Market Basket Analysis

Introduction to Market Basket Analysis and a SAS Implementation of the Apriori Algorithm

Bill Qualls

DePaul University, Spring 2013

ECT584 – Web Data Mining for Business Intelligence

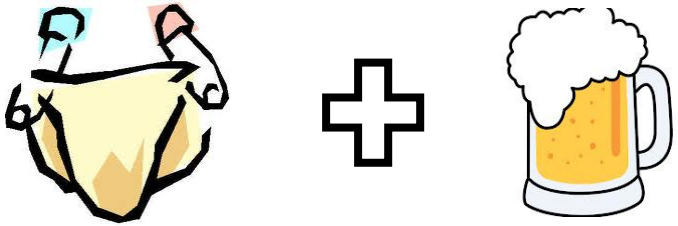
Professor Jonathan Gemmell

Outline

- Introduction
- Sample market basket
- Support
- Pairs
- Iterations
- Association Rules
- Confidence
- Lift
- SAS macro

Introduction

"An apocryphal early illustrative example for this was when one super market chain discovered in its analysis that customers that bought diapers often bought beer as well, have put the diapers close to beer coolers, and their sales increased dramatically. Although this urban legend is only an example that professors use to illustrate the concept to students, the explanation of this imaginary phenomenon might be that fathers that are sent out to buy diapers often buy a beer as well, as a reward." (Retrieved May 5, 2013 from http://en.wikipedia.org/wiki/Market_basket)



Items Sold

bananas
bologna
bread
buns
butter
cereal
cheese
chips
eggs
hotdogs
mayo
milk
mustard
oranges
pickles
soda

Sales Transactions

Our customers are trained to shop alphabetically. ☺

#1
bread
butter
eggs
milk

#2
bologna
bread
cheese
chips
mayo
soda

#3
bananas
bread
butter
cheese
oranges

#4
buns
chips
hotdogs
mustard
soda

#5
buns
chips
hotdogs
mustard
pickles
soda

#6
bread
butter
cereal
eggs
milk

#7
bananas
cereal
eggs
milk
oranges

#8
bologna
bread
buns
cheese
chips
hotdogs
mayo
mustard
soda

#9
bananas
bologna
bread
cheese
milk
oranges
soda

#10
bread
butter
cereal
eggs
milk

#11
bananas
chips
soda

#12
bread
butter
eggs
milk
oranges

#13
bananas
bologna
bread
cheese
mayo
mustard

#14
bread
cereal
eggs
milk

#15
bologna
bread
cheese
chips
mayo
mustard
soda

#16
bread
butter
eggs
milk
oranges

#17
buns
chips
hotdogs
soda

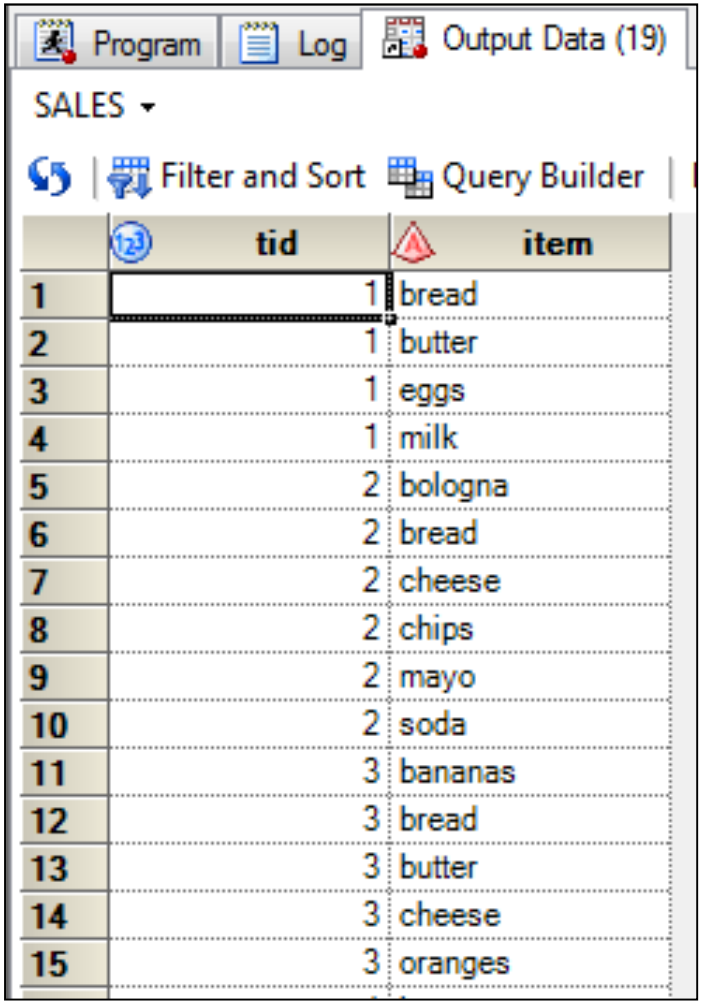
#18
buns
cheese
chips
hotdogs
mustard
soda

#19
chips
pickles
soda

#20
bologna
bread
cheese
chips
mayo
mustard
soda

Creating the Sales dataset

```
data work.sales;
input tid item $;
datalines;
1 bread
1 butter
1 eggs
1 milk
2 bologna
2 bread
…
20 mustard
20 soda
;
run;
```



	tid	item
1	1	bread
2	1	butter
3	1	eggs
4	1	milk
5	2	bologna
6	2	bread
7	2	cheese
8	2	chips
9	2	mayo
10	2	soda
11	3	bananas
12	3	bread
13	3	butter
14	3	cheese
15	3	oranges

Support

- The **support** of an item is the number of transactions containing that item.
- **Minimum support** is one of the parameters to the MBA macro.
- Items not meeting the minimum support criteria are excluded from further analysis.
- Support can be expressed as a count, or as a percentage of all tranactions.
- For our purposes, we will assume a minimum support requirement of four.

Support

pickles do not meet our minimum support requirement (4).

#1
bread
butter
eggs
milk

#2
bologna
bread
cheese
chips
mayo
soda

#3
bananas
bread
butter
cheese
oranges

#4
buns
chips
hotdogs
mustard
soda

#5
buns
chips
hotdogs
mustard
pickles
soda

#6
bread
butter
cereal
eggs
milk

#7
bananas
cereal
eggs
milk
oranges

#8
bologna
bread
buns
cheese
chips
hotdogs
mayo
mustard
soda

#9
bananas
bologna
bread
cheese
milk
oranges
soda

#10
bread
butter
cereal
eggs
milk

#11
bananas
chips
soda

#12
bread
butter
eggs
milk
oranges

#13
bananas
bologna
bread
cheese
mayo
mustard

#14
bread
cereal
eggs
milk

#15
bologna
bread
cheese
chips
mayo
mustard
soda

#16
bread
butter
eggs
milk
oranges

#17
buns
chips
hotdogs
soda

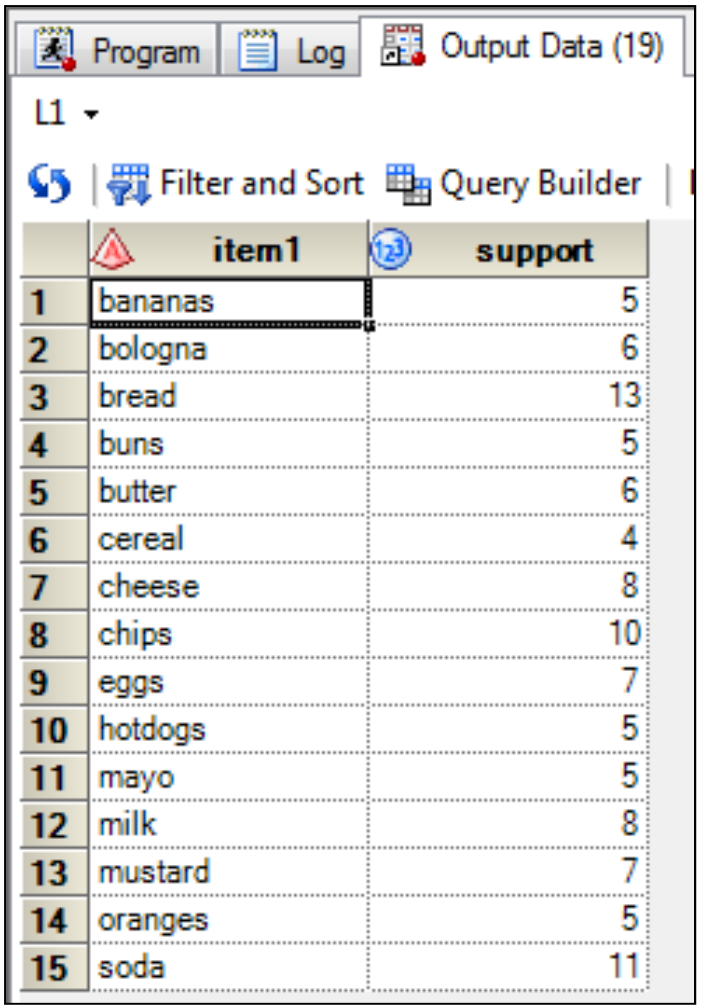
#18
buns
cheese
chips
hotdogs
mustard
soda

#19
chips
pickles
soda

#20
bologna
bread
cheese
chips
mayo
mustard
soda

Support

- Items meeting the minimum support requirement are included in subsequent processing.



	item1	support
1	bananas	5
2	bologna	6
3	bread	13
4	buns	5
5	butter	6
6	cereal	4
7	cheese	8
8	chips	10
9	eggs	7
10	hotdogs	5
11	mayo	5
12	milk	8
13	mustard	7
14	oranges	5
15	soda	11

Pairs

- We then create all possible pairings of the **surviving** items and see if the pair of items meets the minimum support requirement.
- This limiting ourselves to the surviving items is the key point of the **apriori algorithm**.
- The **support** of each pair of items is the number of transactions containing that pair.
- Pairs of item not meeting the minimum support criteria are excluded from further analysis.
- Consider the following pairings…

Pairs

bananas → support = 5
oranges → support = 5
bananas, oranges → support = 3
insufficient support

#1
bread
butter
eggs
milk

#2
bologna
bread
cheese
chips
mayo
soda

#3
bananas
bread
butter
cheese
oranges

#4
buns
chips
hotdogs
mustard
soda

#5
buns
chips
hotdogs
mustard
pickles
soda

#6
bread
butter
cereal
eggs
milk

#7
bananas
cereal
eggs
milk
oranges

#8
bologna
bread
buns
cheese
chips
hotdogs
mayo
mustard
soda

#9
bananas
bologna
bread
cheese
milk
oranges
soda

#10
bread
butter
cereal
eggs
milk

#11
bananas
chips
soda

#12
bread
butter
eggs
milk
oranges

#13
bananas
bologna
bread
cheese
mayo
mustard

#14
bread
cereal
eggs
milk

#15
bologna
bread
cheese
chips
mayo
mustard
soda

#16
bread
butter
eggs
milk
oranges

#17
buns
chips
hotdogs
soda

#18
buns
cheese
chips
hotdogs
mustard
soda

#19
chips
pickles
soda

#20
bologna
bread
cheese
chips
mayo
mustard
soda

Pairs

bologna → support = 6
chips → support = 10
bologna, chips → support = 4
sufficient support

#1
bread
butter
eggs
milk

#2 ★
bologna
bread
cheese
chips
mayo
soda

#3
bananas
bread
butter
cheese
oranges

#4
buns
chips
hotdogs
mustard
soda

#5
buns
chips
hotdogs
mustard
pickles
soda

#6
bread
butter
cereal
eggs
milk

#7
bananas
cereal
eggs
milk
oranges

#8 ★
bologna
bread
buns
cheese
chips
hotdogs
mayo
mustard
soda

#9
bananas
bologna
bread
cheese
milk
oranges
soda

#10
bread
butter
cereal
eggs
milk

#11
bananas
chips
soda

#12
bread
butter
eggs
milk
oranges

#13
bananas
bologna
bread
cheese
mayo
mustard

#14
bread
cereal
eggs
milk

#15 ★
bologna
bread
cheese
chips
mayo
mustard
soda

#16
bread
butter
eggs
milk
oranges

#17
buns
chips
hotdogs
soda

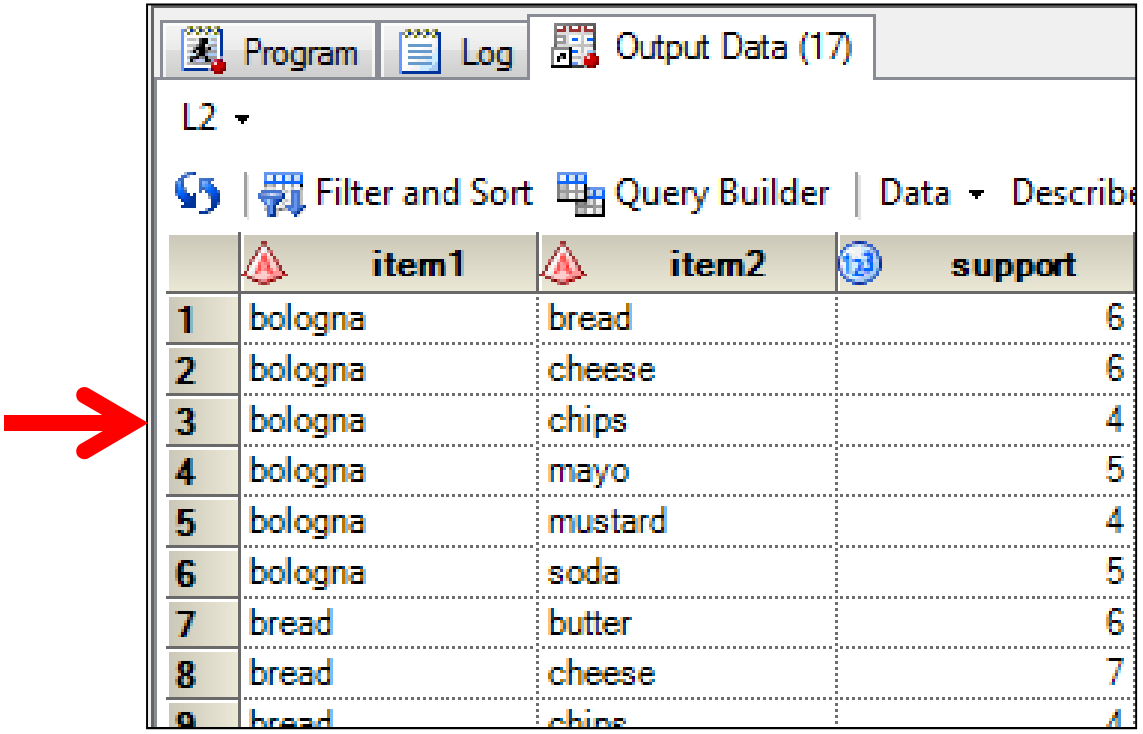
#18
buns
cheese
chips
hotdogs
mustard
soda

#19
chips
pickles
soda

#20 ★
bologna
bread
cheese
chips
mayo
mustard
soda

Pairs

- Pairs of items meeting the minimum support requirement are included in subsequent processing.



	item1	item2	support
1	bologna	bread	6
2	bologna	cheese	6
3	bologna	chips	4
4	bologna	mayo	5
5	bologna	mustard	4
6	bologna	soda	5
7	bread	butter	6
8	bread	cheese	7
9	bread	chips	4

Iterate

- We then repeat the process, iterating with itemsets of size 3, itemsets of size 4, etc., until:
  - we are unable to find any itemsets with sufficient support, or
  - we reach the indicated maximum number of iterations (this is one of the parameters to the MBA macro).

Association rules

- Our final results are expressed using association rules:

$$\{LHS\} \rightarrow \{RHS\}\ [support, confidence]$$

- LHS stands for Left Hand Side
- RHS stands for Right Hand Side
- Example:
  - {bologna} → {chips} [0.2, (discussed next)]
  - {chips} → {bologna} [0.2, (discussed next)]

0.2 = 4 / 20

Confidence

- Confidence is defined as the **conditional probablity** that a transaction containing the LHS will also contain the RHS.

$$\frac{support(LHS \cup RHS)}{support(LHS)}$$

- Confidence for {bologna} → {chips}

$$\frac{support(bologna, chips)}{support(bologna)} = \frac{4/20}{6/20} = 0.67$$

- Confidence for {chips} → {bologna}

$$\frac{support(chips, bologna)}{support(chips)} = \frac{4/20}{10/20} = 0.40$$

Lift

- Lift is a measure of the **improvement** in occurrence of the RHS given the association rule over the occurence of the RHS regardless. We'd like to see a lift value greater than one.

$$\frac{confidence(LHS \rightarrow RHS)}{support(RHS)}$$

- Lift for {bologna} → {chips}

$$\frac{confidence(bologna \rightarrow chips)}{support(chips)} = \frac{0.67}{10/20} = 1.33$$

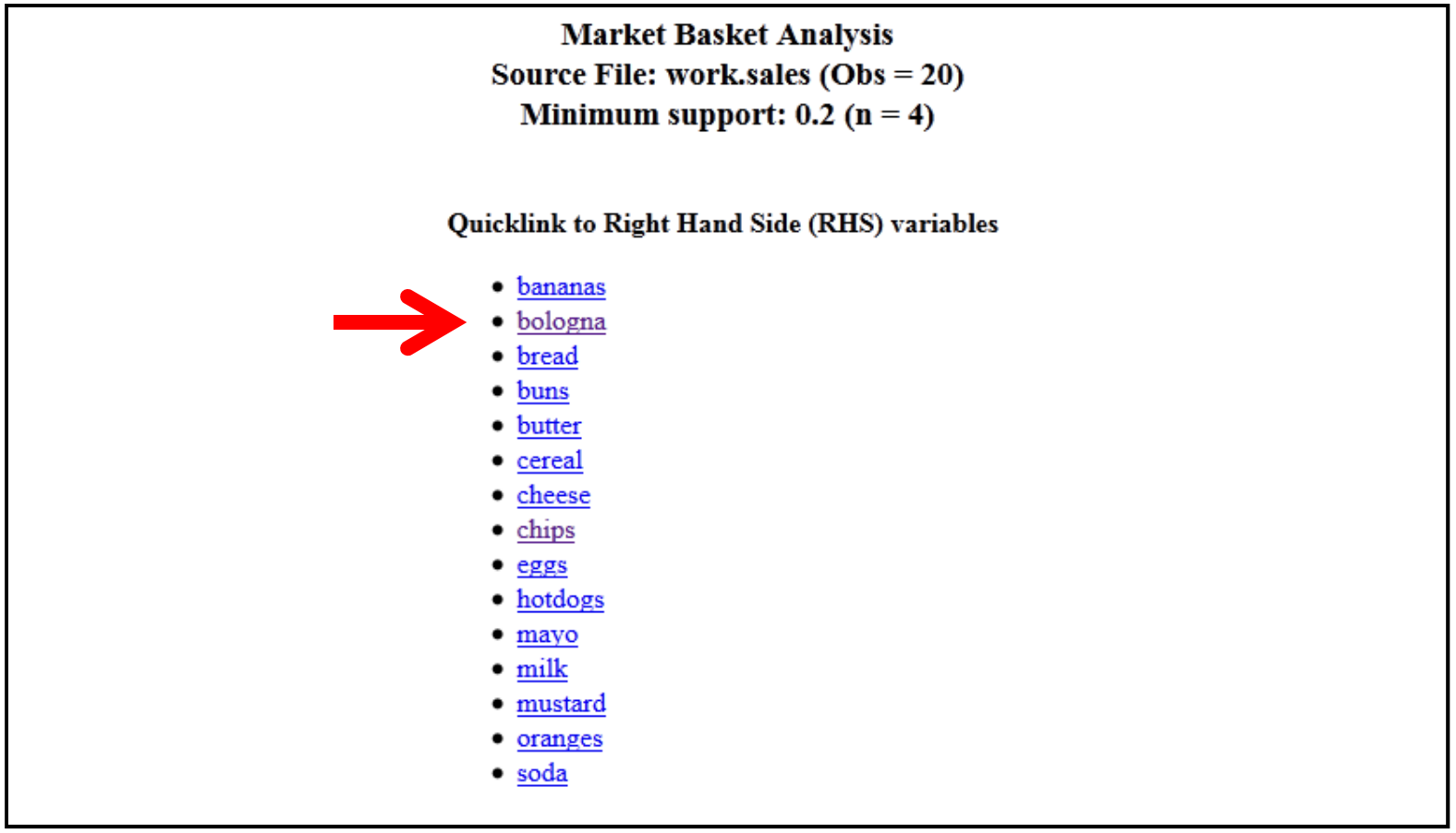
- Lift for {chips} → {bologna}

$$\frac{confidence(chips \rightarrow bologna)}{support(bologna)} = \frac{0.40}{6/20} = 1.33$$

Running the SAS MBA macro

```
%mba(work.sales,                    ← SAS transactions dataset
     "Y",                           ← is item id a string? "Y" or "N".
     work.Results,                  ← name of SAS results dataset
     5,                             ← maximum iterations
     0.2,                           ← minimum support (0.2 = 20%)
     "C:\temp\mba.html");           ← web page output
run;
```

Running the SAS MBA macro



Running the SAS MBA macro

RHS: bananas

LHS 1	LHS 2	LHS 3	LHS 4	Support	Confidence	Lift
bananas				5	.	.

RHS: bologna

LHS 1	LHS 2	LHS 3	LHS 4	Support	Confidence	Lift
bread	cheese	chips	mayo	4	1.00	3.33
bread	cheese	chips	soda	4	1.00	3.33
bread	cheese	mayo	mustard	4	1.00	3.33
bread	cheese	mayo	soda	4	1.00	3.33
bread	chips	mayo	soda	4	1.00	3.33
cheese	chips	mayo	soda	4	1.00	3.33
bread	cheese	chips		4	1.00	3.33
bread	cheese	mayo		5	1.00	3.33
bread	cheese	mustard		4	1.00	3.33
bread	cheese	soda		5	1.00	3.33

Running the SAS MBA macro

(RHS: bologna (continued)

mayo	soda			4	1.00	3.33
bread				6	0.46	1.54
cheese				6	0.75	2.50
chips				4	0.40	1.33
mayo				5	1.00	3.33
mustard				4	0.57	1.90
soda				5	0.45	1.52
bologna				6	.	.

RHS: bread

LHS 1	LHS 2	LHS 3	LHS 4	Support	Confidence	Lift
bologna	cheese	chips	mayo	4	1.00	1.54
bologna	cheese	chips	soda	4	1.00	1.54
bologna	cheese	mayo	mustard	4	1.00	1.54
bologna	cheese	mayo	soda	4	1.00	1.54
bologna	chips	mayo	soda	4	1.00	1.54
cheese	chips	mayo	soda	4	1.00	1.54

Questions?