1. (15 pts) Consider the market basket transactions shown in Table 1 below,
Table 1: Market basket transactions.

Transaction ID	Items
1	Milk, Beer, Diapers
2	Bread, Butter, Milk
3	Milk, Diapers, Cookies
4	Bread, Butter, Cookies
5	Beer, Cookies, Diapers
6	Milk, Diapers, Bread, Butter
7	Bread, Butter, Diapers
8	Beer, Diapers
9	Milk, Diapers, Bread, Butter
10	Beer, Cookies

(a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support, but not rules involving empty itemsets)?
Answer -
Maximum number of association rules for d items is given by -
$R = 2^d - 2$

d = 6 (Beer, Bread, Butter, Cookies, Diapers and Milk)

$R = 2^6 - 2 = 62$

(b) What is the maximum size of frequent items that can be extracted (assuming minsup >0)?
Answer -
Since item sets 6 and 9 have a maximum size of 4. Maximum size of frequent items that can be extracted (assuming minsup >0).

(c) Write an expression for the maximum number of 3-itemsets that can be de-
rived from this data set.
Answer-
Since 6 itemsets have 3 items which include all 6 unique items. Counting n
(d) Find an itemset (of size 2 or larger) that has the largest support max number of 3-itemsets is equal to nCr.
6C3 = 6! / 3! * 3! = 20

(e) Find a pair of items, a and b, such that rules {a} ⟶ {b} and {b} ⟶ {a} have
the same confidence.
Answer-
Support of bread is 5 and support of butter is also 5.
Bread -> butter and butter -> bread have same confidence c = support of bread and butter / support of bread
C = 5/5 = 1

2. (15 pts) The Apriori algorithm uses hash tree data structure to store and count
support of candidate itemsets. Consider the hash tree for candidate 3-itemsets
shown in the Figure 1,
a)Answer-
{1,3,6}-L5
{1,3,8}-L5
{1,3,9}-L5
{1,6,8}-L5
{1,6,9}-L5
{1,8,9}-L4

{3,6,8}-L12
{3,6,9}-L12
{3,8,9}-L11

{6,8,9}-L11

Therefore Candidates traveled would be L4, L5, L11, L12.

b) {1,6,8}, {6,8,9} are supported by the transaction {1,3,6,8,9}.

3. (10 pts) Given a contingency table for Tea and Coffee shown in Figure 2, compute the chi-squared statistic. Show the steps (including the expected counts assuming these two variables are independent to each other). You are welcome to use online tools to check if your answer is correct.
Answer -

	Coffee	Coffee’	
Tea	150	50	200
Tea’	650	150	800
	800	200	1000

Expected counts -

	Coffee	Coffee’	
Tea	200*800/1000=160	200*200/1000=40	200
Tea’	800*800/1000=640	800*200/1000=160	800
	800	200	1000

chi-squared statistic = $\sum_{i=1}^{rc} (obs_i - exp_i)^2 / exp_i$

chi-squared statistic =$(150 - 160)^2$/ 160 + $(50 - 40)^2$/40 + $(650 - 640)^2$/640 + $(150 - 160)^2$/160

chi-squared statistic = 10/16 + 2.5 + 10/64 + 10/16 = 3.9

4. (10 pts) Use an example to explain Simpson’s Paradox. If you get your idea from somewhere, please make sure that you include proper citations. Use your own words to describe. Don’t copy and paste.
Answer-

Lets assume we are treating a medical condition which infects dogs and humans -

Lets assume that among 1 dog and 4 humans which are treated 1 dog recovers and 1 of 4 humans recovers.
Let's assume that among 4 dogs and 1 human which are not treated 3 dogs recover and the human does not recover

	Dogs	Humans
Treatment received	1/1 = 100%	¼ = 25%
Treatment not given	¾ = 75%	0

This concludes that if treatment is given then the recovery chance increases
For humans it becomes 0 -> 25%
For dogs it becomes 75 -> 100%
Therefore showing that the treatment is effective

But if we combine the data

Among 1 dog and 4 humans together who are treated only ⅖ - 40% recovered
While when treatment is not given to 4 dogs and 1 human ⅗ - 60% recovered

This gives a conclusion that when treatment is not given more % of recovery was observed
Showing that treatment is ineffective

This shows 2 opposite conclusions on the same data depending upon how we aggregate the data. This is called Simpson's Paradox.

If we know that humans get this hypothetical infection more seriously than dogs therefore they are more likely to be prescribed treatment then it makes sense that few humans survived after treatment. Because the people receiving treatment are more likely to not recover. Thus giving an explanation as to why less humans recover even with treatment.