# Homework 2

## Satwik Singh satwiks2

## Question 1

1. 
   - D1:{13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70}
     - Bin1: [13,15,16] Bin4: [22,25,25] Bin7: [35,35,35]
       Bin2: [16,19,20] Bin5: [25,25,30] Bin8: [36,40,45]
       Bin3: [20,21,22] Bin6: [33,33,35] Bin9: [46,52,70]
     - Mean1: 14.667 Mean4: 24 Mean7: 35
       Mean2: 18.333 Mean5: 26.667 Mean8: 40.333
       Mean3: 21 Mean6: 33.667 Mean9: 56
     - Bin1: [14.667,14.667,14.667] Bin4: [24,24,24] Bin7: [35,35,35]
       Bin2: [18.333,18.333,18.333] Bin5: [26.667,26.667,26.667] Bin8: [40.333,40.333,40.333]
       Bin3: [21,21,21] Bin6: [33.667,33.667,33.667] Bin9: [56,56,56]
   - D2:{5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215}
     - Bin1: [5,10,11] Bin2: [13,15,35] Bin3: [50,55,72] Bin4: [92,204,215]
     - Mean1: 8.667 Mean2: 21 Mean3: 59 Mean4: 170.333

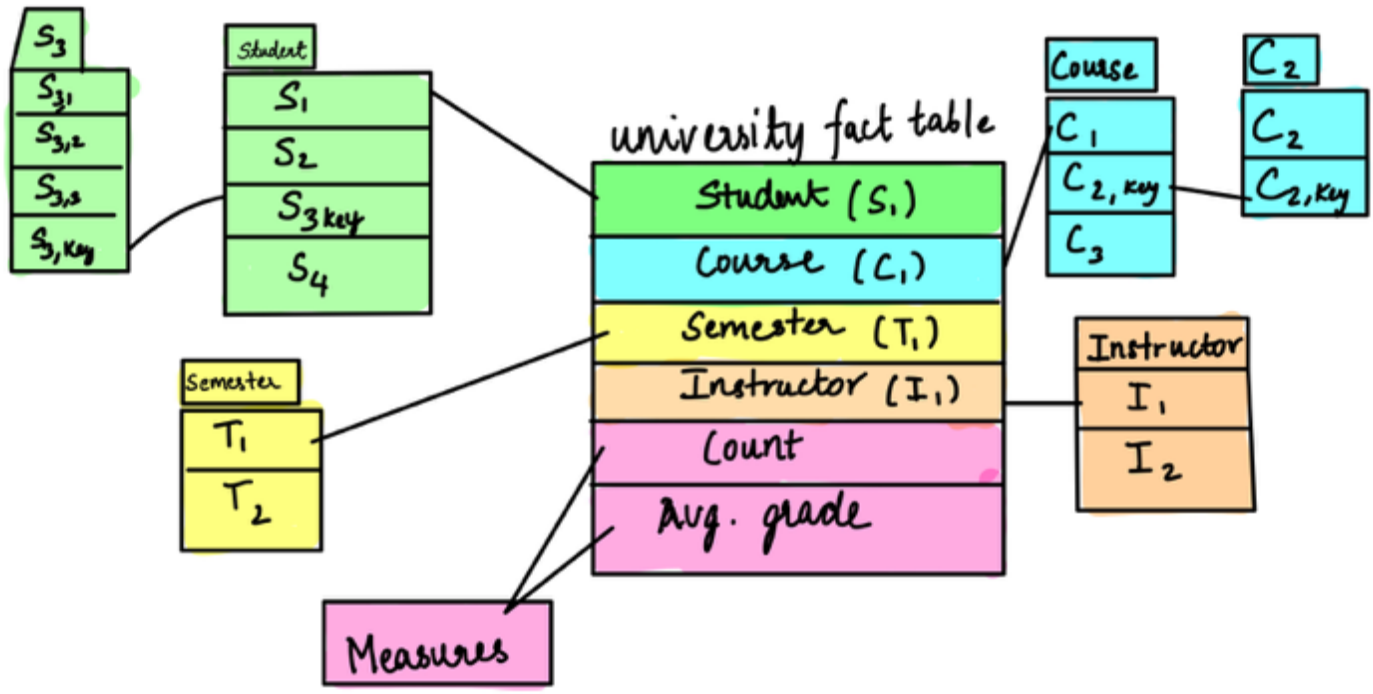
This method performs local smoothing, i.e takes neighboring data points into account when smoothing instead of the entire dataset.

2. 
   - Equal Frequency:
     - D1:{13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70}
       - Bin1: [13, 15, 16, 16, 19, 20, 20, 21, 22, 22]
       - Bin2: [25, 25, 25, 25, 30, 33, 33]
       - Bin3: [35, 35, 35, 35, 36, 40, 45, 46, 52, 70]

    - D2:{5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215}
        - Bin1: [5, 10, 11, 13]
        - Bin2: [15, 35, 50, 55]
        - Bin3: [72, 92, 204, 215]
  - Equal width $w = (max - min)/k$
    - D1:{13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70}
        - Bin1: [13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25]
        - Bin2: [30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52]
        - Bin3: [70]
    - D2:{5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215}
        - Bin1: [5, 10, 11, 13, 15, 35, 50, 55, 72]
        - Bin2: [92]
        - Bin3: [204, 215]

# Question 2

1. Snowflake schema diagram



2. OLAP operations
    1. Roll-up on course from courses to department.
    2. Roll-up on semester from semesters to all.
    3. Slice for CS course.
3. Cube has $5^4 = 625$ cuboids
    - Total number of cuboids = $\prod_{i=1}^{n}(L_i + 1)$ hence 4 dimensions and 5 (including all) levels.

# Question 3

1. min sup = 1
    - 5 Closed Patterns
      $P_1 = \{a_1, .., a_{12}\} : 3, P_2 = \{a_1, .., a_{20}\} : 2, P_3 = \{a_1, .., a_{30}\} : 1, P4 = \{a_{10}, a_{11}, a_{20}\} : 3, P5 = \{a_{10}, a_{11}\} : 4$
    - 1 Max patterns
      $P_3 = \{a_1, .., a_{30}\} : 1$
2. min sup = 2
    - 4 Closed Patterns
      $P_1 = \{a_1, .., a_{12}\} : 3, P_2 = \{a_1, .., a_{20}\} : 2, P_3 = \{a_{10}, a_{11}, a_{20}\} : 3, P_4 = \{a_{10}, a_{11}\} : 4$
    - 1 Max patterns
      $P_2 = \{a_1, .., a_{20}\} : 2$
3. min sup = 4
    - 1 Closed Patterns
      $P_5 = \{a_{10}, a_{11}\} : 4$
    - 1 Max patterns
      $P_5 = \{a_{10}, a_{11}\} : 4$

# Question 4

1. $Support(A \Rightarrow B) = P(A \cup B)$ and $confidence(A \Rightarrow B) = P(B|A)$
   thus Support = (number of transactions with A and B)/total transactions = 4/11 = 0.364,
   confidence = sup(A,B)/sup(A) = (4/11) / (8/11) = 0.5

TID Items
T1 A,B,C

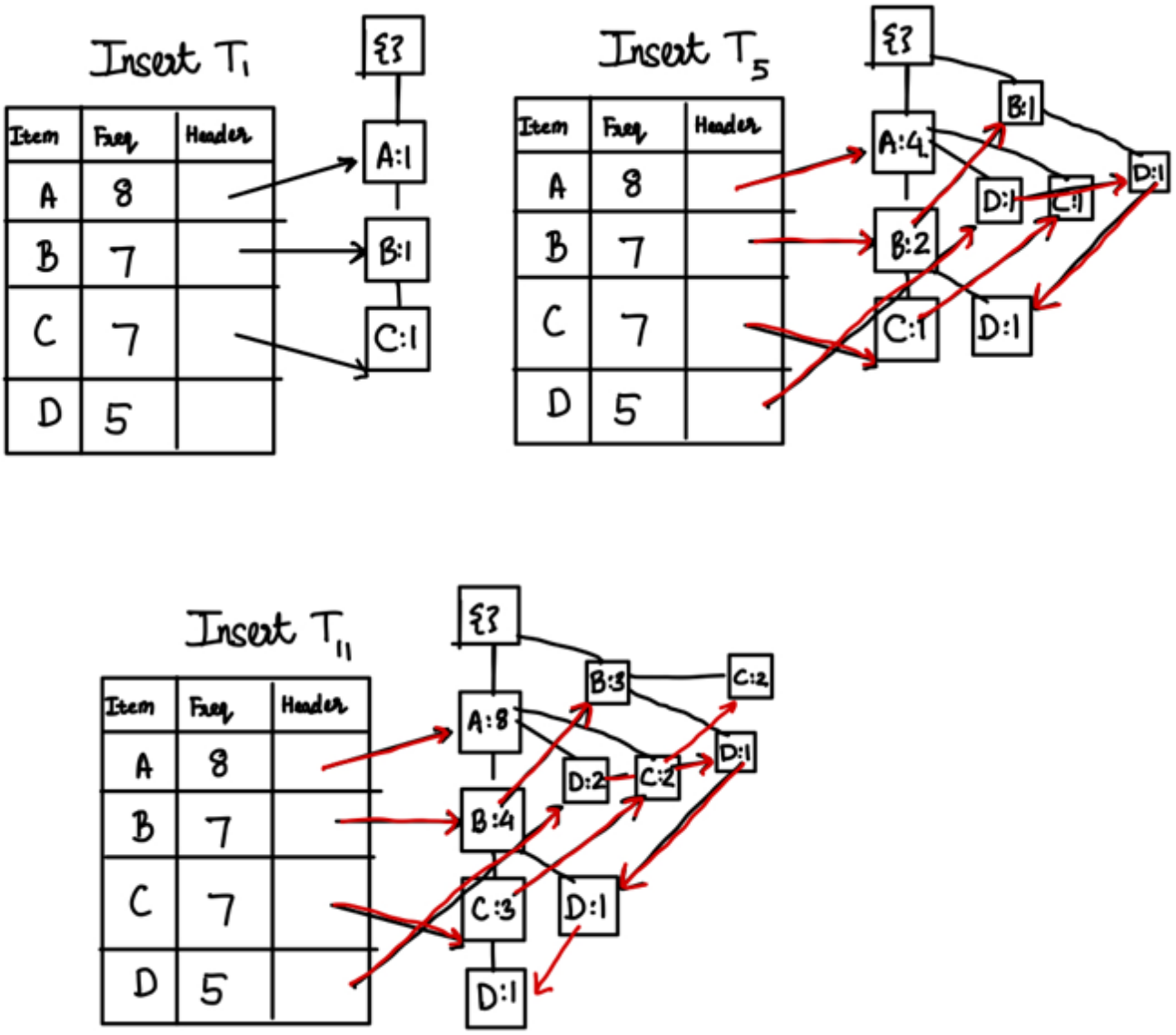
T2 A,D,E
T3 B,D
T4 A,B,D
T5 A,C
T6 B,C
T7 A,C
T8 A,B,C,D,E
T9 B,C
T10 A,D
T11 A,B,C

2. Using apriori on {A,B,C,D,E}:
    - First scan 1 item sets (filtering for minimum support = 3):
        - {A}:8, {B}:7, {C}:7, {D}:5
    - Second scan 2 item sets (filtering for minimum support = 3):
        - {A,B}:4, {A,C}: 5, {A,D}: 4, {B,C}: 5, {B,D}: 3
    - Third scan 3 items sets (filtering for minimum support = 3):
        - {A,B,C}:3

- Using formulae $Support(A \Rightarrow B) = P(A \cup B)$ and $confidence(A \Rightarrow B) = P(B|A)$:
  For A,B ⇒ C(s,c), s = 3/11 = 0.273,c = 3/4 = 0.75
  For A,C ⇒ B(s,c), s = 3/11 = 0.273,c = 3/5 = 0.60
  For B,C ⇒ A(s,c), s = 3/11 = 0.273,c = 3/5 = 0.60

3. FP tree
    - Scan the table and get the freq of items that meet the minimum support 3 and sort in ascending order, {A}:8, {B}:7, {C}:7, {D}:5
    - Scan DB again, find the ordered frequent itemlist for each transaction
      TID Items ------|| Ordered freq itemlist
      T1 A,B,C ------ || A,B,C
      T2 A,D,E -------|| A,D
      T3 B,D -------- || B,D
      T4 A,B,D ------ || A,B,D
      T5 A,C -------- || A,C
      T6 B,C -------- || B,C

T7 A,C -------- || A,C
T8 A,B,C,D,E - || A,B,C,D
T9 B,C -------- || B,C
T10 A,D ------- || A,D
T11 A,B,C ----- || A,B,C




4. Take the paths from the FP tree and calculating the support for each path. The conditional FP-tree can also be generated by taking the most common path from the conditional patterns. Therefore, for D, the frequent pattern generated is: {< A, D : 5 >}; for C, the frequent pattern generated is:{< A,C : 5 >< B,C : 2 >< A,B,C : 2 >}; for B, the frequent pattern generated is: {< A,

B : 5 >}.

# Question 5

1. $Kulc(A,B) = (P(a|b) + P(b|a))/2 = \frac{1}{2}(\frac{\frac{a}{(a+b+c+d)}}{\frac{(a+c)}{(a+b+c+d)}} + \frac{\frac{a}{(a+b+c+d)}}{\frac{(a+b)}{(a+b+c+d)}}) = \frac{1}{2}(\frac{a}{a+c} + \frac{a}{a+b})$
   - It can be seen that the denominator of each support calculation is canceled out and not affect the calculations.
   - This allows the count for the null transactions to drop out, thus this calculation is null invariant.
   - The count for the null transactions, the transactions that do not contain A or B, is represented by d, and this measure is not dependent on d, so it is null invariant.
2. $Lift(A,B) = \frac{s(A \cup B)}{(s(A)s(B))} = \frac{a(a+b+c+d)}{(a+c)(a+b)}$
   - Since d is present is in this equation, the equation will always be dependent to the amount of null transactions. Hence it is not null invariant.
3. $Cosine(A,B) = \frac{s(A \cup B)}{\sqrt{(s(A)s(B))}} = \frac{a}{\sqrt{(a+c)(a+b)}}$
   - since the equation doesn't depend on d hence it is null invariant.