ASSOCIATION RULES
MARKET BASKET ANALYSIS

DATA SCIENCE
INSTITUTE

Market Basket Analysis

DATA SCIENCE
INSTITUTE

Introduction to Market Basket Analysis

- Def: Market Basket Analysis (Association Analysis) is a mathematical modeling technique based upon the theory that if you buy a certain group of items, you are likely to buy another group of items.

- It is used to analyze the customer purchasing behavior and helps in increasing the sales and maintain inventory by focusing on the point of sale transaction data.

Using Market Basket Analysis

The analysis can be applied in various ways:

- Develop combo offers based on products sold together
- Organize and place associated products/categories nearby inside a store
- Determine the layout of the catalog of an ecommerce site
- Control inventory based on product demands and what products sell together

- Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns.

- It involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together.

- Amazon is a great example that leverages this analysis to cross-sell products.
- These are the products that come under the suggested item list which might interest you along with your current purchase.
- Your browsing history, what other customers have bought with a given product, and other factors determine which products appear in the suggested category.

MBA

- Basket-Set of item (whether you go to supermarket or buy online products)
- Every basket has transaction ID. Not at Customer level (Data has transaction IDs and the list of Items Purchased, and not the amount of transaction)
- Customers may have multiple transactions data and they are not really clubbed for customers. (Customer analytics requires data at customer level but MBA requires the data at Transaction level. So, Focus of MBA is on the items and not on the customers.)
- Data has Transaction ID and List of Items.

Contents

Rules…

- Support
- Confidence
- Lift

Support

Support simply emphasizes how popular an itemset is. Support, despite being simple, is an important metric in the Affinity Analysis that is used to determine the strength of association between items.

Take 5 transactions, for instance. If you purchase bread in 3 transactions, you can tell the support of bread is equal to 3/5.

$$Support = \frac{freq(i_1, i_2)}{N}$$

Confidence

While the support emphasizes how popular an itemset is, confidence denotes the likelihood of certain items are purchased together.

For instance, how likely butter is purchased when item bread is purchased.

Confidence is typically notated as ***Bread ⇒ Butter*** (Proportion of transactions containing Bread that also contain Butter.)

$$P(butter \mid bread) = \frac{Support(Bread, Butter)}{P(Bread)}$$

Definition of confidence

Confidence, as you can see above, is a probability and so its range is [0,1]. If the confidence of ***Bread ⇒ Butter*** is equal to 1, we can say every time a customer purchases bread, also purchases butter.

DATA SCIENCE INSTITUTE

Lift

Lift

Like confidence, the lift is notated as *Bread ⇒ Butter*. It says how likely Butter is purchased when Bread is purchased while controlling for how popular Butter is.

$$Lift(Bread \Rightarrow Butter) = \frac{Support(Bread, Butter)}{Support(Bread) * Support(Butter)} = \frac{P(Bread, Butter)}{P(Bread) * P(Butter)}$$

Definition of Lift

Lift's range is [0, +∞]. When lift equal to one, bread and butter are independent and, thus, no inferences can be made about butter when the bread is purchased. However, when the lift is greater than 1, it means that the butter is likely to be purchased together with bread.

About Association Rules

Association Rule Learning

Method for discovering interesting relations between variables in large databases

- Based on the **concept of strong rules**, Rakesh Agrawal introduced association rules concept.
- **Association rules are a popular technique in data mining and machine learning used to find interesting relationships (associations) between variables in large datasets. These rules are often used in market basket analysis to identify products that frequently co-occur in transactions.**

DATA SCIENCE INSTITUTE

Introduction to Market Basket Analysis

- The most widely used area of application for association rules is **Market Basket Analysis**

> Market Basket Analysis (Association Analysis) is a **mathematical modeling technique based upon the theory that if you buy a certain group of items, you are likely to buy another group of items**

- It is used to **analyze the customer purchasing behavior and helps in increasing the sales and maintain inventory** by focusing on the point of sale transaction data

Market Basket Analysis – Uses

Product Building	• Develop combo offers based on products bought together
Optimisation	• Organise and place associated products/categories nearby inside a store
Advertising and Marketing	• Determine the layout of the catalog of an ecommerce site
Inventory Management	• Control inventory based on product demands and what products sell together

Definitions and Terminology

Term	Definition
Transactions	A set of items (Item set)
Support	Ratio of **number of times two or more items occur together** to **the total number of transactions** Support can be thought of as P(A and B)
Confidence	Conditional probability that **a randomly selected transaction will include Item B given Item A** **P(B|A)** (written as A => B)
Lift	Ratio of the **probability of Items A and B occurring together (Joint probability)** to **the product of P(A) and P(B)**

Get an Edge!

The Famous Story

An article in The Financial Times of London (Feb. 7, 1996) stated,

"The example of what data mining can achieve is the case of a large US supermarket chain which discovered a strong association for many customers between a brand of babies nappies (diapers) and a brand of beer. Most customers who bought the nappies also bought the beer. The best hypothesisers in the world would find it difficult to propose this combination but data mining showed it existed, and the retail outlet was able to exploit it by moving the products closer together on the shelves."

DATA SCIENCE INSTITUTE

Rule Evaluation – Support

Transaction No.	Item 1	Item 2	Item 3	…
100	**Beer**	**Diaper**	Chocolate	
101	Milk	Chocolate	Shampoo	
102	Beer	Wine	Vodka	
103	**Beer**	Cheese	**Diaper**	
104	Ice Cream	**Diaper**	**Beer**	

A ⬇ B ⬇

Support of {Diaper, Beer}

$$\text{Support} = \frac{\text{No.of transactions containing both A and B}}{\text{Total no.of transactions}} = \frac{3}{5} = 60\%$$

Support of {Diaper, Beer} is 3/5

DATA SCIENCE INSTITUTE

Rule Evaluation – Confidence

Transaction No.	Item 1	Item 2	Item 3	…
100	**Beer**	**Diaper**	Chocolate	
101	Milk	Chocolate	Shampoo	
102	**Beer**	Wine	Vodka	
103	**Beer**	Cheese	**Diaper**	
104	Ice Cream	**Diaper**	**Beer**	

$$\text{Confidence for } \{A\} \Rightarrow \{B\} = \frac{\text{No. of transactions containing both A and B}}{\text{No. of transactions containing A}}$$

Confidence for $\{Diaper\} \Rightarrow \{Beer\}$ is 3/3

When Diaper is purchased, the likelihood of Beer purchase is 100%

Confidence for $\{Beer\} \Rightarrow \{Diaper\}$ is 3/4

When Beer is purchased, the likelihood of Diaper purchase is 75%

$\{Diaper\} \Rightarrow \{Beer\}$ is a more important rule according to Confidence

Rule Evaluation – Lift

Transaction No.	Item 1	Item 2	Item 3	Item 4
100	Beer	Diaper	**Chocolate**	
101	**Milk**	**Chocolate**	Shampoo	
102	Beer	**Milk**	Vodka	**Chocolate**
103	Beer	**Milk**	Diaper	**Chocolate**
104	**Milk**	Diaper	Beer	

A B

Consider $\{\text{Chocolate}\} \Rightarrow \{\text{Milk}\}$

$$\text{Lift} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{3/5}{\left(4/5\right)\left(4/5\right)} = 0.9375$$

Lift < 1 indicates Chocolate is decreasing the chance of Milk purchase
Support and confidence are high but lift is low

DATA SCIENCE INSTITUTE

Case Study – Groceries Purchase Data

Background

- A typical grocery outlet records point-of-sale transaction data

Objective

- To mine association rules and information about item sets

Available Information

- **Total number of transactions is 9835**
- **Items are aggregated to 169 categories**
- **Data is collected for 1 month (30-days)**

Data Snapshot

Groceries

```
     items
[1] {citrus fruit,
     semi-finished bread,
     margarine,
     ready soups}
[2] {tropical fruit,
     yogurt,
     coffee}
[3] {whole milk}
[4] {pip fruit,
     yogurt,
     cream cheese ,
     meat spreads}
[5] {other vegetables,
```

Columns	Description	Possible values
id	Transaction Id	Positive Integers
items	Set of Items purchased in a transaction	Subset from 169 categories of items

DATA SCIENCE INSTITUTE

Market Basket Analysis in R

```
#Market Basket Analysis Using Apriori Recommendation

install.packages("arules")
library(arules)

install.packages("arulesViz")
library(arulesViz)



data("Groceries")
```

- We will be using two packages for performing Market Basket Analysis in R.
- Package **"arules"** stands for 'Association Rules' and it contains functions for mining association rules and frequent itemsets.
- Package **"arulesViz"** is used for visualisation.
- Install and load these two packages.

- Load the dataset.
- The **Groceries** data set is provided for package **arules** by Michael Hahsler, Kurt Hornik and Thomas Reutterer.*
- The data is of class 'transaction' supported by package **arules**.

DATA SCIENCE INSTITUTE

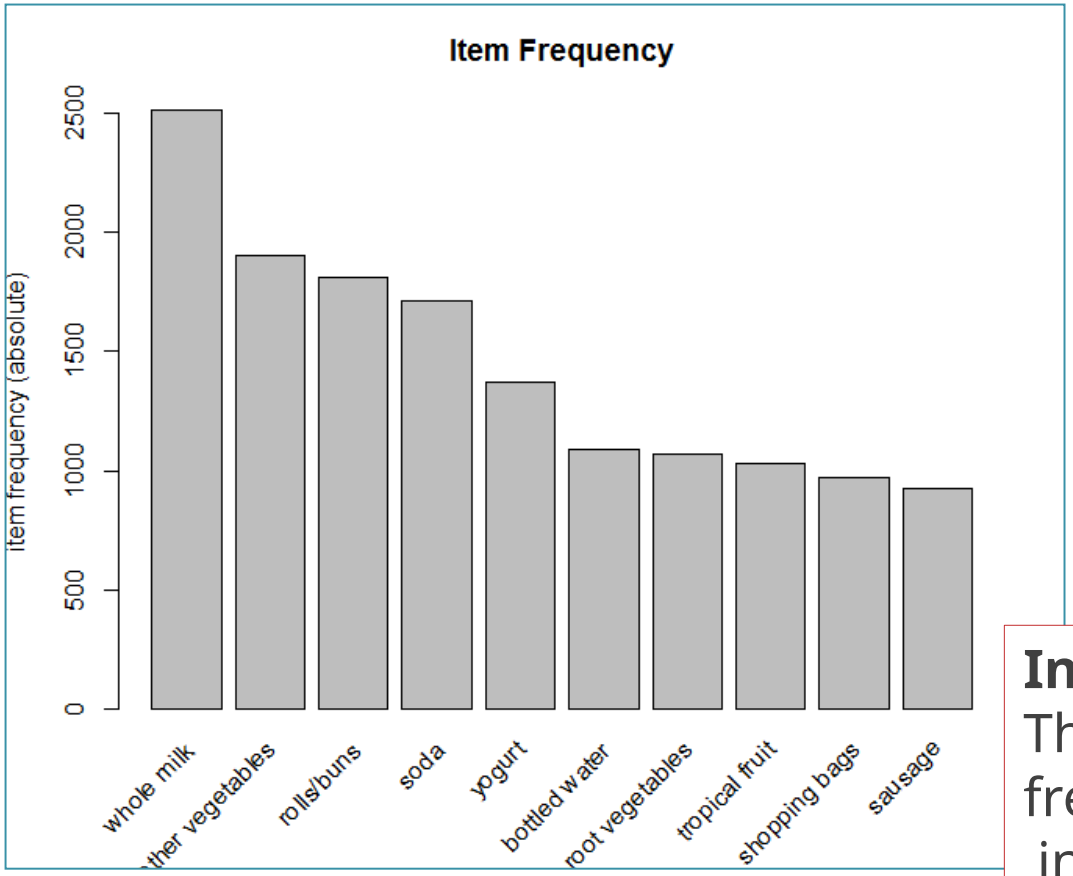
Visualise Item Frequency

```
#Item Frequency Plot
itemFrequencyPlot(Groceries,topN=10,type="absolute",
                  main="Item Frequency")
```

- **itemFrequencyPlot()** calculates item frequency and returns a barplot.
- **topN=** instructs R to plot only top N highest item frequency or lift (Logical, if **lift=TRUE**). It plots values in decreasing order.
- **type=** is a character string indicating whether item frequencies should be displayed relative or absolute. Default is relative.

Item Frequency Plot

```
# Output
```

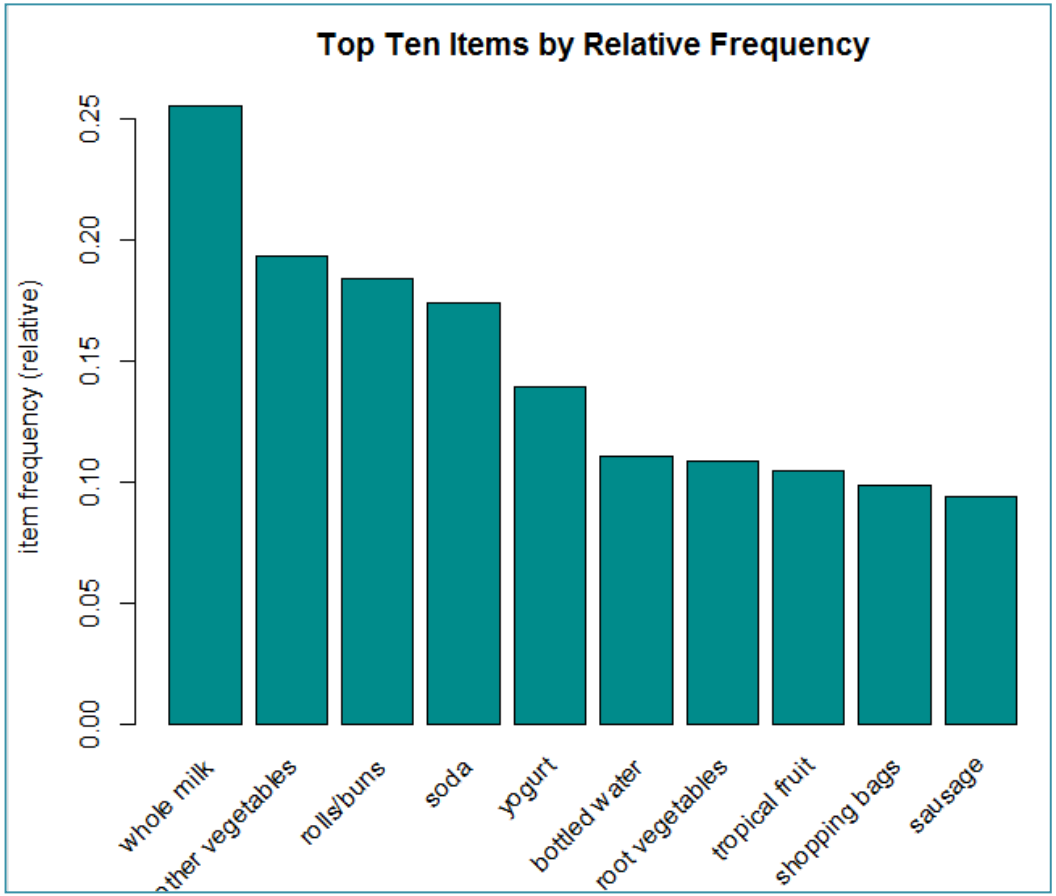


Interpretation:
The plot shows items by frequency
in a descending order

Item Frequency Plot

```
itemFrequencyPlot(Groceries,topN=10,type="relative",
col="darkcyan",main="Top Ten Items by Relative Frequency")
```

```
# Output
```



- **type= "relative"** displays barplot with the relative frequency
- **col=** specifies the colour of the bars

Interpretation:

- The plot shows items by relative frequency in a descending order.

DATA SCIENCE INSTITUTE

Get and Display the Rules

```
#Get the Rules
rules<-apriori(Groceries,parameter=list(supp=0.001,conf=0.8))
```

- The Apriori algorithm employs level-wise search for frequent itemsets.
- **apriori()** is used to mine frequent itemsets, association rules or association hyperedges using this algorithm.
- The default is to mine rules with **support 0.1, confidence 0.8**.
- **Here, we have used threshold of 0.001 for support.**
- **apriori()** returns an object of class rules or itemsets.

```
#Show Top 5 Rules But Only 2 Digits
options(digits=2)

inspect(rules[1:5])
```

global options which affect the way in which R computes and displays results. We have set **digits=2** to display results

inspect in package **arules** displays association and plus additional information formatted for online inspection.

Get and Display the Rules

```
# Output of Rules
```

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.8    0.1    1 none FALSE            TRUE       5   0.001      1
 maxlen target   ext
     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 9

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.01s].
sorting and recoding items ... [157 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.03s].
writing ... [410 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Interpretation:

- The output displays parameter specification, algorithmic control and absolute minimum support count.
- It also lists down tasks performed and time taken to complete them.
- We are interested in knowing how many rules were created; 410 in our case.

Get and Display the Rules

```
# Output of inspect
```

	lhs		rhs	support	confidence	lift	count
[1]	{liquor,red/blush wine}	=>	{bottled beer}	0.0019	0.90	11.2	19
[2]	{curd,cereals}	=>	{whole milk}	0.0010	0.91	3.6	10
[3]	{yogurt,cereals}	=>	{whole milk}	0.0017	0.81	3.2	17
[4]	{butter,jam}	=>	{whole milk}	0.0010	0.83	3.3	10
[5]	{soups,bottled beer}	=>	{whole milk}	0.0011	0.92	3.6	11

Interpretation:

- **inspect()** returns list of lhs and rhs items, their support, confidence and lift values.

Manage How the Rules are Displayed

```
#Sort the Rules
rules<-sort(rules,by="lift",decreasing=TRUE)
```

- sort() from package arules is used
- by="lift" indicates sort by values of Lift
- decreasing= logical, specifies the direction of sorting. Default is decreasing=TRUE.

```
#Show Top 5 Rules (Sorted)

options(digits=2)

inspect(rules[1:5])
```

DATA SCIENCE INSTITUTE

Top Five Rules (Sorted)

```
# Output
```

	lhs		rhs	support	confidence	lift	count
[1]	{liquor, red/blush wine}	=>	{bottled beer}	0.0019	0.90	11.2	19
[2]	{citrus fruit, other vegetables, soda, fruit/vegetable juice}	=>	{root vegetables}	0.0010	0.91	8.3	10
[3]	{tropical fruit, other vegetables, whole milk, yogurt, oil}	=>	{root vegetables}	0.0010	0.91	8.3	10
[4]	{citrus fruit, grapes, fruit/vegetable juice}	=>	{tropical fruit}	0.0011	0.85	8.1	11
[5]	{other vegetables, whole milk, yogurt, rice}	=>	{root vegetables}	0.0013	0.87	8.0	13

Interpretation:

- The rules are now sorted based on lift. Sorting ensures that most relevant rules appear first.

Targeting Items

Association rules should be used to explain decision making and further utilised to form effective strategies.

Continuing with the example of consumer's buying preferences, the following two questions can be of interest. **Reference item is Whole Milk.**

What are customers likely to buy if they purchase whole milk?

Targeting Items

```
library(arules)
data("Groceries")

rules<-apriori(Groceries,parameter=list(supp=0.001,conf=
0.15,minlen=2),appearance=list(default="rhs",lhs="whole
milk"),control=list(verbose=FALSE))
```

- We have already seen the basic arguments used in **apriori().**
- **minlen=** in **parameter=** is used to specify how many items to be considered

  Default is **minlen=1** which means that rules with only one item will be created.
- **appearance=** is used to restrict item appearance.
- **control=** controls the algorithmic performance of mining algorithm.
- **verbose=FALSE** ensures R does not show progress report.

```
rules<-s

inspect(
```

DATA SCIENCE INSTITUTE

* Here we continue with previous ppt & use the same inbuilt dataset “Groceries”

supp=0.001: Minimum support threshold. Rules with a support less than 0.1% are excluded.

conf=0.15: Minimum confidence threshold. Rules with a confidence less than 15% are excluded.

minlen=2: Minimum length of the rules. This ensures that rules contain at least 2 items.

Appearance:

This specifies constraints on the items appearing in the rules.

default="rhs": By default, items are allowed only in the right-hand side (consequent) of the rules.

lhs="whole milk": Specifies that "whole milk" must appear in the left-hand side (antecedent) of the rules.

verbose=FALSE: Suppresses the printing of progress and summary messages during the rule mining process.

Targeting Items

```
# Output
```

	lhs		rhs	support	confidence	lift	count
[1]	{whole milk}	=>	{other vegetables}	0.075	0.29	1.5	736
[2]	{whole milk}	=>	{rolls/buns}	0.057	0.22	1.2	557
[3]	{whole milk}	=>	{yogurt}	0.056	0.22	1.6	551
[4]	{whole milk}	=>	{root vegetables}	0.049	0.19	1.8	481
[5]	{whole milk}	=>	{tropical fruit}	0.042	0.17	1.6	416

Interpretation:

- Based on confidence, customers are most likely to move to other vegetables immediately after buying whole milk.

DATA SCIENCE INSTITUTE

Visualise Rules

```
#Creating Interactive Graph
library(arulesViz)
rules<-
apriori(Groceries,parameter=list(supp=0.001,conf=0.15,minlen=2),
appearance=list(default="rhs",lhs="whole
milk"),control=list(verbose=FALSE))

plot(rules,method="graph",interactive=TRUE,shading=NA)
```

- **plot()** is the default function to visualise association rules and itemsets in package **arulesViz**. If no argument is specified other than the object, then the function returns a simple scatter plot.
- The package offers different plot styles. **method=** tells R which style to use for visualisation. Here, we are using "**graph**"
- **interactive=** is a logical, specifying whether to return plot in interactive mode.
- **shading=** is used to enhance the interpretability of the graph, by using colour

Interactive Graphs

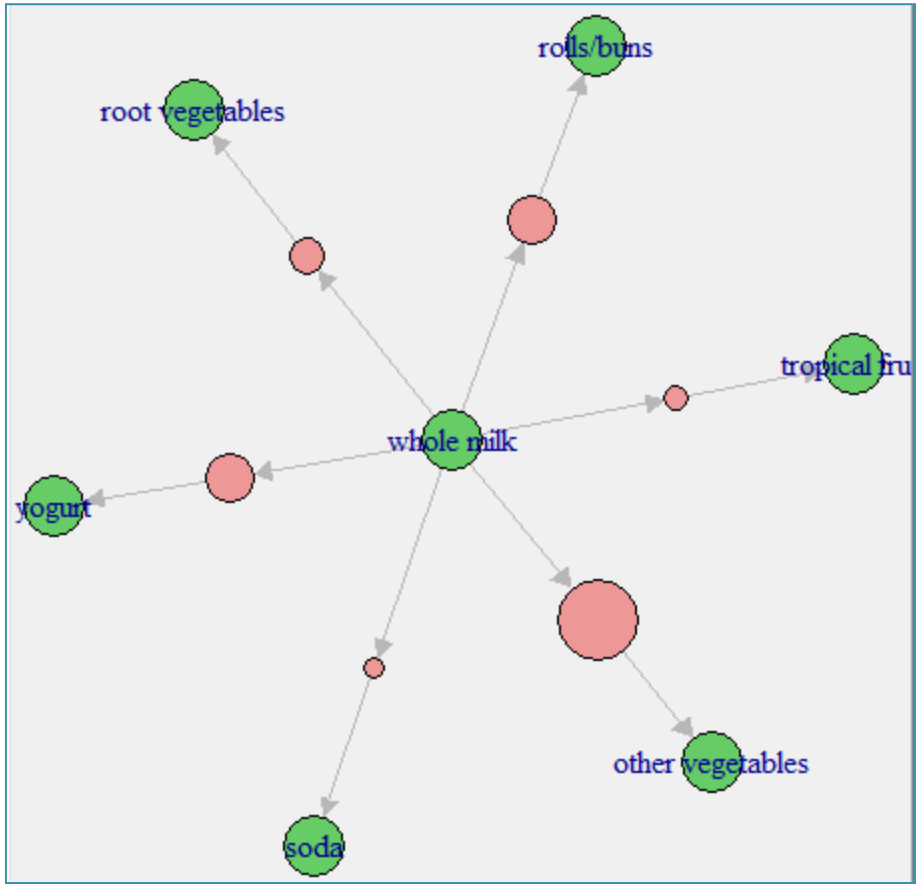
Upon running the command for interactive graph in **`arulesViz`**, a new window opens

Graph uses vertices and edges to visualise rules

- Vertices are itemsets or items
- Edges indicate relationship in rules
- Labels on the edges or colour or width of the arrows displaying edges represent interest measures

Note that graph-based visualisations are viable only for a small set of rules, as more number of rules would make the graph cluttered and difficult to interpret

Interpreting the Interactive Graph



Out target item is whole milk, which is at the centre of the graph

- The coloured vertex represent confidence. As seen before, confidence for {whole milk, other vegetables} is the highest, followed by yogurt and rolls/buns.
- Lowest confidence in the top five is for {whole milk, tropical fruit}

Using MBA for Recommendations

- Support can be used for initial recommendations or to determine the layout of the catalog of an ecommerce site
- Confidence can be used to provide recommendations based on first product purchase.
- Use rules only if lift is greater than one.

THANK YOU!!

DATA SCIENCE
INSTITUTE