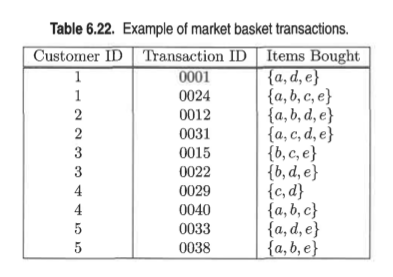
Data Mining Assignment 3

1) Read Chapter 6 (only sections 6.1 and 6.7).  
  
2) Do Chapter 6 textbook problem #2 (parts a,b,c,d only) on page 404.

Consider the data set shown in Table 6.22

(a) Compute the support for itemsets {e}, {b, d}, and {b, d, e} by treating each transaction ID as a market basket.  
Answer:



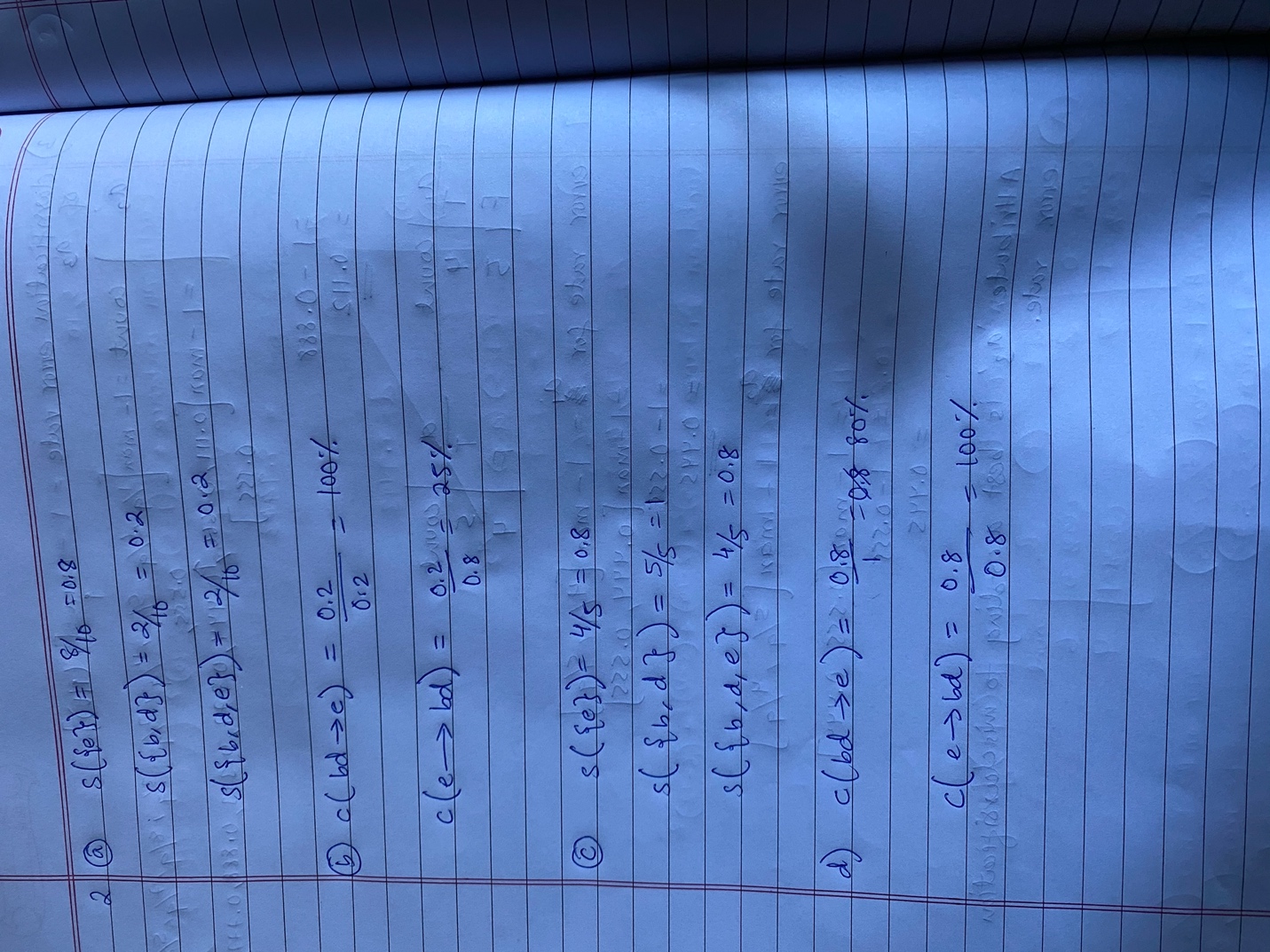
1. Compute the support for itemsets {e}, {b, d}, and {b, d, e} by treating each transaction ID as a market basket.
2. Use the results in part (a) to compute the confidence for the association rules {b,d} −→ {e} and {e} −→ {b,d}. Is confidence a symmetric measure?

No, confidence is not a symmetric measure.

1. 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.)
2. Use the results in part (c) to compute the confidence for the association rules {b, d} −→ {e} and {e} −→ {b, d}.

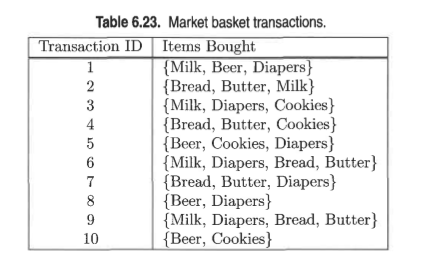
Part c

|  |  |
| --- | --- |
| Customer ID | Items |
| 1 | A, b ,c d, e |
| 2 | A, b c, d ,e |
| 3 | B, c , d, e |
| 4 | A, b, c ,d |
| 5 | A, b , d, e |



3) Do Chapter 6 textbook problem #6 (parts d,e only) on page 406.

Consider the market basket transactions shown in T able 6.23.



R = 3d −2d+1 +1. (R=rules)

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

Answer: {Bread, Butter}.

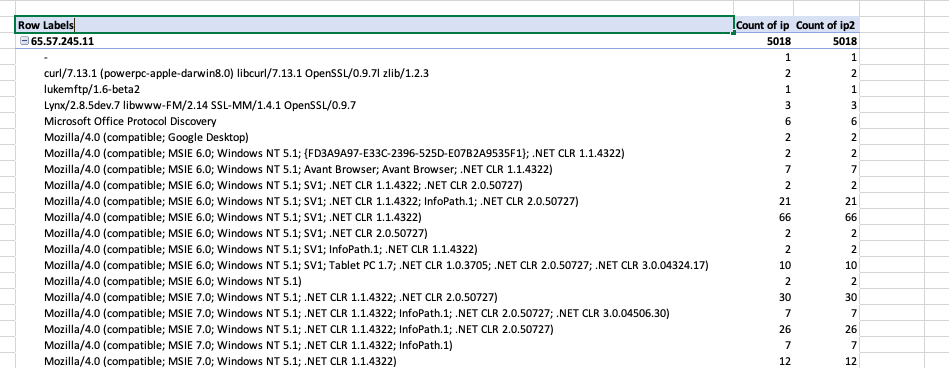
(e)  Find a pair of items, a and b, such that the rules {a} −→ {b} and {b} −→ {a} have the same confidence.  
Answer: (Beer, Cookies) or (Bread, Butter).

4) Using the data at [www.stats202.com/more\_stats202\_logs.txt](http://www.stats202.com/more_stats202_logs.txt) and treating each row as a "market basket" compute the support and confidence for the rule ip=65.57.245.11 → "Mozilla/5.0 (X11; U; Linux i686 (x86\_64); en-US; rv:1.8.1.3) Gecko/20070309 Firefox/2.0.0.3".

support(65.57.245.11) = 5018

support= 1385/14803 = 0.09356211578 = 9.3%

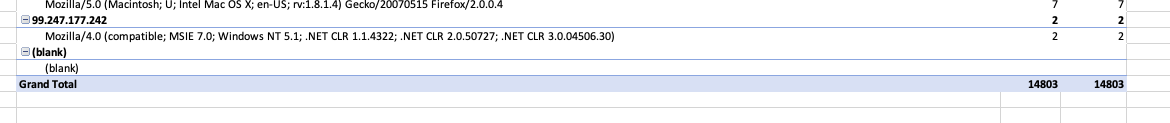
confidence = 1385/5018 = 0.27600638 = 27.6 %



according to the table the support



Total count



State what the support and confidence values mean in plain English in this context.

**Support** is an indication of how frequently the items appear in the **data**.

**Confidence** can be interpreted as the likelihood of occurring both IP address and link together.