Diamond Price Estimation

Sahba Salarian (Student ID:251001238)

December 2018

Introduction

Diamond is a solid form of the element carbon which is atomically arranged in a crystal structure and is considered among the most valuable materials on the earth. This is due to the fact that the most natural diamonds have ages between 1 billion and 3.5 billion years and most are formed at depths of 150-250 or even 800 kilometers in the Earth. The process of formation of diamonds has created special features in these materials, including their extreme hardness in certain orientations, their toughness and high yield opposing breakage or deformation, their high electrical conductivity and their chemical stability. Besides the unique chemical and electro-mechanical characteristics of diamonds, their high value and dazzling appearance have classified them as gems.

The dispersion of white light into spectral colors is the primary gemological characteristic of gem diamonds. Experts in gemology developed methods of grading diamonds based on four characteristics which are now commonly used as the basic descriptors of diamonds. These characteristics include, diamonds mass in carats, diamond's cut considering the proportions, symmetry and polish, diamonds color and how close to white or colorless they are, and the diamond's clarity.

Due to the high value of diamonds as gems and the high price paid to own them, I believe it will be an interesting analysis to be capable of estimating their monetory value. In this report, firstly, I develop a model for diamond prices based on the available gemological characteristics and the diamond dimensions to understand and describe the relationship between the diamond price and the mentioned characteristics. Then, I examine the accuracy of the model to predict the price for other diamonds.

Data Exploration

In this study, I used a dataset which consists the diamond prices and their characteristics. The description of diamond dataset is presented as follow by using str() command in R software. It illustrates that the data frame contains 1000 observations with 10 variables.The diamond price in USD, is selected as the dependent variable and carat, cut, color, clarity, depth (%) and table (%), and x, y and z dimensions in mm, are considered as explanatory variables.

```
## 'data.frame':    1000 obs. of  10 variables:
##  $ carat  : num  0.33 1.52 0.41 1.01 0.51 0.31 0.32 0.77 0.9 0.8 ...
##  $ cut    : int  3 4 5 4 3 5 5 4 5 5 ...
##  $ color  : int  3 7 4 2 7 4 4 3 3 4 ...
##  $ clarity: int  10 4 10 6 5 10 7 5 7 7 ...
##  $ depth  : num  62.9 62.8 61.7 62.5 63.5 61.8 60.7 60.7 62.2 61.1 ...
##  $ table  : num  58 58 56 58 56 54 56 58 58 58 ...
##  $ x      : num  4.38 7.37 4.8 6.44 5.08 4.36 4.46 5.95 6.14 5.96 ...
##  $ y      : num  4.43 7.28 4.77 6.39 5.06 4.38 4.43 5.91 6.17 6.03 ...
##  $ z      : num  2.77 4.6 2.95 4.01 3.22 2.7 2.7 3.6 3.83 3.66 ...
##  $ price  : int  787 8631 1431 4676 1574 891 773 2369 4428 3578 ...
```

Color, cut and clarity are qualitatively classified values which have been quantified for this study. The reference to quantify these explanatory variables are respectively showed in the following tables.

Table 1: Color Grading

Color	D	E	F	G	H	I	J
Value	7	6	5	4	3	2	1

Table 2: Cut grading

Cut	Ideal	Premium	Very Good	Good	Fair
Value	5	4	3	2	1

Table 3: Clarity Grading

Clarity	IF	VVS1	VVS2	VS1	VS2	SI1	SI2	I1	I2	I3
Value	10	9	8	7	6	5	4	3	2	1

Data Visualization

The relation between the diamond price and explanatory variables is investigated by pairs scatter plot matrix. The results are illustrated in Figure 1. This figure indicates that there is a strong relationship between the price and some variables such as the carat, dimensions, clarity and color. However, all these variables have to be checked for statistical significance. In addition, as illustrated, some explanatory variables are strongly correlated (the dimensions and the carat are highly correlated) which is the sign for probable multicollinearity. So, the Variance Inflation factor of the models should be checked to ensure that there are not any multicollinearity effects in the final fitted model.

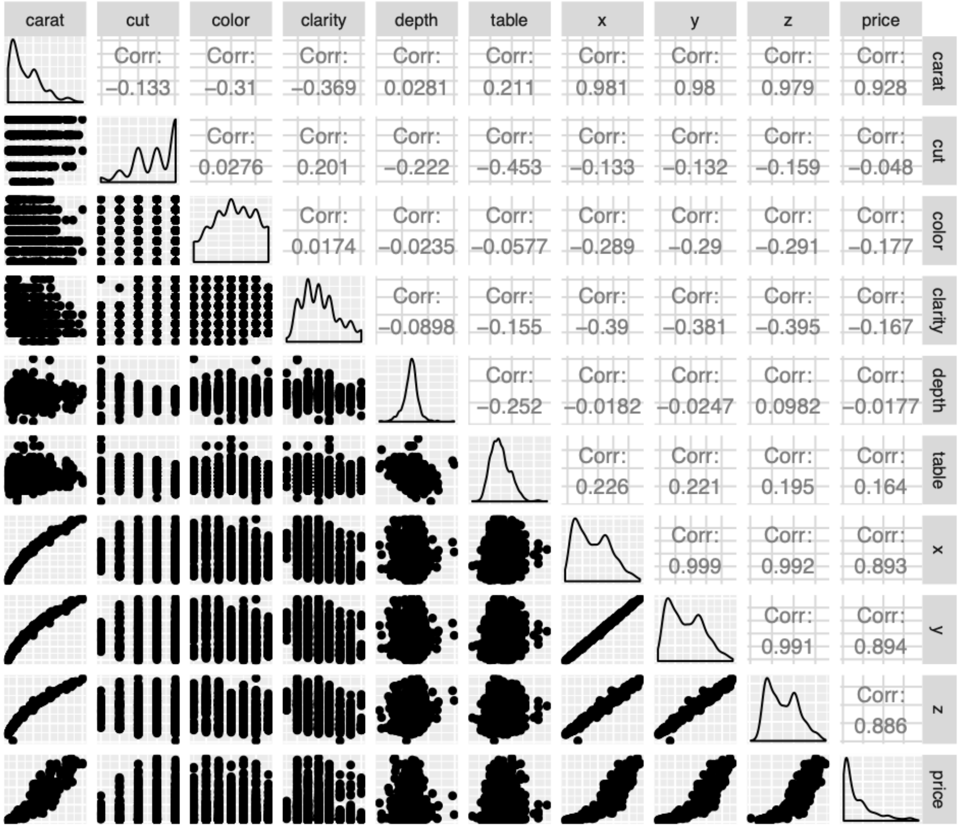
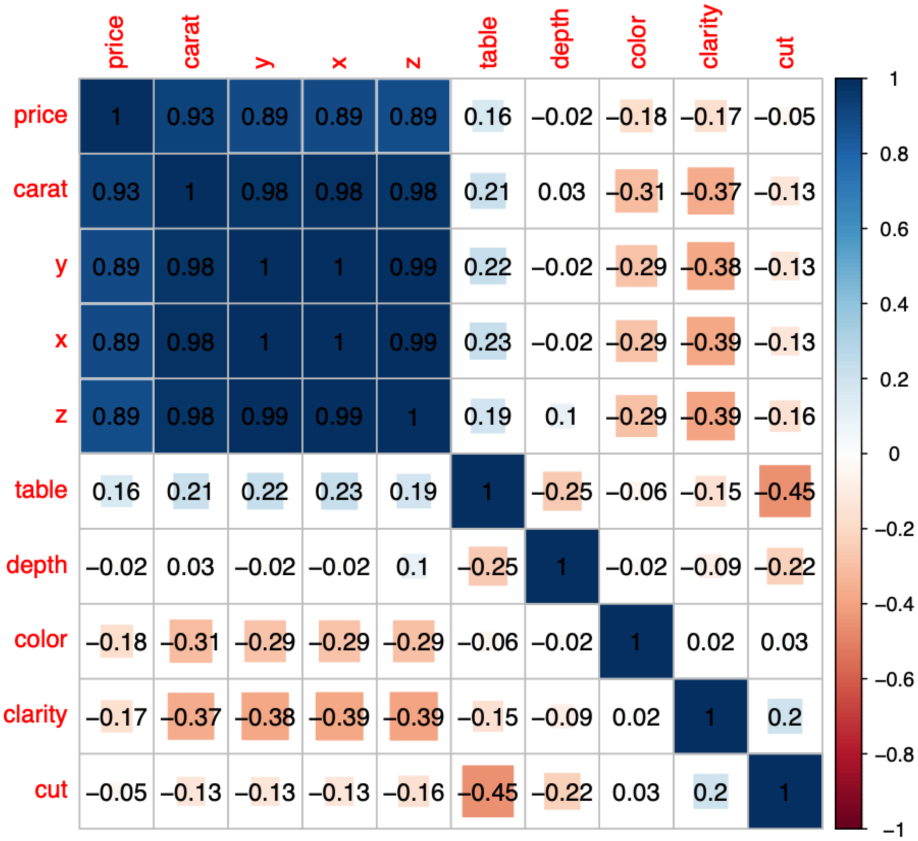


Figure 1: Pairs Scatterplot



	price	carat	y	x	z	table	depth	color	clarity	cut
price	1	0.93	0.89	0.89	0.89	0.16	−0.02	−0.18	−0.17	−0.05
carat	0.93	1	0.98	0.98	0.98	0.21	0.03	−0.31	−0.37	−0.13
y	0.89	0.98	1	1	0.99	0.22	−0.02	−0.29	−0.38	−0.13
x	0.89	0.98	1	1	0.99	0.23	−0.02	−0.29	−0.39	−0.13
z	0.89	0.98	0.99	0.99	1	0.19	0.1	−0.29	−0.39	−0.16
table	0.16	0.21	0.22	0.23	0.19	1	−0.25	−0.06	−0.15	−0.45
depth	−0.02	0.03	−0.02	−0.02	0.1	−0.25	1	−0.02	−0.09	−0.22
color	−0.18	−0.31	−0.29	−0.29	−0.29	−0.06	−0.02	1	0.02	0.03
clarity	−0.17	−0.37	−0.38	−0.39	−0.39	−0.15	−0.09	0.02	1	0.2
cut	−0.05	−0.13	−0.13	−0.13	−0.16	−0.45	−0.22	0.03	0.2	1

Figure 2: The correlation between the diamond price and other explanatory variables.

Correlation Plot

Figure 2 is another representation of the correlation between the variables via the correlation matrix. It can be seen that the positive correlations are displayed in blue and negative correlations in red. The intensity of the color and the size of the inner square are proportional to the correlation coefficients. The figure shows that the diamond price has a strong positive correlation with the diamond's carat and diamond's dimensions. It shows milder positive correlation with table and negative correlations with color and clarity. It is illustrated that depth and cut have the minimum corellation with the diamond price based on the provided data. All the explanatory variables shall be analized regarding the statistical significance. In this report p-value of 0.05 is considered as the acceptable significance level. It is worth mentioning that the considerable correlation between x, y, z dimensions should not be neglected. As mentioned before the issue of multicolliniarity and how to remove them is an important topic which will be discussed later in this report.

Split Data into Train/Test Samples

In order to develop a model and check the reliability of the fitted regression, the available data frame is divided into two categories of Train and Test sets. The selection of data from the main data base for these two groups is done via a random procedure. The Train sample is the data set used for generating the linear model while the predictability of the finalized model will be investigated via the Test sample. From the available 1000 observations, the Train sample contains 70% of the total observations, and the remainig 300 observations are classified in the Test sample for validation analysis.

Full Multiple Linear Regression Model

At this stage a full linear regression model is built over the Train sample, based on all the 9 explanatory variables of carat, color, cut, clarity, depth, table and x, y, z dimensions. The ANOVA test of the full linear regression model is showed in Table 4. The results show that the P-values of the intercept, along with,carat, cut, color, clarity, depth percentage and y, z dimensions are less than 0.01 demonstrating their high statistictal significane in diamond pricing. However, the results shows less significanse for other parameters (table percentage and x diemension). The R^2 and adjusted R^2 of the full-model regression are 0.926 and 0.925, respectively. Meaning that 92.6 percent of the variability has been captured and explained by the full model regression.

Table 4: ANOVA test of the Fullmodel.

	Dependent variable:
	price
carat	12,156.900***
	(461.930)
cut	148.213***
	(44.344)
color	320.636***
	(24.835)
clarity	441.009***
	(26.923)
depth	656.967***
	(156.473)
table	16.152
	(23.805)
x	1,675.167
	(1,112.099)
y	4,797.304***
	(1,168.181)
z	−12,880.370***
	(2,566.411)
Constant	−43,423.710***
	(10,084.940)
Observations	693
R^2	0.926
Adjusted R^2	0.925
Residual Std. Error	1,053.168 (df = 683)
F Statistic	945.682*** (df = 9; 683)
Note:	*p<0.1; **p<0.05; ***p<0.01

Variable Selection

The ANOVA test of the full model and the resulted P-values showed that some variables are more significant in predicting the diamond price. In order to check the redundancy of the explanatory variables and find a balanced set of variables, two distinct methods of backward stagewise regression and best subset regression is used in this report by AIC/BIC approach. The lowest AIC and BIC, obtained from each method are presented in Table 5.

AIC picked 8 variables while BIC picked 7 variables as more effectives. Table 6 shows the parameters chosen for each method. In this report, the best variables chosen via the BIC method is used since the BIC method is more parsimonious than AIC.

	AIC	BIC
Backward-stagewise	11622.99	11664.26
Best subset	-2679.28	-2639.90

Table 5: AIC/BIC Comparisons

	The.selected.variables
Backward-stagewise AIC	carat+ cut + color + clarity + depth+ x + y + z
Backward-stagewise BIC	carat+ cut + color + clarity + depth + y + z
Best subset AIC	carat+ cut + color + clarity + depth + x + y + z
Best subset BIC	carat+ cut + color + clarity + depth + y + z

Table 6: The list of the selected variables by using backward-stagewise and best subset methods.

Best Model Based on BIC Variable Selection

Based on the variable selection section and the BIC method, a new model is generated using linear regression with explanatory variables of carat, cut, color, clarity, depth, y and z. Table 7 summarizes the results of ANOVA test for this model. The results show that the diamond price is significantly affected by all the seven explanatory variables and the intercept and all have P-values less than 0.01. Also, the R^2 and adjusted R^2 of the model are 92.5%. meaning that 92.5 percent of the variability is explained by this model. Interestingly the value of R^2 is similar to previous analysis with all the 9 explanatory variables being considered, meaning that the model is not missing the explanation of any variabilities although the number of explanatory variables are reduced. This also confirms the redundancy of the omitted explanatory variables.

Table 7: ANOVA test of the Bestmodel based on backward stepwise (BIC).

	Dependent variable:
	price
carat	12,226.150***
	(460.323)
cut	130.514***
	(38.049)
color	320.122***
	(24.847)
clarity	434.059***
	(26.583)
depth	481.088***
	(116.000)
z	−10,053.780***
	(1,893.477)
y	4,706.630***
	(1,145.476)
Constant	−31,461.360***
	(7,121.973)
Observations	693
R^2	0.925
Adjusted R^2	0.925
Residual Std. Error	1,053.804 (df = 685)
F Statistic	1,214.006*** (df = 7; 685)
Note:	*p<0.1; **p<0.05; ***p<0.01

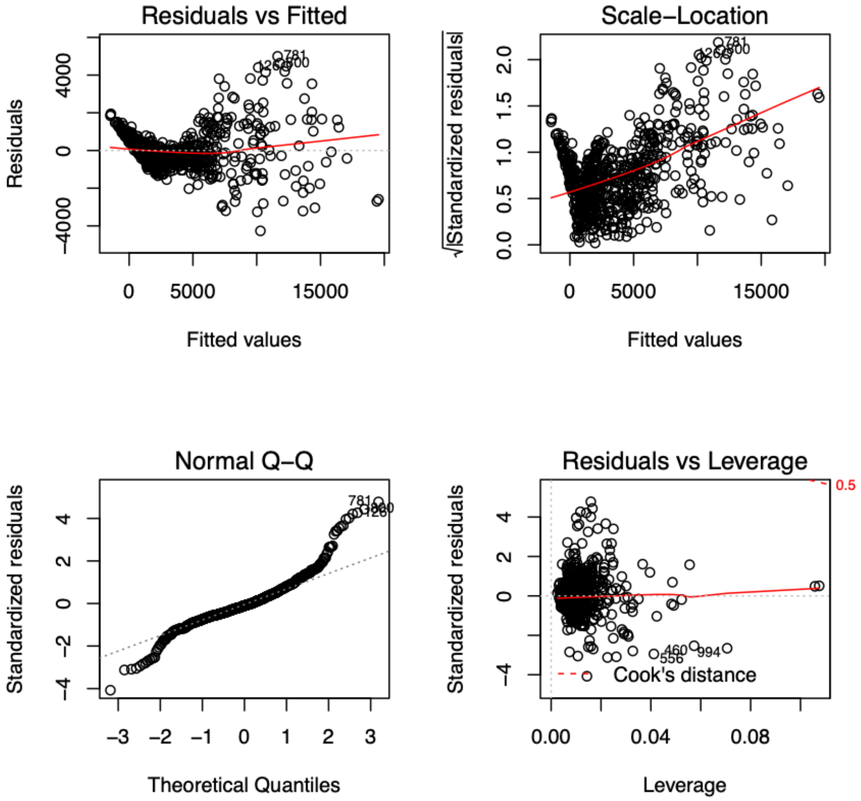


Figure 3: Diagnostic checks for Full Model

Diagnostics Plot of the Best Model

The basic tool for examining the fitted model is checking the residuals. To check the heteroscedasticity, normality, and influential observations for this model, residual diagnostic plots are presented in Figure 3. The first two plots, residual vs. fitted model and scale location plot, check the assumption of linearity and homoscedasticity.Some heteroscedastic behaviour can be observed in residuals due to the existence of very small and large values at the same time in our data set. The scale location diagram which shall have the average of 1, a range between 0.5 to 1.75, which is not exactly what we expect from an ideal standaradized residual behaviour. The presence of influential outliers is checked by the residual vs. Leverage plot. The normality assumption for the residuals is also examined via the Q-Q plot, comparing the residuals to “ideal” normal observations. The Q-Q plot reveals that the observations are not well alligned with the 45-degree line, meaning that the behaviour of the original population is not completely close to the normal distribution. The Box-cox transformation of the response variable is built to check if there is a room to improve the normality.

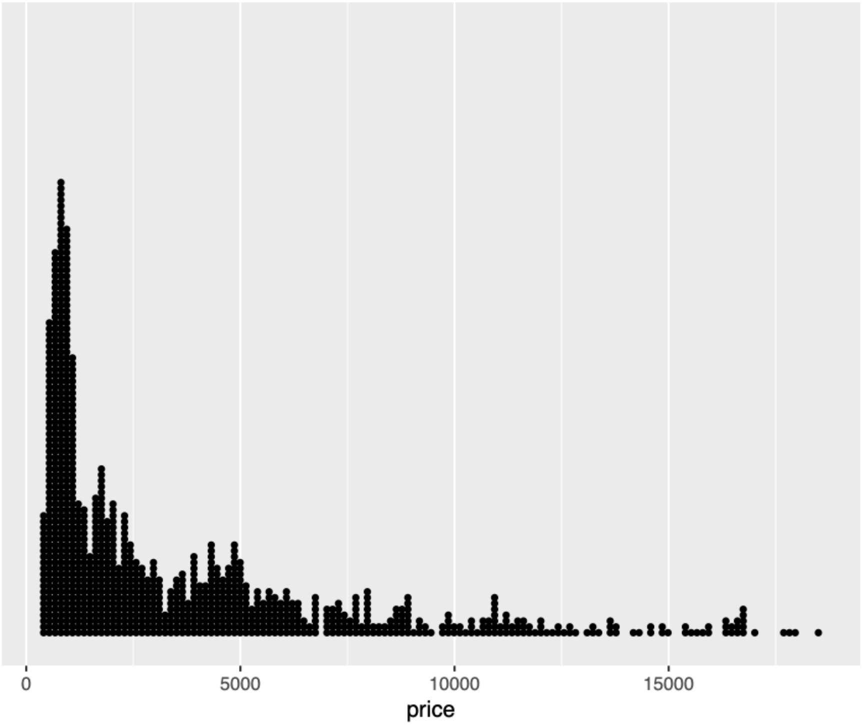


Figure 4: Dotplot of diamond prices

Histogram of the data

Investigating the histogram of the data set gives us a clue about the range of diamond prices in our data frame. As illustrated in Figure 4., the range of diamond prices provided in data set is very wide from small values, under 2000 to high values of over 15000. Due to this variety the heteroscedastic resuals are more likely to happen.

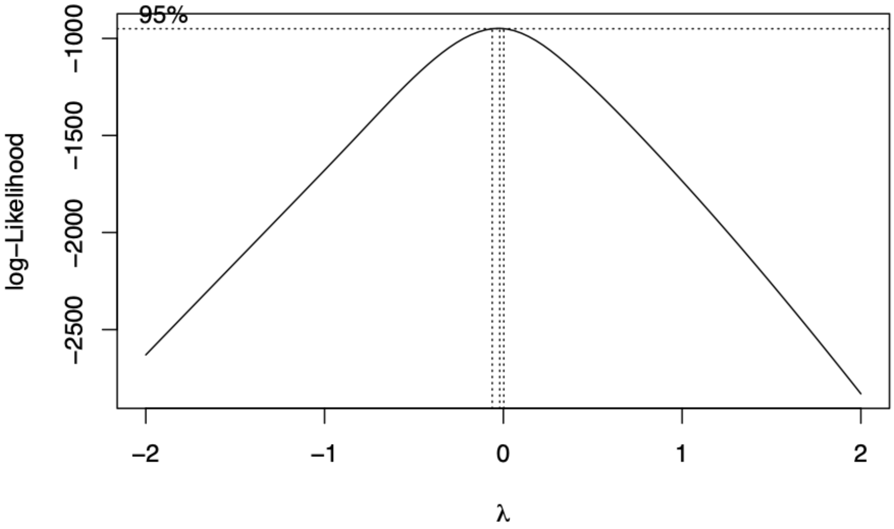


Figure 5: The Box-Cox Transformation Coefficient Value.

Box-Cox Transforming

Considering the fact that real data sets are not always following the normal distribution behaviour, as also shown in Figure 4, an appropriate transformation may improve the statistical model. In this study, BoxCox command is used to check if there is a room for a power transformation to better explain the variabilities and improve the normality. Figure 5 displays the Box-Cox transformation coefficient (lambda) value for the model with explantory variables of carat, cut, color, clarity, depth, y and z. As shown, the optimal value of lambda is close to zero. The lambda= 0 corresponds to a logarithmic transformation of response (diamond price). So a new model based on the logarithmic value of the price of the diamond is developed and the regarding diagnostic plots are checked again.

Model after Log Transformation

Since the optimum lambda is 0, a logarithmic transformation of the diamond price has been developed. Table 8 presents the ANOVA results for this logarithmic model with explanatory variables of carat, cut, color, clarity, depth, y and z. In the new logarithmic model where R^2 is 0.978, the explanatory variable of depth and the intercept are statistically less significant in this model.

Table 8: ANOVA test of the Model after log transformation.

	Dependent variable:
	log(price)
carat	-0.999^{***}
	(0.066)
cut	0.025^{***}
	(0.005)
color	0.085^{***}
	(0.004)
clarity	0.111^{***}
	(0.004)
depth	-0.028^{*}
	(0.017)
y	0.556^{***}
	(0.163)
z	1.364^{***}
	(0.270)
Constant	1.132
	(1.016)
Observations	693
R^2	0.978
Adjusted R^2	0.978
Residual Std. Error	0.150 (df = 685)
F Statistic	$4{,}382.490^{***}$ (df = 7; 685)
Note:	*p<0.1; **p<0.05; ***p<0.01

Modified Model after Log Transformation

As mentioned before, ANOVA calculations of the logartithmic model showed inadequate significance level with higher P-values than the 0.05 for depth. A new model after removal of this variable is developed whose ANOVA table is illudtrated in Table 9. As illustrated in this table Rˆ2 is 97.8%, and all the six explanatory variables of carat, cut, color, clarity, y and z, have sufficient statistical significance.

Table 9: ANOVA test of modified Model after log transformation.

	Dependent variable:
	log(price)
carat	-0.984^{***}
	(0.065)
cut	0.026^{***}
	(0.005)
color	0.085^{***}
	(0.004)
clarity	0.110^{***}
	(0.004)
y	0.819^{***}
	(0.044)
z	0.929^{***}
	(0.072)
Constant	-0.553^{***}
	(0.116)
Observations	693
R^2	0.978
Adjusted R^2	0.978
Residual Std. Error	0.151 (df = 686)
F Statistic	$5{,}099.139^{***}$ (df = 6; 686)
Note:	*p<0.1; **p<0.05; ***p<0.01

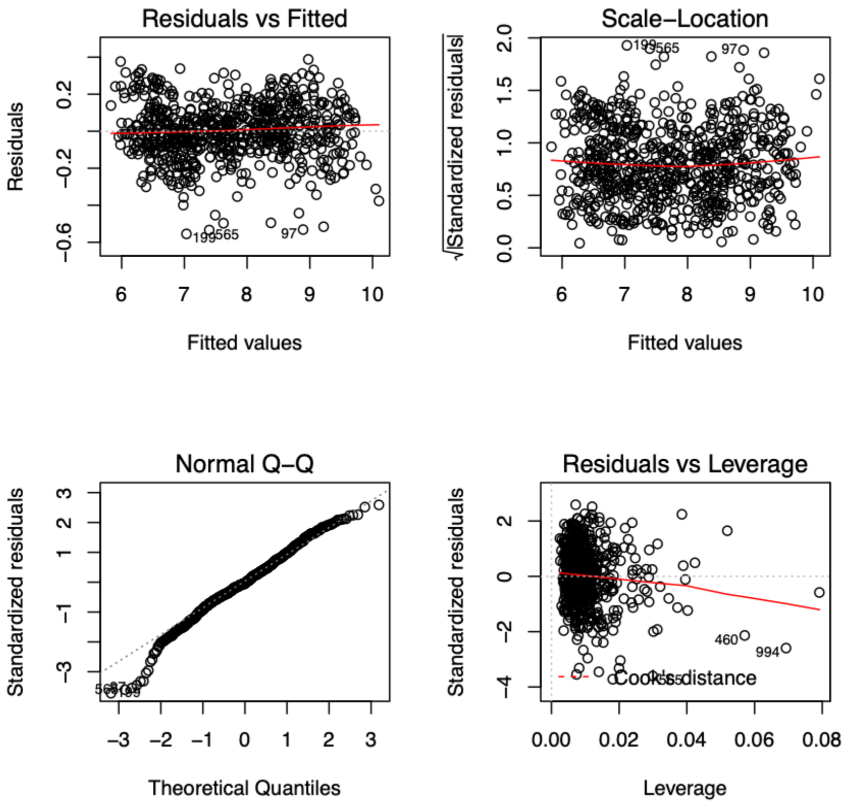


Figure 6: Diagnostic checks for Modified Model after Log Transformation.

Diagnostics Plot of Modified Model after Log Transformation

Figure 6 shows the diagnostic plots for the modified logarithmic price model. The plots show a more random behaviour for residuals than the previous non-logarithmic price model and the hetereoscedastic behaviour is gone. By the logarithmic transformation the scale location plot has improved with a more uniform mean, closer to one. The Q-Q plot also shows notable improvements in normality, although there are still some data points deviating from the ideal normal line. However, the standardized residual vs. leverage plot shows the influence of some outliers and high influencing points in our data set affecting the average line of this plot.

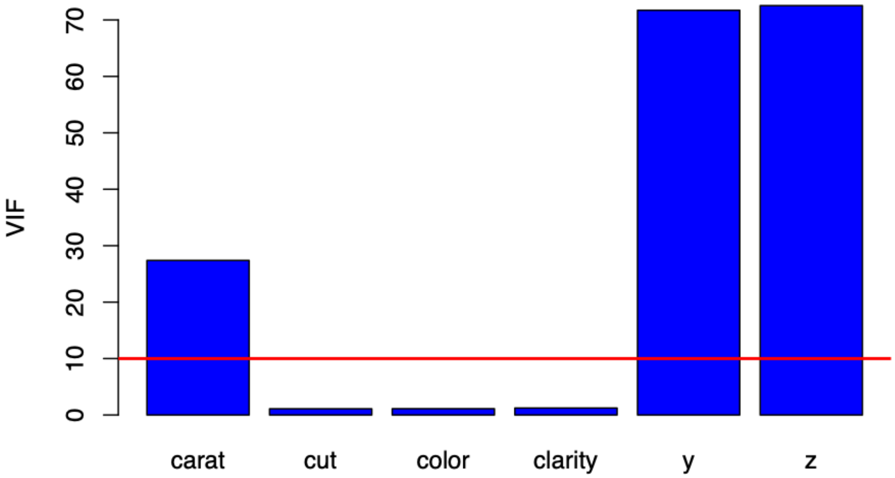


Figure 7: Variance Inflation Factor (VIF) of the Modified Model after Log Transformation.

Variance Inflation Factor (VIF)

At the beginning of this report we noticed the high correlaion between the explanatory variables, which makes it more essential to check for probable issues regarding multicollinearity. Multicollinearity is an important issue which may lead to huge errors in fitted models. The multicolliniearity of our logarithmic model is checked via the vif() command and the obtained results are illustrated in Figure 7. It is evident based on Figure 7 that the VIF of carat, y and z are huge, much larger than the critical value of 10. This means that the model has major problems due the correlations between the chosen explanatory variables.

To solve this problem, one of the highly correlated variables is removed and the VIF is rechecked for the revised fitted model. Another way to handle the problematic correlations would have been to investigate the partial correlation between the variables and decide for variable ommissions based on the calculated values.

Removing the multicollinearity (1)

To solve the issue of multicollinearity the explanatory variable z, has been removed from the model. The ANOVA table regarding the new fitted model with carat, cut, color, clarity and y as the explanatory variables is provided in table 10. The value of R^2 is 97.2% in this model and the variables are statistically significant except for the intercept. It is also surprising that the carat has a negative coefficient in this model.

Table 10: ANOVA test for the model after first step multicollinearity removal

	Dependent variable:
	log(price)
carat	-0.737^{***} (0.069)
cut	0.010^{*} (0.006)
color	0.083^{***} (0.004)
clarity	0.103^{***} (0.004)
y	1.283^{***} (0.029)
Constant	-0.012 (0.121)
Observations	693
R^2	0.973
Adjusted R^2	0.972
Residual Std. Error	0.168 (df = 687)
F Statistic	$4{,}895.060^{***}$ (df = 5; 687)
Note:	*p<0.1; **p<0.05; ***p<0.01

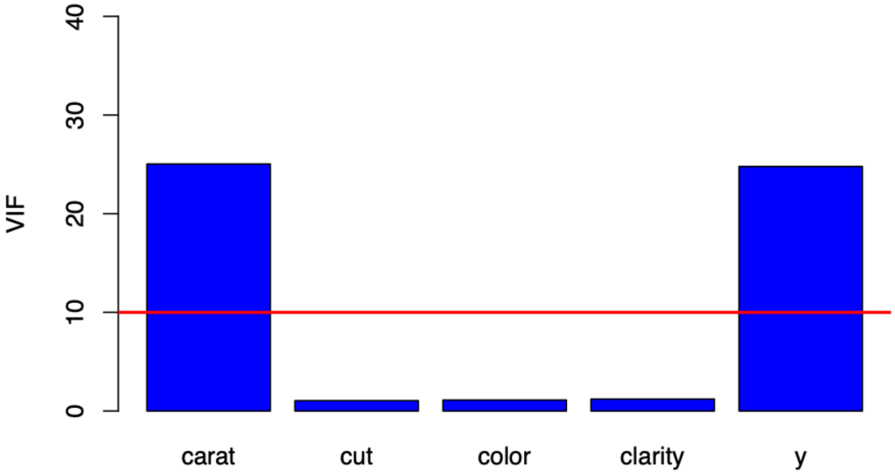


Figure 8: Variance Inflation Factor (VIF) after first step muliticollinearity removal.

Rechacking the VIF (1)

Before moving forward, the VIF of the revised fitted logarithmic model with explanatory variables of carat, cut, color, clarity and y has to be checked again. Figure 8, shows the VIF for this model, demonstarting that explanatory variables of carat and y are still highly correlated. Similar to previous step one of the two variables should be removed to solve the persisting multicollinearity issue.

Removing the multicollinearity (2)

Carat is removed and the revised logarithmic price model is developed with explanatory variables of cut, color, clarity and y. The new model's ANOVA table is shown in table 11. It is representing the R^2 value of 96.8%. However the "cut" variable has P-value of larger than 0.05, showing that it is not statistically significant in this model.

Table 11: ANOVA test of the modified model after log transformation and second step multicollinearity removal.

	Dependent variable:
	log(price)
cut	0.011* (0.006)
color	0.089*** (0.004)
clarity	0.104*** (0.004)
y	0.983*** (0.007)
Constant	1.087*** (0.067)
Observations	693
R^2	0.968
Adjusted R^2	0.968
Residual Std. Error	0.181 (df = 688)
F Statistic	5,239.984*** (df = 4; 688)
Note:	*p<0.1; **p<0.05; ***p<0.01

Final Model after Log Transformation, multicollinearity correction and insignificant explanatory variables removal

The logarithmic final model is developed with color, clarity and y as its major explanatory variables. Table 12. presents the ANOVA table for this model. In the final model, all the explanatory variables are statistically significant and have positive coefficients as expected. Also, the model explains the 96.8 % of the total variabilities.

Table 12: ANOVA test of the Final Model.

	Dependent variable:
	log(price)
color	0.089^{***}
	(0.004)
clarity	0.106^{***}
	(0.004)
y	0.982^{***}
	(0.007)
Constant	1.124^{***}
	(0.064)
Observations	693
R^2	0.968
Adjusted R^2	0.968
Residual Std. Error	0.181 (df = 689)
F Statistic	$6,967.214^{***}$ (df = 3; 689)
Note:	*p<0.1; **p<0.05; ***p<0.01

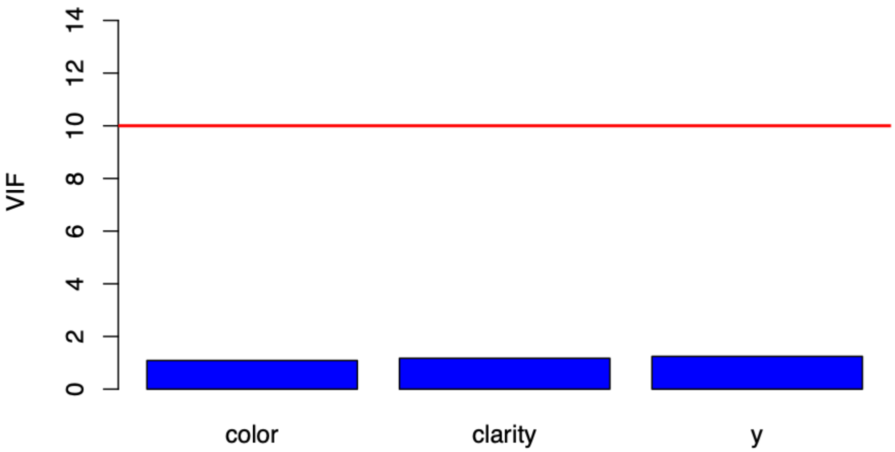


Figure 9: Variance Inflation Factor (VIF) of the Final Model.

Rechacking the VIF (2)

The final logarithmic fitted model with explanatory variables of color, clarity and y, has VIF of less than 2, as shown in Figure 9. This means that the multicollinearity has been removed from our model after corrections.

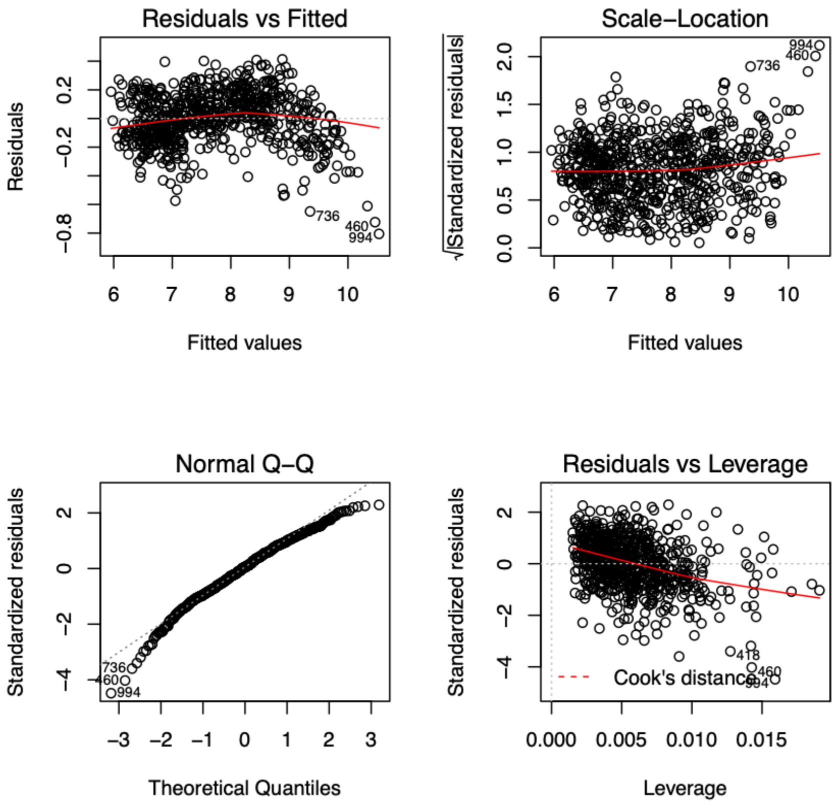


Figure 10: Diagnostic checks for Final Model.

Residual analysis of the Final Model

The residual diagnostic plot of the final model is shown in Figure 10. Although they show that the model is struggling with some patterns due to the variabilities from the data set, the randomness of the residuals, the scale location plot, the residual alignment with the normal 45-degree line in the Q-Q plot and the Residual vs. Leverage plot are all acceptable, if the model be capable of predicting the test sample to the acceptable extent.

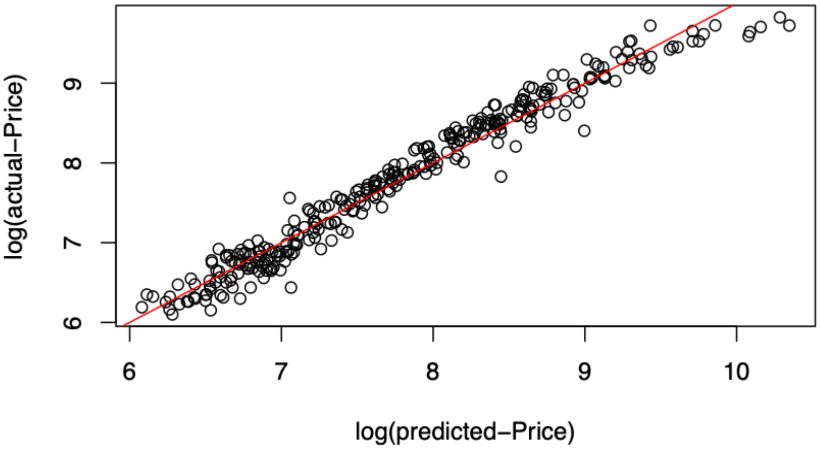


Figure 11: Prediction plot: Comparing the Actual Value and the Fitted Value of the Test Data

Prediction

Now that the Final model has passed the residual and multicollinearity investigations, we need to validate the model to see how accurately it can predict the diamond prices in future. In other words, the fitted model which was generated based on the Train sample set, should work for the test sample set. Figure 11. presentes the relationship between the actual value of the diamond price from the Test sample set and their fitted values based on the developed logarithmic model in this report. It seems that there are acceptable agreements between the actual values and the fitted ones, so the final model can work well in estimation of the future diamond prices.

As another tool for checking the the model regarding its prediction capabilities, the root mean square error (RMSE) analysis is also conducted. Table 13 presents the obtained RMSE values for the fitted model based on the Train and Test data sets, which are both equal to 0.03 confirming that the final fitted model has a good prediction ability.

	RMSE
Train Data	0.03
Test Data	0.03

Table 13: RMSE Comparisons

Conclusions

The influence of 9 various explanatory variables on diamond price was investigated in this study. The variable selection analyses (backward stagewise and best subset), omitted 2 explanatory variables and the remaining 7 variables of carat, cut, color, clarity, depth, y, z were picked as the most statistically significant predictors. The diagnostic checks and histogram dot plot of the data set showed that the original model has issues with normality and covering the wide range of data from the dataset. Logarithmic power transformation was then applied based on the optimum transformation coefficient of zero calculated from the BoxCox command. The developed logarithmic transformation was then improved by removal of less significant variable of depth. The developed model at this stage improved the residual behaviour and also the explanation of variabilities, however the high correlation between the model's explanatory variables led to problematic VIF values. To solve the issue the highly correlated explanatory variables of z and carat were omitted step by step to save the model from multicullinearity. After the final ommission of cut as the last non-significant variable from the new model, the finalized fitted linear regression model was developed as follows:

$\log(\text{Diamond Price}) = 1.1245 + 0.089\ \text{color} + 0.106\ \text{clarity} + 0.982\ \text{y}$

The final model explains 96.8% of the total variabilities, and has shown good ability in predicting the diamond price in future.