Market Basket Analysis Report on Retail Dataset

1 Objective

The objective of this Market Basket Analysis (MBA) is to uncover patterns in customer purchasing behavior using association rule learning techniques. By analyzing retail sales data, we can identify relationships between products, helping retailers make informed decisions on sales, marketing, and inventory strategies.

2 Dataset Overview

The retail dataset includes sales transactions with the following columns:

- **Date**: The transaction date.
- **Gender**: Customer gender.
- **Product Category**: The category of products purchased.
- **Quantity**: Number of units purchased.
- **Price per Unit**: Price of each product.
- **Total Amount**: Total amount spent in the transaction.

3 Exploratory Data Analysis (EDA)

3.1 Monthly Sales Trend

We generated a time series plot of the *Total Amount* over months, showing how sales fluctuated over time. This can help identify seasonal trends, spikes, or declines in sales related to marketing campaigns or events.

3.2 Total Spent by Gender

A bar plot comparing the total amount spent by male and female customers. This provides insights into gender-specific purchasing behavior, which can inform targeted marketing strategies.

3.3 Distribution of Total Amount

A histogram plot revealed the distribution of total amounts spent in each transaction, showing common spending ranges among customers.

3.4 Age Distribution by Gender

Using a box plot, we visualized the age distribution by gender, identifying any significant demographic differences between male and female customers.

3.5 Sales by Product Category

A count plot showed the number of transactions per product category, highlighting which categories are the most popular. This information can help optimize inventory management and promotions.

3.6 Correlation Heatmap

A correlation matrix was used to show the relationships between numerical variables, such as *Age*, *Quantity*, and *Total Amount*. We observed a strong correlation between *Quantity* and *Total Amount*, as expected.

4 Association Rule Learning - Market Basket Analysis

4.1 Methodology

1. **Data Preparation:** The dataset was one-hot encoded, transforming each transaction into a row of binary indicators for purchased product categories. If a product was purchased, the corresponding cell was marked with 1.

2. **Frequent Itemsets:** We applied the Apriori algorithm to find frequent itemsets with a minimum support threshold of 1%. The itemsets represent combinations of products that frequently appear together in transactions.

   $$\text{frequent\_itemsets} = \text{apriori}(basket, \text{min\_support} = 0.01, \text{use\_colnames} = True)$$

3. **Association Rules:** From the frequent itemsets, association rules were extracted using confidence, lift, and leverage metrics to evaluate the strength of each rule. We filtered the rules with a lift greater than 1, indicating meaningful relationships between products.

   $$\text{rules} = \text{association\_rules}(frequent\_itemsets, \text{metric} = "lift", \text{min\_threshold} = 1)$$

4.2 Key Results

Example Rule:

- **Rule:** Customers who buy Product A are likely to buy Product B.
- **Support:** 0.05 (5% of all transactions include both Product A and Product B).
- **Confidence:** 0.8 (80% of customers who bought Product A also bought Product B).
- **Lift:** 1.5 (Customers who bought Product A are 1.5 times more likely to buy Product B compared to the general population).

5 Model Evaluation

We built a linear regression model using *Age*, *Quantity*, and *Price per Unit* to predict *Total Amount* spent in each transaction. The model's performance was evaluated using the following metrics:

- **Mean Squared Error (MSE):** Measures the average of the squared differences between the actual and predicted values. A lower MSE indicates better model accuracy.
- **R-squared (R^2):** This metric indicates the proportion of variance in the dependent variable (*Total Amount*) that can be explained by the independent variables. An R^2 value closer to 1 indicates a stronger model.

$$\text{mse} = \text{mean\_squared\_error}(y\_test, y\_pred)$$

$$\text{r2} = \text{r2\_score}(y\_test, y\_pred)$$

6 Real-World Applications of Rules Discovered

1. **Product Placement Optimization:** Identifying products that are often bought together can help retailers optimize store layout by placing related items near each other to encourage cross-selling.
2. **Personalized Recommendations:** The rules with high confidence and lift can be used to create personalized recommendations, suggesting products based on a customer's past purchases.
3. **Promotions and Bundling:** Retailers can use the association rules to create product bundles for promotions. For example, if chips and soda are frequently bought together, retailers can offer a discount when both are purchased as a bundle.

4. **Inventory Management:** The discovered patterns can help retailers manage stock more effectively by ensuring that frequently bought-together products are stocked in adequate amounts to avoid shortages.

7 Conclusion

This Market Basket Analysis on the retail dataset revealed important insights into customer purchasing behavior. By analyzing association rules, retailers can optimize product placement, personalize recommendations, run targeted promotions, and improve inventory management. These data-driven strategies will help retailers enhance customer satisfaction and increase sales.