Data Spark: Illuminating Insights for Global Electronics

Table of Contents

1 CUSTOMER ANALYSIS:

Analysis Summary of the Datasets

Demographics: The dataset contains customer demographics, including gender, age, location (city, state, country), and birthday. The gender distribution shows a fairly balanced ratio, with slightly more male customers (7748) than female customers (7518).

Age Distribution: The analysis of customer age indicates a range of ages, which can be visualized through a histogram. This distribution can inform marketing strategies and product targeting based on age groups.

Handling Missing Values:

The dataset had missing values in the 'State Code' column, which were imputed using the mode (most frequent value). Post-imputation, there were no missing values left in the dataset.

Location Analysis:

State and Country Distribution: The analysis revealed a diverse customer base across various states and countries. The United States had the highest number of customers (6828), followed by the United Kingdom (1944) and Canada (1553). This insight can aid in regional marketing efforts.

Sales Data Integration:

Merging the sales data with customer data allowed for a comprehensive view of customer behavior. It enabled the calculation of total sales and the number of orders for each customer, which is critical for understanding customer value.

Customer Segmentation:

Customers were segmented based on their Average Order Value (AOV):

Low Value: AOV < 10

Medium Value: $10 \leq AOV < 20$

High Value: AOV ≥ 20

This segmentation helps identify high-value customers and tailor marketing strategies accordingly.

Average Age by Segment:

The average age of customers varied by segment, providing additional demographic insights. High-value customers had an average age of approximately 54, while low-value customers had a slightly higher average age of 55.4.

Data Storage:

The combined analysis was successfully saved into a CSV file for further use, and the data was also stored in a MySQL database for easy access and querying.

Database: Datas

This documentation provides an overview of the SQL queries used for analyzing customer, sales, product, and store data within the Global Electronics dataset.

Customer Demographics Distribution

Analyzes the total number of customers, average age, and segments them by demographic groups.

Purchase Patterns

Examines average order values and total orders for each customer segment.

Customer Segmentation

Segments customers into low, medium, and high-value groups based on total sales.

POWER BI

Demographic Distribution

Topic Name: Customer Demographics

X Label: Age or Customer Segment

Y Label: Number of Customers

Type: Bar Chart or Pie Chart

Purchase Patterns

Topic Name: Purchase Patterns

X Label: Customer Key

Y Label: Average Order Value

Type: Line Chart

Customer Segmentation

Topic Name: Customer Segmentation

X Label: Customer Segment

Label: Total Sales

Type: Bar Chart

---

2 SALES ANALYSIS:

Analysis Summary of the Datasets

Sales Data Analysis:

Sales Overview: The sales dataset includes details such as order numbers, order dates, delivery dates, product keys, quantities, and currency codes. The total number of sales records is 62,884, indicating a substantial volume of transactions.

Sales Trends: An analysis of monthly sales indicates seasonal trends, with noticeable peaks during certain months. This insight can guide inventory and marketing strategies to maximize sales during high-demand periods.

Handling Missing Values:

The dataset initially had significant missing values in the 'Delivery Date' column (49,719). These were addressed by filling in missing entries with "Not Delivered," resulting in a clean dataset with no remaining missing values.

Product Performance Analysis:

Top Products: The analysis revealed the top 10 products by total sales and quantity sold, identifying key drivers of revenue. This information is crucial for inventory management and targeted promotions.

Product Categories: Grouping sales data by product category provided insights into which categories perform best, aiding in strategic product placements and marketing efforts.

Store Performance Analysis:

Store Comparison: Sales were analyzed across different stores, revealing performance disparities. The analysis showed that certain stores consistently outperform others, indicating potential for focused resource allocation and promotional efforts in underperforming locations.

Sales by Currency:

Currency Distribution: The analysis identified the quantity of sales by currency, with the majority of sales conducted in USD. This insight assists in understanding market dynamics and pricing strategies in different regions.

Sales Insights and Visualizations:

Visual Reports: Various visualizations, including line charts for monthly sales trends and bar charts for top products and stores, were created to provide clear and actionable insights for stakeholders.

Data Storage:

The combined analysis results were saved into a CSV file for further use, and the data was successfully stored in a MySQL database under the table name `sales_analysis`, facilitating easy access and future queries.

Database: Datas

This documentation provides an overview of the SQL queries used for analyzing customer, sales, product, and store data within the Global Electronics dataset.

Sales Data Overview

Displays all records from the sales analysis table and describes its structure.

Overall Sales Performance

Aggregates total sales on a monthly basis.

Sales by Product

Identifies the top products by quantity sold and calculates total revenue.

Sales by Store

Evaluates sales performance across different stores, including total sales and order count.

Sales Count by Currency Code

Counts the number of transactions per currency code.

POWER BI

Overall Sales Performance

Topic Name: Sales Over Time

X Label: Order Date (Month/Year)

Y Label: Total Sales

Type: Line Chart

Sales by Product

Topic Name: Product Sales Performance

X Label: Product Name

Y Label: Total Sales

Type: Bar Chart

Sales by Store

Topic Name: Store Sales Performance

X Label: StoreKey

Y Label: Total Sales

Type: Column Chart

Sales by Currency

Topic Name: Currency Impact on Sales

X Label: Currency Code

Y Label: Total Sales

Type: Pie Chart or Bar Chart

---

3 Product Analysis Summary

Data Loading:

The product data was loaded from a CSV file into a DataFrame using Pandas. The dataset comprises 2,517 products with attributes such as `ProductKey`, `Product Name`, `Brand`, `Color`, `Unit Cost USD`, `Unit Price USD`, `Subcategory`, and `Category`.

Data Cleaning:

Missing Values: Initial checks revealed no missing values in key columns.

Data Types: Columns for `Unit Cost USD` and `Unit Price USD` were initially strings and were converted to numeric values by removing non-numeric characters.

Duplicates: No duplicate rows were found in the dataset.

Summary Statistics:

- A summary of statistics was generated, revealing average unit costs of approximately $147.66 and average unit prices of around $356.83.

- Unique brands (11) and categories (8) were also identified.

Sales Data Integration:

- Sales data was loaded from another CSV file and merged with the product data on `ProductKey`.

- Post-merge, the merged DataFrame contained 62,909 entries, with some missing values in sales-related columns.

Handling Missing Values in Merged Data:

- Missing values in critical columns like `Order Number`, `Line Item`, and `Order Date` were removed.

- For categorical columns like `Delivery Date`, missing values were filled with "Not Delivered".

Sales Analysis:

Total Sales Calculation: The total sales amount was calculated as $55,755,479.59.

Sales by Category: Sales were grouped by category, revealing the highest sales in "Computers" ($19,301,595.46) and "Home Appliances" ($10,795,478.59).

Product Popularity: The most popular products were identified, with the top seller being "WWI Desktop PC2.33 X2330 Black" at $505,450. The least popular product recorded only $15.20 in sales.

Profitability Analysis: Profit was calculated for each product, identifying the most profitable product as "Adventure Works Desktop PC2.33 XD233 Black" with a profit of $97,840.45.

Category and Subcategory Analysis:

Sales were analyzed by both category and subcategory, showing a detailed breakdown of sales performance.

Data Visualization:

- Visualizations were created for total sales by category and most popular products, providing insights into sales performance through bar charts.

Data Storage:

The final combined analysis was saved as a CSV file and also stored in a MySQL database, ensuring easy access for future queries.

Database: Datas

This documentation provides an overview of the SQL queries used for analyzing customer, sales, product, and store data within the Global Electronics dataset.

Products Data Overview

Displays all records from the product analysis table.

Product Profitability

Calculates the profit margin percentage for each product based on sales and cost data.

POWER BI

Product Popularity

Topic Name: Most/Least Popular Products

X Label: Product Name

Y Label: Quantity Sold

Type: Bar Chart

Profitability Analysis

Topic Name: Product Profit Margins

X Label: Product Name

Y Label: Profit Margin (Calculated Field)

Type: Scatter Plot

Category Analysis

Topic Name: Sales by Category

X Label: Category

Y Label: Total Sales

Type: Bar Chart

---

4 STORE ANALYSIS:

Analysis Summary of the Datasets

Store Data Analysis:

Store Distribution: The dataset includes information about various stores, including location (country and state), size (in square meters), and opening date. A significant concentration of stores is located in the United States (24 stores), followed by Germany (9) and the United Kingdom (7).

Size Distribution: The analysis of store sizes shows a range of square meters, allowing for insights into store capacity and potential sales performance based on size. This can be visualized using box plots.

Handling Missing Values:

The 'Square Meters' column had one missing value, which was filled with the mean size of the stores. After this imputation, there were no missing values left in the dataset.

Sales Data Integration:

Merging the sales data with store data provided insights into store performance metrics. This integration allowed for the calculation of total sales quantity per store, which is critical for assessing each store's contribution to overall sales.

Store Performance Metrics:

Stores were evaluated based on total sales quantity, average size, and operational years. For instance, the store with the highest sales quantity sold 41,311 items, while others showed significantly lower performance. This analysis can guide resource allocation and strategic planning.

Geographical Sales Analysis:

The analysis identified sales performance by country and state, revealing that the United States generated the highest total sales quantity (83,638 units), followed by Online (41,311) and the United Kingdom (20,625). This information can inform targeted marketing efforts and inventory management.

Correlation Analysis:

A correlation matrix between store size and sales quantity was created, highlighting potential relationships that could exist between the physical size of a store and its sales performance. This analysis aids in understanding operational efficiencies.

Data Storage:

The results of the store analysis, including performance metrics and geographical sales data, were saved into CSV files for further analysis and were also stored in a MySQL database for easy access and querying.

Database: Datas

This documentation provides an overview of the SQL queries used for analyzing customer, sales, product, and store data within the Global Electronics dataset.

Stores Data Overview

Displays all records from the store analysis table.

Store Size Efficiency

Evaluates sales efficiency per square meter for each store.

Geographical Sales Analysis

Provides insights into sales across different geographical locations by joining store and sales data.

POWER BI

Store Performance

Topic Name: Store Performance Metrics

X Label: StoreKey

Y Label: Total Sales Quantity

Type: Column Chart

Geographical Sales

Topic Name: Geographical Sales Distribution

X Label: Country or State

Y Label: Quantity Sold

Type: Map Visualization (Choropleth)

Conclusion

The comprehensive analysis of the Global Electronics dataset has yielded valuable insights across customer, sales, product, and store dimensions.

Customer Insights: The demographic distribution highlights a balanced gender ratio and diverse age groups, allowing for targeted marketing strategies. The segmentation of customers based on Average Order Value (AOV) enables focused efforts on high-value customers, enhancing customer retention and loyalty.

Sales Performance: An analysis of sales trends reveals seasonal patterns that can guide inventory management and promotional campaigns. Understanding sales by product and currency provides insights into market dynamics, enabling tailored pricing and marketing strategies for different regions.

Product Effectiveness: The identification of top-performing products and categories informs strategic decisions on inventory management and marketing initiatives. The profitability analysis aids in recognizing which products drive the highest margins, allowing for optimized product offerings.

Store Operations: Evaluating store performance reveals disparities that can guide resource allocation and strategic planning. The geographical analysis of sales performance informs

targeted marketing efforts and helps identify opportunities for expansion in underperforming areas.

Overall, the findings underscore the importance of data-driven decision-making in optimizing business operations, enhancing customer engagement, and ultimately driving revenue growth. By leveraging these insights, Global Electronics can develop more effective marketing strategies, improve product offerings, and enhance overall performance in a competitive market.