Diamond Price Prediction Using Machine Learning (Based on Custom Implementation)

1. Objective: The primary objective of this project is to develop a robust machine learning model to predict diamond prices based on key characteristics. This solution, created using Python, aims to assist jewellers, dealers, and consumers in estimating fair prices using data-driven insights.

2. Dataset Overview:

- **Source:** Diamond dataset
- **Total Entries:** ~54,000 diamonds
- **Features (Attributes):**
  - carat: Weight of the diamond; a strong indicator of price
  - cut: Quality of cut (Fair, Good, Very Good, Premium, Ideal)
  - color: Color grade (D = best, J = worst)
  - clarity: Clarity rating (from I1 = worst to IF = best)
  - depth: Total depth percentage
  - table: Width of the top of the diamond relative to the widest point
  - price: Target variable; price in US dollars
  - x, y, z: Physical dimensions of the diamond in mm

3. Domain Analysis:

- Carat has a very high positive correlation with price, making it the most influential feature.
- Cut, colour, and clarity are categorical variables with a moderate impact on price.
- Depth and table provide shape characteristics but have relatively weak correlations.
- x, y, and z were found to be strongly correlated with carat and were thus removed to avoid redundancy.

4. Data Preprocessing (Based on Custom Code):

- Dropped the Unnamed: 0 index column.
- Replaced corrupted values (e.g., zeros in x, y, z) with NaN, then dropped them.

- Encoded categorical variables (cut, colour, clarity) using Label Encoding.
- Removed x, y, and z due to high multicollinearity with carat to improve model performance.

5. Exploratory Data Analysis (EDA):

- **Univariate Analysis:**
    - Used **histograms** to visualise the distribution of features like carat, depth, table, and price. These plots showed that most diamonds have lower carat values, and prices are right-skewed.
- **Bivariate Analysis:**
    - Applied **scatterplots** to examine relationships between carat, depth, table and price. A strong positive relation was observed between carat and price.
- **Multivariate Analysis:**
    - A **correlation heatmap** was plotted, revealing strong positive correlations among carat, x, y, z, and price. Hence, only the carat was retained.
- **Outlier Detection:**
    - **Box plots** were used to identify and interpret outliers in variables such as carat, price, and depth. This step ensured better model generalisation.

6. Feature Selection:

- **Final features used:** carat, cut, colour, clarity, depth, table
- **Removed features:** x, y, z due to multicollinearity with carat

7. Model Building (Using Custom Code):

- Models implemented:
    - **Linear Regression:** Achieved highest performance with an R^2 score of **0.91**
    - **K-Nearest Neighbours (KNN):** Achieved a lower R^2 score of **0.80**
- Based on results and interpretability, **Linear Regression was chosen** as the most suitable model.

8. Final Model Summary:

- **Best Model:** Linear Regression
- **R^2 Score:** 0.91

- **Adjusted R²:** ~0.9076
- **Performance:** High accuracy, low complexity, and interpretable results
- **Tools Used:** pandas, matplotlib, seaborn, scikit-learn

9. Conclusion: The diamond price prediction model built using custom Python code demonstrates a high degree of accuracy and efficiency. The linear regression model, with an R^2 of 0.91, emerged as the most reliable and interpretable approach. EDA techniques such as histograms, scatterplots, and box plots provided critical insights during feature engineering and outlier detection. Removing highly correlated dimensions (x, y, z) in favour of using only carat improved model performance and reduced redundancy. This predictive model can be effectively used in real-world scenarios for fair price estimation and can be further enhanced with ensemble learning techniques.