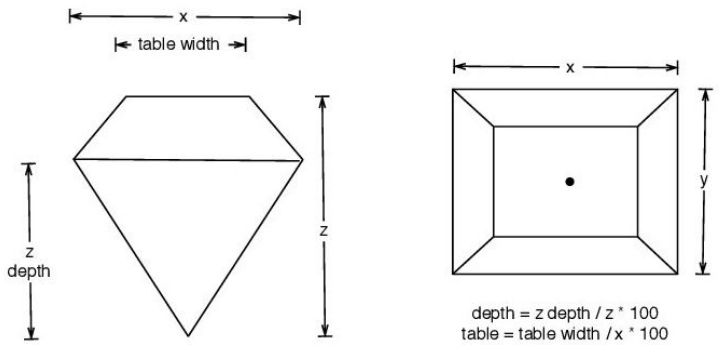
Diamond Analysis

To see the code behind this analysis please click on this link.

Introduction

Diamond Prices can be impacted by various things, such as inflation, carat, etc. I will be examining a data set with 53,943 diamonds which includes the variables, carat, cut, color, clarity, x, y, z, depth, table, and price. The images below give a visual explanation of each variable.



The purpose of this analysis is to answer the following questions:

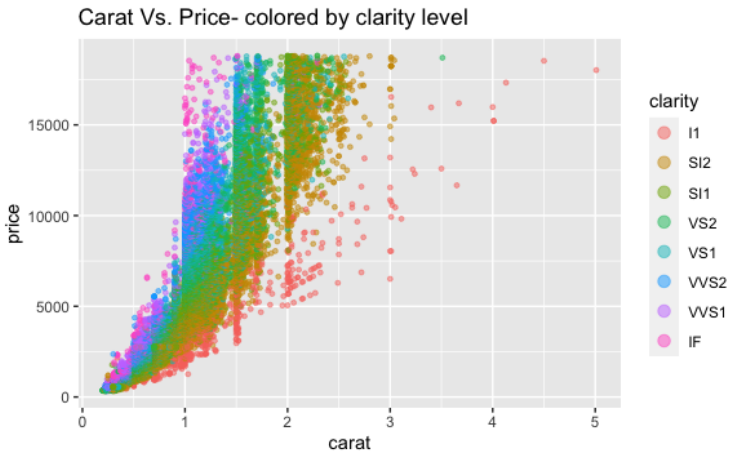
1. Are all the variables in this data set important in explaining diamond prices? If, not which variables are important?

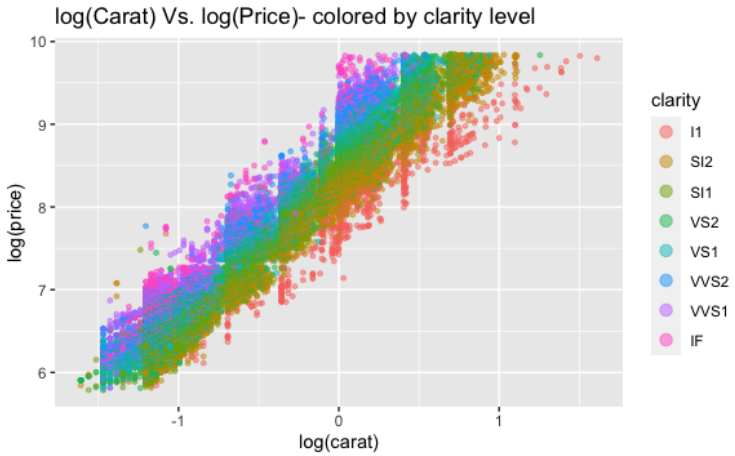
2. Is there an interaction between diamond clarity and carat weight?

3. Does the fitted model explain diamond prices well?

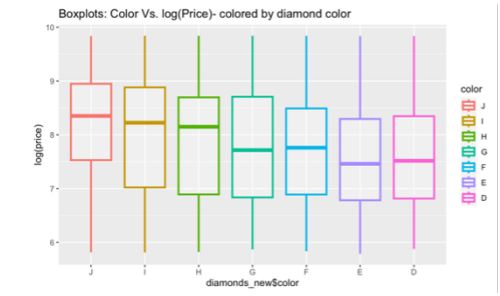
4. What are the differences in prediction intervals and confidence intervals, for diamonds of different carat weights, with the best cut, least amount of color, and highest amount to clarity?

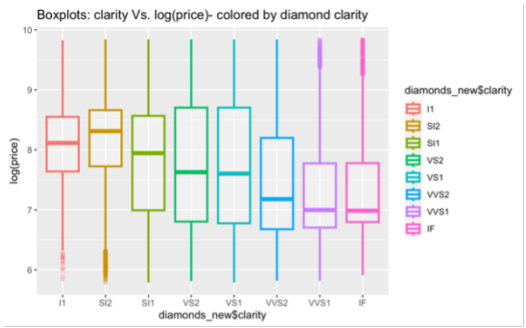
The scatterplot of the price against carat doesn't appear to follow a linear pattern. This would require a transformation to make the data follow a linear trend. I will transform my data by examining both price and carat on a log scale. After transforming the data, the scatterplot follows a more linear pattern. Based on the scatter plot here is not an interaction between diamond clarity and carat weight. If there was an interaction, I would be able to assign different slopes for each clarity level.

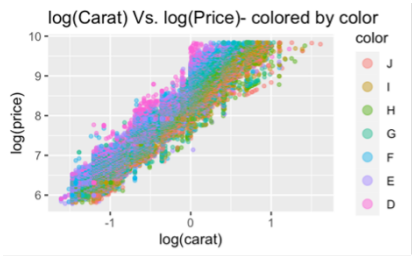


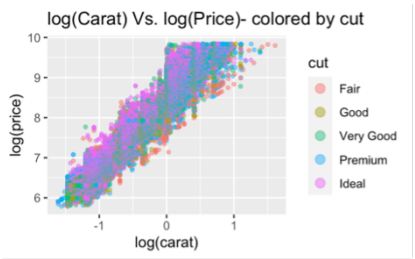


Using boxplots I checked if there is a constant effect on log of price due to clarity, color, or cut. On the color versus log of price boxplot the median price seems to decrease as the diamonds move towards D which is a practically colorless diamond level. The same trend happens in the clarity versus log of price boxplots. The cut versus log of price graph shows that the boxplot as have different median prices depending on cut. It may be possible to remove an indicator for cut, but due to the fair cut begin significantly different from the others and different cuts being an important factor to many diamond experts I included cut as an idicator variable. Since none of the other boxplots for clarity and color seem to have a constant effect, I will include indicator variables for those as well.This choice to include indicator variable is further confirmed by the scatterplots for carat and price on a log scale colored by cut, clarity or color. Each of these scatterplots seems to have at least two distinct clusters for each claraity, cut or color.

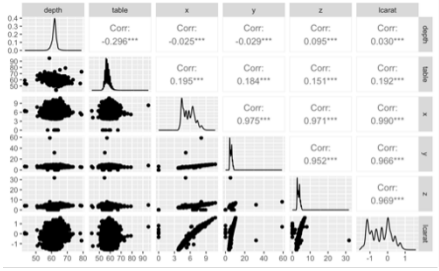








I constructed a scatterplot matrix of my numerical variables to check for collinearity. Based on the high R-squared values I suspect there may be collinearity in variables x, y, z, and log of carat. I used VIF, eigenvalues, and eigenvectors to find where the collinearity lied. The log of carat, x, y, and z all had VIF values larger then 10, this means there is collinearity. The condition index for K6 had a value greater than the threshold of 15. I went to the 6th eigen vector and saw that log of carat, x, and z have the issues with collinearity. Since x, y, and z are all related I removed them from the model and reran the VIF. There were no VIF values that exceeded the threshold.



	GVIF
log(carat)	64.070
depth	1.903
table	1.803
x	94.368
y	20.625
z	23.617
as.factor(color)	1.167
as.factor(clarity)	1.344
as.factor(cut)	1.964

Eigen Vectors

```
$vectors
             [,1]         [,2]         [,3]         [,4]         [,5]         [,6]
[1,]  0.0003570259  0.734537762 -0.67039173 -0.086349778  0.03392066 -0.049215669
[2,] -0.1212775965 -0.669313364 -0.73269072 -0.008767564 -0.01856533 -0.007357665
[3,] -0.4997977137  0.009034135  0.07294808  0.089918747  0.33157668 -0.791688534
[4,] -0.4940856460  0.010499806  0.08937037 -0.790307718 -0.33972268  0.088228552
[5,] -0.4928796321  0.102206884 -0.00429563  0.589138068 -0.62080564  0.118834603
[6,] -0.4984404731  0.042846099  0.02030524  0.112745342  0.62269614  0.590632718
```

Eigen Values

```
1.000000  1.755377  2.406501  9.897937 11.524265 23.441654
```

The equation for our current model is.

$$Y = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\gamma_1G_1+\gamma_2G_2+\gamma_3G_3+\gamma_4G_4+\gamma_5G_5+\gamma_6G_6+\delta_1H_1+\delta_2H_2+\delta_3H_3+\delta_4H_4+\delta_5H_5+\delta_6$$

Y: Actual log(diamond price)

β_0: When all other values are at zero the log(price) of a fair cut, J color, I1 clarity diamond is β_0 after adjusting for log(carat).

β_1: On average, holding all other variables constant there is a β_1 change in log(price) for every 1 unit increase in log(carat).

β_2: On average, holding all other variables constant there is a β_2 change in log(price) for every 1 unit increase in depth.

β_3: On average, holding all other variables constant there is a β_3 change in log(price) for every 1 unit increase in table.

β_4: On average, holding all other variables constant there is a β_4 change in log(price) for every 1 unit increase in x variable.

β_5: On average, holding all other variables constant there is a β_5 change in log(price) for every 1 unit increase in y variable.

β_6: On average, holding all other variables constant there is a β_6 change in log(price) for every 1 unit increase in z variable.

γ_1: Measures the differential log(price) for diamonds in the I color group relative to countries in the J color group after adjusting for log(carat).

γ_2: Same as γ_1 but for color group H.

γ_3: Same as γ_1 but for color group G.

γ_4: Same as γ_1 but for color group F.

γ_5: Same as γ_1 but for color group E.

γ_6: Same as γ_1 but for color group D.

δ_1: Measures the differential log(price) for diamonds in the SI2 clarity group relative to diamonds in the I1 clarity group after adjusting for log(carat).

δ_2: Same as $\delta_1$but for clarity group SI1.

δ_3: Same as δ_1 but for clarity group VS2.

δ_4: Same as δ_1 but for clarity group VS1.

δ_5: Same as δ_1 but for clarity group VVS2.

δ_6: Same as δ_1 but for clarity group VVS1.

δ_7: Same as δ_1 but for clarity group IF.

ζ_1: Measures the differential log(price) for diamonds in the "good" cut group relative to diamonds in the fair cut group after adjusting for log(carat).

ζ_2: Same as ζ_1 but for cut group "very good."

ζ_3: Same as ζ_1 but for cut group "premium."

ζ_4: Same as ζ_1 but for cut group "fair."

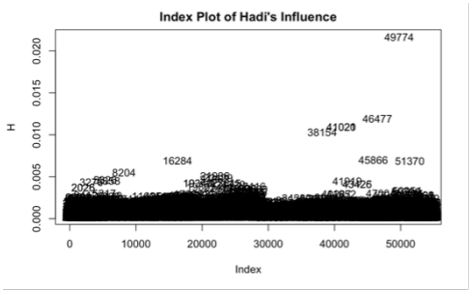
$$G_{ti} = \begin{cases} 1 & \text{if } t^{\text{th}} \text{ diamond is in } i^{\text{th}} \text{ color group} \\ 0 & \text{otherwise} \end{cases}$$

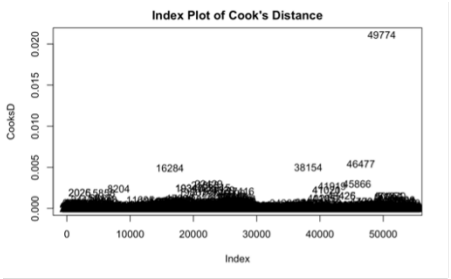
$$H_{ti} = \begin{cases} 1 & \text{if } t^{\text{th}} \text{ diamond is in } i^{\text{th}} \text{ clarity group} \\ 0 & \text{otherwise} \end{cases}$$

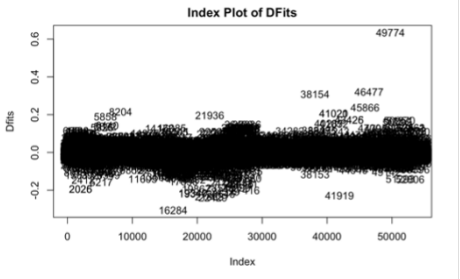
$$J_{ti} = \begin{cases} 1 & \text{if } t^{\text{th}} \text{ diamond is in } i^{\text{th}} \text{ cut group} \\ 0 & \text{otherwise} \end{cases}$$

$$\varepsilon \sim N(0, \sigma^2)$$

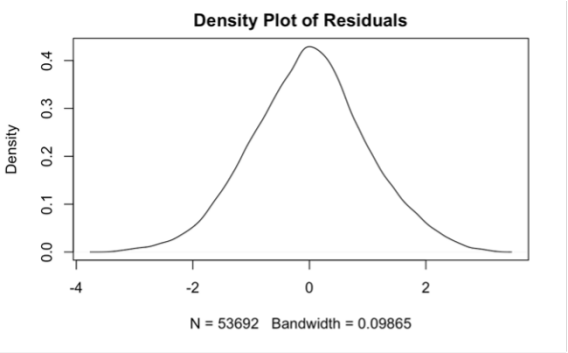
The model summary showed that depth and table were insignificant variables. After removing those variables, I found several influential points through the Dfits, Cook's D and Hadi's influence plots. The influential points were 49774, 46477, 41020, 38154, and 16284. The most noticeable values were row 49774. When I examined the data point, there didn't seem to be anything that was significantly different form the other data points around it.

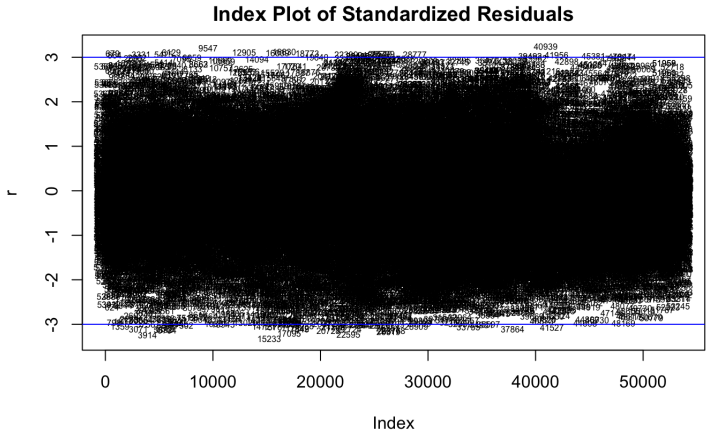


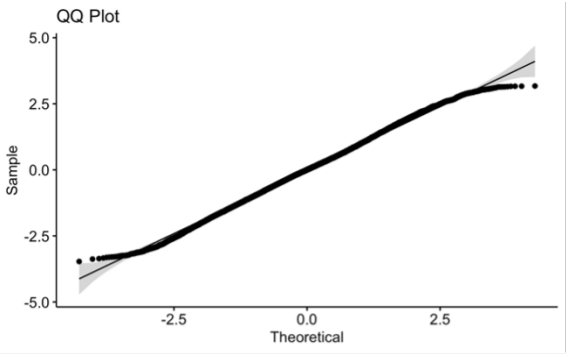




The data also showed dozens of outliers. I removed any standardized residuals that were above 3 or below -3. After remove the outliers I checked the regression assumptions using the residual plot, qq-plot, and density plot. Based on the residual plot the equal variances assumption appears to be met. The density plot is approximately normal, and the qq-plot is approximately with a few points that go briefly outside the error bands.







The regression analysis found that the variables y, z, and table are not significant. I removed those form the model. The new fitted model for the data after the non-significant variables and outliers are removed is:

$$\widehat{\text{Log(price)}} = 7.359521 + 1.886517\text{Log}\,(\text{carat}) + 0.138941G_1 + 0.262054G_2 + 0.351932G_3 + 0.416919G_4 + 0.458401G_5 + 0.509588G_6 +$$

The R^2after transforming back to the original data is 0.9644476. Almost 97% of variability in the log of diamond prices can be explained by the five variables in our model. This means our model is great at explaining variation in the log of diamond prices.

Now that the analysis is complete, the research questions have been answered below.

1. Are all the variables in this data set important in explaining diamond prices? If, no which variables are important?
    a. Based on the summary of the regression model we see that cut, color, clarity, and log(carat) have a significant impact on predicting diamond log(prices). Using confidence intervals, I quantified the relationship between the price the remaining variables. On average for every unit increase in carat weight we are 95% confident there is an increase in diamond price between $6.58 and $6.61. The confidence intervals for clarity show that as the clarity improves the impact on diamond price gets larger. For diamonds with the lowest level of clarity, SI2, the 95% confidence interval shows that on average for every unit change in color diamond prices will increases between $1.49 and $1.52. The 95% confidence interval for the highest level of clarity, IF, is between $2.92 and $2.98. The confidence intervals for color and cut increase as cut improves and color improves. For example, the 95% confidence interval for the worst cut, good cut, is $1.08 to $1.10. The 95% confidence interval for the best cut, ideal cut, is $1.17 to $1.19.

	2.5 %	97.5 %
(Intercept)	1552.838755	1589.544578
lcarat	6.582202	6.610530
as.factor(color)I	1.142287	1.155866
as.factor(color)H	1.292372	1.306861
as.factor(color)G	1.414122	1.429543
as.factor(color)F	1.508920	1.525685
as.factor(color)E	1.572780	1.590354
as.factor(color)D	1.654918	1.674349
as.factor(clarity)SI2	1.485771	1.516026
as.factor(clarity)SI1	1.751699	1.787158
as.factor(clarity)VS2	2.033766	2.075151
as.factor(clarity)VS1	2.182434	2.227488
as.factor(clarity)VVS2	2.495340	2.548370
as.factor(clarity)VVS1	2.679777	2.738296
as.factor(clarity)IF	2.915506	2.984396
as.factor(cut)Good	1.083862	1.100036
as.factor(cut)Very Good	1.125545	1.141171
as.factor(cut)Premium	1.152412	1.168237
as.factor(cut)Ideal	1.177720	1.193754

2. Is there an interaction between diamond clarity and carat weight?

a. Based on the initial scatterplot, there does not appear to be an interaction between diamond clarity and carat weight. If there was an interaction between the variables, we would be able to assign several different slopes for each clarity level. One slope seems to be able to fit the data. Therefore, there is not an interaction.

3. Does the fitted model explain variability in diamond prices well?

a. The fitted model has an r-squared value of about .97. This means that our model explains about 97% of variability in the log of diamond prices. The residual standard error for any diamond price is about $723.84. This residual standard error is quite large.

4. What are the differences in prediction intervals and confidence intervals, for diamonds of different carat weights, with the best cut, least amount of color, and highest amount to clarity?

a. It appears that for prediction intervals as the carat weight increases so does the price, even when keeping cut, color, and clarity constant. i. We are 95% confident that a future .25 carat diamond with an ideal cut, D grade color, and IF clarity will be priced between $519.60 and $861.56. ii. We are 95% confident that a future .50 carat diamond with an ideal cut, D grade color, and IF clarity will be priced between $1921.17 and $3185.56. iii. We are 95% confident that a future .75 carat diamond with an ideal cut, D grade color, and IF clarity will be priced between $4128.22 and $6845.22. iv. We are 95% confident that a future 1 carat diamond with an ideal cut, D grade color, and IF clarity will be priced between $7103.30 and $11778.46.

b. The confidence intervals show the same trend as the prediction intervals, as the carat weight increases so does the price. i. We are 95% confident that the average price of .25 carat diamonds with an ideal cut, D grade color, and IF clarity will be between $664.46 and $673.74. ii. We are 95% confident that the average price of .50 carat diamonds with an ideal cut, D grade color, and IF clarity will be between $2456.71 and $2491.15. iii. We are 95% confident that the average price of .75 carat diamonds with an ideal cut, D grade color, and IF clarity will be between $5278.16 and $5353.87. iv. We are 95% confident that the average price of 1 carat diamonds with an ideal cut, D grade color, and IF clarity will be between $9080.42 and $9213.88.

c. The prediction interval is a lot wider than the confidence interval. Since prediction intervals attempt to create an interval for a specific new observation, there's more uncertainty in our estimate and thus prediction intervals are always wider than confidence intervals.