Abstract

This project aims to build accurate predictive models for diamond prices using regression models. It is crucial for diamond sellers and buyers to determine competitive prices and make informed decisions, respectively. Factors such as carat weight, cut quality, color grade, clarity, and physical dimensions impact a diamond's price. The project uses a dataset from Kaggle, with data on these characteristics for 53,909 diamonds.

The analysis encompasses three regression models and handles multi-collinearity among predictors and prevents model overfitting. A range of techniques like visualization, exploratory data analysis, model diagnostics, and train-test split are used to ensure the models' reliability. The selected model achieves 0.9848 in R^2 and 449681.2 in MSE.,

The results demonstrate high accuracy in predicting diamond prices based on these properties. However, there's an acknowledgment that the models might not account for all potential factors, like market demand or economic conditions. Future work could explore additional variables, advanced regression techniques like Lasso or Elastic Net regression, broader datasets, or sentiment analysis of customer reviews to further improve the model.

1 Introduction

Diamonds have long been prized for their rarity and beauty, making them a valuable commodity in the jewelry industry. The diamond industry is highly competitive, and pricing decisions can have a significant impact on sales and profits. Historically, pricing decisions were often based on intuition and market trends, which might be not fair enough. The price of a diamond depends on complex factors, including its carat weight, cut quality, color grade, clarity, and physical dimensions. Accurately predicting diamond prices is crucial for both buyers and sellers in the diamond market.

The primary objective of this project is to build a model that can predict diamond prices accurately. This predictive model will enable diamond sellers to set competitive prices, and buyers to make informed purchasing decisions. Moreover, understanding the factors that influence diamond prices can provide valuable insights into market trends and consumer preferences.

To address the problem of predicting diamond prices, we will build three regression models. The dataset used in this project contains information about diamonds and can be accessed through Kaggle [1]. The linear regression models will serve as the baseline models, capturing the relationships between the dependent variable (price) and the independent variables (diamond characteristics). The ridge regression model will provide a regularized approach to handle multicollinearity among predictor variables and prevent overfitting.

By analyzing a comprehensive dataset containing diamond attributes and prices, we will develop and compare the performance of the three regression models. Through visualization, exploratory data analysis, model diagnostics, and cross-validation, we will ensure that our models adhere to the assumptions of linear regression and provide reliable price estimates.

1.1 Dataset Introduction

This dataset has 53909 samples, including the diamond prices (response variable) and 9 corresponding predictors:

- carat: Weight of the diamond.
- cut: Quality of the cut (Fair, Good, Very Good, Premium, Ideal).
- color: Diamond color, from J (worst) to D (best).
- clarity: A measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)).
- depth: Total depth percentage = $\frac{z}{mean(x,y)} = \frac{2z}{x+y}$.
- table: Width of top of diamond relative to widest point.
- x: Length in mm.
- y: Width in mm.
- z: Depth in mm.

2 Data Preprocessing

2.1 Data Cleaning

In this project, all data should be greater than zero. Therefore, invalid data points are filtered out by removing observations with non-positive values for the dimensions (x, y, z) and carat weight.

2.2 Feature Engineering

- **Categorical Variables**

  In this dataset, we employ one-hot encoding for the three categorical variables ("cut", "color", "clarity"). One-hot encoding converts each category within a variable into a binary variable. For instance, "cut" had five categories ("Fair", "Good", "Very Good", "Premium", "Ideal"), we created five binary variables for each category. Each binary variable takes a value of 1 if the data belongs to that category and 0 otherwise.

- **Numerical Variables**

  Considering the definition of "depth" $\frac{2z}{x+y}$, we create a new term "x_y" by adding "x" and "y", which allows the model to capture the joint effect of the length and width of the diamond. Then we visualize the distribution of all numerical variables via histograms as is shown in Fig. 1.

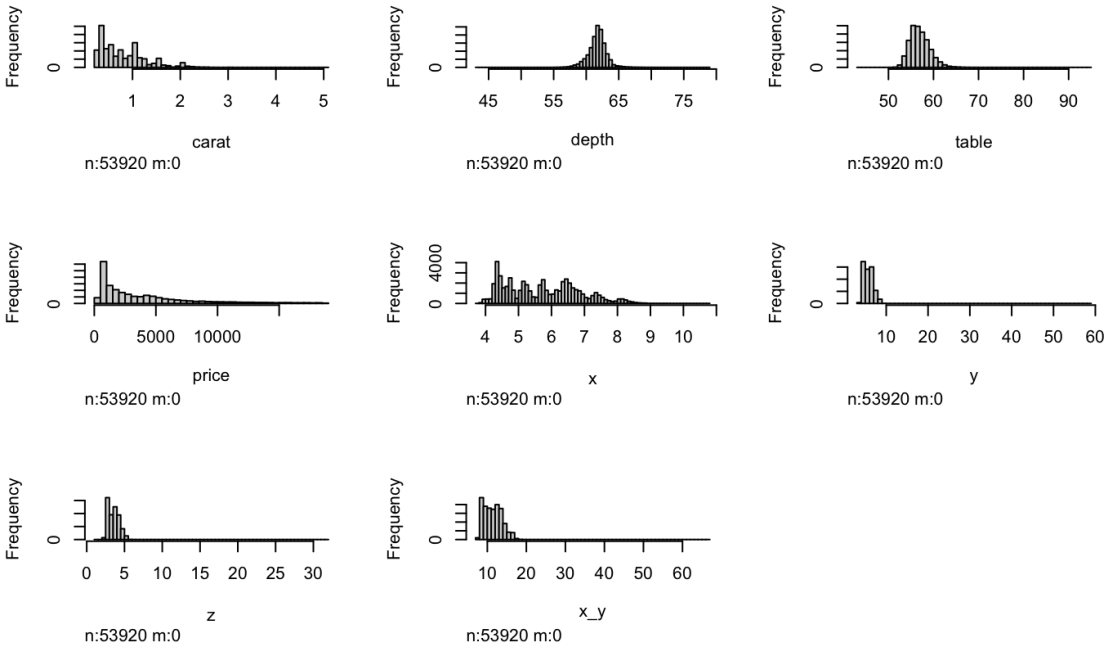

Fig. 1. Distribution of all numerical features.

To address skewness in the data and enhance the linearity assumption for the regression model, we apply the Box-Cox transformation to the numerical variables "price", "carat", "depth", "table", "x", "y" and "z". This transformation is particularly useful in handling variables with skewed distributions and aims to stabilize variance and achieve normality.

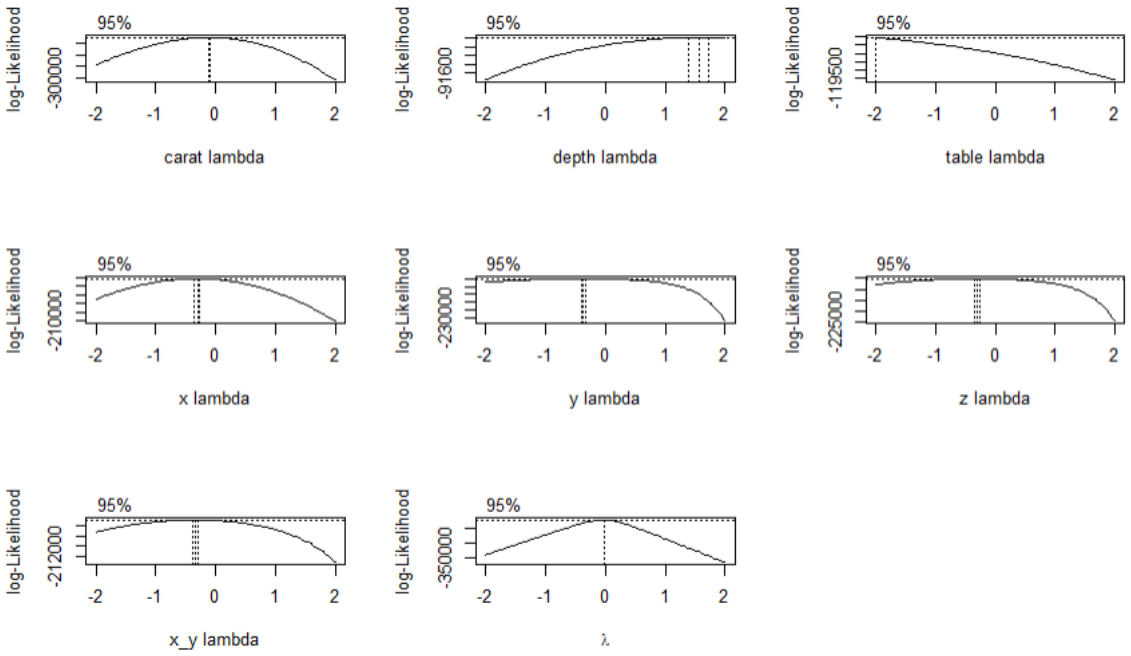

Fig. 2. Box-Cox plots of all numerical features.

Based on the Box-Cox plot, we perform log transformations on the independent variables whose λ is close to 0 ("carat", "x", "y", "z" and "x_y") to further mitigate skewness. After that, we conduct a Box-Cox transformation on the regression model that relates the "price" of diamonds to the log-transformed features and the plot is shown in Fig. 2. Since λ is close to 0, we apply log transformation to the response variable "price".

3 Model Design and Analysis

3.1 Feature Selection Based on Multi-collinearity Detection

To reduce the error resulting from Multi-collinearity, redundant features are required to be removed.

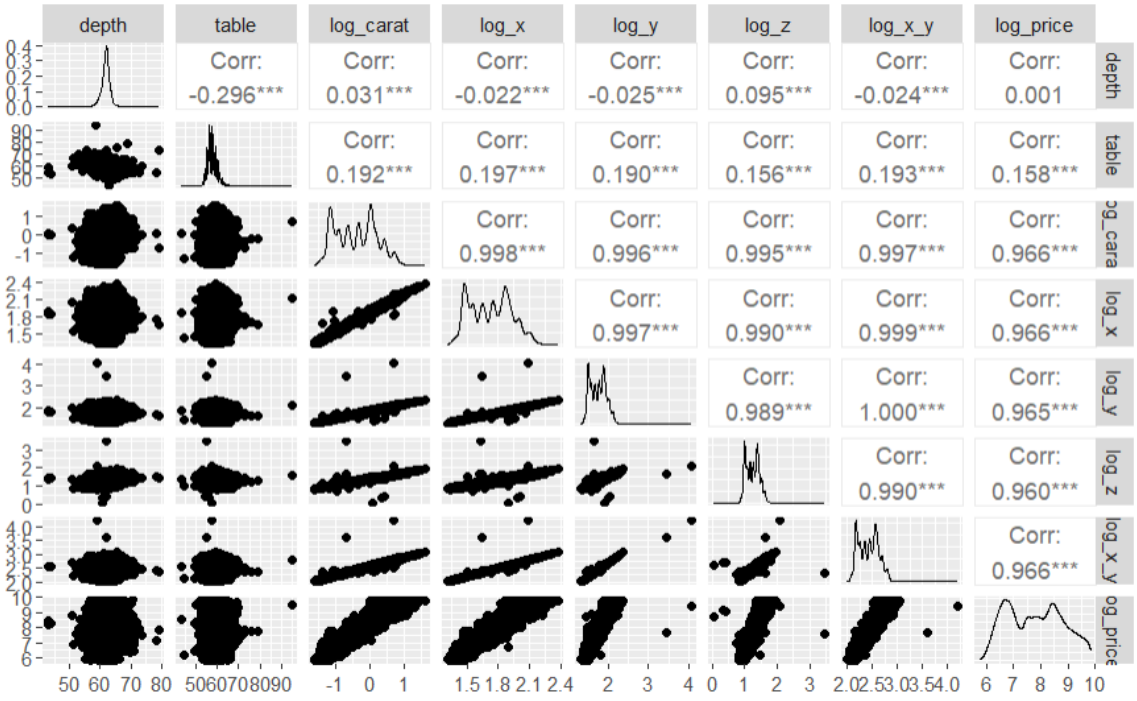

Fig. 3. Pair plots of numerical features.

According to the pair plot and the correlation calculated shown in Fig. 3, there is a strong correlation between log_carat, log_x, log_y, log_z, and log_x_y. To be more precise, we use the variance inflation factor to quantify it in Tab. 1. Moreover, with the idea of principal component analysis, the scatter plot between residuals of log_price, log_x, log_y, and log_z after regression with respect to log_carat can show whether log_x, log_y, and log_z are necessary for predicting the diamond price shown in Fig. 4.

Tab. 1. VIF in $log\_price \sim log\_carat + log\_x + log\_y + log\_z$

Variable	VIF
log_carat	515.8976
log_x	395.6284
log_y	164.8624
log_z	111.1384

3.2 Feature Selection Based on BIC

However, even after investigating collinearity, there are still unexplored variables. To address this, we employ the Bayesian Information Criterion (BIC) to identify significant features, as defined in Eq. 1. Our approach begins with a simple regression using the variable log_carat, and subsequently, we use stepwise selection (forward or backward) to include or exclude model features from the following set: log_carat, carat, depth, table, color, clarity, and cut, ensuring collinearity is mitigated. The results obtained from the BIC analysis indicate that we can eliminate two variables, namely depth and table, from our linear model.

$$BIC = n \log(\frac{SSE_p}{n}) + p \log n \tag{1}$$

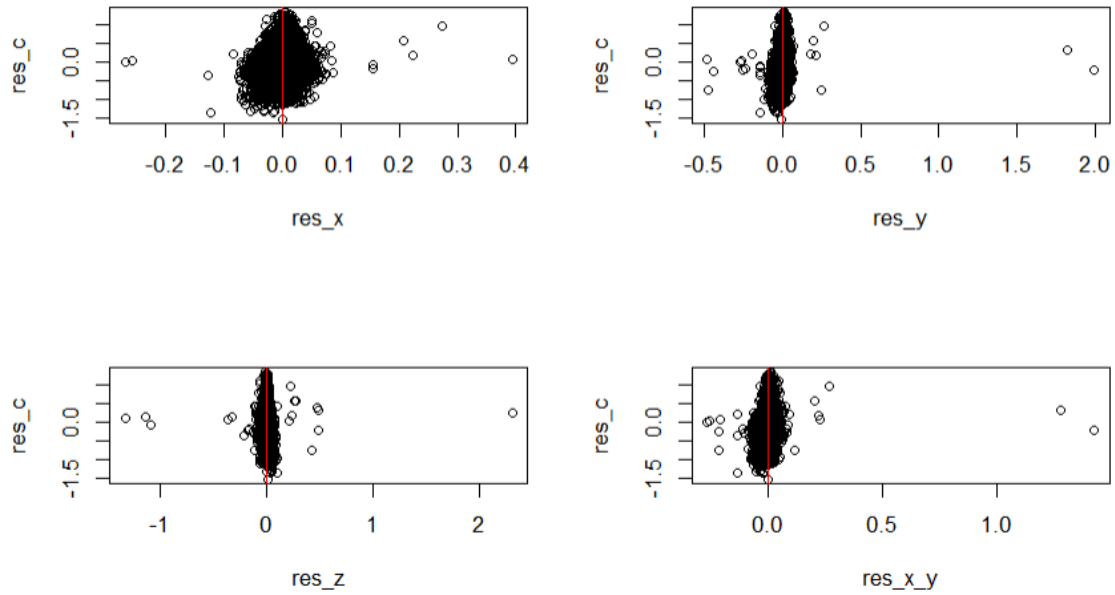

Fig. 4. Pair plots of numerical features.

3.3 Boxplots for Categorical Variables

As is mentioned above, we employ one-hot encoding for categorical variables. Since we are looking for the relationship between a continuous variable (price) and categorical variables (cut, color, and clarity), we can use boxplots to observe the relationship graphically (Fig. 5).

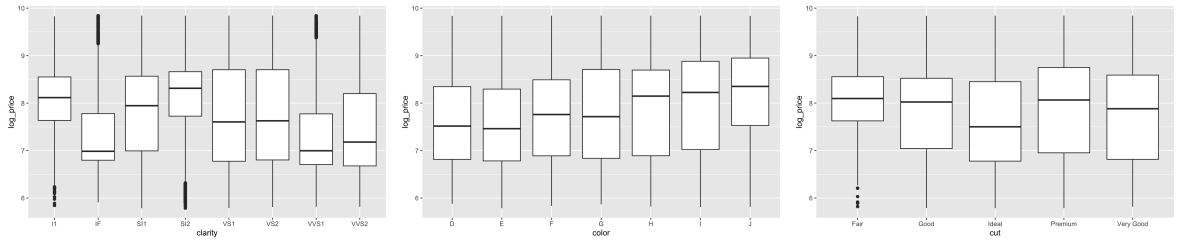

Fig. 5. Boxplots of price for different categories in clarity, color and cut.

Boxplots for the categorical variables (cut, color, and clarity) provide valuable insights into the relationship between each categorical variable and the target variable (diamond prices). Since the medians of different categories vary from each other, we can conclude that these categorical variables have an influence on the price of diamonds. Surprisingly, we find the distribution of price is opposed to our expectations. For instance, diamonds with higher cut grades, higher color grades, and higher clarity grades should tend to have higher median prices. However, we find that the higher the grades are, the lower the price is. So we can suppose that there are interactions between these categorical variables and other numerical variables.

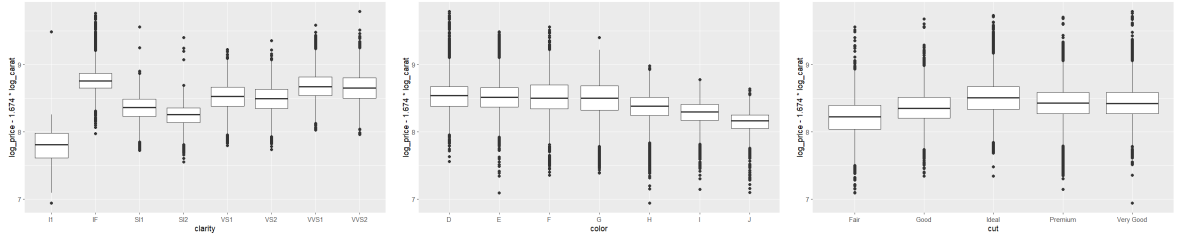

Fig. 6. Adjusted Boxplots of price for categorical variables.

To further investigate the counter-intuitive tendency, the linear regression model "log_price" ∼ "log_carat" is fitted, and the fitted linear parameter for "log_carat" turns out to be 1.674. By substituting the y-axis of Fig. 6 with "log_price" -1.674"log_carat", the tendency of "log_price" on the categorical variables now fits the intuition, where the higher the grades are, the higher the price is. One possible reason for the phenomenon is that diamonds with higher quality hold lower carats overall.

3.4 Interaction Terms between Categorical and Numerical Variables

After exploring the categorical variables, the following analysis intends to discover the interaction between two variables. Since the carat has been identified as the most important factor in the price of diamonds, the analysis will start with the interaction between the categorical variables including cut, clarity, and color and the numerical variable carat.

In order to examine the interaction, we will draw the relationship between the log value of price and the log value of carat respectively with respect to cut, color, and clarity shown in Fig. 7.

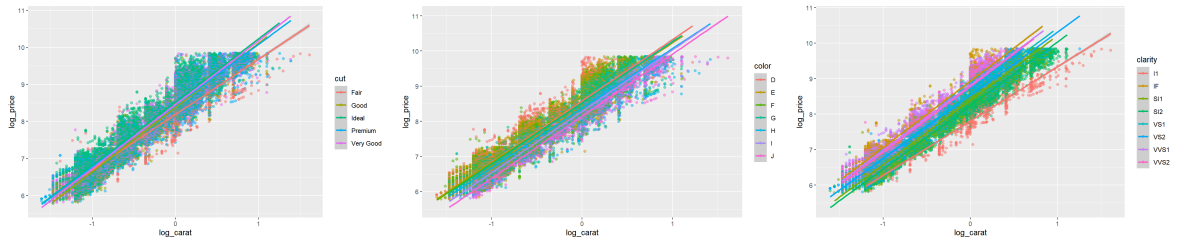

Fig. 7. Interaction between log_carat and categorical variables.

As for the cut, the figure buttresses that the cut quality and clarity of diamonds will lead to the variance in the intercept and the slope in the linear relationship between the price and carat in their log scale. However, the diamond color doesn't seem significantly affect the slope but may change the intercept term. Based on the information extracted from Fig. 7, we can state that it is wise to the interaction term between carat and cut and that between carat and clarity, which provides graphical evidence for the models trained later. We may use the performance to determine whether to include the interaction between color and clarity.

From the cut and clarity, we can also identify that the diamonds with the fair cut quality seem to significantly differ from diamonds with other cut qualities. Simultaneously, the diamonds with the I1 clarity also deviate from diamonds with those with other degrees of clarity. In this way, we can generate two new identity variables based on the analysis. The tall and thin shape of the residual comparison indicates that y, z, and x_y may not contribute to the model performance, but we may preserve the conservative attitude towards x. To clarify, res in the figure label is shortened to "residual".

3.5 Interaction Terms between Categorical Variables

Further exploration focuses on the interaction within the categorical variables. To simplify the graph, we choose to visualize the relationship between price and carat in different clarity and cut in each group of color shown in Fig. 8, which can significantly decrease the visualization burden. It can be noticed that lines with different colors are positioned in different orders in each sub-figure in Fig. 8. For example, we can see that figures in the same column in Fig. 8 have a similar order to those colored lines. However, the figures in the same row may share similar patterns if neglecting those with few data points. In this case, one may consider the interaction between clarity and color in the linear regression model to improve performance.

3.6 Model Fitting and Evaluation

To better evaluate the performance of the models, we split the diamonds dataset into training (70%) and testing (30%) sets. Both value of R^2 and the mean squared error on the test set will be applied to evaluate the models.

The full linear model before log transformation is fitted as a starting point. The mean squared error on the testing set for the model is 1227713. The other models are evaluated with the mean squared error percentage decreasement compared with the full model.

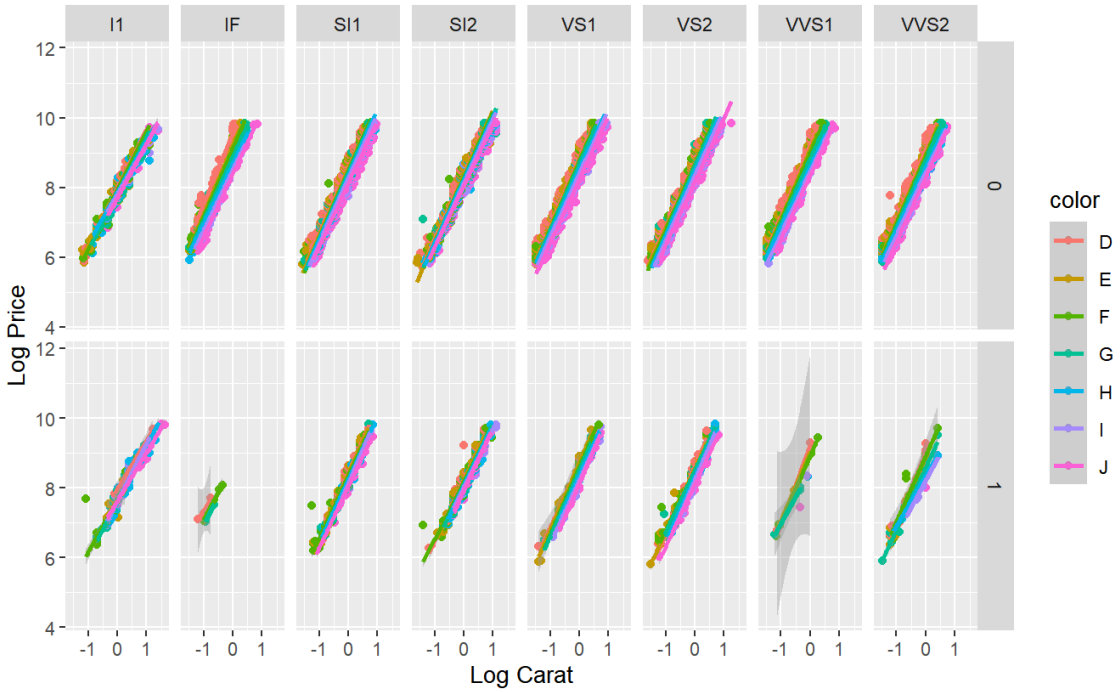

Fig. 8. Interaction within categorical variables.

Based on the previous explorations of feature interactions, we discovered that the logged price is correlated to the interaction between "log_cat" with "clarity", "log_cat" with "cut", "color" with "clarity", and possibly "log_cat" with "color". In this way, along with the full linear model after log transformation, three following models are trained:

$$M1 : log\_price \sim color * clarity + log\_carat * clarity + log\_carat * cut$$
$$M2 : log\_price \sim color * clarity + log\_carat * clarity + log\_carat * cut + log\_carat * color$$
$$M3 : log\_price \sim log\_carat + cut + color + clarity + depth + table^{-2} + log\_x + log\_y + log\_z + log\_x\_y$$

Tab. 2. Model performance analysis

Model	R^2	R^2_{adj}	**Test MSE**	**Performance Improvement**
M1	0.9846	0.9845	487594.9	60.28%
M2	0.9848	0.9848	449681.2	63.37%
M3	0.9827	0.9827	625240.4	49.07%

The performance of each model is shown in Tab. 2. According to the MSE value calculated based on the test dataset, it can be observed that all 3 models perform far better than the original full model before log transformation. However, within the 3 models, $M1$ and $M2$ perform far better than $M3$, which fits the previous analysis of the data. Moreover, Fig. 8 shows the scatter plot of fitted values on actual values. It can be observed that the scatters concentrate closely around the $y = x$ line, which indicates the model fits well.

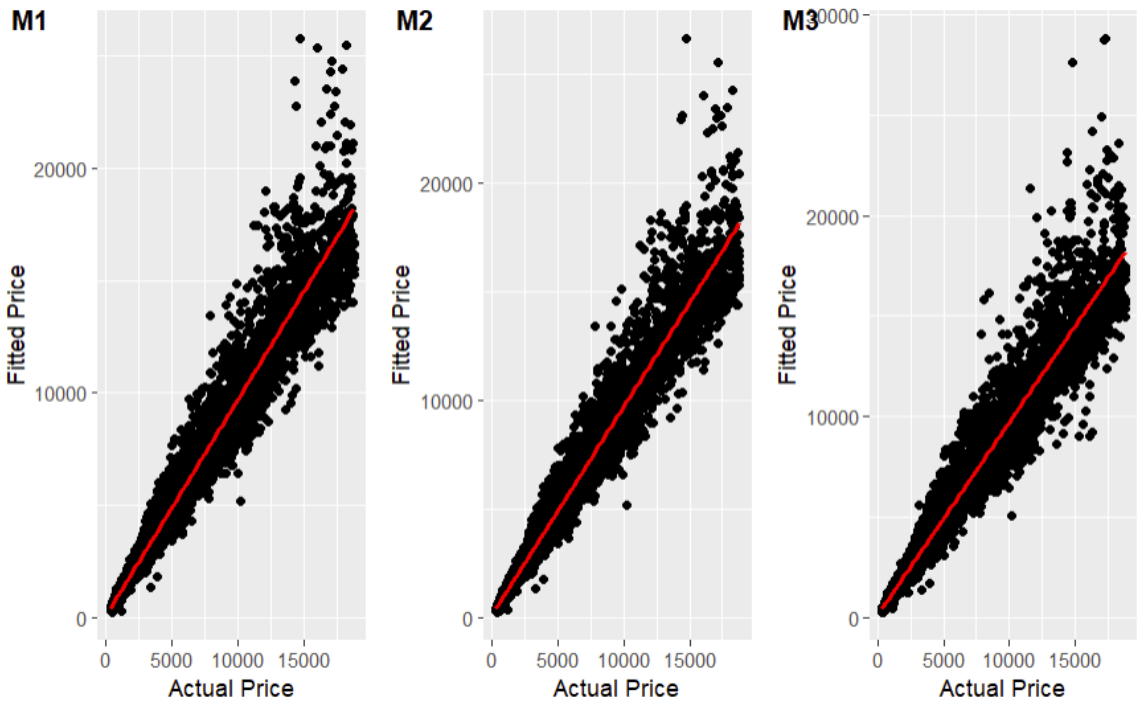

Fig. 9. Fitted price v.s. actual price.

The basic regression assumptions for the linear regression models are independent, identical, and normally distributed. In Fig. 10, the residual plots of the models show that the residuals of the fitted model distributed uniformly around 0, independent from the fitted values, and the QQ-plots of the models show that the distribution of the standardized residuals of the fitted model is nearly normal, but slightly heavy-tailed. It can thus be concluded that the model basically fit the assumptions. Concentrating on $M1$ and $M2$, it is obvious that $M1$ is a reduced model of $M2$. By conducting F-test, with confidence level $\alpha = 0.05$ and $p - value < 2.2e - 16$, it can be concluded that we cannot safely reduce model $M2$ to $M1$, and thus model $M2$ should be selected.

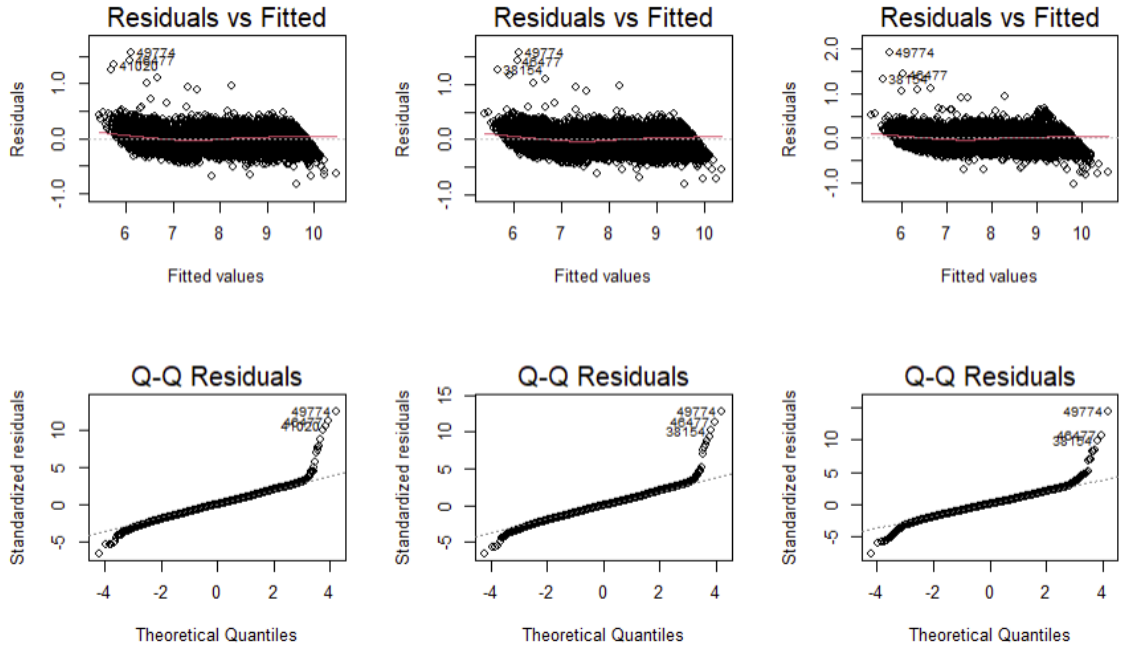

Fig. 10. Residual and QQ plots of models.

4 Results

The statistical information of residuals indicates differences between the actual log_price values and the predicted log_price values. The statistics show that the residuals have a relatively small spread, with the majority of values centered around zero (median close to zero) and exhibiting a fairly symmetrical distribution. However, there are some outliers with larger positive and negative residuals, as indicated by the maximum and minimum values (1.56809 and -0.83817).

The summary of the model displays the coefficients of the model, along with their estimates, standard errors, t-values, and p-values. The coefficients represent the amount of change in the log_price for a one-unit change in each predictor, while holding all other predictors constant. The "Intercept" term represents the baseline log_price when all predictors are zero. All the coefficients, except for the "colorE" coefficient, are highly significant with p-values much smaller than 0.05 (indicating strong evidence against the null hypothesis of no effect). This suggests that all predictors, including "log_carat", "clarity", and "cut", have a significant impact on the log_price of diamonds.

Specifically, the "log_carat" coefficient has a positive estimate of 1.529665, meaning that for every one-unit increase in the logarithm of carat weight, the log_price of the diamond is expected to increase by approximately 1.53 units.

The "clarity" coefficients represent the effect of different clarity levels compared to the reference category "IF". For example, diamonds with clarity "SI1" have an estimated log_price that is approximately 0.91 units lower than the clarity "IF" diamonds.

Similarly, the "cut" coefficients indicate the effect of different cut categories compared to the reference category "cutFair". For instance, diamonds with "cutIdeal" have an estimated log_price that is approximately 0.18 units higher than "cutFair" diamonds.

The "color" coefficients show the effect of different color grades compared to the reference category "colorD". For example, diamonds with "colorJ" have an estimated log_price that is approximately 0.305 units lower than "colorD" diamonds, indicating that higher color grades tend to have higher log_prices.

The third section presents interaction terms between "log_carat" and "clarity", as well as "log_carat" and "cut". These interaction terms capture how the relationship between "log_carat" and "log_price" varies across different levels of clarity and cut. All interaction terms are highly significant (p-values much smaller than 0.05), indicating that the effect of "log_carat" on "log_price" is not uniform across all levels of clarity and cut.

In conclusion, this multiple linear regression model provides valuable insights into the factors influencing diamond prices. "log_carat", "clarity", "cut" and "color" all play significant roles in determining the log_price of diamonds, and their effects may interact with each other, leading to variations in price relationships across different combinations of these predictors. The model seems to be a good fit, as most of the coefficients are highly significant and consistent with what we would expect based on diamond market knowledge.

5 Conclusion and Discussion

In this project, we explore the task of predicting diamond prices using linear regression and ridge regression. The dataset contains valuable information about various diamond characteristics, including carat weight, cut quality, color grade, clarity grade, depth percentage, table size, and physical dimensions. Leveraging this rich dataset, we perform exploratory data analysis (EDA), build three regression models, and evaluate their performance. Finally, our model can predict the price of diamonds based on their basic properties with high accuracy.

With log transformation, variable interaction, and model reduction, the eventually derived models are $M1 : log\_price \sim color * clarity + log\_carat * clarity + log\_carat * cut$ and $M2 : log\_price \sim color * clarity + log\_carat * clarity + log\_carat * cut + log\_carat * color$. With the models, we managed to reduce the mean squared error of the model from 1227713 to less than 500000, with a drop of over 60%.

With F-test, we reached the conclusion that model $M2$ cannot be safely reduced to $M1$, and $M2$ should be selected. And the inference for the fitted parameters meets the intuition on diamond prices. With a higher quality of color, cut, and clarity, and with higher carats, the diamonds will hold a higher price.

While our models exhibit promising performance, they may not account for all potential factors influencing diamond prices. Other external factors, such as market demand and economic conditions, could also play a role. Future work could involve feature engineering to create additional informative

variables and explore other advanced regression techniques beyond linear regression, such as Lasso regression or Elastic Net regression, to compare their performance. Additionally, exploring more extensive datasets or incorporating sentiment analysis of customer reviews might help gain insights into subjective factors influencing diamond prices.

References

[1] Diamond Dataset. Kaggle. https://www.kaggle.com/datasets/shivam2503/diamonds.