Buckley Dowdle (bd6fr)
Xun Liu (xl4xw)
Michael Pajewski (mtp9k)
Jordan Machita (jm8ux)

Stat 6021: Project 1

Executive Summary

Cavalier Pawn Shop has recently been approached by a potential client looking to offload ten diamonds. In order to make an offer on the diamonds, Cavalier Pawn Shop needs to know at what price they will be able to resell them in order to build sufficient margin into their offer. However, the owners have never previously dealt with diamonds and thus need expert advice to make a competitive but profitable offer. Cavalier Pawn Shop hired our team of data scientists to provide this expertise with multiple goals in mind First and foremost, they need a reliable valuation of the ten diamonds under consideration. Secondly, they want to learn more about diamonds as they hope to expand their shop to higher end items.

Their experience in the pawn industry has taught them that a key skill is the ability to identify rare items and value them accordingly. The owners of Cavalier Pawn Shop know that the price of the diamond increases with carat size but are unaware of the specifics of this relationship, such as if it is linear. Additionally, they suspect that color, clarity, and cut also play an important role in determining the value as some are certain to be more common than others. We believe that if they have a clear understanding of these variables of color, clarity, and cut, they can better recognize the rarity of diamonds. They also want us to produce a model that then allows them to account for each of these factors in addition to carats to estimate the diamonds' value with a high degree of accuracy. With this heightened knowledge and our regression model at hand, Cavalier Pawn Shop will be equipped to make a significant profit with the ten diamonds brought to them and to become a local leader in the diamond sector of the pawn industry.

In order to meet the needs of Cavalier Pawn Shop, we obtained a dataset that includes the price, carat, color, clarity, and cut of over 210,000 diamonds. However, the owners have further informed us that they will almost exclusively be dealing with diamonds of two carats or less, so we are able to restrict our analyses to those parameters. From our dataset, we are still left with just shy of 200,000 diamonds. We have used this data to price each of the ten diamonds and to provide analyses on the rarity and characteristics of diamonds that are two carats or less.

In Table 1, we can see the characteristics of each of the ten diamonds offered by Cavalier Pawn Shop's client. In addition, we have included the recommended price of each diamond that was generated by our model. In the following sections, we provide a more detailed description of how we generated and evaluated our model that resulted in our price estimates. The resulting valuation of our analyses is the price at which we believe Cavalier Pawn Shop will be able to resell the diamond. It is important to note that this is not the price that Cavalier Pawn Shop should offer or show the seller of the diamonds as they need to acquire the diamonds below the price we have generated in order to make a profit.

	Carat	**Color**	**Clarity**	**Cut**	**Recommended Price**
Diamond 1	0.53	F	VS1	Very Good	$1,450.04
Diamond 2	0.32	H	VVS2	Ideal	$601.96
Diamond 3	1.75	G	IF	Good	$17,302.87
Diamond 4	0.95	H	SI1	Ideal	$4,075.48
Diamond 5	1.34	D	VS2	Ideal	$11,473.14
Diamond 6	0.24	G	SI1	Good	$228.22
Diamond 7	1.14	E	SI2	Very Good	$5,061.27
Diamond 8	1.89	J	VS1	Ideal	$13,391.93
Diamond 9	0.76	I	FL	Very Good	$3,675.65
Diamond 10	1.52	F	VVS1	Ideal	$15,548.27

Table 1: Ten Diamonds

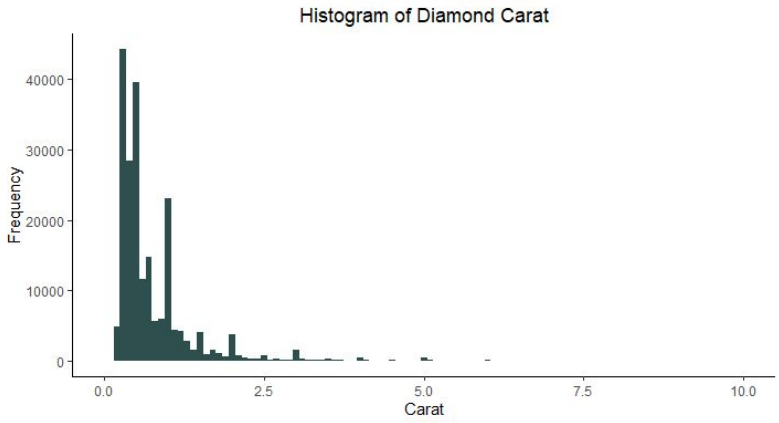
Data Description

The data used in our analysis consists of five attributes describing over 210,000 diamonds. The attributes are carat, clarity, color, cut, and price. For the purposes of our client, we limited our analysis to diamonds no greater than 2.0 carats. This resulted in an analysis of 199,598 diamonds from .23 carats to 2.0 carats. Carat is a description of the volume of each diamond. Clarity is a categorical factor that measures how clear the diamond is ranging from I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best). Color is another categorical factor measuring how colorless a diamond is. It is considered ideal for a diamond to be completely

colorless, represented by “D”, but our dataset extends to rating “J”. The final categorical variable measures the quality of the diamonds’ cut. The categories in our filtered dataset include Good, Very Good, Astor Ideal, and Ideal. Finally, price is a continuous variable describing the price in USD for which the diamond was sold. For diamonds under 2.0 carats, prices range from $229 to $66,380.

Data Background

Our team started with a high level analysis of the data based on graphical and statistical summaries to obtain a better understanding of the basic data structure and how the variables interact with the price of the diamond. Plot 1 shows the frequency of weight in the dataset is heavily skewed toward diamonds lighter than 2.5 carats. Table 2 represents the R summary breakdown of the carat variable. The largest diamond in the dataset has a weight of 20.34 carats and is significantly outside the third quartile of 1 carat. Our team made the decision to limit the dataset to diamonds under the 2 carat weight because 75% of the data is for diamonds under the 1 carat weight and the ten diamonds in the sample set provided by the client are all under the 2 carat weight.

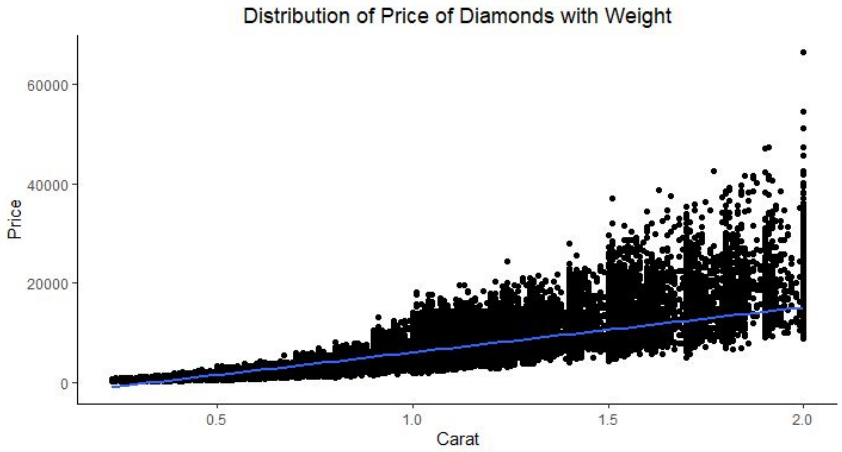


Plot 1: Histogram of Diamond Weight in Carat

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
0.23	0.37	0.51	0.7621	1	20.34

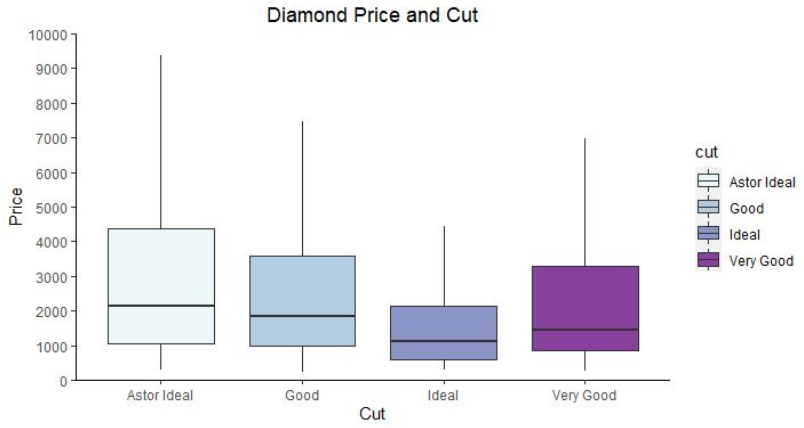
Table 2: Summary of Carat in Diamonds Dataset

After limiting the dataset to diamonds weighing under 2 carats, we looked at the distribution of price and weight for the diamonds in the dataset. Plot 2 shows that there is a positive relationship between the weight of the diamonds and their price. We can see the linear trend line does not fit the data and it suggests potentially an exponential relationship.The data also shows a dispersion of the relationship between price and weight as the weight increases. It should also be noted that the dense areas in the graph that create vertical stripes occur because carat weights take on a discrete value.



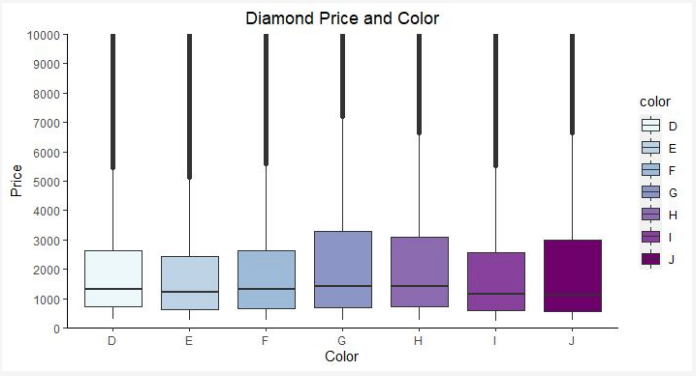
Plot 2: Distribution of Price of Diamonds with Weight

We then looked at how the variables cut, color, and clarity influenced the price of the diamonds to gain some insight into which were the most important variables to focus on in our model. Looking at the box plot for price by the four different cuts, Plot 3, we see that cut appears to influence the price of the diamond. It is interesting to note that the mean value for the least ideal cut, astor ideal, is higher than any of the other more premium cuts. We can not confirm that using cut in will influence the model without further analysis that will be discussed later in the report.



Plot 3: Diamond pierce according to cut

Comparing color to diamond price also shows a correlation. Plot 4 shows there is a difference between the mean price of each color type. The mean price increases by category from color D though G, peaks at G, and then decreases form G to J. For context the diamond colors scale ranges from D, a colorless diamond, to J, nearly colorless. It is interesting to note that our data shows perfect color diamonds on average are cheaper than the average diamond with a color value of G. This does not confirm that color will be useful in our price model but it gives us context for making decisions in our model and leaves us to consider it importances.

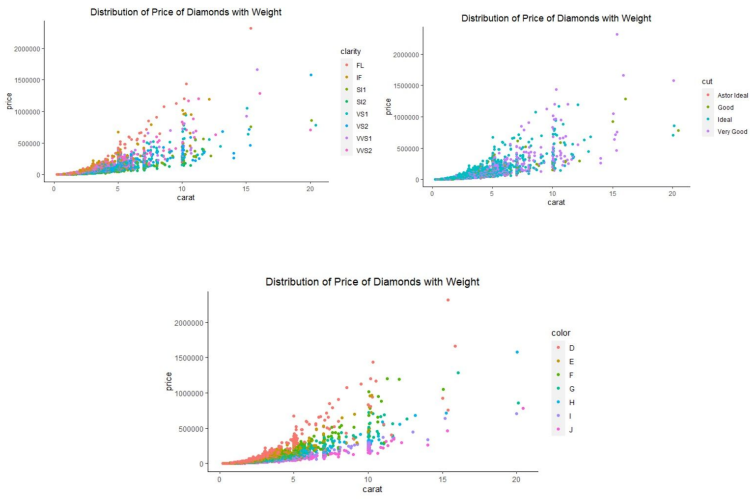


Plot 4: Diamond price according to color

Comparing price to clarity again shows a correlation between the variable and price. The clarity range spans from FL being the highest quality clarity to WS2 being the lowest quality. Plot 5 shows that the highest clarity diamonds have the highest mean price per category but there is not a strictly negative relationship from highest quality to lowest quality in terms of price.

Plot 5: Diamond Price according to Clarity

From the analysis of carat, cut, color, and clarity influence on price, it is clear that carat has the greatest direct impact on the price of diamonds but the other factors do individually influence the model. Plot 6 includes clarity, cut, and color individually overlaid on the distribution of price of diamonds with weight. Looking at the clarity graph at a given carat for almost all instances, the classification FL is priced higher than any other classification. A similar pattern is seen in cut and color, respectively, where Ideal is priced highest and color D is priced highest. Carat is the strongest driving factor in determining price but based on the exploratory analysis, clarity, cut, and color also may have an influence on price. In the following sections of the report, we use multiple linear regression techniques to further assess these relationships and determine the basis for our final model.



Plot 6: Clarity, Cut, and Color, on Distribution of Price of Diamonds with Weigh

Detailed Analysis

Model Selection

After conducting an exploratory analysis of the data, we decided to use carat, color, clarity, and cut as predictors for price. To aid in our decision making and identify other possible models, we used automated search procedures, including forward selection, backward elimination, and stepwise regression. The search procedures also confirmed that using all available predictors was an ideal model to begin with initially. However, they do not lend any insight into whether or not any interaction effects should be included.

In order to access the validity of using all of these variables simultaneously, we fit a linear regression model using all four predictors and examined the p-value of each. All p-values were far less than 0.05, showing that all added value to our model, even with the other predictors present. The evidence was now overwhelming that we should begin testing a model containing all predictors. Before testing the model to see if it met the necessary assumptions for linear regression, we assessed the value of adding interaction between our categorical variables. After assessing several models, we concluded that including a single interaction, between carat and color, produced the highest adjusted R^2 value and also had no insignificant predictors without adding too much complexity. The summary of this model (before any transformations) can be seen in Figure 1. While it is possible introducing multiple interactions could have produced a slightly more accurate model, we did not feel the added complexity was in the best interest of our client.

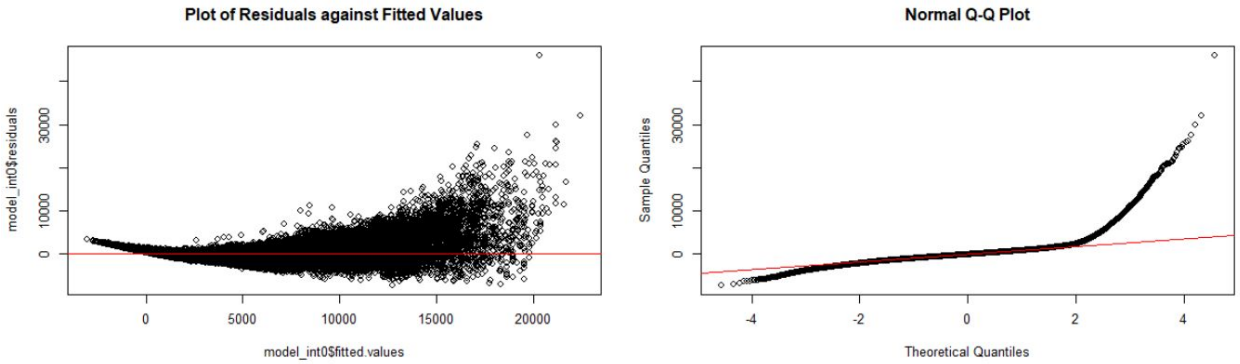
```
Call:
lm(formula = price ~ carat * color + clarity + cut)

Residuals:
   Min     1Q Median     3Q    Max
 -7170   -664    -68    554  46051

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  -1704.21      51.47  -33.114   <2e-16 ***
carat        12067.55      20.70  582.937   <2e-16 ***
colorE         738.31      20.36   36.260   <2e-16 ***
colorF         831.38      20.45   40.646   <2e-16 ***
colorG        1065.71      21.01   50.727   <2e-16 ***
colorH        1322.74      22.33   59.246   <2e-16 ***
colorI        1864.54      22.27   83.742   <2e-16 ***
colorJ        2147.74      26.39   81.380   <2e-16 ***
clarityIF    -1273.31      46.71  -27.262   <2e-16 ***
claritySI1   -2627.51      45.25  -58.067   <2e-16 ***
claritySI2   -2937.94      45.62  -64.397   <2e-16 ***
clarityVS1   -2068.99      45.29  -45.679   <2e-16 ***
clarityVS2   -2270.34      45.33  -50.086   <2e-16 ***
clarityVVS1  -1666.27      45.46  -36.656   <2e-16 ***
clarityVVS2  -1952.09      45.44  -42.961   <2e-16 ***
cutGood      -1171.08      25.04  -46.775   <2e-16 ***
cutIdeal        20.70      22.73    0.911    0.362
cutVery Good  -828.52      23.04  -35.962   <2e-16 ***
carat:colorE -1577.24      29.03  -54.334   <2e-16 ***
carat:colorF -2018.43      28.66  -70.433   <2e-16 ***
carat:colorG -2803.97      28.56  -98.182   <2e-16 ***
carat:colorH -3631.12      30.16 -120.402   <2e-16 ***
carat:colorI -4829.36      30.83 -156.643   <2e-16 ***
carat:colorJ -5782.45      35.26 -163.998   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1293 on 199574 degrees of freedom
Multiple R-squared:  0.8727,    Adjusted R-squared:  0.8727
F-statistic: 5.949e+04 on 23 and 199574 DF,  p-value: < 2.2e-16
```

Figure 1: Regression Summary without Transformations

Model Assessment

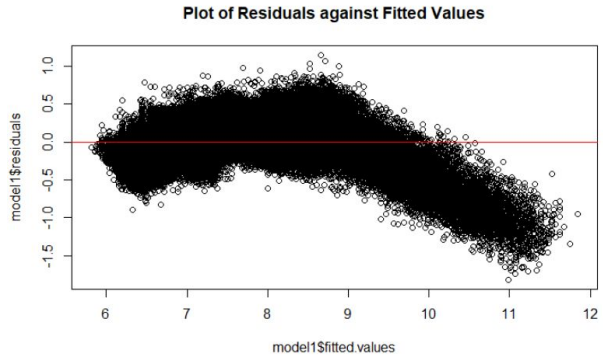
Our next step was to assess if our model met all the assumptions of a linear model. We began by plotting the residuals against the estimated value of y. We observed our model did not meet at least two of the assumptions of linear models. The residuals were not evenly scattered

around the horizontal axis and they did not have a similar vertical variation. The results of this scatter plot showed that the error terms did not have a mean of zero and that the variance was not constant as can be seen in Plot 7. In addition, from the Q-Q plot in Plot 8, the errors for each fixed value of x, do not follow a normal distribution (qq plot 1). As these assumptions are not met, the results from the corresponding hypothesis test are not reliable.



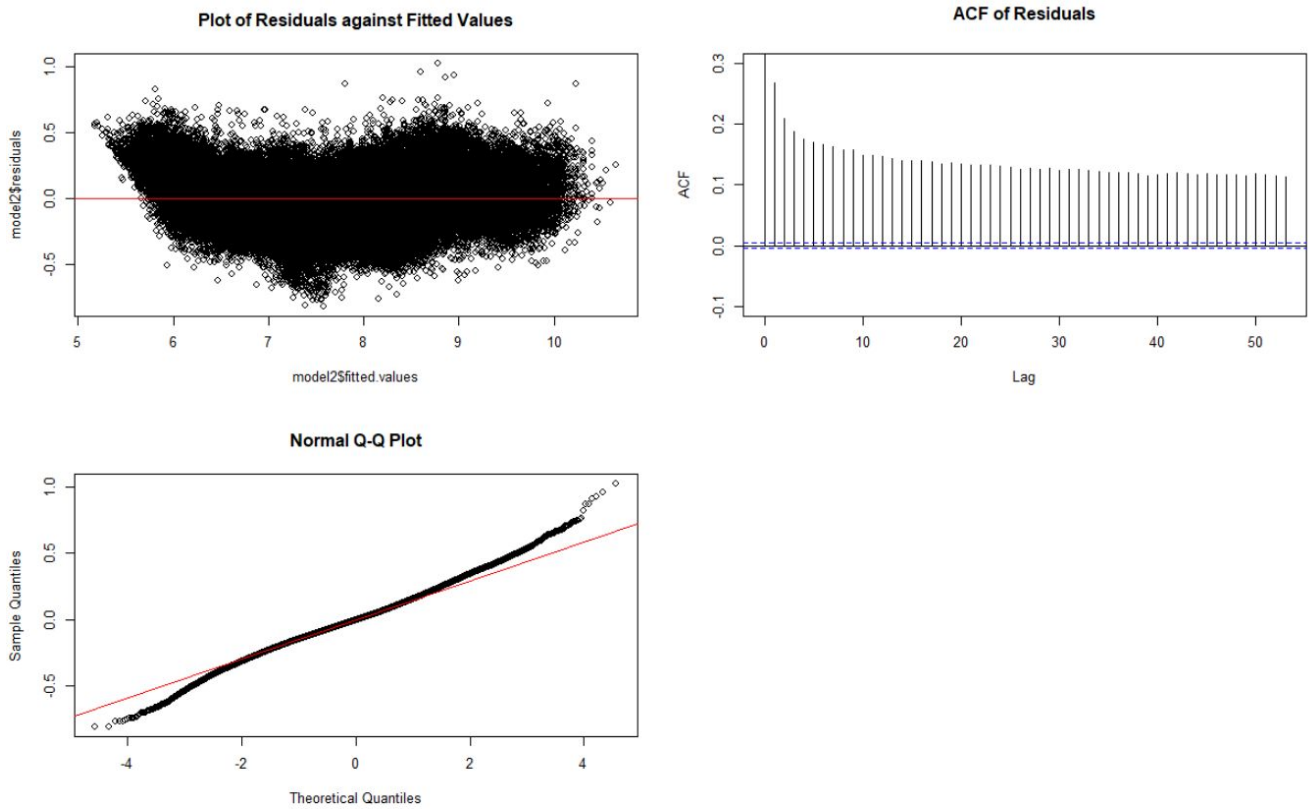
Plot 7 & Plot 8: Testing Regression Assumptions

To assess how to transform our variables in order to resolve these issues, we used a Box-Cox plot. The plot indicated that a natural logarithmic transformation of our response variable, price, would yield the best results. However, after implementing this transformation, our model still violated at least two assumptions as can be seen in Plot 9.



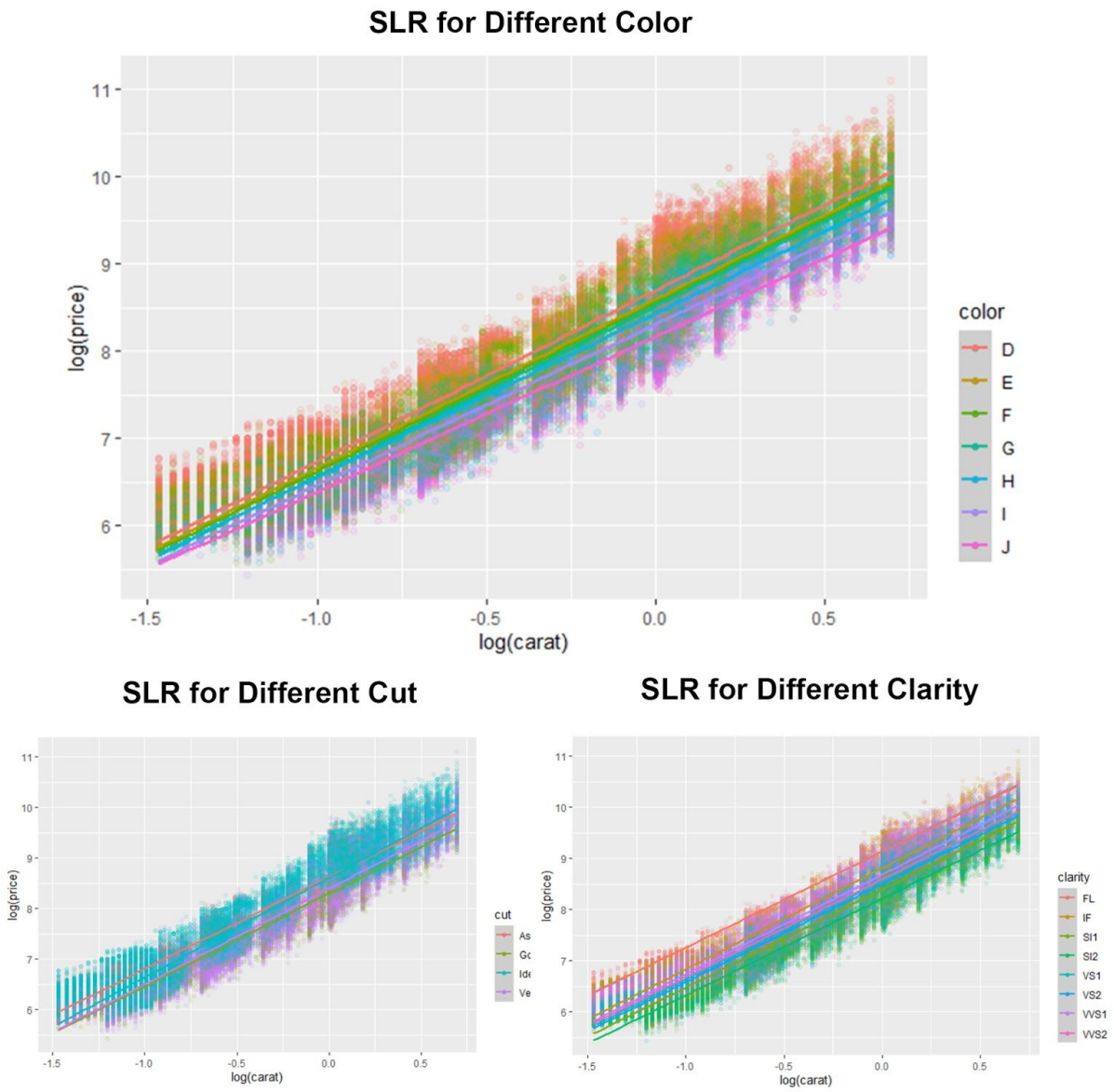
Plot 9: Residuals after Transformation of Response

We then decided to transform the predictor value carat, also with a natural logarithmic transformation. The resulting model shows a linear relationship between the predictor variables and the response variable, error terms with a mean of zero, and a constant variance; all of which are illustrated in Plot 10. The residuals also follow a normal distribution as can be seen in Plot 12. However, our model does show error terms to be significantly correlated, which is represented in Plot 11.



Plot 10, Plot 11, & Plot 12: Testing Regression Assumptions after Both Transformations

The scatterplots after both transformations seem to have strong linear correlations. As the spread of the response variable is wide, we conducted separate plots and overlay the regression lines for each category of the color, cut and clarity in Plot 13. We observe that the slopes of the regression lines for each category of Cut are similar, but the intercepts for the regression lines show significant difference. Similar conclusions can be drawn from the regression lines of the different categories of Clarity. However, the slopes of the regression line for each category of Color show significant differences, indicating the necessity of adding interactions between the variable Color and log(carat).



Plot 13: Scatterplots after Both Transformations

Finally, we conducted an analysis of variance (ANOVA) test and calculated the predicted residual error sum of squares (PRESS) statistic again to ensure our model was superior to a simple linear regression that had undergone the same transformations. In the ANOVA test, our p-value was significantly less than the 0.05, so we needed to reject the null hypothesis that our additional predictor's slopes were equal to zero. Additionally, our multiple linear regression model had a significantly better PRESS statistic. Both of these results confirmed that our model was indeed the best predictor of price based on available data. We show the summary of our final multiple linear regression model in Figure 2.

```
Call:
lm(formula = log(price) ~ log(carat) * color + clarity + cut)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81414 -0.10218 -0.00516  0.09783  1.02436

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          9.345690   0.006305 1482.34  < 2e-16 ***
log(carat)           2.005191   0.001765 1136.03  < 2e-16 ***
colorE              -0.076345   0.001986  -38.44  < 2e-16 ***
colorF              -0.107790   0.001942  -55.50  < 2e-16 ***
colorG              -0.155737   0.001912  -81.47  < 2e-16 ***
colorH              -0.239186   0.002025 -118.14  < 2e-16 ***
colorI              -0.357407   0.002118 -168.74  < 2e-16 ***
colorJ              -0.499720   0.002373 -210.61  < 2e-16 ***
clarityIF           -0.247836   0.005797  -42.75  < 2e-16 ***
claritySI1          -0.610574   0.005616 -108.71  < 2e-16 ***
claritySI2          -0.745378   0.005663 -131.63  < 2e-16 ***
clarityVS1          -0.447639   0.005622  -79.63  < 2e-16 ***
clarityVS2          -0.501609   0.005626  -89.16  < 2e-16 ***
clarityVVS1         -0.332013   0.005642  -58.85  < 2e-16 ***
clarityVVS2         -0.401509   0.005640  -71.19  < 2e-16 ***
cutGood             -0.300504   0.003107  -96.71  < 2e-16 ***
cutIdeal            -0.083176   0.002822  -29.48  < 2e-16 ***
cutVery Good        -0.253936   0.002859  -88.81  < 2e-16 ***
log(carat):colorE   -0.025917   0.002453  -10.57  < 2e-16 ***
log(carat):colorF   -0.025317   0.002452  -10.33  < 2e-16 ***
log(carat):colorG   -0.009153   0.002480   -3.69 0.000224 ***
log(carat):colorH   -0.055426   0.002651  -20.91  < 2e-16 ***
log(carat):colorI   -0.092681   0.002684  -34.53  < 2e-16 ***
log(carat):colorJ   -0.140134   0.003079  -45.51  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1604 on 199574 degrees of freedom
Multiple R-squared:  0.9744,    Adjusted R-squared:  0.9744
F-statistic: 3.299e+05 on 23 and 199574 DF,  p-value: < 2.2e-16
```

Figure 2: Summary of Final Regression Equation

Final Regression Equation

After selecting, assessing, and transforming the model as needed, we found that the best model was a natural log transformation of price regressed against a natural log transformation of carat that has an interaction effect with color and additive effects with clarity and cut. We represent the final regression equation as:

$$\log(y\hat{}) = B_0 + B\hat{}_1*\log(x_1) + B\hat{}_2*I_1 + B\hat{}_3*I_2 + B\hat{}_4*I_3 + B\hat{}_5*I_4 + B\hat{}_6*I_5 + B\hat{}_7*I_6 + B\hat{}_8*J_1 + B\hat{}_9*J_2 + B\hat{}_{10}*J_3 + B\hat{}_{11}*J_4 + B\hat{}_{12}*J_5 + B\hat{}_{13}*J_6 + B\hat{}_{14}*J_7 + B\hat{}_{15}*K_1 + B\hat{}_{16}*K_2 + B\hat{}_{17}*K_3 + B\hat{}_{18}*\log(x_1)*I_1 + B\hat{}_{19}*\log(x_1)*I_2 + B\hat{}_{20}*\log(x_1)*I_3 + B\hat{}_{21}*\log(x_1)*I_4 + B\hat{}_{22}*\log(x_1)*I_5 + B\hat{}_{23}*\log(x_1)*I_6$$

$$\log(y\hat{}) = 9.3457 + 2.0052*\log(x1) - 0.0763*I_1 - 0.1078*I_2 - 0.1557*I_3 - 0.2392*I_4 - 0.3574*I_5 - 0.4997*I_6 - 0.2478*J_1 - 0.6106*J_2 - 0.7454*J_3 + - 0.4476*J_4 - 0.5016*J_5 - 0.3320*J_6 - 0.4015*J_7 - 0.3005*K_1 - 0.0832*K_2 - 0.2539*K_3 - 0.0259*\log(x_1)*I_1 - 0.0253*\log(x_1)*I_2 - 0.0092*\log(x_1)*I_3 - 0.0554*\log(x_1)*I_4 - 0.0927*\log(x_1)*I_5 - 0.1401*\log(x_1)*I_6$$

We use the indicator variables $I_1 \ldots I_6$ to account for the seven classes associated with color as indicated in Table 3. Similarly, $J_1 \ldots J_7$ represent the eight classes of clarity and are displayed in Table 4. Lastly, K_1, K_2, and K_3 are associated with the four classes of cut as seen in Table 5.

	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	$\mathbf{I_4}$	$\mathbf{I_5}$	$\mathbf{I_6}$
D	0	0	0	0	0	0
E	1	0	0	0	0	0
F	0	1	0	0	0	0
G	0	0	1	0	0	0
H	0	0	0	1	0	0
I	0	0	0	0	1	0
J	0	0	0	0	0	1

Table 3: Indicator Variables for Color

	J_1	J_2	J_3	J_4	J_5	J_6	J_7
FL	0	0	0	0	0	0	0
IF	1	0	0	0	0	0	0
ST1	0	1	0	0	0	0	0
ST2	0	0	1	0	0	0	0
VS1	0	0	0	1	0	0	0
VS2	0	0	0	0	1	0	0
VVS1	0	0	0	0	0	1	0
VVS2	0	0	0	0	0	0	1

Table 4: Indicator Variables for Clarity

	K_1	K_2	K_3
Astor Ideal	0	0	0
Good	1	0	0
Ideal	0	1	0
Very Good	0	0	1

Table 5: Indicator Variables for Cut

Model Prediction

Based on our final model, we are able to input the characteristics of the ten diamonds under consideration by Cavalier Pawn Shop and generate their estimated value, or price. It is important to note that our model generates only an estimated price for the diamond based on the characteristics given. This is the price at which we believe Cavalier Pawn Shop will be able to resell the diamond given sufficient time to do so. We can see the recommendations in Table 6.

There are instances within the dataset where diamonds with the exact same carat, cut, clarity, and color have different prices. For example, diamonds with 1.01 carats, SI2 clarity, color J, and cut of Very Good, range in price from $1,921 to $3,993. This variance occurs because the categories of cut, clarity, and color used to value a diamond are broad range categorical variables instead of continuous numerical values like carat. This explains some of the error in our model. Also note that the data was obtained from an online diamond retailer;

however, other factors such as location may influence the price of diamonds being sold at brick and mortar stores all over the world. Additionally, the data does not account for seasonal changes in diamond prices, but major holidays may have an influence on the demand and thus the price of diamonds. (International Gem Society, 2020)

We emphasize that our valuation should be used with caution because of the effect of other predictors on price described above that are not accounted for in our model. For instance, our dataset contains 30 diamonds with the exact same characteristics as Diamond 1, but at various prices. The prices range from $882 to $1,750 with a mean of $1,214. Our estimated value of such a diamond is $1,450 based on the relationships price has with each of the predictor variables for ranges that extend beyond the narrow scope of any singular diamond. However, we do note that the mean of $1,214 is within our prediction interval for Diamond 1 and our estimated price is within the range of prices of similar diamonds from the dataset. We hope that the owners at Cavalier Pawn Shop will use our estimated prices as a tool to help them make sound business decisions but that they recognize the influence of other factors and rely on context and personal judgement to guide them as well.

	Carat	**Color**	**Clarity**	**Cut**	**Recommended Price**
Diamond 1	0.53	F	VS1	Very Good	$1,450.04
Diamond 2	0.32	H	VVS2	Ideal	$601.96
Diamond 3	1.75	G	IF	Good	$17,302.87
Diamond 4	0.95	H	SI1	Ideal	$4,075.48
Diamond 5	1.34	D	VS2	Ideal	$11,473.14
Diamond 6	0.24	G	SI1	Good	$228.22
Diamond 7	1.14	E	SI2	Very Good	$5,061.27
Diamond 8	1.89	J	VS1	Ideal	$13,391.93
Diamond 9	0.76	I	FL	Very Good	$3,675.65
Diamond 10	1.52	F	VVS1	Ideal	$15,548.27

Table 6: Price Recommendations

References

International Gem Society (2020). “The Fifth C: What Determines Diamond Cost?” Retrieved from https://www.gemsociety.org/article/what-determines-diamond-cost/