Optimizing Buying Strategies for Diamonds

By

Tomiwa Omotesho

December 2021

Chapter 1

1.1 Introduction

There are many aspects when it comes to pricing a diamond, and it is not only important to look at each one of these elements by itself, but also to look at them in tandem. The four main elements of categorizing a diamond are Carat, Cut, Color, Clarity. These are known as the 4 Cs of the diamond. There are also five lessor known factors that could cause a diamond's value to increase or decrease. They are the depth, table, x, y and z of the diamond; all of which have to do with the shape or size of the diamond.

If we look at the past few years we can see that roughly 133 million carats of rough diamonds where produced. The two largest producers are Russia and Botswana with about half of the world's production combined. Angola, Australia, Namibia and Canada produce most of the remaining diamonds.

1.2 Objective

1. To make a comprehensive chart showing the value of diamonds by each of there individual elements as well as how the value change when multiple elements are combined.

2. To find the point of diminishing returns in terms of an increase in quality vs the increase in price.

3. To see how strong of an association there is between diamond prices and each of its regressor variables.

1.3 Research questions

1. What regressors variable will be in the final model?

2. What regressor will have the largest impact on the cost of the Dimond?

Chapter 2

Data and Methodology

2.1 Data

The data that we are performing the regression on came from the website Kaggle and more specifically came from the article intitled Diamonds. It started out with 9 regressor variables and a response variable. The response variable was the price of the diamond while the regressor variable was broken down into two main categories. The first category being qualitative in this case, means it cannot be measured and the diamonds are placed into groups. The second group is quantitative meaning the observation can be measured. The qualitative variables within the data are as follows:

1. The cut which is broken down into 5 different groups from least desirable to most desirable (Fair, Good, Very Good, Premium, Ideal). The cut describes the general shape symmetry proportion and polish of the diamond.
2. Color which has to do with how much yellowing the diamond has. Going from worst to best is labeled (J, I, H, G, F, E, D).
3. Clarity is how many internal inclusions and external blemishes there are on the diamond. The scale going from worst to best is as follows (I2, SI2, SI1, VS2, VS1, VVS2, VVS1, IF

The next group of 6 variables are quantitative variables are as follows:

1. Carat which is the weight of the diamond. For example, 1 carat is equivalent to .2 grams.
2. Depth the height of a diamond measured from the culet to the table divided by its average girdle diameter
3. Table is the width of the top of the diamond divided by its average girdle diameter.
4. X is the length in MM
5. Y is the width in MM
6. Z is the depth in MM

2.2 Methodology

2.2.1 Full Model

The first model we created contained all the variables with the price being the response variable (y) and the remaining 9 being the regressor variables (x).

Linear Regression Model: y = $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$

2.2.2 Assessment of Model Assumptions

We assessed the assumptions for linear regression in the model. We checked for following assumptions.

- The relationship between response y and the regressors is linear, at least approximately
- The random error term ε has zero mean and constant variance σ^2
- The errors are uncorrelated
- The errors are normally distributed

2.2.3 Transformation

From the model assumptions assessment, we determined the regressor variables that had a non-linear relationship with the response variable and performed necessary transformations to satisfy the model assumptions.

2.2.4 Multicollinearity

We checked for multicollinearity among the regressor variables and selected the regressor variables with variance inflation factor (VIF) above 4. A VIF greater than 4 indicated moderate multicollinearity, while a VIF greater than 10 indicates severe multicollinearity.

2.2.5 Pearson Correlation Coefficient

We computed the Pearson correlation coefficients for the regressor variables to determine the variables that were highly correlated. The Pearson Correlation Coefficient ranges from -1 to 1. A Pearson Correlation Coefficient of -1 indicates a strong negative linear relationship, 0 indicates no linear relationship and +1 indicated a strong positive linear relationship.

2.2.6 Analysis of Variance (ANOVA)

We analyzed the type II ANOVA table to determine the variables to drop from the model until a model without multicollinearity was achieved. The type II ANOVA table gives you the contributions of each regressor variable to the model.

2.2.7 Variable Selection

We carried out variable selection using the Stepwise selection (bi-direction) method to remove variables without significant contributions to the model.

2.2.8 Identification of Outliers, Leverage and Influential Points

We used internally and externally studentized residuals to identify outliers while we used hat matrix to identify leverage points. To identify influential points, we used Difference in Fits (DIFFITS), Cook's distance and COVRATIO. Observations common to all three categories were identified and removed from the model.

Chapter 3

Data Analysis and Results

3.1 Full Model

3.1.1 Construction of Full Model (Model 1)

The first model we created contained all the variables with the price being the response variable (y) and the remaining 9 being the regressor variables (x).

hat(price)= 2184.477+11256.978(carat)+ 579.751(factor(cut)Good)+ 832.912(factor(cut)Ideal)+ 762.144(factor(cut)Premium)+ 726.783(factor(cut)Very Good) -209.118(factor(color)E) -272.854(factor(color)F) -482.039(factor(color)G) -980.267(factor(color)H) -1466.244(factor(color)I) -2369.398(factor(color)J)+ 5345.102(factor(clarity)IF)+ 3665.472(factor(clarity)SI1)+ 2702.586(factor(clarity)SI2)+ 4578.398(factor(clarity)VS1)+ 4267.224(factor(clarity)VS2)+ 5007.759(factor(clarity)VVS1)+ 4950.814(factor(clarity)VVS2) -63.806(depth) -26.474(table) -1008.261(x)+ 9.609(y) -50.119(z)

This model had a p-value of 2.2e-16, and an adjusted R^2 value of 0.9198. This mean that 91.98% of the data's correlation can be explained in the model. See Appendix A for the summary of the model

3.1.2 Assessment of Model Assumptions

The scatter and residual plots of the response variables against the quantitative regressor variables was used to check for linearity. The plots revealed a non-linear relationship between the response variable and most of the quantitative regressor variables. See figures 1 2 and 3. The plot for zero mean and constant variance was not satisfactory as the plot was biased, did not have a zero vertical mean and did not have constant variance of random error. See figure 4. The normal probability plot revealed a heavy tailed distribution with most of the observation not on the normal probability line. See figure 5. We could not perform a Shapiro-Wilk test because the number of observations in our dataset exceeded the range for the test (3 to 5000 observations).

3.2.1 Construction of New Model (Model 2)

We used the box cox transformation to improve the linear relationship between the response variable and the regressor variables. We obtained an lambda value of 0 from the box cox transformation and performed a log transformation on the response variable price. The new model was fitted with the regression line equation as follows:

hat(log(price))= -3.1460012 -0.6120540 (carat)+ 0.0911177 (factor(cut)Good)+ 0.1557466 (factor(cut)Ideal)+ 0.1102122 (factor(cut)Premium)+ 0.1244075 (factor(cut)Very Good) - 0.0581373 (factor(color)E) - 0.0894535 (factor(color)F) - 0.1574203 (factor(color)G) - 0.2582798 (factor(color)H) - 0.3846735 (factor(color)I) - 0.5243942 (factor(color)J)+ 1.0953087 (factor(clarity)IF)+ 0.6078471 (factor(clarity)SI1)+ 0.4409340 (factor(clarity)SI2)+

0.8184076 (factor(clarity)VS1)+ 0.7503510 (factor(clarity)VS2)+ 1.0047321 (factor(clarity)VVS1)+ 0.9380828 (factor(clarity)VVS2) + 0.0521543 (depth) + 0.0089978 (table) +1.1646945 (x)+ 0.0325386 (y) - 0.0427049 (z)

This model had a p-value of 2.2e-16, showing that at least one of the variables is associated with the price variable. The adjusted R^2 value was 0.9701. This mean that 97.01% of the data's correlation can be explained in the model.

3.2.2 Assessment of Model Assumptions

Improvements were noticed in the scatter plots, residual plots, zero mean and constant variance plot and normal probability plot. However, most of the plots were still not satisfactory. See figures 6, 7, 8.

3.3.1 Construction of Better Model (Model 3)

We performed a log transformation on price and carat. The model was fitted with the regression line equation as follows:

The Fitted model is:

hat(log(price))= 7.385e+00+ 1.771e+00 log(carat)+ 7.935e-02 (factor(cut)Good)+ 1.579e-01 (factor(cut)Ideal)+ 1.348e-01 (factor(cut)Premium)+ 1.153e-01 (factor(cut)Very Good) -5.460e-02 (factor(color)E) -9.441e-02 (factor(color)F) -1.607e-01 (factor(color)G) -2.527e-0 (factor(color)H) -3.754e-01 (factor(color)I) -5.147e-01 (factor(color)J)+ 1.115e+00 (factor(clarity)IF)+ 5.967e-01 (factor(clarity)SI1)+ 4.302e-01 (factor(clarity)SI2)+ 8.153e-01 (factor(clarity)VS1)+ 7.452e-01 (factor(clarity)VS2)+ 1.020e+00 (factor(clarity)VVS1)+ 9.490e-01 (factor(clarity)VVS2) + 1.362e-03 (depth) +7.028e-05 (table) + 5.685e-02 (x- 1.487e-03 (y) - 6.624e-03 (z)

This model had a p-value of 2.2e-16, showing that at least one of the variables is associated with the price variable. The adjusted R^2 value was 0.9827. This mean that 98.27% of the data's correlation can be explained in the model. See Appendix B for the summary of the model.

3.3.2 Assessment of Model Assumptions

The model assumptions are somewhat satisfied. We noticed a major improvement in the scatter and residual plots, especially for the plot between "price" and "carat". The exponential relationship between these variables seen in the initial model (model 1) changed to a linear relationship after the log transformation on "price" and "carat". For the normal probability plot, we also see an improvement. See figures 9,10,11,12.

3.3.3 Check for Multicollinearity

Regressor variables "carat", "x", "y" and "z" had VIF values greater than 4. This indicated that the model was suffering from multicollinearity. See Appendix C.

3.3.4 Pearson Correlation Coefficient, ANOVA and Variable Selection

The Pearson Correlation Coefficients of the regressor variables were computed with the cor() function. We analyzed the correlation table and noticed that variables "x", "y" and "z" were highly correlated with the "carat" variable. We looked at the type II ANOVA table for the model and determined that the variables "x", "y" and "z" were to be dropped from the model as they were not significant contributors to the model. See Appendix D, E and F.

Regressor variables "depth and "table" were also dropped from the model with the Bi-direction stepwise selection process, as these variables were not significant contributors to the model

3.3.5 Check for Outliers, Leverage and Influential Points

Using the methods mentioned in the methodology, we identified 215 observations that were common to all three categories. See Appendix G.

3.4 Final Model (Model 4)

the scatter plots of the transformed variables. We then fit r model with the transformed data:

The Fitted model is:

Hat(log(price))= 7.869425+ 1.886598*log(carat)+ 0.088245 (factor(cut)Good)+ 0.170505 (factor(cut)Ideal)+ 0.148842 (factor(cut)Premium)+ 0.125335 (factor(cut)Very Good) - 0.051362 (factor(color)E) - 0.092831 (factor(color)F) - 0.157825 (factor(color)G) - 0.247643 (factor(color)H) -0.3724385 (factor(color)I) - 0.509736 (factor(color)J)+ 1.081322 (factor(clarity)IF)+ 0.570303 (factor(clarity)SI1)+ 0.405671 (factor(clarity)SI2)+ 0.790374 (factor(clarity)VS1)+ 0.719612 (factor(clarity)VS2)+ 0.996315 (factor(clarity)VVS1)+ 0.924772 (factor(clarity)VVS2)

This model also has a P-value of 2.2e-16 showing that at least one of the variables is associated with the price variable. The basic summary shows a R value of 0.9838 and an adjusted R^2 value of 0.9838, which mean that 98.38% of the data's correlation can be explained in the model. The final model has the highest adjusted R^2 value out of the 4 models built.

3.4.1 Assessment of Model Assumptions

The model assumptions are satisfied. The scatter plot of the one quantitative regressor variable left in the model as a linear relationship with the response variable. Also the probability normal plot is very good with almost all the observations on the normal probability line. See figures 13, 14, 15, 16.

3.5 Standardized Beta Coefficients

Based on the price prediction model, 4c, carat, color, clarity and cut are most important feature to determine the diamond price.

```
> lm.beta(mlr1)

Call:
lm(formula = log(price) ~ log(carat) + cut + color + clarity,
    data = diamonds)

Standardized Coefficients::
 (Intercept)   log(carat)      cutGood     cutIdeal   cutPremium cutVery Good       colorE       colorF       colorG
  0.00000000   1.08574439   0.02268512   0.07782550   0.05991044   0.04816351  -0.02062403  -0.03557538  -0.06430676
      colorH       colorI       colorJ    clarityIF    claritySI1   claritySI2   clarityVS1   clarityVS2  clarityVVS1
 -0.08930403  -0.11041312  -0.11187382   0.19661260   0.25036980   0.15857227   0.28701508   0.30652140   0.25235069
 clarityVVS2
  0.27234720
```

To understand which feature has the most impact in price prediction, standardized beta coefficients above showed that *carat* carries is the most important feature.

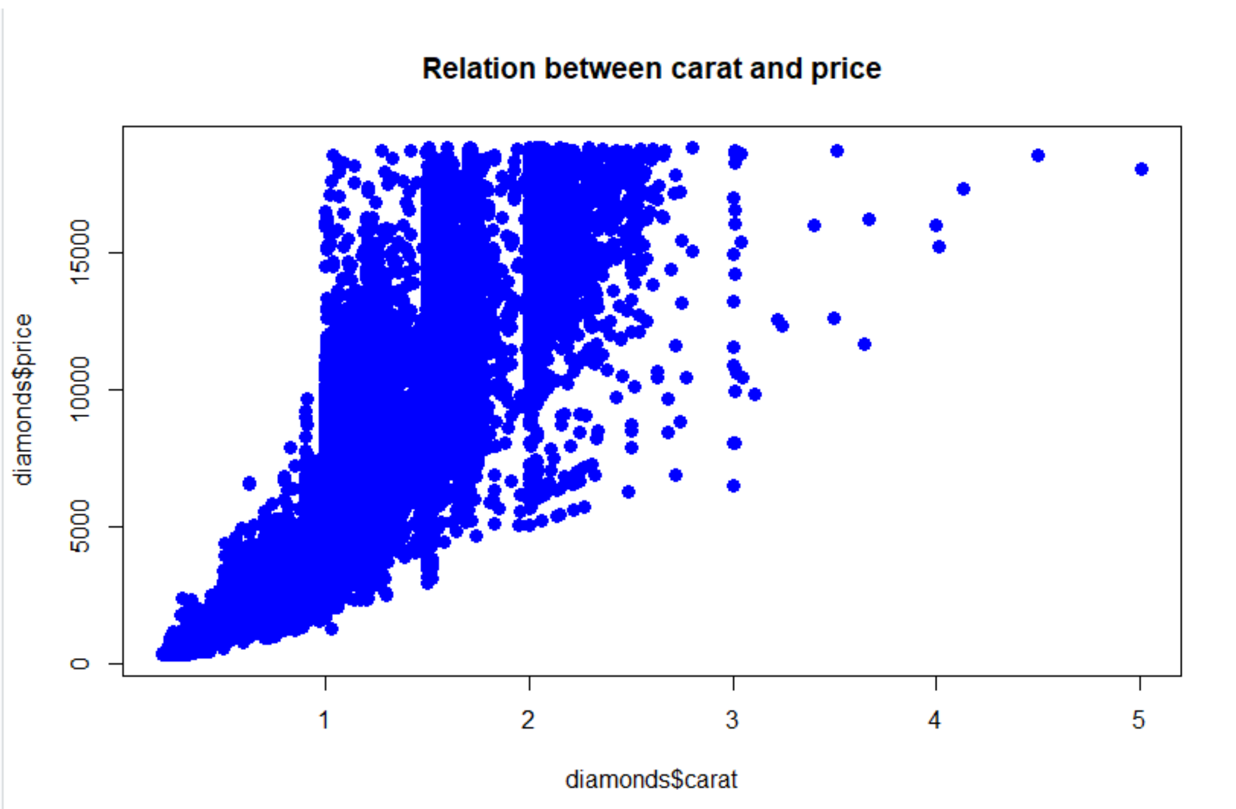
> lm.beta(mlr1)

Call:
lm(formula = log(price) ~ log(carat) + cut + color + clarity,
    data = diamonds)

Standardized Coefficients::
 (Intercept)   log(carat)   cutGood   cutIdeal   cutPremium cutVery Good   colorE   colorF
colorG
  0.00000000   1.08574439   0.02268512   0.07782550   0.05991044   0.04816351 -0.02062403
-0.03557538 -0.06430676

colorH colorI colorJ clarityIF claritySI1 claritySI2 clarityVS1 clarityVS2
clarityVVS1
-0.08930403 -0.11041312 -0.11187382 0.19661260 0.25036980 0.15857227 0.28701508
0.30652140 0.25235069
clarityVVS2
0.27234720

First, we could figure out the relationship between carat and price below. Even though it is not a perfect liner relationship, we could easily find out the diamond price will go up with the diamond's carat increase. We could see when the carat arrived 1, 1.5 or 2, the price will increase rapidly. Combined with the personal experience, if you would like to choose a 1, 1.5 or 2 diamond, the best choice is to choose a little bit smaller one, such as 0.95, 1.48 or 1.96, since you will not figure out the difference by eyes but the prices will be huge difference.

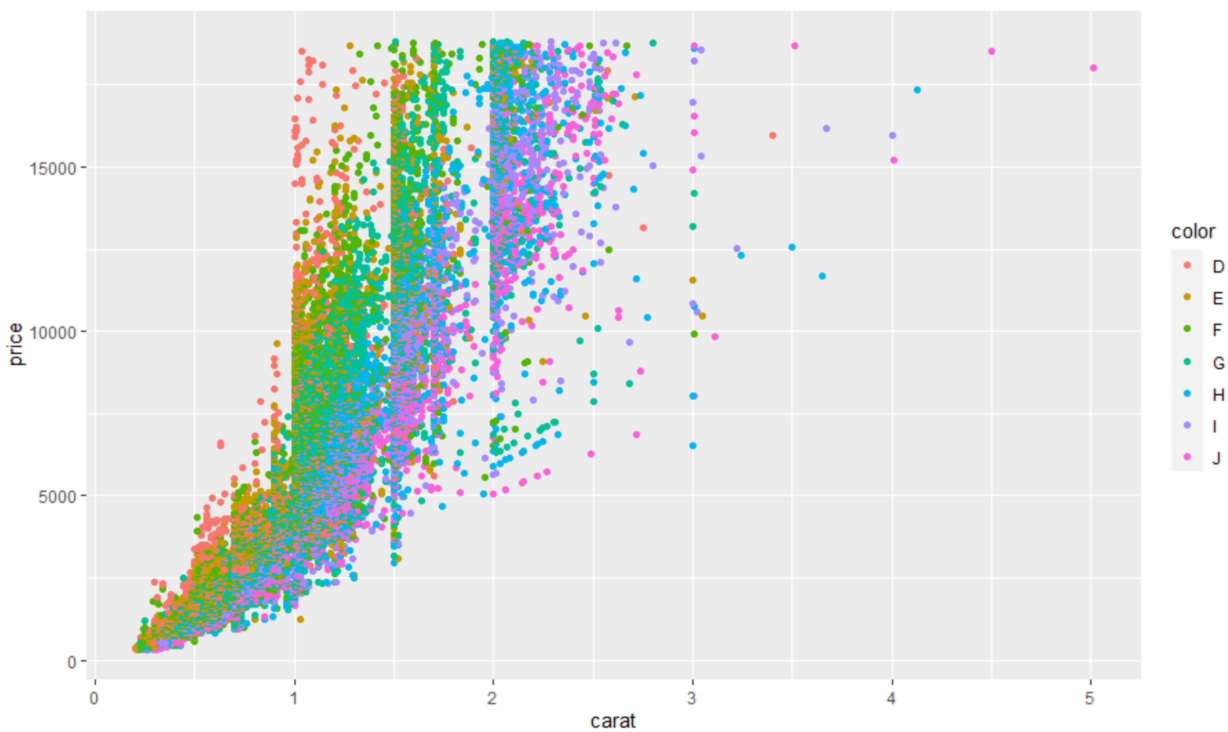


R code:

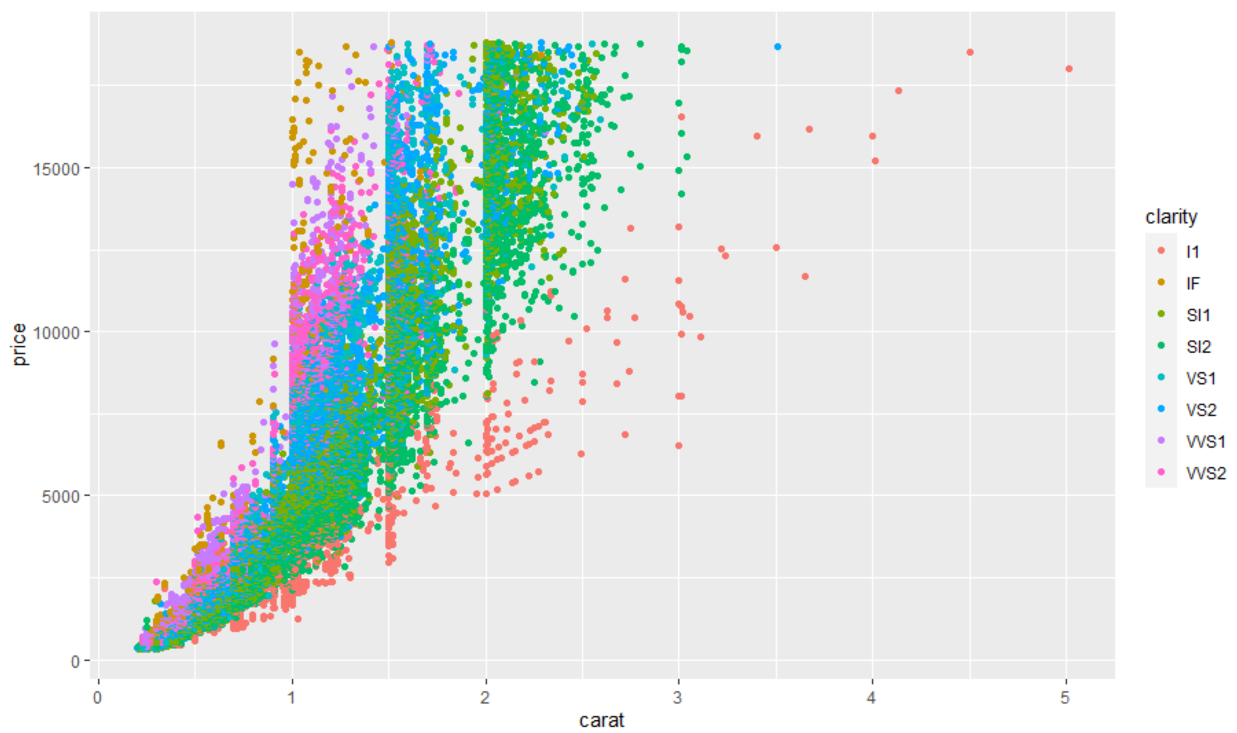
```
plot(diamonds$carat,diamonds$price, pch = 16, cex = 1.3, col = "blue", main = "Relation
between carat and price", )
```

The next three figures showed the relationship between price, carat with color, clarity and cut. We could easily figure out when the carat is constant, color D which is the best color has the highest price and color J, the worst color has the lowest price. Clarity and cut has the same relationship. When the carat does not change, IF, the best color has the highest price and I1, the worst color has the lowest price; the best cut ideal is most expensive compared with the worst quality of cut, fair.

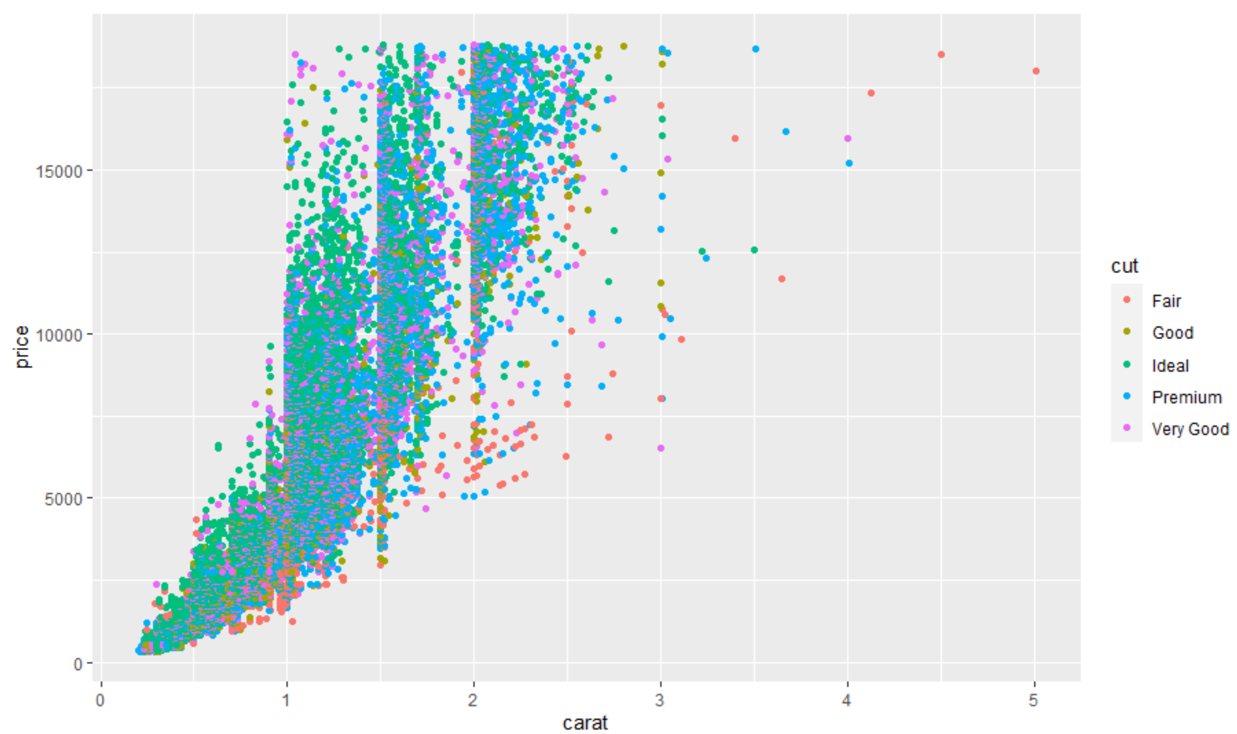
Price, Carat VS Color



Price, Carat VS Clarity



Price, Carat VS Cut



R code:

```
ggplot(diamonds, aes(x=carat,y=price, color=clarity))+ geom_point()
ggplot(diamonds, aes(x=carat,y=price, color=color))+ geom_point()
ggplot(diamonds, aes(x=carat,y=price, color=cut))+ geom_point()
```

Conclusion

Our final comprehensive buying strategy is very subjective. It totally depends on buyers' budget and preference. If buyer has budget is large enough, the larger the carat, the better the color, clarity and cut, the diamond will be better. However, if the buyers' budget is limited, then buyer's preference will decide the best value of the diamond. For example, if buyers prefer to buy a larger diamond, carat will be the most valuable feature and first consideration, but they must give up cut, color and clarity to meet the budget limitation. Same as any C as the buyers' favorite, or maybe buyers could balance all 4Cs to find the most valuable diamond in their hearts.

Appendices

Appendix A: Summary of the Full Model (Model 1)`

```
Call:
lm(formula = price ~ carat + factor(cut) + factor(color) + factor(clarity) +
    depth + table + x + y + z, data = Dat)

Residuals:
     Min       1Q   Median       3Q      Max
-21376.0   -592.4   -183.5    376.4  10694.2

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           2184.477    408.197   5.352 8.76e-08 ***
carat                11256.978     48.628 231.494  < 2e-16 ***
factor(cut)Good        579.751     33.592  17.259  < 2e-16 ***
factor(cut)Ideal       832.912     33.407  24.932  < 2e-16 ***
factor(cut)Premium     762.144     32.228  23.649  < 2e-16 ***
factor(cut)Very Good   726.783     32.241  22.542  < 2e-16 ***
factor(color)E        -209.118     17.893 -11.687  < 2e-16 ***
factor(color)F        -272.854     18.093 -15.081  < 2e-16 ***
factor(color)G        -482.039     17.716 -27.209  < 2e-16 ***
factor(color)H        -980.267     18.836 -52.043  < 2e-16 ***
factor(color)I       -1466.244     21.162 -69.286  < 2e-16 ***
factor(color)J       -2369.398     26.131 -90.674  < 2e-16 ***
factor(clarity)IF     5345.102     51.024 104.757  < 2e-16 ***
factor(clarity)SI1    3665.472     43.634  84.005  < 2e-16 ***
factor(clarity)SI2    2702.586     43.818  61.677  < 2e-16 ***
factor(clarity)VS1    4578.398     44.546 102.779  < 2e-16 ***
factor(clarity)VS2    4267.224     43.853  97.306  < 2e-16 ***
factor(clarity)VVS1   5007.759     47.160 106.187  < 2e-16 ***
factor(clarity)VVS2   4950.814     45.855 107.967  < 2e-16 ***
depth                  -63.806      4.535 -14.071  < 2e-16 ***
table                  -26.474      2.912  -9.092  < 2e-16 ***
x                    -1008.261     32.898 -30.648  < 2e-16 ***
y                        9.609     19.333   0.497    0.619
z                      -50.119     33.486  -1.497    0.134
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1130 on 53916 degrees of freedom
Multiple R-squared:  0.9198,    Adjusted R-squared:  0.9198
F-statistic: 2.688e+04 on 23 and 53916 DF,  p-value: < 2.2e-16
```

Appendix B: Summary of Model 3

```
Call:
lm(formula = log(price) ~ log(carat) + factor(cut) + factor(color) +
    factor(clarity) + depth + table + x + y + z, data = Dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05041 -0.08575 -0.00009  0.08301  1.93916

Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)          7.385e+00  6.005e-02  122.969   <2e-16 ***
log(carat)           1.771e+00  7.870e-03  225.022   <2e-16 ***
factor(cut)Good      7.935e-02  3.969e-03   19.990   <2e-16 ***
factor(cut)Ideal     1.579e-01  3.947e-03   40.002   <2e-16 ***
factor(cut)Premium   1.348e-01  3.810e-03   35.383   <2e-16 ***
factor(cut)Very Good 1.153e-01  3.809e-03   30.262   <2e-16 ***
factor(color)E      -5.460e-02  2.114e-03  -25.820   <2e-16 ***
factor(color)F      -9.441e-02  2.138e-03  -44.157   <2e-16 ***
factor(color)G      -1.607e-01  2.093e-03  -76.739   <2e-16 ***
factor(color)H      -2.527e-01  2.225e-03 -113.570   <2e-16 ***
factor(color)I      -3.754e-01  2.497e-03 -150.355   <2e-16 ***
factor(color)J      -5.147e-01  3.080e-03 -167.078   <2e-16 ***
factor(clarity)IF    1.115e+00  6.029e-03  184.881   <2e-16 ***
factor(clarity)SI1   5.967e-01  5.149e-03  115.882   <2e-16 ***
factor(clarity)SI2   4.302e-01  5.175e-03   83.125   <2e-16 ***
factor(clarity)VS1   8.153e-01  5.258e-03  155.064   <2e-16 ***
factor(clarity)VS2   7.452e-01  5.177e-03  143.949   <2e-16 ***
factor(clarity)VVS1  1.020e+00  5.572e-03  183.077   <2e-16 ***
factor(clarity)VVS2  9.490e-01  5.416e-03  175.240   <2e-16 ***
depth                1.362e-03  5.536e-04    2.460   0.0139 *
table                7.028e-05  3.455e-04    0.203   0.8388
x                    5.685e-02  4.979e-03   11.416   <2e-16 ***
y                   -1.487e-03  2.286e-03   -0.651   0.5154
z                    6.624e-03  3.960e-03    1.673   0.0943 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1335 on 53916 degrees of freedom
Multiple R-squared:  0.9827,    Adjusted R-squared:  0.9827
F-statistic: 1.33e+05 on 23 and 53916 DF,  p-value: < 2.2e-16
```

Appendix C: VIF of Regressor Variables

	GVIF	Df	GVIF^(1/(2*Df))
log(carat)	64.073039	1	8.004564
factor(cut)	1.964391	4	1.088062
factor(color)	1.166590	6	1.012923
factor(clarity)	1.344045	7	1.021345
depth	1.902735	1	1.379397
table	1.803199	1	1.342832
x	94.370322	1	9.714439
y	20.624657	1	4.541438
z	23.617250	1	4.859758

Appendix D: Pearson Correlation Coefficient

	price	carat	depth	table	x	y	z
price	1.0000000	0.92159130	-0.01064740	0.1271339	0.88443516	0.86542090	0.86124944
carat	0.9215913	1.00000000	0.02822431	0.1816175	0.97509423	0.95172220	0.95338738
depth	-0.0106474	0.02822431	1.00000000	-0.2957785	-0.02528925	-0.02934067	0.09492388
table	0.1271339	0.18161755	-0.29577852	1.0000000	0.19534428	0.18376015	0.15092869
x	0.8844352	0.97509423	-0.02528925	0.1953443	1.00000000	0.97470148	0.97077180
y	0.8654209	0.95172220	-0.02934067	0.1837601	0.97470148	1.00000000	0.95200572
z	0.8612494	0.95338738	0.09492388	0.1509287	0.97077180	0.95200572	1.00000000

Appendix E: VIF of Regressor Variables after Dropping "x", "y", "z"

	GVIF	Df	GVIF^(1/(2*Df))
log(carat)	1.326904	1	1.151913
factor(cut)	1.924454	4	1.085272
factor(color)	1.144628	6	1.011320
factor(clarity)	1.330205	7	1.020590
depth	1.378248	1	1.173988
table	1.789348	1	1.337665

Appendix F: Type II ANOVA

```
Anova Table (Type II tests)

Response: log(price)
                 Sum Sq    Df    F value  Pr(>F)    
log(carat)       902.96     1 50634.7096 < 2e-16 ***
factor(cut)       39.94     4   559.9398 < 2e-16 ***
factor(color)    893.70     6  8352.5387 < 2e-16 ***
factor(clarity) 1826.76     7 14634.0291 < 2e-16 ***
depth              0.11     1     6.0515 0.01390 *  
table              0.00     1     0.0414 0.83883    
x                  2.32     1   130.3338 < 2e-16 ***
y                  0.01     1     0.4232 0.51537    
z                  0.05     1     2.7987 0.09434 .  
Residuals        961.48 53916                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix G: Removing Outliers, Leverage and Influential Points

```
integer [215]                15 660 1642 2025 2026 2412 ...
```

```
  [1]    15   660  1642  2025  2026  2412  3276  4051  5079  5217  5245  5326  5822  5858  5931  6140  6349  6418  6440  6721  6736  7111  7150  7303  7735
 [26]  7744  7745  8204  8239  8593  8790  9180  9712 10130 10188 10660 11407 11408 11515 11561 11605 11660 12150 12563 12687 12954 13224 13563 14078 14110
 [51] 14239 14435 14576 14928 15471 15906 15942 16085 16284 16297 16342 16405 16541 16603 17404 17467 17898 19059 19082 19092 19180 19253 19315 19340 19347
 [76] 19554 19604 19867 19922 21139 21725 21863 21887 21936 22143 22287 22400 22429 22441 22494 22525 22701 22732 23410 23540 23645 23775 23878 23940 24132
[101] 24276 24298 24329 24448 24655 24785 24817 24884 24946 25068 25069 25315 25332 25461 25526 25580 25626 25779 25925 25926 25998 25999 26000 26004 26078
[126] 26091 26101 26106 26130 26199 26238 26312 26408 26432 26445 26484 26492 26550 26635 26658 26661 26966 26999 27017 27050 27131 27197 27227 27350 27355
[151] 27416 27456 27458 27508 27531 27636 27674 27680 27835 27923 28711 28712 28713 29549 29946 30165 30722 30805 30948 31809 31996 32525 34044 34612 35874
[176] 36652 36653 36984 37449 38057 38154 39502 39566 40287 40806 41020 41021 41052 41754 41919 43425 43426 43910 43990 45866 46345 46477 47004 47950 48378
[201] 49774 50672 50673 50792 50954 50959 51174 51175 51293 51370 52378 52806 52861 52862 53599
```

Appendix B: Summary of Final Model

```
Call:
lm(formula = log(price) ~ log(carat) + factor(cut) + factor(color) +
    factor(clarity), data = newdat)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.42810 -0.08542 -0.00017  0.08269  0.41739 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           7.869425   0.005691 1382.77   <2e-16 ***
log(carat)            1.886598   0.001093 1725.48   <2e-16 ***
factor(cut)Good       0.088245   0.003775   23.38   <2e-16 ***
factor(cut)Ideal      0.170505   0.003446   49.48   <2e-16 ***
factor(cut)Premium    0.148842   0.003476   42.82   <2e-16 ***
factor(cut)Very Good  0.125335   0.003514   35.67   <2e-16 ***
factor(color)E       -0.051362   0.002050  -25.06   <2e-16 ***
factor(color)F       -0.092831   0.002074  -44.77   <2e-16 ***
factor(color)G       -0.157825   0.002031  -77.73   <2e-16 ***
factor(color)H       -0.247643   0.002154 -114.97   <2e-16 ***
factor(color)I       -0.370782   0.002411 -153.78   <2e-16 ***
factor(color)J       -0.509736   0.002975 -171.34   <2e-16 ***
factor(clarity)IF     1.081322   0.005951  181.72   <2e-16 ***
factor(clarity)SI1    0.570303   0.005105  111.71   <2e-16 ***
factor(clarity)SI2    0.405671   0.005135   79.00   <2e-16 ***
factor(clarity)VS1    0.790374   0.005205  151.85   <2e-16 ***
factor(clarity)VS2    0.719612   0.005132  140.23   <2e-16 ***
factor(clarity)VVS1   0.996315   0.005503  181.04   <2e-16 ***
factor(clarity)VVS2   0.924772   0.005357  172.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.129 on 53706 degrees of freedom
Multiple R-squared:  0.9838,    Adjusted R-squared:  0.9838 
F-statistic: 1.813e+05 on 18 and 53706 DF,  p-value: < 2.2e-16
```

Figures

Figure 1

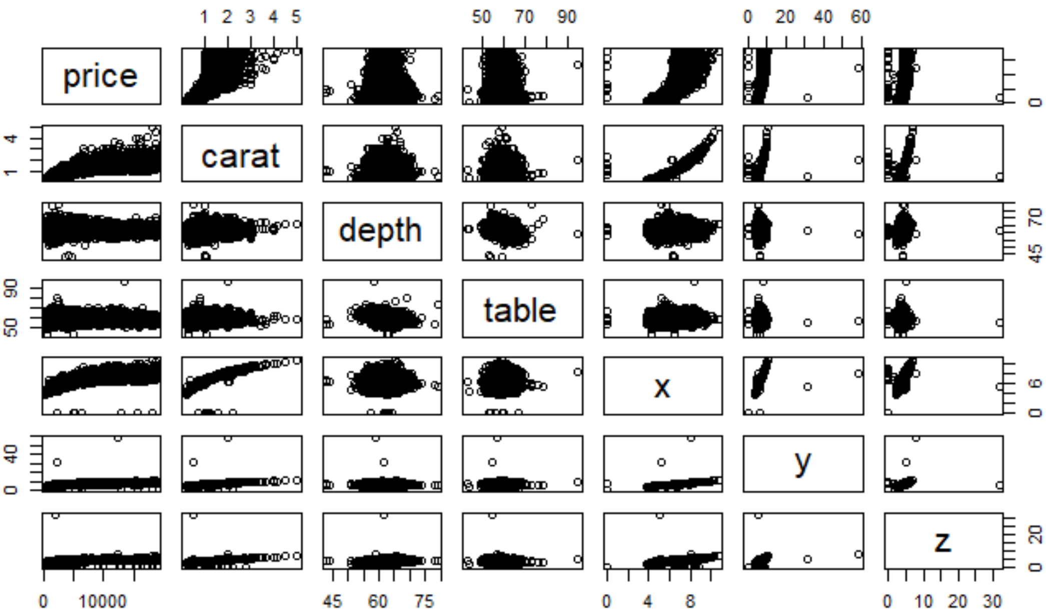


Figure 2

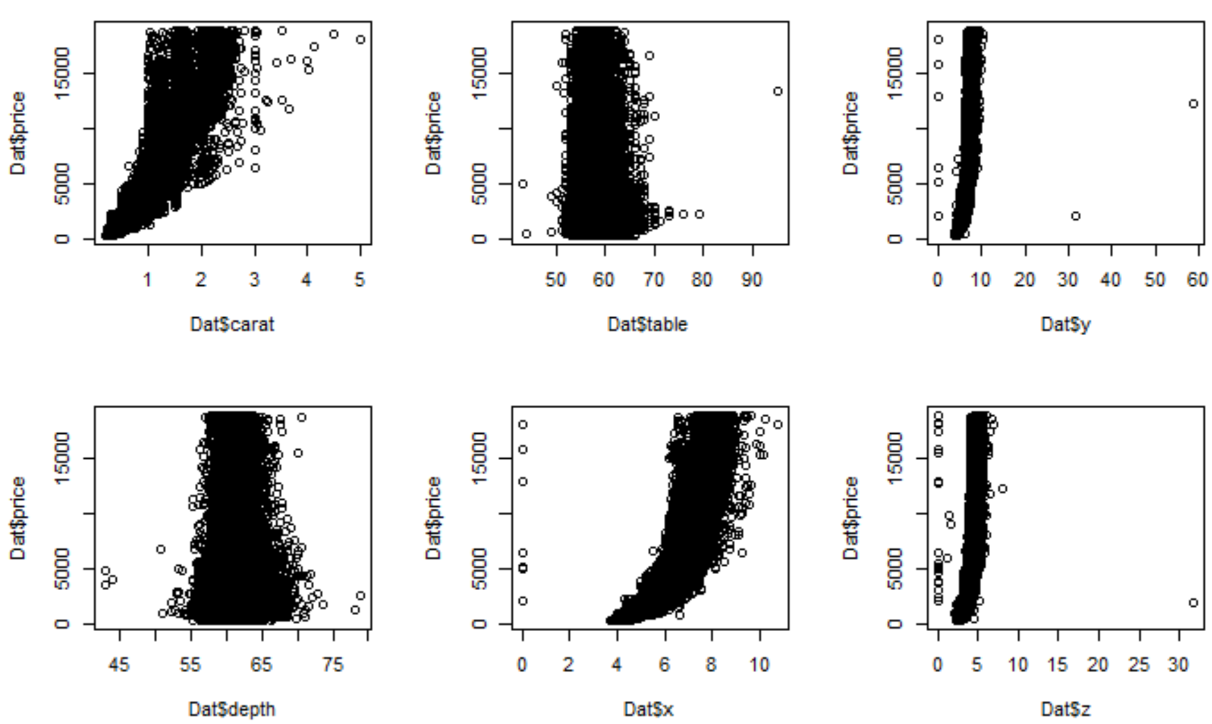


Figure 3

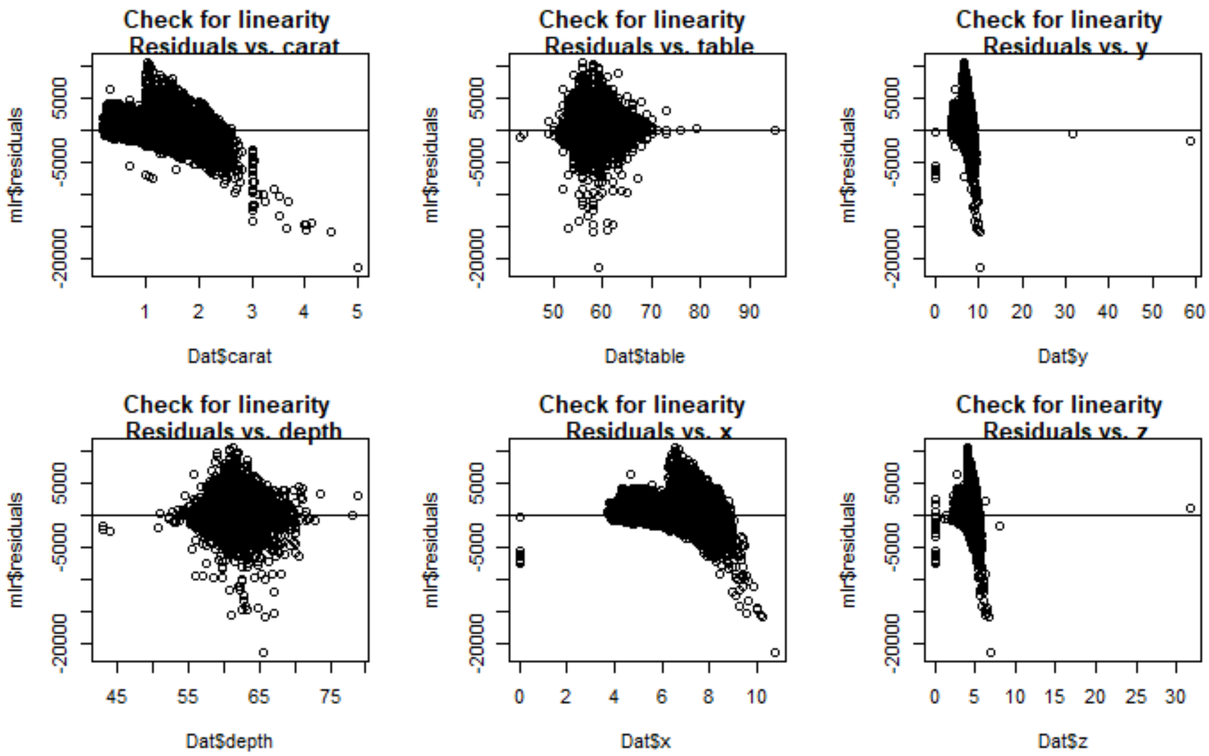
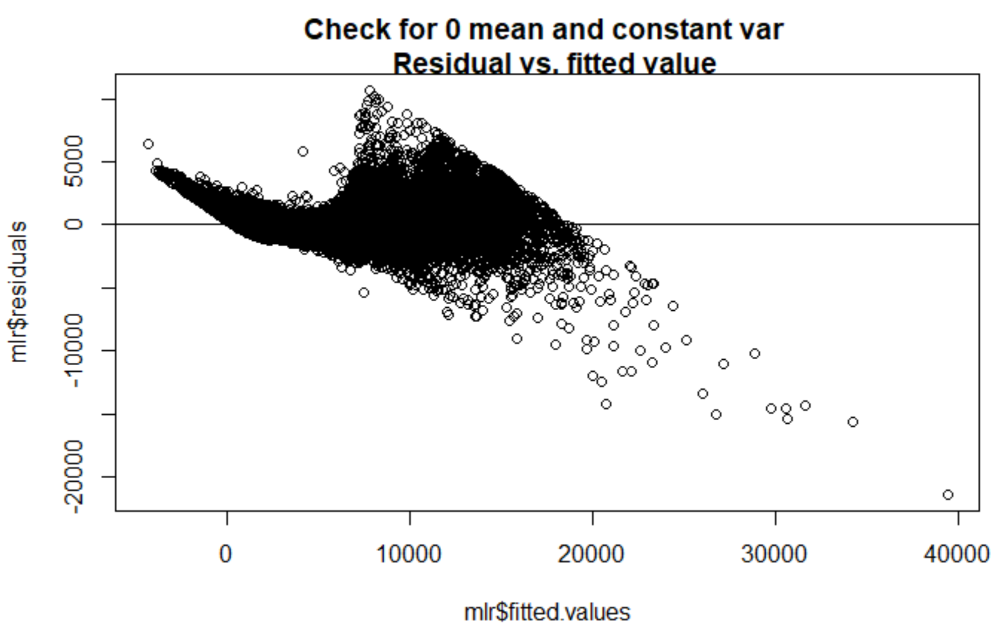


Figure 4



1.6

Figure 5

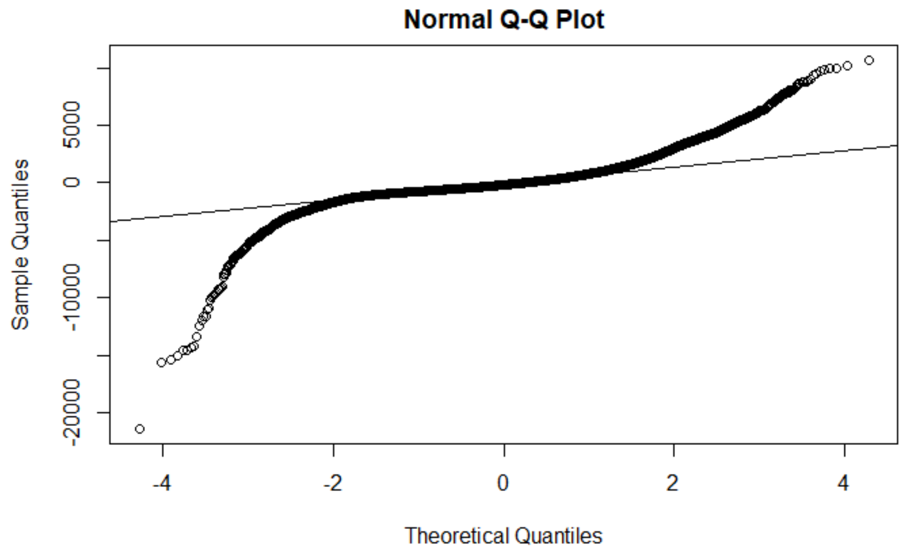


Figure 6

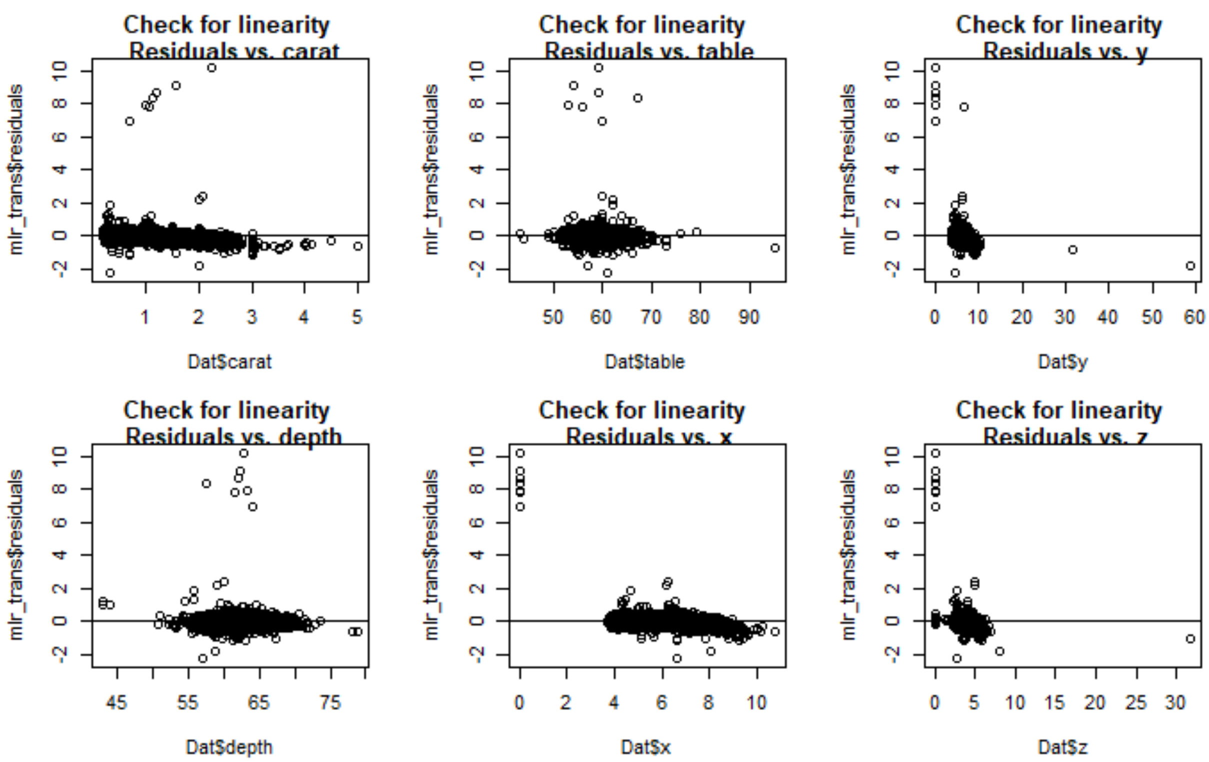


Figure 7

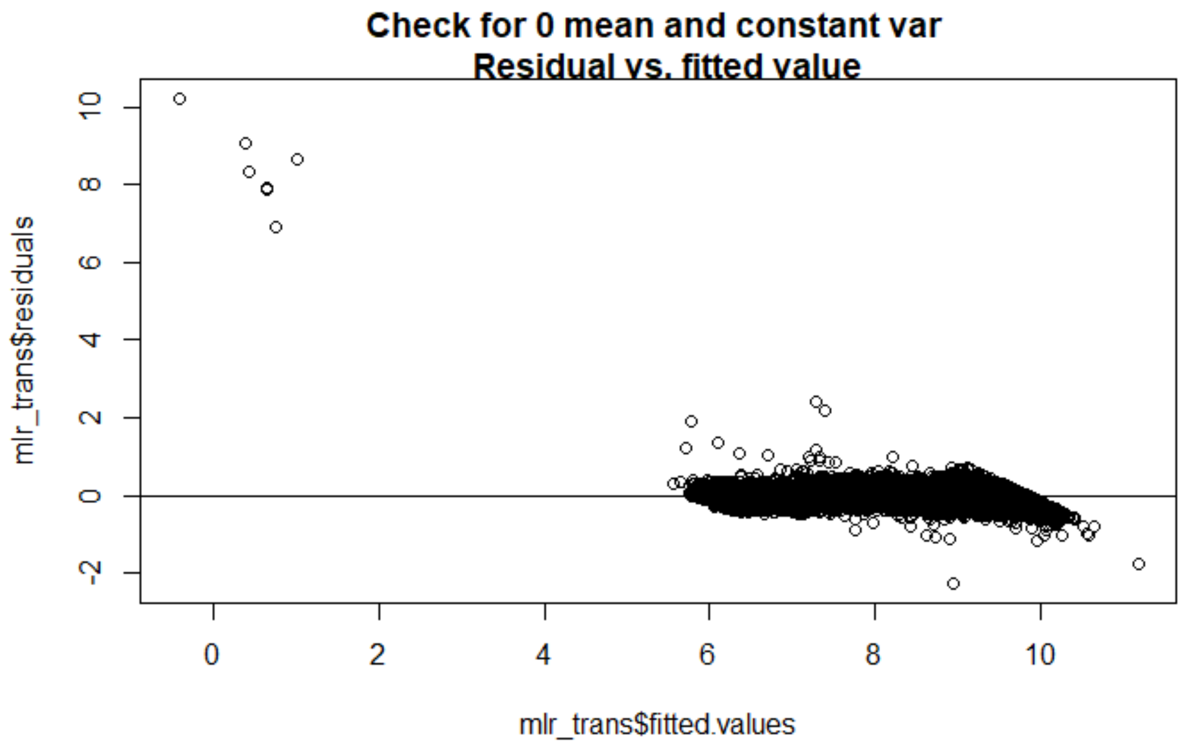


Figure 8

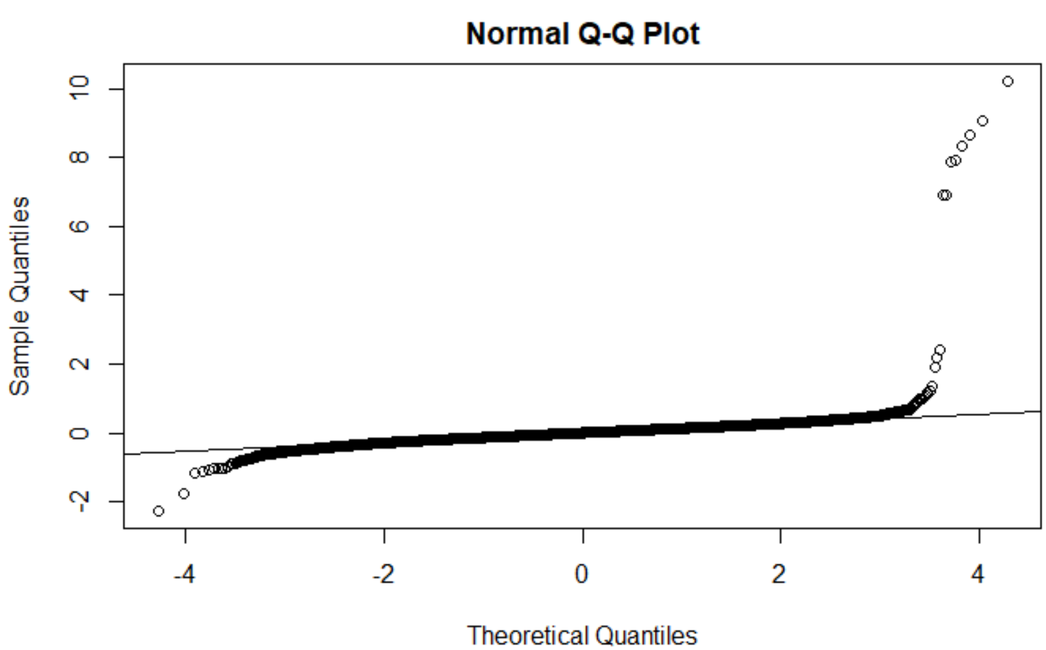


Figure 9

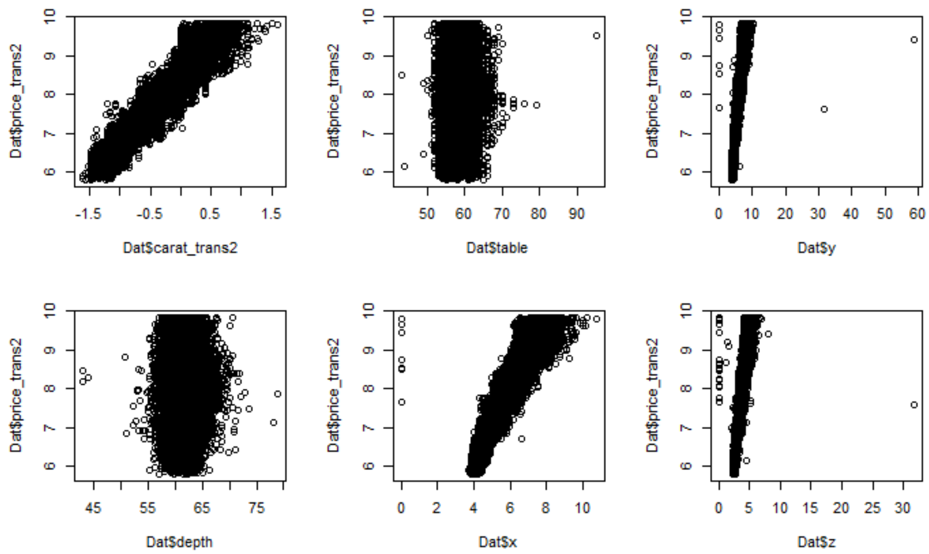
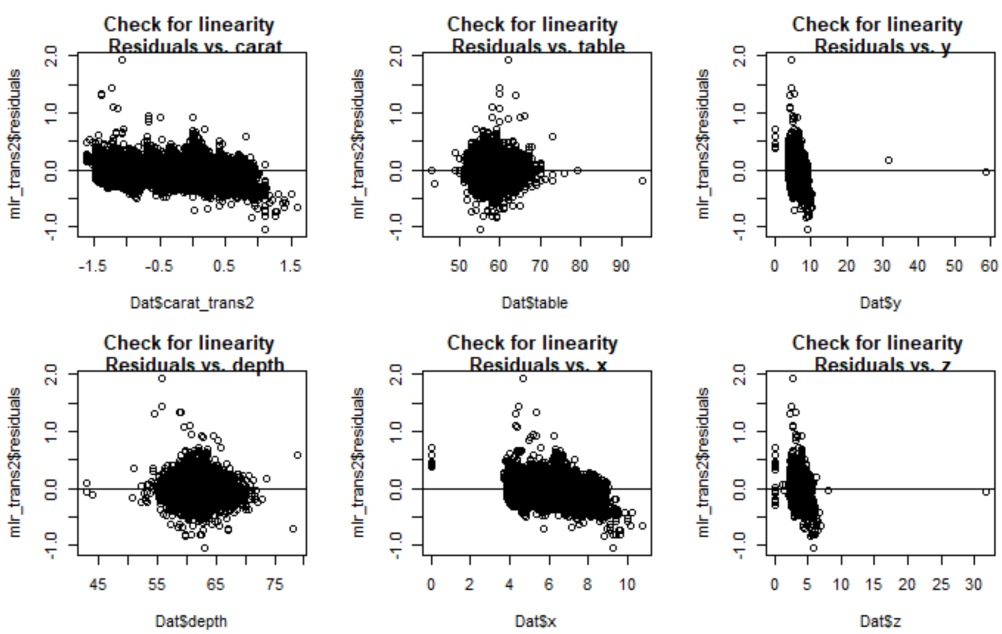


Figure 10

3.5



3.6

Figure 11

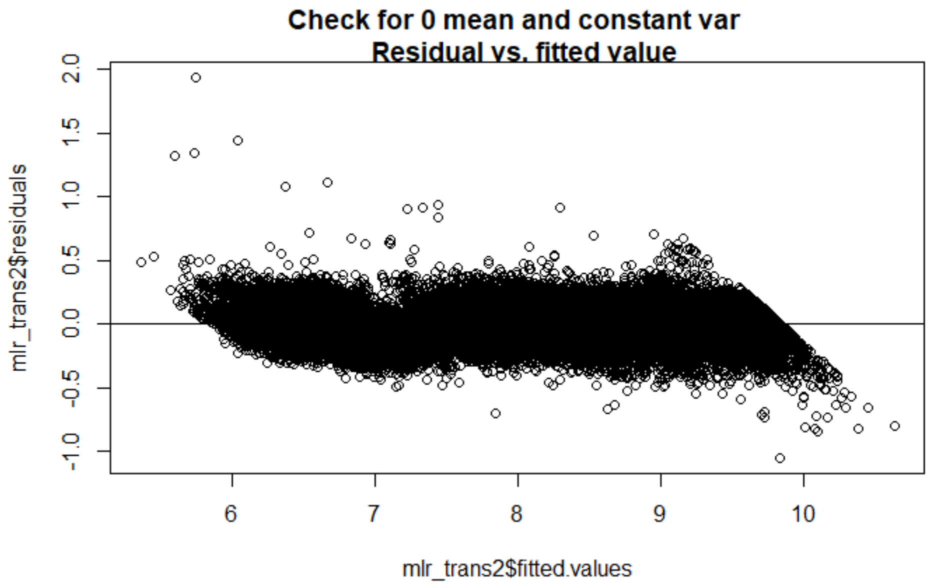


Figure 12

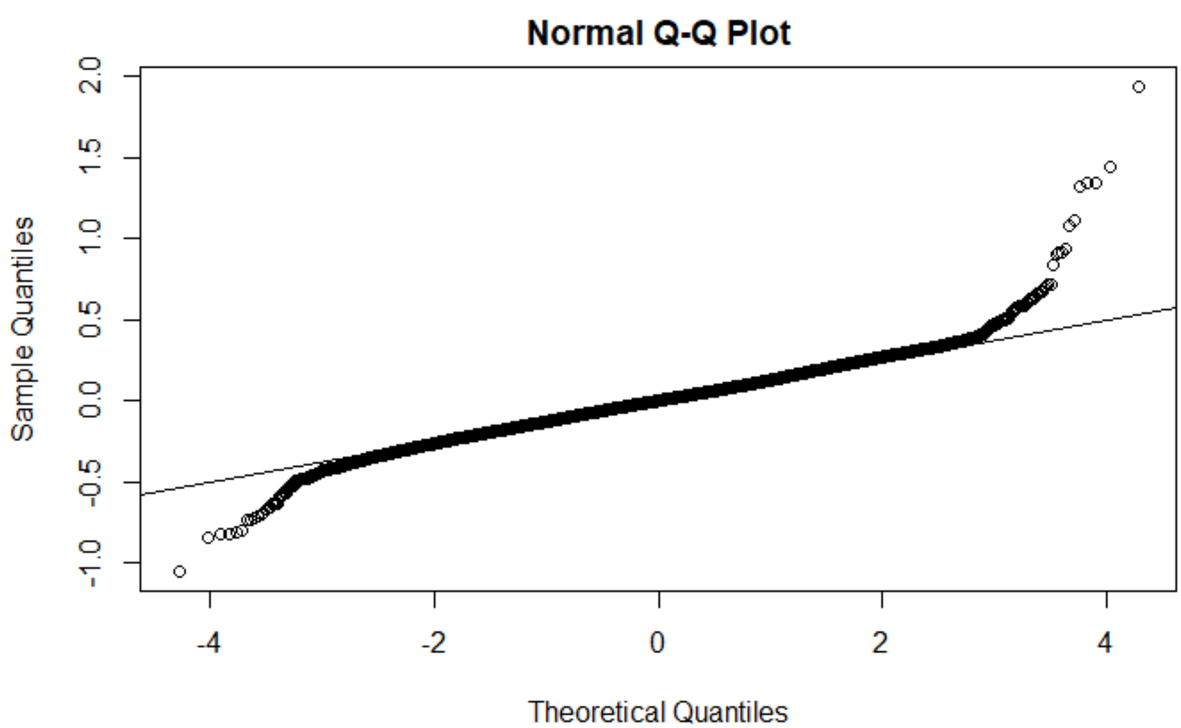


Figure 13

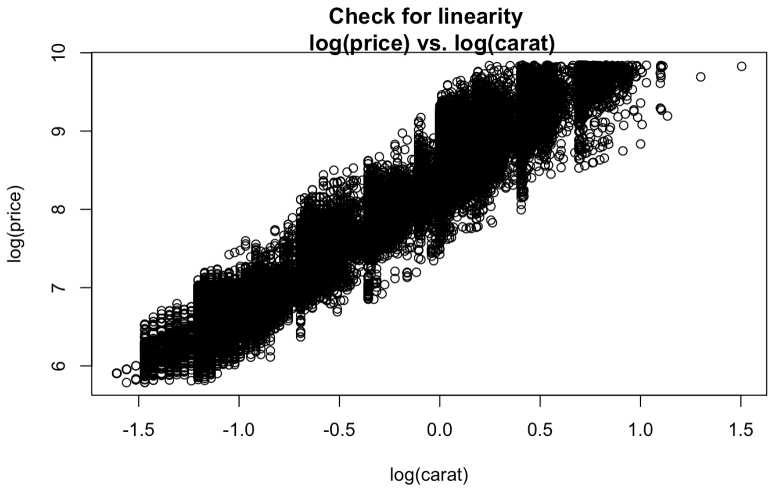


Figure 14

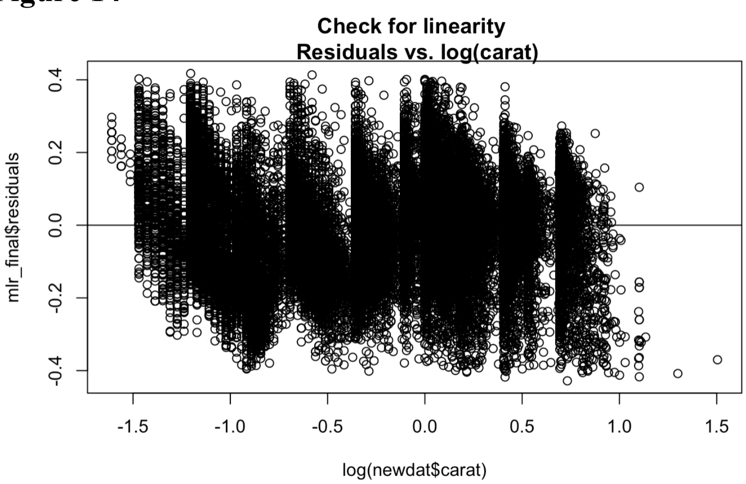


Figure 15

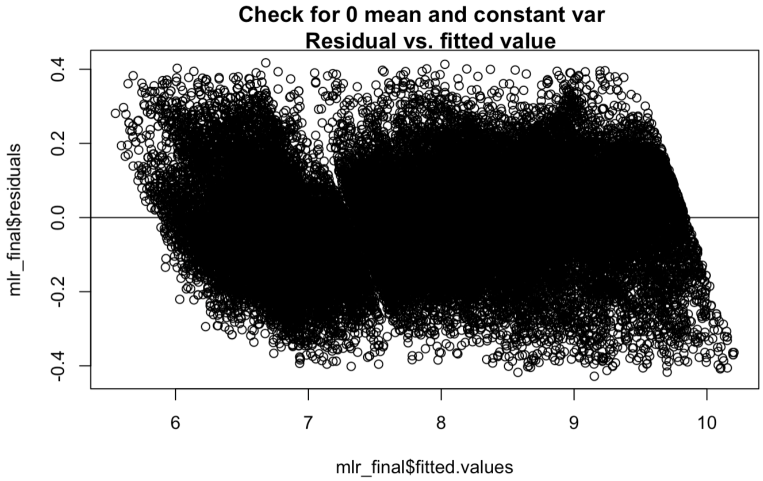
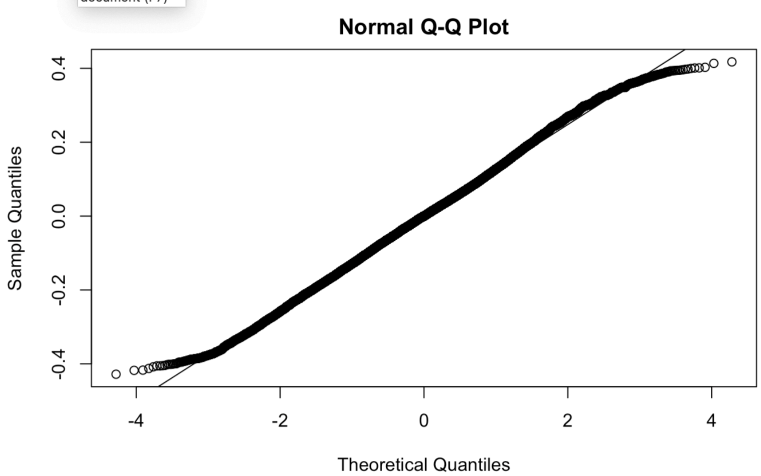


Figure 16



References

Agrawal, S. (2017, May 25). *Diamonds*. Kaggle. Retrieved December 9, 2021, from https://www.kaggle.com/shivam2503/diamonds.