**Assignment 8**

**Due**: 4/3  
**Note: Show all your work.**

**Problem 1 (20 points)**. Consider the following transactional database.

|  |  |
| --- | --- |
| **TID** | **Items** |
| 100 | 2, 3, 4, 5, 7 |
| 200 | 1, 2, 5, 6 |
| 300 | 1, 4, 5, 7, 8 |
| 400 | 1, 2, 3, 5, 6, 7 |
| 500 | 2, 3, 5, 6, 7, 8 |

**(1)** Mine all frequent itemsets using Apriori. Show all candidate itemsets and frequent itemsets. You should follow the process described in the book and lecture (i.e., C1 → L1 → C2 → L2 → ...). Minimum support = 60% (or 3 or more transactions). To save your time, L1 and L2 are given below:

**C1:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Itemset** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| **Count** | **3** | **4** | **3** | **2** | **5** | **3** | **4** | **2** |

**L1:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Itemset** | 1 | 2 | 3 | 5 | 6 | 7 |
| **Count** | 3 | 4 | 3 | 5 | 3 | 4 |

**C2:** **C2:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Itemset** |  | **Itemset** | **Count** |
| {1,2} |  | {1,2} | 2 |
| {1,3} |  | {1,3} | 1 |
| {1,5} |  | {1,5} | 3 |
| {1,6} |  | {1,6} | 2 |
| {1,7} | second scan —> | {1,7} | 2 |
| {2,3} |  | {2,3} | 3 |
| {2,5} |  | {2,5} | 4 |
| {2,6} |  | {2,6} | 3 |
| {2,7} |  | {2,7} | 3 |
| {3,5} |  | {3,5} | 3 |
| {3,6} |  | {3,6} | 2 |
| {3,7} |  | {3,7} | 3 |
| {5,6} |  | {5,6} | 3 |
| {5,7} |  | {5,7} | 4 |
| {6,7} |  | {6,7} | 2 |

**L2:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Itemset** | {1,5} | {2,3} | {2,5} | {2,6} | {2,7} | {3,5} | {3,7} | {5,6} | {5,7} |
| **Count** | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 4 |

**C3:** **C3:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Itemset** |  | **Itemset** | **Count** |
| {2,3,5} |  | {2,3,5} | 3 |
| {2,3,6} |  | {2,3,6} | 2 |
| {2,3,7} |  | {2,3,7} | 3 |
| {2,5,6} | second scan —> | {2,5,6} | 3 |
| {2,5,7} |  | {2,5,7} | 3 |
| {2,6,7} |  | {2,6,7} | 2 |
| {3,5,7} |  | {3,5,7} | 3 |
| {5,6,7} |  | {5,6,7} | 2 |

**L3:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Itemset** | {2,3,5} | {2,3,7} | {2,5,6} | {2,5,7} | {3,5,7} |
| **Count** | 3 | 3 | 3 | 3 | 3 |

**C4:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Itemset** |  | **Itemset** | **Count** |
| {2,3,5,7} | scan —> | {2,3,5,7} | 3 |
| {2,5,6,7} |  | {2,5,6,7} | 2 |

**L4:**

|  |  |
| --- | --- |
| **Itemset** | {2,3,5,7} |
| **Count** | 3 |

**(2)** Sort all frequent 4-itemsets by their item number. Then, select the first frequent 4-itemset from 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%.

|  |  |  |
| --- | --- | --- |
| **Rule#** | **Rule** | **Confidence** |
| **R1** | {2} 🡺 {3,5,7} | (sup({2,3,5,7}))/sup({2}) = 3/4 = 75% |
| **R2** | {3} 🡺{2,5,7} | (sup({2,3,5,7}))/sup({3}) = 3/3 = 100% |
| **R3** | {5} 🡺{2,3,7} | (sup({2,3,5,7}))/sup({5}) = 3/5 = 60% |
| **R4** | {7} 🡺{2,3,5} | (sup({2,3,5,7}))/sup({7}) = 3/4 = 75% |
| **R5** | {2,3} 🡺 {5,7} | (sup({2,3,5,7}))/sup({2,3}) = 3/3 = 100% |
| **R6** | {2,5} 🡺 {3,7} | (sup({2,3,5,7}))/sup({2,5}) = 3/4 = 75% |
| **R7** | {2,7} 🡺 {3,5} | (sup({2,3,5,7}))/sup({2,7}) = 3/3 = 100% |
| **R8** | {3,5} 🡺 {2,7} | (sup({2,3,5,7}))/sup({3,5}) = 3/3 = 100% |
| **R9** | {3,7} 🡺 {2,5} | (sup({2,3,5,7}))/sup({3,7}) = 3/3 = 100% |
| **R10** | {5,7} 🡺 {2,3} | (sup({2,3,5,7}))/sup({5,7}) = 3/4 = 75% |
| **R11** | {2,3,5} 🡺 {7} | (sup({2,3,5,7}))/sup({2,3,5}) = 3/3 = 100% |
| **R12** | {2,3,7} 🡺 {5} | (sup({2,3,5,7}))/sup({2,3,7}) = 3/3 = 100% |
| **R13** | {2,5,7} 🡺 {3} | (sup({2,3,5,7}))/sup({2,5,7}) = 3/3 = 100% |
| **R14** | {3,5,7} 🡺 {2} | (sup({2,3,5,7}))/sup({3,5,7}) = 3/3 = 100% |

**Rules satisfying minimum confidence: R2, R5, R7, R8, R9, R11, R12, R13, R14**

**Problem 2 (10 points)**. Consider the following transactional database for sequential pattern mining.

|  |  |  |
| --- | --- | --- |
| CID | Day | Items |
| 1 | 1  14  24  31 | B, D, H  A, C, E  B, C, H  D, F, G |
| 2 | 4  9  14 | A, B, G, H  B, D, E, G  C, E, H |
| 3 | 1  24  51 | B, G, H  A, C, D, E  A, B, G, H |
| 4 | 2  12  25 | B, G A, B, E, H  B, C, D, E, G |

Determine the supports of the following sequences:

<{A}, {D}>, <{B, D}, {G}>, <{A}, {D, E}>

Note: below the transaction ID is considered to be **CID**.

|  |  |  |
| --- | --- | --- |
| **Rule#** | **Rule** | **Support** |
| R1 | {A} 🡺 {D} | s(R1) = support({A,D}) = 1/4 = 25% |
| R2 | {B,D} 🡺 {G} | s(R2) = support({B,D,G}) = 2/4 = 50% |
| R3 | {A} 🡺 {D,E} | s(R3) = support({A,D,E}) = 1/4 = 25% |

**Problem 3 (20 points)**. Consider the following contingency table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | C (buys coffee = Yes) | C (buys coffee = No) | Sum (row) |
| T (buys tea = Yes) | 238 | 197 | 435 |
| T (buys tea = No) | 82 | 85 | 167 |
| Sum (col) | 320 | 282 | 602 |

**(1)**  Compute the *lift*, *all-confidence*, *cosine*, *Kulczynski* and *imbalance ratio* measure, and determine whether buying coffee and buying tea are positively correlated, negatively correlated, or not correlated.

**lift**

lift(C,T) = = 1.03

lift(C,¬T) = = 0.92

lift(¬C, T) = = 0.97

lift(¬C,¬T) = = 1.09

**Tea & coffee** purchased together has a lift>1 and neither tea nor coffee purchased has a lift>1, while both of other combinations are lift<1, suggesting that there is a **positive correlation** between the purchases of tea and coffee.

**all-confidence**

all\_conf(C,T) = = = 0.55

all\_conf(C,¬T) = = = 0.26

all\_conf(¬C, T) = = = 0.45

all\_conf(¬C,¬T) = = = 0.30

All Confidence > 0.5 only for **coffee and tea** purchased together suggesting a **positive** **correlation**. All values associated with other combinations suggest negative correlations.

**cosine**

cosine(C,T) = = 0.64

cosine(C,¬T) = = 0.35

cosine(¬C, T) = = 0.56

cosine(¬C,¬T) = = 0.39

Cosine > 0.5 for **coffee and tea** purchased together, as well as not coffee and tea, suggesting a **positive** **correlation**. All values associated with other combinations suggest negative correlations.

**Kulczynski**

Kulc(C,T) = \* ( + ) = 0.65

Kulc(C,¬T) = \* ( + ) = 0.37

Kulc(¬C, T) = \* ( + ) = 0.58

Kulc(¬C,¬T) = \* ( + ) = 0.41

Kulczynski> 0.5 for **coffee and tea** purchased together, as well as not coffee and tea, suggesting a **positive** **correlation**. All values associated with other combinations suggest negative correlations.

**Imbalance ratio**

IR(C,T) = = 0.20

IR(C,¬T) = = 0.38

IR(¬C, T) = = 0.29

IR(¬C,¬T) = = 0.32

IR<0.5 for all combinations of purchases suggesting negative correlations.

Overall, coffee and tea purchased together almost always show a positive correlation suggesting that one one is purchased, the other is also purchased.

**(2)**  Perform the chi-square test with 95% significance level and determine whether they are correlated or not. degrees of freedom = (num\_rows – 1) \* (num\_cols – 1) = 1, and *α* = 0.05

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **C** | **¬C** | **Sum (r)** |  |  | **C** | **¬C** | **Sum (r)** |
| **T** | 238 | 197 | 435 |  | **T** | **231** | **204** | **435** |
| **¬T** | 82 | 85 | 167 |  | **¬T** | **89** | **78** | **167** |
| **Sum (c)** | 320 | 282 | 602 |  | **Sum (c)** | **320** | **282** | **602** |

**Round(435\*320/602) = 231**

**Round(167\*320/602) = 89**

**Round(435\*282/602) = 204**

**Round(167\*282/602) = 78**

χ2 = + + + = 1.63; χ20.05,1 = 3.84. Computed<referenced, so we fail to reject the null hypothesis, and conclude that we do not have enough evidence to conclude that there is a correlation between buying coffee and buying tea.

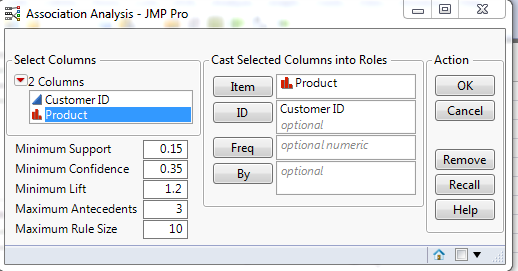
**Problem 4 (20 points).** This problem has two parts. Problem 2-1 uses Oracle and Problem 2-2 uses JMP Pro. Choose one of the two.

**Problem 2-1 (Oracle)** Follow the instructions in *oracle-association-rule-assignment.pdf* file. The submission requirements are indicated with **“Required.”**

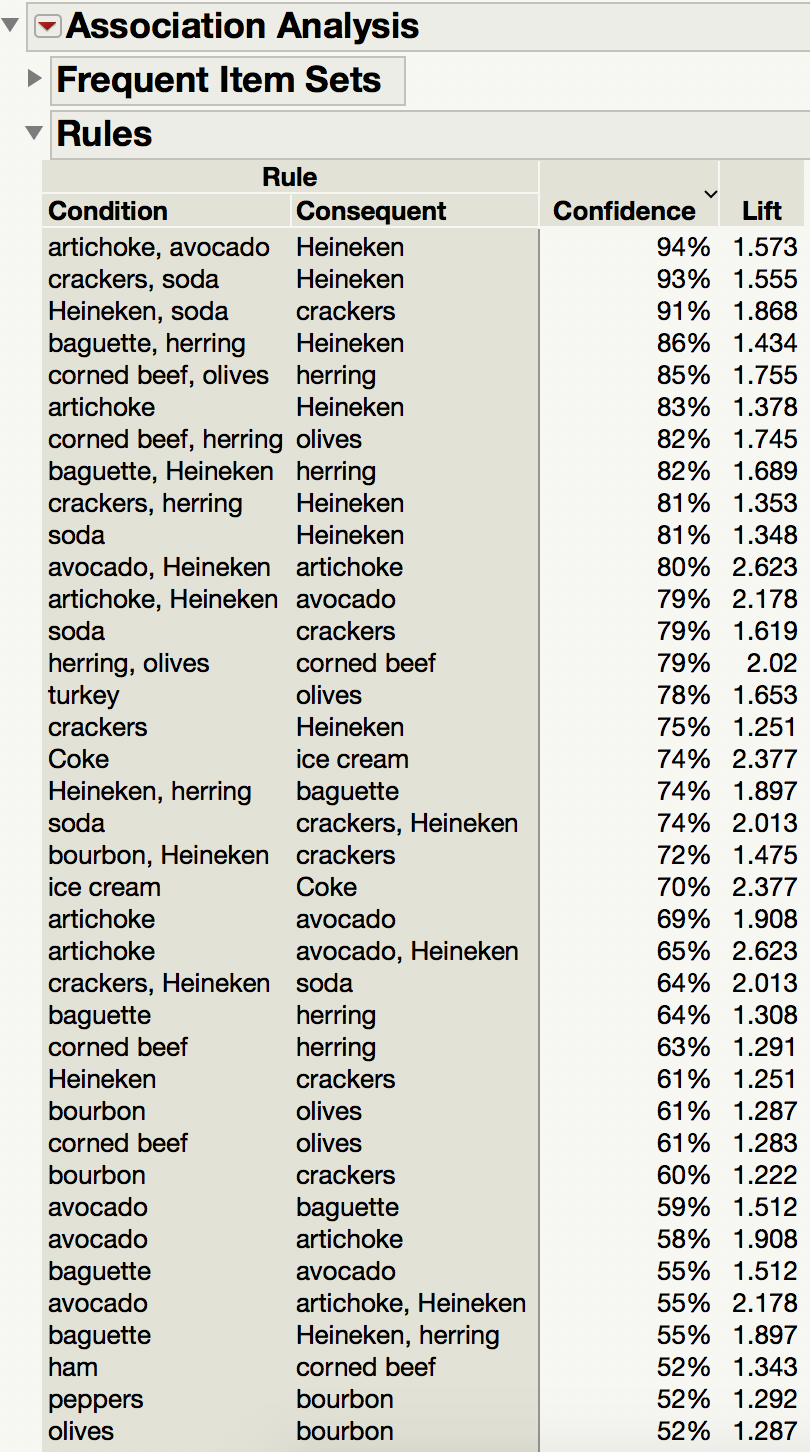
**Problem 2-1 (JMP Pro)** You will perform association analysis using JMP Pro. There is a secion in *Predictive and Specialized Modeling.pdf* documentation that shows how to do association analysis. You may want to read this section before starting the assignment. For this assignment, follow the instructions given below.

1. Start JMP Pro  
2. Select Help > Sample Data Library and open Grocery Purchases.jmp. 3. Select Analyze > Screening > Association Analysis.  
4. Select Product and click Item.  
5. Select Customer ID and click ID.  
6. Set the following parameters

Minimum Support: 0.15 Minimum Confidence 0.35 Lift: 1.2  
Maximum Antecedents: 3 Maximum Rule Size: 10

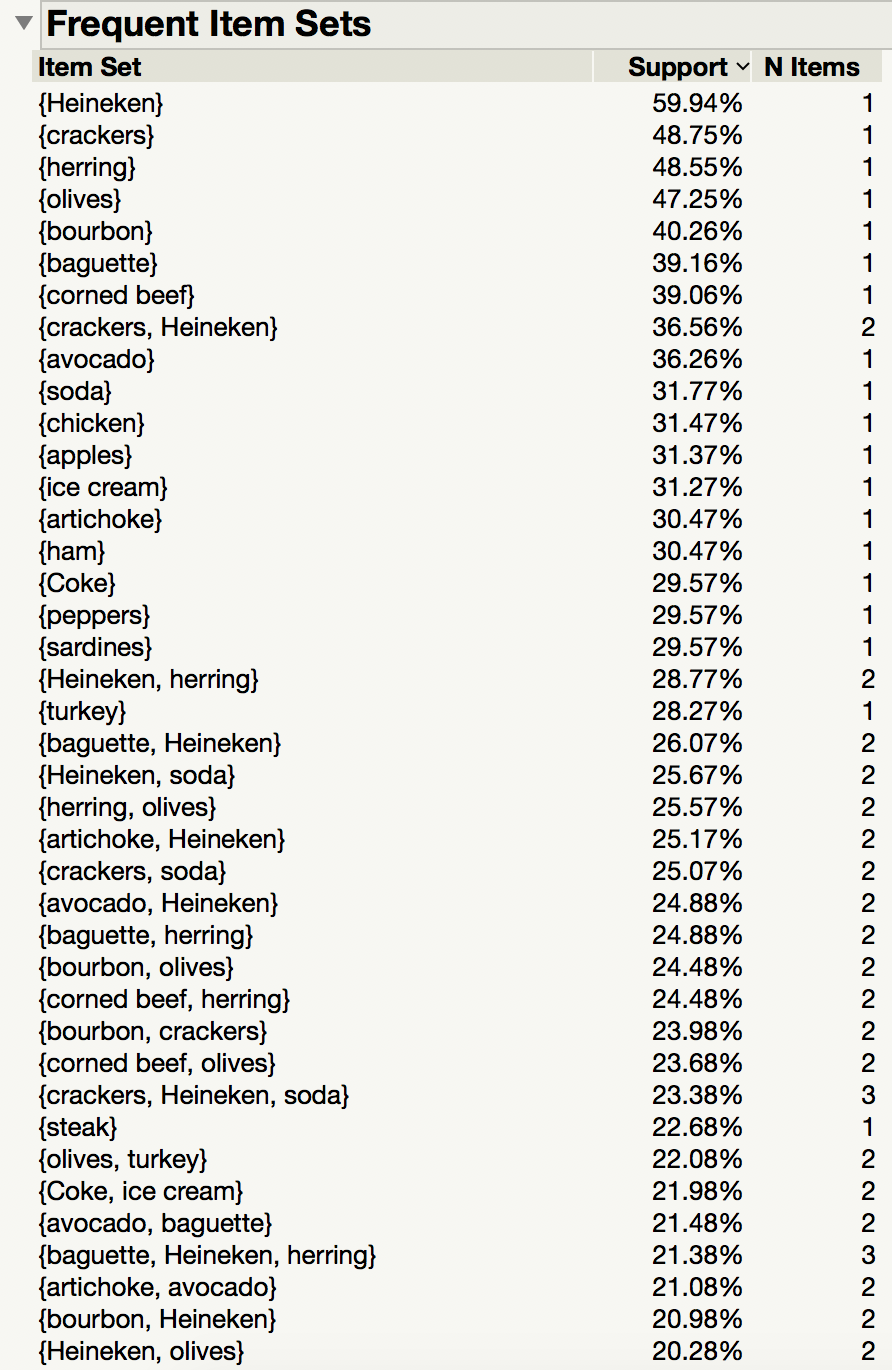


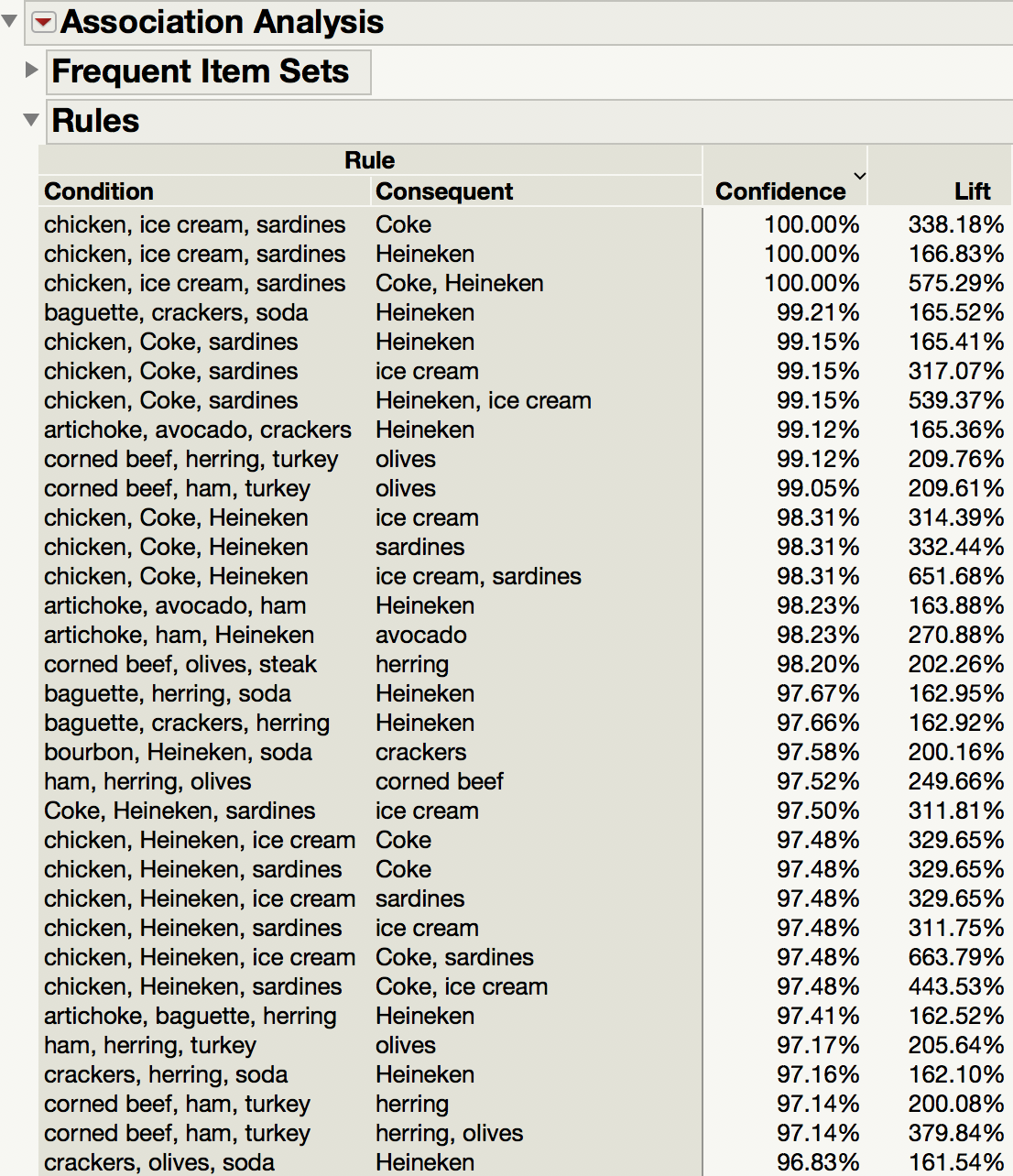
7. Click OK.  
Rules report will appear. Capture a top part of the report, including about 20 rules, and include this screenshot in your report.



8. Right-click Confidence > select Format Column Change Dec from 0 to 2  
Select Support, Confidence, and Lift

Click OK. Rules report will now show 2 digits below the decimal point for those three measures. Capture a top part of the report, including about 20 rules, and include this screenshot in your report.





9. Find the 3-itemset {crackers, Heineken, soda} under Frequent Item Sets.

sup({crackers, Heineken, soda}) = 0.2338  
10. Manually mine all rules from this 3-itemset and calculate the confidences of these

rules using the method discussed in the class. Show, in your report, all rules and their

confidences. You need to show how you calculated the confidences.

|  |  |  |
| --- | --- | --- |
| **Rule#** | **Rule** | **Confidence** |
| **R1** | {crackers} 🡺 {Heineken, soda} | (sup({crackers, Heineken, soda}))/sup({crackers}) = 0.2338/0.4875 = 0.480 |
| **R2** | {Heineken} 🡺{crackers, soda} | (sup({crackers, Heineken, soda}))/sup({Heineken}) = 0.2338/0.5994 = 0.390 |
| **R3** | {soda} 🡺{Heineken, crackers} | (sup({crackers, Heineken, soda}))/sup({soda}) = 0.2338/0.3177 = 0.736 |
| **R4** | {crackers, Heineken} 🡺{soda} | (sup({crackers, Heineken, soda}))/sup({crackers, Heineken}) = 0.2338/0.3656 = 0.639 |
| **R5** | {crackers, soda} 🡺 {Heineken} | (sup({crackers, Heineken, soda}))/sup({crackers, soda}) = 0.2338/0.2507 = 0.933 |
| **R6** | {soda, Heineken} 🡺 {crackers} | (sup({crackers, Heineken, soda}))/sup({soda, Heineken}) = 0.2338/0.2567 = 0.911 |

11. If the minimum confidence is 70%, which ones are strong rules?

**Rules satisfying minimum confidence: R3, R5, R6**