my title*

subtitle

Yawen Tan

December 3, 2024

4 sentence

Table of contents

*Code and data are available at: [https://github.com/YawennnnnnTan/Analysis-of-Diamond-Price]

1 Introduction

4C

2 Data

2.1 Overview

The data set has 2690 data and 5 variables. Variables include response variable diamond price, predictor variables Diamond Color, Diamond Carat Size, Diamond Clarity, Diamond Cut. Table 1 summarizes the range, variable type and examples of each variable. The detailed description of each variables will be presented in outcome variable and predictor variables section.To ensure data quality and clarity, all missing values were removed, and column names were standardized for consistency and readability. Additionally, key categorical variables, including color, cut, and clarity, were converted into factors to appropriately represent their categorical nature in the analysis.

We use the statistical programming language R (R Core Team 2023), alongside several key libraries and datasets to analyze and visualize data effectively. Our data (Data and (DASL) 2024) is sourced from the Data and Story Library, providing detailed attributes of diamonds for multiple regression analysis. Following Alexander (2023), we consider how storytelling principles can enhance data interpretation and presentation. The tidyverse ecosystem (Wickham et al. 2019) underpins our data manipulation and visualization, with dplyr (Wickham et al. 2023) for data wrangling and ggplot2 integrated packages to create visualizations. The here package (Müller 2020) simplifies file path management, while knitr (Xie 2023) and patchwork

(Pedersen 2020) streamline report generation and composite plotting. Bayesian modeling is conducted with rstanarm (Goodrich et al. 2024), leveraging robust statistical frameworks. Additionally, caTools (**caTools?**) aids in statistical operations such as ROC curve analysis, and ggcorrplot (Kassambara 2018) facilitates correlation visualization.

Table 1: Summary of Diamond Dataset

Attribute	Range or Levels	Data Type	Example
Diamond Price (USD)	1000 to 10,000 (USD)	Numeric	5000
Diamond Color	D to K (D: Best, K: Worst)	Categorical	G
Diamond Carat Size	0.30 to 2.02	Numeric	0.8
Diamond Clarity	SI2, SI1, VS2, VS1, VVS2, VVS1, IF (Low to High)	Categorical	VS1
Diamond Cut	Good, Very Good, Excellent, Ideal (Worst to Best)	Categorical	Very Good

2.2 Measurement

Gemological Institute of America (n.d.) illustrates the measurement of diamond features. Color is conducted by comparing the diamond against GIA's master stones (graded from D to Z) under controlled lighting and background conditions, as light sources significantly affect appearance. At least two graders assess each diamond's color, with additional graders involved if discrepancies arise, until a consensus is reached. Clarity is performed under 10x magnification in standard observation conditions, where the grader examines internal and external characteristics, documenting inclusions, blemishes, or treatments such as laser drilling or fracture filling. Cut utilizes high-precision instruments to measure cutting proportions, angles, symmetry, and polish quality, following international standards set by organizations like GIA or AGS, with grades such as Excellent and Very Good. Carat size is determined using highly accurate electronic microbalances, capable of measuring to five decimal places, or one ten-thousandth of a carat, complemented by optical devices to assess the diamond's proportions, dimensions, and facet angles.

2.3 Outcome Variable: Price

The diamond price is a continuous numerical variable, ranging from close to 1000 USD to 10,000 USD. Through the histogram of Figure 1, we can find that the price of diamonds has obvious right skewness, and with the increase of price, the number of diamonds decreases gradually. In addition, combining the histogram and box chart in Figure 1, we can find that there is no obvious abnormal value of diamond prices. Most diamond prices are concentrated

in the range of 2,500 USD to 5,000 USD, with an average price of about 4,000 USD and a median of 3,500 USD, indicating that a few high-priced diamonds have raised the overall average.

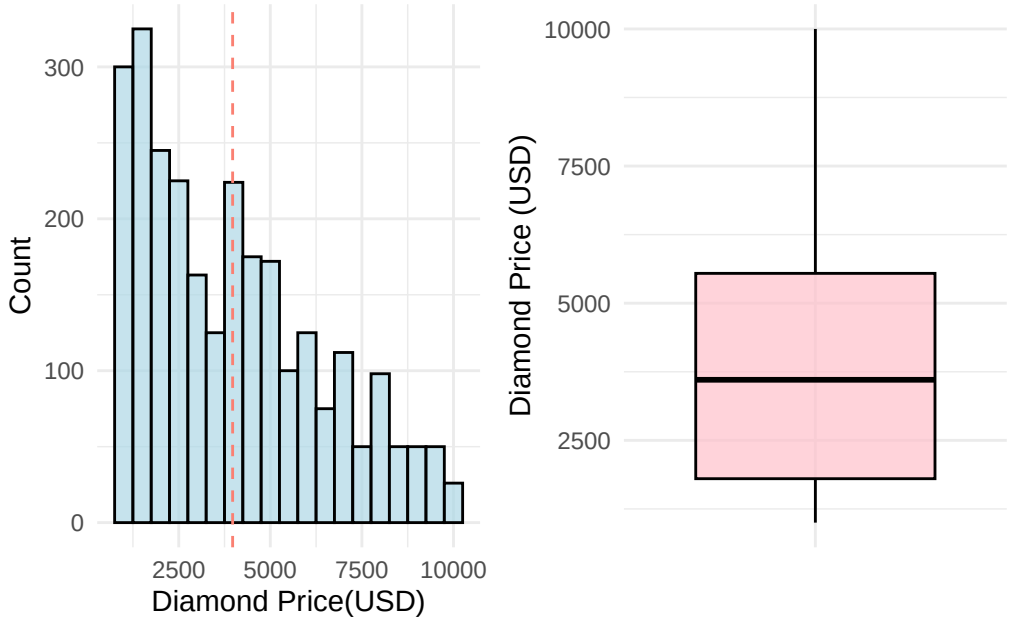

Figure 1: Graphs of Diamond price

2.4 Predictor Variables

2.4.1 Color

The diamond color, a categorical variable, represents the degree of colorlessness of a diamond. It has seven categories in the dataset, ranging from the letter D (completely colorless) to the letter K (faint yellow or brown tint), based on the GIA (Gemological Institute of America) grading system.

- D is the highest grade, completely colorless and the rarest and most expensive color in diamonds.
- E and F grades are almost colorless, and extremely weak tones can only be detected under professional instruments.
- G and H grades still look almost colorless, and only when compared with higher grades in bright light can a slight yellow or brown tone be observed.
- I and J began to show slightly yellow tint visible, especially in larger diamonds.
- K shows obvious yellow tone or brown tone.

The barplot on the left of Figure 2 shows the distribution of diamond color from D to K. The barplot shows that the number of diamonds with color grades E and F is the largest, indicating that these nearly colorless diamonds are more popular in the market, probably because they have high quality and cost performance. In contrast, the number of completely colorless D-class diamonds and K-class diamonds with slight yellow is small, which may be due to their rarity and color deviation, respectively, resulting in low demand. This picture directly reflects the supply of diamonds of different color grades in the market and their potential demand trends. The violin chart on the right of Figure 2 shows the distribution of diamond prices in different color grades. The violio chart shows that the prices of diamonds with color grades D to F are generally higher, especially those with color grades D and E, which shows a wider price distribution range, indicating that the prices of high-end diamonds with these color grades can increase significantly. In contrast, the price distribution of diamonds with color grades I to K is more concentrated and the overall price is lower, reflecting that the price fluctuation of diamonds close to yellow is less in the market. In addition, there are significant outliers in top colors such as D and E, which may be due to the unusually high price of diamonds with large carats or other high-quality characteristics. This figure clearly reveals the influence of color grade on price and its market value distribution characteristics.

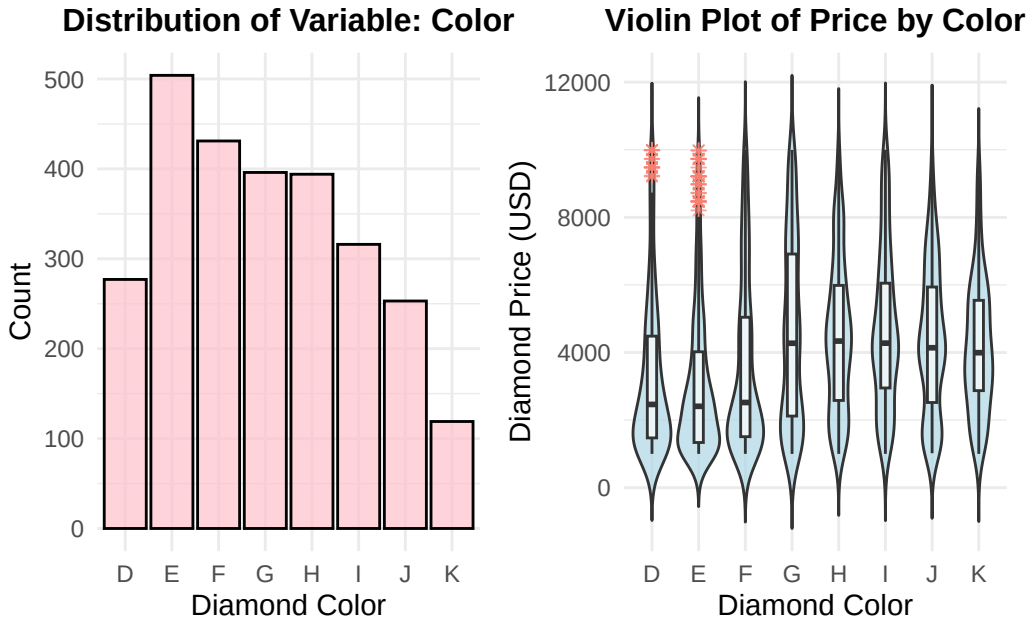

Figure 2: Graphs of Predictor Variables: Color

2.4.2 Carat Size

The diamond carat size, a continuous numerical variable representing the weight of the diamond, ranges from approximately 0.30 to 2.02 carats in the dataset. One carat is equivalent to 200 milligrams, making it a key factor in determining the diamond's size and price. The histogram on the left of Figure 3 shows the distribution of the Carat Size of diamonds. On the whole, it presents a right-skewed distribution, indicating that the number of small carats (such as 0.5 to 1.0 carats) is the largest in the market, while the number of larger carats is gradually decreasing. The red dotted line in the figure indicates the average carat size, and it can be seen that the carat size of most diamonds is concentrated on the left side of the average, which further reflects the dominant position of small carat diamonds in the market. This reflects that consumers have higher demand for small carats, while large carats are scarce because of their rarity and high price. The scatter chart on the right of Figure 3 shows the relationship between the carat size of diamonds and the price. It can be clearly seen that the price rises rapidly with the increase of carat size, showing a nonlinear growth trend. Especially in the range of close to 1.0 carats and larger carats, the price increase is more significant. This shows that the carat size has an important influence on the diamond price, but when the Dancla size exceeds a certain value (such as 1.5 carats), the price growth rate tends to be flat. In addition, the fitted black curve further clearly shows this nonlinear relationship, indicating that the marginal effect of carat size on price decreases.

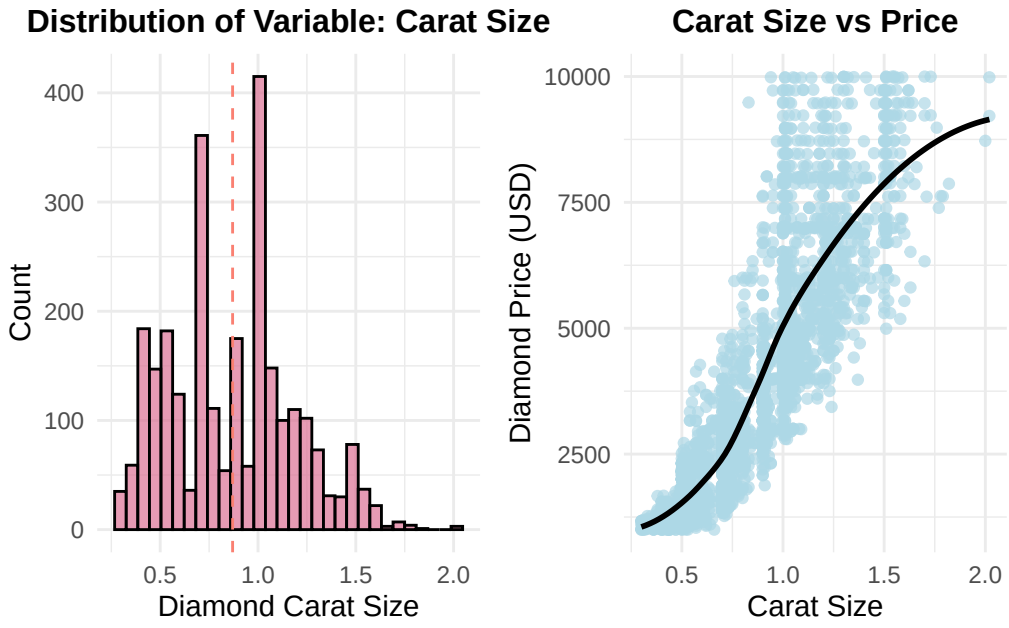

Figure 3: Graphs of Predictor Variable: Carat Size

2.4.3 Clarity

The diamond clarity, representing the number, size, location, and visibility of internal inclusions and external blemishes in a diamond, is a categorical variable with seven categories in the dataset: IF, VVS1, VVS2, VS1, VS2, SI1, and SI2. According to the GIA grading system, diamond clarity decreases progressively from IF (Internally Flawless) to SI (Slightly Included).

- Internally Flawless (IF) diamonds are characterized by having no visible inclusions under 10x magnification, which is nearly flawless. The only possible imperfections might be extremely fine surface blemishes, such as polishing marks, which can typically be removed with re-polishing.
- Very Very Slightly Included (VVS) diamonds are divided into two subgrades: VVS1 and VVS2. VVS1 diamonds have extremely small inclusions that are barely detectable under 10x magnification, typically located at the pavilion (the bottom part of the diamond). VVS2 diamonds may have slightly more inclusions, such as tiny feather-like marks or minute blemishes, but these still require professional tools to identify.
- Very Slightly Included (VS) diamonds are classified into two subgrades: VS1 and VS2. VS1 diamonds have very small inclusions that are detectable under 10x magnification but careful observation to identify. VS2 diamonds contain relatively minor inclusions such as small feather-like marks or pinpoints, have more noticeable imperfections under magnification, but these imperfections have minimal impact on the diamond's overall appearance.
- Slightly Included (SI) diamonds are categorized into two subgrades: SI1 and SI2. SI1 diamonds have inclusions that are more noticeable under 10x magnification and can be easily detected using professional tools. SI2 diamonds contain more inclusions, which are larger or more prominent, but these imperfections are typically not visible to the naked eye in most conditions.

The barplot on the left of Figure 4 shows the distribution of diamond Clarity. It shows that the number of diamonds with clarity grades SI1 and SI2 is the largest, indicating that these grades of diamonds are the most widely available in the market, probably because they have higher cost performance and greater market demand. In contrast, the number of diamonds with clarity grade of IF(Internally Flawless) is the least, which reflects the rarity of completely flawless diamonds. This shows that diamonds with lower definition grades (such as SI1 and SI2) are the main ones in the market, while high definition diamonds are scarce. The violin chart on the right of Figure 4 shows the distribution of diamond prices at different clarity levels. With the improvement of clarity, the price rises significantly, especially for diamonds with high definition such as IF and VVS1. The price distribution is wider and there are obvious outliers, which may be caused by the carat size or other quality characteristics of these diamonds. On the other hand, the price distribution of grades with low definition (such as SI1 and SI2) is relatively concentrated, and the median is obviously lower than that of high-grade diamonds, but their price range shows their stable demand and high cost performance in the market.

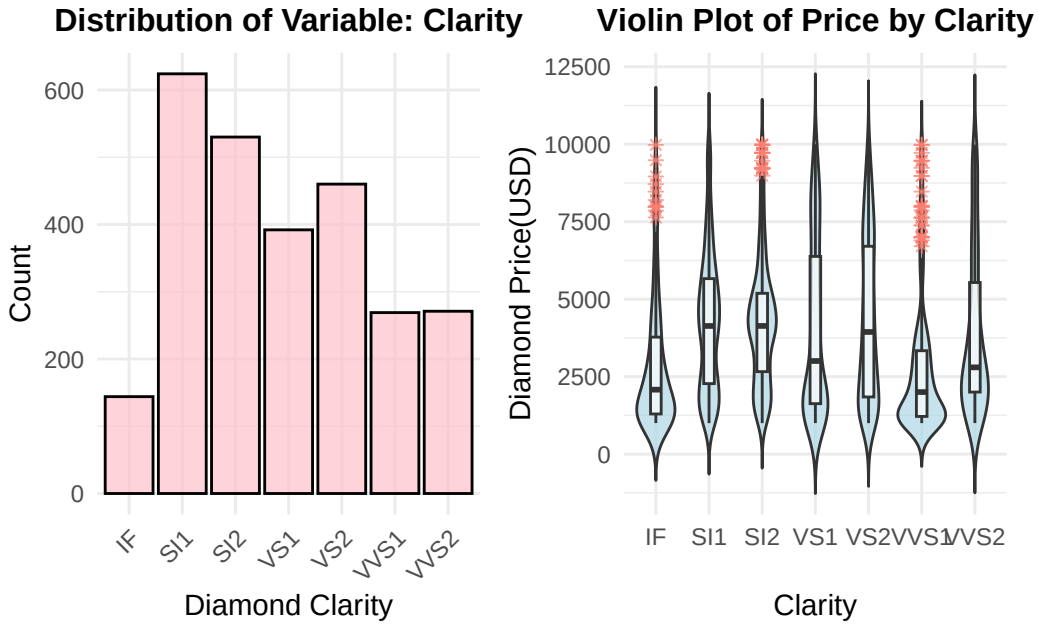

Figure 4: Graphs of Predictor Variable: Clarity

2.4.4 Cut

The diamond Cut measures the quality of diamond cutting, which determines how a diamond reflects light, thus affecting its brightness, fire and scintillation.The diamond cut in the data set is a categorical variable with four categories: Excellent, Very Good, Good, and Ideal. According to the grading systems of GIA and AGS, cut quality decreases from Ideal to Good.

- The Ideal cut with nearly all incoming light reflecting through the diamond's top to maximize brilliance and sparkle. It features balanced and prominent fire and scintillation, with precisely cut crown and pavilion angles to achieve optimal light refraction. Proportional standards, such as table percentage and pavilion depth percentage, meet ideal criteria. The facets are perfectly aligned with no visible deviations, and the surface is finely polished, free from any scratches or blemishes.
- Excellent cut diamonds reflect nearly all light through the top, showcasing maximum brilliance and fire. As the high cut grade, they exhibit optimal optical performance and exceptional visual appeal.
- Very Good cut diamonds reflect most of the light through the top, though a small amount may escape from the sides or bottom and their brilliance and fire are slightly less than those of Excellent cut diamonds.

- Good cut diamonds exhibit noticeably reduced light refraction, with some light escaping from the sides or bottom and their brilliance and fire are not as strong as higher-grade cuts.

The barplot on the left of Figure 5 shows the distribution of diamonds with different cutting grades. It indicates that the number of diamonds with Excellent and Very Good grades is obviously more, indicating that most diamonds on the market are concentrated in these high-cut grades, which may be because they have better visual effects and higher market demand. In contrast, the number of Good and Ideal diamonds is small, especially the rarity of Ideal cutting may reflect strict cutting ratio requirements and high quality standards. The violin chart on the right of Figure 5 shows the price distribution of diamonds with different cutting grades. It shows the price range of each cutting grade is very close to the median, which shows that the cutting grade has little direct influence on the price. However, Ideal and Excellent diamonds are more obviously distributed in the high-end price range, which may be because these two grades of diamonds are usually combined with other high-quality characteristics, such as high carat number or clarity. At the same time, the overall price range is relatively large, indicating that other factors (such as carats, colors, etc.) may play a more important role in determining the price.

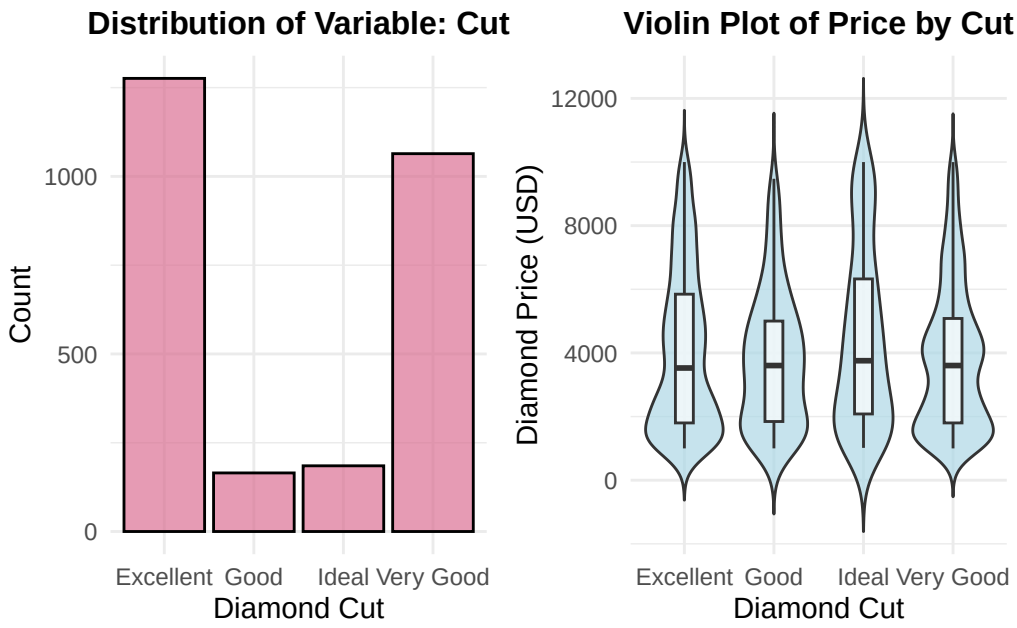

Figure 5: Graphs of Predictor Variable: Cut

2.5 Analysis of Correlation

Figure 6 shows several key insights about the relationships between diamond attributes and price. Carat size shows a strong positive correlation with price, indicating that it is one of the primary factors influencing diamond value—larger diamonds tend to be more expensive. In contrast, cut (e.g., cutExcellent, cutVery Good) demonstrates a weaker correlation with price, suggesting that cut grade alone has a limited direct impact on diamond cost but may work in conjunction with other factors such as carat size or color. Similarly, color (e.g., colorD to colorK) shows a weak correlation with price, with diamonds closer to colorless (e.g., colorD and colorE) potentially commanding higher prices, but the overall effect is minimal. Clarity (e.g., clarityIF to claritySI2) also exhibit a modest relationship with price, reflecting their role as a contributing but less dominant factor in determining diamond value. Overall, carat size emerges as the most influential attribute, while other variables like cut, color, and clarity play supporting roles. Figure 6 also presents the relationships between the predictor variables, showing that the correlations among them are generally weak. For example, attributes like cut quality, clarity, color, and carat size do not exhibit strong intercorrelations, as most of the corresponding cells are closer to white or light blue. This suggests that the predictor variables are relatively independent, reducing the risk of multicollinearity in the analysis. The weak correlations among these variables ensure that each contributes uniquely to the model, providing a robust foundation for predicting diamond price.

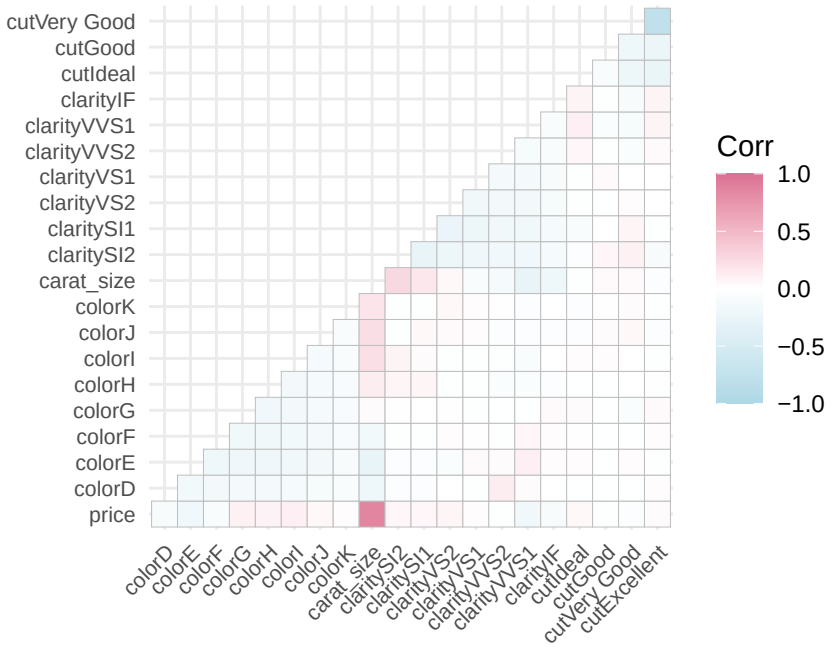

Figure 6: Correlation: Heat Graph

3 Model

The goal of our modeling strategy is twofold: to assess the contributions of key factors—such as carat size, diamond color, cut, and clarity—to variations in the standardized and transformed diamond price, and to provide accurate predictions of diamond prices based on these features. We use generalized linear model (GLM) implemented using the `glm` function in R Core Team (2023). The GLM model investigates the relationship between standardized, transformed diamond prices and predictors like carat size, color, cut, and clarity, using the Gamma family with a log link function. The dataset is divided into training and testing sets, with 70% of the data allocated for model fitting and parameter estimation, and the remaining 30% reserved for evaluating predictive accuracy.Background details, model evalution, diagnostics and model predictions are included in Appendix A.

3.1 Data Preprocession

Before building the model, the data underwent several preprocessing steps to ensure compatibility and improve model performance. The response variable `price` was standardized because its values were significantly larger compared to other continuous numeric variables 'carat_size', which could disproportionately influence the model. After scaling, the response variable was transformed using the exponential function to ensure that `price` remained positive, making it more interpretable and suitable for subsequent analysis.Then, categorical variables, including `color`, `clarity`, and `cut`, were explicitly converted into factors to enable proper handling in the GLM model. Additionally, the levels of these categorical variables were reordered to define reference levels: - Variable 'color' have reference level 'K'(worst color) - Variable 'clarity' have reference level 'SI2' (worst clarity) - Variable 'cut' have reference level 'Good' (worst cut)

3.2 Model set-up

The GLM relies on several key assumptions. Firstly, we assume the response variable y_i (diamond prices) is assumed to follow a Gamma distribution with mean μ_i and dispersion parameter ϕ. Then, we assume there is a linear relationship between the predictors and the logarithm of the mean price $\log(\mu_i)$ in the transformed space. The assumptions underlying this GLM model are as follows (1) and (2):

$$y_i \mid \mu_i, \phi \sim \text{Gamma}(\mu_i, \phi), \tag{1}$$

where μ_i is the mean price and ϕ is the dispersion parameter.

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \sum_j \beta_{2j} \cdot \text{color}_{ij} + \sum_k \beta_{3k} \cdot \text{cut}_{ik} + \sum_l \beta_{4l} \cdot \text{clarity}_{il}. \quad (2)$$

Then, based on these assumptions of GLM, we get the final generalized linear model (3):

$$\begin{aligned} \log(\text{standardized and exponentiated price}) = \beta_0 + \beta_1 \cdot \text{carat\_size} \\ + \sum_{i=\text{E}}^{\text{K}} \gamma_i \cdot I(\text{color} = i) \\ + \sum_{j=\text{Good}}^{\text{Ideal}} \delta_j \cdot I(\text{cut} = j) \\ + \sum_{k=\text{SI2}}^{\text{IF}} \theta_k \cdot I(\text{clarity} = k). \end{aligned} \quad (3)$$

- β_0: Intercept term.
- β_1: Coefficients of carat size, indicating the influence of carat weight of diamonds on the logarithmic space of price.

- γ_i: Coefficient of color in category i, where i can be D, E, F, G, H, I, J, K.
- δ_j: Coefficient of cut in category j, where j can be Ideal, Excellent, Very Good, Good.
- θ_k: Coefficient of clarity in category k, where k can be IF, VVS1, VVS2, VS1, VS2, SI1, SI2.
- I: Indicator function, used for dummy variable coding of classified variables. Its function is to judge whether the condition in parentheses is true. If the condition is true, the function takes the value of 1; otherwise, it takes the value of 0.

Since the response variable `price` is standardized and exponential-transformed and log-transformed during modeling, predicting the price requires reversing these transformations by only applying the inverse of the standardization because exponential-transformed and log-transformed cancel each other. Thus, the formula for predicting the price is (4):

$$\text{price} = \log(\text{price}) \cdot \sigma + \mu. \quad (4)$$

- σ: the standard deviation of price in the training data, which is 2442.597.
- μ: the mean of price in the training data, which is 4010.706.

3.3 Model justification

We expect a positive relationship between predictors (carat size, cut, color and clarity) and response variable price. To test this relationship, we chose the Generalized Linear Model (GLM) with a Gamma distribution and a log link function. For categorical predictors like color, clarity, and cut, by setting the lowest quality levels (e.g., K for color, SI2 for clarity, and Good for cut) as reference levels, we can directly interpret the coefficients as the incremental effect of higher-quality levels on diamond prices. For continuous predictor carat size, we can interpret the coefficients in the model as how it influence diamond prices. This setup also aligns with our expectation that higher-quality diamonds (in terms of color, clarity, and cut) will positively influence price.

The Generalized Linear Model (GLM) was chosen for its flexibility and ability to handle diverse response variable distributions. Unlike traditional linear models, which require normally distributed response variables, the GLM accommodates non-normal distributions, making it particularly well-suited for our data where the response variable, diamond price, follows a Gamma distribution, as shown in Figure 1. By specifying the log link function, the GLM captures the nonlinear relationship between the response variable and predictors, allowing for better modeling of the price variability while maintaining interpret ability and precision.

Despite its advantages, the GLM has potential limitations and may not be suitable in certain scenarios. When the distribution of response variable is hard to observe, we may not able to assume its distribution. Furthermore, highly nonlinear relationships may not be adequately captured by the specified link function, requiring alternative modeling approaches.

Furthermore, to validate and evaluate the model, a test dataset was used to calculate key metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Error (MSE). These metrics quantify the model's predictive accuracy and help identify potential areas of improvement. And the predicted values and the actual values are visualized and diagnostics are presented. These details are provided in Section A.

3.4 Model Comparision: Generalized Linear Model vs. Linear Model vs.Bayesian regression

In addition to the generalized linear model, we also explore and compare alternative models, including linear regression and Bayesian regression.

a. Linear Regression Model:

The advantage of Linear model is that the results are easy to explain, and it can provide accurate parameter estimation under the condition of satisfying basic assumptions (such as linear relationship, normality, independence and variance homogeneity). However, its disadvantage is that the assumptions of the data are strict, and when the data violates these assumptions, the estimation of the model may be invalid. We use linear regression model implemented using

the `lm` function in R Core Team (2023) and use training dataset for model fitting and test dataset for model evaluation.

b. Bayesian regression Model:

The advantage of Bayesian model lies in its flexibility, which is especially suitable for scenes with less data, complex distribution or the need to combine prior knowledge. However, the disadvantage of Bayesian model lies in its high computational complexity, and improper selection of prior distribution may significantly affect the inference results. We also use the Bayesian Regression model to analyze the relationship between diamond prices and various predictors, implemented using the `stan_glm` function from the rstanarm package (Goodrich et al. (2024)) in R Core Team (2023). The model assumes a Gamma family with a log link function, and the priors are specified as weakly informative: normal distributions (mean = 0, standard deviation = 10) for both the coefficients and intercept, and an exponential distribution (rate = 3) for the auxiliary parameter.

c. Comparision Result:

Table 2 shows that the Generalized Linear Model (GLM) outperforms the other models in both predictive accuracy and model parsimony. The GLM achieved the lowest Mean Squared Error (MSE) of 1.627, Mean Absolute Error (MAE) of 0.517, and Root Mean Squared Error (RMSE) of 1.276, indicating superior performance in predicting the scaled and transformed diamond prices. Furthermore, it also has the lowest Akaike Information Criterion (AIC) of 962.513 and Bayesian Information Criterion (BIC) of 1068.332, suggesting that it provides a more parsimonious fit to the data compared to the other models. In contrast, the Linear Model (LM) and Bayesian model demonstrated notably poorer performance. The LM exhibited an MSE of 2.231, an MAE of 0.954, and an RMSE of 1.494, while the Bayesian model displayed an MSE of 6.255, an MAE of 1.775, and an RMSE of 2.501, all significantly higher than the GLM's metrics. Additionally, the AIC and BIC scores for the LM were 6685.589 and 6791.408, respectively, reflecting both higher complexity and inferior predictive power. The Bayesian model, however, lacked calculable AIC and BIC values. The Bayesian model's AIC and BIC values are `NA` because `rstanarm` does not natively calculate these metrics. AIC and BIC are traditionally based on maximum likelihood estimation (MLE), which is not directly applicable to Bayesian models that rely on posterior distributions rather than MLE. Instead, Bayesian models are typically evaluated using metrics like the Deviance Information Criterion (DIC) or the Watanabe-Akaike Information Criterion (WAIC), which are designed to work with posterior distributions.

Table 2: Model Comparison Metrics for GLM, LM, and Bayesian Models

Model	MSE	MAE	RMSE	AIC	BIC
GLM	1.627	0.517	1.276	962.513	1068.332
LM	2.231	0.954	1.494	6685.589	6791.408

Table 2: Model Comparison Metrics for GLM, LM, and Bayesian Models

Model	MSE	MAE	RMSE	AIC	BIC
Bayesian	6.272	1.780	2.504	NA	NA

4 Results

Figure 7 shows the relative contribution of carat weight, color, cut and clarity to the diamond price after standardization and exponential transformation and logarithmic transformation. And Figure 7 also visualizes the estimated regression coefficient of each variable and its 95% confidence interval. We can find: - Carat Size: Carat weight has the most significant positive impact on diamond price, with a regression coefficient of 3.73. This indicates that as carat weight increases, the diamond price rises, which aligns with expectations.

- Color: Higher-grade colors (e.g., D) have the largest positive impact on diamond prices, with a coefficient of 1.343 for color D compared to the baseline (K, the lowest grade). Lower-grade colors (e.g., J and I) have progressively smaller positive coefficients, suggesting that price increases are more obvious for diamonds with higher color grades.
- Clarity: Similar to color, higher clarity grades (e.g., IF and VVS1) positively influence diamond prices. For instance, the coefficient for clarity IF is 0.926, while VVS1 is 0.896, compared to the baseline clarity grade (SI2). Lower clarity grades (closer to the baseline) have relatively smaller positive coefficients, showing their lesser contribution to price increases.
- Cut: Cut quality has a smaller impact compared to other factors. For example, the "Ideal" cut has a coefficient of 0.216, which indicates a slight positive impact compared to the baseline ("Good" cut). Similarly, "Excellent" and "Very Good" cuts contribute marginally to price increases.
- Intercept: The intercept of -4.805 represents the baseline value of diamond price (after transformation) when all predictors are at their reference levels (i.e., K for color, SI2 for clarity, and Good for cut).

In summary, carat weight has the most significant impact on diamond prices, followed by clarity and color, while cut quality has a relatively smaller effect. These results suggest that buyers prioritize carat weight and the diamond's visual attributes (color and clarity) over the precision of the cut.

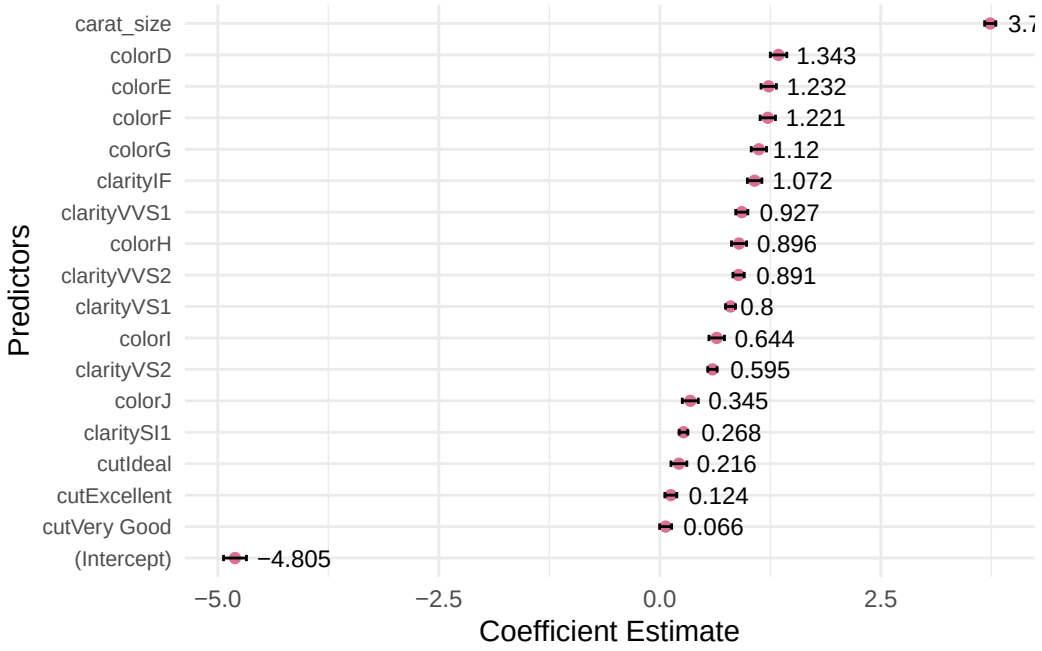

Figure 7: Model Results

4.1 Example of Prediction

5 Discussion

5.1 Exploration of Model Result

This study examines how key features of diamonds, such as carat size, color, clarity, and cut, influence their prices. The results show that carat weight has the most significant impact on diamond prices, followed by clarity and color, while cut quality has a relatively smaller effect. These findings highlight the preferences and values that drive consumer behavior and shape industry practices. By understanding the role of each predictor, we can explore the deeper social and cultural meanings behind these choices, such as the importance of visibility, aesthetics, and societal expectations in defining value.

Carat size has the strongest impact on price. Larger diamonds are often associated with wealth, success, and love, making them highly desirable for significant events like engagements and weddings. This focus on size shows how people value physical attributes that are easily noticeable, often linking material possessions to social status and personal achievements. Economically, the preference for size drives demand for larger diamonds, encouraging industries to focus on producing and marketing them as rare and valuable.

Color and clarity also have a strong effect on price. Higher grades in these features lead to much higher prices, which shows that buyers are willing to pay more for diamonds that look perfect or flawless. Even though these differences might not always be noticeable to the average person, marketing and cultural standards make these traits seem important. This focus influences how diamonds are graded and priced, creating a market where beauty and perfection are key drivers of value.

Cut, while important for a diamond's brilliance, has a smaller impact on price compared to the other features. This suggests that buyers may undervalue the craftsmanship behind the cut, focusing instead on features like size and color that are easier to see and compare. This pattern is common in luxury markets, where visible traits often matter more than technical quality or skill.

In conclusion, diamond prices are influenced by a mix of consumer behavior, cultural values, and industry practices. Buyers tend to value features that are visible and easy to compare, driving demand and shaping market trends. This study shows how societal values influence pricing and production, not only in the diamond market but also in other areas of the luxury goods industry.

5.2 Limitations

5.2.1 Data Limitations

Although the data set is valuable, there are several limitations, which affect the universality and depth of the research results. A key limitation is the limited price range, which only includes diamonds with a price between $1,000 and $10,000. This does not include low-priced and high-end luxury diamonds, which may miss the unique pricing model of these market segments. In addition, the price distribution is seriously tilted to the right, and most of the prices are concentrated between 2500 and 5000 dollars. This imbalance may lead to deviation in the analysis, because diamonds with higher prices may be affected by different market factors and their representation is insufficient.

Another limitation is that the data set narrowly focuses on physical properties, such as carat size, color, clarity and cutting, while ignoring background factors such as brand reputation, certification sources and market trends. These external factors will significantly affect pricing, but they are not taken into account in the data. In addition, the data set lacks geographical and cultural details, which can provide insights into regional differences in preferences and needs. This omission limits the global applicability of the research results, because diamond pricing often varies with the market and consumer base.

The dataset also simplifies some predictor variables, particularly cut quality, by classifying them into broad groups (Ideal, Excellent, Very Good, and Good) without accounting for finer details like proportions or symmetry. These details, such as the arrangement of facets or crown angles, are crucial for evaluating a diamond's brilliance and could provide a deeper explanation

for price variations. Similarly, the color and clarity variables are not fully represented. For instance, diamond color is graded from D (colorless) to Z (light yellow or brown), but the dataset only includes colors from D to K, leaving out lower-grade colors that could provide additional insights. Moreover, the absence of time-related factors limits the ability to analyze how diamond prices change over time due to shifts in consumer preferences, market trends, or seasonal effects.

Finally, the data set shows the potential problems of imbalance and lack of diversity. For example, a lower quality cut (good) and a lower color grade (k) are insufficient, which may lead to distorted results and limit the ability to draw conclusions about these groups. In addition, if the data set comes from a single retailer or region, the results may reflect localized pricing strategies or consumer preferences, rather than broader market trends, thus reducing its generalization to other environments.

5.2.2 Model Limitations

One limitation of the GLM model is its inability to fully capture the complexity of diamond price variability due to its fixed distributional assumptions. The use of the Gamma distribution with a log link function, while appropriate for positive and skewed data, may oversimplify the true underlying patterns in the data. This is particularly evident in higher price ranges, where the model underestimates prices.

Another limitation is in how the predictors are handled. The model treats clarity, color, and cut as categorical variables divided into broad categories, which simplifies their complexity. For example, subtle variations within a clarity grade or interactions between attributes, like how cut quality might impact price differently depending on carat size, are not accounted for. These oversimplifications may prevent the model from capturing more detailed relationships, potentially reducing its predictive accuracy.

5.3 Futurer Direction

To improve the depth and reliability of future research, several enhancements to the dataset and model can be considered. Expanding the price range to include both low-cost and luxury diamonds would provide a broader view of pricing patterns across the market. Additionally, addressing imbalances by incorporating underrepresented categories, such as lower-quality cuts and color grades, would reduce bias and improve generalizability. We could also include additional predictors, such as cut proportions, symmetry, and time, to enable a more detailed analysis of pricing factors and facilitate the study of price trends over time, including seasonal effects, and market fluctuations. The model can be improved by adopting more flexible approaches, such as generalized additive models (GAMs) or Bayesian hierarchical models, which can better handle non-linear relationships and interactions between predictors. These methods are particularly useful for capturing complex patterns in data, especially in higher price

ranges. And to address variability and outliers, techniques like robust regression or weighted least squares can help manage non-constant variance in residuals, while carefully identifying and examining influential points can reduce their impact on predictions.

Appendix

A Additional Model details

A.1 Predicted Diamond Prices

Figure 8 compares the predicted prices from the GLM model (blue line) with the scaled and transformed prices from the test dataset (red points). It is important to note that the prices shown in the plot are not the actual diamond prices. The response variable, `price`, was first scaled using z-score normalization and then exponentiated. Additionally, the predictions from the GLM model were transformed using a log link function, which is consistent with the assumptions of the Gamma distribution.

The GLM model captures the general trend of the transformed prices well, especially in the lower and middle ranges. The predicted values closely follow the transformed prices in these areas, indicating that the GLM model performs well for most of the data. However, at higher transformed price levels, the predictions tend to underestimate the values, as indicated by the wider spread of red points compared to the smoother blue line. The variability of the transformed prices increases at higher indices, but the model does not fully capture this. This suggests that the GLM model, with its fixed Gamma distribution and log link function, may oversimplify the behavior of the response variable in these regions. The underestimation at higher levels highlights potential limitations in the model's assumptions.

In summary, while the GLM model effectively models the scaled and transformed price data for most observations, it struggles to account for the increasing variability in higher ranges. Further refinements, such as addressing variability or investigating the suitability of the distributional assumptions, could improve the model's performance.

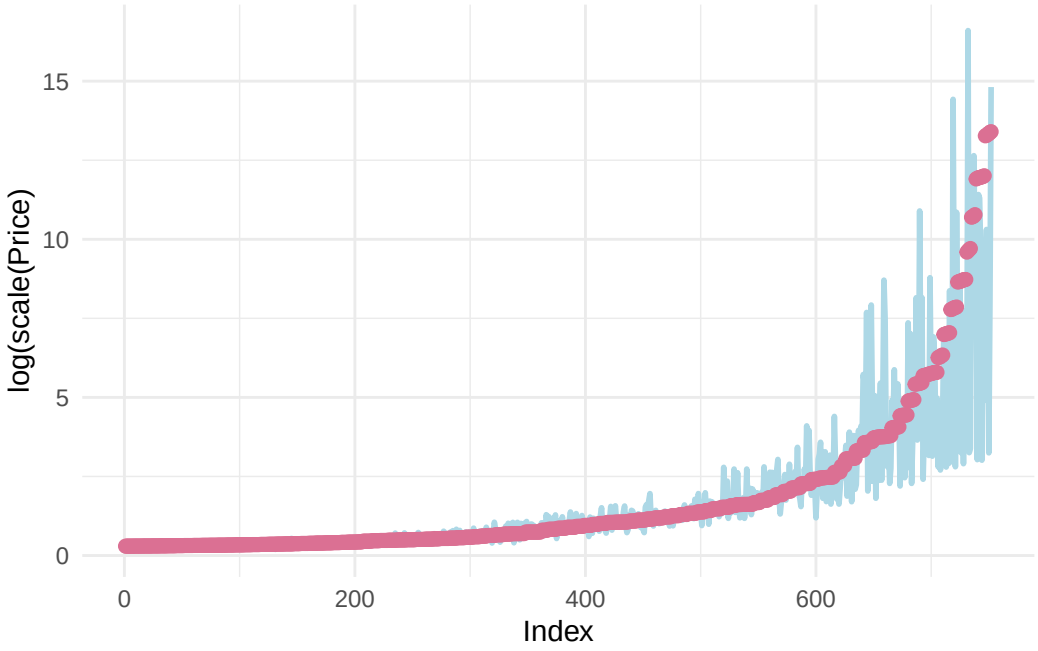

Figure 8: Actual vs Predicted Values

A.2 Evaluation Metrics

Figure 9 shows how the Generalized Linear Model (GLM) was evaluated on the test dataset using three key metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The MSE of 1.627 indicates that the overall squared differences between the predicted and actual diamond prices are relatively small, suggesting that the model captures the relationship between the predictors and the response variable reasonably well. However, since MSE amplifies the impact of larger errors, it is complemented by RMSE for better interpret ability.The MAE of 0.517 shows that, on average, the absolute difference between the predicted and actual values is low. This highlights the model's strong ability to provide accurate predictions in most cases and reflects that its performance is not heavily skewed by outliers. Compared to MSE, MAE is less sensitive to extreme errors, offering a more balanced perspective on prediction accuracy. The RMSE of 1.276, being the square root of MSE, provides a direct measure of the standard deviation of prediction errors in the original data's units. While RMSE tends to be slightly higher than MAE because it gives more weight to larger errors, its relatively low value confirms the model's stability and reliable performance on the test data. However, the gap between RMSE and MAE suggests the presence of some larger errors or outliers that may require further investigation.Overall, these metrics indicate that the GLM performs well in predicting diamond prices, with reasonable accuracy and controlled error levels.

Metric	Value
MSE	1.627
MAE	0.517
RMSE	1.276

Figure 9: GLM Model Metrics

A.3 Diagnostics

Figure 10 from the GLM model shows how well the model performs and whether it follows the key assumptions of the Gamma distribution with a log link.

The "Residuals vs. Fitted" plot indicates a slight curvature in the residuals, particularly at lower fitted values, suggesting potential non-linearity or slight misspecification in the link function. Additionally, the spread of residuals increases slightly for higher fitted values, hinting at heteroscedasticity (non-constant variance). Ideally, residuals should be randomly scattered around zero with no discernible pattern.

The Q-Q plot assesses whether the residuals conform to the theoretical Gamma distribution. While most points follow the diagonal line, deviations are observed in the upper quantiles, with several extreme residuals identified (e.g., points 1930, 1899). These deviations shows that the model may not fully capture the response variable's distribution or that outliers are present.

The Scale-Location plot further supports evidence of heteroscedasticity, as the variance of residuals increases with higher fitted values. Ideally, the points should be evenly spread around the horizontal line, but the observed increase in variance indicates that the model might benefit from additional adjustments to better handle variability in the response variable.

The "Residuals vs. Leverage" plot highlights a few potentially influential points (e.g., 1930, 1899, 1924) with higher leverage. Although none of these points exceed Cook's distance thresholds, their presence warrants closer examination to ensure they do not disproportionately affect the model.

In summary, the model performs well but shows slight heteroscedasticity and deviations from the Gamma distribution. Addressing these issues and checking outliers may improve reliability and accuracy.

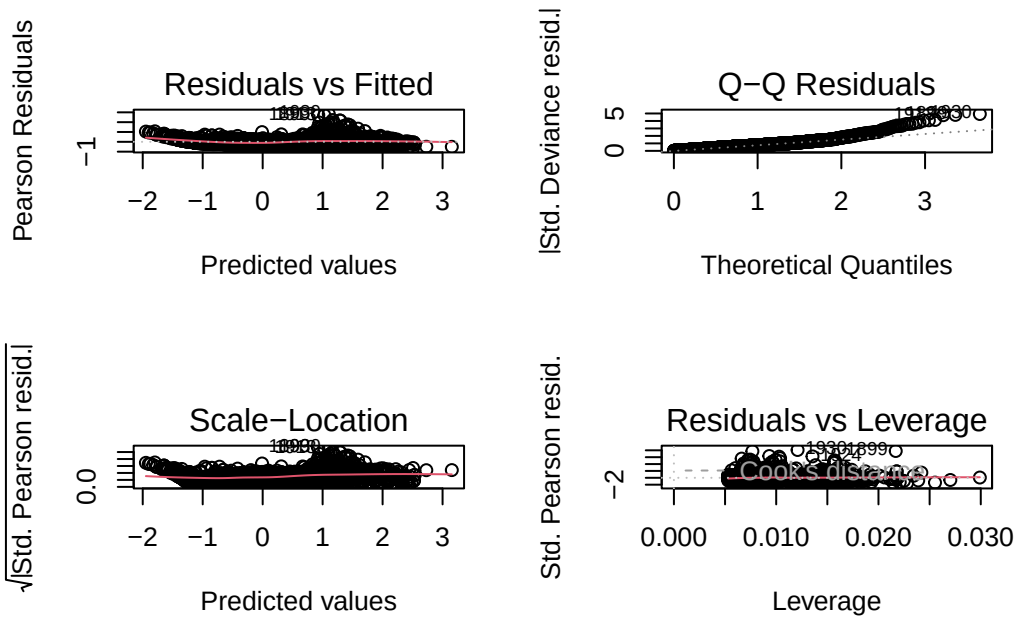

Figure 10: Diagnostics

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Data, and Story Library (DASL). 2024. "Diamonds Dataset." https://dasl.datadescription.com/datafile/diamonds/?_sfm_methods=Multiple+Regression&_sfm_cases=1000+59943.

Gemological Institute of America. n.d. "Grading the Diamond 4Cs." https://4cs.gia.edu/en-us/grading-diamond-4cs/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Kassambara, Alboukadel. 2018. "Ggcorrplot: Visualization of a Correlation Matrix Using 'Ggplot2'." https://github.com/kassambara/ggcorrplot.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Pedersen, Thomas Lin. 2020. "Patchwork: The Composer of Plots." https://patchwork.data-imaginist.com/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan,

Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.