CSCI 5502 - Data Mining - Homework 3
Name: Krishna Chaitanya Sripada
Student ID: 104375417
Honor Pledge: On my honor, as a University of Colorado at Boulder student, I have neither given nor received unauthorized assistance on this work.

Ans 1.

a. lift(ski, football) = (1500/4000) /((2000/4000) * (2500/4000)) = 0.375/0.3125 = 1.2
Since the lift measure is greater than 1, the correlation relationship between ski and playing football is positive.

b. Since the association rule "ski ⇒ football" is mined, let us assume X = ski and Y = football.
Support = P(X∪Y) = 1500/4000 = 0.375 = 37.5%.
Confidence = P(Y|X) = 1500/2000 = 0.75 = 75%.
Since the support value is 37.5% and Confidence value is 75% and they satisfy the minimum support and minimum confidence thresholds, the association rule "ski ⇒football" is strong.

Ans 2.

a. Since min_support = 60%, the maximum number of possible frequent itemsets = 3:

b. Apriori Algorithm for finding the frequent Itemsets:

i. By scanning the table, we find the set of frequent 1-itemsets (C_1):

Itemset	Count	Support
{B}	4	80%
{D}	1	20%
{E}	3	60%
{F}	1	20%
{G}	3	60%
{I}	3	60%
{N}	4	80%
{O}	1	20%
{S}	1	20%
{T}	1	20%
{Z}	3	60%

ii. After pruning this data, itemsets {D}, {F}, {O}, {S}, {T} are removed. So L_1 is:

Itemset	Count	Support
{B}	4	80%
{E}	3	60%
{G}	3	60%
{I}	3	60%
{N}	4	80%
{Z}	3	60%

iii. By scanning the table, we find the set of frequent 2-itemsets (C_2):

Itemset	Count	Support
{B,E}	2	40%
{B,G}	2	40%
{B,I}	3	60%
{B,N}	4	80%
{B,Z}	2	40%
{E,G}	1	20%
{E,I}	2	20%
{E,N}	2	40%
{E,Z}	1	20%
{G,I}	1	20%
{G,N}	2	40%
{G,Z}	3	60%
{I,N}	3	60%
{I,Z}	1	20%
{N,Z}	2	40%

iv. After pruning this data, itemsets {B,E}, {B,G}, {B,Z}, {E,G}, {E,I}, {E,N}, {E,Z}, {G,I}, {G,N}, {I,Z}, {N,Z} are removed. So L_2 is:

Itemset	Count	Support
{B,I}	3	60%
{B,N}	4	80%
{G,Z}	3	60%
{I,N}	3	60%

v. By scanning the table, we find the set of frequent 3-itemsets (C_3):

Itemset	Count	Support
{B,I,N}	3	60%

v. After pruning this data, no itemset is removed and since frequent 4-itemsets cannot be found, the algorithm ends here. L_3 is:

Itemset	Count	Support
{B,I,N}	3	60%

c. The number of rounds of database scans is 3. The total number of candidates =11 (6+4+1).

d. In the first approach, since all frequent k-itemsets are part of the candidate k-itemsets, it produces $\mathcal{O}(|F_{k-1}| * |F_1|)$ where $|F_k|$ is the number of frequent k-itemsets. The computational complexity is $\mathcal{O}(\sum_{k=1}^{n} |F_{k-1}||F_1|)$.
In the second approach, since there are 'n' items, the number of candidates itemsets generated at level k is equal to $\binom{n}{k}$. Given that the amount of computations needed for each candidate is $\mathcal{O}(k)$, the computational complexity would be $\mathcal{O}(\sum_{k=1}^{n} k * \binom{n}{k}) = \mathcal{O}(n * 2^{n-1})$.
The first approach has a substantial improvement over the second approach however, the first approach still produces a large number of unnecessary candidates. To avoid this, we have to make sure that for every candidate k-itemset that survives the pruning step, every item in the candidate must be contained in atleast k-1 of the frequent (k-1)-itemsets.

Ans 3.

a. The largest value of k = 3 and the data containing the frequent itemset is {(Bread, Milk, Cheese), (Bread, Milk, Pie)}.
The non-empty subsets for this frequent itemset are
1. {Bread, Milk}, {Bread, Cheese}, {Milk, Cheese}, {Bread}, {Milk}, {Cheese}.
2. {Bread, Milk}, {Bread, Pie}, {Milk, Pie}, {Bread}, {Milk}, {Pie}.

For 1. ,the association rules are:
Bread $\wedge$ Milk $\Rightarrow$ Cheese [support = 3/4 = 75%, confidence = 3/4=75%]
Bread $\wedge$ Cheese $\Rightarrow$ Milk [support = 3/4 = 75%, confidence = 3/3=100%]
Milk $\wedge$ Cheese $\Rightarrow$ Bread [support = 3/4 = 75%, confidence = 3/3=100%]

Since min_support = 60% and min_confidence = 80%, the rules that satisfy are:
Bread $\wedge$ Cheese $\Rightarrow$ Milk [75%,100%]
Milk $\wedge$ Cheese $\Rightarrow$ Bread [75%,100%]

For 2. ,the association rules are:
Bread $\wedge$ Milk $\Rightarrow$ Pie [support = 3/4 = 75%, confidence = 3/4=75%]
Bread $\wedge$ Pie $\Rightarrow$ Milk [support = 3/4 = 75%, confidence = 3/3=100%]
Milk $\wedge$ Pie $\Rightarrow$ Bread [support = 3/4 = 75%, confidence = 3/3=100%]

Since min_support = 60% and min_confidence = 80%, the rules that satisfy are:
Bread $\wedge$ Pie $\Rightarrow$ Milk [75%,100%]
Milk $\wedge$ Pie $\Rightarrow$ Bread [75%,100%]

b. Since min_support = 60% and min_confidence = 80%, the largest value of k=3 and the frequent dataset is {(Wonder-Bread, Sweet-Pie, Sunset-Milk), (Wonder-Bread, Sweet-Pie, Dairyland-Milk), (Wonder-Bread, Dairyland-Cheese, Sunset-Milk)}