Walmart Project Objective and Attributes

Dataset Overview: The dataset provided by Walmart Inc. contains detailed information related to customer transactions, product details, and logistics. It has 1000 rows and 30 columns, each representing a different aspect of the company's operations. Below is a detailed explanation of the columns:

Attribute Details:

Customer_ID: A unique identifier for each customer. Region: The geographical region where the transaction took place (e.g., North, South, East, West).

Product_Category: The category of the product purchased (e.g., Electronics, Clothing, Groceries).

Sales_Amount: The total sales amount for the transaction.

Quantity_Sold: The number of units sold in the transaction.

Customer_Satisfaction: A satisfaction score provided by the customer (0 to 5).

Stock_Level: The stock level of the product at the time of the transaction.

Date_of_Purchase: The date when the transaction occurred.

Product_Price: The price per unit of the product.

Discount_Applied: The discount applied to the product in the transaction.

Shipping_Cost: The cost incurred for shipping the product.

Total_Revenue: The total revenue generated from the transaction (after discount). Profit: The profit made from the transaction (Total Revenue - Product Price - Shipping Cost).

Customer_Age: The age of the customer.

Customer_Gender: The gender of the customer (Male or Female).

Payment_Method: The method used for payment (e.g., Credit Card, Debit Card, PayPal, Cash).

Order_Status: The status of the order (e.g., Completed, Pending, Cancelled).

Shipping_Time: The time taken to ship the product (in days).

Product_Returned: Whether the product was returned by the customer (Yes or No).

Loyalty_Points_Earned: The loyalty points earned by the customer for this transaction.

Promotion_Applied: Whether a promotion was applied to the transaction (Yes or No).

Order_Source: The source of the order (e.g., Online, In-Store, Mobile App).

Customer_Segment: The customer segment (e.g., Regular, VIP, New).

Warehouse_Location: The warehouse from which the product was shipped (e.g., Warehouse A, B, C).

Supplier_Rating: The rating of the supplier who provided the product (0 to 5).

Product_Weight: The weight of the product (in kg).

Product_Dimensions: The dimensions of the product (in cm^3).

Delivery_Time: The time taken to deliver the product (in days).

Return_Reason: The reason for returning the product (if returned).

Customer_Feedback_Score: The feedback score provided by the customer (0 to 5).

Project Overview:

The project is divided into four sprints, each focusing on different aspects of data analysis using NumPy and Pandas. The team will work collaboratively to achieve the goals set for each sprint.

Sprint 1: Data Collection and Cleaning

Objective: The management team at Walmart Inc. has provided raw data that needs to be cleaned and organized before any meaningful analysis can be performed. The team will focus on loading the data, identifying and correcting errors, handling missing values, and ensuring data integrity.

Tasks:

1. Load Data: Use Pandas to load the provided Excel file into a DataFrame.

2. Data Cleaning: Clean the data by handling missing values, removing duplicates, and correcting data types.

3. Initial Data Exploration: Perform basic exploratory data analysis (EDA) to understand the distribution and relationships between variables.

4. Documentation: Document the data cleaning process, including the steps taken and any issues encountered.

Deliverables:

• Cleaned DataFrame.

• Data cleaning documentation.

Sprint 2: Advanced Data Manipulation with NumPy

Objective: With the data cleaned, the next step involves performing advanced data manipulation using NumPy. The management team is interested in statistical properties, trends, and detailed analysis across different regions and product categories.

Scenario 1: Sales Performance Matrix

1. Create and Manipulate Sales Performance Matrix: Create a 2D NumPy array representing sales data, where rows correspond to regions and columns correspond to product categories.

2. Implement Advanced Indexing and Slicing: Extract specific data points from the matrix using advanced indexing and slicing techniques.

Deliverables:

• Transposed matrix, summary statistics, top-performing regions for each product category.

• Extracted data for electronics, comparative analysis between North and South regions, percentage change in sales.

Scenario 2: Inventory Optimization

3. Calculate Stock Turnover Rates: Calculate the stock turnover rate for each product category using NumPy.

4. Simulate Inventory Scenarios: Simulate different inventory scenarios using NumPy to understand the impact of varying stock levels.

Deliverables:

• Stock turnover rates, interpretation of high and low turnover products.

• Simulated turnover rates for increased and decreased stock levels, analysis of inventory impact.

Scenario 3: Sales Trend Analysis

5. Analyze Regional Sales Trends: Use NumPy to analyze sales trends across different regions over time.

6. Evaluate Discount Impact on Sales: Analyze the impact of discounts on sales volumes using NumPy.

Deliverables:

• Time series analysis of regional sales trends, identification of peak and low seasons.

• Correlation analysis between discounts and sales volumes, identification of optimal discount levels.

Documentation:

Document the NumPy operations performed and the insights gained.

Output: Detailed report of NumPy operations and findings.

Sprint 3: Advanced Data Analysis with Pandas

Objective: After manipulating the data with NumPy, the next step is to dive deeper into the data using Pandas for more detailed analysis. The management team seeks to understand customer behavior, product performance, and sales patterns to make data-driven decisions.

Scenario 1: Customer Segmentation Analysis

1. Segment Customers Using Pandas: Segment customers based on their purchasing behavior and demographics.

2. Correlation Analysis Using Pandas: Analyze the correlation between customer satisfaction scores and total spending.

Deliverables:

• Customer segments, analysis of product preferences by segment.

• Correlation coefficient, interpretation of results.

Scenario 2: Product Bundling Strategy

3. Identify Product Bundles: Identify products frequently purchased together and recommend potential bundles.

4. Optimize Bundling Strategy: Simulate different bundling scenarios using Pandas and recommend the most profitable strategy.

Deliverables:

• Product bundles, estimated impact on sales from bundling.

• Simulated revenue estimates, recommended bundling strategy.

Scenario 3: Sales Channel Analysis

5. Analyze Sales by Order Source: Analyze sales performance across different order sources (Online, In-Store, Mobile App).

6. Evaluate Payment Method Preferences: Analyze customer payment method preferences and their impact on sales.

Deliverables:

• Sales trends by order source, identification of the most profitable channel.

• Analysis of payment methods, recommendations for payment options to promote.

Documentation:

Document the analysis process, key insights, and visualizations. Output: Comprehensive analysis report with visualizations.