Data Analysis with Pandas: Problems 02

This assignment focuses on integrating and analyzing data from multiple sources. You will use advanced Pandas techniques like merging DataFrames, handling missing data, and performing detailed time-series analysis to solve a new set of business problems. Two datasets have been provided for this assignment.

The Use Case: Multi-Source Data Analysis

> The business owner, Ms. Kavita, now has two separate datasets: `sales_data.csv` (containing daily order details) and `customer_info.csv` (containing customer names and their city of residence). She needs to combine these datasets to gain deeper insights into her customer base and product sales. Your task is to use Pandas to link these two data sources and answer her questions.

Instructions

For each problem, write and execute the Python code using Pandas. The problems are designed to be solved sequentially. Load both datasets and use them as needed.

Problem 1: Data Loading and Merging

Your first task is to load both CSV files and merge them into a single DataFrame. This combined dataset will be the foundation for all subsequent analysis.

Write Python code to:

- Load `sales_data.csv` into a DataFrame named `df_sales`.
- Load `customer_info.csv` into a DataFrame named `df_customers`.
- Merge the two DataFrames on a common column. Choose the correct join type to ensure no sales records are lost.
- Display the first 5 rows and the column information of the new, merged DataFrame.

Hint: Look for a common identifier column in both datasets to perform the join.

Problem 2: Advanced Analytical Questions with Merged Data

Now that you have the combined dataset, answer Ms. Kavita's more complex questions that require customer and sales information together.

Write Python code to:

- Find the total sales revenue generated from customers in each city.
- Identify the top 3 best-selling products by quantity.
- Determine the city with the highest total revenue.
- Find the customer who has spent the most money in total.

Problem 3: Cleaning and Filtering for Specific Insights

Ms. Kavita needs to prepare a report for her marketing team. This requires some data cleaning and specific filtering.

Write Python code to:

- Identify and handle any missing values in the merged DataFrame. Explain your chosen method.
- Filter the DataFrame to show all orders made by customers from 'Mumbai' for the 'Pistachio Delight' product.
- Create a new DataFrame containing only the columns: `customer_name`, `city`, `product_name`, and `total_price_inr` for all orders that have a revenue of more than INR 300.

Problem 4: Time-Series and Product-Specific Analysis

The business wants to understand how sales of specific products trend over time and how they perform on different days.

Write Python code to:

- Convert the 'order_date' column to a proper datetime format.
- Create a new column named 'day_of_week' that shows the day name (e.g., 'Monday', 'Tuesday').
- Calculate the total revenue for the 'Vanilla Dream' product each day.
- Find the average daily revenue for each product.