Task 1: Answer the following questions

Question 1: Which of the following statements best describes a dataset?
A) A collection of software tools for data analysis.
B) A group of data visualization techniques.
C) A structured collection of data points representing some aspect of the real world.
D) A set of algorithms used for machine learning.

Question 2: Why is data preprocessing an important step in data analysis?
A) It helps to generate random data points for analysis.
B) It increases the complexity of the analysis.
C) It reduces noise and inconsistencies in the data, improving the quality of analysis.
D) It allows for visualizing data without any modifications.

Question 3: Which of the following is considered categorical data?
A) Temperature in degrees Celsius.
B) Height of individuals in centimeters.
C) Colors of flowers (e.g., red, blue, yellow).
D) Prices of products in dollars.

Question 4: What is one common method for handling missing data in a dataset?
A) Ignoring the missing values and proceeding with the analysis.
B) Removing the entire row or column containing missing values.
C) Creating new random values to replace missing data.
D) Rearranging the dataset to fill in missing values.

Question 5: What does feature engineering involve in data analysis?
A) It refers to removing all features from the dataset to simplify analysis.
B) It focuses on selecting only numerical features for analysis.
C) It involves creating or transforming new features to improve the model's performance.
D) It refers to preprocessing data without considering feature transformation.

Question 6: Why is splitting a dataset into training and testing sets important?
A) To reduce the size of the dataset for faster analysis.
B) To create multiple copies of the dataset for different types of analysis.
C) To ensure that the model's performance is evaluated on unseen data.
D) To combine the training and testing data for better accuracy.

Question 7: What is a common technique to handle categorical data before feeding it into a machine learning model?
A) Removing all categorical data from the dataset.
B) Converting categorical data into strings for better representation.
C) One-Hot Encoding, where each category becomes a binary column.
D) Replacing categorical data with the mean value of the entire dataset.

Question 8: Why might scaling numerical features in a dataset be necessary?
A) Scaling has no impact on numerical features.
B) To convert numerical features into categorical ones.
C) To ensure that all numerical features have the same unit of measurement.
D) Scaling only affects the model's training time, not its performance.

Question 9: What is an outlier in the context of data analysis?
A) A type of categorical variable.
B) Data points that are missing from the dataset.
C) Unusual or extreme data points that significantly differ from the rest.
D) A subset of data that is used for validation.

Question 10: What does data imputation involve?
A) Replacing all categorical data with numerical values.
B) Filling missing values with arbitrary values.
C) Creating entirely new datasets to replace the original one.
D) Filling in missing values with estimated or calculated values.

Question 11: What is a consideration when dealing with time-series data in data analysis?
A) Time-series data cannot contain missing values.
B) Time intervals between data points are irrelevant.
C) The order and timing of data points matter.
D) Time-series data should only contain numerical values.

Question 12: What is the primary goal of dimensionality reduction techniques in data analysis?
A) To increase the dimensionality of the dataset.
B) To transform categorical features into numerical ones.
C) To decrease the amount of missing data in the dataset.
D) To reduce the number of features while preserving relevant information.

Question 13: Why is addressing imbalanced classes important when building models?
A) Imbalanced classes do not affect the model's performance.
B) Imbalanced classes lead to faster model training.
C) Imbalanced classes can bias the model towards the majority class.
D) Imbalanced classes are only relevant when dealing with categorical data.

Question 14: Which preprocessing step is commonly used for text data before analysis?
A) Converting text data to numerical values using encoding techniques.
B) Remove all punctuation marks and capitalization from the text.
C) Converting text data into categorical variables.
D) Text data does not require any preprocessing.

Task 2: Data Analysis and Machine Learning Preprocessing

Task Description:
You have been given a dataset about customer orders and their interactions with an e-commerce platform. Your task is to perform advanced data analysis, preprocessing, and feature engineering using NumPy and Pandas. Additionally, you'll prepare the data for machine learning by transforming categorical variables and splitting it into training and testing sets.

Dataset: `e_commerce_data.csv` (contains columns: CustomerID, Timestamp, ProductID, Category, Price, Quantity, Action)

Part 1: Data Analysis and Preprocessing

1. Load the CSV dataset into a Pandas DataFrame.
2. Handle missing values, considering different strategies for different columns.
3. Analyze customer interactions by calculating the total number of actions (purchases, views, etc.) for each customer.

Part 2: Feature Engineering and Analysis

1. Create a new feature TotalSpent by calculating the total amount spent by each customer.
2. Group the data by Category and analyze the most popular categories.
3. Calculate the average price of products in each category.

Part 3: Machine Learning Preprocessing

1. Convert categorical variables (Category, Action) into numerical representations using one-hot encoding.
2. Standardize numerical features (Price, Quantity, TotalSpent) using Z-score normalization.
3. Split the dataset into training and testing sets (80% training, 20% testing) for machine learning.

Part 4: Insights and Data Preparation Summary

1. Write a summary of your data analysis, feature engineering, and preprocessing steps.
2. Highlight any trends or patterns you observed in the data.
3. Discuss the rationale behind your choices for feature engineering and preprocessing techniques.

Submission Guidelines:

- Provide a well-structured Python script or a Jupyter Notebook.
- Include explicit comments explaining each step of your code.
- Include visualizations and plots to support your analysis.
- Summarize your findings and insights in a detailed manner.
- Discuss your approach to handling missing data and preprocessing techniques.
- Highlight how your preprocessing steps contribute to preparing the data for machine learning.