CST8390 - Business Intelligence and Data Analytics

Lab 9 - *Association Rule*

Name: Min Li - Id: 040930563

Due Date: Week 12 in corresponding lab sessions.

Introduction

The goal of this lab is to perform **Association Rule Mining** on Super Market dataset.

Steps

1. Open **Weka** and load the file **supermarket.arff** from "data" directory of **Weka**.

2. From the preprocess tab, click on the Edit button to view the instances. The "**t**" letters show which items were purchased.

3. Close Edit window and look at the **attributes**.

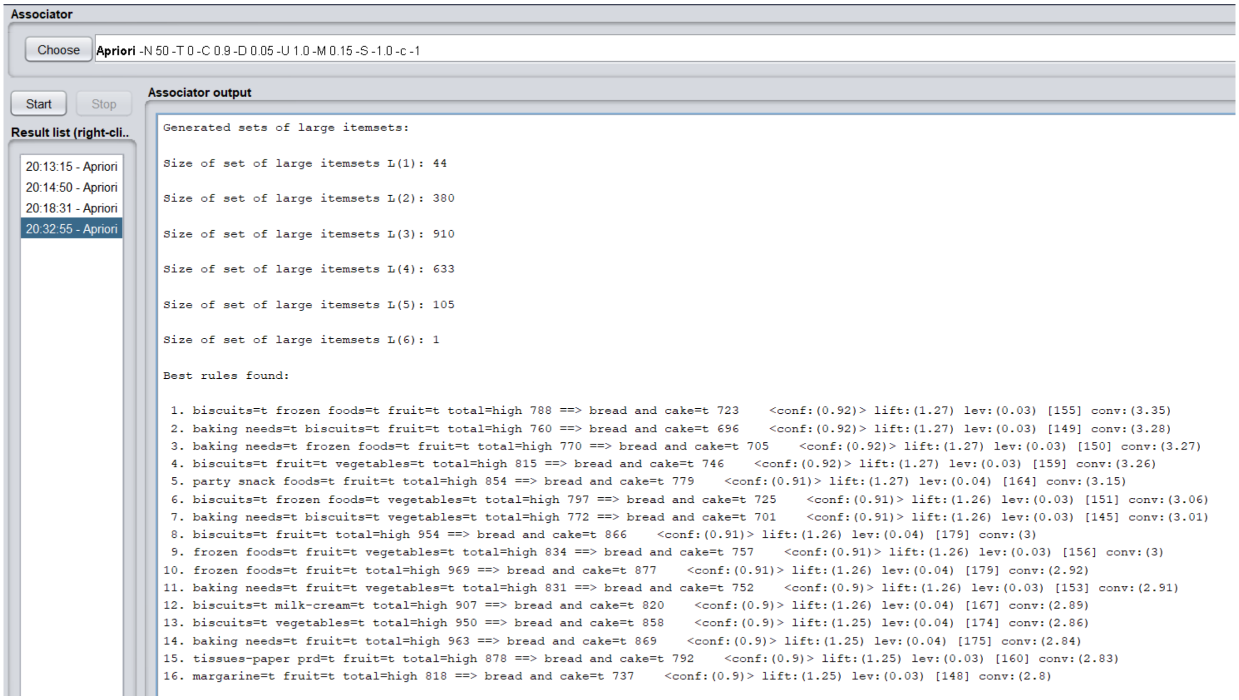
   a. Number of attributes: 217.

   b. Number of instances: 4627.

4. **Find** tea, coffee, medicines and flowers and see **how many times** each of the item was purchased?
   Tea: 896.
   Coffee: 1094.
   Medicines: 204.
   Flowers: 0.

5. Click on the "Associate" tab. The **Apriori algorithm** should already be selected but click on the **text field** to **edit** the parameters. Find the **lowerBoundMinSupport**. This is the minimum support percentage that is required to create the rule sets. Set it to 0.25 (i.e., 25%). Set the "numRules" to 15, to print out the **top 15 rules** that are found. Click "Ok" to close the window and then click "Start" to run the algorithm.

6. The algorithm should run for a number of seconds and then return with **no rules**. That means that no rules were found that have a minimum support of 20%. Lower the support to 15% and run it again. Set numRules to 50. How many **rules** were generated this time? 16.



Apriori -N 50 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.15 -S -1.0 -c -1

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 44

Size of set of large itemsets L(2): 380

Size of set of large itemsets L(3): 910

Size of set of large itemsets L(4): 633

Size of set of large itemsets L(5): 105

Size of set of large itemsets L(6): 1

Best rules found:

 1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723    <conf:(0.92)> lift:(1.27) lev:(0.03) [155] conv:(3.35)
 2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696    <conf:(0.92)> lift:(1.27) lev:(0.03) [149] conv:(3.28)
 3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705    <conf:(0.92)> lift:(1.27) lev:(0.03) [150] conv:(3.27)
 4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746    <conf:(0.92)> lift:(1.27) lev:(0.03) [159] conv:(3.26)
 5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779    <conf:(0.91)> lift:(1.27) lev:(0.04) [164] conv:(3.15)
 6. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725    <conf:(0.91)> lift:(1.26) lev:(0.03) [151] conv:(3.06)
 7. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701    <conf:(0.91)> lift:(1.26) lev:(0.03) [145] conv:(3.01)
 8. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866    <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(3)
 9. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757    <conf:(0.91)> lift:(1.26) lev:(0.03) [156] conv:(3)
10. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877    <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(2.92)
11. baking needs=t fruit=t vegetables=t total=high 831 ==> bread and cake=t 752    <conf:(0.9)> lift:(1.26) lev:(0.03) [153] conv:(2.91)
12. biscuits=t milk-cream=t total=high 907 ==> bread and cake=t 820    <conf:(0.9)> lift:(1.26) lev:(0.04) [167] conv:(2.89)
13. biscuits=t vegetables=t total=high 950 ==> bread and cake=t 858    <conf:(0.9)> lift:(1.25) lev:(0.04) [174] conv:(2.86)
14. baking needs=t fruit=t total=high 963 ==> bread and cake=t 869    <conf:(0.9)> lift:(1.25) lev:(0.04) [175] conv:(2.84)
15. tissues-paper prd=t fruit=t total=high 878 ==> bread and cake=t 792    <conf:(0.9)> lift:(1.25) lev:(0.03) [160] conv:(2.83)
16. margarine=t fruit=t total=high 818 ==> bread and cake=t 737    <conf:(0.9)> lift:(1.25) lev:(0.03) [148] conv:(2.8)
```

7. The rules are **sorted** from highest lift to lowest. The lift tells you **how often** the rules are related, or the strength of the rule. Which rules have the **highest lift**? Rule1,2,3,4,5.

1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 <conf:(0.92)> lift:(1.27) lev:(0.03) [155] conv:(3.35)

2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696 <conf:(0.92)> lift:(1.27) lev:(0.03) [149] conv:(3.28)

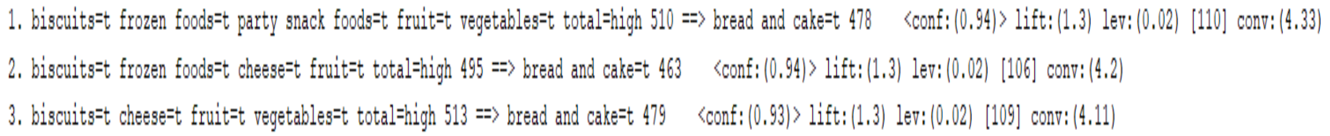
3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705 <conf:(0.92)> lift:(1.27) lev:(0.03) [150] conv:(3.27)

4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746 <conf:(0.92)> lift:(1.27) lev:(0.03) [159] conv:(3.26)

5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779 <conf:(0.91)> lift:(1.27) lev:(0.04) [164] conv:(3.15)

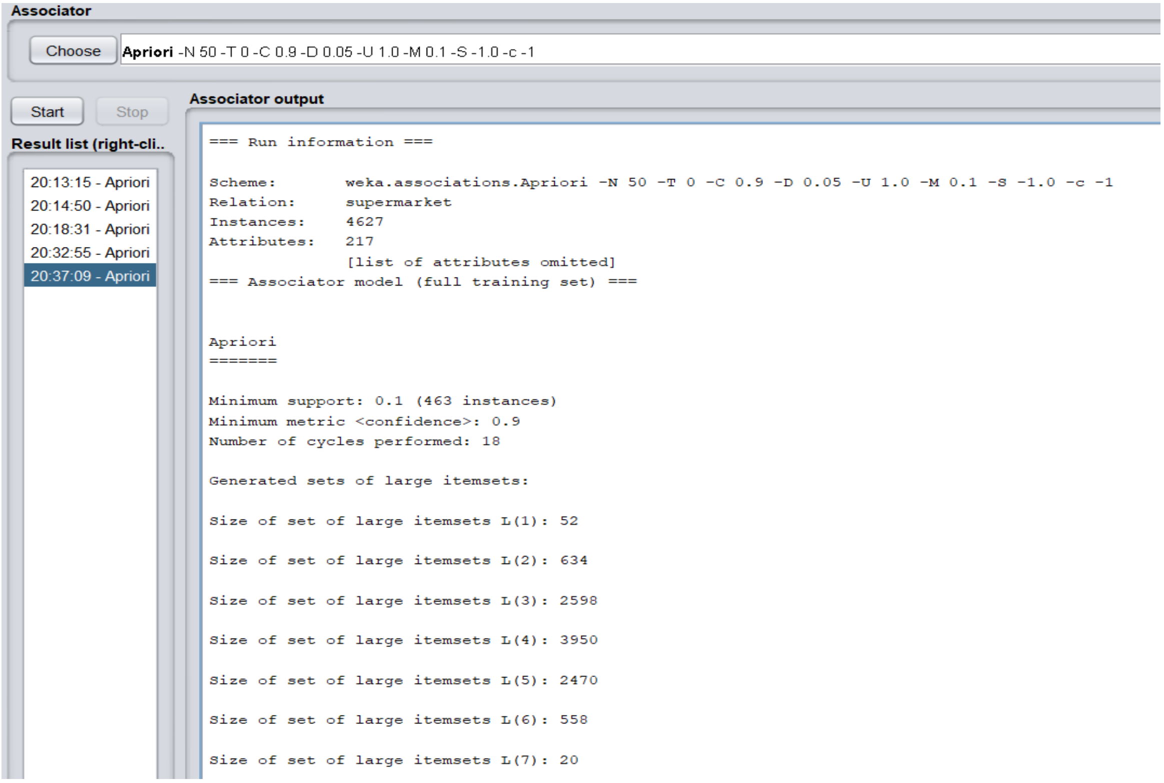
8. Lower the support now to 10% and re-run the algorithm. Since more rules are included in the search, this time it should take a long time to run. What is the **highest lift** now that was found and what are the rules? The highest lift is 1.3. Rule1,2,3.

Example:
If you get:
frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(2.92),
you need to writer the rule as:
frozen foods, fruit ==> bread and cake (conf: 0.91, lift: 1.26).



```
1. biscuits=t frozen foods=t party snack foods=t fruit=t vegetables=t total=high 510 ==> bread and cake=t 478    <conf:(0.94)> lift:(1.3) lev:(0.02) [110] conv:(4.33)

2. biscuits=t frozen foods=t cheese=t fruit=t total=high 495 ==> bread and cake=t 463    <conf:(0.94)> lift:(1.3) lev:(0.02) [106] conv:(4.2)

3. biscuits=t cheese=t fruit=t vegetables=t total=high 513 ==> bread and cake=t 479    <conf:(0.93)> lift:(1.3) lev:(0.02) [109] conv:(4.11)
```

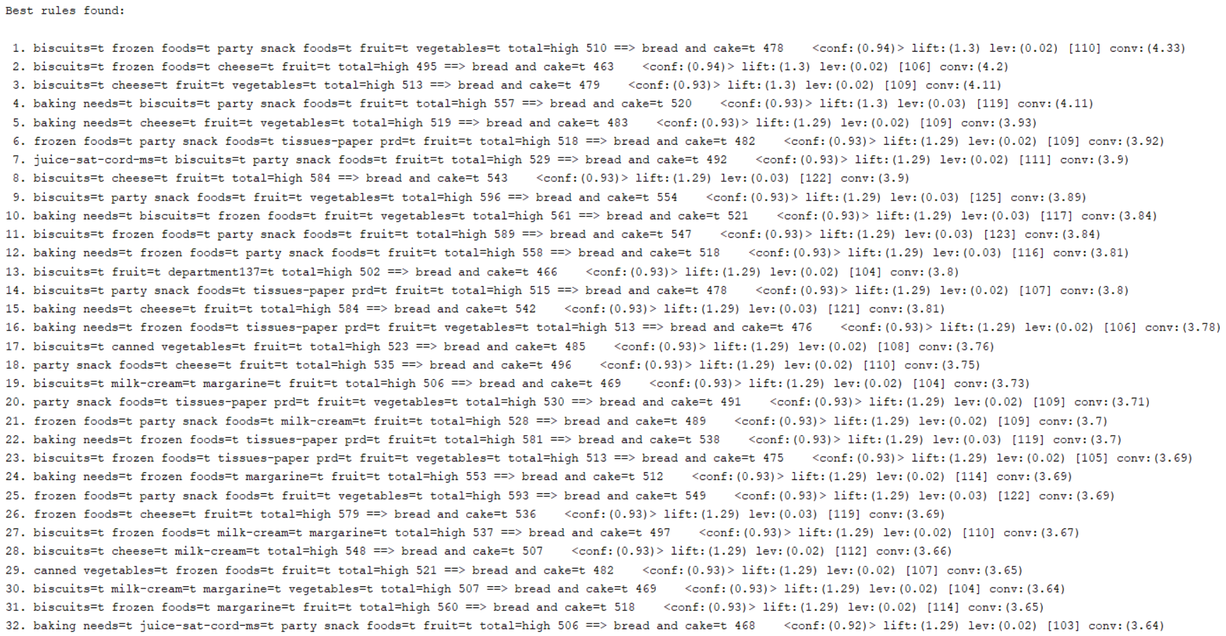
1. biscuits, frozen foods, party snack foods, fruit, vegetables ==> bread and cake (conf: 0.94, lift:1.3).
2. biscuits, frozen foods, cheese, fruit ==> bread and cake (conf: 0.94, lift:1.3).
3. biscuits, cheese, fruit, vegetables ==> bread and cake (conf: 0.93, lift:1.3).



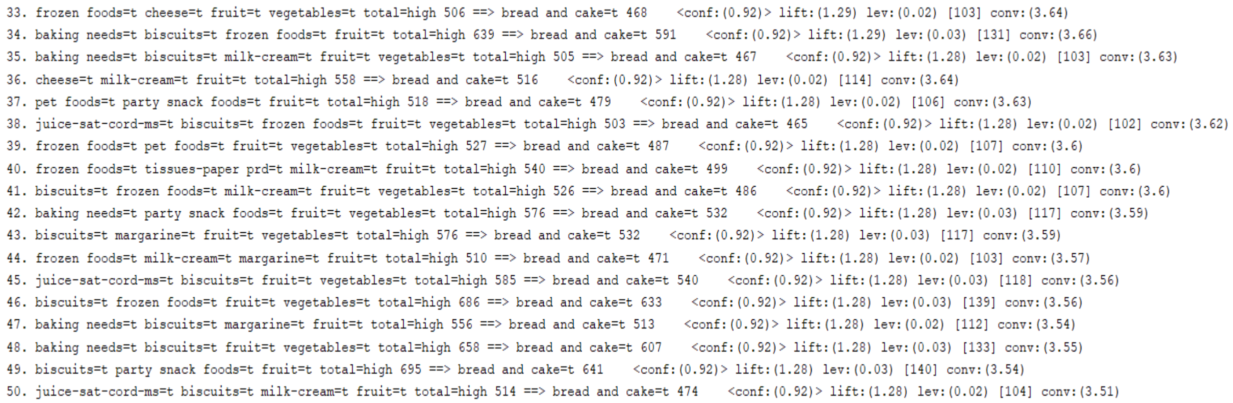
Apriori -N 50 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

20:13:15 - Apriori
20:14:50 - Apriori
20:18:31 - Apriori
20:32:55 - Apriori
20:37:09 - Apriori

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 50 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     supermarket
Instances:    4627
Attributes:   217
              [list of attributes omitted]
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (463 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 52

Size of set of large itemsets L(2): 634

Size of set of large itemsets L(3): 2598

Size of set of large itemsets L(4): 3950

Size of set of large itemsets L(5): 2470

Size of set of large itemsets L(6): 558

Size of set of large itemsets L(7): 20
```



```
Best rules found:

 1. biscuits=t frozen foods=t party snack foods=t fruit=t vegetables=t total=high 510 ==> bread and cake=t 478    <conf:(0.94)> lift:(1.3) lev:(0.02) [110] conv:(4.33)
 2. biscuits=t frozen foods=t cheese=t fruit=t total=high 495 ==> bread and cake=t 463    <conf:(0.94)> lift:(1.3) lev:(0.02) [106] conv:(4.2)
 3. biscuits=t cheese=t fruit=t vegetables=t total=high 513 ==> bread and cake=t 479    <conf:(0.93)> lift:(1.3) lev:(0.02) [109] conv:(4.11)
 4. baking needs=t biscuits=t party snack foods=t fruit=t total=high 557 ==> bread and cake=t 520    <conf:(0.93)> lift:(1.3) lev:(0.03) [119] conv:(4.11)
 5. baking needs=t cheese=t fruit=t vegetables=t total=high 519 ==> bread and cake=t 483    <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.93)
 6. frozen foods=t party snack foods=t tissues-paper prd=t fruit=t total=high 518 ==> bread and cake=t 482    <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.92)
 7. juice-sat-cord-ms=t biscuits=t party snack foods=t fruit=t total=high 529 ==> bread and cake=t 492    <conf:(0.93)> lift:(1.29) lev:(0.02) [111] conv:(3.9)
 8. biscuits=t cheese=t fruit=t total=high 584 ==> bread and cake=t 543    <conf:(0.93)> lift:(1.29) lev:(0.03) [122] conv:(3.9)
 9. biscuits=t party snack foods=t fruit=t vegetables=t total=high 596 ==> bread and cake=t 554    <conf:(0.93)> lift:(1.29) lev:(0.03) [125] conv:(3.89)
10. baking needs=t biscuits=t frozen foods=t fruit=t vegetables=t total=high 561 ==> bread and cake=t 521    <conf:(0.93)> lift:(1.29) lev:(0.03) [117] conv:(3.84)
11. biscuits=t frozen foods=t party snack foods=t fruit=t total=high 589 ==> bread and cake=t 547    <conf:(0.93)> lift:(1.29) lev:(0.03) [123] conv:(3.84)
12. baking needs=t frozen foods=t party snack foods=t fruit=t total=high 558 ==> bread and cake=t 518    <conf:(0.93)> lift:(1.29) lev:(0.03) [116] conv:(3.81)
13. biscuits=t fruit=t department137=t total=high 502 ==> bread and cake=t 466    <conf:(0.93)> lift:(1.29) lev:(0.02) [104] conv:(3.8)
14. biscuits=t party snack foods=t tissues-paper prd=t fruit=t total=high 515 ==> bread and cake=t 478    <conf:(0.93)> lift:(1.29) lev:(0.02) [107] conv:(3.8)
15. baking needs=t cheese=t fruit=t total=high 584 ==> bread and cake=t 542    <conf:(0.93)> lift:(1.29) lev:(0.03) [121] conv:(3.81)
16. baking needs=t frozen foods=t tissues-paper prd=t fruit=t vegetables=t total=high 513 ==> bread and cake=t 476    <conf:(0.93)> lift:(1.29) lev:(0.02) [106] conv:(3.78)
17. biscuits=t canned vegetables=t fruit=t total=high 523 ==> bread and cake=t 485    <conf:(0.93)> lift:(1.29) lev:(0.02) [108] conv:(3.76)
18. party snack foods=t cheese=t fruit=t total=high 535 ==> bread and cake=t 496    <conf:(0.93)> lift:(1.29) lev:(0.02) [110] conv:(3.75)
19. biscuits=t milk-cream=t margarine=t fruit=t total=high 506 ==> bread and cake=t 469    <conf:(0.93)> lift:(1.29) lev:(0.02) [104] conv:(3.73)
20. party snack foods=t tissues-paper prd=t fruit=t vegetables=t total=high 530 ==> bread and cake=t 491    <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.71)
21. frozen foods=t party snack foods=t milk-cream=t fruit=t total=high 528 ==> bread and cake=t 489    <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.7)
22. baking needs=t frozen foods=t tissues-paper prd=t fruit=t total=high 581 ==> bread and cake=t 538    <conf:(0.93)> lift:(1.29) lev:(0.03) [119] conv:(3.7)
23. biscuits=t frozen foods=t tissues-paper prd=t fruit=t vegetables=t total=high 513 ==> bread and cake=t 475    <conf:(0.93)> lift:(1.29) lev:(0.02) [105] conv:(3.69)
24. baking needs=t frozen foods=t margarine=t fruit=t total=high 553 ==> bread and cake=t 512    <conf:(0.93)> lift:(1.29) lev:(0.02) [114] conv:(3.69)
25. frozen foods=t party snack foods=t fruit=t vegetables=t total=high 593 ==> bread and cake=t 549    <conf:(0.93)> lift:(1.29) lev:(0.03) [122] conv:(3.69)
26. frozen foods=t cheese=t fruit=t total=high 579 ==> bread and cake=t 536    <conf:(0.93)> lift:(1.29) lev:(0.03) [119] conv:(3.69)
27. biscuits=t frozen foods=t milk-cream=t margarine=t total=high 537 ==> bread and cake=t 497    <conf:(0.93)> lift:(1.29) lev:(0.02) [110] conv:(3.67)
28. biscuits=t cheese=t milk-cream=t total=high 548 ==> bread and cake=t 507    <conf:(0.93)> lift:(1.29) lev:(0.02) [112] conv:(3.66)
29. canned vegetables=t frozen foods=t fruit=t total=high 521 ==> bread and cake=t 482    <conf:(0.93)> lift:(1.29) lev:(0.02) [107] conv:(3.65)
30. biscuits=t milk-cream=t margarine=t vegetables=t total=high 507 ==> bread and cake=t 469    <conf:(0.93)> lift:(1.29) lev:(0.02) [104] conv:(3.64)
31. biscuits=t frozen foods=t margarine=t fruit=t total=high 560 ==> bread and cake=t 518    <conf:(0.93)> lift:(1.29) lev:(0.02) [114] conv:(3.65)
32. baking needs=t juice-sat-cord-ms=t party snack foods=t fruit=t total=high 506 ==> bread and cake=t 468    <conf:(0.92)> lift:(1.29) lev:(0.02) [103] conv:(3.64)
```



```
33. frozen foods=t cheese=t fruit=t vegetables=t total=high 506 ==> bread and cake=t 468    <conf:(0.92)> lift:(1.29) lev:(0.02) [103] conv:(3.64)
34. baking needs=t biscuits=t frozen foods=t fruit=t total=high 639 ==> bread and cake=t 591    <conf:(0.92)> lift:(1.29) lev:(0.03) [131] conv:(3.66)
35. baking needs=t biscuits=t milk-cream=t fruit=t vegetables=t total=high 505 ==> bread and cake=t 467    <conf:(0.92)> lift:(1.28) lev:(0.02) [103] conv:(3.63)
36. cheese=t milk-cream=t fruit=t total=high 558 ==> bread and cake=t 516    <conf:(0.92)> lift:(1.28) lev:(0.02) [114] conv:(3.64)
37. pet foods=t party snack foods=t fruit=t total=high 518 ==> bread and cake=t 479    <conf:(0.92)> lift:(1.28) lev:(0.02) [106] conv:(3.63)
38. juice-sat-cord-ms=t biscuits=t frozen foods=t fruit=t vegetables=t total=high 503 ==> bread and cake=t 465    <conf:(0.92)> lift:(1.28) lev:(0.02) [102] conv:(3.62)
39. frozen foods=t pet foods=t fruit=t vegetables=t total=high 527 ==> bread and cake=t 487    <conf:(0.92)> lift:(1.28) lev:(0.02) [107] conv:(3.6)
40. frozen foods=t tissues-paper prd=t milk-cream=t fruit=t total=high 540 ==> bread and cake=t 499    <conf:(0.92)> lift:(1.28) lev:(0.02) [110] conv:(3.6)
41. biscuits=t frozen foods=t milk-cream=t fruit=t vegetables=t total=high 526 ==> bread and cake=t 486    <conf:(0.92)> lift:(1.28) lev:(0.02) [107] conv:(3.6)
42. baking needs=t party snack foods=t fruit=t vegetables=t total=high 576 ==> bread and cake=t 532    <conf:(0.92)> lift:(1.28) lev:(0.03) [117] conv:(3.59)
43. biscuits=t margarine=t fruit=t vegetables=t total=high 576 ==> bread and cake=t 532    <conf:(0.92)> lift:(1.28) lev:(0.03) [117] conv:(3.59)
44. frozen foods=t milk-cream=t margarine=t fruit=t total=high 510 ==> bread and cake=t 471    <conf:(0.92)> lift:(1.28) lev:(0.02) [103] conv:(3.57)
45. juice-sat-cord-ms=t biscuits=t fruit=t vegetables=t total=high 585 ==> bread and cake=t 540    <conf:(0.92)> lift:(1.28) lev:(0.03) [118] conv:(3.56)
46. biscuits=t frozen foods=t fruit=t vegetables=t total=high 686 ==> bread and cake=t 633    <conf:(0.92)> lift:(1.28) lev:(0.03) [139] conv:(3.56)
47. baking needs=t biscuits=t margarine=t fruit=t total=high 556 ==> bread and cake=t 513    <conf:(0.92)> lift:(1.28) lev:(0.02) [112] conv:(3.54)
48. baking needs=t biscuits=t fruit=t vegetables=t total=high 658 ==> bread and cake=t 607    <conf:(0.92)> lift:(1.28) lev:(0.03) [133] conv:(3.55)
49. biscuits=t party snack foods=t fruit=t total=high 695 ==> bread and cake=t 641    <conf:(0.92)> lift:(1.28) lev:(0.03) [140] conv:(3.54)
50. juice-sat-cord-ms=t biscuits=t milk-cream=t fruit=t total=high 514 ==> bread and cake=t 474    <conf:(0.92)> lift:(1.28) lev:(0.02) [104] conv:(3.51)
```

REMEMBER:

In order to get grades, you need to upload filled-in answer document and screenshots from steps 6-8.

FOR YOUR ANALYSIS:

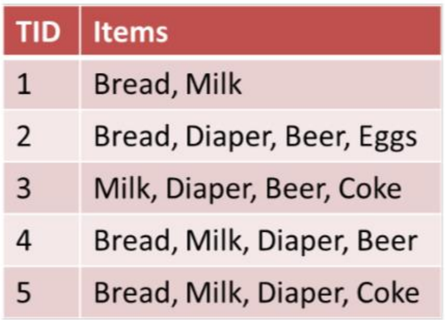
**** Option 1**: Use your own words to explain **Association Rule Mining** and situations where apply it.*

**** Option 2**: What is the strategy to identify association rules in a specific scenario?*

Option 1: Association Rule Mining is one of the ways to find patterns in data. It finds: 1. features (dimensions) which occur together 2. features (dimensions) which are "correlated". What does the value of one feature tell us about the value of another feature? For example, people who buy diapers are likely to buy baby powder. Or we can rephrase the statement by saying: If (people buy diaper), then (they buy baby powder). Note the if, then rule. This does not necessarily mean that if people buy baby powder, they buy diaper. In General, we can say that if condition A tends to B it does not necessarily mean that B tends to A. Watch the directionality!

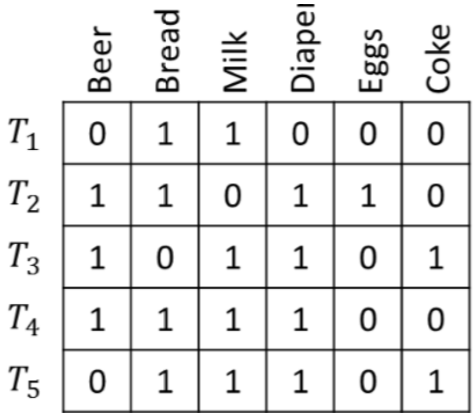
We can use Association Rules in any dataset where features take only two values i.e., 0/1. Some examples are here: 1. **Market Basket Analysis** is a popular application of Association Rules. 2. People who visit webpage X are likely to visit webpage Y. 3. People who have age-group [30,40] & income [>$100k] are likely to own home.

The measures of effectiveness of the rule are as Follows: 1. Support 2. Confidence 3. Lift 4. Others: Affinity, Leverage. We are going to discuss Support and Confidence in more detail using an example dataset.



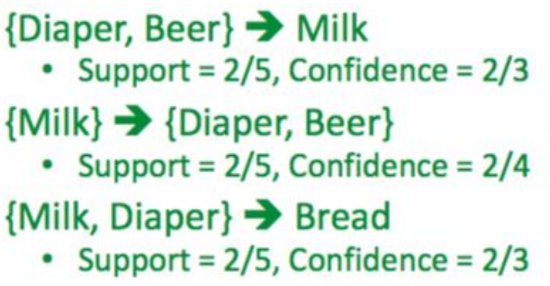
TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

The above dataset can also be represented like this:



	Beer	Bread	Milk	Diapei	Eggs	Coke
T_1	0	1	1	0	0	0
T_2	1	1	0	1	1	0
T_3	1	0	1	1	0	1
T_4	1	1	1	1	0	0
T_5	0	1	1	1	0	1

Support means how much historical data supports your rule and **Confidence** means how confident are we that the rule holds. Support can be calculated as the fraction of rows containing both A and B or joint probability of A and B. Among rows containing A, Confidence is the fraction of rows containing B or conditional probability of B given A.



{Diaper, Beer} ➔ Milk

- Support = 2/5, Confidence = 2/3

{Milk} ➔ {Diaper, Beer}

- Support = 2/5, Confidence = 2/4

{Milk, Diaper} ➔ Bread

- Support = 2/5, Confidence = 2/3

Lift is the ratio Confidence is to Support. If the lift is < 1 then A and B are negatively correlated else positively correlated and if it is equal to 1 it is not correlated.

Reference: https://towardsdatascience.com/association-rule-mining-be4122fc1793