Data Manipulation Assignment: Combining and Analyzing Retail Data

This assignment is designed to test your ability to work with multiple datasets using Pandas. You will be provided with separate files for customer information and sales transactions. Your task is to clean, combine, and analyze these datasets to gain a complete view of our business performance.

The Datasets

You will work with the following files:

- `customers.csv`: Contains customer-level information (`CustomerID`, `CustomerName`, `CustomerAge`, `Gender`, `City`).
- `online_sales.csv`: Contains online sales transactions (`TransactionID`, `CustomerID`, `ProductCategory`, `TransactionAmount`, `SalesChannel`, `PurchaseDate`).
- `in_store_sales.csv`: Contains in-store sales transactions (`TransactionID`, `CustomerID`, `ProductCategory`, `TransactionAmount`, `SalesChannel`, `PurchaseDate`).

Part 1: Data Loading and Initial Inspection

Objective: Load the three datasets and perform a preliminary inspection of their structure.

1. Load `customers.csv`, `online_sales.csv`, and `in_store_sales.csv` into separate Pandas DataFrames.
2. Display the first 5 rows and a summary (`.info()`) for each DataFrame to understand its structure.

Part 2: Data Concatenation (`pd.concat`)

Objective: Combine the sales data from different channels into a single, unified DataFrame.

1. Concatenate the `online_sales` and `in_store_sales` DataFrames to create a new DataFrame called `all_sales_df`.
2. Check the shape and the first 5 rows of `all_sales_df` to ensure the concatenation was successful.
3. Check for and handle any duplicate entries in the `all_sales_df`.
4. Display the counts of each `SalesChannel` to confirm all data was included.

Part 3: Data Merging (`pd.merge`) and Analysis

Objective: Join the combined sales data with the customer data to enable a rich, cross-dataset analysis.

1. Merge the `all_sales_df` with the `customers` DataFrame on the common column `CustomerID`. Store the result in a new DataFrame called `combined_data`.
2. Check the shape and `info()` of `combined_data` to ensure the join was successful.
3. Find the top 5 cities with the highest total sales.
4. Calculate the average transaction amount for each `ProductCategory`.
5. Determine the number of unique customers who made a purchase in each `SalesChannel`.

Part 4: Data Visualization

Objective: Create informative plots to visually represent key insights from the combined data.

1. Create a bar plot showing the total `TransactionAmount` for each `City`.
2. Visualize the total sales amount by `ProductCategory` using a bar plot.
3. Create a histogram of the `CustomerAge` distribution for customers who made a purchase.
4. Use a box plot to show the distribution of `TransactionAmount` across different `ProductCategory`s.
5. Create a bar plot to show the average `TransactionAmount` by `Gender`.