Assignment: Predict Diamond Price Using Machine Learning

🔍 Objective:

You are given a small dataset containing information about diamonds, including details like **carat weight**, **cut**, **color**, and **clarity**. Your task is to build a **machine learning model** that can **predict the price** of a diamond using this information.

📊 Column Descriptions:

💎 *Cut*

- **Definition**: Refers to how well the diamond has been cut and shaped from its rough form. This affects how well it reflects light (sparkle).
- **Categories**: Common grades include:

Fair	Lowest quality cut. Reflects less light, appears dull.
Good	A decent cut, but doesn't reflect light as well as higher grades.
Very Good	Reflects most light well; high quality.
Ideal	Excellent proportions and symmetry. Very bright and sparkly.
Signature-Ideal	Even better than Ideal — top-tier, often used by premium brands.

- **Why it matters**: A well-cut diamond appears brighter and more attractive. Even if two diamonds have the same carat and color, the one with a better cut can look more beautiful and be priced higher.

🎨 *Color*

- **Definition**: Measures how colorless a diamond is. The less color, the higher the value (except for fancy-colored diamonds).
- **Grading Scale** (by GIA):
  - **D–F**: Colorless
  - **G–J**: Near colorless
  - **K–M**: Faint yellow
  - **N–Z**: Noticeable yellow or brown
- **Why it matters**: Diamonds with less visible color are rarer and more valuable. Color impacts the overall brightness and appeal of the diamond.

🔎 *Clarity*

- **Definition**: Indicates how many imperfections (called inclusions or blemishes) are present inside or on the surface of the diamond.
- **Grades**:
  - **FL** (Flawless)
  - **IF** (Internally Flawless)
  - **VVS1, VVS2** (Very, Very Slightly Included)
  - **VS1, VS2** (Very Slightly Included)
  - **SI1, SI2** (Slightly Included)

  - **I1, I2, I3** (Included)
- **Why it matters**: Fewer inclusions mean higher clarity and therefore a higher price. Some inclusions can also affect durability and sparkle.

Polish

- **Definition**: Refers to the smoothness of the diamond's surface after cutting. Affects how clearly light reflects off the surface.
- **Grades**:
  - **Excellent (EX)**
  - **Very Good (VG)**
  - **Good (G)**

    Ideal(ID)
- **Why it matters**: Better polish can enhance a diamond's shine and brilliance, making it more valuable.

Symmetry

- **Definition**: Measures how well the diamond's facets are aligned and proportioned.
- **Grades**: Same as Polish — EX, VG, G, F, P.
- **Why it matters**: Good symmetry improves light reflection, which affects brilliance and beauty. Poor symmetry can make a diamond look uneven or dull.

Report (Certification Lab)

- **Definition**: Identifies which gemological lab evaluated and certified the diamond.

- **Examples**:
  - **GIA** – Gemological Institute of America (most trusted and strict)
  - **AGSL** – American Gem Society Laboratories
- **Why it matters**: Certification from a reputable lab ensures consistent grading. A diamond certified by GIA, for example, might cost more than one graded the same by a less strict lab.

💰 *Price*

- **Definition**: The actual price (in US dollars) the diamond is sold for.
- **Why it matters**: This is the **target variable** you're trying to predict using machine learning. All other features influence this value.

✅ Your Tasks:

◇ Part 1: Data Preparation

1. Load the dataset using pandas.
2. Show the first 5 rows.
3. Check if there are any missing values, clean if there is any.
4. Find how many have an ideal cut and have reports from GIA
5. What is the highest diamond price? Why do you think it is that costly? Note the answer in a markdown
6. Draw a pie chart to show polish distribution
7. Convert the categorical columns (Cut, Color, Clarity, Polish, Symmetry, Report) into numbers using **Label Encoding**.

◇ Part 2: Exploratory Data Analysis (EDA)

1. Plot **histograms** to show the distribution of each column
2. Create a **scatter plot** between Carat Weight and Price to see if there's a pattern.
3. Create a **correlation heatmap** to understand which features relate to price.

◇ Part 3: Model Building

1. Split your data:
    a. Features (X) = All columns **except** Price
    b. Target (y) = Price
2. Use train_test_split() to split your data into training and testing sets.
3. Train a **Linear Regression** model using scikit-learn.
4. Predict on the test set and evaluate using:
    a. R^2 Score
5. Try your own diamond data and see what you price one will pay