Simulating MapReduce for Customer Purchase Pattern Analysis

1. Objective

The objective of this mini project is to simulate the MapReduce framework using a retail transaction dataset to derive meaningful business insights. This simulation demonstrates how large-scale data processing techniques can be applied in spreadsheet environments like Excel. The primary goals of this project include:

- Identifying the top 5 selling products
- Calculating total revenue by product category
- Determining the most frequent buyers

2. Dataset Description

The dataset contains 1000 transaction records representing purchases made by customers. Each transaction includes details such as:

- Transaction ID (Order ID)
- Customer ID
- Product Category
- Product Sub-Category
- Quantity
- Unit Price
- Transaction Date
- Payment Mode
- Region

Data preparation steps included:

- Cleaning the dataset by removing blank Customer IDs
- Filtering out negative or zero Quantity values
- Creating a Total Sales column using Quantity * Unit Price
- Using XLOOKUP to ensure accurate Category mapping from Sub-Category
- Adding Month column by using TEXT function

- XLOOKUP formula used:

=XLOOKUP([@Subcategory], CategoryMap!A:A, CategoryMap!B:B, "Not Found")

- Total Sales formula:

=Quantity * Unit Price

- Month formula:

=TEXT([@Transaction Date],"mmmm")

3. Methodology: Simulating MapReduce

The project simulates two key phases of the MapReduce model:

Map Phase

- Product → Quantity (to calculate total sold quantity)
- Category → Revenue (Quantity × Unit Price)
- Customer ID → Frequency (count of transactions)

Reduce Phase

- Aggregated total quantities for each product
- Aggregated total revenue for each product category
- Aggregated number of transactions per customer

These computations were performed using Excel Pivot Tables.

4. Key Insights and Findings

- Top 5 Selling Products: Shirts, Laptops, Fiction books, Phones, and T-shirts appeared most frequently in customer transactions.
- Top Revenue-Generating Categories: Electronics and Clothing were the highest revenue contributors.
- Most Frequent Buyers: Customers like Customer_977 and Customer_4294 placed the highest number of orders, indicating strong brand loyalty.

5. Visual Representations

To support the findings, the following charts were created using Excel:

- A. Bar Chart: Top 5 Products by Quantity Sold
- B. Pie Chart: Revenue Distribution by Category
- C. Bar Chart: Top 5 Customers by Purchase Frequency
- Bar Chart: Regional Sales
- Bar Chart: Frequency of Payment Mode

- Combo Chart: Revenue Distribution by Month

These visuals were dynamically linked to the pivot tables and filtered to show only the top results.

6. Conclusion

This mini project successfully demonstrates the power of MapReduce concepts in a simplified spreadsheet environment. By mapping and reducing the dataset through Pivot Tables and Excel functions, critical insights were extracted that can guide inventory, marketing, and customer relationship strategies. The use of basic tools to simulate big data techniques underlines the versatility and practicality of analytical thinking in business intelligence.