Project Title: Analyzing Factors Influencing Diamond Prices

Short Description: This project guides students to analyze how various attributes like carat, cut, colour, and clarity influence diamond prices using Python, focusing on data analysis and visualization techniques.

Difficulty Level: Beginner

AI-ML Concepts Used: Data Analysis, Visualization

Learning Outcome: Understand factors affecting diamond pricing

Programming Skills Mastered: Pandas, Plotly

Introduction

Diamonds are precious gemstones whose prices are influenced by various factors such as carat weight, cut quality, color, and clarity. Understanding these factors is essential for both buyers and sellers. In this project, we will analyze how these attributes affect diamond prices using Python. This analysis is tailored for Class 9 and 10 CBSE/ICSE students to introduce them to data analysis and visualization techniques.

Dataset Description

We will use a dataset containing information about diamonds, including:

- **Carat:** Weight of the diamond.
- **Cut:** Quality of the diamond's cut (e.g., Ideal, Premium, Good).
- **Color:** Diamond color, with categories ranging from D (best) to J (worst).
- **Clarity:** Measurement of diamond clarity (e.g., SI2, VS1).
- **Depth:** Total depth percentage (z / mean(x, y)).
- **Table:** Width of the top of the diamond relative to the widest point.
- **Price:** Price of the diamond.
- **X:** Length in mm.
- **Y:** Width in mm.
- **Z:** Depth in mm.

Implementation Steps

1. **Import Libraries and Load Dataset**

   We will start by importing the necessary Python libraries and loading the dataset.

   ```python
   CopyEdit
   import pandas as pd
   import plotly.express as px

   # Load the dataset
   data = pd.read_csv("diamonds.csv")

   # Display the first few rows of the dataset
   print(data.head())
   ```

2. **Data Preprocessing**

   We will remove unnecessary columns and handle any missing values.

   ```python
   CopyEdit
   # Drop the 'Unnamed: 0' column if present
   if 'Unnamed: 0' in data.columns:
       data = data.drop(columns=['Unnamed: 0'])

   # Check for missing values
   print(data.isnull().sum())
   ```

3. **Analyze Carat vs. Price**

   We will examine the relationship between the carat weight and the price of diamonds.

   ```python
   CopyEdit
   # Scatter plot of carat vs. price
   fig = px.scatter(data_frame=data, x="carat", y="price", color="cut",
                    title="Carat Weight vs. Price of Diamonds",
                    labels={"carat": "Carat Weight", "price": "Price
   (USD)"})
   fig.show()
   ```

4. **Analyze Cut Quality vs. Price**

We will analyze how the cut quality affects diamond prices.

```python
CopyEdit
# Box plot of cut quality vs. price
fig = px.box(data_frame=data, x="cut", y="price",
             title="Cut Quality vs. Price of Diamonds",
             labels={"cut": "Cut Quality", "price": "Price (USD)"})
fig.show()
```

5. **Analyze Color vs. Price**

We will explore the impact of diamond color on its price.

```python
CopyEdit
# Box plot of color vs. price
fig = px.box(data_frame=data, x="color", y="price",
             title="Color vs. Price of Diamonds",
             labels={"color": "Color Grade", "price": "Price (USD)"})
fig.show()
```

6. **Analyze Clarity vs. Price**

We will investigate how clarity influences diamond prices.

```python
CopyEdit
# Box plot of clarity vs. price
fig = px.box(data_frame=data, x="clarity", y="price",
             title="Clarity vs. Price of Diamonds",
             labels={"clarity": "Clarity Grade", "price": "Price
(USD)"})
fig.show()
```

7. **Calculate and Analyze Diamond Size**

We will calculate the size of each diamond and analyze its effect on price.

```python
CopyEdit
# Calculate size (volume) of each diamond
data["size"] = data["x"] * data["y"] * data["z"]

# Scatter plot of size vs. price
fig = px.scatter(data_frame=data, x="size", y="price", color="cut",
                 title="Diamond Size vs. Price",
                 labels={"size": "Size (mm³)", "price": "Price
(USD)"})
fig.show()
```

Conclusion

In this project, we analyzed how various attributes such as carat weight, cut quality, color, clarity, and size influence diamond prices. We utilized Python libraries like Pandas for data manipulation and Plotly for data visualization. Through this analysis, students learned how to handle datasets, perform exploratory data analysis, and visualize relationships between different features and the target variable.