**Introduction**

This report is based on the Online Retail Dataset, which contains transactions from a UK-based online store between 2010 and 2011. The data includes product descriptions, invoice numbers, quantities, and customer IDs. Our goal is to analyze which products are often bought together and create association rules to help in making better marketing and sales decisions.

**Data Preparation & Exploration**

Column Overview:

The dataset has several important columns:

• InvoiceNo: The unique number for each transaction.

• StockCode: Product code.

• Description: Name of the product.

• Quantity: Number of units purchased.

• InvoiceDate: The date and time of the transaction.

• UnitPrice: The price of one unit.

• CustomerID: A unique ID for the customer.

• Country: The country where the customer is located.

Data Inspection:

Before we start the analysis, we checked for:

* Duplicates: We removed duplicate rows to make sure each transaction is counted once.
* Missing Values: Some rows had missing CustomerID, and we removed those because we need complete customer information.
* Negative Quantities: These were canceled orders, so we also removed them.

Data Cleaning

* Removing Missing Values: We deleted rows with missing CustomerID.
* Removing Canceled Orders: Any rows with negative quantities were removed because they represent canceled transactions.
* Removing Duplicates: Duplicated rows were cleaned to avoid repeat data.

After cleaning, the data was ready for analysis.

**Feature Engineering**

Creating a Basket (One-Hot Encoding):

We needed to transform the data into a “basket” format where:

• Rows represent transactions (using InvoiceNo).

• Columns represent products (using Description).

• The values are 1 if the product was purchased in that transaction, or 0 if not.

This transformation helps us to see which items are bought together.

Modeling: Frequent Itemset Mining

Apriori Algorithm Attempt:

We first tried to use the Apriori algorithm to find frequent itemsets. However, it was too slow for this dataset because and it crash the RAM, it needs more time and memory to process large data. Since the Apriori algorithm did not perform well, we switched to the FP-Growth algorithm, which is faster for larger datasets.

FP-Growth Algorithm:

The FP-Growth algorithm was used to find frequent combinations of products. It is faster and works well with large datasets. We set the support threshold at 0.02 (meaning the itemset must appear in at least 2% of the transactions).

Association Rule Generation

After finding frequent itemsets with the FP-Growth algorithm, we created association rules using the lift metric. We looked at three main measurements for each rule:

* + Support: How often the items appear together in the dataset.
  + Confidence: How likely it is that the second product is bought when the first one is bought.
  + Lift: How strong the relationship is between the products compared to random chance.

For example, if we have the rule {Tea Cup} -> {Saucers}, the confidence might tell us that when “Tea Cup” is bought, “Saucers” are also bought 80% of the time.

**Insights & Conclusions**

* + Frequent Itemsets: The analysis showed that some products are bought together often, like “Plates” and “Tea Cups.” These combinations can be used to create product bundles.
  + Association Rules: We found several strong association rules that suggest which items are often bought together. These insights could help with marketing promotions and improving customer recommendations.
  + Data Quality: The dataset had some missing values and canceled orders, but after cleaning, it was ready for analysis. In the future, adding more details about the customers could provide deeper insights.