Scenario

You are a data analyst for a retail company. Your task is to analyze customer and sales data to generate meaningful insights while handling real-world data issues.

Assignment Tasks

Task 1

Load the given datasets into Pandas DataFrames. Inspect the datasets and perform the following:

- Display the first few rows of each dataset.
- Show the total number of rows and columns.
- Check for missing values in each dataset and handle them appropriately.

Task 2

Using the `customers.csv` file, convert its data into a Python dictionary. Use the dictionary to filter customers from a specific city. Repeat the operation using a DataFrame and compare the efficiency of both approaches.

Task 3

Identify duplicate rows, if any, in the datasets. Remove these duplicates to ensure clean data. After cleaning, verify that there are no duplicates left.

Task 4

Create a new column in the `sales.csv` data that reflects the total amount after applying a 10% discount on the `Amount` column. Group the data by `Product` and calculate the total sales for each product. Present the results in a well-structured format.

Task 5

Filter the data in the `customers.csv` file to retain only those customers whose age falls in the range of 25 to 35. Save the filtered data in a new structure and analyze how many customers belong to each city within this age range.

Task 6

Merge the `customers.csv` and `sales.csv` datasets on `CustomerID`. From the merged dataset:

- Identify the city that generated the highest total sales.
- Find the product with the most units sold.

Task 7

Explore the merged dataset to derive insights:

- Display the unique values in the `City` and `Product` columns.
- Calculate the mean and median of the `Amount` column.

Submission Guidelines

- Submit your work as a Jupyter Notebook (`.ipynb`) file.
- Ensure all steps, observations, and results are with appropriate comments.