Apriori Algorithm Group Assignment

Group 2

Group Members

Name	Grade
Jave Anthony Espiritu	100
Yuan Haley Enriquez	100
James Daniel Esguerra	100
Dale Andrew Fajardo	100
Marc Daniel Falic	100
Emmanuel Joseph Garlando	100
China Mae Gonzales	100
John Pol Jalapan	100

Dataset

- **Name:** Groceries Dataset
- **Source:** Kaggle Groceries Dataset
- **Size:** Over 9,800 transactions
- **Format:** CSV file with columns: `Member_number`, `Date`, `itemDescription`

Source Code

The full source code is available at: GitHub - apriori.ipynb

Explanation of Dataset

The Groceries Dataset records individual items purchased by customers in a supermarket. Each row represents a single item bought by a customer on a specific date. The main columns are:

- `Member_number`: Unique customer ID.
- `Date`: Date of purchase.
- `itemDescription`: Name of the purchased item.

Grouping Strategies Considered:

1. **By customer on a specific date:** Items bought by a customer on a specific date are grouped as one transaction.
2. **By customer in a month:** Items bought by a customer within a month are grouped as one transaction.
3. **By customer (all-time basket):** All items ever bought by a customer are grouped as one transaction.

Chosen Grouping:
For this analysis, we selected **grouping by customer (all-time basket)**. This approach reveals long-term purchasing patterns and overall customer preferences, which are valuable for understanding customer loyalty and persistent product associations.

Explanation of Source Code

- **Data Loading:** The code loads the groceries dataset using pandas.
- **Transaction Preparation:** Transactions are formed by grouping all items ever bought by each customer.
- **Apriori Algorithm:**
  - `get_frequent_itemsets`: Finds all itemsets with support above the minimum threshold using the Apriori principle.
  - `generate_association_rules`: Generates association rules from frequent itemsets with confidence above the threshold.
- **Parameter Selection:** Minimum support and confidence are set based on the distribution of item frequencies.
- **Execution:** The algorithm is run to find frequent itemsets and strong association rules.

Visualization

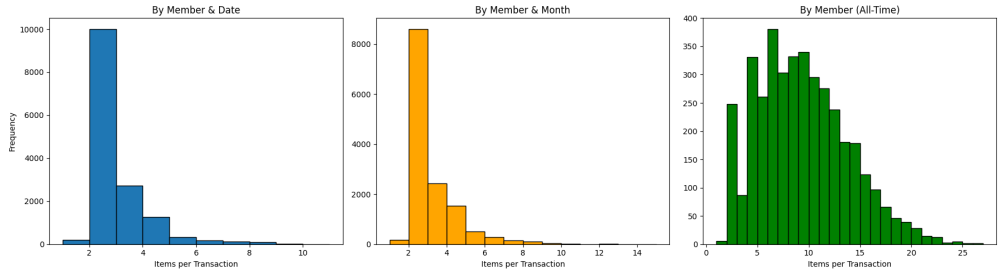

Figure 1: Distribution of transaction sizes for each grouping: (a) by customer per date, (b) by customer per month, (c) by customer all-time.

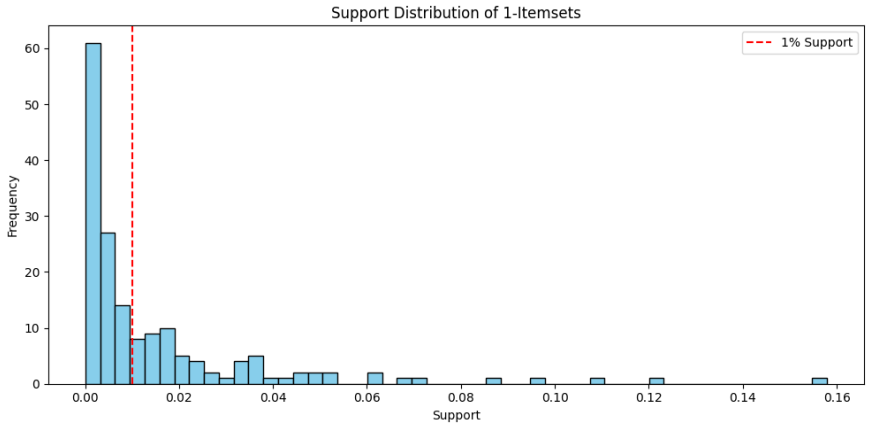

Figure 2: Support Distribution of 1-Itemsets

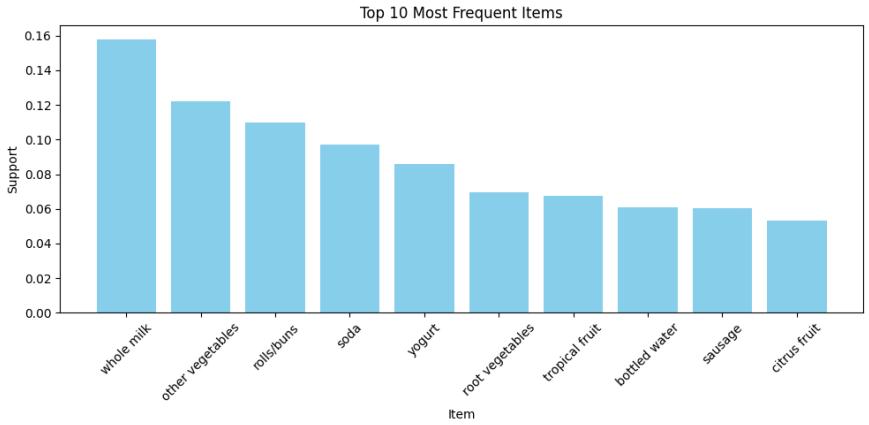

Figure 3: Top 10 Most Frequent Items

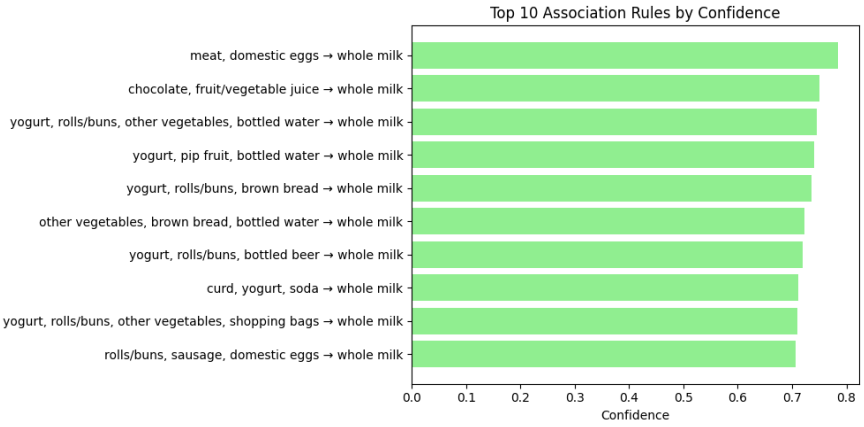

Figure 4: Top 10 Association Rules by Confidence

Conclusion

The Apriori algorithm was successfully applied to the Groceries Dataset using the all-time basket grouping. This approach enabled the discovery of strong, persistent associations between products, providing valuable insights for long-term marketing and customer analysis.