**Objective**

To identify the consumers, purchase patterns and insights

**Theory:**

Before implementing the project let’s first understand what is apriori algorithm

**How Does the Apriori Algorithm Work?**

Most store customers have purchased popcorn, milk, and cereal together. Therefore, {popcorn, milk, cereal} is a frequent itemset as it appears in a majority of purchases. So, if a person grabs popcorn and milk, they will also be recommended cereal.

According to the Apriori algorithm, a subset of the frequent itemset is also frequent. Since {popcorn, milk, cereal} is a frequent itemset, this means that {popcorn, milk}, {popcorn, cereal}, and {milk, cereal} are also frequent. Due to this, if a customer only goes for popcorn, they will be recommended both milk and cereal as well.

**What Are the Components of the Apriori Algorithm?**

The Apriori algorithm has three main components:

* Support
* Lift
* Confidence

You can think of these as metrics that evaluate the relevance and popularity of each item combination.

Let’s illustrate. The baskets below contain items purchased by four customers at a grocery store:

Here is a tabular representation of this purchase data:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Milk | Beer | Eggs | Bread | Bananas | Apples |
| Basket 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| Basket 2 | 1 | 0 | 0 | 1 | 0 | 0 |
| Basket 3 | 1 | 0 | 0 | 1 | 0 | 1 |
| Basket 4 | 0 | 0 | 0 | 1 | 1 | 1 |

**Let’s calculate the support, confidence, and lift.**

**Support**

The first component of the Apriori algorithm is support – we use it to assess the overall popularity of a given product with the following formula:

*Support(item) = Transactions comprising the item / Total transactions*

In the purchase data we’re working with, we have support(milk) = ¾ = **0.75**. This means that milk is present in 75% of all purchases.

Similarly, we have support(bread) = 4/4 = **1**. This means that bread is present in 100% of purchases.

A high support value indicates that the item is present in most purchases, therefore marketers should focus on it more.

**Confidence**

Confidence tells us the likelihood of different purchase combinations. We calculate that using the following formula:

*Confidence (Bread -> Milk) = Transactions comprising bread and milk / Transactions comprising bread*

In this case, it can show how many users who purchased bread also bought milk:

*Confidence (Bread -> Milk) = ¾ = 0.75*

This means that 75% of the customers who bought bread**also purchased milk**.

**Lift**

Finally, lift refers to the increase in the ratio of the sale of milk when you sell bread:

*Lift = Confidence (Bread -> Milk) / Support(Bread) = 0.75/1 = 1.3.*

This means that customers are **1.3 times more likely to buy milk** if you also sell bread.

**IMPLEMENTATION:**

**Loting data**

from matplotlib import pyplot as plt

df=pd.read\_excel("/kaggle/input/market-basket-analysis/Assignment-1\_Data.xlsx")

**DATA EXPLORATION**

* Thus, high-quality data will provide better prediction and accuracy. It’s important to improve the quality of raw data.
* At first, explore data and identify all extra spaces and rows that we didn’t need.
* Then we can Strip those extra spaces and remove the rows that we don’t need.
* Each country has different types of purchasing patterns, So we can segment the data with respect to the country to identify the purchasing patterns of the consumers in that country.

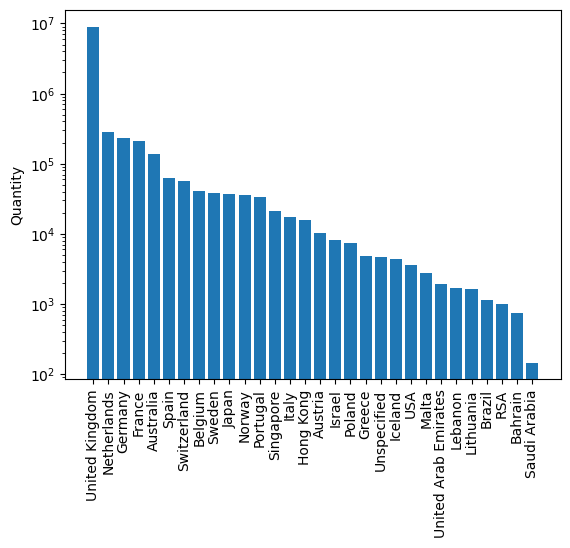
*#Dropping rows where ItemName isn't available*

df.dropna(subset=["Itemname"],inplace=True)

*#Dropping rows where Quantity <=0*

df = df[df["Quantity"]>0]

df.isnull().sum()

****

**PRUNING**

Reducing the number of item sets to only the important ones using pruning by the Apriori algorithm is crucial for several reasons:

It works by generating a large number of candidate item sets and then pruning those that do not meet minimum support thresholds. Pruning helps in reducing the computational complexity and runtime of the algorithm.

**1.   Efficiency:**

**2.   Scalability:**

**3.   Memory Efficiency:**

**4.   Identification of Important Patterns:**

**5.   Reducing Noise:** In many datasets, there may be a large number of infrequent or rare item sets that are not meaningful for analysis. Pruning eliminates such noise, ensuring that the patterns discovered are more relevant and actionable.

**6.   Improved Interpretability:** A smaller set of frequent item sets is more interpretable and manageable for further analysis and decision-making. It allows analysts and domain experts to focus their attention on a manageable subset of patterns rather than sifting through a vast number of irrelevant item sets.

**7.   Better Insights:** Pruning ensures that you are not overwhelmed by an abundance of patterns. Instead, you can concentrate on the most important item sets, gaining deeper insights into the relationships and associations between items in your data.

Reducing the number of item sets to only the important ones using pruning by the apriori algorithm.

**ASSOCIATION**

· Finding out the most rules b/w different items using association rules and apriori algorithm.

· Then we can try to visualize the trend for the current country using a scatterplot

· We can identify low support values would also mean that we would discard rules that are strong according to confidence.

**INSIGHTS**

By reducing the association rules, we can identify insights.

We can gather insights for other countries. By applying the above methods each one of them individually.