Assignment: Data Preprocessing - Categorical Data Handling, Missing Values, Ordinal Data Handling, and Outlier Removal

---

Dataset Description

You are provided with a dataset containing customer information from an e-commerce platform. The dataset includes categorical variables, missing values, ordinal data, and possible outliers. Your task is to perform necessary preprocessing steps to clean and prepare the data for analysis.

Dataset Columns:

1. **Customer_ID** (Unique Identifier)
2. **Age** (Numerical, may contain missing values)
3. **Gender** (Categorical: Male, Female, Other)
4. **City** (Categorical: New York, Los Angeles, Chicago, Houston, Miami)
5. **Annual_Income** (Numerical, may contain outliers)
6. **Education_Level** (Ordinal: High School, Bachelor's, Master's, PhD)
7. **Purchase_Frequency** (Numerical: Number of purchases in a year)
8. **Subscription_Status** (Categorical: Active, Inactive)
9. **Credit_Score** (Numerical, may contain missing values)
10. **Product_Category** (Categorical: Electronics, Clothing, Home, Beauty, Sports)

The dataset contains **50 records** with some missing values and outliers.

Assignment Questions

Section 1: Handling Categorical Data

1. Load the dataset and display the first five records.
2. Identify all categorical columns in the dataset.
3. Convert the **Gender** column into numerical values using One-Hot Encoding.
4. Apply Label Encoding to the **City** column.
5. Convert the **Product_Category** column using Frequency Encoding.
6. Perform Target Encoding on the **Subscription_Status** column using Purchase_Frequency as the target.

7. Compare Label Encoding and One-Hot Encoding results on the **Education_Level** column.

Section 2: Handling Missing Values

8. Identify missing values in the dataset.
9. Fill missing values in the **Age** column using the median.
10. Replace missing values in the **Credit_Score** column with the mean.
11. Analyze the impact of missing data on model performance and suggest strategies to handle it.

Section 3: Handling Ordinal Data

12. Define an appropriate mapping for the **Education_Level** column and transform it into numerical values.
13. Verify whether the transformed **Education_Level** column retains the correct order (High School < Bachelor's < Master's < PhD).
14. What is the advantage of using ordinal encoding for the **Education_Level** column instead of One-Hot Encoding?

Section 4: Outlier Detection and Removal

15. Detect outliers in the **Annual_Income** column using the Interquartile Range (IQR) method.
16. Remove outliers in the **Annual_Income** column based on the IQR method.
17. Detect outliers in the **Purchase_Frequency** column using the Z-score method.
18. Remove outliers in the **Purchase_Frequency** column based on the Z-score threshold ($|z| > 3$).
19. What impact does outlier removal have on the dataset's mean and standard deviation?
20. Compare and contrast the IQR and Z-score methods for detecting outliers.
