SOC - Time Series Analysis

Assignment 1 : Exploratory Data Analysis

Deadline - 7th June

Objective:

Perform comprehensive exploratory data analysis (EDA) on the given dataset to extract meaningful insights and demonstrate your proficiency in using pandas, numpy, and other relevant tools.

Dataset:

For this assignment, you will work with the "Global Superstore" dataset. This dataset contains sales data for a fictional global retail store and includes information on orders, customers, products, and shipping details. You can download the dataset from here.

Tasks:

1. **Data Loading and Initial Inspection:**
   - Load the dataset using pandas.
   - Display the first few rows of the dataset.
   - Provide a summary of the dataset, including data types, missing values, and basic statistical details.

2. **Data Cleaning:**
   - Identify and handle missing values appropriately.
   - Remove any duplicate rows.
   - Ensure that data types are appropriate for analysis (e.g., convert date columns to datetime type).

3. **Data Exploration:**
   - **Descriptive Statistics:**
     - Calculate and display descriptive statistics for numerical columns.
     - Generate and interpret summary statistics for categorical columns.
   - **Univariate Analysis:**
     - Create histograms, box plots, and count plots for key variables (e.g., Sales, Quantity, Discount, Category, Sub-Category).

- Identify and interpret any patterns or outliers.
- **Bivariate Analysis:**
  - Create scatter plots, correlation matrices, and pair plots to explore relationships between numerical variables.
  - Use bar plots and box plots to explore relationships between numerical and categorical variables (e.g., Sales by Category, Profit by Region).
- **Multivariate Analysis:**
  - Perform and interpret the pivot-table and group-by operation to analyze data across multiple dimensions (e.g., average Sales and Profit by Category and Sub-Category).

4. **Data Visualization:**
   - Use matplotlib and seaborn to create visually appealing and informative plots.
   - Include at least one visualization for each type of analysis (univariate, bivariate, multivariate).
   - Make sure to add titles, labels, and legends to your plots for better readability.

5. **Feature Engineering:**
   - Create at least two new features that could be useful for further analysis (e.g., profit margin, sales per customer, order processing time).