1. Apriori Algorithm
    a. Support Priority in Apriori algorithm
        i. The Apriori algorithm is based on the downward closure property, also known as the anti-monotonicity of support. This property states:
            1. If an itemset is frequent, then all of its subsets must also be frequent.
        ii. This can be understood mathematically as follows:
            1. Let s be a frequent itemset, and let s′ be any non-empty subset of s
            2. The support of an itemset is the proportion of transactions in the dataset that contain that itemset.
                a. The support of any subset s′ of the frequent itemset s is at least as large as the support of s
                b. This is because if a transaction contains s, it must also contain every subset of s, including s′
            3. Thus, the relative support of s′ is at least as great as that of s
                a. This property is a cornerstone of the Apriori algorithm, which prunes the search space by eliminating itemsets whose subsets are infrequent, since if any subset of an itemset is infrequent, the itemset itself cannot be frequent.
    b. Confidence of Association rules
        i. Let L be a frequent itemset, and sss be a subset of L
        ii. The confidence of an association rule is defined as the ratio of the support of the union of the itemsets in the rule to the support of the antecedent
            1. confidence(s ⇒ (L − s)) = support(L) / support(s)
            2. Let s′ be a subset of s. We need to prove that:
                a. confidence(s′ ⇒ (L−s′)) ≤ confidence(s ⇒ (L−s))
        iii. Proof:
            1. The confidence of s ⇒ (L−s) is:
                a. confidence(s ⇒ (L − s)) is:
                    i. confidence(s ⇒ (L−s)) = support(L) / support(s)
                b. The confidence of s′ ⇒ (L − s′) is:
                    i. confidence(s’ ⇒ (L−s’)) = support(L) / support(s’)
                c. Since s′ ⊆ s by the downward closure property of support:
                    i. support(s′) ≥ support(s)
                d. Thus, the denominator of confidence(s′⇒(L−s′)) is greater than or equal to the denominator of confidence(s⇒(L−s)).
                e. Since the numerator support(L) is the same in both cases, we conclude:
                    i. confidence(s′ ⇒ (L−s′)) ≤ confidence(s ⇒ (L−s))
                f. Hence, the confidence of the rule s′⇒(L−s′) cannot exceed the confidence of s ⇒ (L−s).
2. Frequent itemsets
    a. Find all frequent itemsets using Apriori

i. You need to calculate the frequent itemsets based on the minimum support threshold (60%) using the Apriori algorithm. Here's how to proceed manually:

    1. List of transactions:
        a. T100: {M, O, N, K, E, Y}
        b. T200: {D, O, N, K, E, Y}
        c. T300: {M, A, K, E}
        d. T400: {M, U, C, K, Y}
        e. T500: {C, O, K, I, E}
    2. Step 1: Calculate the support for individual items (1-itemsets). You need to find the support of each item. The support is the proportion of transactions that contain the item. The minimum support threshold is 60%, which means an item must appear in at least 3 out of 5 transactions.
        a. M: appears in T100, T300, T400 → Support = 3/5 = 60%
        b. O: appears in T100, T200, T500 → Support = 3/5 = 60%
        c. N: appears in T100, T200 → Support = 2/5 = 40%
        d. K: appears in T100, T200, T300, T400, T500 → Support = 5/5 = 100%
        e. E: appears in T100, T200, T300, T500 → Support = 4/5 = 80%
        f. Y: appears in T100, T200, T400 → Support = 3/5 = 60%
        g. D: appears in T200 → Support = 1/5 = 20%
        h. A: appears in T300 → Support = 1/5 = 20%
        i. U: appears in T400 → Support = 1/5 = 20%
        j. C: appears in T400, T500 → Support = 2/5 = 40%
        k. I: appears in T500 → Support = 1/5 = 20%
        l. Frequent 1-itemsets (support ≥ 60%):
            i. {M}, {O}, {K}, {E}, {Y}
    3. Step 2: Generate candidate 2-itemsets and calculate support. Pair up frequent 1-itemsets and check their support.
        a. {M, O}: appears in T100 → Support = 1/5 = 20%
        b. {M, K}: appears in T100, T300, T400 → Support = 3/5 = 60%
        c. {M, E}: appears in T100, T300 → Support = 2/5 = 40%
        d. {M, Y}: appears in T100, T400 → Support = 2/5 = 40%
        e. {O, K}: appears in T100, T200, T500 → Support = 3/5 = 60%
        f. {O, E}: appears in T100, T200, T500 → Support = 3/5 = 60%
        g. {O, Y}: appears in T100, T200 → Support = 2/5 = 40%
        h. {K, E}: appears in T100, T200, T300, T500 → Support = 4/5 = 80%

        i. {K, Y}: appears in T100, T200, T400 → Support = 3/5 = 60%
        j. {E, Y}: appears in T100, T200 → Support = 2/5 = 40%
        k. Frequent 2-itemsets (support ≥ 60%):
            i. {M, K}, {O, K}, {O, E}, {K, E}, {K, Y}
    4. Step 3: Generate candidate 3-itemsets and calculate support
        a. {M, K, E}: appears in T100, T300 → Support = 2/5 = 40%
        b. {O, K, E}: appears in T100, T200, T500 → Support = 3/5 = 60%
        c. {K, E, Y}: appears in T100, T200 → Support = 2/5 = 40%
        d. {O, K, Y}: appears in T100, T200 → Support = 2/5 = 40%
        e. Frequent 3-itemsets (support ≥ 60%):
            i. {O, K, E}
        f. Frequent itemsets (final):
            i. {M}, {O}, {K}, {E}, {Y}, {M, K}, {O, K}, {O, E}, {K, E}, {K, Y}, {O, K, E}

ii. FP-growth

1. This part would involve building an FP-tree based on the frequency of items. I can guide you through the steps or explain the construction of the FP-tree if you'd like!

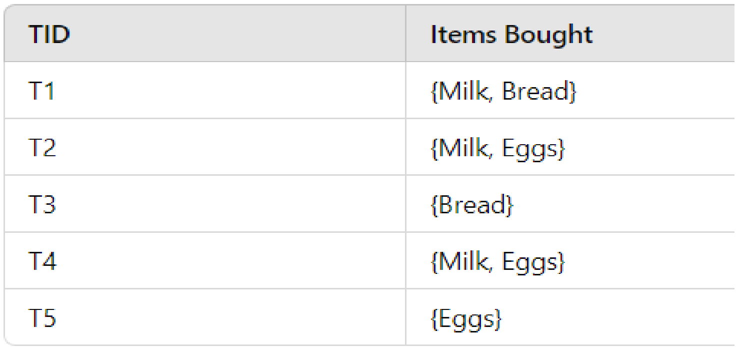
iii. Strong Association Rules

1. Once frequent itemsets are identified, the strong rules are those with confidence ≥ 80%. Would you like to proceed with rule generation from frequent itemsets?

b. Association rule mining

i. In association rule mining, a strong association rule is one that meets the minimum support and confidence thresholds. However, this does not always imply a positive correlation between items. Items can still be negatively correlated even if they form a strong rule.

ii. Example: Let's assume the following transactions in a database:



TID	Items Bought
T1	{Milk, Bread}
T2	{Milk, Eggs}
T3	{Bread}
T4	{Milk, Eggs}
T5	{Eggs}

iii. **Min Support** = 10%
Min Confidence = 50%

iv. Now, consider the rule:
{Milk} ⇒ {Bread}

v. Step 1: Calculate support and confidence
   1. Support of {Milk}: Appears in T1, T2, T4 → Support = 3/5 = 60%
   2. Support of {Bread}: Appears in T1, T3 → Support = 2/5 = 40%
   3. Support of {Milk, Bread}: Appears only in T1 → Support = 1/5 = 20%
   4. Confidence of {Milk} ⇒ {Bread} = Support({Milk, Bread}) / Support({Milk}) = (1/5) / (3/5) = 33.33%

vi. Since the confidence is below 50%, this rule is not strong in this case

vii. Example of Negative Correlation (with a strong rule):
   1. Now, consider the rule {Milk} ⇒ {Eggs}:
      a. upport of {Milk, Eggs}: Appears in T2, T4 → Support = 2/5 = 40%
      b. Confidence of {Milk} ⇒ {Eggs} = Support({Milk, Eggs}) / Support({Milk}) = (2/5) / (3/5) = 66.67%
      c. This rule has confidence above 50%, so it is a strong rule

viii. Negative Correlation:
   1. To check if there is negative correlation, we calculate the Lift of the rule:
      a. Lift = Confidence({Milk} ⇒ {Eggs}) / Support({Eggs})
      b. Support of {Eggs} = 3/5 (appears in T2, T4, T5)
   2. Lift = (2/3) / (3/5) = 10/9 ≈ 1.11

ix. In conclusion, while the rule {Milk} ⇒ {Eggs} is strong, it is not necessarily positively correlated—it's nearly independent based on the lift value.

c. To list all strong association rules that satisfy the given condition (support s and confidence c meeting the minimum requirements), we will use the results from part (a) (frequent itemsets) and generate rules of the form
   i. ∀X ∈ transaction, buys(X,item1) ∧ buys(X,item2) ⟹ buys(X,item3)

Given Conditions:

- **Min Support** s = 60%
- **Min Confidence** c = 80%

d. We will generate rules from the frequent itemsets we obtained earlier:
   i. Frequent 2-itemsets: {M, K}, {O, K}, {O, E}, {K, E}, {K, Y}
   ii. Frequent 3-itemset: {O, K, E}

e. For each rule, we'll check if the confidence is greater than or equal to 80%.
   i. Step 1: Generate Rules from the Frequent 3-itemset {O, K, E}

Rule 1: {O, K} ⇒ {E}

- Support({O, K, E}) = 3/5 = 60%
- Confidence({O, K} ⇒ {E}) = Support({O, K, E}) / Support({O, K}) = (3/5) / (3/5) = 100%
- **Strong rule** because confidence = 100% ≥ 80%

Rule 2: {O, E} ⇒ {K}

- Support({O, K, E}) = 3/5 = 60%
- Confidence({O, E} ⇒ {K}) = Support({O, K, E}) / Support({O, E}) = (3/5) / (3/5) = 100%
- **Strong rule** because confidence = 100% ≥ 80%

Rule 3: {K, E} ⇒ {O}

- Support({O, K, E}) = 3/5 = 60%
- Confidence({K, E} ⇒ {O}) = Support({O, K, E}) / Support({K, E}) = (3/5) / (4/5) = 75%
- **Not a strong rule** because confidence = 75% < 80%

Step 2: Check Rules from Frequent 2-itemsets

We now check if any rules can be derived from frequent 2-itemsets that meet the confidence threshold.

- No rules can be generated from the 2-itemsets since they don't have enough other items in the same frequent itemset to form a rule of the form

Final List of Strong Association Rules:

1. **{O, K}** ⇒ **{E}** (Support = 60%, Confidence = 100%)
2. **{O, E}** ⇒ **{K}** (Support = 60%, Confidence = 100%)

3. Association rule

- Given:
  - Min Support = 25%
  - Min Confidence = 50%
- We need to check if the rule "hot dogs ⇒ hamburgers" meets these thresholds.

  From the contingency table:

  - Transactions where both hot dogs and hamburgers are bought: 2000
  - Total transactions: 5000
  - Transactions with hot dogs: 3000

Step 1: Calculate Support

Support is the proportion of transactions that contain both hot dogs and hamburgers:

Support = 2000 / 5000 = 0.4 = 40%

- Since 40% > 25%, the rule satisfies the support threshold.

Step 2: Calculate Confidence

Confidence is the proportion of transactions with hot dogs that also contain hamburgers:

Confidence = 2000 / 3000 = 0.67 = 67%

Since 67% > 50%, the rule satisfies the confidence threshold.

Conclusion for Part (a):

Yes, the association rule "hot dogs ⇒ hamburgers" is strong because it meets both the minimum support (40% > 25%) and minimum confidence (67% > 50%).

Part (d): Is the purchase of hot dogs independent of the purchase of hamburgers?

We will use the **Chi-square test** to determine whether the purchase of hot dogs is independent of the purchase of hamburgers.

Step 1: Expected Values

To calculate the expected values, we use the formula:

Eij = (row total x grand total) / (grand total)

E11 (Hot dogs and hamburgers) = 2000 x 3000 / 5000 = 1200

E12 (Hot dogs and no burgers) = 3000 x 3000 / 5000 = 1800

E21 (No hot dogs and burgers) = 2000 x 2000 / 5000 = 800

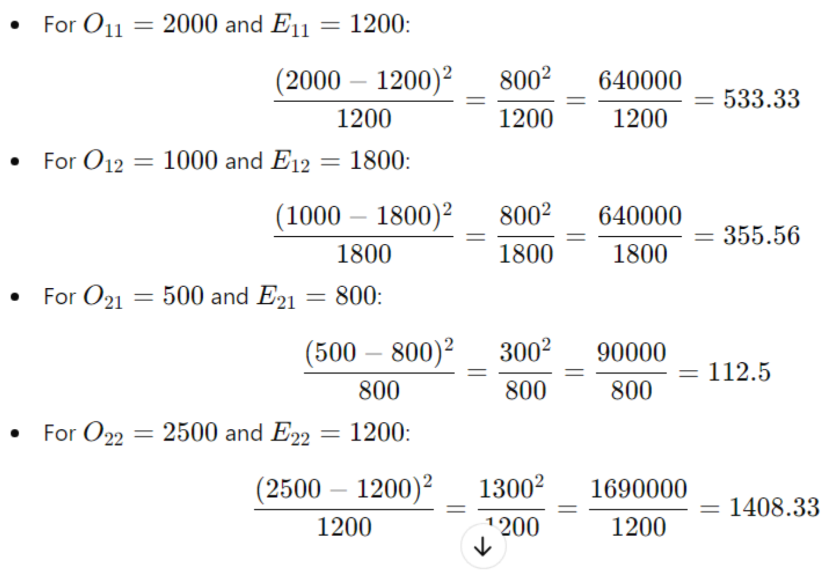
E22 (No hot dogs and no burgers) = 2000 x 3000 / 5000 = 1200

Step 2: Calculate Chi-square Statistic

The Chi-square statistic is calculated using:

X^2 = ∑ (Oij - Eij)^2 / Eij

Where Oij is observed value and Eij is expected value



- For $O_{11} = 2000$ and $E_{11} = 1200$:

$$\frac{(2000 - 1200)^2}{1200} = \frac{800^2}{1200} = \frac{640000}{1200} = 533.33$$

- For $O_{12} = 1000$ and $E_{12} = 1800$:

$$\frac{(1000 - 1800)^2}{1800} = \frac{800^2}{1800} = \frac{640000}{1800} = 355.56$$

- For $O_{21} = 500$ and $E_{21} = 800$:

$$\frac{(500 - 800)^2}{800} = \frac{300^2}{800} = \frac{90000}{800} = 112.5$$

- For $O_{22} = 2500$ and $E_{22} = 1200$:

$$\frac{(2500 - 1200)^2}{1200} = \frac{1300^2}{1200} = \frac{1690000}{1200} = 1408.33$$

Adding all values together:

X^2 = 533.33+355.56+112.5+1408.33 = 2409.72

Step 3: Compare with Critical Value

Using the chi-square table for 1 degree of freedom (df = 1) and a significance level of alpha = 0.05 the critical value is 3.84

Since 2409.72 >> 3.84, we reject null hypothesis of independence

Conclusion for Part (d):

The purchase of hot dogs is **not independent** of the purchase of hamburgers. Based on the contingency table, they are **positively correlated** because the observed co-occurrence of hot dogs and hamburgers is much higher than expected under independence.

e. Association rule metrics

- All Confidence, Max Confidence, Kulczynski, Cosine, and Lift for the rule "hot dogs ⇒ hamburgers" using the given contingency table

Contingency Table Recap:

	Hot Dog	No Hot Dog	Row Total
Hamburgers	2000	500	2500
No Hamburgers	1000	1500	2500
Col Total	3000	2000	5000

Basic Measures:

- Support (hot dogs & hamburgers) = $\frac{2000}{5000} = 0.4$ (40%)
- Support (hot dogs) = $\frac{3000}{5000} = 0.6$ (60%)
- Support (hamburgers) = $\frac{2500}{5000} = 0.5$ (50%)
- Confidence (hot dogs ⇒ hamburgers) = $\frac{2000}{3000} = 0.67$ (67%)
- Confidence (hamburgers ⇒ hot dogs) = $\frac{2000}{2500} = 0.8$ (80%)

-

1. **All Confidence**

The **All Confidence** measure is the minimum of the two confidence values for the rule and its inverse. It is defined as:

$$\text{All Confidence} = \frac{|X \cap Y|}{\max(|X|, |Y|)}$$

In this case:

$$\text{All Confidence} = \frac{2000}{\max(3000, 2500)} = \frac{2000}{3000} = 0.67$$

-

2. **Max Confidence**

The **Max Confidence** measure is the maximum of the two confidence values for the rule and its inverse. It is defined as:

$$\text{Max Confidence} = \max\left(\frac{|X \cap Y|}{|X|}, \frac{|X \cap Y|}{|Y|}\right)$$

In this case:

$$\text{Max Confidence} = \max\left(\frac{2000}{3000}, \frac{2000}{2500}\right) = \max(0.67, 0.8) = 0.8$$

-

3. **Kulczynski Measure**

The **Kulczynski** measure is the average of the two confidence values for the rule and its inverse. It is defined as:

$$\text{Kulczynski} = \frac{1}{2}\left(\frac{|X \cap Y|}{|X|} + \frac{|X \cap Y|}{|Y|}\right)$$

In this case:

$$\text{Kulczynski} = \frac{1}{2}\left(\frac{2000}{3000} + \frac{2000}{2500}\right) = \frac{1}{2}(0.67 + 0.8) = \frac{1}{2} \times 1.47 = 0.735$$

-

4. **Cosine Measure**

The **Cosine** measure is the geometric mean of the two confidence values, or equivalently, the normalized support. It is defined as:

$$\text{Cosine} = \frac{|X \cap Y|}{\sqrt{|X| \cdot |Y|}}$$

In this case:

$$\text{Cosine} = \frac{2000}{\sqrt{3000 \times 2500}} = \frac{2000}{\sqrt{7500000}} = \frac{2000}{2738.61} \approx 0.73$$

-

5. Lift

The **Lift** measure shows how much more likely it is for the items to co-occur than if they were independent. It is defined as:

$$\text{Lift} = \frac{\text{Support}(X \cap Y)}{\text{Support}(X) \times \text{Support}(Y)}$$

In this case:

$$\text{Lift} = \frac{\frac{2000}{5000}}{\frac{3000}{5000} \times \frac{2500}{5000}} = \frac{0.4}{0.6 \times 0.5} = \frac{0.4}{0.3} = 1.33$$

-

Summary of Measures:

Measure	Value
All Confidence	0.67
Max Confidence	0.8
Kulczynski	0.735
Cosine	0.73
Lift	1.33

-