**Practicum Problems**

**2.1 Problem 1**

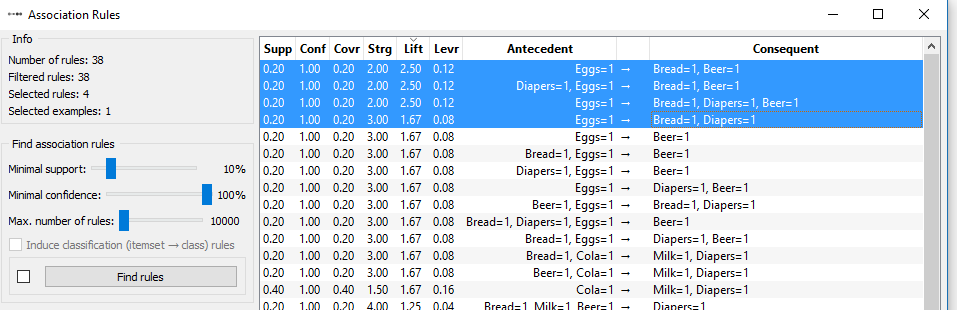
Load the market-basket sample dataset into the Orange application, and run both frequent item set as well as association rule modules. Set the support threshold to 10% and observe the antecedent in the rules with the highest lift.

What item is observed to be there, and what is its support? Is this a valuable association rule? Why or why not?

**Answer:**

When the support threshold is set to 10% we obtain Eggs and Eggs, Diaper as antecedents with highest lift=2.5 and support=20%.

Yes, the association rule is valuable because lift highlights rules which are rare but informative. Also, high-confidence rules can sometimes be misleading because the confidence measure ignores the support of the itemset appearing in the rule consequent. One way to address this problem is by applying a metric known as lift.

****

**2.2 Problem 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 Simpson's Paradox section of the text.

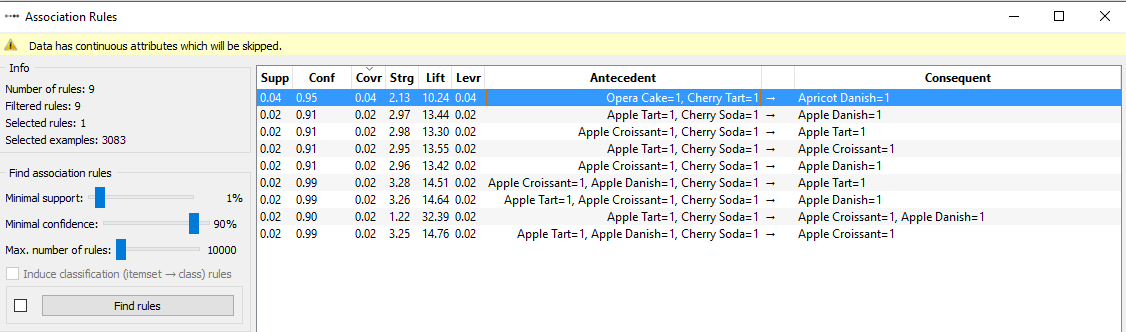
**Answer:**

**Case 1:**

When the support threshold is set to 1% and the confidence threshold is set to 90%.

Opera Cake appears along with Cherry Tart as an antecedent.

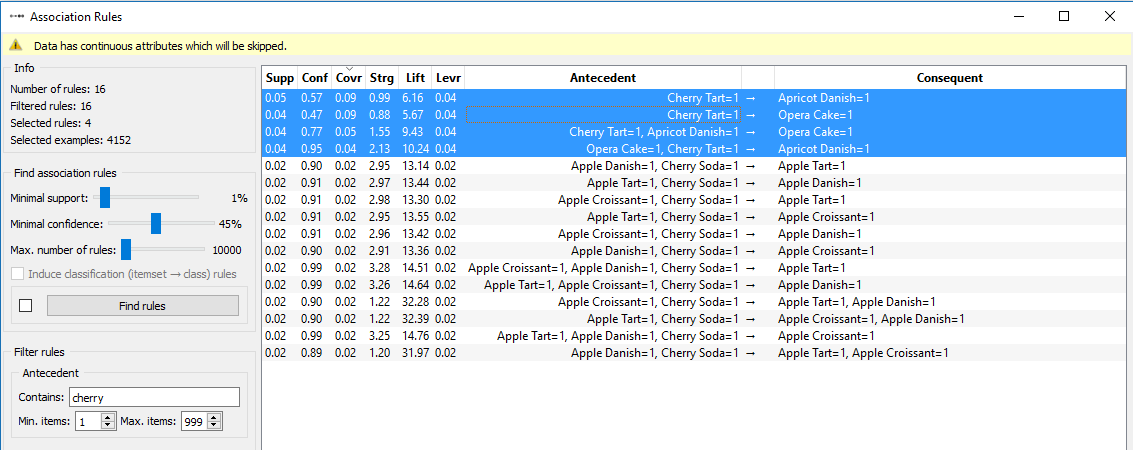
Apricot Danish is observed as consequent.

****

**Case 2:**

When the confidence threshold is lowered to 45%

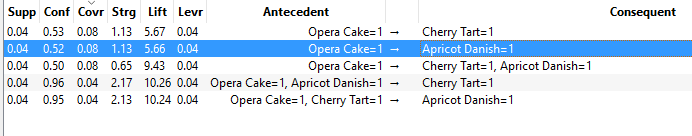
Cherry Tart appears with Opera Cake and Apricot Danish as an antecedent and twice without any other antecedents. The consequents have changed to Opera Cake and Apricot Danish.



When Opera Cake and Cherry Tart are considered as antecedents with Apricot Danish as consequent then the confidence is 95%. However, when Opera Cake to Apricot Danish is considered the confidence is 52% and for Cherry Tart to Apricot Danish we get 57% confidence. We find that the Cherry Tart, Opera Cake and Apricot Danish is positively correlated in the combined data but is negatively correlated in the stratified data.

Here the presence of hidden variables may have caused the observed relationship between Cherry Tart, Opera Cake and Apricot Danish to reverse its direction, due to the phenomenon known as Simpson's paradox.

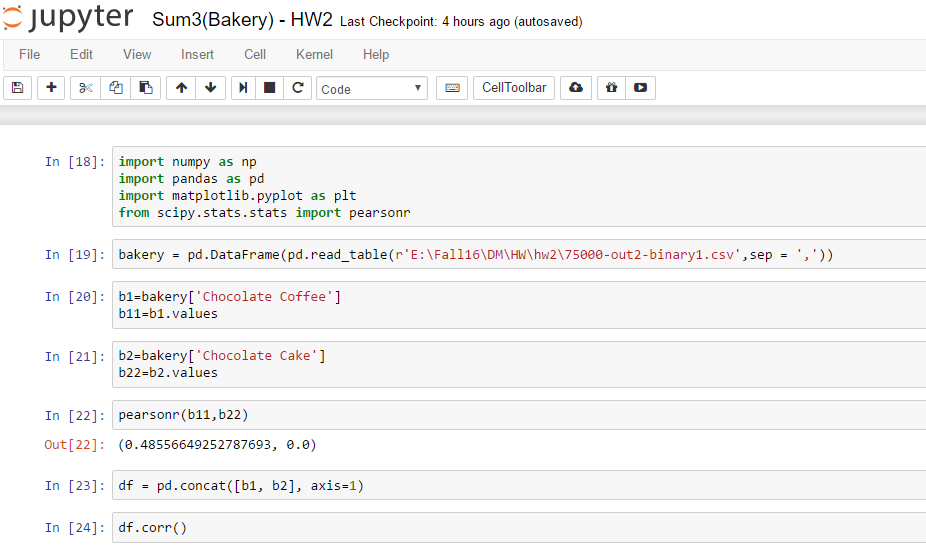
Hence, proper stratification is needed to avoid generating spurious patterns resulting from Simpson's paradox.

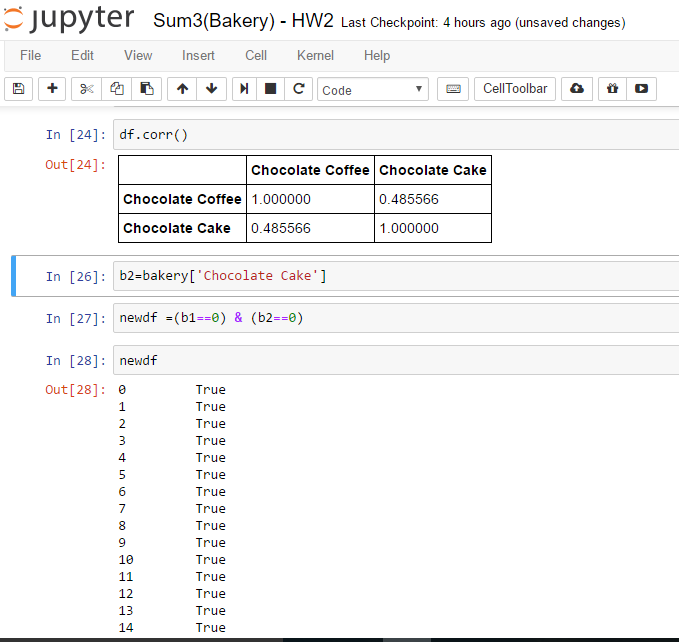
****

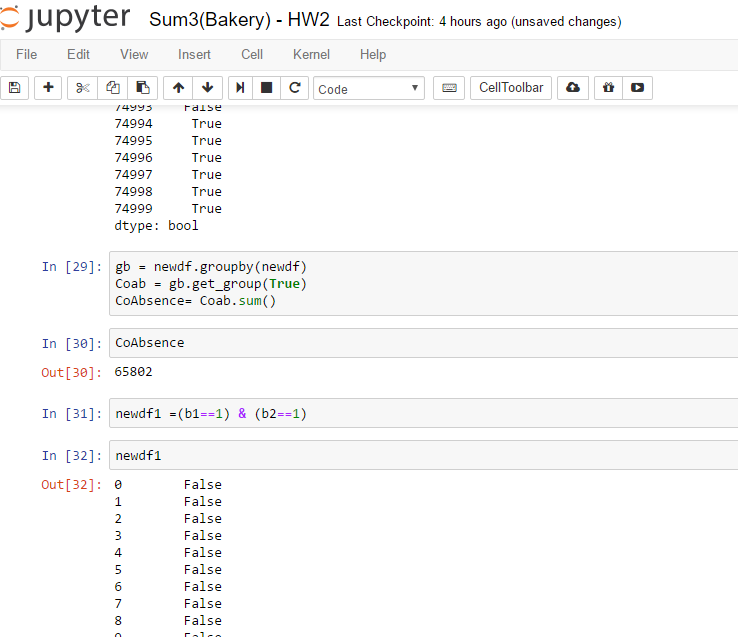
**2.3 Problem 3**

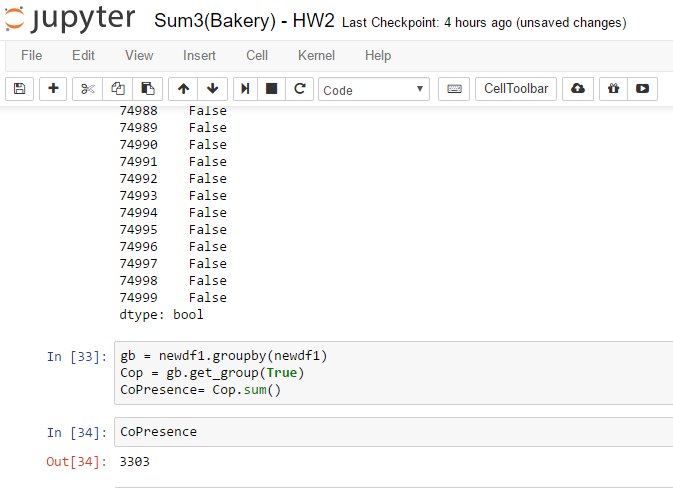
Load the Extended Bakery dataset (75000-out2-binary.csv) into Python using a Pandas dataframe. Calculate the binary correlation coefficient for the Chocolate Coffee and Chocolate Cake items. Show whether the two items are symmetric binary variables via their co-presence and co-absence. Would an association rule between these items as antecedent and consequent have a high confidence level? Why or why not?

**Answer:**

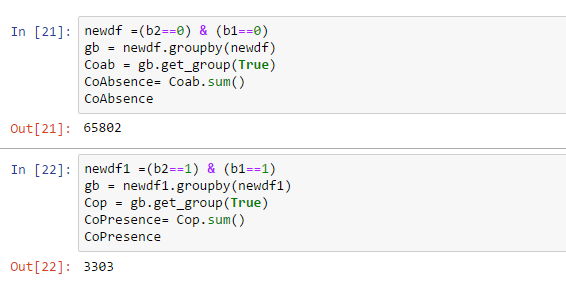
****

****

****

****

After swapping columns Chocolate Cake and Chocolate Coffee -

****

Yes, they are symmetric variables because even if the columns are swapped the correlation value and co-presence and co-absence values do not change.

No, the association rule between Chocolate Coffee and Chocolate Cake won’t have high confidence because of high value of co-absence.