A Diamond Is Forever: The Characteristics That Determines the Sales Price of Diamonds

Zuojun Gong
Data Analysis Exam 2

Abstract

Diamonds are the one of the most precious type of gemstones and they are usually very expensive. What qualities determines the price of the diamonds, are heavier diamonds more expensive? Does the appearance and cut of the diamond matter? If we can find out what characteristics makes diamonds more expensive, we can determine how to process the diamonds for a higher sales price. We analyzed a sample of 336 round cut diamonds by using a multivariate linear regression model under the normal error assumption to predict the sales price of the diamonds. From our regression model, we conclude weight is not the only factor that makes diamonds more expensive – its color, clarity and how it shapes matters. Diamonds with larger width is associated with a higher sales price than the diamonds with a smaller width and larger length of the same weight. Diamonds with better coloring and clarity are likely to be more expensive. We concluded that the depth of the diamond, cutting grade of the diamond, width of top part of diamond relative to widest point, and depth of diamond from the widest point relative to depth are not useful predicators. Since our conclusion is limited to the diamonds of round cut type, we can look into diamonds with different cutting styles and maybe more characteristics of the diamonds to reach a well-rounded conclusion.

Introduction

Diamond, a symbol of eternity, a precious gemstone, and a beautiful creation from nature. Since they are one of the most expensive jewelry in the market, it is important to understand what determines the price of the diamonds, whether it's the size, shape or color. In the following report we have given the information of a sample of round-cut diamonds, including several physical characteristics as well as quality measurements. Specifically, we believe that there is a multivariate linear regression model between the sales price of the diamonds and the predicator variables. We want to test the hypothesis that the weight of the diamond is positively correlated to the sales price and the size measurement pf importance is of the width. It is also believed that there is a relationship between body size and sales price depends on clarity, particularly that if a diamond has low clarity it needs to have a large body to have a high sales price, and a diamond with high clarity are expensive regardless of body size.

Exploratory Data Analysis

We have information of 336 round cut diamonds, and their respective sales price in dollars, weight in carat, length, width and depth in millimeters, width of top part of diamond relative to widest point, and depth of diamond from the widest point relative to depth, both in percentages. In Table 1 and Figure 1, we listed the characteristics and showed histograms of the variables. The distribution of sales price of our diamonds is right skewed, and it ranges from $361.0 to $18800.0 dollars with a median of $3102.0 and mean of $4695.0. The weight of the diamonds in our dataset is right skewed and ranges from 0.2300 to 2.5200 carats, and the median is 0.8847 and the mean is 0.7700 carats. The length of the diamonds ranges from 3.930mm to 8.780mm, the width ranges from 3.960mm to 8.700mm, and the height of the diamonds ranges from 2.430mm to 5.350mm. The width of top part relative to widest point has a right skew ranges from 53% to 68%, and the depth from the widest point relative total depth is normally distributed ranges from 56.90% to 68.70%.

	Mean	Median	1st Quantile	3rd Quantile	SD
Price (in $)	4695.0	3102.0	980.2	6658.0	4674.414
Weight (in Carat)	0.8847	0.7700	0.4000	1.1600	0.528088
Length (in mm)	5.911	5.890	4.748	6.712	1.197635
Width (in mm)	5.911	5.920	4.738	6.730	1.188749
Depth / Height (in mm)	3.657	3.620	2.928	4.175	0.741953
Top (in %)	57.52	57.00	56.00	59.00	2.301697
Body (in %)	61.89	62.00	61.27	62.60	1.319192

Table 1 Summary table for Price, Carat, Length, Width, Height, Top and Body in respective units.

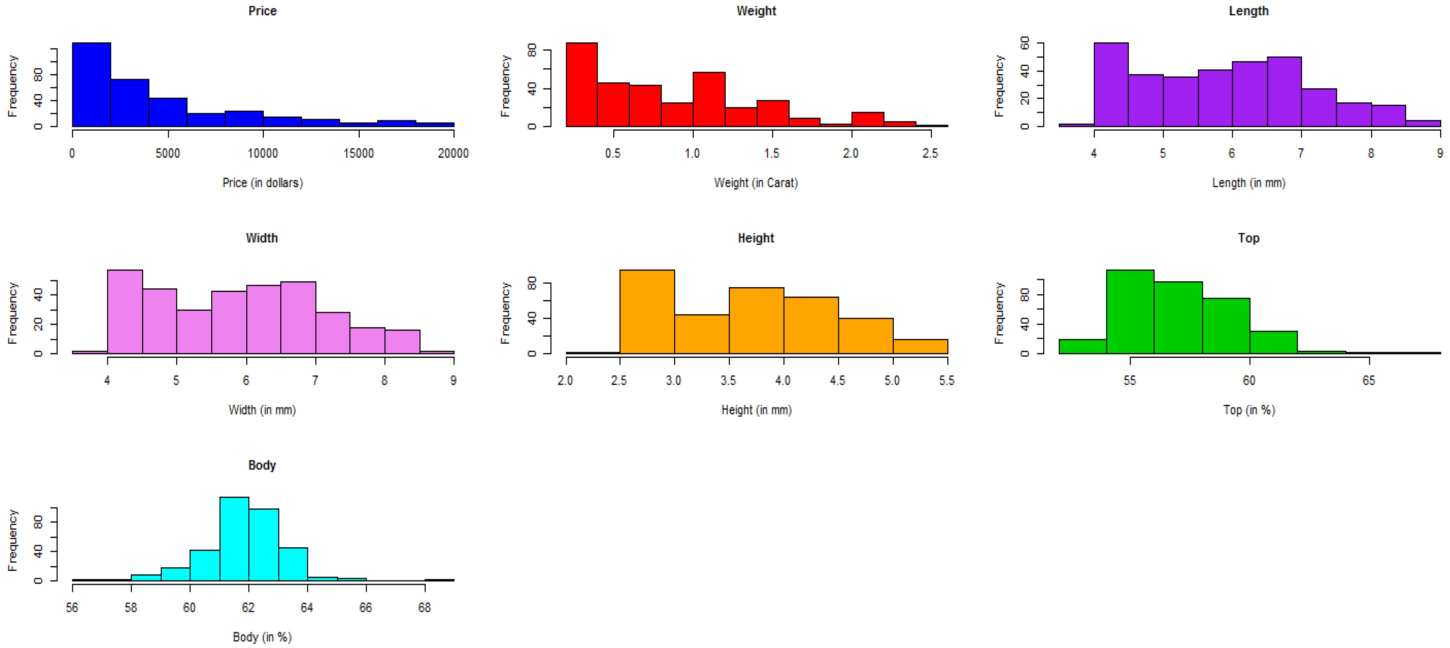

Figure 1: Histogram for Price, Carat, Length, Width, Height, Top and Body in respective units

We observe from Table 2 that the Ideal cutting quality is about 37.5% of all diamonds and premium and very good are about 26.2% and 24.7%. We have most of the diamonds in the middle range of graded color between H-F (55.1%). Most of the diamonds have clarity level between SI1-VS2 (80.3%) and about 35 of VVS2 the top category (10.4%).

Graded Quality of the diamond Cut		**Graded color of the diamond**		**Graded measurement of diamond's clarity**	
				I1	4 (1.2%)
		J	26 (7.7%)	SI1	88 (26.2%)
Fair	14 (4.2%)	I	30 (8.9%)	SI2	73 (21.7%)
Good	25 (7.4%)	H	57 (17.0%)	VS1	43 (12.8%)
Very God	83 (24.7%)	G	73 (21.7%)	VS2	66 (19.6%)
Premium	88 (26.2%)	F	55 (16.4%)	VVS1	17 (5.1%)
Ideal	126 (37.5%)	E	57 (16.9%)	VVS2	35 (10.4%)
		D	38 (11.3%)	IF	10 (3.0%)

Table 2 Summary table for the variables

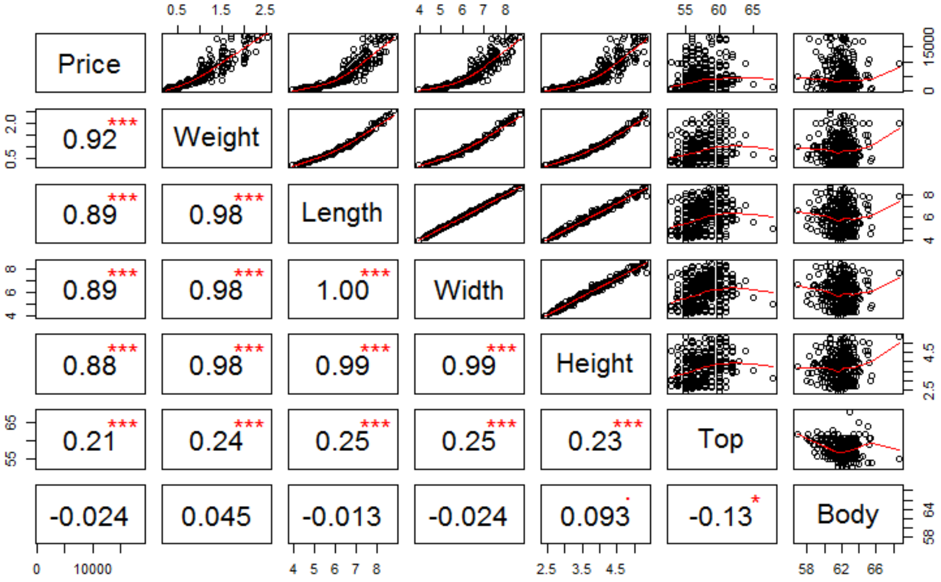

Figure 2: Bivariate Relationship between Price, Carat, Length, Width, Height top and body.

Figure 2 shows us the biviarate scatterplots and corrleation between price, weight, length, width, height, width of top part relative to widest point, and depth of the widest point relative to total depth. We observe strong corrleation of posotive trend between price and weight (0.92), length (0.89), width (0.89), and height (0.88). These peridator variables may have non-linear realtionships with the response varaible. In addition, width of top part relative to widest point is moderately positively corrleated with price (0.21).

There are evidence of muticollinearity between several predicator variables. Length, Width and height are strongly positively correlated with the weight of the diamond (all 0.98). Also, we observe strong correlation between length, width and height. Top is weakly correlated with Length (0.25), Width (0.25), and Height (0.23), and top is also weakly correlated with the weight of diamond (0.24). Also, Top and Body are weakly negatively correlated.

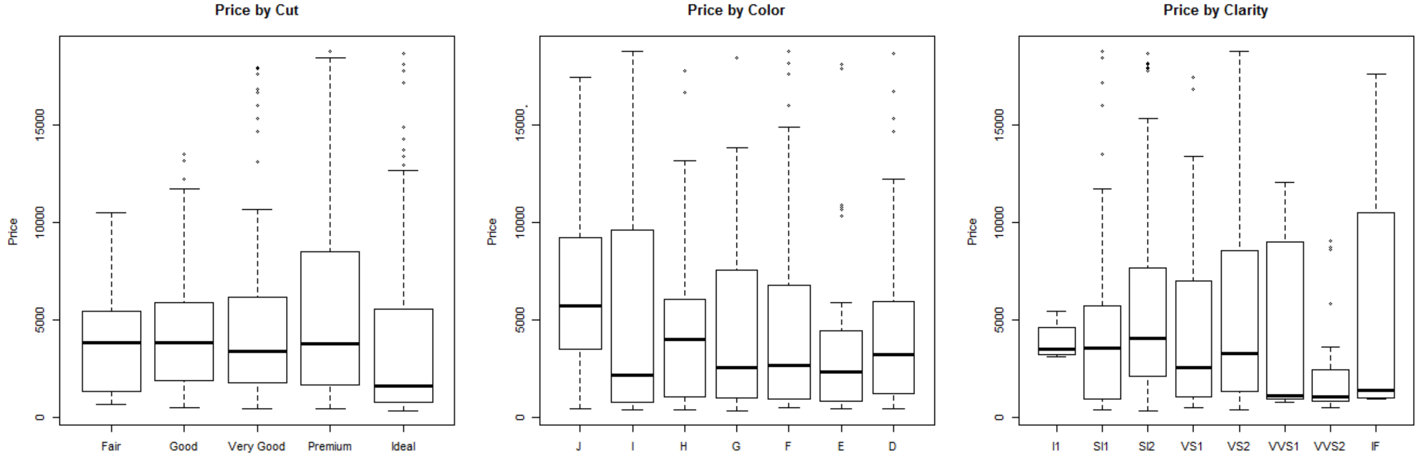

Figure 3: Conditional Distribution box-plot between Price and Cut, Color and Clarity.

Figure 3 provides us the conditional distribution between diamond's price and the graded quality of diamond's cut, graded measurement of clarity, and graded color of the diamond. From observation, all three variables does not seem to have linear relationship with sales price. The Ideal graded cut diamond seems to have the lowest median sales price across all other graded cuts, lowest graded of color has the highest median sales price across all other graded color, and the top three highest graded clarity has lowest median sales price across all graded clarity. We do no observe any consistent pattern between sales prices and graded cut, sales price and graded color, and sales price and graded clarity. They do not seem to be good predicators for the price of diamond by themselves, as there maybe potential interactions between these variables and other predicator variables. We will explore and discuss them later on in the report.

Initial Modeling

We first look at the relationship between the categorical variables, graded cut, color and clarity and the sales price of the diamonds. Table 3 shows us the possible way of collapsing the categorical variables and the modeling comparisons. In terms of Cut, if we were to leave it as an ordered categorical variable, then we expect the same amount of changes of sales price across different levels of grades of cut, which has potential evidence. We can also choose to use each grade level of cut as an individual category, using ideal as reference group, and we observe consistency across all levels in terms of decrease in coefficient in lower levels and similar standard errors. However, we can also collapse Cut into two categories, Ideal and non-Ideal, to increase model stability by collapsing the categories with smaller sizes together. Even if the result is not significant, it may change after we adjust our models. In terms of graded color, if we were to leave it as an ordered categorical variable, it is significant and with each level of increase in graded color, and the price of diamond has an unadjusted increase of $328.74. However, some of the lower levels are underrepresented in our ordered categorical variable model. Then, when we use each graded level of color as a categorical variable with level G as reference group, but we noticed that there is a decrease in estimated price in level E with both increase in estimated price in the previous(F) and next level(D). We then explore the option of collapsing graded color into two categorical variables, graded color between level J-G and level F-D, the diamonds with average graded color versus the diamonds with good graded color, using level J-G as reference group. We observe significance in our model while we increased the stability of the model by collapsing the variables. In terms of graded clarity, our ordered categorical variable is significant, but from the inconsistency of changes median estimated sales price that we observed from Figure 3, we decided to explore further options of categorizing the variables. When we use every level of clarity as a categorical variable, we observe significance across all levels but the changes in estimated sales price between different levels can be inconsistent. Therefore, we decided to collapse them into three categories, Clarity level I1-SI2, VS1-VS2, and VVS1-IF, using I1-SI2 as reference group. This allows us to compare the difference between lower, medium, and high grades of clarity and the estimated sales price of diamonds. We have a more stable model due to the collapsed categories and we observe significance in our categories. Therefore, we decide to model Cut (Ideal vs Non-Ideal), Color (J-G vs F-D) and Clarity (I1-SI2, VS1-VS2,VVS1-IF) as collapsed categorical variables described in Table 3.

	Coefficient (St.Err)		**Coefficient (St.Err)**		**Coefficient (St.Err)**
Cut	221.84 (85.81)**	**Fair**	-1113.73 (480.10)*	Ideal	217.85 (199.51)
		Good	-739.76 (361.08)*	Non-Ideal	Reference
		Very Good	-242.94 (225.88)		
		Premium	-99.74 (244.30)		
		Ideal	Reference		
Color	328.74 (46.89)***	**Level** J	-2320.77 (325.98)***	**Level** J-G	Reference
		I	-736.25 (305.25)*	F-D	757.79 (167.55)***
		H	-501.54 (251.65)*		
		G	Reference		
		F	318.46 (246.54)		
		E	-61.09 (248.44)		
		D	428.43 (291.08)		
Clarity	464.00 (45.66)***	**Level** I1	-4131.66 (665.90)***	I1-SI2	Reference
		SI1	Reference	VS1-VS2	1563.48 (174.26)***
		SI2	-707.55 (203.81)***	VVS1-IF	2118.22 (220.19)***
		VS1	1090.81 (239.63)***		
		VS2	1223.62 (206.32)***		
		VVS1	2158.56 (342.33)***		
		VVS2	1169.61 (261.43)***		
		IF	3188.31 (426.72)***		

p-value < 0.001 ***, <0.01 **, <0.05*, <0.1 . All data adjusted for other predicator variables in their original forms.

Table 3: Comparison of modeling the graded levels of Cut, Color, and Clarity.

We then move on to find the potential confounders in our model. We will potentially adjust for the variables that we mentioned in our research hypothesis, such as the weight and the width of the diamods. We may potentially remove graded cut by using t-tests, and we can only do so after our model assumptions are met after diagnostics. Also, Height, Width and Depth apperars to provide us similar information, and we can use partial F-test to determine whether they should stay in our model after we adjust our model to meet the linear regression model normal error assumiton.

In temrs of interactions between the variables, we explored the relationship between cut and color with sales price, the relationship between claraity and body with sales price, and the relationship between weight and Length, Width and Height. From Figure 4 we can observe the interaction plots that we will use for evidence of interaction between the variables. From our graph in the left we have the plot of collapsed category graded cut variable (Idea vs Non-Ideal) and sales price, with the collpased graded color variable (color <= G vs Color > G). We ovserve an potential reinforcment interaction between the two variables, better coloring have a stronger effect in perdicting an lower estimated price. However, this is subject to change after we construct our full model. The relationship between categories of clarity and body appears to have potential interaction. There appears to be an interaction between the all three categories of clarity of diamonds. But since we use clarity <= 3 as a reference group, we will discuss the interaction between 4< clarity < 6, clarity >= 6, and body. In terms the interaction between weight and length, width height, we considered the added residual plots for these variables. Even if there is no obvious evidence of interaction, we will still consider the interactions in our initial model. Therefore, we will include all interactions into our model for diagnostics.

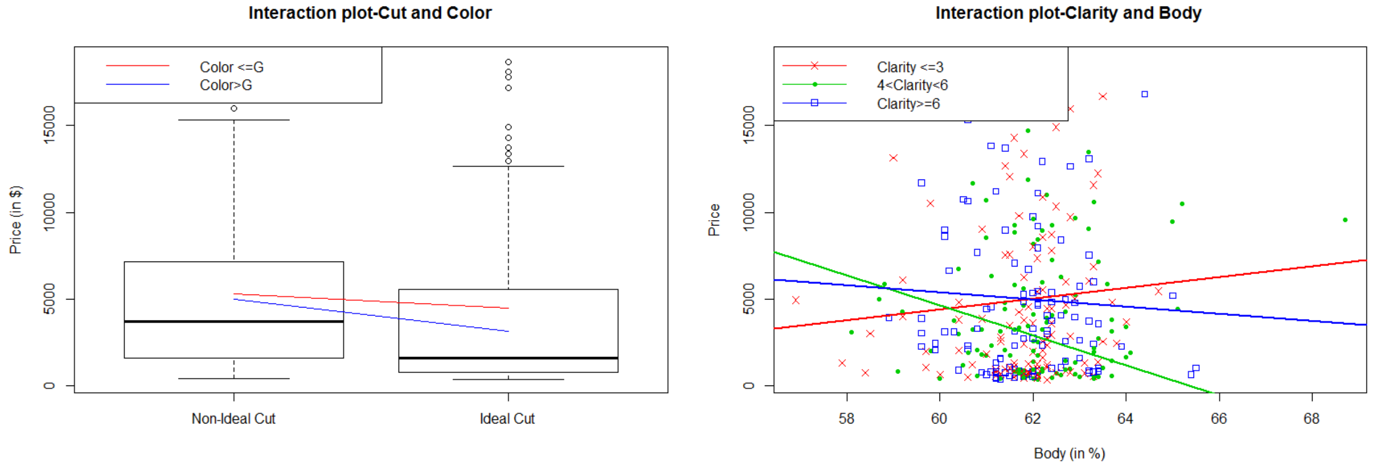

Figure 4: Interaction plots between Price, cut and color, clarity and body.

Our initial model has a response variable sales price of diamond, and the predicator variable of the weight of the diamond, Cut of the diamond (Ideal vs Non-Ideal), Color of the diamond (J-G vs F-D), clarity of the diamond (Level I1-SI2 as reference), length, width, height of the diamond, width of top part of diamond relative to widest point, depth of diamond from the widest point relative to total depth, the interaction between cut and color, the interaction terms between clarity and body, the interaction between weight and length, the interaction between weight and width, and the interaction between weight and height. We then proceed to assess the model for linear regression normal error assumption.

<u>Diagnostics</u>

Before we attempt to examine our model for our model assumptions, we can remove the influential outliers in sales price in diamond since our previous analysis suggests we have potential outliers in our initial model. First, we want to identify the outliers in our response variable, and thus we choose the deleted standardized residual technique, and also with mean leverage to identify the predicator variables' outliers, we decided that diamond #13 is our influential outlier since it is the only potential that overlaps with the two techniques. Thus, we decided to keep all other potential outliers due to lack of evidence.

We first examine if our initial model after outlier removal satisfies multivariate linear regression model under normal error assumption. In Figure 5 we have the residual plots of fitted price against residuals, residuals against index values, residuals against predicator variables, residual against the interaction of cut and color, and the residual against clarity and body. We observe that there most of our residual plots satisfies zero expectation and our residual vs index indicates that the independence assumption is fulfilled. There may be a violation in constant variance from fitted values vs residuals, length, width and height vs residuals. Also, from the pre-transformed normality plot in Figure 6, we observe that the normal residual plot is skewed on both sides, and thus we conclude that there may be a violation to normal error assumption, and thus we will apply transformation. We would also apply log transformation on Price, Length, Width, and Height. After transformation on the given variables, we reexamine our model assumption from the residual plots. From Figure 6 we can observe that our log-fitted values vs residuals satisfies the assumption of zero-expectation and constant variance. Also, residual vs index still suggests model independence, and from our residuals vs perdicator variables we conclude that the zero expectation and constant varaiance are met in most cases (besides body vs residuals). Our normality plot shows a left skew and right truncation. Even if our transformed model is not perfect, we still decided to continune with it.

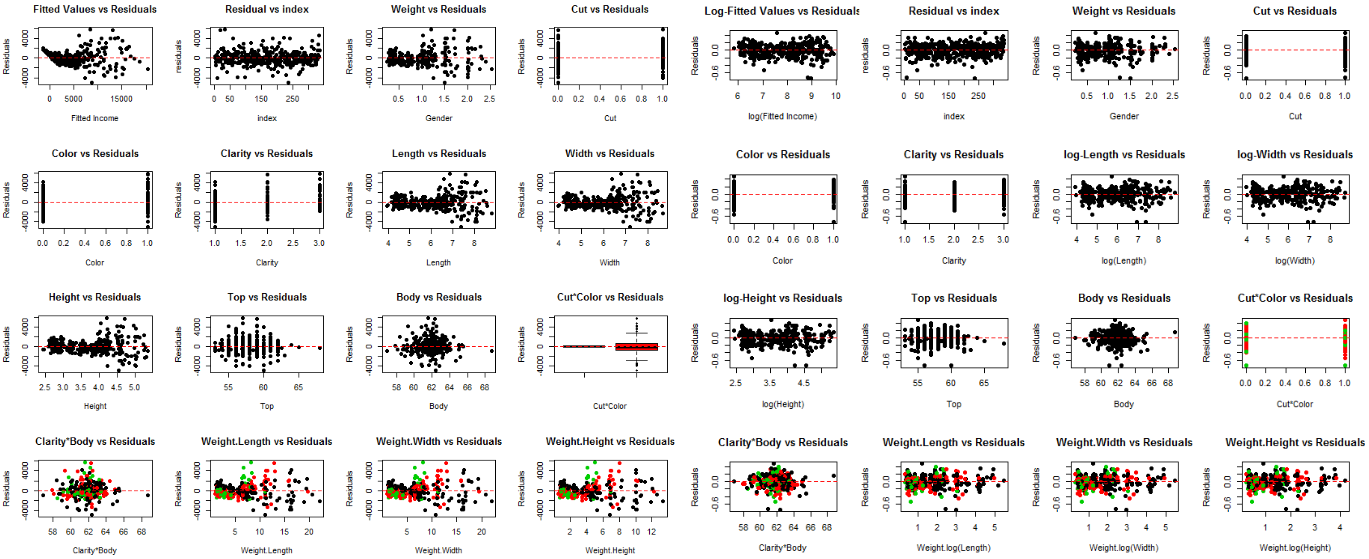

Figure 5: Residual plots for the initial model (left), and the transformed model (right)

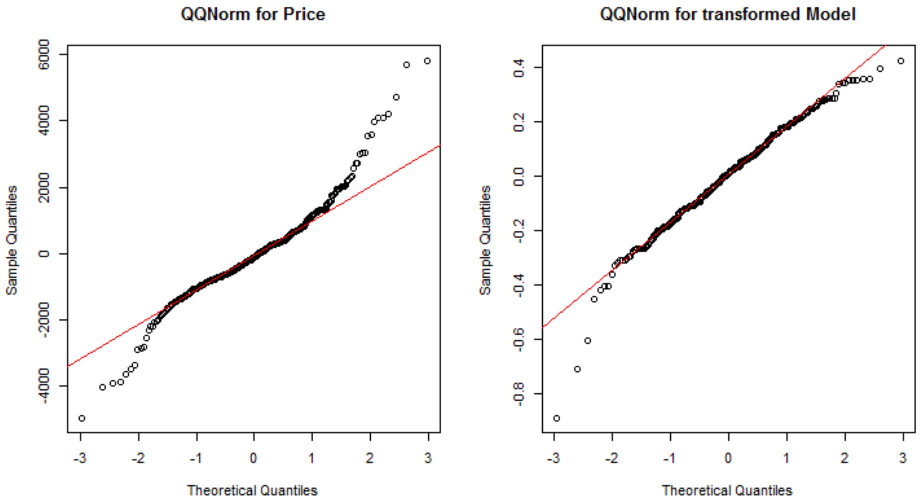

Figure 6: The normality plot comparison

Final model Inference / Results

We then proceed to variable selection by using partial F-tests. We decided to remove graded cut of the diamond (F*=0.0838 vs F(0.90,2,318) = 2.32), Top and Body (F*=0.41662 vs F(0.9,4,318)=1.9626), and Height (F*=1.4372 vs F(0.90,2,318)=2.319). After the removal of the variables, our final model suggests that there is a multivariate linear regression model under normal error assumption, with (F*=1651, df = 8,326) and p-value <0.001. The adjusted R squared for our final model is 0.9753, meaning that in the penalized version of coefficient of determination suggests that 97.53% of the data variation is explained in our model.

	Coeff (Sd.err)	95% CI	P-value
Weight (in carat)	5.3988 (0.98012)	(3.469, 7.3288)	< 0.001
Color			
J-G	Reference		
F-D	0.15327 (0.01929)	(0.1153 , 0.1912)	< 0.001
Clarity			
I1-SI2	Reference		
VS1-VS2	0.26529 (0.02143)	(9.1013 , 16.4146)	< 0.001
VVS1-IF	0.48558 (0.02716)	(0.43215 , 0.53900)	< 0.001
Log (Length)	12.75798 (1.85874)	(9.1013 , 16.4146)	< 0.001
Log (Width)	- 8.93852 (1.88157)	(-12.6400, -5.2370)	< 0.001
Weight * log(Length)	-13.316 (1.87025)	(-0.1679 , 0.0215)	< 0.001
Weight * log(Width)	11.13393 (1.83419)	(-0.1088 , 0.1221)	< 0.001

Table 4: Multivariate Linear Regression results for estimating log (price) for diamonds

Our final model includes weight, color (J-G vs F-D), Clarity (I1-SI2, VS1-VS2, VVS1-IF), log transformed length and Width, interaction between weight and log transformed Length and interaction between weight and log transformed width. Our response variable is in log transformed dollars, and all our interpretation of coefficients of estimate are all adjusted for the model. A one carat increase in weight is associate with to \$5.3988 – 13.316*log (Length) + 11.13393 log (Width) increase in sales price. Diamonds with graded color in category of F-D is estimated to have \$0.15327 more sales price than graded color in category J-G. Diamonds of clarity level VS1 – VS2 are expected to to have a higher sales price of \$0.26529 compared to level I1-SI2, and diamonds with clarity level VVS1-IF are expected to have a higher sales price of estimated \$0.48558. The an increase in one unit of log(length) is associated with an increase of \$12.75798 + -13.316*Weight, and an increase in one unit of log(Width) is associated with an increase of \$ - 8.93852 + 11.13393*Weight. All of the coefficient of estimates have a p-value that is less than 0.001.

Therefore, we conclude that heavier diamonds are not necessarily more expensive unless they have a larger width than length, and graded cut is not a useful predictor for the sales price because it was founded to be not significant. Therefore, different cut grades does not have relationship with the sales price. The depth / Height, as well as the width of top part of diamond relative to widest point, and depth of diamond from the widest point relative to depth of the diamond re also not significant and thus it is not a useful predictor of the sales price. Since we removed the variable of diamond's body, we conclude that the interaction between the diamond's body and clarity is not significant. Therefore, the useful predicator of diamond's sales price are weight, length, width, color, clarity and the interactions between weight, length and width.

Discussion and Results

From our final model we conclude that the useful predicator of a round-cut diamond's sales price are its weight, size and appearance. A diamond with a heavier weight and larger width is associated with a higher sales price, and a diamond with better graded color and graded clarity is associated with a higher sales price. We concluded that the graded cut, body, top, and the diamond's depth are not useful predictors of the sales price. We discovered that bigger and heavier diamonds are not always more expensive. Its width and length are just as important as its weight and size, in the sense that people prefer diamonds with larger width. The cut of the diamonds are not useful predicators of its sales price, possibly because round-cut diamonds are similar to each other compare to other style of cut, and therefore the differences of cutting quality is not necessarily very different from each other. With regards to our original research hypothesis, we conclude that heavier diamond is more expensive, however it is also related to not only the width of the diamond but also the length of the diamond. We also discovered that increasing the quality of cut grade does not always correspond to an increased sales price. Since we removed body and top from our model because they were not significant, we cannot make any specific measure on the interaction between body and clarity, but we can conclude that the interaction is not significant. In terms of improvement, we can look for diamonds across different cutting styles, and it will be interesting to compare the different desirable characteristics in different cutting styles of the diamonds.