Analysis and Modeling of Diamond Prices

1. Data Preprocessing

- The dataset contains 53,940 entries with the following columns: `carat`, `cut`, `color`, `clarity`, `depth`, `table`, `price`, `x`, `y`, and `z`.
- No missing values were found in the dataset.
- The `Unnamed: 0` column was dropped as it only contained indices.
- Diamonds with dimensions (`x`, `y`, `z`) equal to 0 were removed, eliminating 20 logically invalid data points.
- Outliers were detected and removed based on the following criteria:
  - `depth` between 45 and 75.
  - `table` between 40 and 80.
  - `x`, `y`, `z` less than 30, and `z` greater than 2.
- After preprocessing, the dataset contains **53907** entries.

2. Exploratory Data Analysis

- Pair plots were created to visualize relationships between variables. Outliers were detected in `y`, `z`, `depth`, and `table`.
- Regression plots confirmed the presence of outliers, which were subsequently removed.
- Violin plots showed the distribution of `price` by `cut`, `color`, and `clarity`.
- **Observation:** The price does not vary significantly with `cut`, `color`, and `clarity`, as the dataset contains predominantly lower-quality diamonds. If the dataset included higher-quality diamonds, these variables would likely have a stronger influence.

3. Feature Encoding

- Categorical variables (`cut`, `color`, `clarity`) were encoded using `LabelEncoder`.

4. Correlation Analysis

- A heatmap revealed strong correlations between `price` and physical dimensions (`carat`, `x`, `y`, `z`).
- Weak correlations were observed between `price` and `color` or `clarity`, reflecting the dataset's composition of lower-quality diamonds.
- **Observation:** The parameters (`cut`, `color`, `clarity`) do not significantly affect the `price` because the dataset contains lower-quality diamonds, all with low color grades. This explains the weak correlations. If the dataset included diamonds with higher-quality colors, `cut`, `color`, and other parameters would exhibit stronger and more coherent relationships with `price`.

5. Regression Modeling

- A linear regression model was trained on the preprocessed dataset.
- The data was split into training (75%) and test (25%) sets.
- Model performance on the test set:
  - R^2: **0.8890**
  - Mean Absolute Error (MAE): **849.3507**
  - Mean Squared Error (MSE): **1741183.6678**

6. Conclusion

- The model demonstrates a strong relationship between diamond `price` and physical dimensions (`carat`, `x`, `y`, `z`).
- Weak correlations with `color` and `clarity` are consistent with the dataset's lower-quality composition.
- Further analysis including higher-quality diamonds may yield different results.