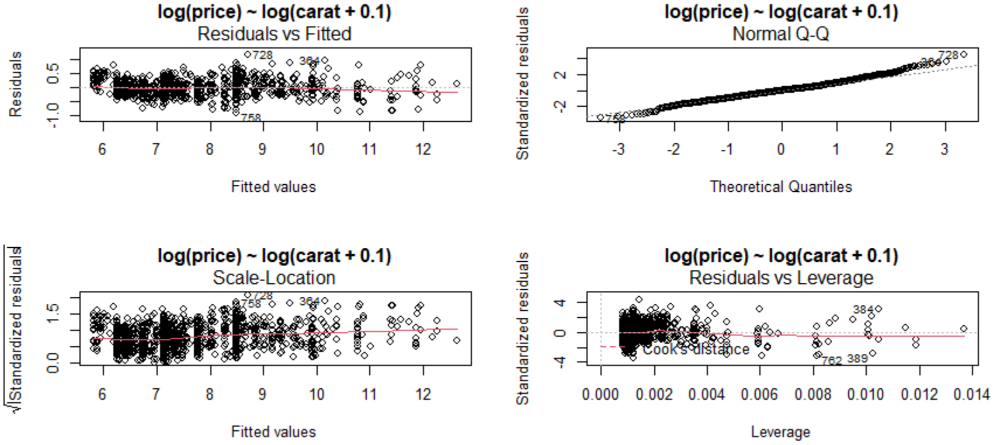
The Residuals vs Fitted plot shows that the residuals do not have constant variance, so we decided to transform the response variable, Price, to fit a linear model. The Box-Cox method was applied to determine how to transform Price.



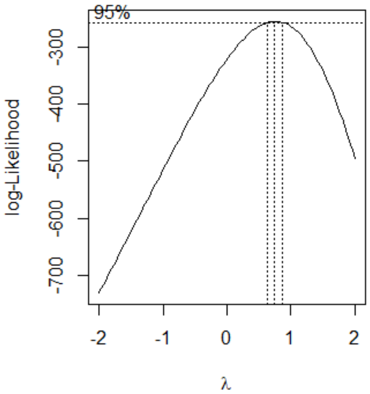
The Box-Cox plot gives a lambda value of approximately 0.3, however, that is relatively close to 0, and using a log transformation results in more interpretable results. Thus, we decide to use a log transformation.

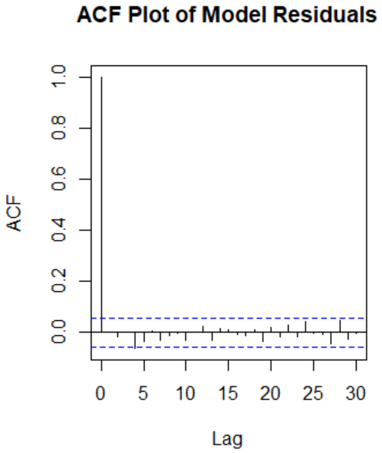


After transformation, the constant variance assumption does not appear to be completely fixed. There are several data points hovering in the lower section (0) of our data, so we decide to test small steps on the log-transformed Carat variable, leading us to choose log(Carat + 0.1) for the predictor variable.



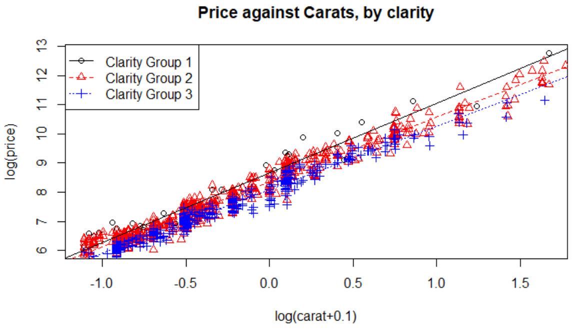
Viewing these revised assumption plots, the constant variance assumption has improved significantly. As an added bonus, the normality assumption also has become more valid. Finally, we check the following two plots for the other linear regression assumptions.





The Box-Cox plot shows a lambda confidence interval very close to 1, and the ACF plot only has one small significant value. These two small discrepancies end up improving and fitting the assumptions closer as we add categorical variables to the model. Thus we decide to move forward in our model selection process using the transformed quantitative variables of log(Price) and log(Carat + 0.1).

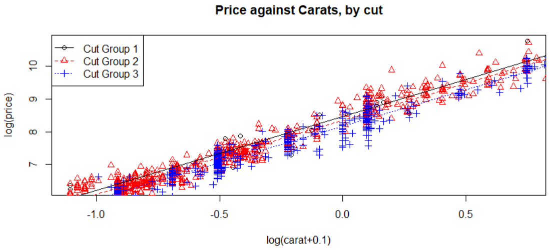
At this point, we want to add in some of the categorical variables. We can create the following series of plots, that are similar to plots in the EDA section of the data to help potentially group some of the levels of the categorical variables.



After the transformations, we are able to group the levels of the Clarity group into 3 distinct groups. This is done by the following groups: Clarity Group 1: FL + IF, Clarity Group 2: VS1 + VVS1 + VS2 + VVS2 and Clarity Group 3: SI1 + SI2.

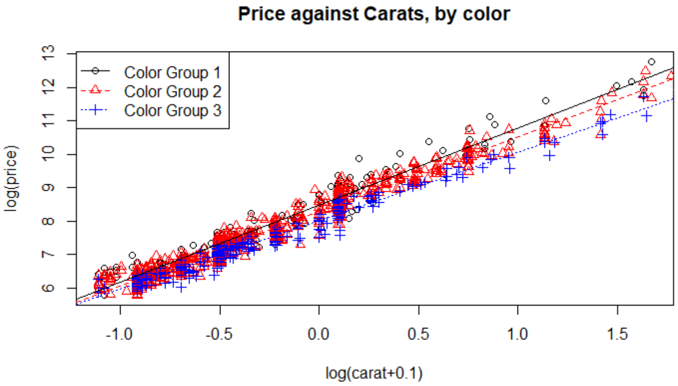
In the plot above we can see that these groupings appear to work well on the transformed data, and they help with the concern of FL only having 3 observations as mentioned in the EDA portion above. All of the slopes for the different Clarity groupings appear to be essentially the same, thus suggesting that interaction terms most likely will not be needed.

We follow a similar pattern for the other variables condensing down the levels into 3 particular groupings. The groupings for Cut are as follows: Cut Group 1: Astor Ideal, Cut Group 2: Ideal, and Cut Group 3: Good + Very Good .The plot below shows the scatterplot with the groups aforementioned.



The new levels of Cut score also appear to have very similar slopes suggesting that interaction terms are not necessarily needed following the quantitative transformations. The following process is also performed for the variable Color.

The levels of Color are broken down into the following: Color Group 1: D, Color Group 2: E + F + G + H, and Color Group 3: I + J.



The story for Color follows the same pattern as the previous two variables. The groupings all appear to have similar slopes, thus the need for interaction terms has become less likely for our model.

With the knowledge from these plots of the transformed model and consolidated levels of the categorical variables, we are ready to begin testing which categoricals we should potentially use in the model. From the plots above we also can choose to select “group 3” as the reference levels for each of the respective variables. To begin this process we ran a step function using both forward and backwards selection with the following small and full models:

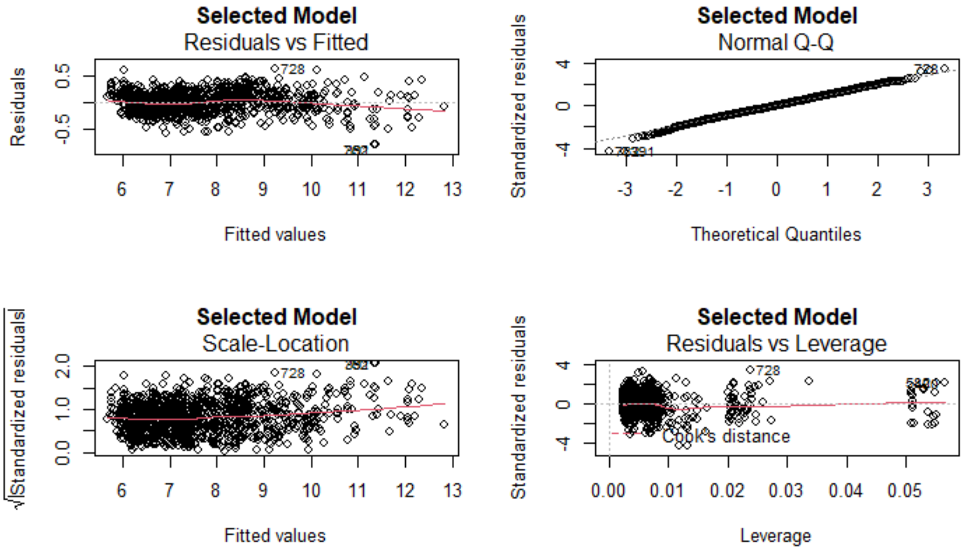
$$small\,model\colon\ log(Price)\ =\ log(Carat\ +\ 0.1)$$

$$full\,model\colon\ log(Price)\ =\ log(Carat\ +\ 0.1)\ +Clarity\ +\ Color\ +\ Cut$$

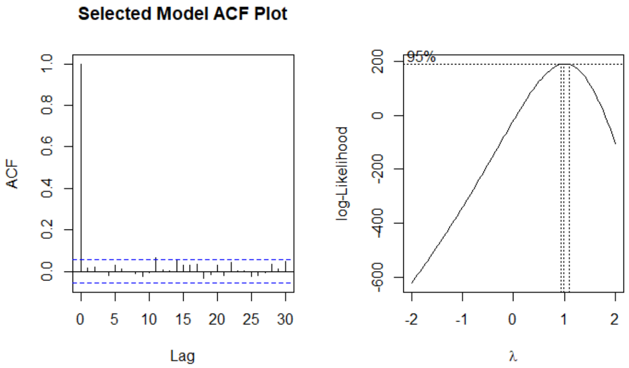
Both forwards and backwards selection methods result in the same model:

$$selected\,model\colon\ log(Price)\ =\ log(Carat\ +\ 0.1)\ +Clarity\ +\ Color\ +\ Cut$$

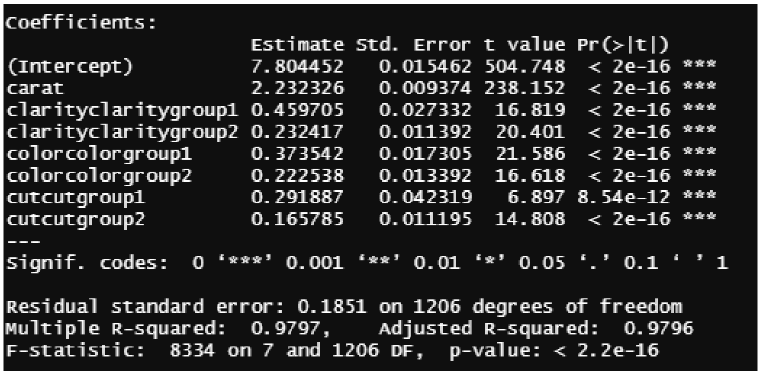
After trying out all interaction terms, some interaction terms seemed to have a small influence on the model, however, when we view the relevant plots above, all groupings have similar slopes. So, we decide to ignore the interaction terms despite the partial F-test leaning towards keeping them. The following assumptions plots for the selected model are below.



The plots all appear to pass the linear regression assumptions. There is one potential concern for our model, which is the one lag that is just barely significant in our ACF plot pictured below.



The ACF plot shows one lag that is slightly significant, suggesting the independent assumption may be slightly violated. The Box-Cox plot shows confidence that no further transformation is needed. Furthermore, with the following summary output of our model showing an R-squared value of 0.9797, we are confident in its ability to predict the Price of diamonds based on the 4C's. The summary output of our model is as follows and leads us to the following regression model.



```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          7.804452   0.015462 504.748  < 2e-16 ***
carat                2.232326   0.009374 238.152  < 2e-16 ***
clarityclaritygroup1 0.459705   0.027332  16.819  < 2e-16 ***
clarityclaritygroup2 0.232417   0.011392  20.401  < 2e-16 ***
colorcolorgroup1     0.373542   0.017305  21.586  < 2e-16 ***
colorcolorgroup2     0.222538   0.013392  16.618  < 2e-16 ***
cutcutgroup1         0.291887   0.042319   6.897 8.54e-12 ***
cutcutgroup2         0.165785   0.011195  14.808  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1851 on 1206 degrees of freedom
Multiple R-squared:  0.9797,    Adjusted R-squared:  0.9796 
F-statistic:  8334 on 7 and 1206 DF,  p-value: < 2.2e-16
```

$$
\begin{aligned}
\mathit{Final\ Model}\colon \log(\mathit{price}) \\
&= 7.80442 + 2.232326(\log(\mathit{carat} + 0.1)) + 0.459705\big(I(\mathit{Clarity\ Group} = 1)\big) \\
&+ 0.232417\big(I(\mathit{Clarity\ Group} = 2)\big) + 0.373542\big(I(\mathit{Color\ Group} = 1)\big) \\
&+ 0.222538\big(I(\mathit{Color\ Group} = 2)\big) + 0.291887\big(I(\mathit{Cut\ Group} = 1)\big) \\
&+ 0.165785\big(I(\mathit{Clarity\ Group} = 2)\big)
\end{aligned}
$$

The interpretation of our transformed model can go as the following as carat size increases:

If we start the process with our regression model (simplified):

$log(price)\ =\ log(carat\ +0.1)\ +\ C\ where\ C\ =\ sum\ of\ intercept\ and\ any\ categorical\ predictors$

When carat is increased by 10% we have the following equation:

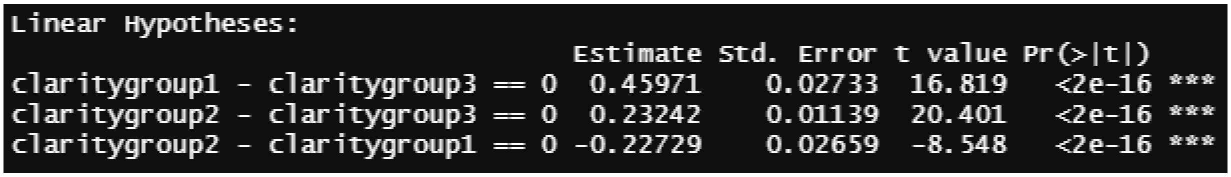
$log(price_{new})\ =\ log(1.1carat\ +\ 0.1)\ +\ C\ where\ C\ =\ sum\ of\ intercept\ and\ any\ categorical\ predic$

Next it follows this process:

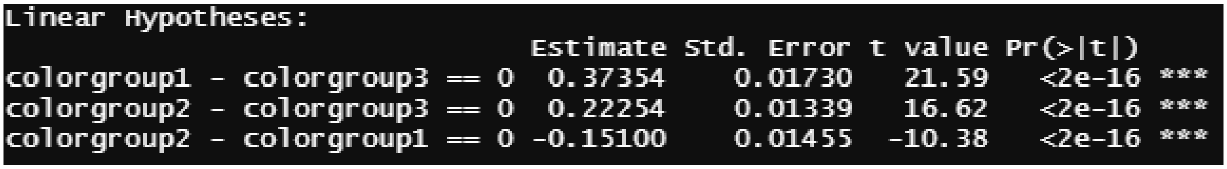
$$log(price_{new})\ -\ log(price)\ =\ Beta\ *\ log(1.1)$$
$$log(price_{new}/price)\ =\ log(1.1)^{Beta}$$
$$price_{new}/price\ =\ 1.1^{Beta}\ =\ 1.1^{2.232326}$$

From the equations above we see that for the increase of 10% in carat the predicted price is increased by a factor of $1.10^{2.232326}$. In general this can be written as for an increase of x% in the carat variable, the predicted price is increased by a factor of $(1\ +\ x/100)^{2.232326}$. This takes into account when all of the categorical predictors are held constant. The constant C will change as the categorical variables change.

To double check that all of the categorical variables groupings are significantly different from one another, we can run the pairwise tukey tests to receive the following results. From the output below we can see while holding all other variables constant, the mean effect changes for different levels of the categorical variables. For example, when clarity is equal to clarity group 1, there is, on average, a 0.45971 increase in log(price).[2]

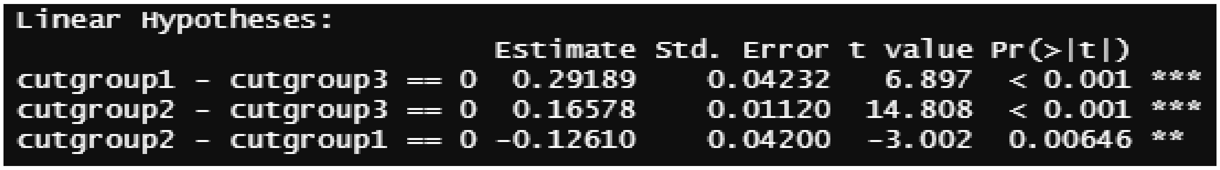


```
Linear Hypotheses:
                                  Estimate Std. Error t value Pr(>|t|)    
claritygroup1 - claritygroup3 == 0  0.45971    0.02733  16.819   <2e-16 ***
claritygroup2 - claritygroup3 == 0  0.23242    0.01139  20.401   <2e-16 ***
claritygroup2 - claritygroup1 == 0 -0.22729    0.02659  -8.548   <2e-16 ***
```



```
Linear Hypotheses:
                              Estimate Std. Error t value Pr(>|t|)    
colorgroup1 - colorgroup3 == 0  0.37354    0.01730   21.59   <2e-16 ***
colorgroup2 - colorgroup3 == 0  0.22254    0.01339   16.62   <2e-16 ***
colorgroup2 - colorgroup1 == 0 -0.15100    0.01455  -10.38   <2e-16 ***
```

[2] (clarity group 3 is the reference class)



```
Linear Hypotheses:
                             Estimate Std. Error t value Pr(>|t|)    
cutgroup1 - cutgroup3 == 0  0.29189    0.04232   6.897  < 0.001 ***
cutgroup2 - cutgroup3 == 0  0.16578    0.01120  14.808  < 0.001 ***
cutgroup2 - cutgroup1 == 0 -0.12610    0.04200  -3.002  0.00646 ** 
```

Model Selection Summary and Further Analysis

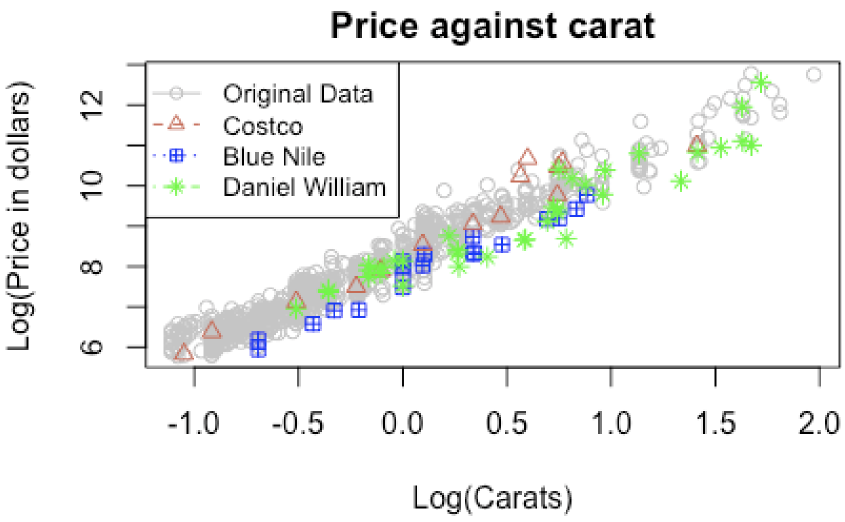
After our EDA and model selection process, we believe that our model can be successful in predicting changes in prices of diamonds within the given data set based on several settings. All of the categorical variables prove to be significant and are different based on the groupings of levels selected. Since the transformation, the model will be useful in showing the increases or decreases in diamond prices as Carat size fluctuates between observations. It also takes into account what Color, Cut rating, and Clarity rating the diamond has and will give a sound estimate of diamond price using the given data set. Next steps of this study would be to potentially develop a diamond pricing model for different brands and see how our model performs on diamond price data from unrelated sources. We performed some of this preliminary research below.

Use of Model as a Consumer Price-Shopping Tool

Can our model predict diamond prices from external data sources? Our expectations are low since we do not know how samples for the original data source were selected, and it is likely that external data will have different selection bias than the original. Additionally, external data may not have the same categorical identifiers as the original dataset. For example, “K” Color diamonds or “Excellent” Cut diamonds are in external data but are absent in the original dataset. Luckily, we know the original data is a few years old, so inflation is unlikely to be a major source of price differences.

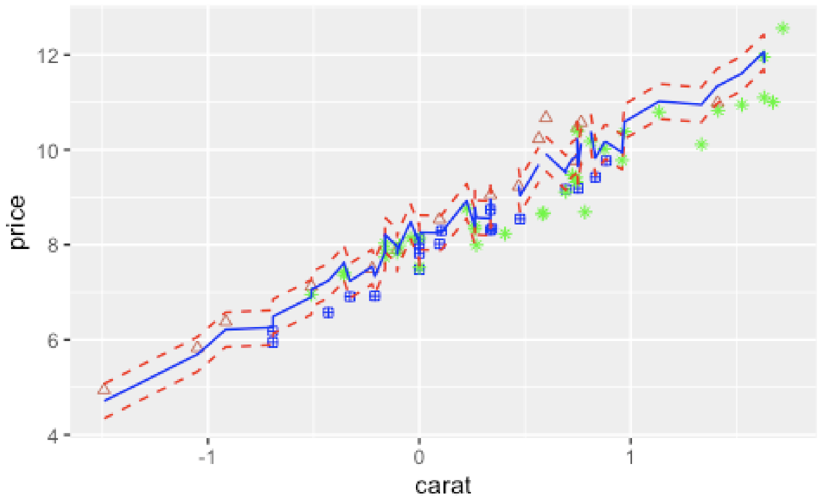
We chose three external data sources: the retail websites of Costco, a Los Angeles diamond seller named Daniel Willam, and the current selection of Blue Nile. Costco’s online diamond inventory does not contain loose diamonds, so to approximate the price of loose stones for comparison, we selected 15 samples from Costco's diamond solitaire rings and single-diamond stud earrings, 'guesstimated' the value of their settings, and subtracted the setting value from the jewelry price. This guesstimation introduces a source of comparison error. The number of such rings and earrings for sale is small, so we recorded data from all Costco listings, eliminating selection bias. The other two sources sell loose diamonds, so they didn't have the setting-cost issue that Costco does. However, they do have huge inventories that raise the issue of sampling bias. We did not have the time to develop an unbiased sampling technique, so instead we manually adjusted the dials of their diamond-selection tools in ways that seemed likely to select a range of stones. Since the results for these two vendors' dial setting are displayed in ascending order of price, however, our sample likely has a bias towards lower price. We collected 19 and 36 round diamond price samples from Blue Nile and Daniel William respectively.

The prices of each externally priced diamond appear visually in line with the original sample:

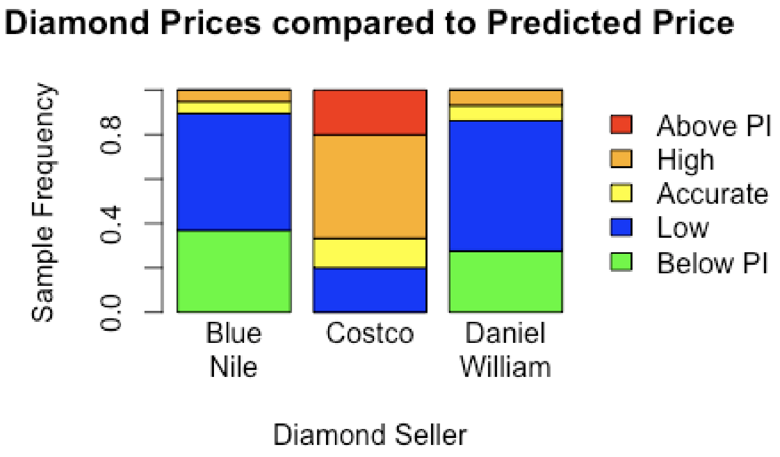


We look further at the points inside and outside the prediction interval using this plot:

log(Price) against log(Carat + 0.1)



The red-dotted lines define the prediction interval, and the blue line represents the mean fitted price result from our model. The red diamonds, small blue grids, and green snowflakes are data points from Costco, Blue Nile, and Daniel William respectively. The plot suggests that the samples from Blue Nile and Daniel William are on the low side of the prediction interval, and the Costco samples are on the high side. We look at the same data using another method.



We compare each price from the externally sampled diamonds against our model's prediction interval (PI) and categorize it as above, below, or within the PI. Within the PI, we further break down the price as "Accurate" (defined informally as within 5% of the fitted mean), "Low" (between the lower PI bound and the Accurate range), or "High". These results reinforce that the majority of stones selected from Blue Nile and Daniel William are priced lower than the fitted mean, whereas the stones from Costco have wider variation from the fitted mean and tend to be higher priced. Although the prediction intervals look narrow when graphed at a log scale, in real-dollar terms, the top of the interval is about twice the price of the bottom. When one goes to purchase a diamond worth thousands of dollars, that range is too large to be useful.

A relevant question for a consumer might be: where can I shop for diamonds and get the best value for my money? Because our sample-selection method differed so greatly between Costco and the other two sources, we cannot say how Costco compares to other sources as a place to purchase inexpensive diamond jewelry. However, the bar charts above suggest that Blue Nile and Daniel William have a comparable range of prices. Since their inventory is available online, and is thus easy for buyers to compare, this is not surprising. The similarity of results between Blue Nile and Daniel William provides evidence that our model can work consistently for purchases from multiple sources.

One caveat to consider is that diamond valuation is particularly subjective. Diamond certification standards can vary by retailer and lab, meaning different appraisers can characterize a diamond's features differently. This, along with other factors such as diamond country of origin and lab-grown vs mined diamonds would be an interesting feature to add to our model to see how it would affect pricing.

Github: tjmcintyre/STAT6021_Project1: Used for project 1 in Stat6021 (github.com)