CS 422

Data Mining

Homework Assignment 2

Practicum Questions 1 and 2

Ronit Rudra
A20379221
rrudra@hawk.iit.edu

Course Instructor
Prof. Jawahar Panchal

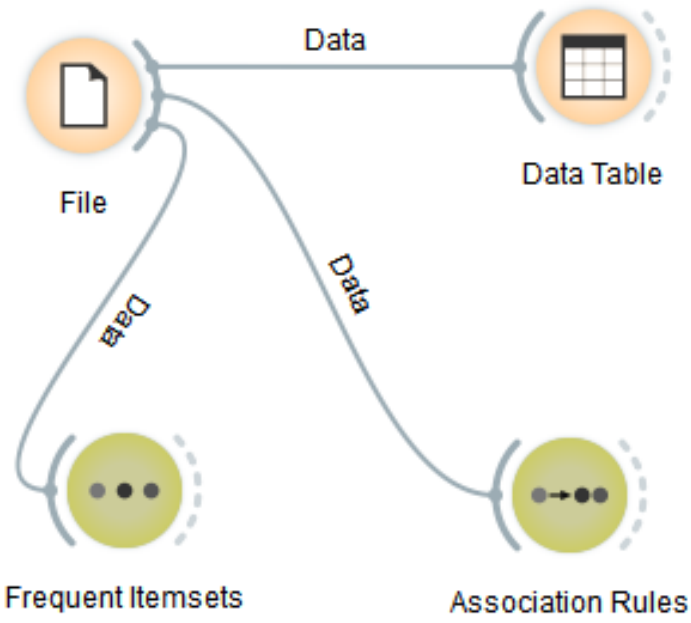
1 Question 1:

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

	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

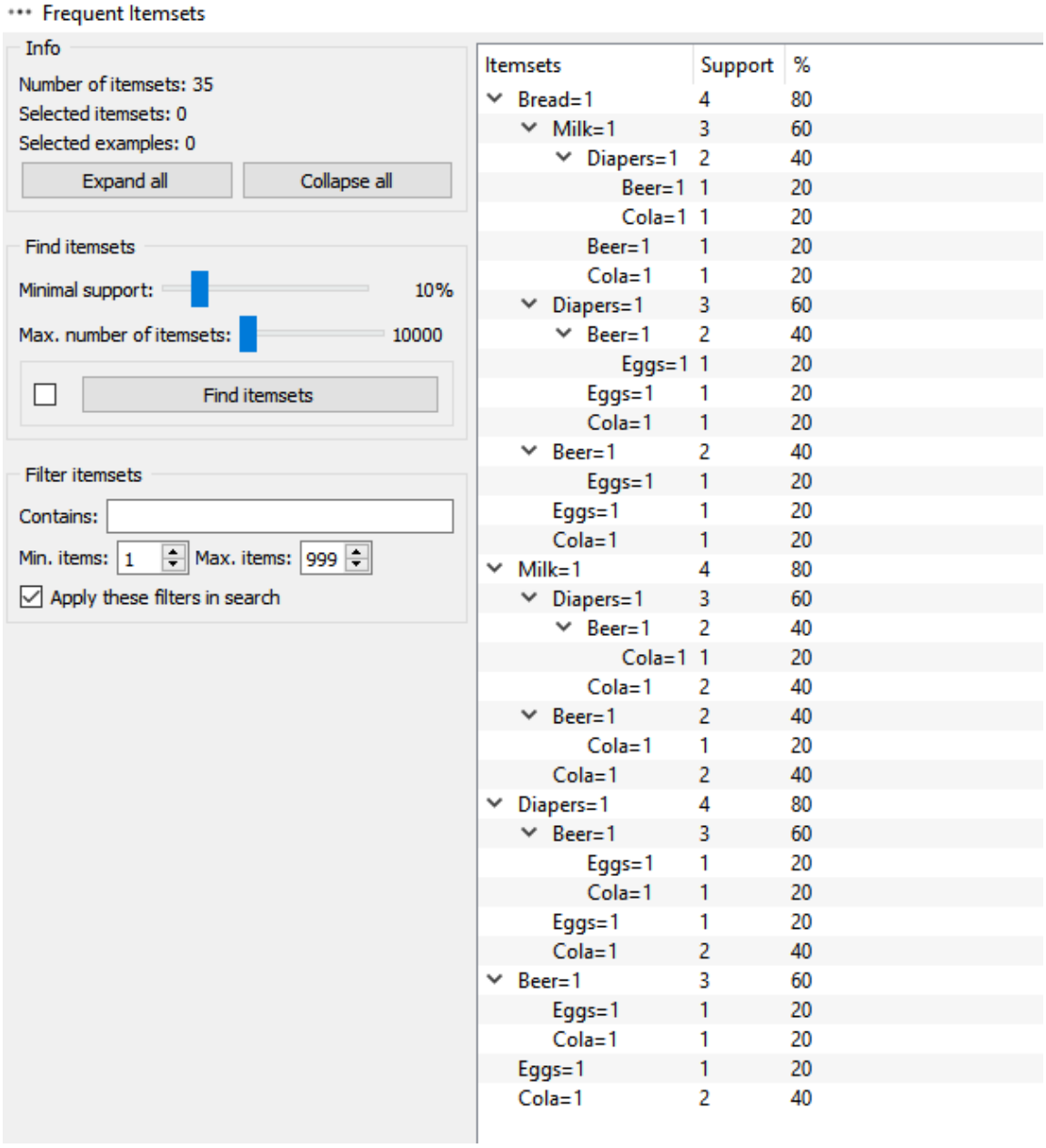
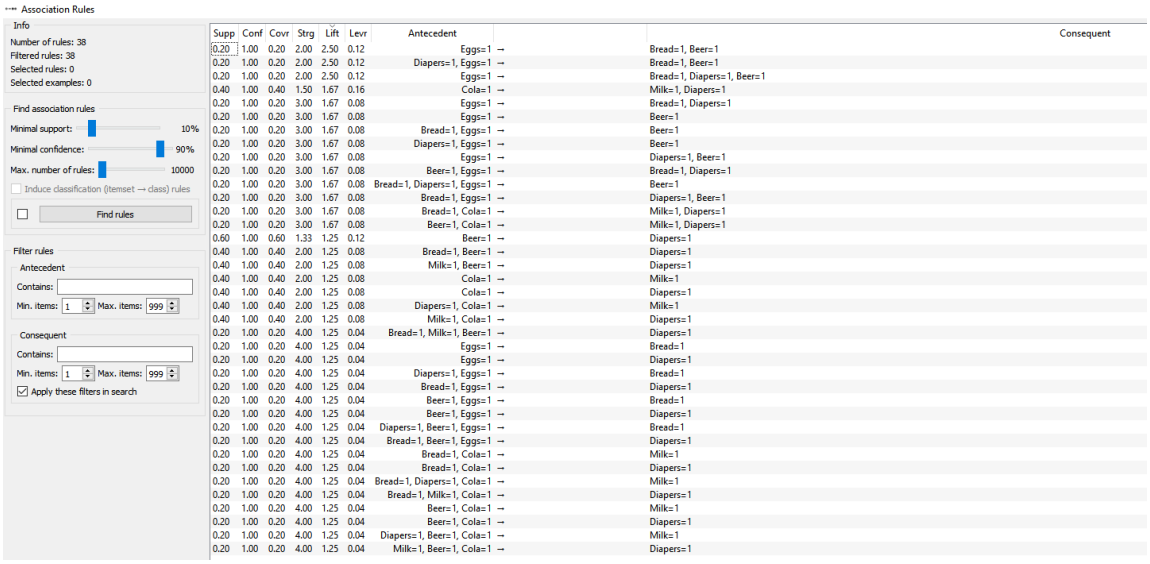


Figure 4: Association Rules in Market-Basket



Supp	Conf	Covr	Strg	Lift	Levr	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	Bread=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:

$$\{eggs \rightarrow \{bread, beer\}\}$$
$$\{\{diapers, eggs\} \rightarrow \{bread, beer\}\}$$
$$\{Eggs \rightarrow \{bread, diapers, beer\}\}$$

– 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.

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

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**.

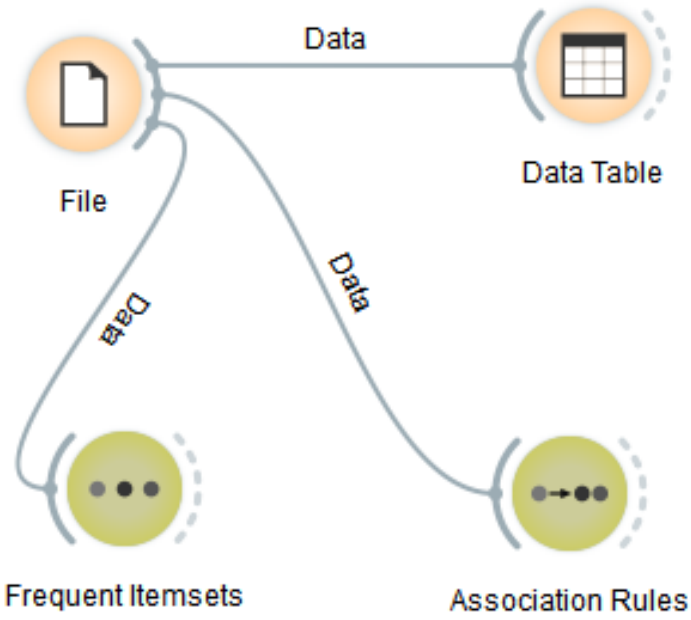
2 Question 2:

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:

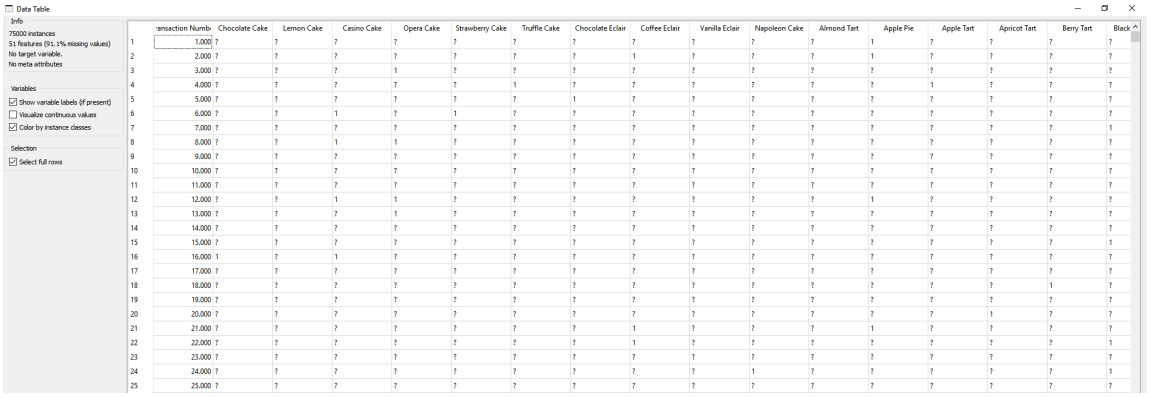
- Creating appropriate workflow with the *75000-out2-final.csv* dataset.

Figure 5: Workflow



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

Figure 6: Extended Bakery Dataset



- 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%

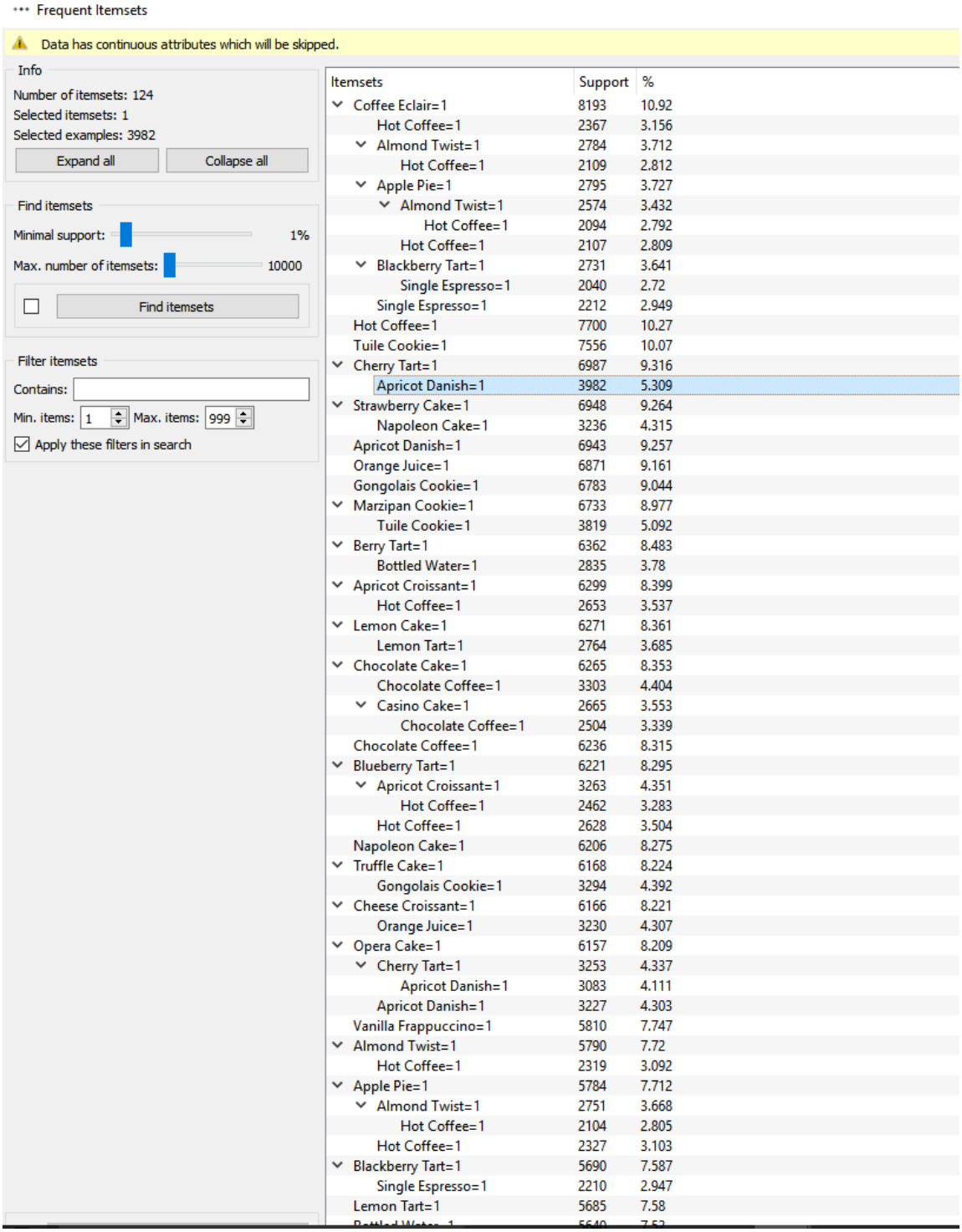
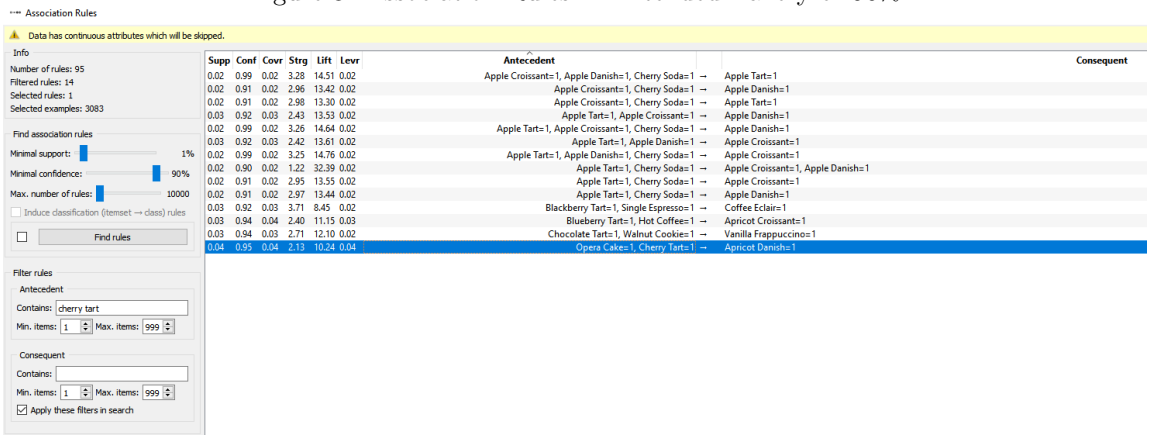


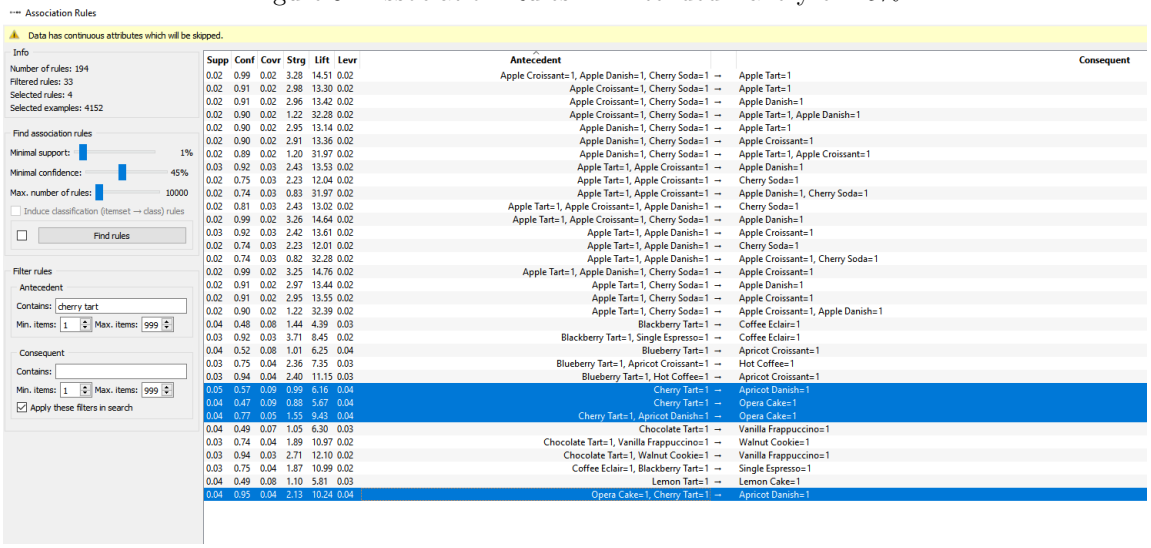
Figure 8: Association Rules in Extended Bakery c=90%



In Figure 8. the antecendents have been filtered by the "*Cherry Tart*" variable and highlighted.

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

Figure 9: Association Rules in Extended Bakery c=45%



In Figure 10. the antecendents have been filtered by the "*Cherry Tart*" variable and highlighted.

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

Table 2: Cherry Tart and Opera Cake

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

Table 3: Opera Cake and Apricot Danish

	y	$\overline{y}$
x	3227	2930
$\overline{x}$	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**.