A.SHIVANI REDDY

2211CS010003(S2)

Sales Data Analysis 2017

Dataset: Sales2017 (1).csv

1.Introduction

- This project focuses on performing data analysis and visualization of 2017 sales data using PySpark.
- The dataset contains transactional details such as OrderDate, ProductKey, CustomerKey, TerritoryKey, and OrderQuantity.
- The main objective is to explore sales trends, analyze order quantities, and understand performance across different products and territories.

2.Initial Analysis of the Dataset

- **Dataset Size**
  - The dataset contains 29,481 records and 8 columns.
  - It was loaded successfully using PySpark's spark.read.csv() with header=True and inferSchema=True.
- **Key Columns**
  - **Numerical Columns:** ProductKey, CustomerKey, TerritoryKey, OrderLineItem, OrderQuantity.
  - **Date/String Columns:** OrderDate, StockDate, OrderNumber.
- **Data Summary**
  - The data covers orders from January 1, 2017, to June 9, 2017.
  - There are 10,502 distinct customers, 102 distinct products, and 10 distinct territories represented in the dataset.

3.Dataset Observations

- **General Observations**
  - The dataset is well-structured, containing transactional sales data for the first half of 2017.

    - The use of PySpark indicates the data is suitable for scalable, big data analysis techniques.

- **Numerical Insights**
    - The average order quantity per line item is approximately 1.54.
    - The dataset shows a broad customer base and a moderate range of products, allowing for analysis of customer purchasing habits.
- **Behavioral & Regional Trends**
    - The presence of TerritoryKey allows for geographical analysis of sales performance.
    - The OrderDate column enables time-series analysis to identify monthly or seasonal sales trends.

4.Graphs

- **1. Histogram (Order Quantity Distribution)**
    - **Purpose:** To show the frequency distribution of order quantities per line item.
    - **Observations:** Most order line items consist of a small quantity (1 or 2), with the frequency decreasing as the quantity increases.
- **2. Bar Chart (Average Order Quantity per Product)**
    - **Purpose:** To compare the average order quantity for each distinct product.
    - **Observation:** The average order quantity varies across different products, suggesting some items are typically bought in larger quantities than others.
- **3. Bar Chart (Total Orders by Territory)**
    - **Purpose:** To visualize and compare the total quantity of items ordered across different sales territories.
    - **Observation:** There is a significant variation in sales volume by territory, highlighting regions with stronger and weaker market performance.
- **4. Line Chart (Monthly Sales Trend - 2017)**
    - **Purpose:** To track the total order quantity on a monthly basis for the first half of 2017.
    - **Observation:** The chart displays fluctuations in sales from January to June, which can be used to identify high-performing months or potential seasonal trends.
- **5. Bar Chart (Top 10 Customers by Order Quantity)**

    - **Purpose:** To identify the top 10 customers based on the total quantity of items they ordered.
    - **Observation:** A small number of customers are responsible for a large volume of orders, indicating key accounts or high-value clients.
- **6. Heatmap (Correlation Heatmap)**
    - **Purpose:** To visualize the correlation between the numerical features in the dataset.
    - **Observation:** The heatmap reveals relationships between variables; for instance, it can show if there is any correlation between territory and order quantity or between customer key and product key.

5.Conclusion

- The analysis of the 2017 sales dataset provided valuable insights into transaction patterns, product performance, and customer behavior.
- **Key findings indicate that:**
    - Typical order quantities are low, with most transactions involving one or two items.
    - Sales performance varies significantly across different products and geographical territories.
    - Monthly trends show noticeable fluctuations in sales volume over the first six months of the year.

6.Future Scope

- **Predictive Modeling:** Build machine learning models to forecast monthly or quarterly sales based on historical data.
- **Customer Segmentation:** Use clustering techniques to group customers with similar purchasing behaviors for targeted marketing campaigns.
- **Time Series Analysis:** If more data were available, a deeper time series analysis could be performed to identify long-term seasonality and growth trends.
- **Recommendation System:** Develop a basic product recommendation model based on customer purchase history.