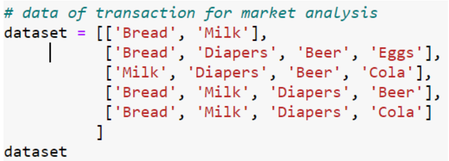
Section 5.10 - 1, 2, 3, 4, 6, 7, 8

1. **For each of the following questions, provide an example of an association rule from the market basket domain that satisfies the following conditions. Also, describe whether such rules are subjectively Interesting.**

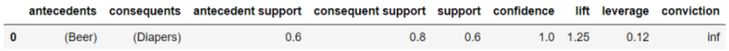
   Data set used to solve this problem:



```
# data of transaction for market analysis
dataset = [['Bread', 'Milk'],
           ['Bread', 'Diapers', 'Beer', 'Eggs'],
           ['Milk', 'Diapers', 'Beer', 'Cola'],
           ['Bread', 'Milk', 'Diapers', 'Beer'],
           ['Bread', 'Milk', 'Diapers', 'Cola']
          ]
dataset
```

a) A rule that has high support and high confidence.

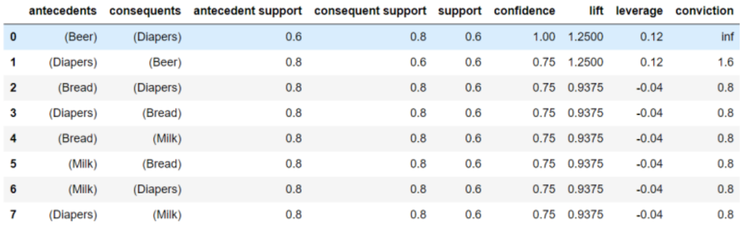
Answer: With minimum support (60%), we are left with one set {bear, diaper, milk} and after considering the threshold confidence (75%), we see an interesting pattern that beer and diapers are bought together.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Beer)	(Diapers)	0.6	0.8	0.6	1.0	1.25	0.12	inf

b) A rule that has reasonably high support but low confidence.

Answer: Here, because of high support we got same set as above {bear, diaper, milk}. With threshold confidence (25%) we see the following patterns.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Beer)	(Diapers)	0.6	0.8	0.6	1.00	1.2500	0.12	inf
1	(Diapers)	(Beer)	0.8	0.6	0.6	0.75	1.2500	0.12	1.6
2	(Bread)	(Diapers)	0.8	0.8	0.6	0.75	0.9375	-0.04	0.8
3	(Diapers)	(Bread)	0.8	0.8	0.6	0.75	0.9375	-0.04	0.8
4	(Bread)	(Milk)	0.8	0.8	0.6	0.75	0.9375	-0.04	0.8
5	(Milk)	(Bread)	0.8	0.8	0.6	0.75	0.9375	-0.04	0.8
6	(Milk)	(Diapers)	0.8	0.8	0.6	0.75	0.9375	-0.04	0.8
7	(Diapers)	(Milk)	0.8	0.8	0.6	0.75	0.9375	-0.04	0.8

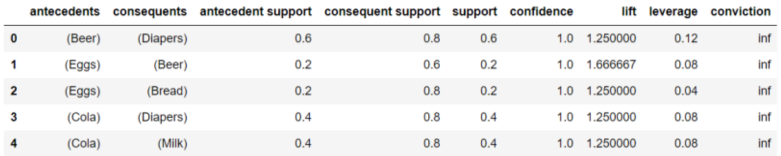
This rule seems uninteresting.

c) A rule that has low support and low confidence.

Answer: Here because of support (20%) and confidence (25%) there are all possible combinations of item sets and all possibilities are available to buy. We don't see any interesting pattern.

d) A rule that has low support and high confidence.

Answer: with support (20%) and confidence (75%), we got the following dataset.



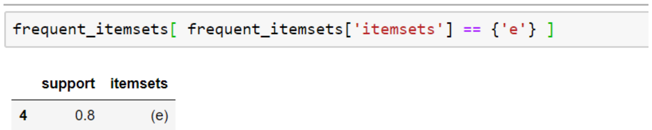
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Beer)	(Diapers)	0.6	0.8	0.6	1.0	1.250000	0.12	inf
1	(Eggs)	(Beer)	0.2	0.6	0.2	1.0	1.666667	0.08	inf
2	(Eggs)	(Bread)	0.2	0.8	0.2	1.0	1.250000	0.04	inf
3	(Cola)	(Diapers)	0.4	0.8	0.4	1.0	1.250000	0.08	inf
4	(Cola)	(Milk)	0.4	0.8	0.4	1.0	1.250000	0.08	inf

Here we see that eggs-> beer can be interesting because it is common to purchase eggs and bread together.

2. **Consider the data set shown in Table 5.20.**

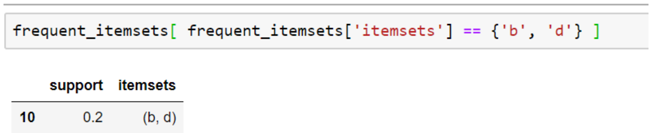
a. Compute the support for item sets{e}, {b, d}, and {b, d, e} by treating each transaction ID as a market basket.

Answer:



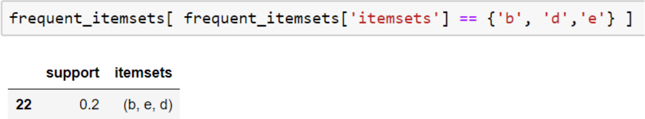
```
frequent_itemsets[ frequent_itemsets['itemsets'] == {'e'} ]
```

	support	itemsets
4	0.8	(e)



```
frequent_itemsets[ frequent_itemsets['itemsets'] == {'b', 'd'} ]
```

	support	itemsets
10	0.2	(b, d)



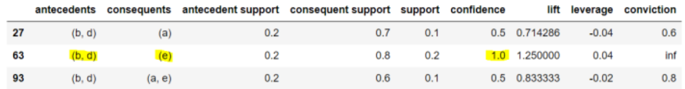
```
frequent_itemsets[ frequent_itemsets['itemsets'] == {'b', 'd','e'} ]
```

	support	itemsets
22	0.2	(b, e, d)

b. **Use the results in part (a) to compute the confidence for the association rules {b, d} →{e} and {e} → {b, d}. Is confidence asymmetric measure?**

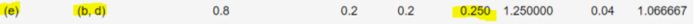
Answer:

Confidence for the association rule {b, d} -> {e}



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
27	(b, d)	(a)	0.2	0.7	0.1	0.5	0.714286	-0.04	0.6
63	(b, d)	(e)	0.2	0.8	0.2	1.0	1.250000	0.04	inf
93	(b, d)	(a, e)	0.2	0.6	0.1	0.5	0.833333	-0.02	0.8

Confidence for the association rule {e} → {b, d}



(e)	(b, d)	0.8	0.2	0.2	0.250	1.250000	0.04	1.066667

No, confidence is not a symmetric measure.

c. **Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise).**

Table – Customer id as transaction and item bought are a, b, c, d, e

Customer id	Item a	b	c	d	e
1	1	1	1	1	1
2	1	1	1	1	1
3	0	1	1	1	1
4	1	1	1	1	0
5	1	1	0	1	1

support for item sets:

S ({e}) = 4/5 = 0.8
S ({b, d}) = 5/5 = 1
S ({b, d, e}) = 4/5 =0.8

d. **Use the results in part (c) to compute the confidence for the association rules {b, d} →{e} and {e} → {b, d}.**

Confidence ({b, d} →{e}) = Support ({b, d}->{e})/ Support {b, d}
= 0.8/1 = 80%

Confidence $(\{e\} \rightarrow \{b, d\}) = 0.8/0.8 = 100\%$

e. **Suppose s1 and c1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket. Also, let s2 and c2 be the support and confidence values of r when treating each customer ID as a market basket. Discuss whether there are any relationships between s1 and s2 or c1 and c2.**

Answer: Although support for {e} remained the same, nothing can be said about support for {b, d} and {b, d, e} (except that it increased significantly by using Customer ID). The increase in support is not reflected in the changes in confidence of the rules. This means that in general, no clear difference in treating transaction IDs or customer IDs as market baskets.

3.

a. **What is the confidence for the rules Ø→A and A→Ø?**

Confidence(A->C) = Support (A ->C)/Support(A)

$C (\emptyset \rightarrow A) = S(\emptyset \cup A)/S (\emptyset) = s(A)$
$C (A \rightarrow \emptyset) = S(\emptyset \cup A)/S(A) = S(A)/S(A) = 1$

b. **Let c1, c2, and c3 be the confidence values of the rules {p}→{q}, {p} → {q, r}, and {p, r} → {q}, respectively. If we assume that c1, c2, and c3 have different values, what are the possible relationships that may exist among c1, c2, and c3? Which rule has the lowest confidence?**

confidence values of the rule {p}→{q}:	$c1= S (p \cup q)/S(p)$
confidence values of the rule {p} → {q, r}:	$c2 = S (p \cup q \cup r)/S(p)$
confidence values of the rule {p, r} → {q}:	$c3 = S (p \cup q \cup r)/S(q)$

Since $S(pq) \geq S(pqr)$, we can say that $c1 \geq c2$ by looking at the denominators. Similarly, since $S(p) \geq S(pr)$, we can say that $c3 \geq c2$. Thus, the rule {p} → {q, r} has the lowest confidence (c2).

c. **Repeat the analysis in part (b) assuming that the rules have identical support. Which rule has the highest confidence?**

In this case: $S(pq) = S(pqr)$, which leads to $c1 = c2$. As we still have $S(p) \geq S(pr)$, we can say that $c3 \geq c1$ and $c3 \geq c2$.

d. **Transitivity: Suppose the confidence of the rules A→B and B→C are larger than some threshold, minconf. Is it possible that A→C has a confidence less than minconf?**

Yes, it depends on the support of items. For Example:

Support (A, B) = 60%	Support (A) = 90%
Support (A, C) = 20%	Support (B) = 70%
Support (B, C) = 50%	Support (C) = 60%

Let's assume minconf= 50% Therefore:

C(A->B) = 6/9 = 66 % > minconf
C(B->C) = 5/7 = 71 % > minconf
C(A->C) = 2/9 = 22% < minconf (Yes)

4. For each of the following measures, determine whether it is monotone, anti-monotone, or non-monotone (i.e., neither monotone nor anti-monotone).
Example: Support, s=σ(x)|T| is anti-monotone because s(X)≥s(Y) whenever X⊂Y.

a. **A characteristic rule is a rule of the form {p} → {q1, q2, …, qn}, where the rule antecedent contains only a single item. An itemset of size *k* can produce up to *k* characteristic rules. Let ζ be the minimum confidence of all characteristic rules generated from a given itemset: ζ({p1, p2, …, pk})=min[c({p1}→{p2, p3, …, pk}), …c({pk}→{p1, p2, …, pk−1})]**
Is ζ monotone, anti-monotone, or non-monotone?

Answer:
ζ is an anti-monotone measure because, ζ ({A1, A2, · · ·, Ak}) ≥ ζ ({A1, A2, · · ·, Ak, Ak+1})
For example, we can compare the values of ζ for {A, B} and {A, B, C}.

Z ({A, B}) = min (c (A → B), c (B → A))
= min (s (A, B)/s(A), s (A, B)/s(B))
= s (A, B)/max(s(A), s(B))
Z ({A, B, C}) = min (c (A → BC), c (B → AC), c (C → AB))
= min (s (A, B, C)/s(A), s (A, B, C)/s(B), s (A, B, C)/s(C))
= s (A, B, C)/max(s(A), s(B), s(C))

Since, s (A, B, C) ≤ s (A, B) and max(s(A), s(B), s(C)) ≥ max(s(A), s(B)). Therefore: ζ ({A, B}) ≥ ζ ({A, B, C}).

b. **A discriminant rule is a rule of the form {p1, p2, …, pn} → {q}, where the rule consequent contains only a single item. An itemset of size k can produce up to k discriminant rules. Let η be the minimum confidence of all discriminant rules generated from a given itemset:**
η ({p1, p2, …, pk}) =min [c ({p2, p3, …, pk} → {p1}), …c ({p1, p2, …, pk−1} →{pk})]
Is η monotone, anti-monotone, or non-monotone?

Answer:
η is non-monotone. We can show this by comparing η ({A, B}) against η ({A, B, C}).

η ({A, B}) = min (c (A → B), c (B → A))
= min (s (A, B)/s(A), s (A, B)/s(B))
= s (A, B)/max(s(A), s(B))
η ({A, B, C}) = min (c (AB → C), c (AC → B), c (BC → A))
= min (s (A, B, C)/s (A, B), s (A, B, C)/s (A, C), s (A, B, C)/s (B, C))
= s (A, B, C)/max (s (A, B), s (A, C), s (B, C))

Since, s (A, B, C) ≤ s (A, B) and max (s (A, B), s (A, C), s (B, C)) ≤ max(s(A), s(B)).
Therefore, η ({A, B, C}) can be greater than or less than η ({A, B}). The measure is non-monotone.

c. **Repeat the analysis in parts (a) and (b) by replacing the min function with a max function.**
Answer:
Let ζ' ({A1, A2, · · ·, Ak}) = max (c (A1 → A2, A3, · · ·, Ak), · · · c (Ak → A1, A3 · · ·, Ak−1))

ζ' ({A, B}) = max (c (A → B), c (B → A))
= max (s (A, B)/s(A), s (A, B)/s(B))
= s (A, B)/min(s(A), s(B))

ζ' ({A, B, C}) = max (c (A → BC), c (B → AC), c (C → AB))
= max (s (A, B, C)/s(A), s (A, B, C)/s(B), s (A, B, C)/s(C))
= s (A, B, C)/min(s(A), s(B), s(C))

Since $s(A,B,C) \le s(A,B)$ and $\min(s(A), s(B), s(C)) \le \min(s(A), s(B))$, $\zeta'\_(\{A,B,C\})$ can be greater than or less than $\zeta'\_(\{A,B\})$. Therefore, the measure is non-monotone.

Let $\eta'(\{A1, A2, \cdots, Ak\}) = \max(c(A2, A3, \cdots, Ak \longrightarrow A1), \cdots c(A1, A2, \cdots Ak-1 \longrightarrow Ak))$

$$\eta'(\{A, B\}) = \max(c(A \rightarrow B), c(B \rightarrow A))$$
$$= \max(s(A, B)/s(A), s(A, B)/s(B))$$
$$= s(A, B)/\min(s(A), s(B))$$
$$\eta'(\{A, B, C\}) = \max(c(AB \rightarrow C), c(AC \rightarrow B), c(BC \rightarrow A))$$
$$= \max(s(A, B, C)/s(A, B), s(A, B, C)/s(A, C), s(A, B, C)/s(B, C))$$
$$= s(A, B, C)/\min(s(A, B), s(A, C), s(B, C))$$

Since $s(A,B,C) \le s(A,B)$ and $\min(s(A,B), s(A,C), s(B,C)) \le \min(s(A), s(B), s(C)) \le \min(s(A), s(B))$, $\eta\_(\{A,B,C\})$ can be greater than or less than $\eta\_(\{A,B\})$. Therefore, the measure is non-monotone.

6. **Consider the market basket transactions shown in Table 5.21.**

   a. **What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?**

   The total number of possible rules, R, extracted from a data set that contains d items is:

   $$R = 3^d - 2^{d+1} + 1$$

   There are d = 6 items in the table (Beer, Bread, Butter, Cookies, Diapers and Milk).

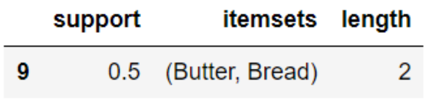
   $$R = 3^6 - 2^7 + 1 = 602$$

   602 association rules can be extracted from this data.

   b. **What is the maximum size of frequent itemsets that can be extracted (assuming minsup>0)?**
   Answer: As the longest transaction contains 4 items, the maximum size of frequent itemset is 4.

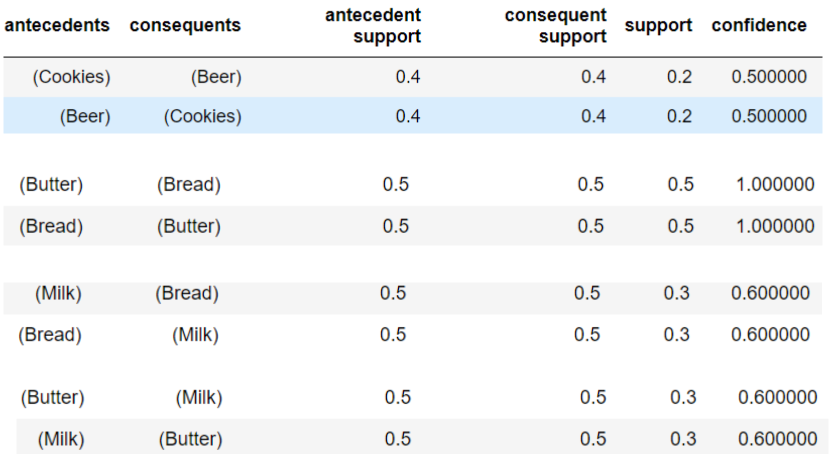
   c. **Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.**
   Disregarding the support threshold, there are 6! /3! possible 3-itemsets (with duplicates).
   The number of distinct 3-itemsets is therefore:
   $$= 6! / (3!)$$
   $$= (6 * 5 * 4)/ (3 * 2 *1)$$
   $$= 20$$

   d. Find an itemset (of size 2 or larger) that has the largest support.



	support	itemsets	length
9	0.5	(Butter, Bread)	2

e. Find a pair of items, a and b, such that the rules {a}→{b} and {b}→{a} have the same confidence.



antecedents	consequents	antecedent support	consequent support	support	confidence
(Cookies)	(Beer)	0.4	0.4	0.2	0.500000
(Beer)	(Cookies)	0.4	0.4	0.2	0.500000
(Butter)	(Bread)	0.5	0.5	0.5	1.000000
(Bread)	(Butter)	0.5	0.5	0.5	1.000000
(Milk)	(Bread)	0.5	0.5	0.3	0.600000
(Bread)	(Milk)	0.5	0.5	0.3	0.600000
(Butter)	(Milk)	0.5	0.5	0.3	0.600000
(Milk)	(Butter)	0.5	0.5	0.3	0.600000

7. Show that if a candidate *k*-itemset *X* has a subset of size less than k−1 that is infrequent, then at least one of the (k−1)-size subsets of *X* is necessarily infrequent.

Answer: Here a data stream arriving as a time ordered series of transaction is considered for analysis and is denoted by K= {k1, k2, k3, ………., kn}.
Each transaction ti contains a set of items ai and a ε I, where I = {a1, a2, ……. a3} is a set of items or objects and tn is called the current transaction arriving on the stream.
Assume that, TS0, TS1, ………., TSi-K+1, …..TSi denote the time periods or time slot which contains multiple transaction arriving in the interval and thus they form a partition of the transaction of the stream. Given an integer K, the time-based sliding window is defined as the set of transaction arriving in the last k times periods and denoted by SW= {TSik+1, ……. TSi -1, TSi}.
TSi is called the latest time slot and TSi-k is called the expiring one is the sliding window. When the time shift to the new slot TSI the effect of all transaction in TSi-K will be eliminated from the mining model.

8. Consider the following set of frequent 3-itemsets:
{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3,4, 5}.
Assume that there are only five items in the data set.

a. List all candidate 4-itemsets obtained by a candidate generation procedure using the Fk−1×F1 merging strategy.

- {1, 2, 3}: {1, 2, 3, 4}, {1, 2, 3, 5}
- {1, 2, 4}: {1, 2, 4, 5}
- {1, 3, 4}: {1, 3, 4, 5}
- {2, 3, 4}: {2, 3, 4, 5}

Other combinations were duplicates or not extendible from 3-itemsets to 4-itemsets.

b. List all candidate 4-itemsets obtained by the candidate generation procedure in *Apriori*.

From the frequent 3-itemsets, we can assume that minsup = 4.
All 4-itemsets from the previous part were generated from frequent 3-itemsets, so we get the same candidates as before:

{1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5}, {1, 3, 4, 5}, {2, 3, 4, 5}.

c. List all candidate 4-itemsets that survive the candidate pruning step of the *Apriori* algorithm.

{1, 2, 3, 4} survives as all its subsets ({1, 2, 3}, {1, 2, 4}, {1, 3, 4}, {2, 3, 4}) are frequent.
{1, 2, 3, 5} survives as all its subsets ({1, 2, 3}, {1, 2, 5}, {1, 3, 5}, {2, 3, 5}) are frequent.
Other 4-itemsets are pruned.