## Association Rule Mining – Using Apriori Algorithm

## Problem Statement

## You are a Data Scientist at a leading Retail firm. The company is looking to enhance its product specific strategies to increase sales. The Board of Management (BOM) has scheduled a meeting to get deeper insights into the sales pattern. You have been provided with a comprehensive dataset. Your task is to analyze the data to uncover patterns in customer purchasing behavior. Based on your findings, you are expected to suggest product bundling strategies and cross-promotional opportunities to increase sales. Your team has been asked to submit a detailed report on the strategies and presentation to the BOM.

## Introduction

Understanding customer purchasing behavior is crucial for businesses looking to optimize their sales strategies and improve customer satisfaction. This project focuses on analyzing customer transactions to identify patterns and relationships among various products. By examining a dataset of transactions, where each row represents a unique purchase and each column corresponds to different products with Boolean values indicating their purchase status, we can uncover valuable insights. The analysis aims to inform product bundling strategies and cross-promotional opportunities, enabling businesses to enhance their marketing efforts and ultimately drive sales growth.

## Objective

## The project seeks to analyze transaction data to uncover product relationships, develop bundling strategies, and provide actionable recommendations to optimize marketing and increase sales.

## Data Overview

The dataset consists of transactions involving various products. Each row corresponds to a transaction, and each column represents a product with **boolean values** (TrueFalse), indicating whether the product was purchased in that transaction. The columns include:

**Products:** Apple, Bread, Butter, Cheese, Corn, Dill, Eggs, Ice cream, Kidney Beans, Milk, Nutmeg, Onion, Sugar, Unicorn, Yogurt, chocolate

**Transactions (ID):** Each transaction has an ID.

**Why Association Rule Mining**:- Association rule mining is generally used for discovering relationships, not for prediction in the traditional accuracy sense.

Here, False refers to not bought and True refers to bought.

**Apple**

False    615

True     384

**Butter**

False    579

True     420

**Bread**

False    616

True     383

**Cheese**

False    595

True     404

**Corn**

False    592

True     407

**Dill**

False    601

True     398

**Eggs**

False    615

True     384

**Kidney Beans**

False    591

True     408

**Milk**

False    594

True     405

**Nutmeg**

False    598

True     401

**Onion**

False    596

True     403

**Sugar**

False    590

True     409

**Unicorn**

False    610

True     389

**Yogurt**

False    579

True     420

**Chocolate**

False    578

True     421

We have dropped the ID column as it has no use in the dataset. We had a total of 16 features in this dataset. Initially, the values in the dataset were either true or False. So, we have converted all the categorical values into numerical values using LabelEncoder. From data type Boolean to Integers.

**Methodology**

Association rule mining was used to discover relationships between products. The Apriori algorithm was applied to filter out frequent item sets with a minimum support of greater than 5%. The support value was varied to generate different numbers of association rules. The top 10 association rules based on confidence were selected for analysis.

We will create association rules using the data available in the dataset.

## Results

The results of the association rule mining are presented below:

At 0.15 minimum support value, 238 association rules were generated.

At 0.10 minimum support value, 436 association rules were generated.

At 0.05 minimum support value, 7086 association rules were generated.

At 0.04 minimum support value, 22630 association rules were generated.

At 0.03 minimum support value, 31902 association rules were generated.

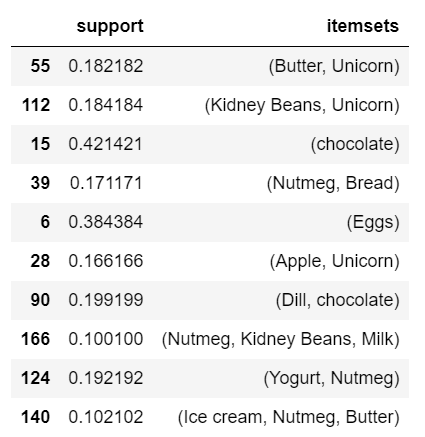
At 0.02 minimum support value, 119418 association rules were generated.

We observed that the number of association rules generated varies with different support values. As the minimum support value decreases, the number of association rules increases.

Higher support values (0.15, 0.10): These generate fewer rules, which might miss out on some potentially valuable insights. Our focus is on the most frequently occurring patterns, these higher support values may work well, though they could overlook less common but still important combinations.

Lower support values (0.05 to 0.02): These values generate a significantly larger number of rules. While this uncovers more insights, it may also lead to an overwhelming number of rules to analyze, many of which might be irrelevant or less interesting.

The apriori algorithm filters out frequent item sets which have minimum support of greater than 5%. From the table, we can infer that Nutmeg, Kidney Beans and Milk appear together in about 10.01% of the baskets. These item sets can be passed to association rules for generating rules and corresponding metrics.



## Top 10 association rules based on confidence sorted from highest to lowest at 0.02 & 0.05 support value

## 

## 

We can infer that the probability that a customer buys (Kidney Beans), given he has bought (Milk, butter, Bread, Corn and Nutmeg) is 0.8750. Now these rules can be used to create strategies like keeping the items together in store shelves or cross selling.

## Analysis and Insights

## The analysis of the association rules reveals several insights into customer purchasing behavior:

## The probability that a customer buys Kidney Beans, given they have bought Milk, Butter, Bread, Corn, and Nutmeg, is 0.8750.The probability that a customer buys Kidney Beans, given they have bought Cheese and Ice Cream, is 0.5561.

## These rules can be used to create strategies like keeping the items together in store shelves or cross-selling.

## Metrics:

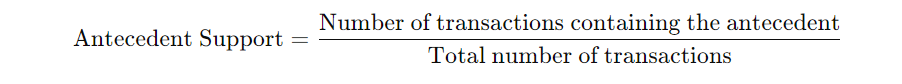
These rules can be used to create strategies like keeping the items together in store shelves or cross-selling.

**Antecedents** refer to the initial items or itemsets that appear in a transaction before certain other items or itemsets. They are the "if" part of an "if-then" association rule, representing the conditions that must be met for the rule to be considered.

**Consequent** refers to the item or itemset that appears as the result or outcome when certain conditions (antecedents) are met. It forms the "then" part of an "if-then" rule.

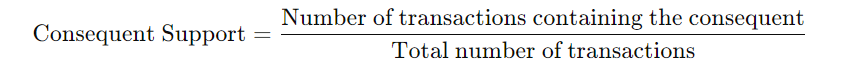
**Antecedent support** in association rule mining refers to the proportion of transactions in the dataset that contain the antecedent item or itemset. It measures how frequently the antecedent appears across all transactions, providing a sense of how common or rare the antecedent is.

It is calculated as:



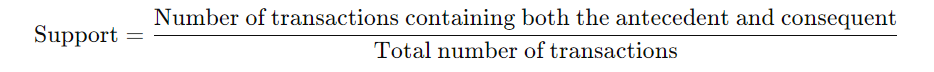
For example, As there are 999 transactions in total, and the antecedent (say, "Unicorn & Dill") appears in 168 of those transactions, the antecedent support would be 16.81% .

**Consequent** **support** in association rule mining refers to the proportion of transactions in the dataset that contain the consequent  item or itemset. It measures how frequently the consequent appears across all transactions, providing a sense of how common or rare the antecedent is.



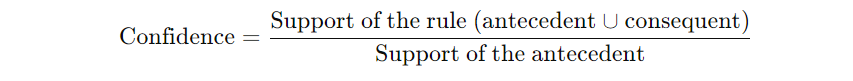
For example, As there are 999 transactions in total, and the consequent (say, "Chocolate") appears in 421 of those transactions, the antecedent support would be 42.14% .

**Support** is a measure of how frequently an itemset appears in the dataset. It represents the proportion of transactions in the dataset that contain a particular combination of items (both antecedent and consequent).



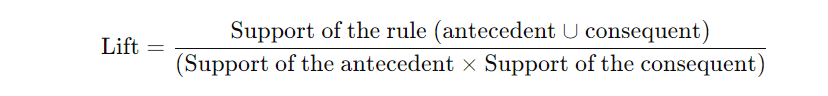
For example, As there are 999 transactions in total, and the consequent (say, "Chocolate") appears in 101 of those transactions, the antecedent support would be 10.11% .

**Confidence** measures the likelihood that the consequent occurs in transactions that contain the antecedent. It indicates the strength of the association between the antecedent and the consequent.



For example, As there are 168 transactions in total of antecedent, ("Unicorn & Dill") and support of the rule, ("Unicorn, Dill & Chocolate") appears in  of those transactions 101 times, the confidence would be 60.11 % .

**Lift** measures the strength of an association between the antecedent and the consequent by comparing the observed co-occurrence of the two to what would be expected if they were independent. It helps to understand whether the presence of the antecedent increases or decreases the likelihood of the consequent occurring.



For example, As the support of the rule(antecedent U consequent) is 10.11% and support values of the antecedent & consequent are 16.81% and 42.14%. Hence the Lift value is 1.4 times.

### Interpretation:

**Lift > 1**: The occurrence of the antecedent increases the likelihood of the consequent occurring, indicating a positive association.

**Lift = 1**: The antecedent and consequent are independent, meaning the antecedent has no effect on the likelihood of the consequent occurring.

**Lift < 1**: The occurrence of the antecedent decreases the likelihood of the consequent occurring, indicating a negative association.

**Leverage** measures the difference between the observed frequency of the antecedent and consequent appearing together and the frequency that would be expected if the antecedent and consequent were independent. It indicates how much more frequently the two items occur together than would be expected by chance.



As the support of the (antecedent U consequent) is 10.11% and support values of antecedent and consequent are 16.81% & 42.14%. Hence the leverage is 0.3.

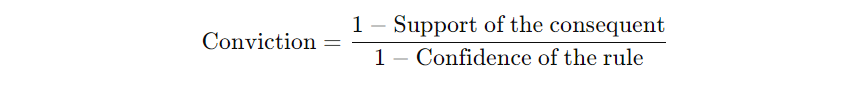
### Interpretation:

**Leverage > 0**: There is a positive association between the antecedent and the consequent, meaning they occur together more frequently than expected by chance.

**Leverage = 0**: There is no association between the antecedent and the consequent; they occur together as often as expected by chance.

**Leverage < 0**: There is a negative association, meaning they occur together less frequently than expected by chance.

**Conviction** measures the strength of an association rule by comparing the probability that the antecedent occurs without the consequent to the probability of the antecedent and consequent occurring together. It gives an indication of how strongly the presence of the antecedent suggests the absence of the consequent when the rule is not satisfied.



As we know from the above calculation the support value of the consequent is 42.14% and the confidence of the rule 60.11%. Hence the conviction value comes to 1.45.

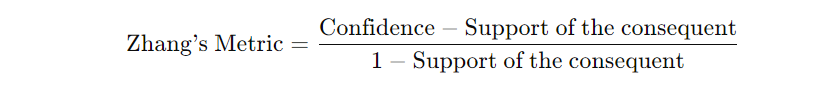
**Interpretation:**

**Conviction > 1**: Indicates a positive association, where the antecedent increases the likelihood of the consequent occurring. The higher the conviction, the stronger the association.

**Conviction = 1**: Suggests no association between the antecedent and the consequent, meaning that the presence of the antecedent does not affect the likelihood of the consequent.

**Conviction < 1**: This is unusual in practice, as it would imply a negative association, where the antecedent decreases the likelihood of the consequent occurring.

**Zhang's Metric** is a measure used in association rule mining to evaluate the interestingness of a rule. It combines elements of support, confidence, and the support of the consequent, aiming to provide a comprehensive view of how strong or significant a particular association rule is.



For example, As the confidence is 0.6011 and support of the consequent is 0.4214. The Zhang’s Metric for the above values will be 0.3 .

### Interpretation

**Zhang's Metric > 0**: This indicates that the rule is considered interesting; the confidence is greater than the support of the consequent, suggesting that the antecedent is indeed providing valuable information about the consequent.

**Zhang's Metric = 0**: The rule is neutral, meaning that the confidence of the rule is equal to the support of the consequent. This suggests that the antecedent does not provide additional predictive power.

**Zhang's Metric < 0**: This implies that the rule is not interesting; the confidence is less than the support of the consequent, indicating that the antecedent does not help in predicting the consequent.

### Conclusion

The analysis of the association rules reveals several insights into customer purchasing behavior, which can be used to inform product bundling strategies and cross-promotional opportunities. The results suggest that customers who buy Milk, Butter, Bread, Corn, and Nutmeg are likely to buy Kidney Beans, and customers who buy Cheese and Ice Cream are also likely to buy Kidney Beans. These insights can be used to increase sales by creating strategies like keeping the items together in store shelves or cross-selling. The metrics used to evaluate the strength of the association rules provide a comprehensive view of the relationships between products. Overall, the report provides valuable insights into customer purchasing behavior and suggests strategies to increase sales.

## Suggestions

This suggestions are based on the minimum support value of 0.05

### 1. Chocolate as a Cross-Promotional Product:

**Cross-Promotion**: The rules indicate a strong association between chocolate and items like **Dill**, **Milk**, and **Ice cream**. We can create promotions where customers get a discount on **chocolate** if they purchase any of these related items (e.g., "Buy Dill or Milk, get 20% off on chocolate").

**Bundling**: Since **Dill and chocolate** and **Milk and chocolate** appear frequently together, We could consider bundling them together as a package (e.g., "Chocolate and Milk combo pack").

### 2. Dill and Cheese Combinations:

**Cross-Promotion**: There's an association between **Dill** and **Cheese** with **Onions**. We could offer a combo deal where a customer purchasing **Cheese and Onions** gets a free or discounted **Dill**.

**Bundling**: Bundle **Dill** and **Cheese** together with **Onions**. This could be an ideal bundle for recipes or a snack pack.

### 3. Ice Cream with Kidney Beans or Butter:

**Cross-Promotion**: The association of **Ice cream with Kidney Beans** or **Butter** suggests an opportunity for bundling unusual yet popular combinations. Offer a deal where buying **Ice cream** gives a discount on **Kidney Beans** or **Butter**.

**Bundling**: We could create a "Gourmet Pack" including **Ice cream**, **Kidney Beans**, and **Butter**, especially appealing to experimental or gourmet shoppers.

### 4. Nutmeg and Milk Combination:

**Cross-Promotion**: The association of **Nutmeg** with **Milk** and **Kidney Beans** suggests potential for cross-promotional deals like, "Buy Nutmeg and get a discount on Milk."

**Bundling**: Consider bundling **Nutmeg**, **Milk**, and **Kidney Beans** in a package for people making specific dishes or beverages like holiday desserts or spiced drinks.

**5. Cheese and Ice Cream Combination:**

**Cross-Promotion**: The rule showing **Cheese** and **Ice cream** frequently occurring with **Kidney Beans** suggests the opportunity for a promotional offer on **Kidney Beans** when customers buy **Cheese and Ice Cream** together.

**Bundling**: We can bundle these three items together, possibly targeting customers interested in unique or gourmet pairings.

### 6. Butter and Unicorn (Uncommon Pairing):

**Cross-Promotion**: The pairing of **Butter** and **Unicorn** (perhaps an unconventional or niche product) suggests a possible market for unique or specialty combinations. Promote them as gourmet or luxury bundles, offering a discount or loyalty points.

**Bundling**: Bundle these with products like **Ice cream** to create luxury or niche gourmet packages.

### 7. Onion Promotions with Cheese and Dill:

**Cross-Promotion**: Since **Onions** are commonly bought with **Cheese** and **Dill**, consider offering an "add-on" discount where customers can get onions at a lower price if they buy **Cheese** or **Dill**.

**Bundling**: This can be marketed as a "cooking essentials" bundle that includes all three items.

### General Strategies:

**Loyalty Rewards**: Offer loyalty points or additional savings for customers who frequently buy bundled products like **Dill and Chocolate** or **Milk and Chocolate**.

**Holiday or Event-Based Bundles**: For example, bundle products like **Nutmeg, Milk, and Chocolate** during the holiday season for making festive drinks and desserts.

**Recipe-Specific Bundles**: Market bundles based on common recipes where multiple products are used together, like **Cheese, Onion, and Dill** for salads or **Butter, Ice cream, and Kidney Beans** for gourmet dishes.

**Other Suggestions:-**

**Targeted Email Campaigns and Recommendations:** Personalization is key in modern marketing. Utilize the data to send personalized emails that recommend products based on previous purchases. For example, after a customer buys *Cheese* and *Dill*, suggest adding *Onion* to their cart. This targeted approach can enhance customer satisfaction and drive additional sales.

**Optimizing Product Placement:** Strategic placement of items can significantly influence purchasing decisions. Position *Milk* near *Dill* and *Chocolate* in stores to encourage impulse buys. Additionally, grouping *Cheese*, *Onion*, and *Dill* together can create attractive displays that catch customers' eyes.

**Loyalty Program Enhancements:** Develop loyalty programs that reward customers for purchasing complementary items. For instance, offer points or discounts for customers buying combinations like *Ice Cream* and *Dill*, which can enhance customer retention and increase overall basket size.

**Content Marketing and Social Media Engagement:** Utilize social media platforms to share recipe ideas featuring combinations such as *Cheese, Dill, and Onion*. Collaborating with influencers can amplify reach and engagement while blogging about recipes can promote featured products organically.

**Seasonal Promotions:** Consider running limited-time promotions that encourage purchases of specific item combinations, particularly around holidays when certain products are more popular.