Assignment: Data Profiling using Pandas and YData

Objective:

This assignment will test your ability to perform data profiling using `pandas` and `ydata-profiling`. You will explore, analyze, and generate insights from a dataset using these libraries.

Dataset: Retail Sales Data

The dataset contains **sales transactions** from an online retail store. It includes customer details, order information, and product attributes.

Dataset Columns:

- `Customer_ID` (Unique ID of the customer)
- `Age` (Customer's age)
- `Gender` (Male/Female)
- `Region` (Geographical region)
- `Product_Category` (Category of purchased product)
- `Product_Name` (Name of product)
- `Quantity` (Number of units purchased)
- `Unit_Price` (Price per unit)
- `Total_Spend` (Total purchase value)
- `Payment_Method` (Credit Card/Debit Card/UPI/Wallet)
- `Discount_Applied` (Yes/No)
- `Review_Rating` (Customer review: 1-5)
- `Purchase_Date` (Date of purchase)
- `Delivery_Days` (Days taken to deliver)
- `Return_Status` (Returned/Not Returned)

Tasks:

Perform the following tasks using **pandas** and **ydata-profiling**.

Exploratory Data Analysis & Data Profiling

1. Load the dataset into a Pandas DataFrame and display its structure.

2. Generate a **pandas-profiling report** using `ydata-profiling`. What are the key observations from the report?
3. Identify **missing values** and suggest possible ways to handle them.
4. Identify **outliers** in the `Total_Spend` column. How can they impact business decisions?
5. Check for **duplicate records** and suggest an approach to remove them.

Statistical Insights & Data Quality Checks

6. Compute the **mean, median, and standard deviation** for `Total_Spend`. What do these metrics indicate?
7. Analyze the **correlation** between `Discount_Applied` and `Total_Spend`. Does offering discounts increase spending?
8. Identify the **most popular product category** using a frequency distribution.
9. What is the **distribution of customer age groups** (e.g., 18-25, 26-35, etc.)?
10. Find the **top 3 regions** where customers spend the most.

Case Study-Based Questions

11. If a business wants to offer discounts, which **customer age group** should be targeted? Justify your answer with data.
12. Analyze the **relationship between Delivery Days and Return Status**. Do longer delivery times lead to more returns?
13. Based on the dataset, what percentage of customers prefer **digital wallets** over other payment methods?
14. Determine the **average review rating** for products that had a return. Do customers give lower ratings for returned products?
15. Which **product category** has the highest return rate? What could be the possible reasons?

Advanced Analysis & Business Recommendations

16. Perform a **hypothesis test** to check if there is a significant difference between average spend for male vs. female customers.
17. Use a **box plot** to analyze the distribution of `Total_Spend` across different payment methods.
18. Identify **high-value customers** (Top 10% based on total spend) and suggest marketing strategies to retain them.
19. If the business plans to launch a **loyalty program**, which metric would be most useful to segment customers?
20. Summarize key **actionable insights** for the business based on the profiling report.