ASSOCIATION RULES
MARKET BASKET ANALYSIS

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

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)**

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

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

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.

Case Study

Background

- Transactions data collected from point of sales is generally in long format. Arules package requires the data to be in wide format.

Objective

- To convert available data to a format suitable for association analysis and conduct analysis via arules package in R

Available Information

- Each transaction is given a unique ID
- Items basket contains five items, items purchased during each transaction are recorded

Data Snapshot

Transactions Data for MBA

id	item
1	B
1	C
1	D
1	E
2	A
2	B
2	C
2	D
2	E
3	A
3	B

Columns	Description	**Measurement**	Possible values
id	Transaction Id	-	Positive Integers
item	Items purchased	A,B,C,D,E	5

Data Conversion

```
#Convert the Data
library(arules)
trans<-read.transactions("Transactions Data for
MBA.csv",format="single",sep=",",cols=c("id","item"),header=TRUE)
```

- **read.transactions()** in package arules reads a transactions data file and creates a transaction object.
- **format=** indicates the format of the dataset. **"single**" implies each line corresponds to a single item, containing at least ids for the transaction and the item. **"basket"** implies each line in the transaction data file represents a transaction where the items (item labels) are separated by the characters specified by **sep**.
- For the **"single"** format, **cols=** is a numeric or character vector of length two giving the numbers or names of the columns (fields) with the transaction and item ids, respectively.

Data Conversion

```
#The converted data looks as follows:
inspect(trans)
```

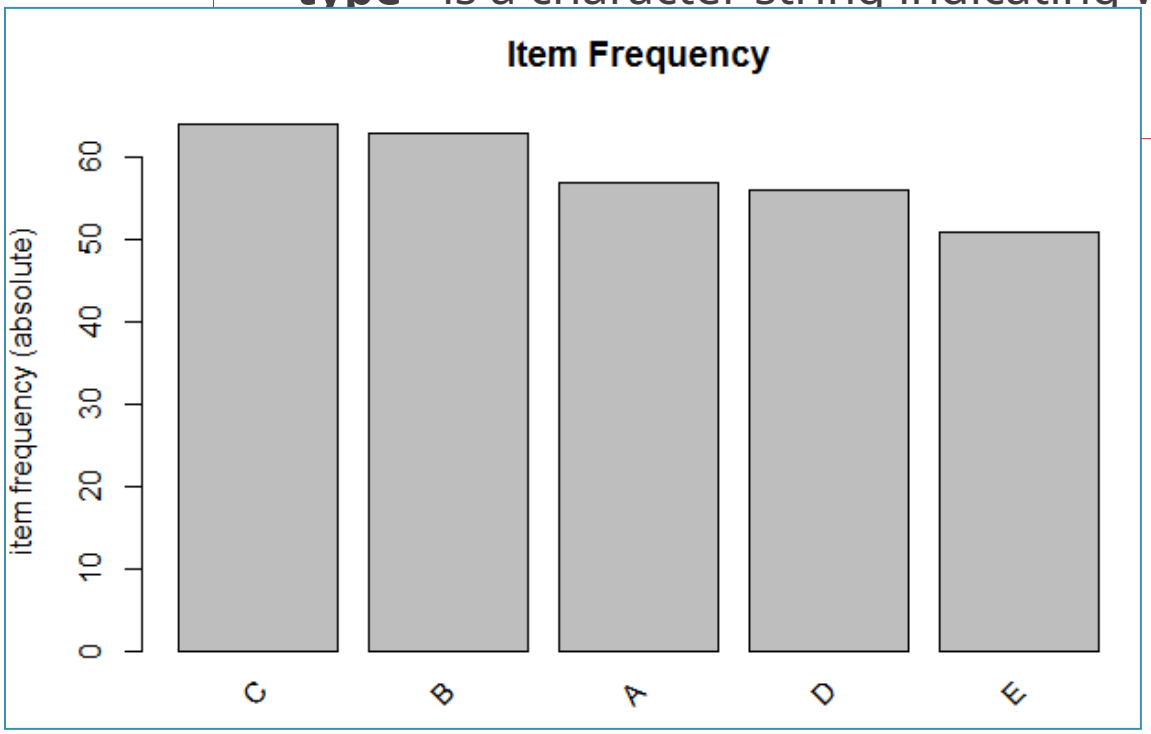
```
      items              transactionID
[1]   {B, C, D, E}       1
[2]   {A, C, D}          10
[3]   {A, B, E}          100
[4]   {B, D, E}          11
[5]   {B, D, E}          12
[6]   {A, B, C, D, E}    13
[7]   {A, C, D, E}       14
[8]   {A, C}             15
[9]   {A, B, D}          16
[10]  {B, C, E}          17
[11]  {A, B, C, D}       18
[12]  {B, C}             19
[13]  {A, B, C, D, E}    2
[14]  {A, B, C, D}       20
[15]  {A, C, D}          21
[16]  {A, E}             22
[17]  {A, C, D}          23
[18]  {A, E}             24
[19]  {A, C, E}          25
```

Association Analysis

```
#Visualise Frequency
itemFrequencyPlot(trans,topN=5,type="absolute")
```

- **itemFrequencyPlot()** calculates item frequency and returns a barplot.
- **topN=** instructs R to plot only top N highest item frequency or lift
- **type=** is a character string indicating whether item

```
# Output
```



Interpretation:

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

Association Analysis

```
#Get the Rules
rules<-apriori(trans,parameter=list(supp=0.001,conf=0.8))
inspect(rules[1:5])
```

- **apriori()** is used to mine frequent itemsets, association rules or association hyperedges using this algorithm with specified support and confidence
- **inspect()** in package **arules** displays association and additional information formatted for online inspection

Association Analysis

```
# Output
```

```
Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support
        0.8    0.1    1 none FALSE            TRUE       5   0.001
 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: 0

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[5 item(s), 100 transaction(s)] done [0.00s].
sorting and recoding items ... [5 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 done [0.00s].
writing ... [5 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 are created; Here 5 rules are

Association Analysis

```
# Output
```

	lhs		rhs	support	confidence	lift	count
[1]	{A,D}	=>	{C}	0.25	0.81	1.3	25
[2]	{A,D,E}	=>	{C}	0.11	0.85	1.3	11
[3]	{A,B,E}	=>	{C}	0.10	0.83	1.3	10
[4]	{A,B,D}	=>	{C}	0.16	0.84	1.3	16
[5]	{A,B,D,E}	=>	{C}	0.06	1.00	1.6	6

Interpretation:

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

- Questions: 1) Find rules any with confidence=1
  2) Find top 3 rules based on highest Lift

THANK YOU!!