ISyE 6414 Final Project Report - Team 13 Gemstone Price Regression

Shuning Tao, Tingyu Liu, Xiaoai Zhu, Yuqi Han

1 Introduction

Gemstone collections are appealing due to their portability, rich histories, potential value, substantial financial returns[1], and symbolic meaning, reflecting the purity of love. Therefore, gemstones are highly sought after in the investment and consumer markets.

This project is part of a Kaggle playground competition series. We aim to assist a gemstone company in increasing business profits by predicting price, distinguishing between more and less profitable gemstones, and identifying the most profitable features.

2 Problem Statement

2.1 Problem statement

Developing a system for gemstone price prediction can be challenging for three main reasons: Large variations in multiple characteristics, Subjectivity, and Market inefficiency.

- Large variations in multiple characteristics. The vast variety in the appearances, sizes, dimensions, and features of gemstones makes the prediction process complex and potentially inaccurate[1].
- Subjectivity. Additionally, marketing gemstones requires specific skills. Pricing a gemstone involves a knowledge battle between the seller and buyer. The price can also depend on whether a buyer views the gemstone in person and inspects it under various lighting conditions.[3]
- Market inefficiency. The gemstone market's inefficiency is another challenge, particularly due to the regulations governing gemstone transactions.[3]

2.2 Variables

To improve the prediction of gemstone prices, we have a tabular dataset containing key attributes of gemstones. The dataset was generated from a deep learning model trained on real-world data. We will predict gemstone prices based on the details given in the dataset so that we can distinguish between higher and lower profitable stones and have a better profit share.

The dependent variable is the gemstone price in dollars, while the independent variables are carat, cut, color, clarity, depth, table, and XYZ dimension.

3 Data Description

3.1 Data source

The data and data dictionary are downloaded from Kaggle, and the direct URL is Regression with a Tabular Gemstone Price Dataset | Kaggle. The data is generated from a deep learning model for gemstone price prediction. It is time-independent so we did not include a time element.

3.2 Data description

The data we will use consists of around 1,000 entries and 10 variables in total. Variables include dependent variable price and independent variables: carat, cut, color, clarity, depth, table, and x,y, and z dimensions. Cut, color, clarity, and carat are widely used measurements, “4Cs”, for the best value of gemstones[7]. Depth and table are height and width ratio indicators for gemstones, while x,y, and z are lengths of gemstones in three dimensions with mm as units.

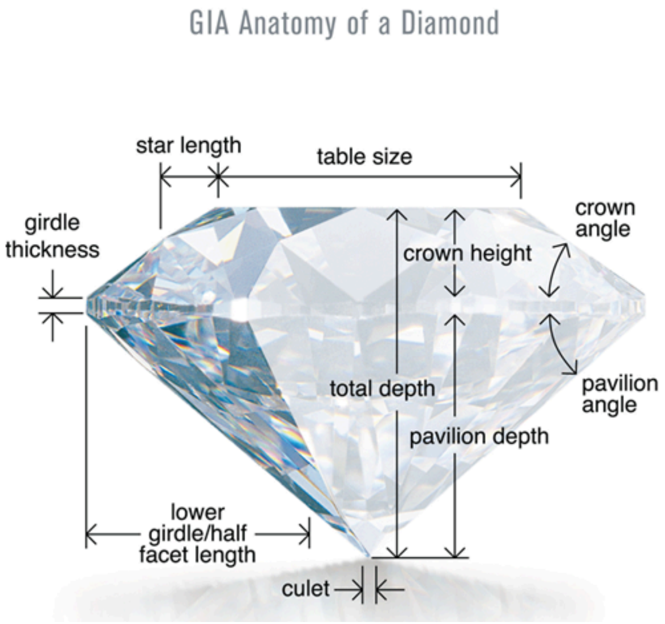


Fig.1 Anatomy of a gemstone
Source: [6]

The variable description is listed as follows. (Full description see Appendix)

Variable	Description	Type
Carat	Carat weight	quantitative
Cut	Describe the cut quality	categorical
Color	The color of the gemstone	categorical
Clarity	The absence of the Inclusions and Blemishes.	categorical

Depth	The height measured from the Culet to the table is divided by its average Girdle Diameter.	quantitative
Table	The width of the Table is expressed as a Percentage of its Average Diameter.	quantitative
X	Length in mm.	quantitative
Y	Width in mm.	quantitative
Z	Height in mm.	quantitative

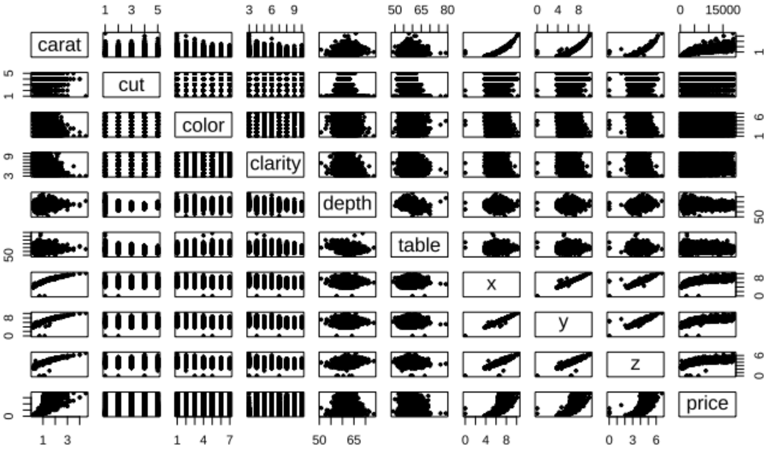
Table 1 Description of Independent Variables

3.3 Data processing

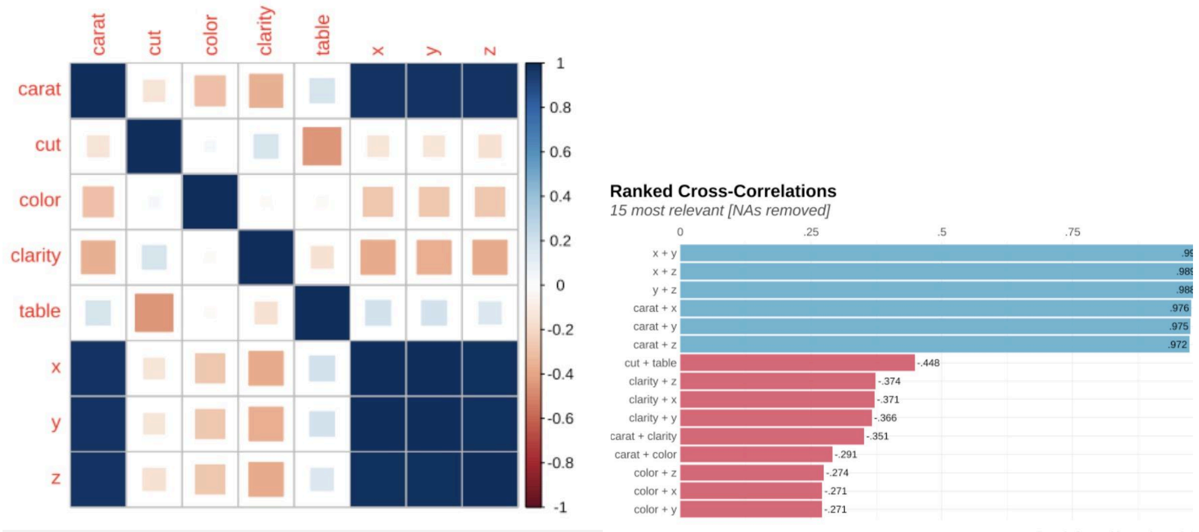
In gemology, the "4Cs" measure the best value of gemstones: cut, color, clarity, and carat. While the carat is quantitative, the other three are categorical and qualitative. We converted these qualitative variables into numerical values using numerical mapping. The ascending order represents the value level of these attributes. For instance, the cut level includes "Fair, Good, Very Good, Premium, Ideal", which we numerically mapped from 1 to 5. The depth variable has more than 20% missing values, so we drop this attribute instead of filing with the mean value.

4 Analyses

4.1 Correlation Analysis and Initial Modeling



Plot 1 Plots of independent and dependent variables

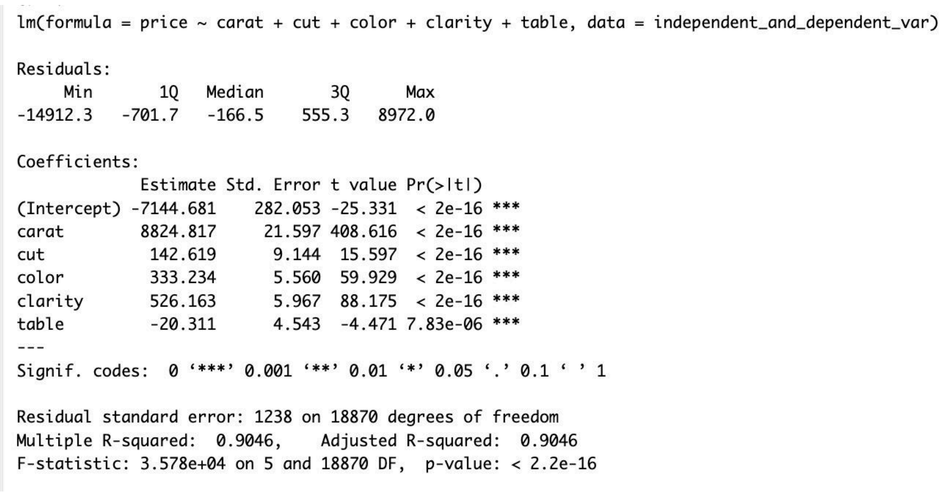


Plot 2 Correlation analysis

In plots of independent and dependent variables, we observe a strong positive relationship between carat and price, suggesting that as the independent variable carat increases, the gemstone price increases proportionally, which is ideal for linear regression models. However, we observe non-linearity in cut, color, clarity, and table. We can observe a strong correlation between Carat and dimension XYZ.

In the initial model selection, we build a Single Linear Regression model with price as the dependent variable and carat, cut, color, clarity, and table as independent variables.

$$price = \beta_0 + \beta_1 * carat + \beta_2 * cut + \beta_3 * color + \beta_4 * clarity + \beta_5 * table$$

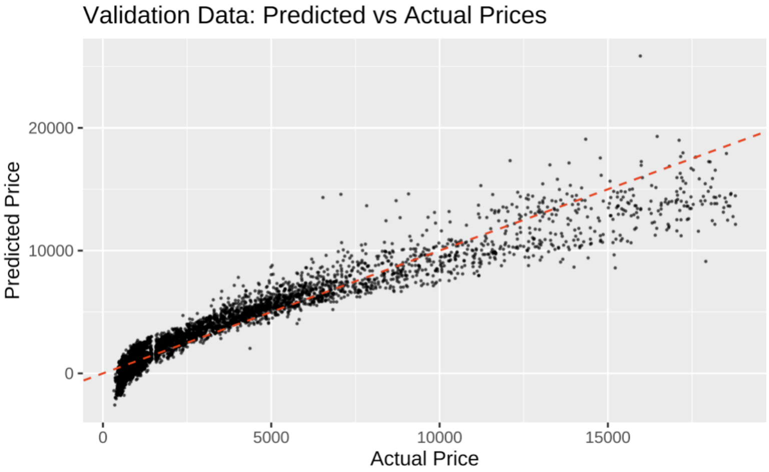


```
lm(formula = price ~ carat + cut + color + clarity + table, data = independent_and_dependent_var)

Residuals:
     Min       1Q   Median       3Q      Max 
-14912.3   -701.7   -166.5    555.3   8972.0 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7144.681    282.053 -25.331  < 2e-16 ***
carat        8824.817     21.597 408.616  < 2e-16 ***
cut           142.619      9.144  15.597  < 2e-16 ***
color         333.234      5.560  59.929  < 2e-16 ***
clarity       526.163      5.967  88.175  < 2e-16 ***
table         -20.311      4.543  -4.471 7.83e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1238 on 18870 degrees of freedom
Multiple R-squared:  0.9046,	Adjusted R-squared:  0.9046 
F-statistic: 3.578e+04 on 5 and 18870 DF,  p-value: < 2.2e-16
```

Fig 2 Initial Model

4.2 Model-Fitting Techniques

Although in our initial SLR, we have an R-squared over 0.90, we still can make some improvements to our model based on the validation result. The scatter plot below displays the actual prices against the predicted prices from the model. Ideally, if a model makes perfect predictions, all points will lie on the dashed red line. The plot shows a general trend where predictions align with actual values but deviate as the price increases, which could indicate the model's predictions are less accurate for higher-priced items. There is a notable spread and a systematic pattern in the residuals, which suggests that the model could be improved, possibly due to non-linear relationships not captured by the current model.

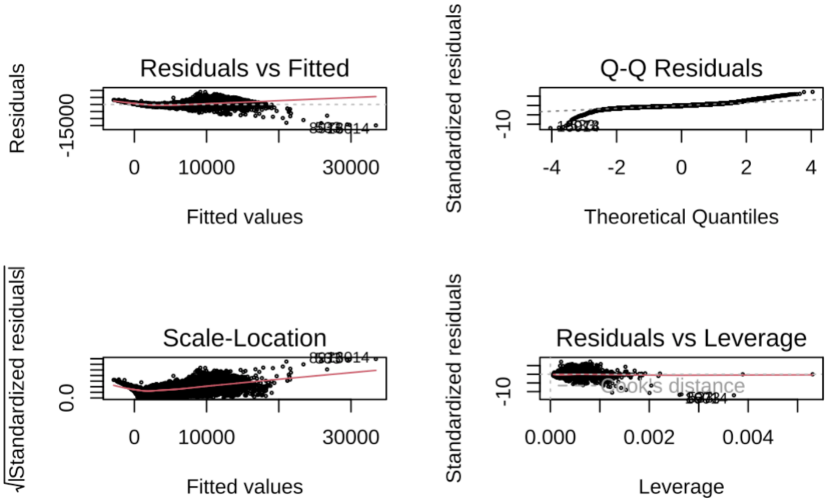


Plot 3 Validation

4.3 Diagnostics and Residual Analysis

The diagnostics and residual analysis of a regression model involves evaluating several diagnostic plots to understand the behavior of the residuals, which can reveal issues with the model such as non-linearity, heteroscedasticity, and influential outliers.

- Residuals vs Fitted: In the plot, there appears to be a pattern, with residuals fanning out as the fitted values increase, indicating potential heteroscedasticity and non-linearity in the model.
- Q-Q (Quantile-Quantile) Residuals: The Q-Q plot indicates that while the middle quantiles seem to align well with the line, the tails deviate significantly, suggesting that the residuals have heavier tails than expected under normality.
- Scale-Location (or Spread-Location): In this plot, the spread of residuals should be consistent across all levels of fitted values. The red line should be horizontal and at a constant level. However, the red line curves upwards, which means the variability of the residuals is increasing with the fitted values, a sign of heteroscedasticity.
- Residuals vs Leverage: This plot helps identify influential cases, that is, observations that have an undue influence on the model's predictions. We will discuss outliers in the later part.



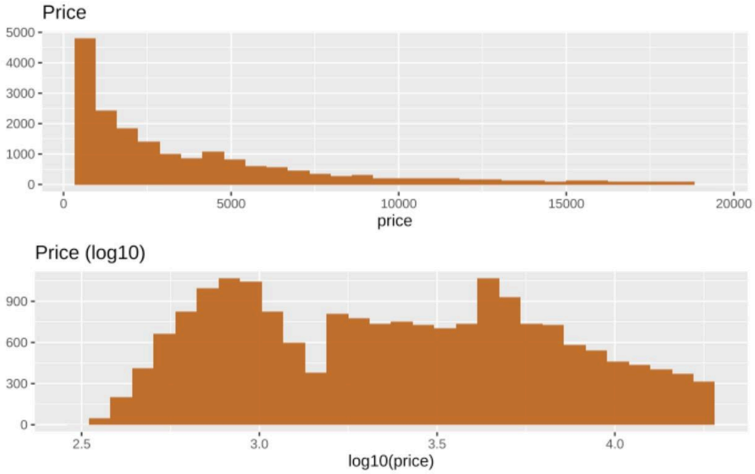
Plot 4 Residual Analysis

In conclusion, the diagnostic plots suggest that there are several potential issues with the initial linear model, including non-linearity, heteroscedasticity, and the presence of influential observations. These findings validate the need for transformations and potential removal or weighting of influential points to improve the model's assumptions and predictive power.

4.4 Model Modifications/Transformations

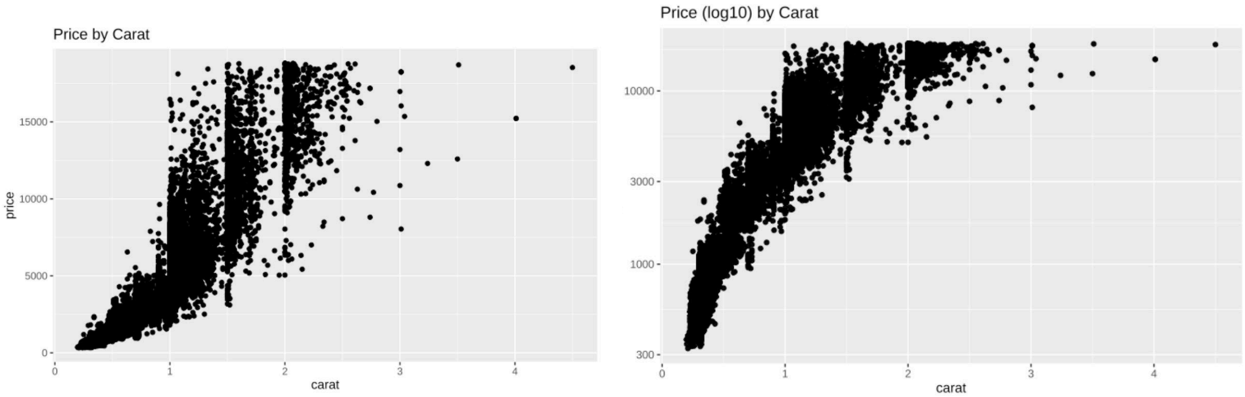
4.4.1 Transformation

Based on the histogram of our price variable, we see a right-skewed distribution, implying the presence of outliers with high prices and a concentration of data points at the lower end of the price spectrum. Such skewness is problematic for linear regression which assumes that the dependent variable is normally distributed. To address the skewness, we applied a logarithmic transformation to 'price', which is a common method for stabilizing variance and making the distribution more symmetric. The log-transformed histogram of 'price' displayed a more bell-shaped distribution, suggesting an improvement towards normality, which is beneficial for the regression analysis.



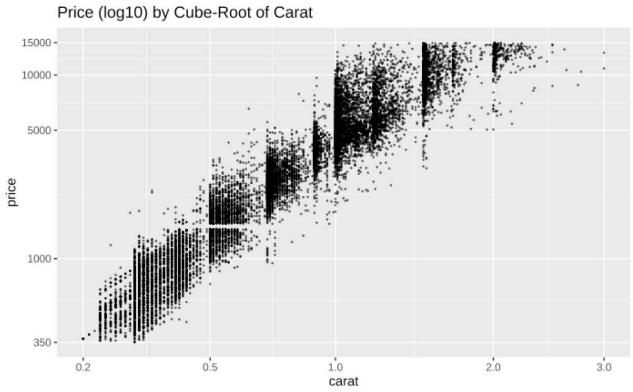
Plot 5 Price transformation

A scatter plot of 'price' against 'carat' was plotted to assess the form of their relationship. It showed a clear non-linear pattern, with a variance in 'price' that increased with 'carat', indicating potential heteroscedasticity. We also plotted the log-transformed 'price' against 'carat'. The resulting plot showed improvements with a trend that was more linear than before but still suggested some residual non-linearity.



Plot 6 Price and Carat

We then proceeded with a dual transformation approach by applying the cube-root transformation to 'carat' in addition to the log transformation on 'price'. The scatter plot of the log-transformed 'price' by the cube root of 'carat' exhibited a markedly linear relationship. This transformation seemed to correct both the non-linearity and heteroscedasticity issues, aligning with the assumptions of linear regression.

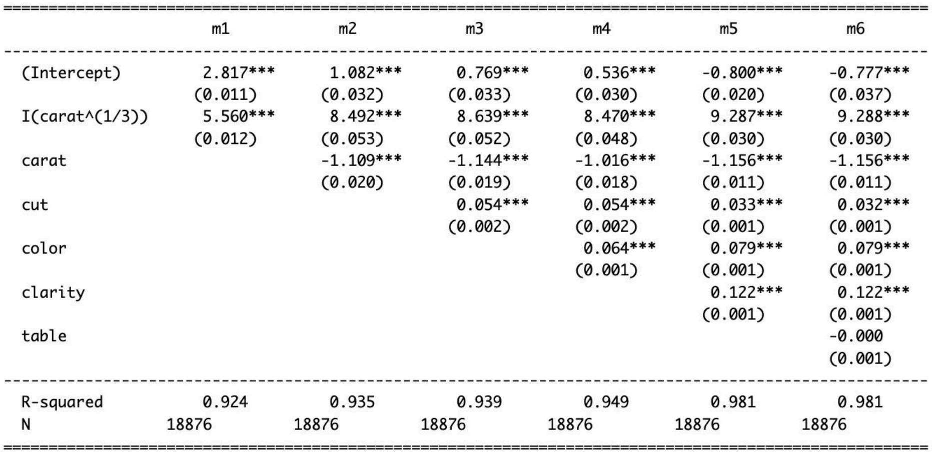


Plot 7 Price(log) and Carat(cube-root)

4.4.2 Modifying the Model

Several models (m1 to m6) were developed, each with one more change aimed at enhancing model performance. This iterative approach allowed us to assess the impact of each modification on the overall model fit.

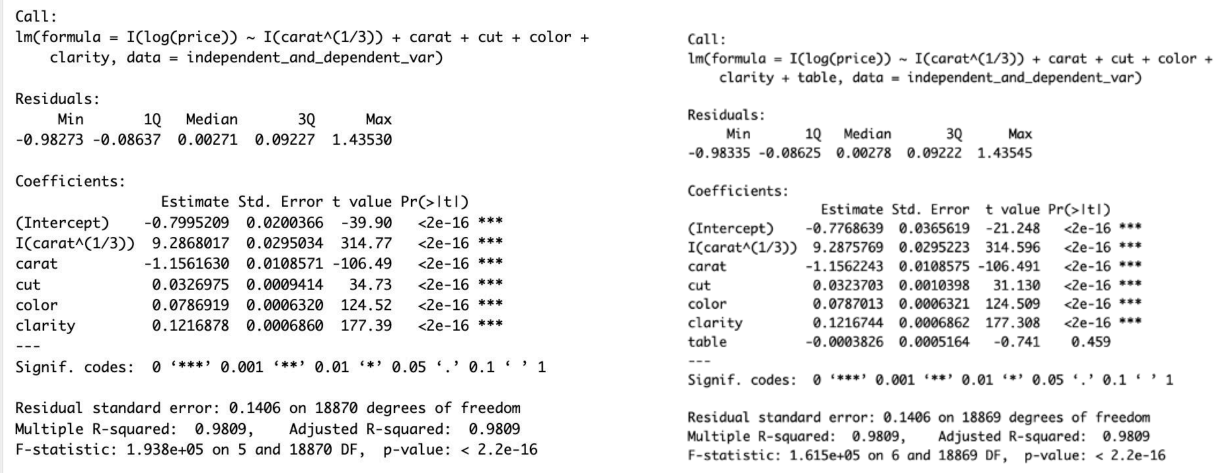
We transitioned to using the natural logarithm of 'price' as the dependent variable to mitigate the effects of skewness and stabilize variance. Moreover, to address the issue of non-linearity between 'price' and 'carat', we introduced the cube root of 'carat' as a new independent variable. The m5 variant emerged as the optimal model, distinguished by its New R-squared value of 0.9809. This metric denotes that 98.09% of the variability in the logarithm of the 'price' is explained by the model, reflecting a significant improvement from the initial models. Across the m5 model, all independent variables displayed highly significant p-values, indicating a robust statistical relevance in predicting the log-transformed 'price'.



```
==============================================================================================
                       m1          m2          m3          m4          m5          m6
----------------------------------------------------------------------------------------------
(Intercept)          2.817***    1.082***    0.769***    0.536***   -0.800***   -0.777***
                    (0.011)     (0.032)     (0.033)     (0.030)     (0.020)     (0.037)
I(carat^(1/3))       5.560***    8.492***    8.639***    8.470***    9.287***    9.288***
                    (0.012)     (0.053)     (0.052)     (0.048)     (0.030)     (0.030)
carat                           -1.109***   -1.144***   -1.016***   -1.156***   -1.156***
                                (0.020)     (0.019)     (0.018)     (0.011)     (0.011)
cut                                          0.054***    0.054***    0.033***    0.032***
                                            (0.002)     (0.002)     (0.001)     (0.001)
color                                                    0.064***    0.079***    0.079***
                                                        (0.001)     (0.001)     (0.001)
clarity                                                              0.122***    0.122***
                                                                    (0.001)     (0.001)
table                                                                           -0.000
                                                                                (0.001)
----------------------------------------------------------------------------------------------
R-squared            0.924       0.935       0.939       0.949       0.981       0.981
N                18876       18876       18876       18876       18876       18876
==============================================================================================
```

Fig 3 New Models

Despite a similar R-squared value, model m6 was excluded from consideration due to the p-value of the variable 'table' exceeding the conventional significance level of 0.05, suggesting its limited contribution to the model.



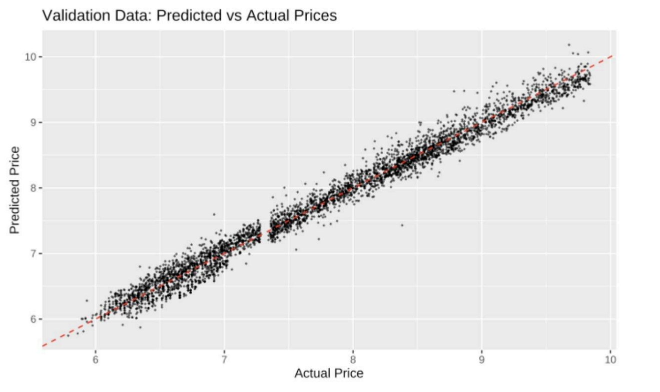
```
Call:
lm(formula = I(log(price)) ~ I(carat^(1/3)) + carat + cut + color +
    clarity, data = independent_and_dependent_var)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98273 -0.08637  0.00271  0.09227  1.43530

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.7995209  0.0200366  -39.90   <2e-16 ***
I(carat^(1/3))  9.2868017  0.0295034  314.77   <2e-16 ***
carat          -1.1561630  0.0108571 -106.49   <2e-16 ***
cut             0.0326975  0.0009414   34.73   <2e-16 ***
color           0.0786919  0.0006320  124.52   <2e-16 ***
clarity         0.1216878  0.0006860  177.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1406 on 18870 degrees of freedom
Multiple R-squared:  0.9809,	Adjusted R-squared:  0.9809
F-statistic: 1.938e+05 on 5 and 18870 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = I(log(price)) ~ I(carat^(1/3)) + carat + cut + color +
    clarity + table, data = independent_and_dependent_var)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98335 -0.08625  0.00278  0.09222  1.43545

Coefficients:
                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)    -0.7768639  0.0365619  -21.248   <2e-16 ***
I(carat^(1/3))  9.2875769  0.0295223  314.596   <2e-16 ***
carat          -1.1562243  0.0108575 -106.491   <2e-16 ***
cut             0.0323703  0.0010398   31.130   <2e-16 ***
color           0.0787013  0.0006321  124.509   <2e-16 ***
clarity         0.1216744  0.0006862  177.308   <2e-16 ***
table          -0.0003826  0.0005164   -0.741    0.459
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1406 on 18869 degrees of freedom
Multiple R-squared:  0.9809,	Adjusted R-squared:  0.9809
F-statistic: 1.615e+05 on 6 and 18869 DF,  p-value: < 2.2e-16
```

Fig 4 Model 5 (left) and Model 6 (right)

$$\log{(price)} = \beta_0 + \beta_1 * \sqrt[3]{carat} + \beta_2 * carat + \beta_3 * cut + \beta_4 * color + \beta_5 * clarity$$

We now present the validation results for model m5, which incorporates log-transformed prices and cube-root transformed carat sizes. The validation plot illustrates the predicted prices (obtained from the m5 model) versus the actual prices. The points closely follow the dashed red line, which represents the line of perfect prediction, where predicted prices exactly match the actual prices. The data points form a dense cloud around this line, indicating a strong linear relationship between the predicted and actual values, with few outliers.



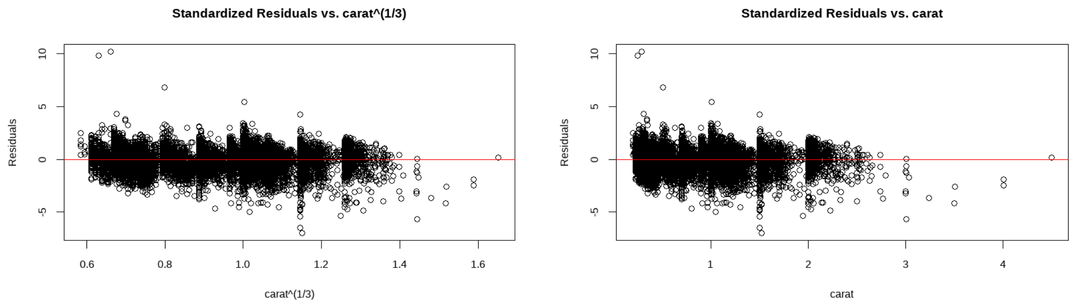
Plot 8 New Validation

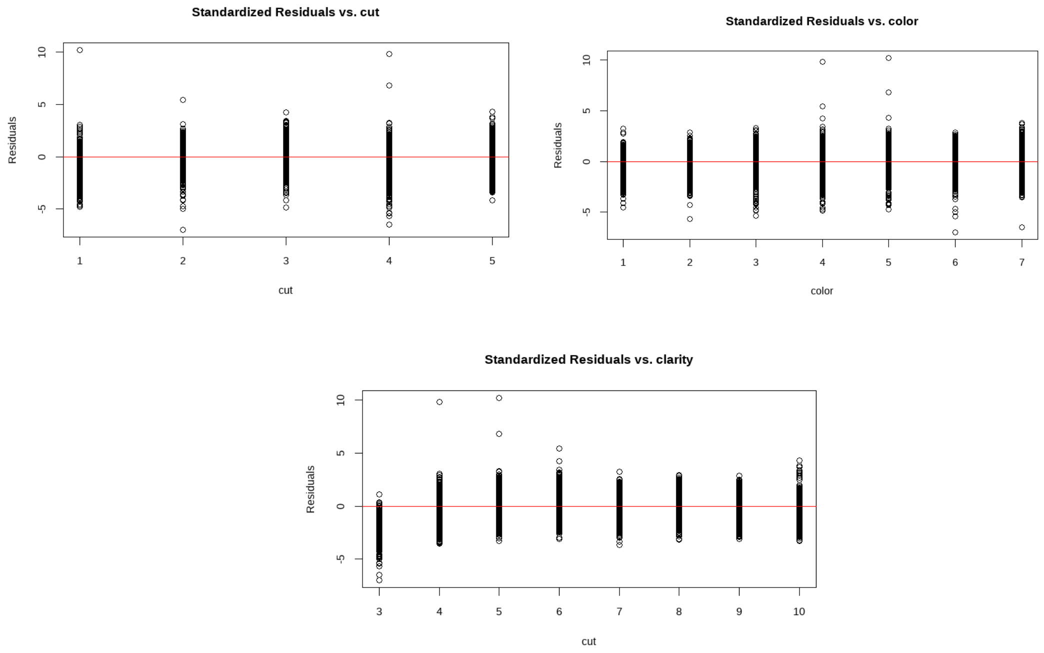
The modified regression model with transformed variables demonstrates a superior fit and predictive accuracy, as evidenced by the high R-squared value, the significance of the coefficients, and the validation result. It effectively addresses the initial concerns of non-linearity and heteroscedasticity, laying the foundation for reliable predictions.

4.5 Assumption Checks and Tests

4.5.1 Linearity Assumption

The linearity assumption tests whether the relationship between the independent variables (predictors) and the dependent variable is linear. Here, we standardized the residuals and plotted the standardized residuals against each predictor. According to Plot 9,the residuals scatter randomly around the zero line without any clear pattern, which indicates that the relationship between the predictors and the residuals is linear. Therefore, the linearity assumption holds.

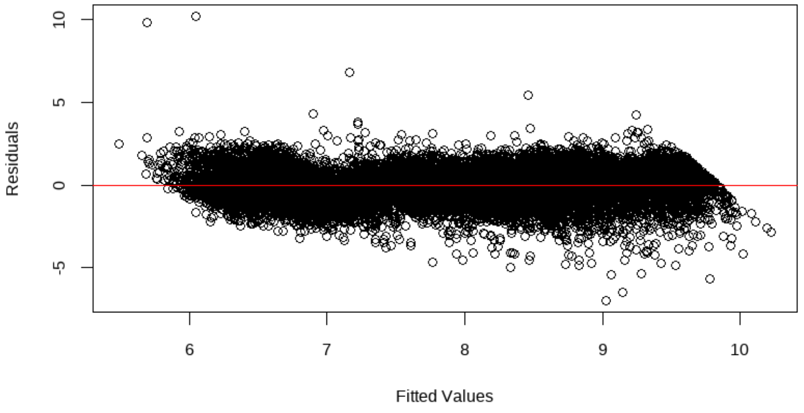




Plot 9 Standardized Residuals against Predictors

4.5.2 Constant Variance Assumption

This assumption checks whether the variance of the residuals is constant. That is to say, the spread of the residuals should remain roughly the same as the values of the independent variables change. Here, we plotted standardized residuals against fitted values. According to Plot 10, there is no discernible pattern or trend, so the constant variance assumption is met.



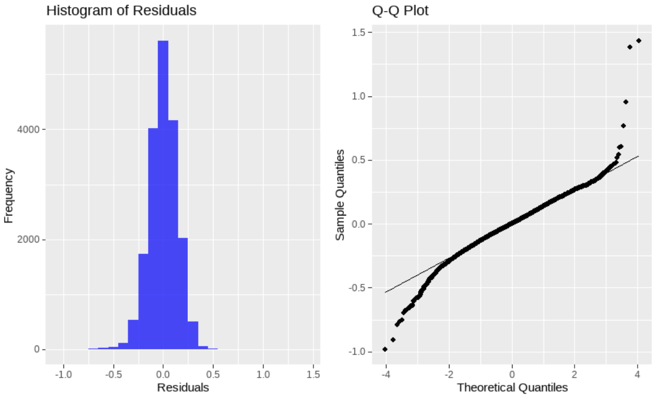
Plot 10 Standardized Residuals against Fitted Values

4.5.3 Independence Assumption

This assumption checks if the residuals resulting from the model's predictions are independent of each other. Knowing the value of one residual should not provide any information about other residuals. Looking into Plot 10, there is no special structure or cluster, which suggests that the residuals are independent of each other and the independence assumption is satisfied.

4.5.4 Normality Assumption

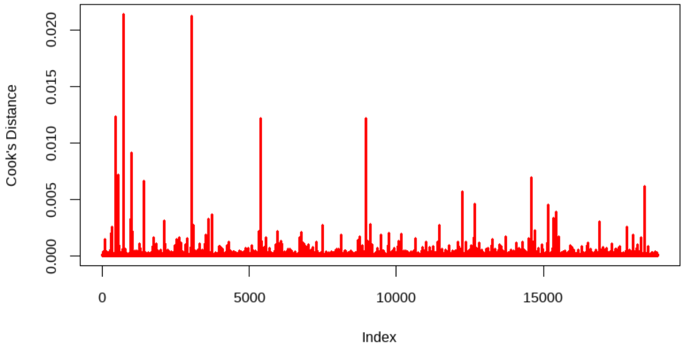
This assumption checks whether the residuals are normally distributed. As is shown in Plot 11, in the histogram the residuals have a unimodal distribution. They are approximately symmetric distributed with no gaps in the data, suggesting a normal distribution. However, in the Q-Q plot, there is curvature at both ends, which indicates non-normality. To sum up, the normality assumption may not be strictly satisfied.



Plot 11 Histogram of Residuals (left) and Q-Q Plot (right)

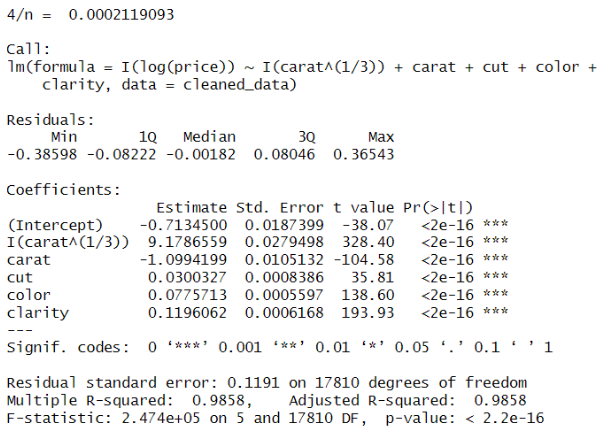
4.6 Identifying and Handling Outliers

To identify the outliers in the data, we used the Cook's distance method. Data points with significantly larger distances could be considered as outliers. In Plot 12, there are a large number of outliers, which seem to be a heavily tailored distribution instead of simply extreme values. We reviewed the original dataset and found that there exists cubic zirconia (an inexpensive diamond alternative with many of the same qualities as a diamond). Thus, these unusual observations could be caused by the lower price of cubic zirconia.



Plot 12 Cook's Distance

To clean the dataset and see if the model could be improved, we used 4/n as the threshold and removed the outliers. According to Fig 5, the model trained on the cleaned dataset can better explain the variance because the R-squared is slightly increased to 0.9858.

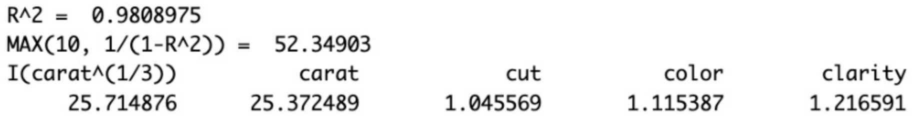


```
4/n =  0.0002119093

Call:
lm(formula = I(log(price)) ~ I(carat^(1/3)) + carat + cut + color + 
    clarity, data = cleaned_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.38598 -0.08222 -0.00182  0.08046  0.36543 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -0.7134500  0.0187399  -38.07   <2e-16 ***
I(carat^(1/3))  9.1786559  0.0279498  328.40   <2e-16 ***
carat          -1.0994199  0.0105132 -104.58   <2e-16 ***
cut             0.0300327  0.0008386   35.81   <2e-16 ***
color           0.0775713  0.0005597  138.60   <2e-16 ***
clarity         0.1196062  0.0006168  193.93   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1191 on 17810 degrees of freedom
Multiple R-squared:  0.9858,	Adjusted R-squared:  0.9858 
F-statistic: 2.474e+05 on 5 and 17810 DF,  p-value: < 2.2e-16
```

Fig 5 Summary of the Model on the Cleaned Dataset

4.7 Multicollinearity Checks

To check for the potential multicollinearity, we calculated the Variance Inflation Factor (VIF) for each predictor. Since the calculated threshold (52.34903) was too large, we chose to ignore it and used 10 as our threshold. The VIFs of carat^(1/3) and carat exceed the limit we set, which indicates that there is multicollinearity between these two variables. However, we decided not to deal with this issue because it would not affect the model performance greatly. In fact, if we remove one of these two variables, the model's R-squared would be decreased slightly.



```
R^2 =  0.9808975 
MAX(10, 1/(1-R^2)) =  52.34903 
I(carat^(1/3))          carat            cut          color        clarity 
     25.714876      25.372489       1.045569       1.115387       1.216591 
```

Fig 6 VIFs for Predictors

5 Conclusions and Recommendations

In this project, we conducted a regression analysis of the determinants of gemstone pricing. We initiated our analysis by addressing the issue of missing data and effectively mapping ordinal to numeric variables. Our exploratory data analysis revealed a strong correlation between the carat of gemstones and the dimensions in x, y, and z, prompting us to simplify our features by focusing solely on the carat variable. The initial model demonstrated satisfactory performance, achieving an R-squared value of 0.9046, which indicates that the model explained a notable proportion of the variance in gemstone prices. However, the diagnostics indicated the presence of a nonlinear relationship in the data, as evidenced by the patterns observed in the residual plots. This led to the model transformation of the dependent variable and a key predictor, resulting in an advanced model that includes both logarithmic and cube-root transformations. The adjusted model demonstrated superior predictive accuracy and model fit, as evidenced by an R-squared value of 0.9809. Additionally, the model was tested using the validation set. Visual observations of the scatter plot revealed a marked improvement in the model's predictive power.

Our analysis has demonstrated that the logarithm of a gemstone's price is predominantly determined by the cube root of the carat, the carat value, and the gemstone's cut, color, and clarity. This model allows for accurate price predictions when these four attributes are known. For instance, applying our model to a gemstone with a carat of 1.5, a cut rating of 4, a color rating of 6, and a clarity rating of 7, we estimated the price to be approximately $14,068.

A significant area for future study is the proactive handling of outliers, particularly the identification and exclusion of synthetic gemstones such as cubic zirconia from the dataset. Furthermore, the inclusion of additional predictors, such as market conditions and geographic factors, could potentially reveal further variations in gemstone pricing. The exploration of these factors would enhance the model’s accuracy and comprehensiveness.

6 Appendix

Data sample

id	carat	cut	color	clarity	depth	table	x	y	z	price
19041	1.17	Good	D	SI2	60.4	65	6.81	6.77	4.1	5567
397	1.2	Ideal	F	VVS1	61.1	55	6.86	6.89	4.2	13088
15627	0.31	Very Good	I	VVS2	61.6	59	4.31	4.33	2.66	544
16598	2.19	Ideal	I	SI2	62.5	56	8.31	8.24	5.18	15254
5812	0.3	Ideal	H	VS2	62.1	57	4.27	4.3	2.66	491
6310	1.01	Very Good	D	SI2	59.1	63	6.59	6.54	3.88	3671

Variable Description

Variable	Description
Carat	Carat weight
Cut	Describe the cut quality, in increasing order Fair, Good, Very Good, Premium, Ideal.
Color	The color of the gemstone,with D being the best and J the worst.
Clarity	The absence of the Inclusions and Blemishes, in order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3
Depth	The height measured from the Culet to the table is divided by its average Girdle Diameter.
Table	The width of the Table is expressed as a Percentage of its Average Diameter.

X	Length in mm.
Y	Width in mm.
Z	Height in mm.

Reference

[1] Waad Alsuraihi, Ekram Al-hazmi, Kholoud Bawazeer, and Hanan Alghamdi. 2020. Machine Learning Algorithms for Diamond Price Prediction. In Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing (IVSP '20). Association for Computing Machinery, New York, NY, USA, 150–154. https://doi.org/10.1145/3388818.3393715

[2]Walter Reade, Ashley Chow. (2023). Regression with a Tabular Gemstone Price Dataset. Kaggle. https://kaggle.com/competitions/playground-series-s3e8

[3]*Supply and Demand: Gemstone Market Dynamics - Gem Society*. (2019, November 19). International Gem Society. https://www.gemsociety.org/article/supply-and-demand-gemstone-market-dynamics/

[4] G. Sharma, V. Tripathi, M. Mahajan and A. Kumar Srivastava, "Comparative Analysis of Supervised Models for Diamond Price Prediction," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 1019-1022, doi: 10.1109/Confluence51648.2021.9377183.

[5] OuYang, Z. (2024). Research on the Diamond Price Prediction based on Linear Regression, Decision Tree and Random Forest. Highlights in Business, Economics and Management, 24, 248-257. https://doi.org/10.54097/13ccwv59

[6]Farhoumand, A. (2014, April 9). Diamond Anatomy, Explained. GIA 4Cs. https://4cs.gia.edu/en-us/blog/diamond-anatomy-explained/

[7] Senarathne K.A.N.S, Epitawatta E.A.E.K, Thennakoon, K. T., Diunugala, M. W., HM Samadhi, C. R., & Madhuhansi, M. P. (2023). “Gemo”: An AI-powered approach to color, clarity, cut prediction, and valuation for gemstones. International Research Journal of Innovations in Engineering and Technology, 7(10), 406-416. doi:https://doi.org/10.47001/IRJIET/2023.710054