Diamond Price Prediction Based on Specifications

A Project Report
Submitted for the Partial Fulfillment of the Requirements of the course

STAT 6021-001

Submitted by-
Chloe Yan - ly6pc
G Haritha - hg5mn
Shengua Wu - sw2hg
Andi Yu- ay6wy

July 2019

Table of Contents:

Abstract:

One of the first things most people learn about diamonds is that not all diamonds are created equal. Diamond professionals use the grading system developed by GIA in the 1950s, which established the use of four important factors to describe and classify diamonds: Clarity, Color, Cut, and Carat. Diamonds with certain qualities are more rare and more valuable than diamonds that lack them. In this project we use multiple linear regression to predict the price of the diamond based on these 4C's. Finally, we created a webapp to let user to input his/her specifications and get an estimated price based on our best MLR model.

1. Introduction

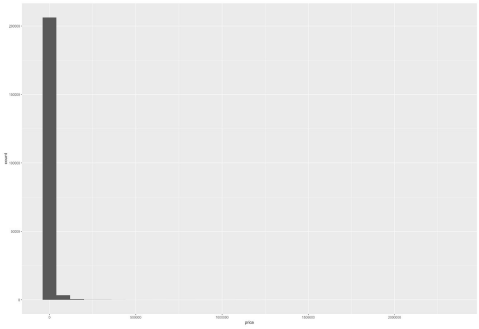
The diamond dataset was sourced from bluenile.com, it contains 4 variables and 210,638 observations of diamonds.

Variables	Descripsion
carat	Weight of the diamond (0.23 - 20.45)
clarity	A measurement of the clearness of the diamond (SI2(worst), SI1, VS2, VS1, VVS2, VVS1, IF, FL(best))
color	The color of diamond (D(worst), E, F, G, H, I, J(best))
cut	Quality of cut (Good, Very good, Ideal, Astor ideal)
price	The price of the diamond(dollars) (229 - 2317596)

After looking at the data structure, we have a total of 5 variables in the data set, price is the price of each diamond, which is our response variable in the project. Carat, Clarity, Color and Cut are four categorical variables. Since we have a numerical dependent variable with both numerical and categorical independent variables, we decided to build a linear model to predict the diamond's price based on the diamond's attributes we have on hand.

2. Model Analysis:

A. Data Visualization



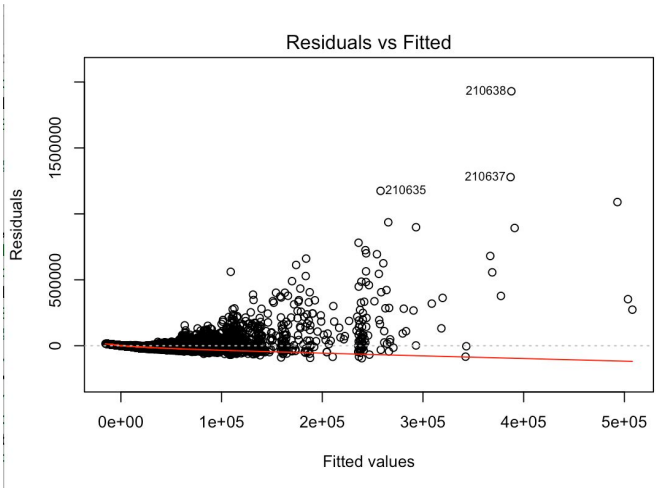
From the histogram of the price, we can see the price's distribution is highly right-skewed. It is because some diamonds have extremely high price compare to most diamonds so we can barely see the right tail from the plot. So we considered to first set up a raw model to figure out what improvements we can do to our prediction. In order to get a better sense of how well our model perform, we split our data into 30% testing and 70% training data set. So we can build in the model in the training data set, and then test it in the testing data set. Then we looked up the correlation between carat and price which is 75% in this case.

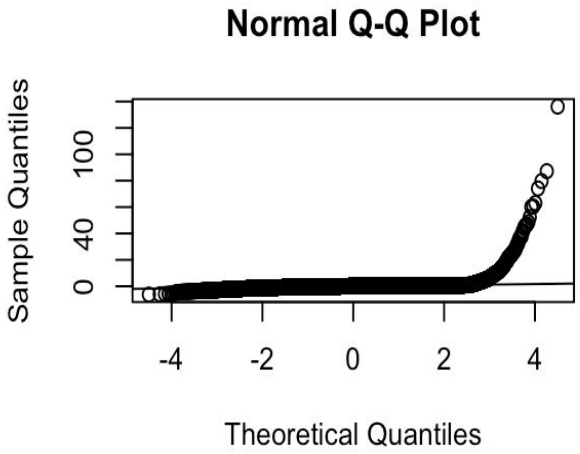
B. Set up a raw model

We decided to build a raw model first to see its performance : Price ~ Carat + Cut + Color + Clarity

Step1 ：The t-test shows that all facotors are significant since our regression model has p-value lower than 0.05 and the model's r-squared is 0.58. It means that 58% of the variance of price can be explained by our raw model. In a general sense, a valid regression model should have over 70% r-square. Hence, even the t-test is significant, the model did not do a good job.

Step2: By looking at the residual plot on the left side below, we find that the residuals spread out as the fitted value increases. The non-constant variance may caused by non-linearity relationship between y and x, or our data violates the linear model assumption. Also, in the normal Q-Q plot on the right side, most points fall in a straight line, but there is a heavy tail at the right hand right. We may have some unusual observations or extreme sample that we didn't count in our model. All of them implies that we have to check the multicollinearity or do some transformation to our model.



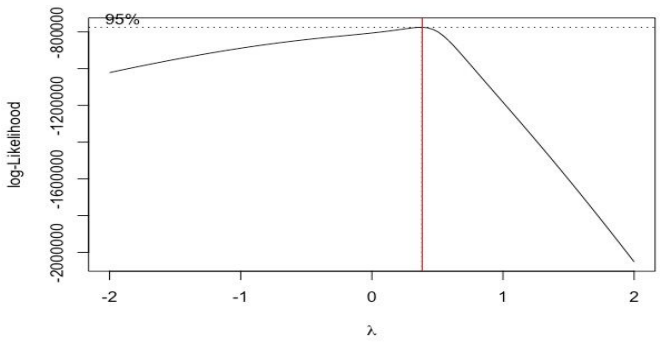


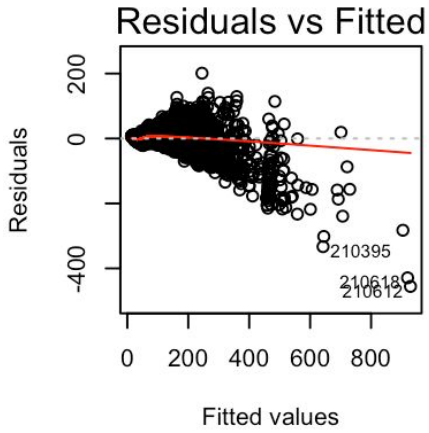
Step3: Then we checked the VIF(Variance Inflation Factor) in our model. Since there is no variable has high VIF, we can conclude that there is no multicollinearity.

Carat	Clarity	Color	Cut
1.03	1.04	1.03	1.03

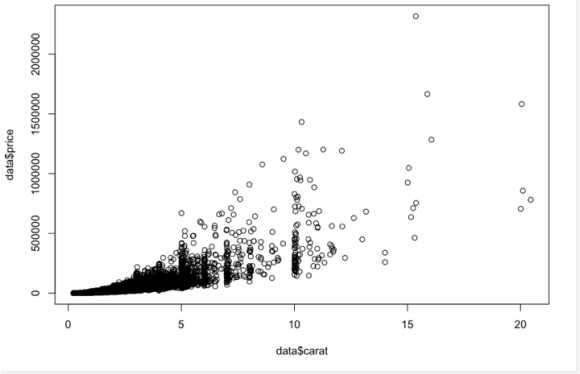
Step4: Then we did conduct a box-cox procedure trying to transform our price, the best lambda the box-cox suggests it 0.38. After we find the best lambda, we fit a new model based on the transformed price. the residuals are still not fit a normal distribution, so we have to change the model. The F-test turns out it is a significant model and the r-square is 92.8%, which is a great

improvement. Then we checked the residual plot who has the similar pattern as the raw model which implies non-constant variance. Hence we tried other ways to improve our model.

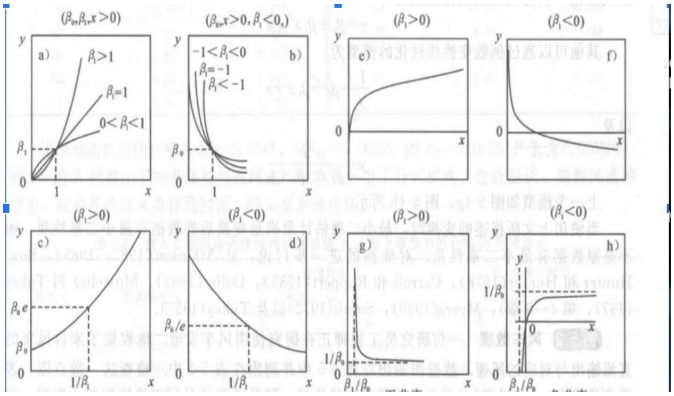




C. Transformation of Price



（a） carat vs price

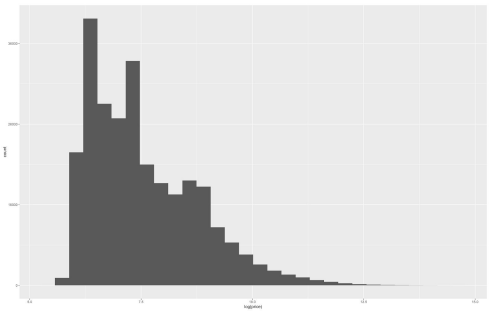


（b) models in the book

Matching the models in the book is a quick way to find the best transformation.
From graph(a), the values of price goes larger where the value of carat goes lager at the range 0 - 5, so we exclude model_b, model_f , model_d, and model_g. The slope of model _c, model _e and model_ h changes too quick at range 0- 5. Finally, we conclude model_a and model_ f probably are good transformation. Referring to table 5.4, we updated our models to log(price) ~ log(carat).

TABLE 5.4 Linearizable Functions and Corresponding Linear Form

Figure	Linearizable Function	Transformation	Linear Form
5.4*a, b*	$y = \beta_0 x^{\beta_1}$	$y' = \log y, x' = \log x$	$y' = \log \beta_0 + \beta_1 x'$
5.4*c, d*	$y = \beta_0 e^{\beta_1 x}$	$y' = \ln y$	$y' = \ln \beta_0 + \beta_1 x$
5.4*e, f*	$y = \beta_0 + \beta_1 \log x$	$x' = \log x$	$y' = \beta_0 + \beta_1 x'$
5.4*g, h*	$y = \frac{x}{\beta_0 x - \beta_1}$	$y' = \frac{1}{y}, x' = \frac{1}{x}$	$y' = \beta_0 - \beta_1 x'$



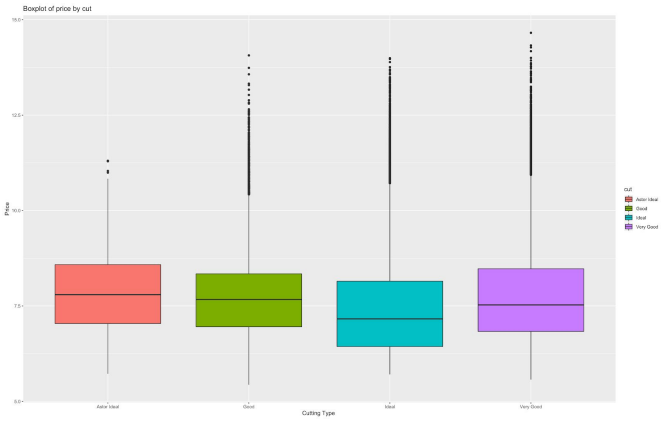
After taking log transformation to price, we plotted out the price again, As we can see, the distribution became much better.

D. Data Visualization after Transformation of Price

After settling down the transformation to price, we plotted out the relationship between Log(price) against our categorical variables: Cut, Color, and Clarity.

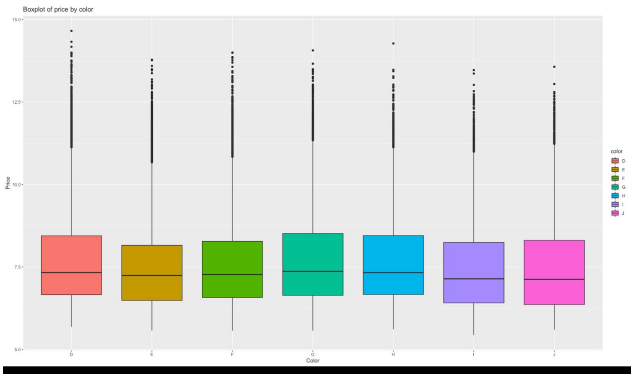
Log(Price) vs. Cut

From this plot, we can see that the diamond with the ideal cut have the lowest price, compared to the diamond with astor ideal cut which have the highest price. However, diamond with good, ideal and very good cut all have a really high upper bound (maximum price).



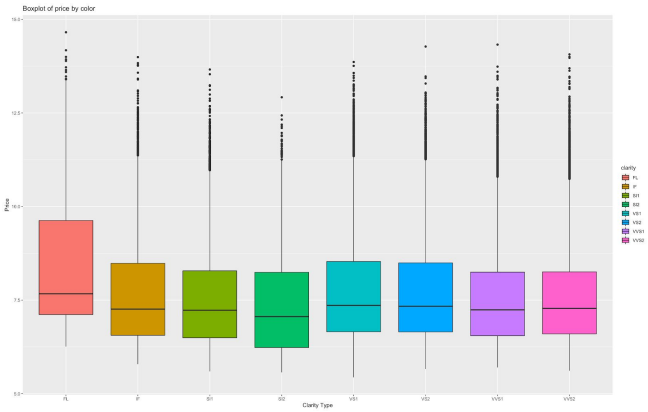
Log(Price) vs. Color

From this plot, the diamond with the color (D to J) have similar prices and they all have a high maximum price. However, as we can see, the worst color D has the maximum price among all of these.

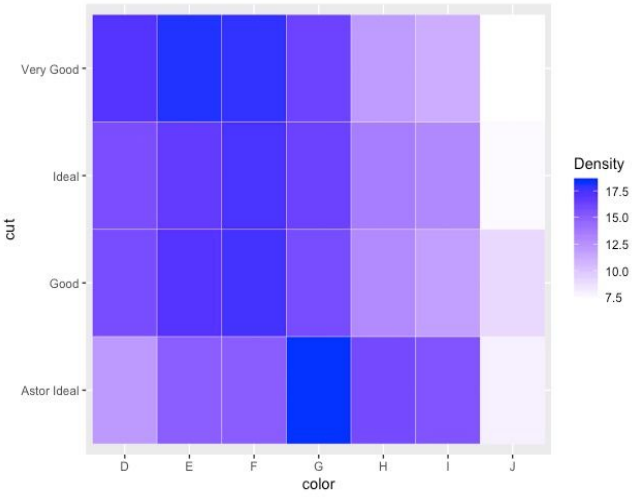


Log(Price) vs. Clarity

From this plot, we can clearly see diamond with flawless(FL) clarity have the highest price and it is skewed right, which means those diamonds have extreme high price compared to others.
Also, the diamond with SI2 clarity have the lowest price. Besides these two, others share a similar price statistics and all have a relatively high maximum price.



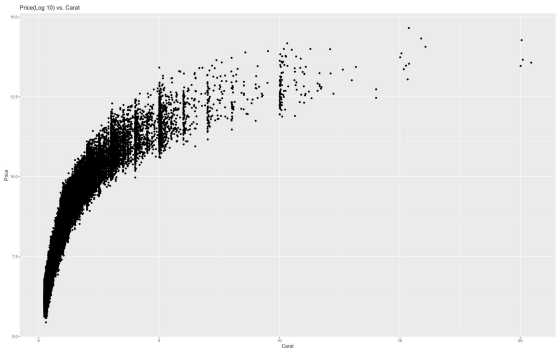
Here's the heat map of cut vs color. From this, we can conclude that most diamonds with ideal and good cuts are from color F. Most of the diamonds with astor idea cut are from color G. And most of diamonds with very good cut is from color E and F.



Then, we plot diagrams of Log(Price) against numerical data(Carat) and transformations of Carat in order to compare them and get the most appropriate linear model.

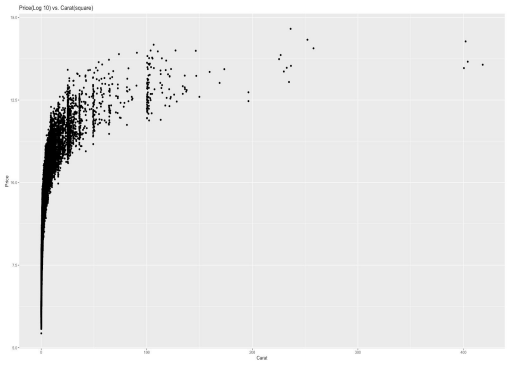
Log(Price) vs. Carat

As we can see, this diagram forms a shape of exponential distribution. To improve it, we tried several transformation to carat and conducted the following three plots of Log(price) against square of carat, square root of carat, the cubic square root of carat.



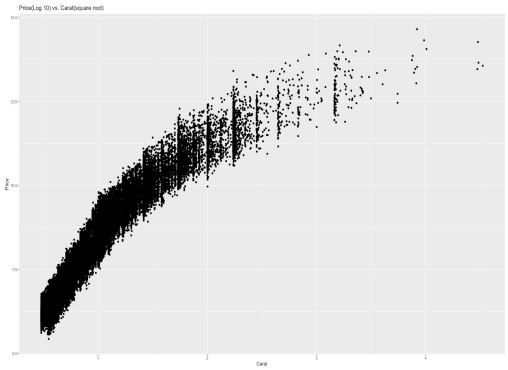
Log(Price) vs. Carat ^ 2

Based on the observation above, we plotted this diagram of log(price) against carat ^ 2. As we can see, the diagram becomes even worse. Therefore, we dropped the transformation and tried to use square root of carat.



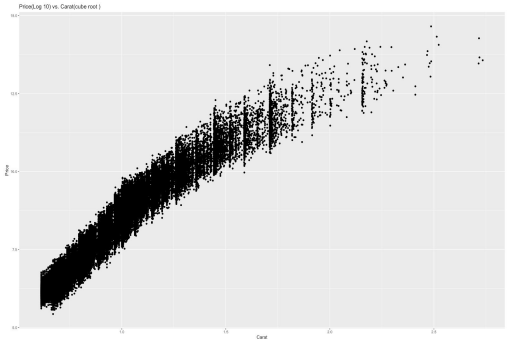
Log(Price) vs. Carat ^ ½

Here, we have the plot of log(price) against the square root of carat. It is better than the two plots before.



Log(Price) vs. Carat ^ ⅓

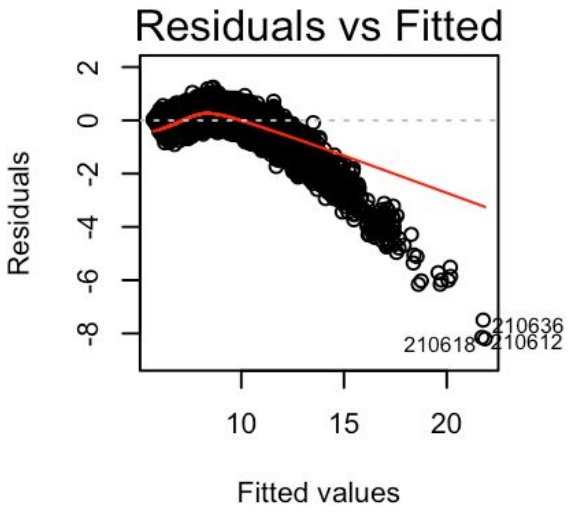
We tried to plot using log(price) against the cubic square root of carat. We clearly see that this plot is better than three plots before. The plot shows a linear shape relationship between log(price) and cube root of carat. Based on the result, we created a new model by adding carat with cube root carat at the same time.

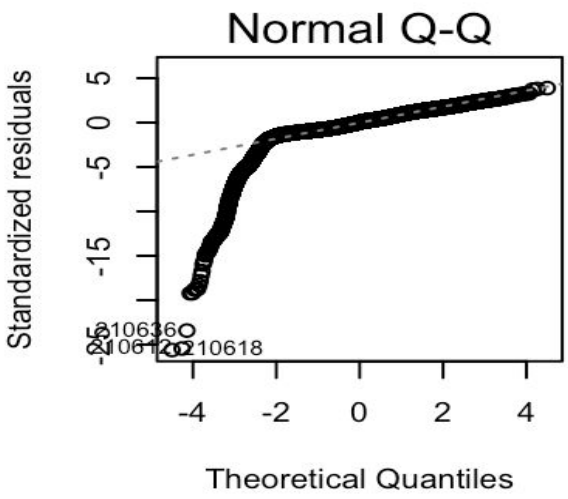


E. Model Selection

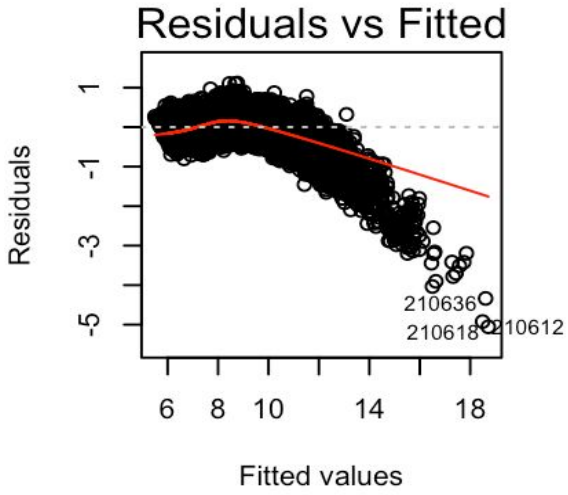
These models are selected and tested based on the data visualization we made. From the summary table, we can see all of these factors are significant and p-values are all pretty close to 0, however, we have different R-square values. The diagnostic plots are different for each model we have chosen.

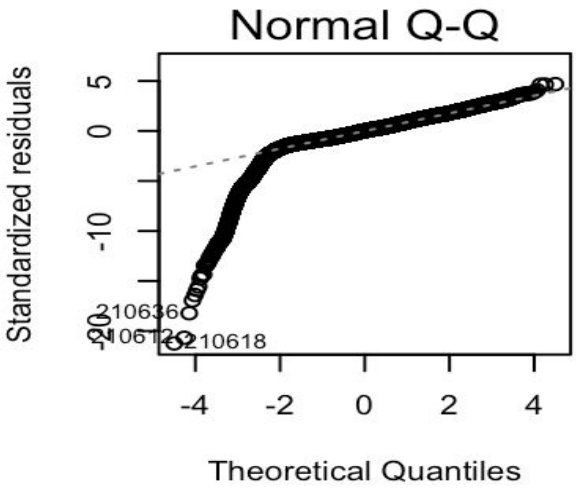
log(price)~cut + clarity + color + I(carat^(1/2))	R-square = 0.9303



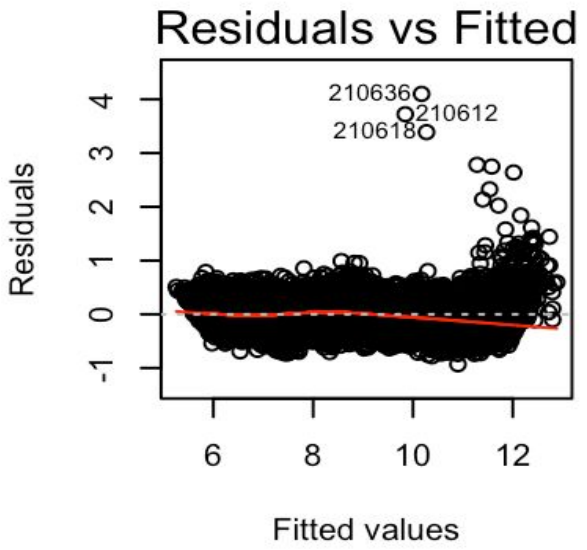


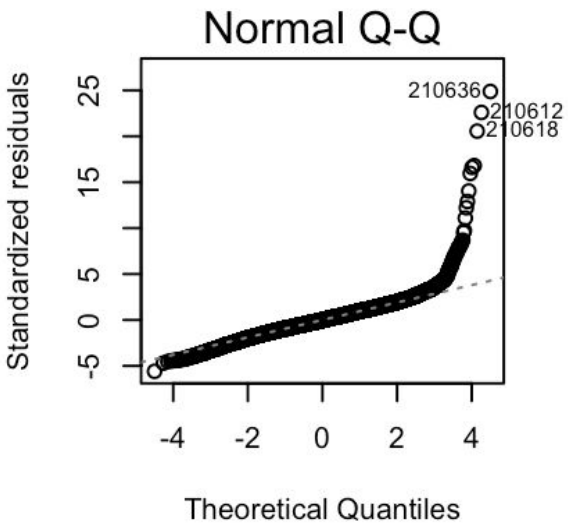
log(price)~cut + clarity + color + I(carat^(1/3)).	R-square = 0.9615



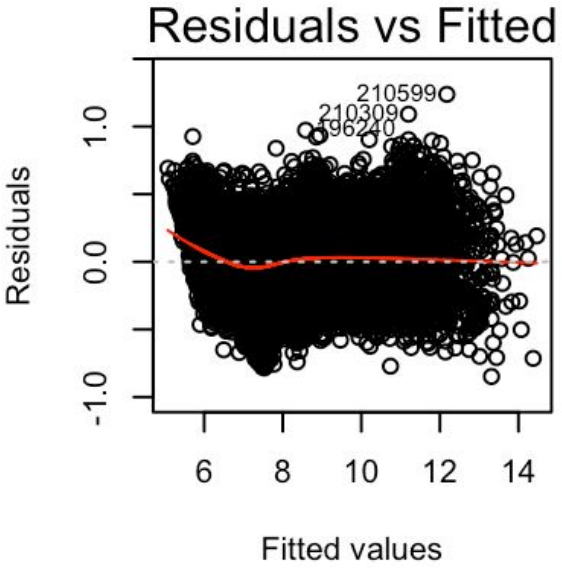


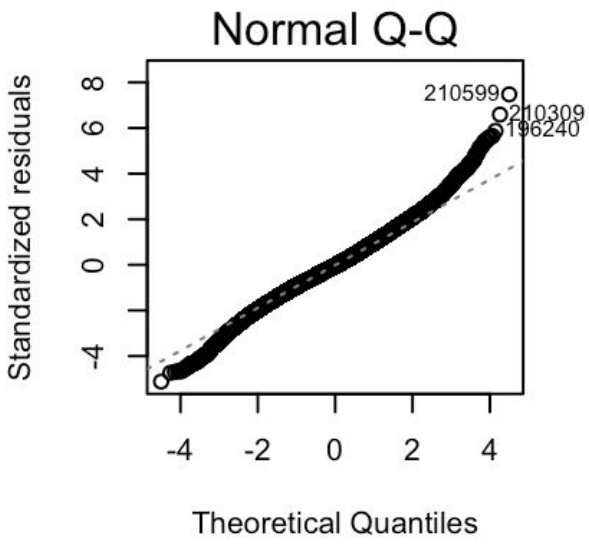
log(price)~cut + clarity + color + I(carat^(1/3) + carat)	R-square = 0.9812





log(price)~cut + clarity + color + log(carat))	R-square = 0.9814





We also checked those models' performance in the testing data set and got their mean square error, which shows the same model accuracy as the r-square suggested and we can't really make comparison between them. After comparison of residual plots and Normal Q-Q plots between these models, we are able to conclude that this model log(price)~log(carat) + cut + clarity + color) is the best model for the diamond data. Because the residuals are evenly distributed around zero, which shows constant variance, and most points fall on the normal qqline. Though there are still some outliers and extreme points fall out of the line. It seems the best plot we can get from the candidate models.

3. Conclusion and Limitation

Even though the mode price = exp(beta1*log(carat) + beta2*cut + beta3*color + beta4*clarity) is not perfect, it still works in most cases. We also have some limitations in the project, if we have more time, we can take a closer look at those outliers and decide whether to drop them from the data set. In addition, the residual and normal Q-Q plot from our final model shows heavy tail, which might be caused by some extreme points like extreme high-quality and high-price or super cheap diamond. We can also set up dummy variables to change the slope and intercept of the regression once price hitting the bound. We extrapolate that there are many types of diamond. Some of them are low-priced just like mirror, but some of them are invaluable selling at auction. In other words, rich people would not buy a diamond on the bluenile.com.

4. Reference

Montgomery, Douglas C., et al. *Introduction to Linear Regression Analysis*. Wiley, 2013.