my title*

subtitle

Yawen Tan

December 2, 2024

4 sentence

Table of contents

1 Introduction **1**

2 Data **1**

2.1 Overview 1
2.2 Measurement 2
2.3 Outcome Variable: Price 2
2.4 Predictor Variables 3
2.4.1 Color 3
2.4.2 Carat Size 4
2.4.3 Clarity 5
2.4.4 Cut 7
2.5 Analysis of Correlation 9

3 Model **10**

3.1 Model set-up 10
3.2 Model justification 11
3.3 Model Comparision: Generalized Linear Model vs. Linear Model 11
3.4 Model Evaluation 11

4 Results **11**

4.1 Model Result 11
4.2 Predictions vs Actual Value 13
4.3 Example of Prediction 13

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

5 Discussion **13**
5.1 First discussion point 13
5.2 Limitations 14
5.2.1 Data Limitations 14
5.3 Further Considerations 14

Appendix **15**

A Additional data details **15**
A.1 Data Cleaning 15

B Additional Model details **15**
B.1 Evaluation Metrics and Diagnostic Table 15
B.2 Feature importance analysis 15
B.3 Diagnostics 15

References **17**

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.

We use the statistical programming language R (R Core Team 2023)…. Our data (Toronto Shelter & Support Services 2024)…. Following Alexander (2023), we consider…

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

Table 1: Summary of Diamond Dataset

Attribute	Range or Levels	Data Type	Example
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.

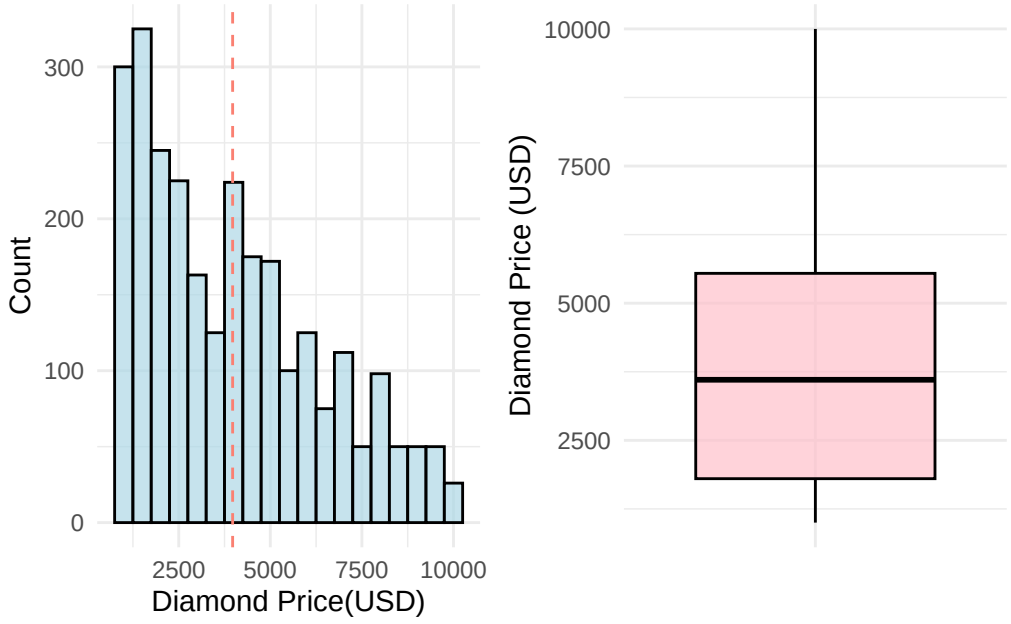

Figure 1: Graphs of Diamond price

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.

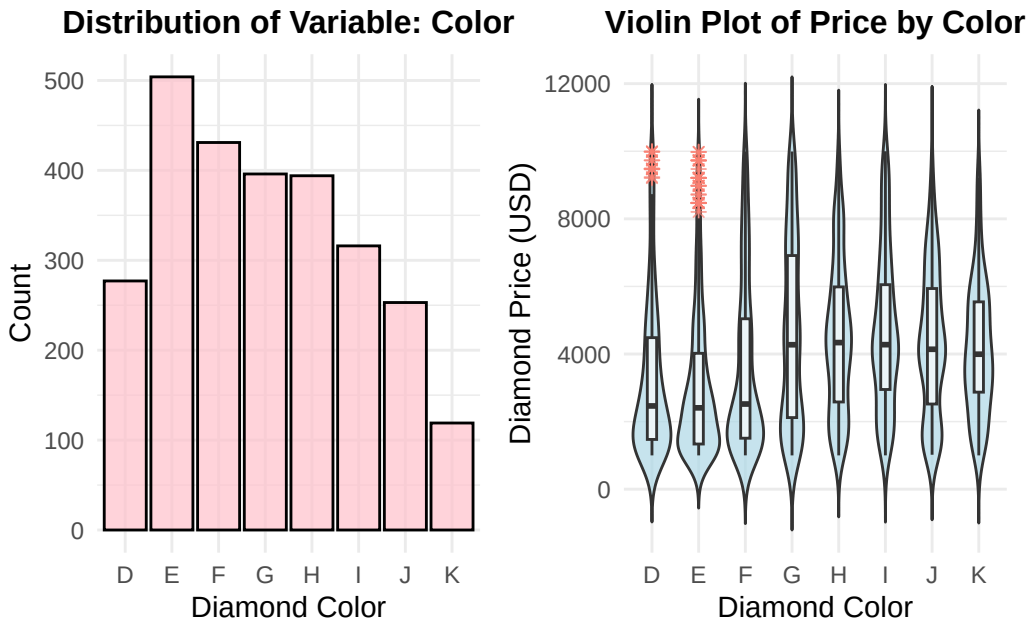

Figure 2: Graphs of Predictor Variables: Color

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.

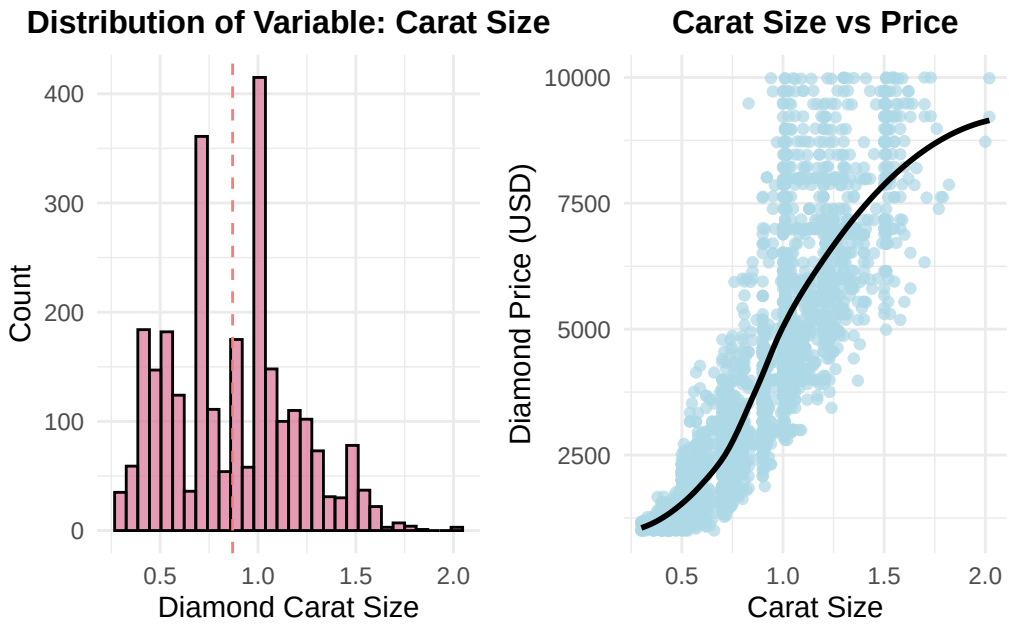

Figure 3: Graphs of Predictor Variable: Carat Size

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.

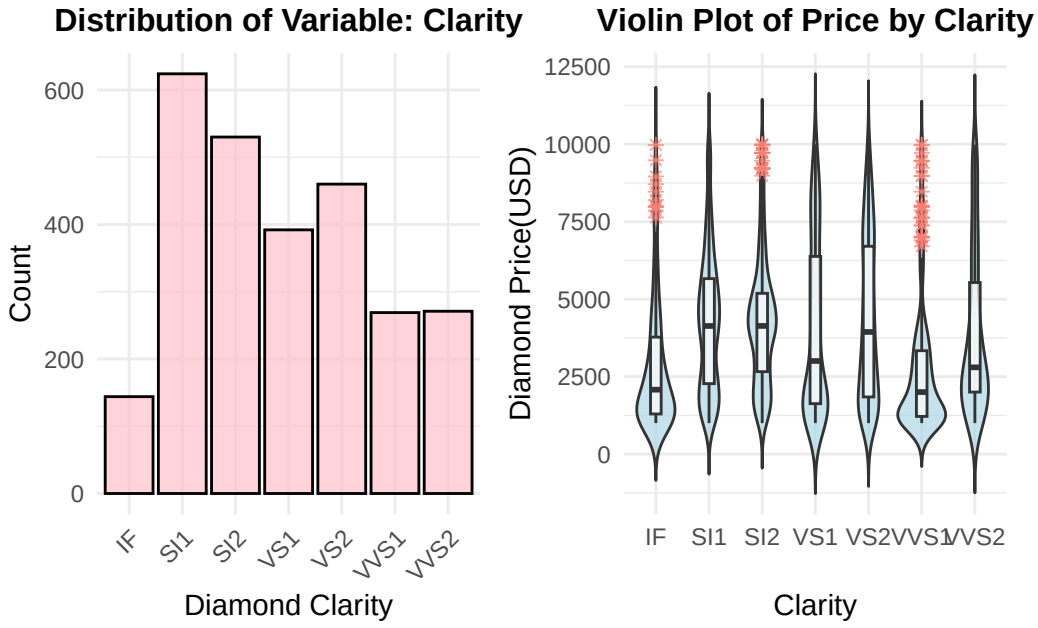

Figure 4: Graphs of Predictor Variable: Clarity

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.

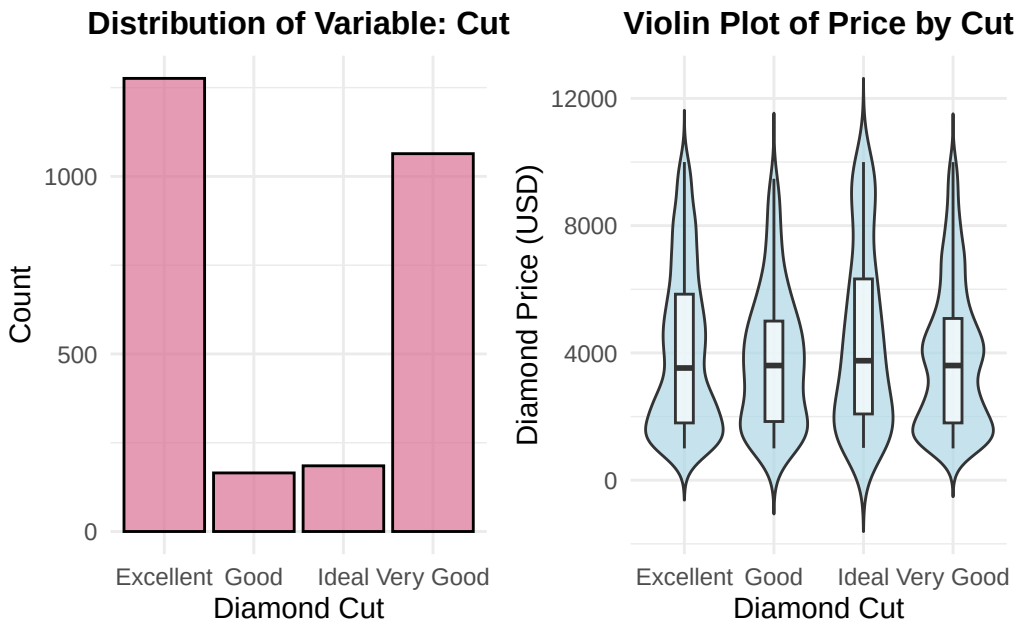

Figure 5: Graphs of Predictor Variable: Cut

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.

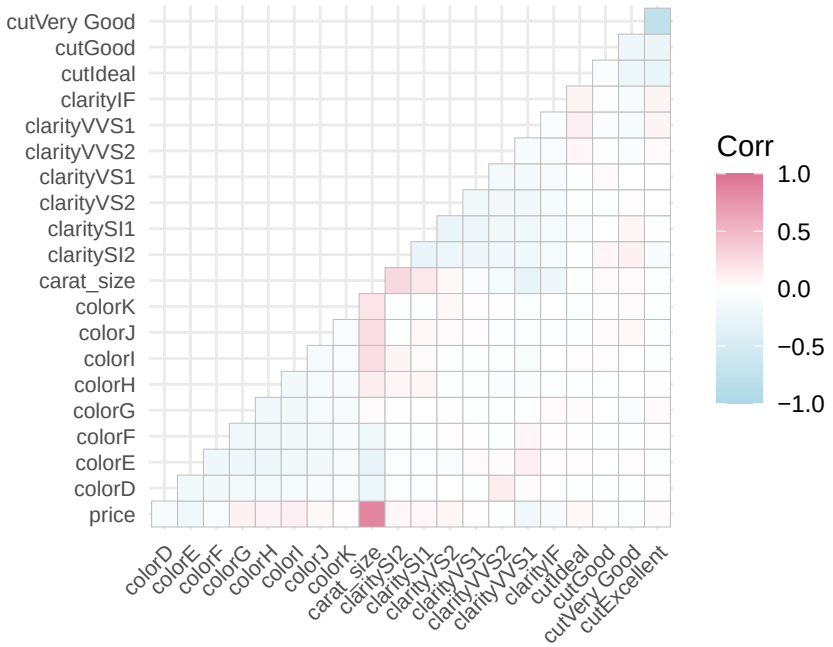

Figure 6: Correlation: Heat Graph

3 Model

3.1 Model set-up

Define y_i as the price of the i th diamond. Predictors include:

x_{1i}: The carat size of the diamond (in carats). x_{2i}: The color grade of the diamond (categorical). x_{3i}: The cut grade of the diamond (categorical). x_{4i}: The clarity grade of the diamond (categorical).

1. **Response Distribution**:

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

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

2. **Link Function**:

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

   - β_0: Intercept term.
   - β_1: Effect of carat size.
   - β_{2j}, β_{3k}, β_{4l}: Effects of j different color, k different cut, and l different clarity, respectively.

3. **Mean Price**:

$$\begin{aligned}\mu_i = \exp\Bigg(\beta_0 + \beta_1 x_{1i} + \sum_j \beta_{2j} \cdot \text{color}_{ij} \\ + \sum_k \beta_{3k} \cdot \text{cut}_{ik} + \sum_l \beta_{4l} \cdot \text{clarity}_{il}\Bigg)\end{aligned} \tag{3}$$

Table 2: Model Reulst: Prediction of Diamond Price based on Color, Cut, Clarity and Carat Size

```
AIC: 962.5128

BIC: 1068.332
```

3.2 Model justification

glm gamma data section reponse

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

3.3 Model Comparision: Generalized Linear Model vs. Linear Model

3.4 Model Evaluation

See appendix

4 Results

4.1 Model Result

```
Call:
glm(formula = price ~ carat_size + color + cut + clarity, family = Gamma(link = "log"),
    data = train_data)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.26556    0.04467 -50.723  < 2e-16 ***
carat_size    3.73719    0.03221 116.036  < 2e-16 ***
colorE       -0.11152    0.03155  -3.534 0.000418 ***
colorF       -0.12186    0.03236  -3.765 0.000171 ***
colorG       -0.22282    0.03372  -6.607 5.06e-11 ***
colorH       -0.44708    0.03377 -13.238  < 2e-16 ***
colorI       -0.69979    0.03620 -19.329  < 2e-16 ***
```

```
colorJ       -0.99835    0.03901 -25.594  < 2e-16 ***
colorK       -1.34329    0.04750 -28.281  < 2e-16 ***
cutGood      -0.12447    0.03443  -3.615 0.000308 ***
cutIdeal      0.09145    0.03356   2.725 0.006490 **
cutVery Good -0.05886    0.01776  -3.313 0.000939 ***
claritySI1   -0.80451    0.04029 -19.966  < 2e-16 ***
claritySI2   -1.07216    0.04176 -25.677  < 2e-16 ***
clarityVS1   -0.27249    0.04124  -6.608 5.03e-11 ***
clarityVS2   -0.47686    0.04085 -11.674  < 2e-16 ***
clarityVVS1  -0.14500    0.04306  -3.368 0.000773 ***
clarityVVS2  -0.18124    0.04334  -4.181 3.03e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.1241234)

    Null deviance: 2194.94  on 1937  degrees of freedom
Residual deviance:  180.03  on 1920  degrees of freedom
AIC: 962.51

Number of Fisher Scoring iterations: 8
```

Table 3: Model Reulst: Prediction of Diamond Price based on Color, Cut, Clarity and Carat Size

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.266	0.045	-50.723	0.000
carat_size	3.737	0.032	116.036	0.000
colorE	-0.112	0.032	-3.534	0.000
colorF	-0.122	0.032	-3.765	0.000
colorG	-0.223	0.034	-6.607	0.000
colorH	-0.447	0.034	-13.238	0.000
colorI	-0.700	0.036	-19.329	0.000
colorJ	-0.998	0.039	-25.594	0.000
colorK	-1.343	0.047	-28.281	0.000
cutGood	-0.124	0.034	-3.615	0.000
cutIdeal	0.091	0.034	2.725	0.006
cutVery Good	-0.059	0.018	-3.313	0.001
claritySI1	-0.805	0.040	-19.966	0.000
claritySI2	-1.072	0.042	-25.677	0.000
clarityVS1	-0.272	0.041	-6.608	0.000

Table 3: Model Reulst: Prediction of Diamond Price based on Color, Cut, Clarity and Carat Size

	Estimate	Std. Error	t value	Pr(>|t|)
clarityVS2	-0.477	0.041	-11.674	0.000
clarityVVS1	-0.145	0.043	-3.368	0.001
clarityVVS2	-0.181	0.043	-4.181	0.000

4.2 Predictions vs Actual Value

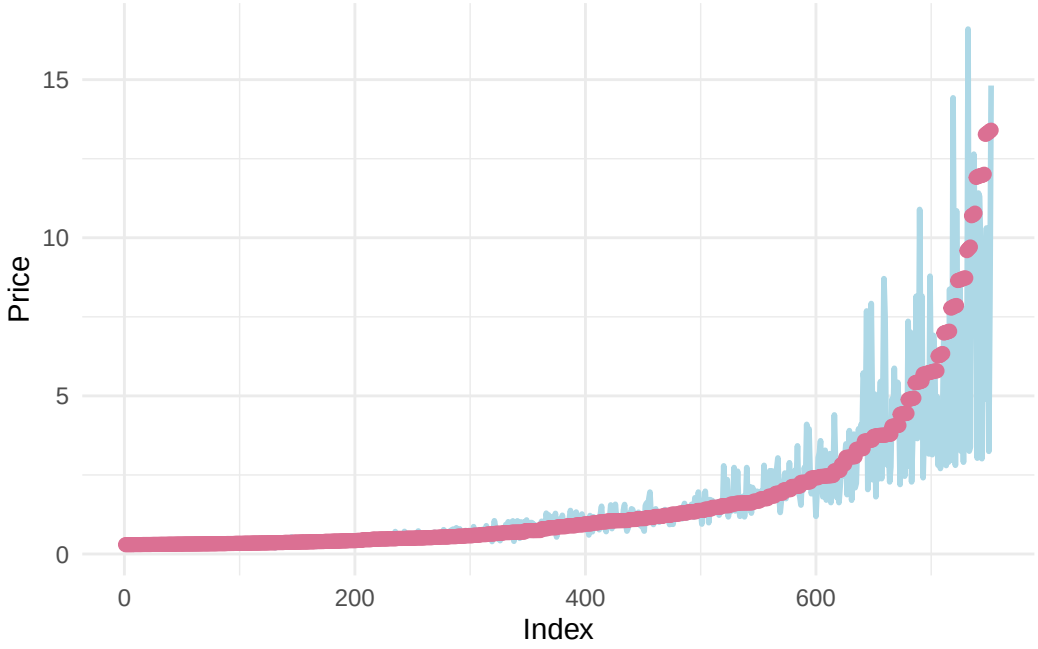

Figure 7: Actual vs Predicted Values

4.3 Example of Prediction

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Limitations

5.2.1 Data Limitations

Model Limitations{#sec-modellimit}

5.3 Further Considerations

Appendix

A Additional data details

A.1 Data Cleaning

data section ## Data Preparation and Data Split{#sec-dataprepare} Model Section response variable train test to fit the mode

B Additional Model details

B.1 Evaluation Metrics and Diagnostic Table

```
### Test Data Results
```

Table 4: Performance Metrics on Test Data

Metric	Value
MSE (Test)	1.627
MAE (Test)	0.517

Figure 8: Actual vs Predicted Values

```
### Dispersion Check
```

Table 5: Overdispersion Check

Metric	Value
Dispersion Ratio	0.094

Figure 9: Actual vs Predicted Values

B.2 Feature importance analysis

B.3 Diagnostics

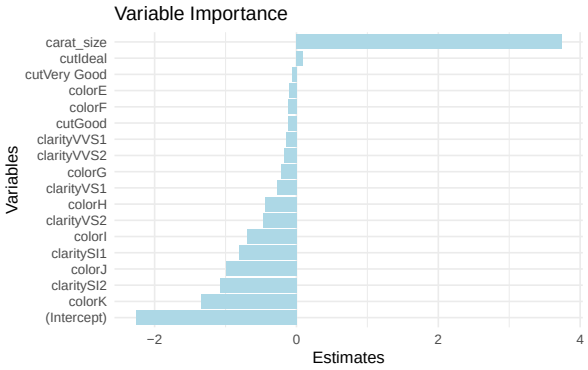

Figure 10: Feature importance analysis

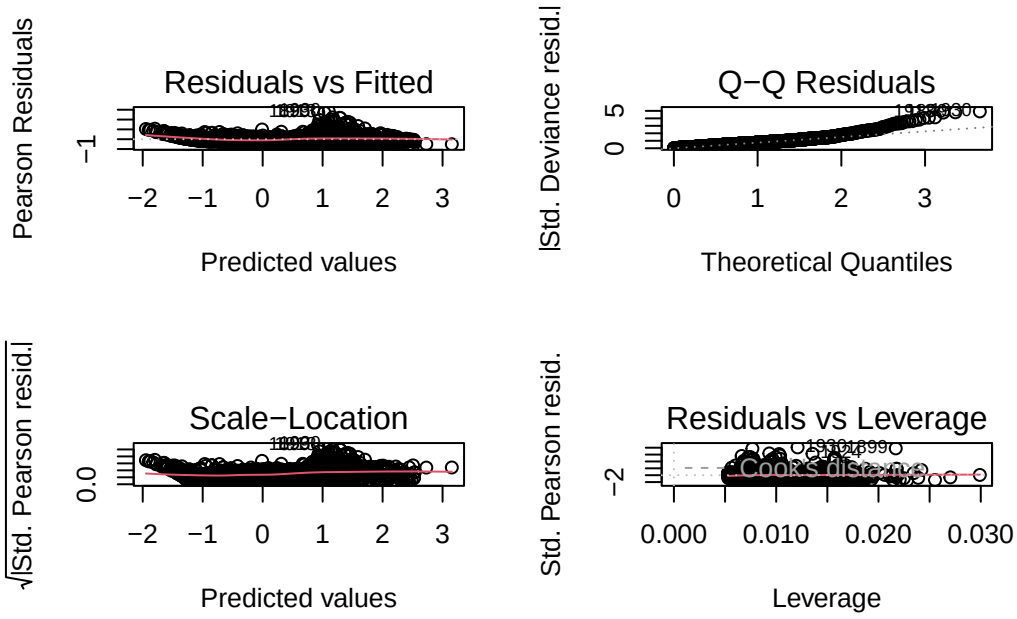

Figure 11: Diagonsis

References

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

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

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

Toronto Shelter & Support Services. 2024. *Deaths of Shelter Residents*. https://open.toronto.ca/dataset/deaths-of-shelter-residents/.