[GSH Online Media] Python Developer Test Exercise

Objective:
Create a Python script that processes and analyzes a dataset, performing specified tasks related to data integration, analysis, and automation.

Scenario:
You are tasked with developing a Python script that processes a CSV file containing data about products and their sales performance. The dataset includes columns like `ProductID`, `ProductName`, `Category`, `Sales`, and `DateSold`. Your script should perform the following tasks:

Tasks:

1. **Data Loading & Preprocessing:**
   - Load the dataset from a CSV file named `sales_data.csv`.
   - Handle missing data appropriately, either by filling with defaults or removing incomplete records.
   - Convert the `DateSold` column into a datetime format and ensure that all dates are consistent.
2. **Data Integration:**
   - Suppose there is another CSV file `product_info.csv` containing additional information about each product, including `ProductID`, `Supplier`, and `CostPrice`.
   - Merge the data from `sales_data.csv` and `product_info.csv` on the `ProductID` field, creating a comprehensive dataset.
3. **Data Analysis:**
   - Calculate the total sales for each product and identify the top 5 best-selling products.
   - Compute the profit for each product (assume `Profit = Sales - CostPrice`).
   - Identify any trends in sales over time (e.g., seasonal trends, monthly sales changes).
4. **Optional: Basic Machine Learning**
   - **Bonus Task:** Use a simple linear regression model to predict future sales based on past data.
   - Split the dataset into training and testing sets, train the model, and evaluate its accuracy.
5. **Automation & Reporting:**
   - Generate a summary report in the form of a CSV file named `sales_summary.csv`. This report should include the following columns: `ProductID`, `ProductName`, `TotalSales`, `TotalProfit`, and `Top5BestSeller` (where `Top5BestSeller` is a boolean indicating whether the product is among the top 5 best-sellers).
6. **Code Documentation:**
   - Write clear and concise comments in your code to explain your logic and the purpose of each major section.
   - Ensure that the code is well-structured and easy to follow.

Submission Requirements:

- Your script should be named `sales_analysis.py`.
- Include the `sales_data.csv` and `product_info.csv` files used for testing, or generate sample data within the script.
- Ensure the script runs without errors and produces the expected output.
- Submit your Python script, along with any necessary data files, in a ZIP archive.

Evaluation Criteria:

- **Code Quality:** Readability, structure, and use of Python best practices.
- **Problem-Solving:** Ability to correctly perform the required data processing and analysis.

- **Data Integration:** Correct and efficient merging of datasets.
- **Optional Task:** (If completed) Accuracy and implementation of the machine learning model.
- **Documentation:** Clarity of comments and code organization.