Project 1: Analysis of Blue Nile Diamond Prices

Abhishek Bada, Christian Schroeder, and Timothy Tyree

School of Data Science, University of Virginia

STAT 6021: Linear Models for Data Science

Dr. Jeffrey Woo

April 5, 2021

Table of Contents

Executive Summary

This report provides an analysis and evaluation of the variety of factors that can drive diamond pricing. The questions answered in this paper are:

- Is it beneficial to group the classes of the categorical variables?
    - Yes, by grouping certain categories in the predictor variables we were able to create a model that accounted for more variation at a slight loss of adjusted R^2 value. (See "Grouping Categorical Variables" for details)
- Which variable, after carat, influences the prediction the most?
    - We found that "Clarity" was the most influential predictor variable after "Carat". (See "Secondary Predictor Analysis" for details)
- Can we create a model that better predicts price than a model that only uses carat as the single variable?
    - Yes, the model `lm(log(price)~ log(carat)+cut+color+clarity)` was better than using "Carat" as our only predictor. (See "Testing the Final Model" for details)

The results of this research show what we believe to be the optimal model for determining the price of diamonds at Blue Nile. Specifically, it was determined that combining several attributes of key variables into larger groups would provide more flexibility for predicting future prices, while retaining accuracy.

The Data

The data used documents four key variables and the price of 1,215 diamonds that were available for purchase from bluenile.com at the time of download. The key variables are carat, the measure of the diamond's weight (1 carat = 0.2 grams), cut, color, and clarity, variables that follow specified rankings of the quality of the diamond. These are used as the predictor variables in the final model with price as the response. The cut of a diamond is a grade given by Blue Nile and denotes the quality of the cut into the following categories defined by Blue Nile:

- Astor Ideal (Astor by Blue Nile^{TM}): Crafted to gather and reflect the most light possible.
- Ideal: Reflects most light that enters the diamond.
- Very Good: Reflects nearly as much light as the idea cut, but for a lower price.
- Good: Reflects most of the light that enters, but not as much as a Very Good cut grade.

Blue Nile does not provide their own grading of the other categorical predictors, rather they rely on color and clarity grading done by the Gemological Institute of America (GIA). The color grades included in this dataset range from D to J and are defined by Blue Nile as the following:

- D: Absolutely Colorless, highest grade and is extremely rare.
- E: Colorless.
- F: Colorless, but differences detectable only by a gemologist.
- G: Near Colorless, very slight warmth to their tone.
- H: Near Colorless, faint yellow hue.
- I: Near Colorless, slight yellow tint more detectable than in H grades.
- J: Near Colorless, slight yellow tone more detectable than in I grades.

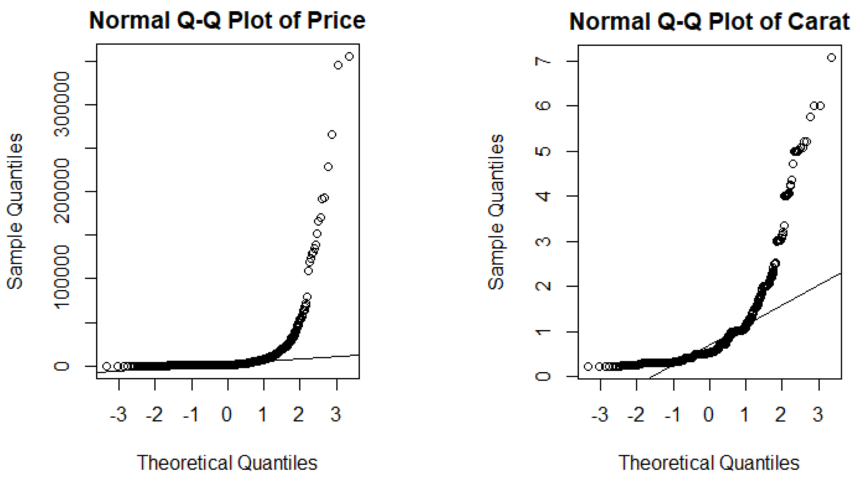
The last variable in the dataset is clarity. The GIA clarity grading scale has a range of 11 clarity grades, but only the following eight are present in this dataset:

- FL: Flawless, no internal or external characteristics.
- IF: Internally Flawless, some small surface blemishes may be visible under a microscope.
- VVS1 and VVS2: Very, Very Slightly Included, minuscule inclusions that are difficult even for a trained eye to see under 10X magnification.
- VS1 and VS2: Very Slightly Included, minuscule inclusions difficult to see at 10x magnification.
- S1 and S2: Slightly Included, inclusions are noticeable at 10x magnification.

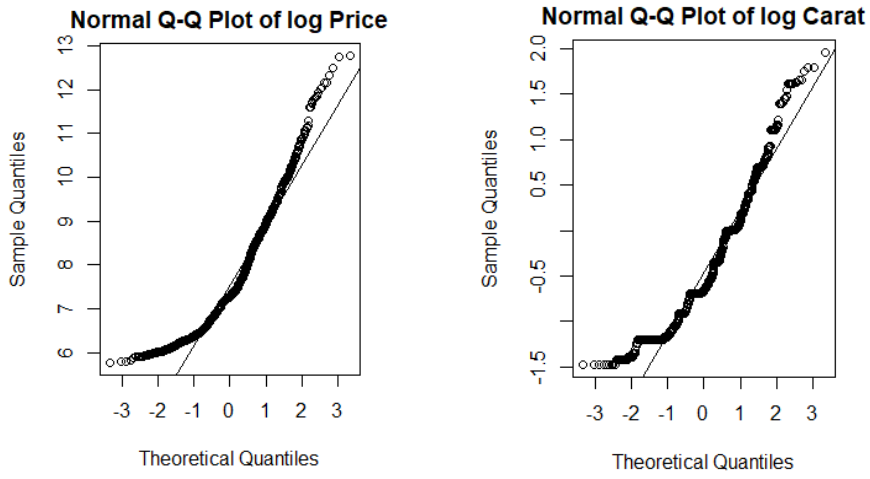
The Exploratory Data Analysis section explains why we decided to combine several of these categories, as well as categories for the other predictors.

Exploratory Data Analysis

In our exploratory analysis we focused on linearizing the model as well as reviewing the correlations between the categorical predictors cut, clarity, and color, their levels, and the price. Blue Nile states on their website that the carat of a diamond is the largest factor when pricing a diamond. This was confirmed in our analysis visually and statistically. The carat of a diamond and its price have a correlation of 0.827. This can be seen in the following QQ plots:

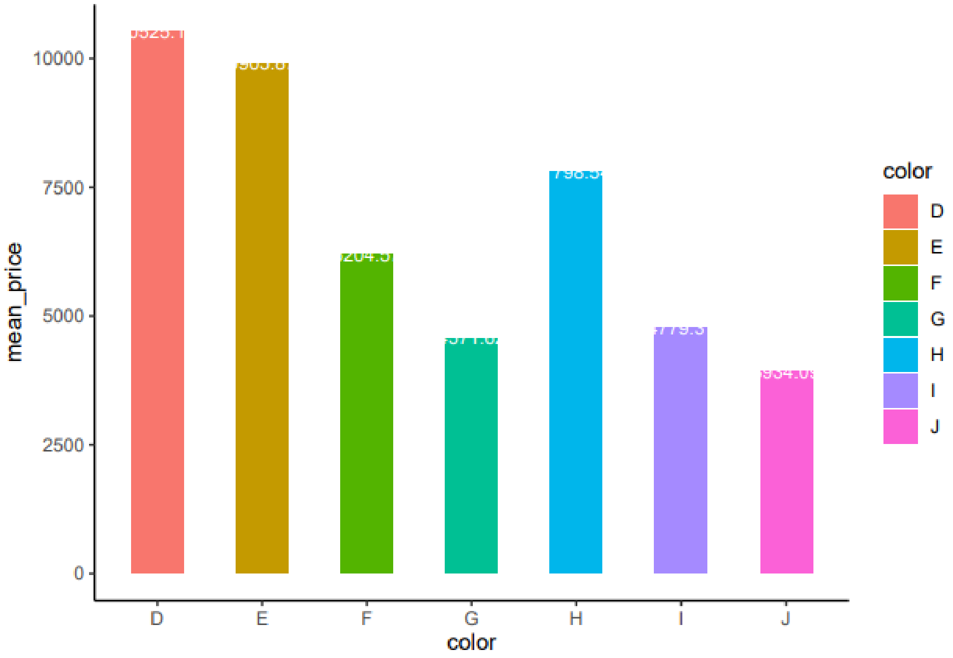


Also visible in these plots is the non-linear form that each variable follows. To fix this, the log of price and log of carat were plotted:

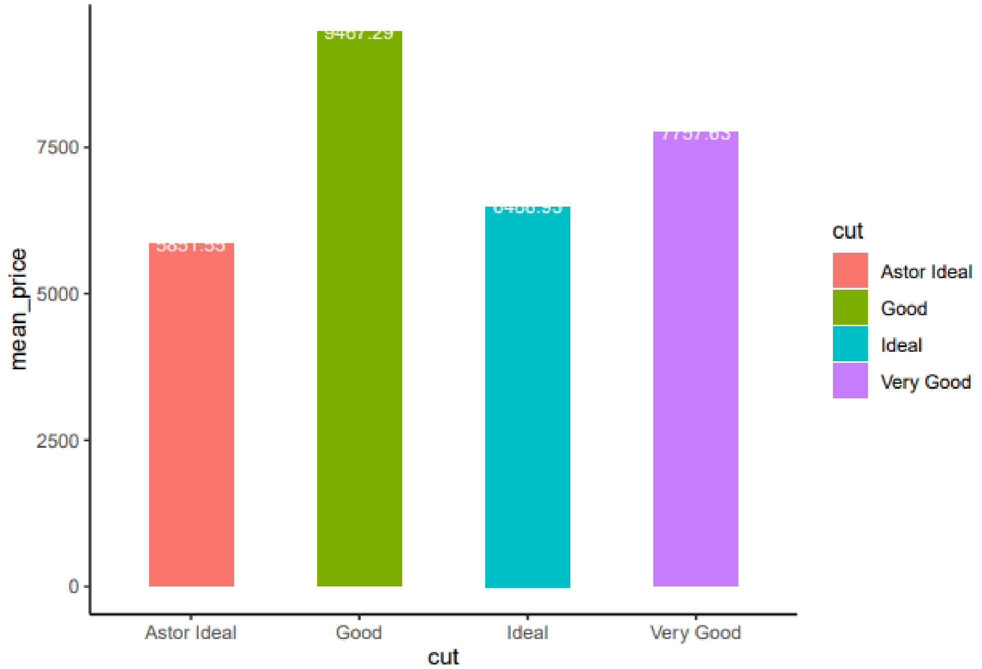


To further compare this transformation we summarized two linear models, one using the data as is, and one using the log of price and log of carat. The as-is model has an adjusted R-squared value of 0.7219, while the transformed model has an adjusted R-squared much higher at 0.9832. Because the transformed model can explain far more variation in the response, it was determined that the log of price and log of carat should be used in the final model to linearize the regression.

Additional analysis was done on the categorical variables to determine if it would benefit the model to merge several classes into larger groups, in turn reducing the complexity of the model. To determine this, each categorical variable was plotted against price to see whether any classes were similar to one another. The first variable inspected was color. A histogram of color classes and mean price shows similarities between the D, E, and F classes, as well as the G, H, I, and J classes.

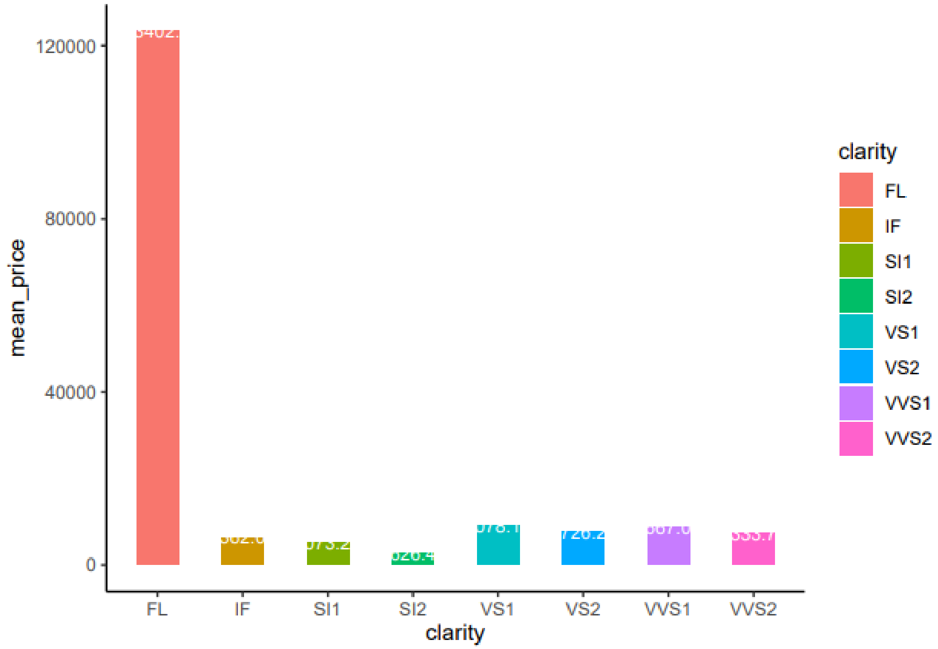


It would most likely be beneficial to the model to group these classes into Nearly Colorless and Near Colorless. The next variable inspected was cut. The histogram of cut versus mean price shows it could be beneficial to group the classes into Ideal and Good.



Something to note is that the mean prices of Astor Ideal and Ideal cuts fell below those of Good and Very Good, even though they are considered rarer and more expensive cuts. This could be attributed to a variety of reasons; however, we suspect it is due to the combination of carat being the strongest driver of price and Good and Very Good cuts being much more prevalent, and more likely to be larger diamonds than the rare Ideal and Astor Ideal cuts.

The last categorical variable is clarity. It was expected that a few classes could be combined for clarity, given their identical definitions provided by Blue Nile. When plotted on a histogram against mean price, it is difficult to determine similarities due to the significant difference between the FL class and the other seven.



The mean price of FL diamonds is above $120,000, while that of the other classes aren't even a quarter of that price. This gives two main options for combining the classes. The first option being to combine the other seven classes, creating FL for flawless and NF for not flawless. The second option being to instead combine the classes based on their grade definitions mentioned earlier, combining FL and IF to F, SI1 and SI2 to SI, VS1 and VS2 to VS, and VVS1 and VVS2 to VVS.

Secondary Predictor Analysis

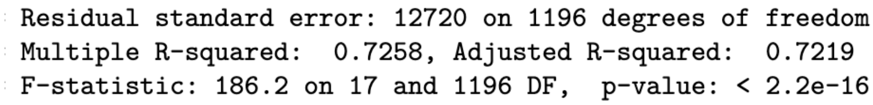
We also questioned which categorical variable was the most influential, after carat, to the price of a diamond. To test this we compared a single variable model of price against carat to three alternate models that each included one of the categorical variables. The models were:

- Reduced Model: `lm(price~carat)`
- Alternate Model 1: `lm(price~carat+cut)`
- Alternate Model 2: `lm(price~carat+clarity)`
- Alternate Model 3: `lm(price~carat+color)`

A partial f-test, `anova(alternate, reduce)`, was run for each alternate model, comparing it to the reduced model. The resulting p-values for adding each variable were 0.006988 for cut, 2.2e-16 for clarity, and 3.547e-05 for color.

Model Building Process

We tested a multiple linear regression model at every step. The first two models compared the predictive power between the data as is, and the data log transformed, the categorical variable classes had not been combined yet. The first, untransformed model offers no additional benefit, and a user would be better able to predict the price of a diamond with a single linear regression model using carat as the predictor variable.

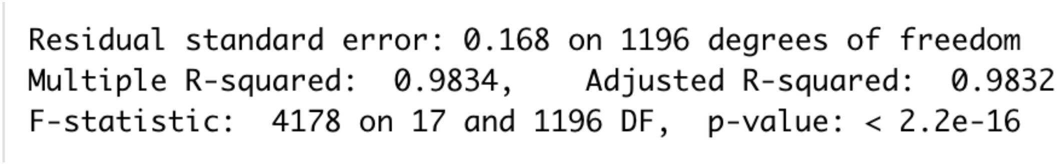


```
Residual standard error: 12720 on 1196 degrees of freedom
Multiple R-squared:  0.7258, Adjusted R-squared:  0.7219
F-statistic: 186.2 on 17 and 1196 DF,  p-value: < 2.2e-16
```

There was a drastic increase in the adjusted R^2 value from 0.7219 to 0.9832 for the transformed model, which tells us that the log transformations were necessary.

Grouping Categorical Variables

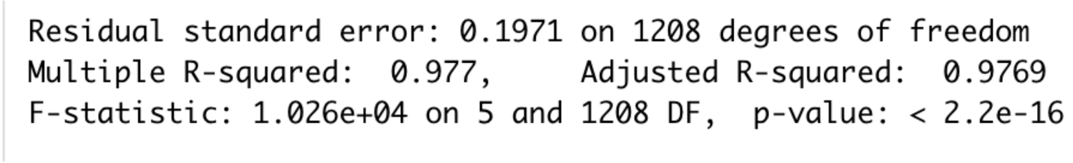
There were a large number of classes within each of the categorical variables, which brought up concerns about whether the model was overfitting the training data and would not be generalizable on predicting unseen test data.



```
Residual standard error: 0.168 on 1196 degrees of freedom
Multiple R-squared:  0.9834,    Adjusted R-squared:  0.9832
F-statistic:  4178 on 17 and 1196 DF,  p-value: < 2.2e-16
```

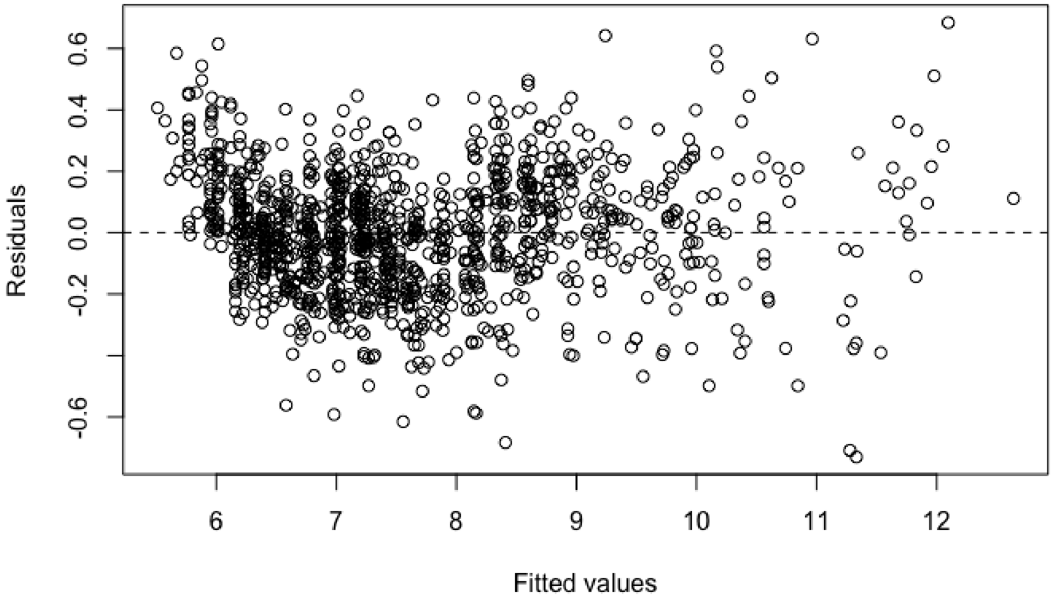
Testing the Final Model

Following the grouping of similarly distributed classes, we ran a final log transformed multiple linear regression model, `lm(log(price)~ log(carat)+cut+color+clarity)`, and compared its predictive power to the previous model. The adjusted R^2 value slightly decreased, but the F-statistic doubled. The F-statistic indicates whether a significant amount (significantly different from zero) of variance was explained by the model. The doubled F-statistic followed with a very slight decrease in adjusted R^2 gave us confidence that the grouped classes model will generalize better than the previous model.

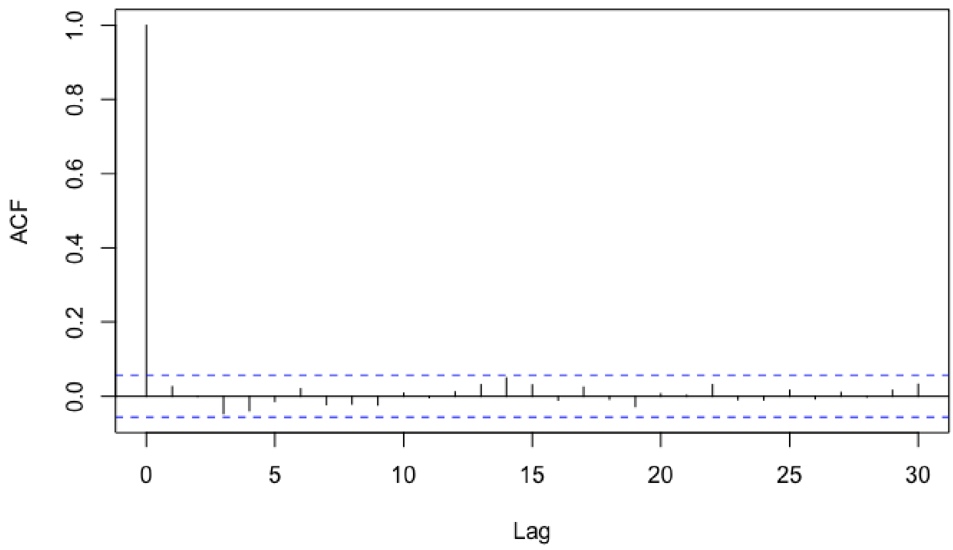


```
Residual standard error: 0.1971 on 1208 degrees of freedom
Multiple R-squared:  0.977,     Adjusted R-squared:  0.9769
F-statistic: 1.026e+04 on 5 and 1208 DF,  p-value: < 2.2e-16
```

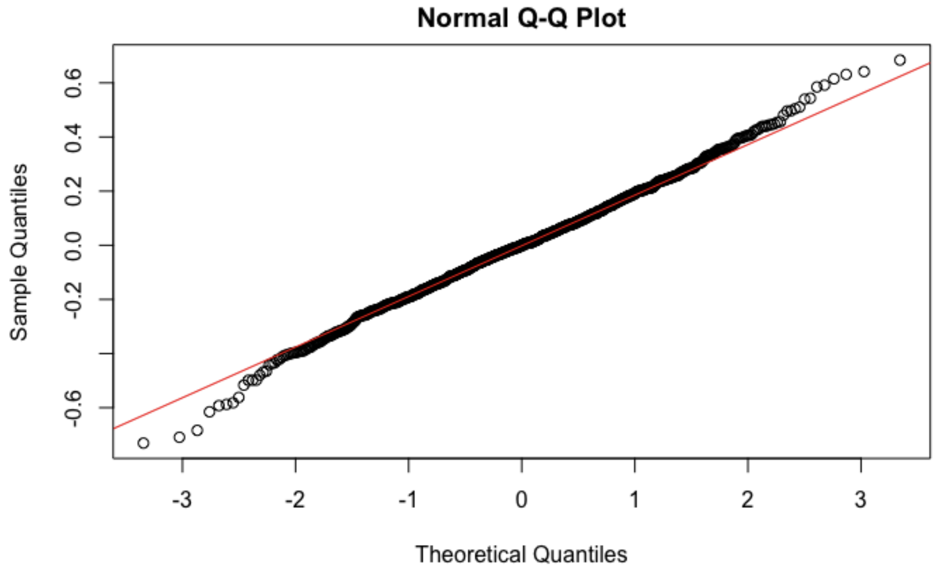
Last we tested the linearity assumptions for our final model. In the residual plot there were no obvious curves present and the data points were evenly distributed above and below the horizontal line, which indicates that the mean is near zero.



We then tested for multicollinearity by performing an auto correlation function and plotting the results. All of the lags were within the boundary, which told us that we do not have an issue with multicollinearity.



Lastly, we compared the probability distribution by plotting their quantiles, and the data points closely followed the regression line. All of our linearity assumptions were then met.



Conclusions

Is it beneficial to group the classes of the categorical variables?

When comparing our final model to the original unfiltered model, we find that grouping classes together in a logical manner accounts for more variance within the model, and allows our model to generalize stronger than the original model. From this we can conclude that grouping the classes of the categorical variables was beneficial to the model while also reducing its complexity.

Which variable, after carat, influences the prediction the most?

To determine which variable was the most influential after carat, we compared a reduced model of price against carat to alternate models that included each categorical variable. From this we found that clarity was the most influential variable to the price of a diamond, after carat. This significance is also present in the original model.

Can we create a model that better predicts price than a model that only uses carat as the single variable?

Using the Partial f-test we compared a model of the log of price against the log of carat to our final filtered and transformed model. The resulting partial F statistic was 292.65, with a p-value of 2.2e-16. From this we can conclude that the addition of the grouped categorical variables did improve the model's ability to explain the price of a diamond.