Data Mining HW4

1.

(a)

Formula for association rules: R = 3^d – 2^(d+1) +1

Given items: Bread, Butter, Milk, Beer, Diapers, Cookies, our distinct values (d) = 6

R = 3^(6) – 2^(6+1) + 1 = 602

(b)

The items have a minimum support of 1.

The maximum number of items in a transaction is 4. Hence, the maximum size of the frequent items that can be extracted is 4.

(c)

Basically we are trying to pick 3 items from a total of 6. Hence, by using the combinations formula, we find the value of 6C3 which is 20

(d)

List of items with a minimum size of 2

Itemset	**Support**
milk, beer	1
milk, diapers	4
milk, bread	3
milk, butter	2
milk, cookies	1
beer, diapers	3
beer, bread	0
beer, butter	0

beer, cookies	2
diapers, bread	2
diapers, butter	3
diapers, cookies	1
bread, butter	5
bread, cookies	1
butter, cookies	1

As we can see, bread and butter itemset combination has the largest count 5

(e)

Confidence formula: conf(A->B) = support{A,B} / support{A}

Let's take bread and butter as an itemset and look at their confidence:

conf(bread -> butter) = (support({bread, butter}))/(support({bread})) = 5/11 = 0.45
conf(butter -> bread) = (support({butter,bread}))/(support({butter})) = 5/11 = 0.45

Hence, "bread, butter" is a pair of items that have the same confidence

2.

(a)

Given itemset: 1,3,6,8,9

Let's find out the total possible combinations: 1 + {3,6,8,9}, 3 + {6,8,9}, 6 + {8,9}, 1,3 + {6,8,9}, 3,6 + {8,9}, 6,8 + {9}, 1,6 + {8,9}, 3,8 + {9}, 1,8 + {9}

Let's look at them along with their terminations as well: 1 + {3,6,8,9} – L5, 3 + {6,8,9} – L12, 6 + {8,9} – L11, 1,3 + {6,8,9} – L5, 3,6 + {8,9} – L12, 6,8 + {9} – L11, 1,6 + {8,9} – L5, 3,8 + {9} – L11, 1,8 + {9} – L4

We can see that we are visiting L5, L12, L11, L5, L12, L11, L5, L11, and L4. The unique ones from these visits are: L4, L5, L11, L12.

(b)

168 and 680 will be the candidate itemsets that are supported by the given transactions.

3.

Let's look at the expected values for all the pairs

E(A, B) = (Total A x Total B)/(Total Observations)

1. E(Tea, Coffee) = (200 x 800)/1000 = 160
2. E(Tea, 'Coffee) = (200 x 200)/1000 = 40
3. E('Tea, Coffee) = (800 x 800)/1000 = 640
4. E('Tea, 'Coffee) = (200 x 200)/1000 = 160

The chi-squared metric for the same are:

1. X^2 (Tea, Coffee) = (150-160)^2/160 = 100/160 = 0.625
2. X^2 (Tea, 'Coffee) = (50-40)^2/40 = 100/40 = 2.5
3. X^2 ('Tea, Coffee) = (650-640)^2/640 = 100/640 = 0.15625
4. X^2 ('Tea, 'Coffee) = (150-160)^2/160 = 100/160 = 0.625

By adding all the values we get our result as 3.90625

4.

Let's look at the example of Hospital mortality rates

Consider that we are contrasting the mortality rates of two hospitals, Hospital A and Hospital B, for two distinct types of surgeries, Surgery X and Surgery Y.

Hospital A

Surgery X: 30 deaths in 200 operations (mortality rate of 15%)
Surgery Y: 5 deaths in 100 operations (mortality rate of 5%)

Hospital B:

Surgery X: 20 deaths out of 100 surgeries (mortality rate of 15%)
Surgery Y: 10 deaths out of 200 surgeries (mortality rate of 5%)

Now Simon’s paradox for this example can be observed in the overall rates:

Hospital A: (30 + 5) / (200 + 100) = 35 / 300 = 11.67% mortality rate
Hospital B: (20 + 10) / (100 + 200) = 30 / 300 = 10% mortality rate

We can see that Hospital A appears to have a higher overall mortality rate than Hospital B. However, we also find that Hospital A has a lower mortality rate for both Surgery X and Surgery Y when you break down the data by type of surgery.

This paradox occurs because Surgery X has a higher risk in operation, however, Hospital A handles more cases in X than Y. With Hospital B, it’s the other way around. It handles more of the surgery with lower risk which is surgery Y.

This is Simon's paradox that rates individually and aggregated can be different and so it’s important to keep this and the context of what you’re trying to achieve for good data analysis.