Assignment Title: Understanding Descriptive Statistics and Sampling for Machine Learning in Python

Assignment Overview:

In this assignment, you will explore the concepts of descriptive statistics and sampling techniques. You will analyze dataset related to an e-commerce platform and perform statistical analysis to understand customer purchasing behavior. By the end of the assignment, you will gain a deeper understanding of how descriptive statistics and sampling techniques play a crucial role in machine learning preprocessing.

Scenario:

You have been hired as a data scientist at an e-commerce company called "ShopSmart." The company wants to understand customer behavior to improve sales and customer satisfaction. They have collected data over the past year from 10 million customers, including transaction details such as purchase amounts, customer demographics, product categories, and timestamps. Since the dataset is enormous, the company wants you to perform a statistical analysis on a representative sample of the dataset.

Your goal is to:

1. Use descriptive statistics to understand the general customer behavior.
2. Implement different sampling techniques to create a manageable dataset for analysis.

Dataset Details:

The dataset consists of the following columns:

- `Customer_ID`: Unique identifier for each customer.
- `Gender`: Gender of the customer (Male/Female).
- `Age`: Age of the customer.
- `Country`: Country of residence.
- `Purchase_Amount`: The amount spent on the purchase.
- `Purchase_Category`: The category of the purchased item (Electronics, Fashion, Groceries, etc.).
- `Transaction_Timestamp`: Date and time of the transaction.

You can either use a synthetic dataset or download an open-source dataset from Kaggle to simulate the scenario.

Tasks:

Task 1: Descriptive Statistics

- **Step 1**: Load the dataset into Python using Pandas.
- **Step 2**: Perform the following descriptive statistics:
    - Calculate the mean, median, and mode for the `Purchase_Amount` column.
    - Find the standard deviation and variance for the `Purchase_Amount`.

- Determine the age distribution of the customers using measures like mean, quartiles, and range.
- Count the frequency of purchases by `Purchase_Category` to identify the most popular categories.
- Visualize the distribution of `Purchase_Amount` using histograms and box plots.

Task 2: Sampling Techniques

- **Step 1**: Since analyzing the entire dataset is computationally expensive, you need to use sampling techniques. Implement the following sampling methods:
    - **Simple Random Sampling**: Randomly select 10% of the customers from the dataset.
    - **Stratified Sampling**: Select a sample based on `Purchase_Category`, ensuring each category is proportionally represented.
    - **Systematic Sampling**: Select every 10th customer in the dataset for analysis.
    - **Cluster Sampling**: Group customers by country and select a few countries to analyze all customers within those countries.
- **Step 2**: For each sampling method, perform the following:
    - Compare the mean and standard deviation of `Purchase_Amount` with the entire dataset.
    - Discuss whether the sample is representative of the overall population.
    - Plot the distribution of `Purchase_Amount` for each sample.

Task 3: Report Findings

- **Step 1**: Summarize your findings in a report. Include the following sections:
    - **Descriptive Statistics Summary**: Explain the key statistics you calculated and what they reveal about customer behavior.
    - **Sampling Techniques Comparison**: Compare the results from different sampling techniques and discuss the trade-offs.
    - **Conclusion**: How can the company use the insights gained from this analysis to improve customer satisfaction and sales?

Expected Deliverables:

1. Python code file (Jupyter Notebook or .py file) with all tasks implemented.