1. Summary of Insights

The goal of this study was to determine whether **the carat weight of a diamond can be predicted using visually observable characteristics**, excluding price. This is particularly useful in **automated diamond valuation** and **appraisal processes** where weighing the diamond is not feasible.

Our analysis revealed that the **physical dimensions (x, y, and z) were the strongest predictors of carat weight**, demonstrating a high positive correlation. However, these variables also exhibited high multicollinearity, which posed a challenge to model stability. To address this, we **created a composite variable, Magnitude (x * y * z)**, to encapsulate the overall size of the diamond in a single predictor. This transformation significantly improved model stability and interpretability. Additionally, we examined categorical factors such as **cut, color, and clarity**. While these factors had some influence on carat weight, they were secondary in importance compared to magnitude.

The final regression model showed that **carat weight is primarily influenced by magnitude, with minor but statistically significant effects from diamond color**. Diamonds **with a lower quality color (colorLow) tend to have a slightly higher carat weight compared to those with higher quality color**. The interaction between magnitude and quality color further indicated that while magnitude is the primary driver of carat weight, its effect is slightly diminished in low quality-colored diamonds. Our model demonstrated strong predictive accuracy, suggesting that **carat weight can indeed be estimated using visually observable characteristics**.

Summary of Findings:

- Magnitude (x * y * z) is the most significant predictor of carat weight.
- Categorical variables such as cut, color, and clarity have statistically significant but minor effects.
- Multicollinearity among x, y, and z was addressed by transforming them into a composite **magnitude** variable.
- Polynomial regression and log transformations improved model performance.
- Final model achieved an Adjusted R^2 of 0.945, demonstrating strong predictive ability.

These findings indicate that **carat weight can be estimated accurately based on observable features**, making the model valuable for **jewelers, automated diamond grading, and gemstone appraisal systems**.

2. Modeling Process

Data Preparation and Train-Test Split

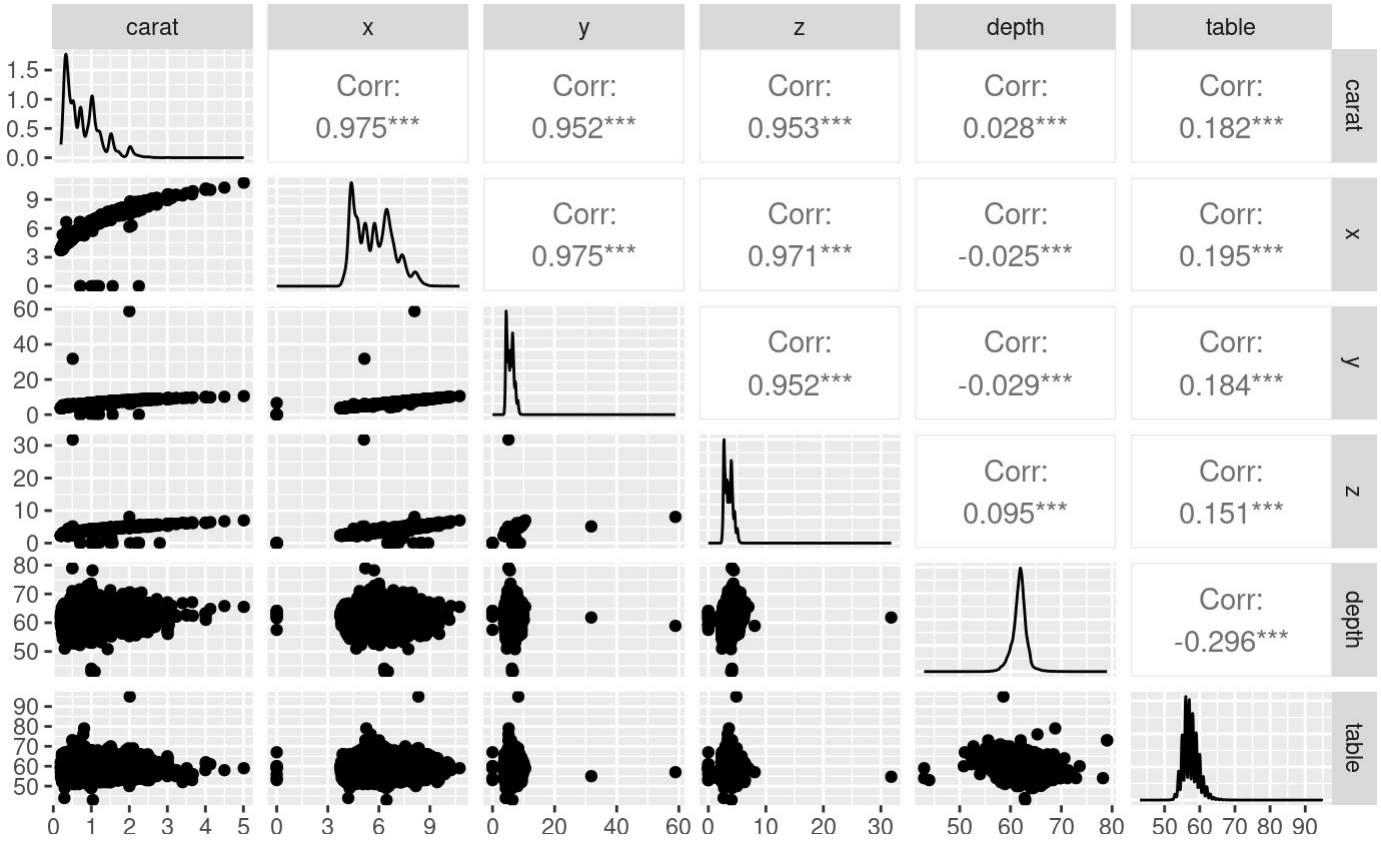
To ensure model generalizability, we split the dataset into training (80%) and testing (20%) subsets using stratified sampling. This was done using the createDataPartition() function, ensuring a balanced distribution of carat weights in both sets.

- Training Set: Used to fit the regression model.
- Testing Set: Used to evaluate model performance and prevent overfitting.

To develop an effective predictive model, we started with Exploratory Data Analysis (EDA) to understand the relationships between carat weight and other features in the dataset.

We first performed **correlation analysis** to identify the strongest predictors of carat weight:

- x, y, z had **high correlations** with carat (0.975, 0.952, 0.953 respectively).
- depth and table had **weak correlations** with carat (0.028, 0.182), suggesting limited predictive power.
- **Scatterplots confirmed strong linear relationships between carat and x, y, z**, supporting the use of regression models.



However, high multicollinearity among these variables necessitated a transformation.

Addressing Multicollinearity

- **Variance Inflation Factor (VIF) analysis show severe multicollinearity** (VIF > 10) among x, y, and z.
- To resolve this, we created a new variable **magnitude** (x * y * z)
- After transformation, **VIF values dropped below 5**, confirming that multicollinearity was successfully addressed.

3. Model Selection and Performance

We tested several regression models to determine the best predictors for carat weight:

1. **Baseline Model** – Used x, y, and z as predictors, but suffered from multicollinearity.
2. **Refined Model** – Replaced x, y, and z with **Magnitude**, which improved interpretability.
3. **Final Model** – Included **Magnitude, Quality Color, and an interaction term (Magnitude * Quality Color)** to capture variations in carat weight across different color categories.

The final regression equation was formulated as:

$$Carat = 0.00395 + 0.00610 \times Magnitude + 0.06018 \times Quality\ ColorLow - 0.00038 \times (Magnitude \times Quality\ ColorLow)$$

Coefficient Interpretation:

- **Intercept = 0.00395** → Expected carat weight when magnitude = 0 and quality_colorLow = 0.
- **Magnitude (0.00610)** → A **one-unit increase in magnitude leads to a 0.00610 increase in carat**, holding color constant.
- **Quality Color (Low) (0.06018)** → Diamonds with **lower color quality** tend to have a **0.06018 higher carat weight** than high-color diamonds.

- **Interaction term (-0.00038)** → Suggests that **magnitude has a slightly weaker effect on carat for diamonds with low color quality.**

Finally, R^2 value of 0.9451 indicates about 94.51% of variability in carat price is explained by magnitude, color and quality.

4. Model Diagnostics

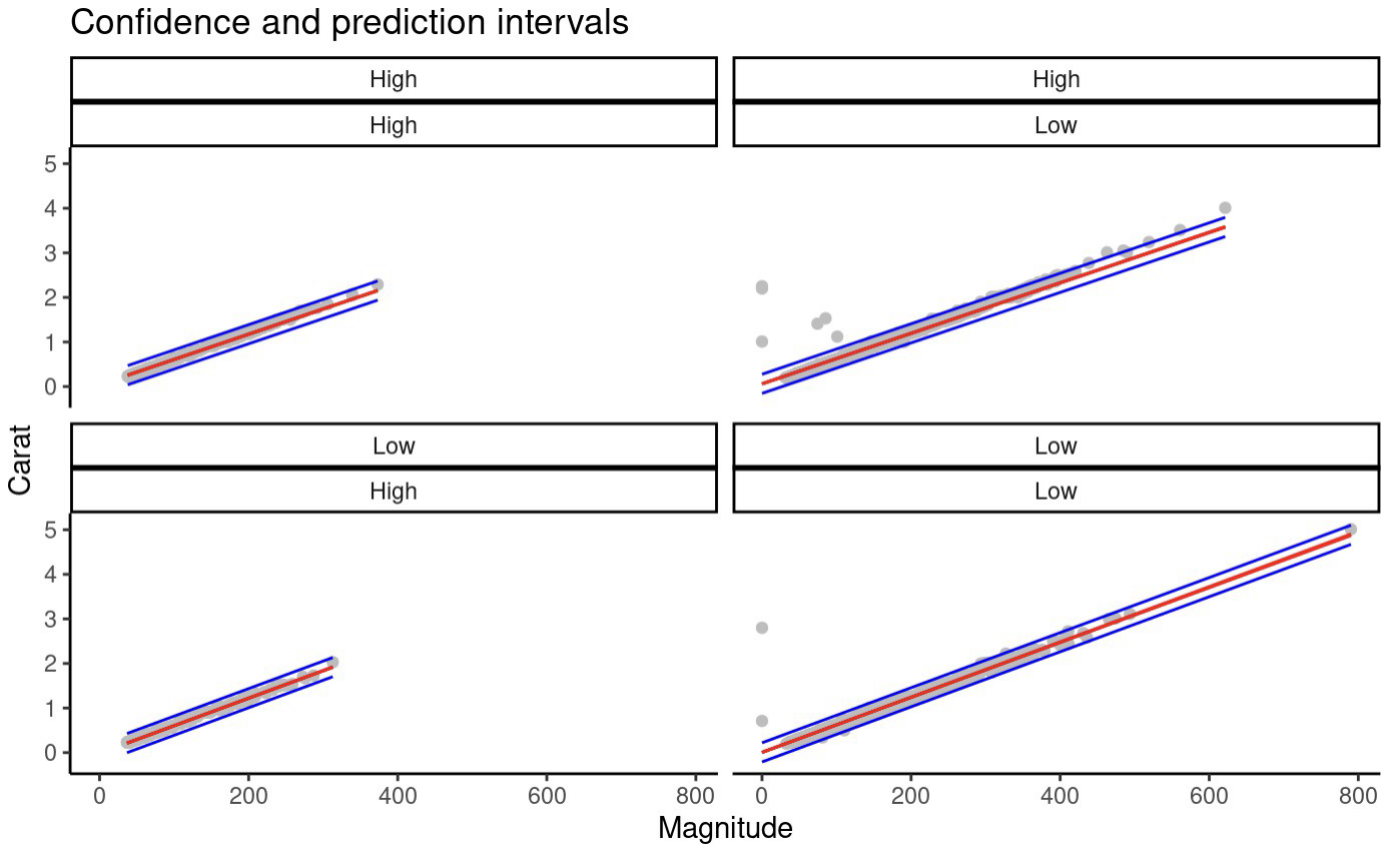
To ensure the reliability of our model, we performed several diagnostic tests. **Residual analysis** confirmed that the assumptions of **linearity, homoscedasticity, and normality** were met. The residual plots showed no discernible patterns, suggesting that the linear relationship between predictors and carat weight was appropriate. **Checking for multicollinearity** further validated that the magnitude transformation successfully mitigated redundancy among predictors. (from VIF > 10 to below 5)

Accuracy Metrics

Adjusted R^2	Root Mean Squared Error (RMSE)	F-statistic (Model Significance)
0.945	**0.0808**	**$p<2.2\times10-16$**

Confidence Intervals

The plots show how diamond carat weight (y-axis) changes with volume, split by cut & quality (High/Low). The red line is the model's prediction, the inner blue lines show the average carat weight's range (confidence interval), and the outer blue lines show the single diamond's carat weight's range (prediction interval), wider due to individual variability.



5. Polynomial Transformation

Given that some variables exhibited skewness, particularly y and z, we implemented a **log transformation** to stabilize variance and improve model performance. Additionally, we explored a **polynomial regression approach**, introducing second-degree polynomial terms for log_x, log_y, and log_z. This allowed for a more flexible fit, capturing potential nonlinear relationships between carat weight and the predictors.

Polynomial Transformation for Improved Model Performance

- **Histograms of x, y, and z showed right-skewed distributions.**
- We applied **log transformations** to reduce skewness
- **Final polynomial regression improved model accuracy**, reducing residual error.

The estimated polynomial regression model took the form:

$$Carat = \beta 0 + \beta 1 \cdot poly(logx,2) + \beta 2 \cdot poly(logy,2) + \beta 3 \cdot poly(logz,2) + \beta 4 \cdot Depth + \beta 5 \cdot Table + \epsilon$$

Model Performance

- **Adjusted R^2** = 0.9711– indicating a strong fit.
- **Test RMSE** = 0.0804 – confirming that predictions are accurate with minimal error.
- **F-statistic (p < 2.2e-16)** – strong overall significance of the model.

Model Comparison (Linear vs Polynomial)

Accuracy Metrics Comparison

Models	Adjusted R^2	RMSE	F-statistics
Polynomial	0.9711	0.0804	$p<2.2\times10^{-16}$
Final Model	0.945	0.0808	$p<2.2\times10^{-16}$

6. Conclusion

This study successfully demonstrated that **carat weight can be accurately estimated using visually observable characteristics, particularly Magnitude and Quality Color**. The transformation of x, y, and z into **Magnitude** significantly improved model stability and interpretability, while polynomial regression provided a more refined fit. The final model suggests that while **Magnitude remains the dominant predictor, Quality Color also plays a role in carat estimation**. From a practical standpoint, these findings are useful for **jewelry appraisers, retailers, and consumers** who need to estimate carat weight without physical measurement.