**Problem 1**

**(1) Mine all frequent itemsets using the Apriori algorithm, which we discussed in the class, with the minimum support = 50% (or 3 or more transactions). Show all candidate itemsets and frequent itemsets. You should follow the process described in the book and lecture (i.e., C1 → L1 → C2 → L2 → …). You don't need to show pruning steps. To save your time, L1 is given below:**

|  |
| --- |
| C2 |
| Itemset |
| {1,2} |
| {1,3} |
| {1,4} |
| {1,5} |
| {1,6} |
| {1,8} |
| {2,3} |
| {2,4} |
| {2,5} |
| {2,6} |
| {2,8} |
| {3,4} |
| {3,5} |
| {3,6} |
| {3,8} |
| {4,5} |
| {4,6} |
| {4,8} |
| {5,6} |
| {5,8} |
| {6,8} |

After pruning(Removed the itemset which support small than 3):

|  |  |
| --- | --- |
| L2 | |
| Itemset | Sup |
| {1,3} | 3 |
| {1,5} | 3 |
| {2,3} | 4 |
| {2,4} | 3 |
| {2,5} | 4 |
| {2,6} | 3 |
| {3,4} | 3 |
| {3,5} | 4 |
| {3,6} | 3 |
| {4,5} | 4 |
| {5,6} | 3 |

|  |
| --- |
| C3 |
| Itemset |
| {1,3,5} |
| {2,3,4} |
| {2,3,5} |
| {2,3,6} |
| {2,4,5} |
| {2,4,6} |
| {2,5,6} |
| {3,4,5} |
| {3,4,6} |
| {3,5,6} |
| {4,5,6} |

After pruning(Removed the itemset which support small than 3):

|  |  |
| --- | --- |
| L3 | |
| Itemset | Sup |
| {2,3,4} | 3 |
| {2,3,5} | 4 |
| {2,3,6} | 3 |
| {2,4,5} | 3 |
| {2,5,6} | 3 |
| {3,4,5} | 3 |
| {3,5,6} | 3 |

|  |
| --- |
| C4 |
| Itemset |
| {2,3,4,5} |
| {2,3,4,6} |
| {2,4,5,6} |
| {3,4,5,6} |

|  |
| --- |
| L4 |

After pruning(Removed the itemset which support small than 3):

|  |  |
| --- | --- |
| L4 | |
| Itemset | Sup |
| {2,3,4,5} | 3 |

So, the following itemsets are frequent itemsets and support greater or equal to 3 mined from the given dataset

|  |  |  |
| --- | --- | --- |
| L1 | | |
| Itemset | Sup | |
| 1 | 4 | |
| 2 | 4 | |
| 3 | 5 | |
| 4 | 4 | |
| 5 | 5 | |
| 6 | 3 | |
| 8 | 3 | |
| L2 | | | |
| Itemset | | Sup | |
| {1,3} | | 3 | |
| {1,5} | | 3 | |
| {2,3} | | 4 | |
| {2,4} | | 3 | |
| {2,5} | | 4 | |
| {2,6} | | 3 | |
| {3,4} | | 3 | |
| {3,5} | | 4 | |
| {3,6} | | 3 | |
| {4,5} | | 4 | |
| {5,6} | | 3 | |
| L3 | | | |
| Itemset | | Sup | |
| {2,3,4} | | 3 | |
| {2,3,5} | | 4 | |
| {2,3,6} | | 3 | |
| {2,4,5} | | 3 | |
| {2,5,6} | | 3 | |
| {3,4,5} | | 3 | |
| {3,5,6} | | 3 | |
| L4 | | | |
| Itemset | | Sup | |
| {2,3,4,5} | | 3 | |

**(2) Sort all frequent 4-itemsets by their item number. Then, select the first frequent 4-itemset form the sorted list of frequent 4-itemsets and mine all strong rules from this itemset that have the format {W, X} => {Y, Z}, where W, X, Y, and Z are individual items. Assume that minimum**

**confidence = 80%**

According to (1), {W,X,Y,Z}={2,3,4,5}

1. Rule: {2, 3} => {4, 5}

Confidence = support({2, 3, 4, 5}) / support({2, 3})

Confidence = 3 / 4 = 0.75=75%

1. Rule: {2, 4} => {3, 5}

Confidence = support({2, 3, 4, 5}) / support({2, 4})

Confidence = 3 / 3 = 1.00=100%

1. Rule: {2, 5} => {3, 4}

Confidence = support({2, 3, 4, 5}) / support({2, 5})

Confidence = 3 / 4 = 0.75=75%

1. Rule: {3, 4} => {2, 5}

Confidence = support({2, 3, 4, 5}) / support({3, 4})

Confidence = 3 / 3= 1.00=100%

1. Rule: {3, 5} => {2, 4}

Confidence = support({2, 3, 4, 5}) / support({3, 5})

Confidence = 3 / 4= 0.75=75%

1. Rule: {4, 5} => {2, 3}

Confidence = support({2, 3, 4, 5}) / support({4, 5})

Confidence = 3 / 4= 0.75=75%

The rules which have confidence >= 80% are:

{2, 4} => {3, 5} and {3, 4} => {2, 5}

**Problem 2**

**You can generate classification rules from the above dataset using the Apriori algorithm, which we discussed in the class.**

1. **. Execute the Apriori algorithm on the above dataset with the minimum support = 40% or 3 transactions. You need to proceed as we discussed in the class, i.e., C1 -> L1 -> C2 -> L2 -> . . . You need to show all candidate itemsets, frequent itemsets, and all rules mined from the dataset.**

|  |  |
| --- | --- |
| L1 | |
| ItemSet | Sup |
| A1=High | 5 |
| A1=Low | 2 |
| A2=On | 4 |
| A2=Off | 3 |
| A3=True | 4 |
| A3=False | 3 |
| Class=Positive | 4 |
| Class=Negative | 3 |

After pruning(Remove itemset which smaller than 3):

|  |  |
| --- | --- |
| L1 | |
| ItemSet | Sup |
| A1=High | 5 |
| A2=On | 4 |
| A2=Off | 3 |
| A3=True | 4 |
| A3=False | 3 |
| Class=Positive | 4 |
| Class=Negative | 3 |

|  |
| --- |
| C2 |
| ItemSet |
| A1=High, A2=On |
| A1=High, A2=Off |
| A1=High,A3=True |
| A1=High,A3=False |
| A1=High,Class=Positive |
| A1=High,Class=Negative |
| A2=On,A3=True |
| A2=On,A3=False |
| A2=Off, A3=True |
| A2=Off, A3=False |
| A2=On,Class=Positive |
| A2=On,Class=Negative |
| A2=Off,Class=Positive |
| A2=Off,Class=Negative |
| A3=True,Class=Positive |
| A3=True,Class=Negative |
| A3=False,Class=Positive |
| A3=False,Class=Negative |

|  |  |
| --- | --- |
| L2 | |
| ItemSet | Sup |
| A1=High, A2=On | 3 |
| A1=High, A2=Off | 2 |
| A1=High,A3=True | 3 |
| A1=High,A3=False | 2 |
| A1=High,Class=Positive | 3 |
| A1=High,Class=Negative | 2 |
| A2=On,A3=True | 1 |
| A2=On,A3=False | 3 |
| A2=Off, A3=True | 3 |
| A2=Off, A3=False | 0 |
| A2=On,Class=Positive | 3 |
| A2=On,Class=Negative | 1 |
| A2=Off,Class=Positive | 1 |
| A2=Off,Class=Negative | 2 |
| A3=True,Class=Positive | 2 |
| A3=True,Class=Negative | 2 |
| A3=False,Class=Positive | 2 |
| A3=False,Class=Negative | 1 |

After pruning(Remove itemset which smaller than 3):

|  |  |
| --- | --- |
| L2 | |
| ItemSet | Sup |
| A1=High, A2=On | 3 |
| A1=High,A3=True | 3 |
| A1=High,Class=Positive | 3 |
| A2=On,A3=False | 3 |
| A2=Off, A3=True | 3 |
| A2=On,Class=Positive | 3 |

|  |
| --- |
| C3 |
| ItemSet |
| A1=High, A2=On,A3=True |
| A1=High, A2=On,Class=Positive |
| A1=High,A3=True,Class=Positive |
| A2=On,A3=False,Class=Positive |

|  |  |
| --- | --- |
| L3 | |
| ItemSet | Sup |
| A1=High, A2=On,A3=True | 1 |
| A1=High, A2=On,Class=Positive | 2 |
| A1=High,A3=True,Class=Positive | 2 |
| A2=On,A3=False,Class=Positive | 2 |

After pruning(Remove itemset which smaller than 3):

|  |  |
| --- | --- |
| L3 | |
| ItemSet | Sup |

So, the following itemsets are frequent itemsets and support greater or equal to 3 mined from the given dataset

|  |  |
| --- | --- |
| L1 | |
| ItemSet | Sup |
| A1=High | 5 |
| A2=On | 4 |
| A2=Off | 3 |
| A3=True | 4 |
| A3=False | 3 |
| Class=Positive | 4 |
| Class=Negative | 3 |

|  |  |
| --- | --- |
| L2 | |
| ItemSet | Sup |
| A1=High, A2=On | 3 |
| A1=High,A3=True | 3 |
| A1=High,Class=Positive | 3 |
| A2=On,A3=False | 3 |
| A2=Off, A3=True | 3 |
| A2=On,Class=Positive | 3 |

**(2). Show only the rules that can be used for classification and calculate their confidences.**

1. Rule: A1=High => A2=On

Confidence = support({A1,High, A2=On}) / support({A1=High})

Confidence = 3 / 5= 0.6=60%

2. Rule: A2=On => A1=High

Confidence = support({A1,High, A2=On}) / support({A2=On})

Confidence = 3 / 4= 0.75=75%

3. Rule: A1=High => A3=True

Confidence = support({A1,High, A3=True}) / support({A1=High})

Confidence = 3 / 5= 0.6=60%

4. Rule: A3=True => A1=High

Confidence = support({A1,High, A3=True}) / support({A3=True})

Confidence = 3 / 4= 0.75=75%

5.Rule: A1=High => Class=Positive

Confidence = support({A1,High, Class=Positive}) / support({A1=High})

Confidence = 3 / 5= 0.6=60%

6.Rule: Class=Positive => A1=High

Confidence = support({A1,High, Class=Positive}) / support({Class=Positive})

Confidence = 3 / 4= 0.75=75%

1. Rule: A2=On => A3=False

Confidence = support({A2=On,A3=False}) / support({A2=On})

Confidence = 3 / 4= 0.75=75%

1. Rule: A3=False => A2=On

Confidence = support({A2=On,A3=False}) / support({A3=False})

Confidence = 3 / 3= 1.00=100%

1. Rule: A2=Off => A3=True

Confidence = support({A2=Off => A3=True}) / support({A2=Off})

Confidence = 3 / 3= 1.00=100%

10 .Rule: A3=True => A2=Off

Confidence = support({A2=Off => A3=True}) / support({A3=True})

Confidence = 3 / 4= 0.75=75%

11 .Rule: A2=On=> Class=Positive

Confidence = support({A2=On => Class=Positive}) / support({A2=On})

Confidence = 3 / 4= 0.75=75%

12 .Rule: Class=Positive => A2=On

Confidence=support({Class=Positive=>A2=On})/ support({Class=Positive})

Confidence = 3 / 4= 0.75=75%

**Problem 3**

**Consider the following contingency table.**

|  |  |  |
| --- | --- | --- |
|  | **C (buys coffee = Yes)** | **‾C (buys coffee = No)** |
| **T (buys tea = Yes)** | **142** | **862** |
| **‾T(buys tea = No)** | **186** | **1859** |

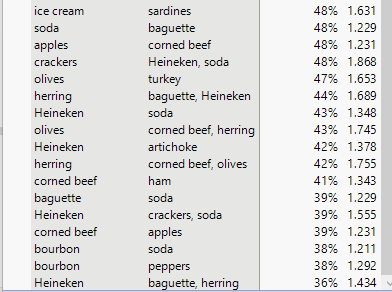
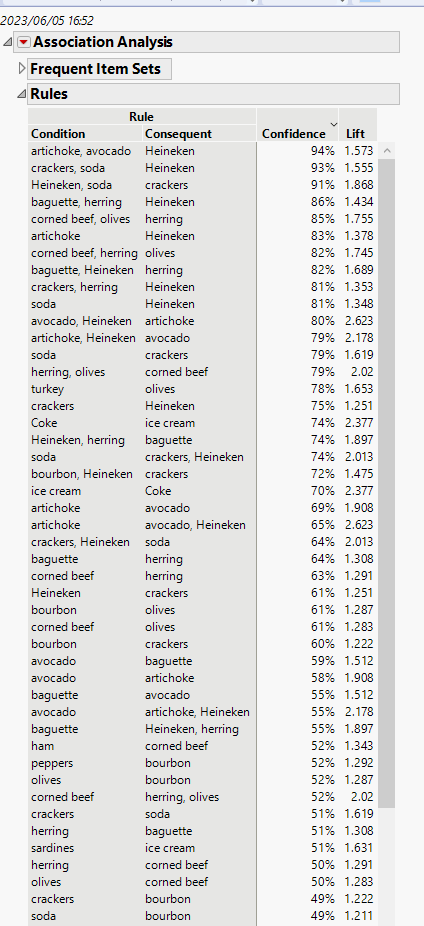
**Compute the lift, all-confidence, cosine, Kulczynski and imbalance ratio measures, and determine whether buying coffee and buying tea are positively correlated, negatively correlated, or not correlated. You must show all calculations.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | C (buys coffee = Yes) | ‾C (buys coffee = No) | Sum |
| T (buys tea = Yes) | 142 | 862 | 1004 |
| ‾T(buys tea = No) | 186 | 1859 | 2045 |
| Sum | 328 | 2721 | 3049 |

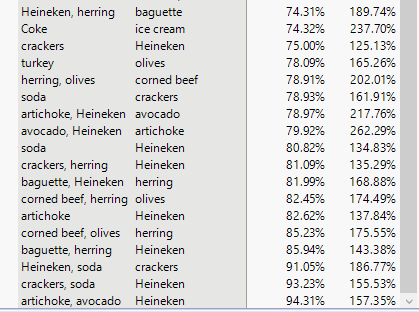
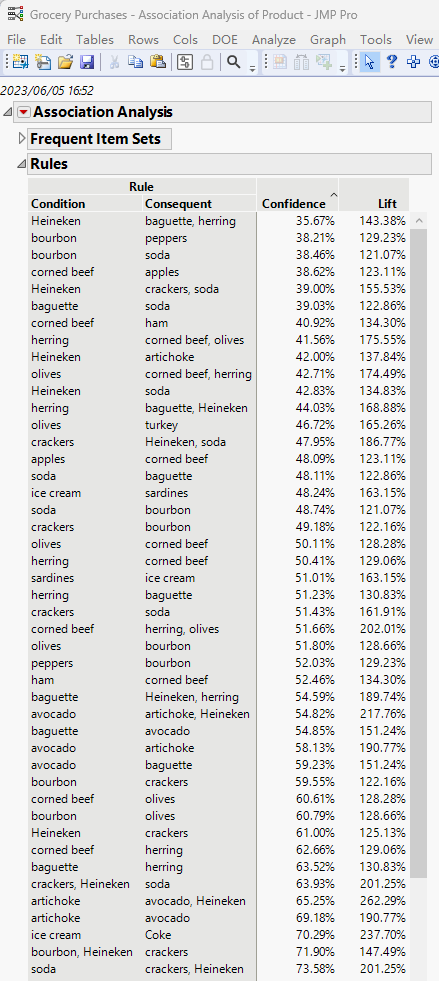
Since all\_confidence measure is smaller than 0.5,cosine measure is smaller than 0.5, and the lift measure is greater than 0.5, we can see that buying tea and coffee are negatively correlated

**Problem 4**

**Step 7**



**Step 8**



**Step 10**

1-Itemset

4-5-10-14-5-10-2

2-Itemset

4-5-10-3

4-5-10-44-5-10-5

3-Itemset

4-5-10-6

So, we can get 6 rules:

1. {crackers} => {Heineken}

Confidence({crackers} => {Heineken}) = Support({crackers, Heineken}) / Support({crackers}) = 0.3656 / 0.4875 = 0.75

1. {crackers} => {soda}

Confidence({crackers} => {soda}) = Support({crackers, soda}) / Support({crackers}) = 0.2507 / 0.4875 = 0.5143

1. {Heineken} => {soda}

Confidence({Heineken} => {soda}) = Support({Heineken, soda}) / Support({Heineken}) = 0.2567 / 0.5994 = 0.4283

1. {Heineken, crackers} => {soda}

Confidence({Heineken, crackers} => {soda}) = Support({Heineken, crackers, soda}) / Support({Heineken, crackers}) = 0.2338 / 0.3656 = 0.64

1. {Heineken, soda} => {crackers}

Confidence({Heineken, soda} => {crackers}) = Support({Heineken, crackers, soda}) / Support({Heineken, soda}) = 0.2338 / 0.2567 = 0.9108

1. {crackers, soda} => {Heineken}

Confidence({crackers, soda} => {Heineken}) = Support({Heineken, crackers, soda}) / Support({crackers, soda}) = 0.2338 / 0.2507 = 0.9326

**Step 11**

The rules in Step 10 greater than 70% are:

1. {crackers} => {Heineken} confidence = 0.75
2. {Heineken, soda} => {crackers} confidence = 0.9108
3. {crackers, soda} => {Heineken} confidence = 0.9326

This 3 rules are strong rule**s**