

CS 422

DATA MINING

HOMework ASSIGNMENT 2  
PRACTICUM QUESTIONS 1 AND 2

RONIT RUDRA  
A20379221  
rrudra@hawk.iit.edu

Course Instructor  
Prof. Jawahar PANCHAL

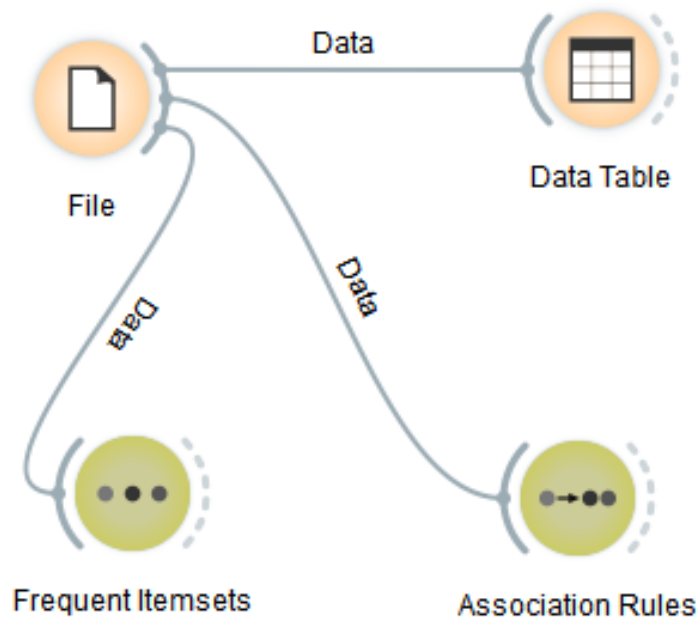
## 1 Question 1:

Load the *market-basket* sample dataset into the Orange application, and run both frequent itemset as well as association rule modules. Set the *support threshold* to 10% and observe the *antecedent* in the rules with the highest lift. What item is observed to be there, and what is its support? Is this a valuable association rule? Why or why not?

### Answer:

- Creating appropriate workflow with the *market-basket.tab* dataset.

Figure 1: Workflow



- Figure 2. shows the market-basket dataset tabulated in *Orange*.

Figure 2: Market Basket Dataset

Info							
5 instances							
6 features (40.0% missing values)							
No target variable.							
No meta attributes							
Variables							
<input checked="" type="checkbox"/> Show variable labels (if present)							
<input type="checkbox"/> Visualize continuous values							
<input checked="" type="checkbox"/> Color by instance classes							
Selection							
<input checked="" type="checkbox"/> Select full rows							
		Bread	Milk	Diapers	Beer	Eggs	Cola
1	1		1	?	?	?	?
2	1		?	1	1	1	?
3	?		1	1	1	?	1
4	1		1	1	1	?	?
5	1		1	1	?	?	1

- Setting up the support threshold to 10% and confidence to 90% in both the Frequent Itemset generation as well as Association Rule generation, we get results as shown in Figure 3 and 4.

Figure 3: Frequent Itemsets in Market-basket

\*\*\* Frequent Itemsets

Info

Number of itemsets: 35

Selected itemsets: 0

Selected examples: 0

Expand all Collapse all

Find itemsets

Minimal support:  10%

Max. number of itemsets:  10000

☐ Find itemsets

Filter itemsets

Contains:

Min. items:  Max. items:

☒ Apply these filters in search

Itemsets	Support	%
▼ Bread=1	4	80
▼ Milk=1	3	60
▼ Diapers=1	2	40
Beer=1	1	20
Cola=1	1	20
Beer=1	1	20
Cola=1	1	20
▼ Diapers=1	3	60
▼ Beer=1	2	40
Eggs=1	1	20
Eggs=1	1	20
Cola=1	1	20
▼ Beer=1	2	40
Eggs=1	1	20
Eggs=1	1	20
Cola=1	1	20
▼ Milk=1	4	80
▼ Diapers=1	3	60
▼ Beer=1	2	40
Cola=1	1	20
Cola=1	2	40
▼ Beer=1	2	40
Cola=1	1	20
Cola=1	2	40
▼ Diapers=1	4	80
▼ Beer=1	3	60
Eggs=1	1	20
Cola=1	1	20
Eggs=1	1	20
Cola=1	2	40
▼ Beer=1	3	60
Eggs=1	1	20
Cola=1	1	20
Eggs=1	1	20
Cola=1	2	40

Figure 4: Association Rules in Market-Basket

Association Rules									
Info									
Number of rules: 38									
Filtered rules: 38									
Selected rules: 0									
Selected examples: 0									
Find association rules									
Minimal support: 10%									
Minimal confidence: 90%									
Max. number of rules: 10000									
<input type="checkbox"/> Induce classification (itemset → class) rules									
Find rules									
Filter rules									
Antecedent									
Contains:									
Min. items: 1 Max. items: 999									
Consequent									
Contains:									
Min. items: 1 Max. items: 999									
<input checked="" type="checkbox"/> Apply these filters in search									
Supp	Conf	Cov	Strg	Lift	Lev	Antecedent		Consequent	
0.20	1.00	0.20	2.00	2.50	0.12	Eggs=1 →		Bread=1, Beer=1	
0.20	1.00	0.20	2.00	2.50	0.12	Diapers=1, Eggs=1 →		Bread=1, Beer=1	
0.20	1.00	0.20	2.00	2.50	0.12	Eggs=1 →		Bread=1, Diapers=1, Beer=1	
0.40	1.00	0.40	1.50	1.67	0.16	Cola=1 →		Milk=1, Diapers=1	
0.20	1.00	0.20	3.00	1.67	0.08	Eggs=1 →		Bread=1, Diapers=1	
0.20	1.00	0.20	3.00	1.67	0.08	Eggs=1 →		Beer=1	
0.20	1.00	0.20	3.00	1.67	0.08	Bread=1, Eggs=1 →		Beer=1	
0.20	1.00	0.20	3.00	1.67	0.08	Diapers=1, Eggs=1 →		Beer=1	
0.20	1.00	0.20	3.00	1.67	0.08	Eggs=1 →		Diapers=1, Beer=1	
0.20	1.00	0.20	3.00	1.67	0.08	Beer=1, Eggs=1 →		Bread=1, Diapers=1	
0.20	1.00	0.20	3.00	1.67	0.08	Bread=1, Diapers=1, Eggs=1 →		Beer=1	
0.20	1.00	0.20	3.00	1.67	0.08	Bread=1, Eggs=1 →		Diapers=1, Beer=1	
0.20	1.00	0.20	3.00	1.67	0.08	Bread=1, Cola=1 →		Milk=1, Diapers=1	
0.20	1.00	0.20	3.00	1.67	0.08	Beer=1, Cola=1 →		Milk=1, Diapers=1	
0.60	1.00	0.60	1.33	1.25	0.12	Beer=1 →		Diapers=1	
0.40	1.00	0.40	2.00	1.25	0.08	Bread=1, Beer=1 →		Diapers=1	
0.40	1.00	0.40	2.00	1.25	0.08	Milk=1, Beer=1 →		Diapers=1	
0.40	1.00	0.40	2.00	1.25	0.08	Cola=1 →		Milk=1	
0.40	1.00	0.40	2.00	1.25	0.08	Cola=1 →		Diapers=1	
0.40	1.00	0.40	2.00	1.25	0.08	Diapers=1, Cola=1 →		Milk=1	
0.40	1.00	0.40	2.00	1.25	0.08	Milk=1, Cola=1 →		Diapers=1	
0.20	1.00	0.20	4.00	1.25	0.04	Bread=1, Milk=1, Beer=1 →		Diapers=1	
0.20	1.00	0.20	4.00	1.25	0.04	Eggs=1 →		Bread=1	
0.20	1.00	0.20	4.00	1.25	0.04	Eggs=1 →		Diapers=1	
0.20	1.00	0.20	4.00	1.25	0.04	Diapers=1, Eggs=1 →		Bread=1	
0.20	1.00	0.20	4.00	1.25	0.04	Diapers=1, Eggs=1 →		Diapers=1	
0.20	1.00	0.20	4.00	1.25	0.04	Beer=1, Eggs=1 →		Bread=1	
0.20	1.00	0.20	4.00	1.25	0.04	Beer=1, Eggs=1 →		Diapers=1	
0.20	1.00	0.20	4.00	1.25	0.04	Diapers=1, Beer=1, Eggs=1 →		Bread=1	
0.20	1.00	0.20	4.00	1.25	0.04	Bread=1, Beer=1, Eggs=1 →		Diapers=1	
0.20	1.00	0.20	4.00	1.25	0.04	Bread=1, Cola=1 →		Milk=1	
0.20	1.00	0.20	4.00	1.25	0.04	Bread=1, Cola=1 →		Diapers=1	
0.20	1.00	0.20	4.00	1.25	0.04	Bread=1, Diapers=1, Cola=1 →		Milk=1	
0.20	1.00	0.20	4.00	1.25	0.04	Bread=1, Milk=1, Cola=1 →		Diapers=1	
0.20	1.00	0.20	4.00	1.25	0.04	Beer=1, Cola=1 →		Milk=1	
0.20	1.00	0.20	4.00	1.25	0.04	Beer=1, Cola=1 →		Diapers=1	
0.20	1.00	0.20	4.00	1.25	0.04	Diapers=1, Beer=1, Cola=1 →		Milk=1	
0.20	1.00	0.20	4.00	1.25	0.04	Milk=1, Beer=1, Cola=1 →		Diapers=1	

- the table in Figure 4. has been sorted based on the descending order of the *lift* metric.

### 1.1 Observations:

- From Figure 4. the association rules with the highest lift are:

$$\begin{aligned} &\{eggs \rightarrow \{bread, beer\}\} \\ &\{\{diapers, eggs\} \rightarrow \{bread, beer\}\} \\ &\{Eggs \rightarrow \{bread, diapers, beer\}\} \end{aligned}$$

- All of the above have a lift of textit2.50.
- The two items in the *antecedent*, **eggs** and **diapers**, have a support of **20%** and **80%** respectively.
- As for the association rules, all of them have a support of **20%**.

### 1.2 Inference:

The Lift metric (also known as Interest), is the measure of how many times items X and Y occur together than expected if they

were statistically independent.

$$Lift(X \rightarrow Y) = \frac{Conf(X \rightarrow Y)}{Supp(Y)} = \frac{P(X \rightarrow Y)}{P(X)P(Y)} \quad (1)$$

From figure 4 it is observed that all associations have a confidence of 1. From the above equation, **lift is inversely proportional to the support of the antecedent**. Thus, **antecedents with a low support count can produce high lift values**. This usually happens when the **dataset is small** like this one and one rare item occurs a minimum of once along with other common items. In this dataset, Eggs appears in only one transaction with 3 other items and results in an enormous lift.

Thus, these association rules are **not valuable at all**.

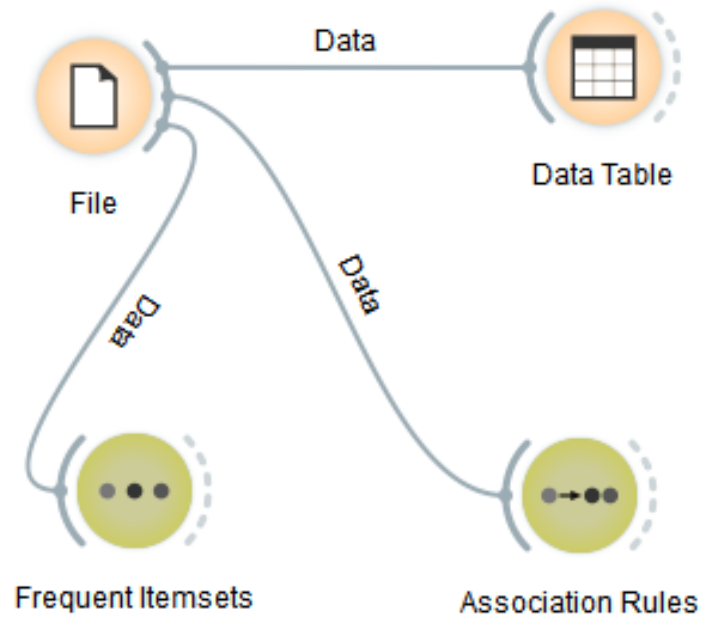
## 2 Question 2:

Load the *Extended Bakery* dataset (**75000-out2-final.csv**) into the Orange application, and run both frequent itemset as well as association rule modules. Set the *support threshold* to 1% and the *confidence threshold* to 90%. Observe the association rules containing the *Cherry Tart* item within the antecedent. What other item appears with it? When the *confidence threshold* is lowered to 45%, does the *Cherry Tart* item now appear without another item in the *antecedent*? Is the same *consequent* observed in both cases? How did lowering the confidence threshold lead to this change? Hint: Reference the Simpsons Paradox section of the text.

### **Answer:**

- Creating appropriate workflow with the *75000-out2-final.csv* dataset.

Figure 5: Workflow



- Figure 6. shows some of the market-basket dataset tabulated in *Orange*.

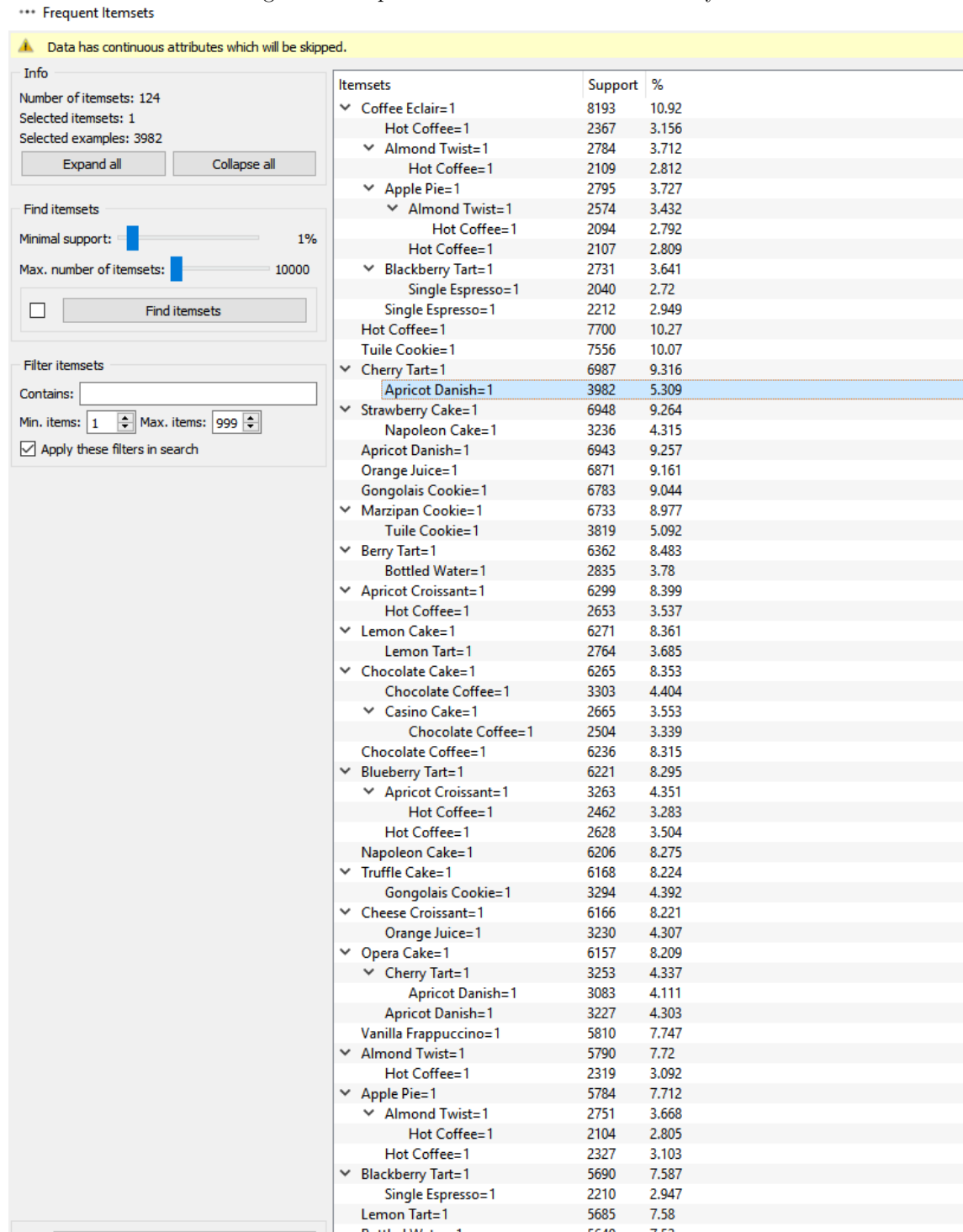
Figure 6: Extended Bakery Dataset

<div> <div>Data Table</div> <div> <div>Info</div> <div> 75500 instances  55 features (91.1% missing values)  No target variable.  No meta-attributes </div> </div> <div> <div>Variables</div> <div> <input checked="" type="checkbox"/> Show variable labels (if present)  <input type="checkbox"/> Visualize continuous values  <input checked="" type="checkbox"/> Color by instance classes </div> </div> <div> <div>Selection</div> <div> <input checked="" type="checkbox"/> Select full rows </div> </div> </div>		Transaction Number	Chocolate Cake	Lemon Cake	Cesino Cake	Opera Cake	Strawberry Cake	Truffle Cake	Chocolate Eclair	Coffee Eclair	Vanilla Eclair	Napoleon Cake	Almond Tart	Apple Pie	Apple Tart	Apricot Tart	Berry Tart	Black
1	1,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
2	2,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
3	3,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
4	4,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
5	5,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
6	6,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
7	7,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
8	8,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
9	9,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
10	10,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
11	11,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
12	12,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
13	13,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
14	14,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
15	15,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
16	16,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
17	17,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
18	18,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
19	19,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
20	20,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
21	21,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
22	22,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
23	23,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
24	24,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
25	25,000	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?

- First setting up the support threshold to 1% and confidence to 90% in both the Frequent Itemset generation as well as Association Rule generation, we get results as shown in Figure 7 and 8.



Figure 7: Frequent Itemsets in Extended Bakery%



**Info**

Number of rules: 95

Filtered rules: 14

Selected rules: 1

Selected examples: 3083

**Find association rules**

Minimal support:

Minimal confidence:

Max. number of rules:

☐ Induce classification (Itemset → class) rules

☐ **Find rules**

**Filter rules**

Antecedent

Contains:

Min. items:  Max. items:

Consequent

Contains:

Min. items:  Max. items:

☒ Apply these filters in search

**Data has continuous attributes which will be skipped.**

	Supp	Conf	Covr	Strg	Lift	Lvr	Antecedent	Consequent
	0.02	0.99	0.02	3.28	14.51	0.02	Apple Croissant=1, Apple Danish=1, Cherry Soda=1	Apple Tart=1
	0.02	0.91	0.02	2.96	13.42	0.02	Apple Croissant=1, Cherry Soda=1	Apple Danish=1
	0.02	0.91	0.02	2.90	13.30	0.02	Apple Croissant=1, Cherry Soda=1	Apple Tart=1
	0.03	0.92	0.03	2.43	13.53	0.02	Apple Tart=1, Apple Croissant=1	Apple Danish=1
	0.02	0.99	0.02	3.26	14.64	0.02	Apple Tart=1, Apple Croissant=1, Cherry Soda=1	Apple Danish=1
	0.03	0.92	0.03	2.42	13.61	0.02	Apple Tart=1, Apple Danish=1	Apple Croissant=1
	0.02	0.99	0.02	3.25	14.76	0.02	Apple Tart=1, Apple Danish=1, Cherry Soda=1	Apple Croissant=1
	0.02	0.90	0.02	3.22	39.99	0.02	Apple Tart=1, Cherry Soda=1	Apple Croissant=1, Apple Danish=1
	0.02	0.91	0.02	2.95	13.55	0.02	Apple Tart=1, Cherry Soda=1	Apple Croissant=1
	0.02	0.91	0.02	2.97	13.44	0.02	Apple Tart=1, Cherry Soda=1	Apple Danish=1
	0.03	0.92	0.03	3.71	8.45	0.02	Blackberry Tart=1, Single Espresso=1	Coffeeclair=1
	0.03	0.94	0.04	2.40	11.15	0.03	Blueberry Tart=1, Hot Coffee=1	Apricot Croissant=1
	0.03	0.94	0.02	2.71	10.10	0.02	Chocolate Tart=1, Walnut Cookies=1	Vanilla Freguccaron=1
	0.04	0.95	0.04	2.13	10.24	0.04	Opera Cake=1, Cherry Tart=1	Apricot Danish=1

- Now, we reduce the confidence level and observe the changes as shown in figure 9.

**Data has continuous attributes which will be skipped.**

### Info

- Number of rules: 194
- Filtered rules: 33
- Selected rules: 4
- Selected examples: 4152

### Find association rules

Minimal support:  1%

Minimal confidence:  45%

Max. number of rules:  10000  
☒ Induce classification (Itemset → class) rules

---

### Filter rules

Antecedent

Contains:  cherry tart

Min. items:  1 Max. items:  999

Consequent

Contains:

Min. items:  1 Max. items:  999

☒ Show all filters in search

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent	Consequent
0.02	0.99	0.02	3.28	14.51	0.02	Apple Croissant=1, Apple Danish=1, Cherry Soda=1	Apple Tart=1
0.02	0.91	0.02	2.98	13.30	0.02	Apple Croissant=1, Cherry Soda=1	Apple Tart=1
0.02	0.91	0.02	2.96	13.42	0.02	Apple Croissant=1, Cherry Soda=1	Apple Danish=1
0.02	0.90	0.02	1.22	32.28	0.02	Apple Croissant=1, Cherry Soda=1	Apple Tart=1, Apple Danish=1
0.02	0.90	0.02	2.95	13.14	0.02	Apple Danish=1, Cherry Soda=1	Apple Tart=1
0.02	0.90	0.02	2.91	13.36	0.02	Apple Danish=1, Cherry Soda=1	Apple Croissant=1
0.02	0.89	0.02	1.20	31.97	0.02	Apple Danish=1, Cherry Soda=1	Apple Tart=1, Apple Croissant=1
0.03	0.92	0.03	2.43	13.53	0.02	Apple Tart=1, Apple Croissant=1	Apple Danish=1
0.02	0.75	0.03	2.23	12.04	0.02	Apple Tart=1, Apple Croissant=1	Cherry Soda=1
0.02	0.74	0.03	0.83	31.87	0.02	Apple Tart=1, Apple Croissant=1	Apple Danish=1, Cherry Soda=1
0.02	0.81	0.03	2.43	13.02	0.02	Apple Tart=1, Apple Croissant=1, Apple Danish=1	Cherry Soda=1
0.02	0.99	0.02	3.26	14.64	0.02	Apple Tart=1, Apple Croissant=1, Cherry Soda=1	Apple Danish=1
0.03	0.92	0.03	2.42	13.61	0.02	Apple Tart=1, Apple Danish=1	Apple Croissant=1
0.02	0.74	0.03	2.23	12.01	0.02	Apple Tart=1, Apple Danish=1	Cherry Soda=1
0.02	0.74	0.03	0.82	32.28	0.02	Apple Tart=1, Apple Danish=1	Apple Croissant=1, Cherry Soda=1
0.02	0.99	0.02	3.25	14.54	0.02	Apple Tart=1, Apple Danish=1, Cherry Soda=1	Apple Croissant=1
0.02	0.91	0.02	2.97	13.44	0.02	Apple Tart=1, Cherry Soda=1	Apple Danish=1
0.02	0.91	0.02	2.95	13.55	0.02	Apple Tart=1, Cherry Soda=1	Apple Croissant=1
0.02	0.90	0.02	1.22	32.39	0.02	Apple Tart=1, Cherry Soda=1	Apple Croissant=1, Apple Danish=1
0.04	0.48	0.08	1.44	4.39	0.03	Blackberry Tart=1	Coffee Eclair=1
0.04	0.92	0.03	3.71	8.45	0.02	Blackberry Tart=1, Single Espresso=1	Coffee Eclair=1
0.04	0.52	0.08	1.01	6.25	0.04	Blueberry Tart=1	Apricot Croissant=1
0.03	0.75	0.04	2.36	7.33	0.03	Blueberry Tart=1, Apricot Croissant=1	Hot Coffee=1
0.03	0.94	0.04	2.40	11.15	0.03	Blueberry Tart=1, Hot Coffee=1	Apricot Croissant=1
0.05	0.57	0.09	0.99	6.16	0.04	Cherry Tart=1	Apricot Danish=1
0.04	0.47	0.09	0.88	5.67	0.04	Cherry Tart=1	Opera Cake=1
0.04	0.77	0.09	1.55	9.43	0.04	Cherry Tart=1, Apricot Danish=1	Opera Cake=1
0.04	0.49	0.07	1.05	6.30	0.03	Chocolate Tart=1	Vanilla Frappuccino=1
0.03	0.74	0.04	1.89	10.97	0.02	Chocolate Tart=1, Vanilla Frappuccino=1	Walnut Cookies=1
0.03	0.94	0.03	2.71	12.10	0.02	Chocolate Tart=1, Walnut Cookies=1	Vanilla Frappuccino=1
0.03	0.75	0.04	1.87	10.99	0.02	Coffee Eclair=1, Blackberry Tart=1	Single Espresso=1
0.04	0.49	0.08	1.10	5.81	0.03	Lemon Tart=1	Lemon Cake=1
0.04	0.95	0.04	2.13	10.24	0.04	Opera Cake=1, Cherry Tart=1	Apricot Danish=1

9

## 2.1 Observations:

- For the 90% confidence threshold, the *antecedent* *Cherry Tart* appears with **Opera Cake** and the *consequent* is **Apricot Danish** and has a confidence of 95%.
- For the 45% confidence threshold, the association rules and their confidences are:

$$\{\{Opera\ Cake, \ Cherry\ Tart\} \rightarrow \{Apricot\ Danish\}\} = 0.95$$

$$\{\{Cherry\ Tart, \ Apricot\ Danish\} \rightarrow \{Opera\ Cake\}\} = 0.77$$

$$\{\{Cherry\ Tart\} \rightarrow \{Apricot\ Danish\}\} = 0.57$$

$$\{\{Cherry\ Tart\} \rightarrow \{Opera\ Cake\}\} = 0.47$$

- On lowering the Confidence Threshold, *Cherry Tart* appears without another item in the *antecedent*.
- The item *Apricot Danish* appears as *consequent* in both cases.

## 2.2 Inference:

*Opera Cake* becomes a *consequent* if the confidence threshold is reduced. Let's calculate the odds ratio from the contingency tables given below.

Table 1: Cherry Tart and Apricot Danish

	y	$\bar{y}$
x	3982	3005
$\bar{x}$	2961	65052

Table 2: Cherry Tart and Opera Cake

	y	$\bar{y}$
x	3253	3734
$\bar{x}$	2904	65109

Table 3: Opera Cake and Apricot Danish

	y	$\bar{y}$
x	3227	2930
$\bar{x}$	3716	65127

The odds ratio for the above contingency tables is 29, 19 and 19. This means that if we look for the odds of finding *Opera Cake* in transactions containing either *Apricot Danish* or *Cherry Tart*, the odds are the same. The probability of buying *Apricot Danish* when a customer purchased *Cherry Tart* is high when *Opera Cake* is included in the transaction. The confidence goes lower when its removed. Thus the item *Opera Cake* behaves as a **confounding variable** and skews the inference from the association leading to **Simpson's Paradox**.