YUMA ENERGY DATA ANALYST ASSIGNMENT

By: Himanshi Garg

1. Project Overview

The goal of this project is to preprocess and analyze a sales dataset to derive meaningful insights. This involves several stages, including data cleaning, validation, aggregation, and visualization. The dataset used in this project is Data Source (sales_transactions), which contains sales transactions with various attributes such as TransactionID, ProductID, Quantity, PricePerUnit, TotalAmount, and CustomerID.

2. Methodology

2.1 Data Preprocessing

Data preprocessing is essential to ensure data quality and consistency before analysis. The steps involved are:

1. **Data Cleaning**:
    - **Remove Missing Values**: Dropped rows with missing `TransactionID` or `ProductID`. Filled missing values in `CustomerID` and `PaymentMethod` with 'Unknown'.
    - **Correct Negative Values**: Converted negative values in `Quantity`, `PricePerUnit`, and `TotalAmount` to their absolute values.
    - **Handle Missing and Placeholder Values**: Replaced missing values in `PricePerUnit`, `Quantity`, and `DiscountApplied` with appropriate placeholders.
    - **Adjust Dates and Times**: Corrected the format of `TransactionDate` and split it into date and time components.
2. **Outlier Detection**:
    - **Outliers in `TotalAmount`**: Used the Interquartile Range (IQR) method to identify and remove outliers from the `TotalAmount` column.

2.2 Data Validation

To verify the accuracy of our preprocessing:

- **Checking for Missing Values**: We examine the dataset to count how many records have missing values in critical fields. This helps ensure that no significant data issues remain.
- **Consistency of Date Formats**: We validate that the `TransactionDate` follows the expected format, ensuring that all dates are standardized.
- **Validating Aggregations**: We compare the counts and summaries from the original and processed datasets to ensure that no data was lost or incorrectly altered during preprocessing.
- **Verifying Calculated Columns**: We cross-check calculated fields like `TotalAmount` to ensure that they match the expected values based on other data fields, such as `Quantity` and `PricePerUnit`.
- **Checking for Duplicates**: We confirm that there are no duplicate records in the dataset, which ensures the uniqueness of each transaction.

2.3 Data Aggregation

Aggregation involves summarizing the data to extract insights:

- **By Product Category**: We aggregate data to determine the number of transactions, total quantity sold, total sales amount, average discounts given, and total trust points used for each product category. This helps in understanding the performance of different product categories.
- **By Customer ID**: We summarize data to analyze customer-specific metrics, including the number of transactions, total spending, average spending, and total trust points used by each customer. This provides insights into customer behavior and spending patterns.
- **By Transaction Date**: We aggregate data by date to analyze daily transaction counts and total sales amounts. This allows us to observe trends and patterns over time.

2.4 Data Visualization

Visualization helps in interpreting the data and deriving insights:

- **Time Series Analysis**: Line charts are used to show trends in sales over time, helping us understand how sales evolve day by day.
- **Categorical Analysis**: Bar charts are employed to compare metrics across different product categories, making it easier to see which categories perform best.
- **Customer Spending**: Scatter plots are used to visualize relationships between total spending and the number of transactions per customer, which helps identify high-value customers.
- **Distribution Analysis**: Histograms are used to visualize the distribution of total sales amounts, providing insights into the variability and spread of sales data.

2.5 Validation of Visualizations

To ensure that visualizations accurately represent the data:

- **Checking Linearity**: Scatter plots are reviewed to assess if there are any linear relationships between variables.
- **Reviewing Aggregated Results**: Aggregated results from visualizations are compared with raw data to confirm their accuracy.
- **Validating Visuals**: Each chart and graph is checked to ensure it correctly represents the data and that there are no misinterpretations.

3. Conclusion

This project involved a comprehensive approach to data preprocessing, validation, aggregation, and visualization. By cleaning the data, ensuring its accuracy, and applying various visualization techniques, we derived valuable insights that help in understanding sales trends and customer behavior.