Experiment No: 4

Advanced Operations on CSV Files using pandas

Objective:

To develop a Python program that processes sales data stored across multiple CSV files, calculates the total and average monthly quantity sold per product, identifies the top 5 best-selling products, and generates a summary report in CSV format.

Task Description:

You are working as a data engineer for a large retail company. Your team is responsible for processing and analyzing sales data from multiple stores across the country. The sales data is stored in CSV files, and each file represents sales data for a specific month and year. Each CSV file has the following columns:

- Date (in the format "YYYY-MM-DD")
- Store ID (a unique alphanumeric code)
- Product ID (a unique alphanumeric code)
- Quantity sold (an integer representing the number of products sold on that date)

The "product_names.csv" file has two columns: "Product ID" and "Product Name," and it contains the mapping for all products in the sales data.

Your task is to write a Python program that performs the following operations:

- Read the sales data from all the CSV files in a given directory and its subdirectories.
- Calculate the total sales (quantity sold) for each product across all stores and all months.
- Determine the top 5 best-selling products in terms of the total quantity sold.

Create a new CSV file named "sales_summary.csv" and write the following information into it:

- Product ID
- Product Name
- Total Quantity Sold
- Average Quantity Sold per month (considering all months available in the data

Steps to Perform the Program:

1. Import necessary libraries:
   - Use os to navigate directories and files.
   - Use csv or pandas to handle CSV data.
2. Load product_names.csv into a dictionary:
   - Map each Product ID to its corresponding Product Name.
3. Traverse all CSV files:

- Use os.walk() to search recursively.
- Read each file that ends with .csv and has a valid sales format (excluding product_names.csv).

4. Parse each sales file:
    - Read rows and extract:
        - Product ID
        - Quantity Sold
    - Maintain a dictionary to accumulate total quantity per product.
5. Count the number of months:
    - Keep track of how many unique files (months) are processed to compute monthly averages.
6. Calculate statistics:
    - Total quantity sold = sum of quantities across all files.
    - Average = total quantity / number of months.
7. Sort products by total quantity sold and extract the top 5.
8. Write to sales_summary.csv:
    - Include Product ID, Product Name, Total Quantity Sold, Average Quantity Sold per Month.