

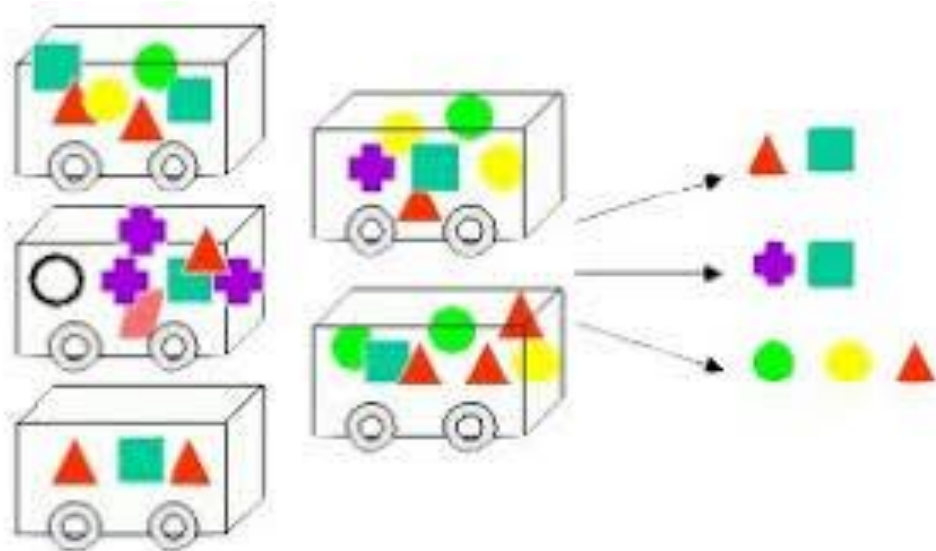
# Chapter#7

## Association Analysis



# What is Association Analysis?

Association analysis uses a set of transactions to discover rules that indicate the likely occurrence of an item based on the occurrences of other items in the transaction



# Association Rule

- An implication expression of the form  $X \Rightarrow Y$ , where  $X$  and  $Y$  are item sets
- Example:  $\{\text{Milk, Diaper}\} \Rightarrow \{\text{Beer}\}$
- Here  $X$  is  $\{\text{Milk, Diaper}\}$   $\Rightarrow Y$  which is  $\{\text{Beer}\}$



TID

Items

1

Chips, Milk

2

Chips, Diaper, Beer, Cornflakes

3

Milk, Diaper, Beer, Pepsi

4

Chips, Milk, Diaper, Beer

5

Chips, Milk, Diaper, pepsi



# Association Rule Evaluation Metrics

- Support (s) = Fraction of transactions that contain both X and Y i.e. how often Milk, Diaper and Beer occur together in the transactions. Milk, Diaper and Beer occur in 2 out of total 5 transactions, hence support =  $2/5=0.4$
- Confidence (c) = Measures how often each item in Y appears in transactions that contain X
- **$C = \text{Support}(X + Y) / \text{Support}(X)$**

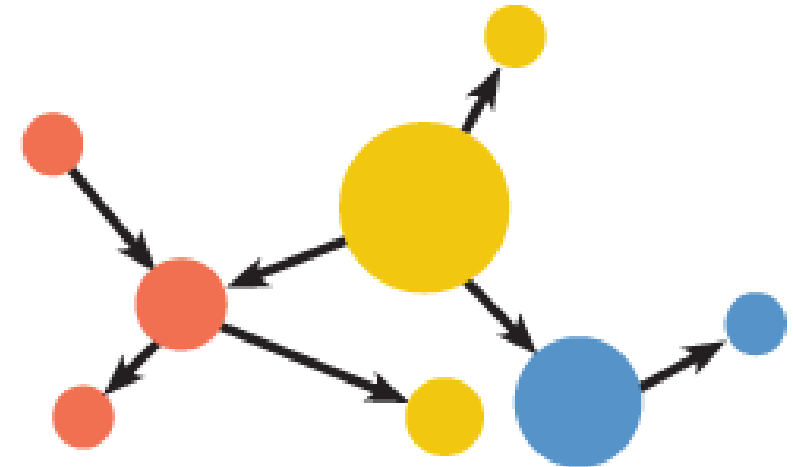


- That is- How often beer occurs in the transactions which contain milk and diaper. Now milk and diaper are together in 3 transactions (TID=3, 4 and 5), and out of the 3, beer is present in 2 of them, hence confidence =  $\frac{2}{3}$  (No. of transactions with Milk, Diaper and Beer/No. of transactions with Milk and Beer) = 0.67
- Lift: The Lift of the rule is  $X \Rightarrow Y$  is the confidence of the rule divided by the expected confidence, assuming that the item sets are independent.

# Interpretation of Lift:

- A lift value greater than 1 indicates that X and Y appear more often together than expected; this means that the occurrence of X has a positive effect on the occurrence of Y or that X is positively correlated with Y.
- A lift smaller than 1 indicates that X and Y appear less often together than expected, this means that the occurrence of X has a negative effect on the occurrence of Y or that X is negatively correlated with Y
- A lift value near 1 indicates that X and Y appear almost as often together as expected; this means that the occurrence of X has almost no effect on the occurrence of Y or that X and Y have Zero Correlation. **Thus, lift is a value between 0 and infinity**

- For all the values of lift which are  $> 1$ , actual lift = Lift value - 1 and
- % Increase in those cases =  $(\text{Lift value} - 1) * 100$
- Coming back to our Example  $\rightarrow$  Lift  $(X \rightarrow Y) = \text{confidence}(X \rightarrow Y) / \text{support}(Y)$
- $= \text{Support}(X+Y) / \text{Support}(X) * \text{Support}(Y)$
- $= 0.67 / (3/5) = 0.67 / 0.60 = 1.1167$



Now, Let us do a bit of Math here->  $((0.67-0.60)/0.60)*100=70/6=11.67$   
i.e. probability of finding beer in the transactions which have Milk and Diaper is greater than the normal probability of finding Beer in the above 5 transactions by 11.67%.

# How? Let's solve further

- Probability= Favorable Number of Cases/Total Sample Space
- Probability of finding beer in the above 5 transactions= $3/5=0.60$
- **Probability of finding beer in the transactions which have milk and diaper**
- Favorable Cases= Beer + Milk + Diaper
- Sample Space=Milk + Diaper
- =number of transactions which have Beer with Milk and Diaper/number of transactions which have
- Milk and Diaper= $2/3=0.67$ . Now 0.67 is 11.67% more than 0.60 i.e. there is a lift or increase of 11.67% of finding beer in the transactions which have Milk and Diaper

# To Summarize:

- **Support:** The support of the rule, that is, the relative frequency of transactions that contain X and Y.
- $\text{Support}(X \rightarrow Y) = \text{support}(X+Y)$
- **Confidence:** The confidence of the rule.  $\text{Confidence}(X \rightarrow Y) = \text{support}(X+Y) / \text{support}(X)$
- **Lift:** The following equation must hold true.  $\text{Lift}(X \rightarrow Y) = \text{confidence}(X \rightarrow Y) / \text{support}(Y)$
- $= \text{Support}(X+Y) / \text{Support}(X) * \text{Support}(Y)$
- **Support of the Rule  $X \Rightarrow Y$  is Symmetric i.e.  $\text{Support}(X \rightarrow Y) = \text{Support}(Y \rightarrow X)$**
- **Lift of the Rule  $X \rightarrow Y$  is Symmetric i.e.  $\text{Support}(X \rightarrow Y) = \text{Support}(Y \rightarrow X)$**

# Drawback of Confidence:

Confidence can sometimes be misleading as is shown in the below example

Credit Card				
Saving's Account		No	Yes	Total
	No	50	350	400
	Yes	100	500	600

Rule:  $X \Rightarrow Y$

$Support = \frac{freq(X,Y)}{N}$

$Confidence = \frac{freq(X,Y)}{freq(X)}$

$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$

- Rule:  $S \Rightarrow C$  (People with Savings Account are likely to have a credit card)
- The interpretation of implication ( $\Rightarrow$ ) in association rules can sometimes be misleading
- As in Above: Support ( $S \Rightarrow C$ ) =  $500/1000 = 50\%$
- Confidence ( $S \Rightarrow C$ ) =  $500/600 = 83\%$
- Expected Confidence ( $S \Rightarrow C$  ( $=350+500$ )/1000) =  $85\%$
- Lift ( $S \Rightarrow C$ ) =  $0.83/0.85 < 1$

- Based on the Support and Confidence, it might be considered a strong rule. However, people without a savings account are even more likely to have a credit card ( $=350/400=87.5\%$ ).
- Savings Account and Credit Card are in fact found to have a negative correlation. Thus, high confidence and support does not imply cause and effect, the two products at times might not even be correlated.
- One has to exercise caution in making any recommendations in such cases and look closely at the lift values.

- **Possible Recommendations for  $X \Rightarrow Y$  Rule (Where X and Y are 2 separate Products and have high support, high confidence and high positive lift > 1)**
- Put X and Y Closer in the Store
- Package X with Y
- Package X and Y with a poorly selling item
- Give Discount on only one of X and Y
- Increase the Price of X and lower the price of Y (or vice versa)
- Advertise only one of X and Y i.e. do not advertise X and Y together
- Example: If X was a toy and Y a form of sweet, then offering sweets in the form of toy X could also be a good option.

## Example: Interpretation of Rules for a sample product transaction set:

The thresholds used were 1.5 % support and 20% confidence.

Product1	==>	Product2	Support (%)	Confidence (%)	Lift
P	==>	Q	2.18	26.33	1.49
R	==>	Q	1.50	23.82	1.35
S	==>	Q	2.42	23.45	1.33
T	==>	U	1.79	21.06	1.23

# Interpretation of the first Rule:

- Products P and Q together appear in 2.18 % of the transactions as indicated by **Support**.
- If there are 100 transactions that contain Product P, then 26 of those also have Q as indicated by the **Confidence**.
- There is 49% more chance of occurrence of Q, given that P is also there as is indicated by the **Lift**.
- Or The Probability of finding Q in all those transactions which have Product P is 49% more than the Probability of finding Product Q in all the transactions

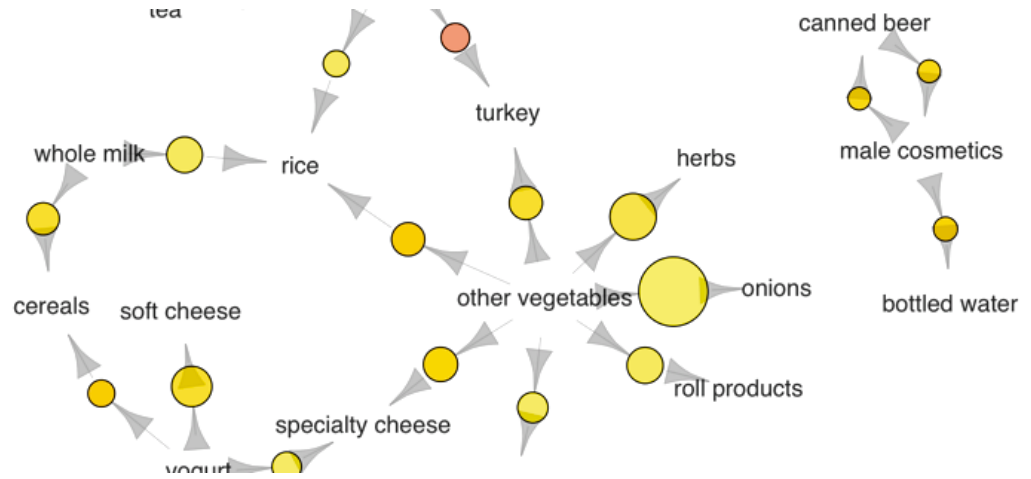
- **Mathematics behind the Rule (Ex B->C):**
- Lift= Support of (B + C)/ Support (B)\*Support (C) = approx 50%
- **The Way Lift has been calculated is as below:**
- Say for Example if total transactions are 100
- C is present in 25=> Probability of finding C in transactions= $25/100=1/4=0.25$
- B is present in 50, but C is present with B in 25 of them. So Probability of finding C in all the transactions with B is = B + C together/ B alone =  $25/50=0.50$
- It implies that Probability of Finding C in all the transactions with B is double the probability of finding C alone in all the transactions

# Example: Interpretation of Rules for a sample Product by region transaction set

- **Summary of association rules:** Min: support = 2.0%, confidence = 20.0%
- Max. Size of an Item Set = 10
- **Support:** Fraction of transactions that contain both X and Y. The threshold has been kept at 2% i.e. at least 2% of the transactions contain both X and Y.
- **Confidence (c):** Measures how often each item in Y appears in transactions that contain X. The threshold has been kept at 20%.

Item Set 1 ( X )	==>	Item Set 2 ( Y )	Support (%)	Confidence (%)	Lift
A1	==>	P1	3.61	88.91	19.41
A2	==>	P2	1.99	65.89	15.11

- Consider the top rule:
- Let  $X = A1$  (Region)
- Let  $Y = P1$  (Product)
- Why the values for Support are same? -> It is just a simple mathematical formula
- Support = Transactions that contain both X and Y / Total Transactions
- Since for both the rules X and Y are same, just that their orientation is different, obviously for both the rules it comes 3.61% i.e. 3.61% of the transactions contain both X & Y



# Interpretation of the Confidence Value:

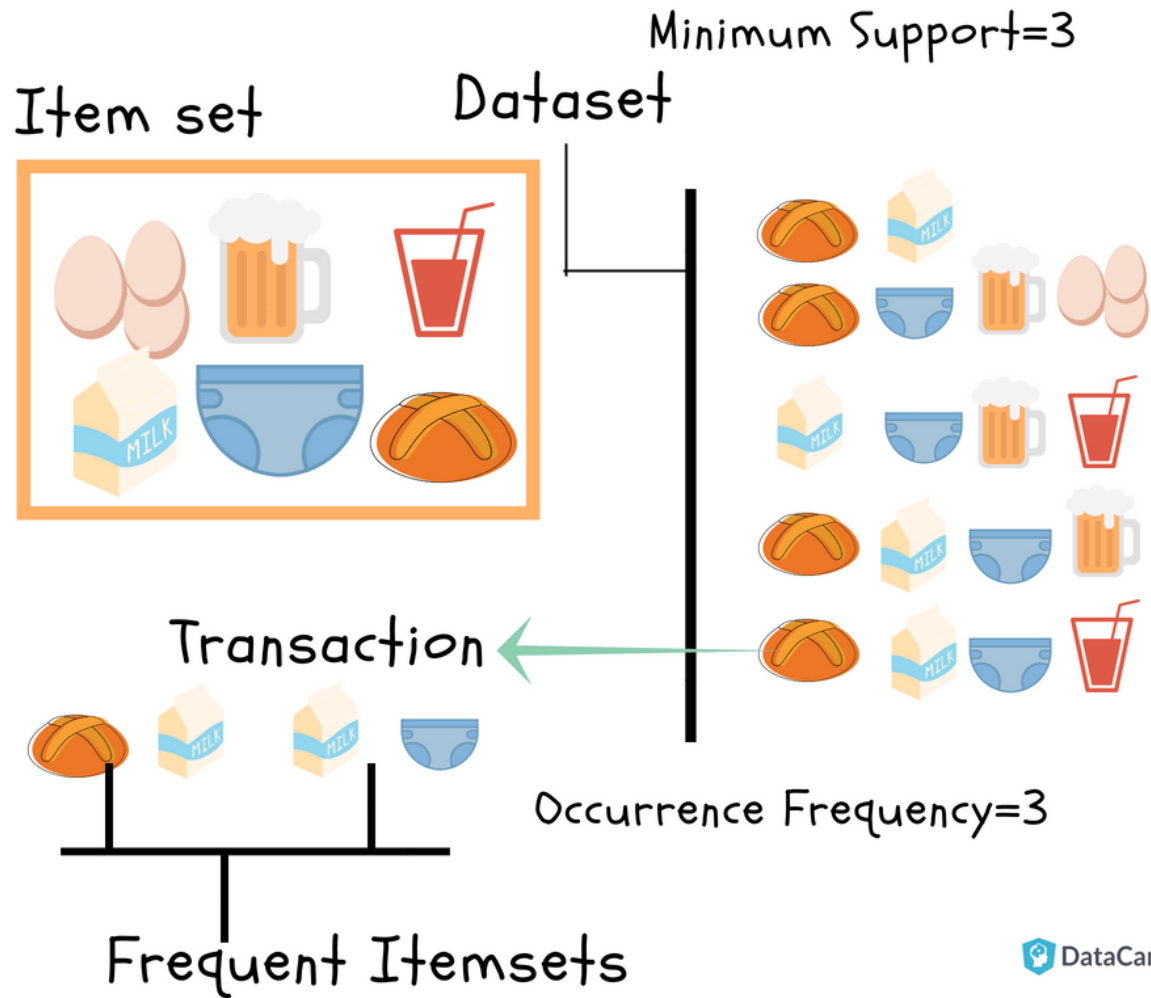
- $X \Rightarrow Y$ , confidence =  $(X \cup Y) / X$  i.e. Support  $(X+Y)$  / Support  $(X)$
- 88.91% of the times, Product P1 occurs in all those transactions which contain A1 as the region.
- Say for example there are 100 transactions which contain region- A1, among them 89 transactions contain the Product P1

# Interpretation of the Lift value:

- $\text{Lift}(X \rightarrow Y) = \frac{\text{confidence}(X \rightarrow Y)}{\text{support}(Y)} = \frac{\text{Support}(X+Y)}{\text{Support}(X) * \text{support}(Y)}$
- For this Rule  $\Rightarrow$  probability of finding Product1 increases 18.4 times in all those transactions where region is A1
- Or
- Probability of P1 in all those transactions which have region A1 is 18.4 times the Probability of Product P1 in all the transactions.

# Mathematics behind the Rule (Ex T->S):

- Say for Example if total transactions are 100
- S is present in 20=> Probability of finding S in transactions= $20/100=0.20$
- T is present in 50, but S is present with T in 20 of them. So Probability of finding S in all the transactions with T is = T + S together/ T alone =  $20/50=0.40$
- It implies that Probability of finding Product S in all the transactions with region T is double the probability of finding S alone in all the transactions



## Conclusion

Congratulations! You have learned APRIORI, one of the most frequently used algorithms in data mining. You have learned all about Association Rule Mining, its applications, and its applications in retailing called as **Market Basket Analysis**. You are also now capable of implementing Market Basket Analysis in R and presenting your association rules with some great plots! Happy learning!

## References:

1. <https://datascienceplus.com/a-gentle-introduction-on-market-basket-analysis%E2%80%8A-%E2%80%8Aassociation-rules/>
2. [https://en.wikipedia.org/wiki/Sparse\\_matrix](https://en.wikipedia.org/wiki/Sparse_matrix)
3. <https://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf>

If you would like to learn more about R, take DataCamp's [Importing and Managing Financial Data in R](#) course.



<https://www.datacamp.com/community/tutorials/market-basket-analysis-r>