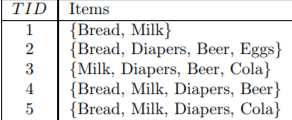
**What is Association Analysis?**

**Market basket transactions**: Customer purchase data collected daily at checkout counters of grocery stores.



**Association analysis** is useful for discovering interesting relationship hidden in large market basket transactions. The uncovered relationships can be represented in the form of association rules. {Diapers} 🡪 {Beer}

The above rule suggest that a strong relationship exist between sale of diapers and beer, many customers who buy diapers also buy beer.

{Diaper} → {Beer},  
{Milk, Bread} → {Eggs, Coke}  
{Beer, Bread} → {Milk}

**Implication means co-occurrence, not causality!**



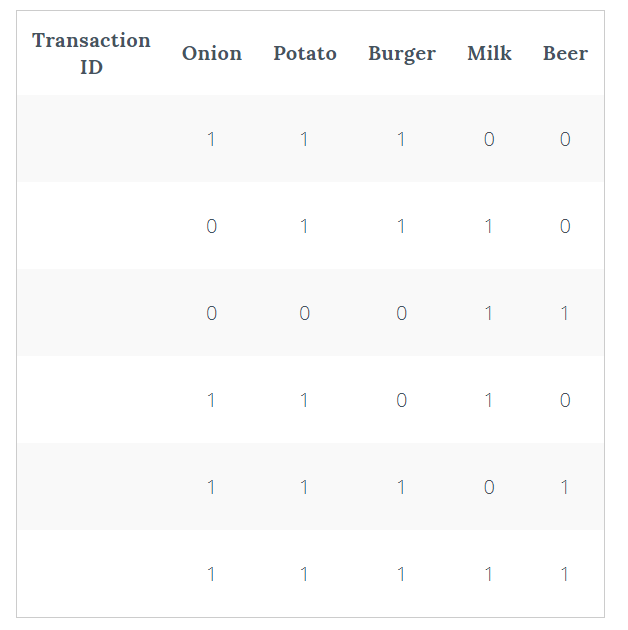
**Apriori Algorithm:**

**Apriori**: Apriori algorithm is a classical algorithm for data science. It is a bottom up approach. It is used for mining frequent itemset and relevant association rules. It is devised to operate on a database containing a lot of transactions, for instance, items brought by customers in a store.

Let a set of n attributes called items and set of transactions. Every transaction has a unique transaction IDs.

Let’s take an easy example from the supermarket sphere. The example that we are considering is quite small and in practical situations, datasets contain millions or billions of transactions. The set of item sets, = {Onion, Burger, Potato, Milk, Beer} and a database consisting of six transactions. Each transaction is a tuple of 0's and 1's where 0 represents the absence of an item and 1 the presence.

An example for a rule in this scenario would be {Onion, Potato} => {Burger}, which means that if onion and potato are bought, customers also buy a burger.



Sample Tuple List

**Basics:**

There are multiple rules possible even from a small database, so in order to select the interesting ones, we use

Constraint on various measures of interest and significance.

1. Support
2. Confidence
3. Lift

**1. Support**: The support of an itemset, is the proportion of transactions in the database in which the item X appears. (Popularity of an itemset).

{If the sales of a product (item) above a certain proportion have a meaningful effect on profits, that proportion can be considered as the support threshold. Furthermore, we can identify itemset that have support values beyond this threshold as significant itemset.}

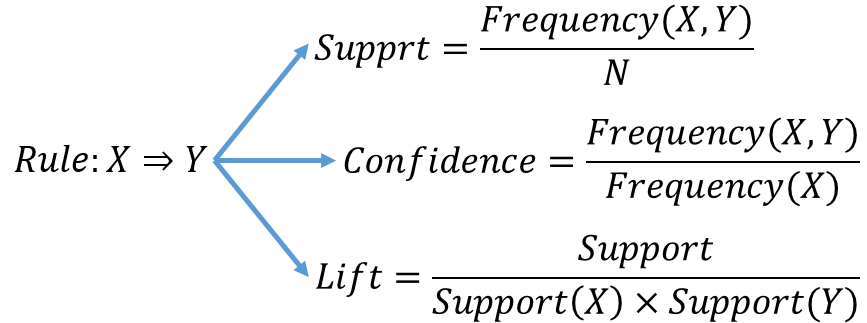
**2. Confidence:** It signifies the likelihood of item Y being purchased when item X is purchased. So, for the rule

{onion, potato} => {Burger}

{This implies that for 75% of the transactions containing onion and potatoes, the rule is correct. It can also be interpreted as the conditional probability, i.e., the probability of finding the itemset in transactions given the transaction already contains.}

**3**. **Lift**: It signifies the likelihood of the itemset being purchased when item is purchased while considering the popularity of itemset.

{If the value is greater than 1, it means that the itemset is likely to be bought with itemset. While value <1 implies that item is unlikely to be bought if itemset is bought}



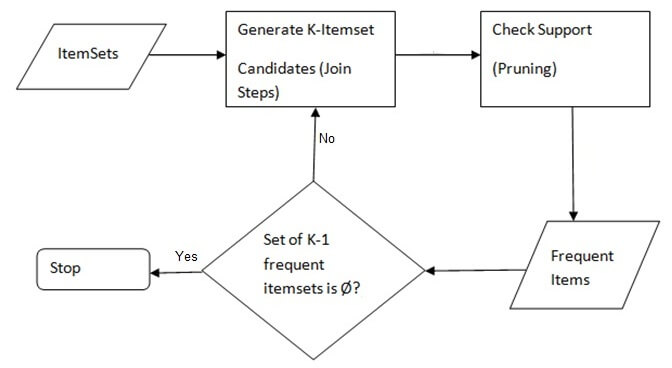
**General Process of the Apriori algorithm**

The entire algorithm can be divided into two steps:

**Step 1:** Apply minimum support to find all the frequent sets with k items in a database.

**Step 2:** Use the self-join rule to find the frequent sets with k+1 items with the help of frequent k-item sets. Repeat this process from k=1 to the point when we are unable to apply the self-join rule.

This approach of extending a frequent itemset one at a time is called the “bottom up” approach.

[](https://blog-c7ff.kxcdn.com/blog/wp-content/uploads/2017/03/Apriori-Algorithm.jpg)