1.

(a)

Each bin has depth of 3. Dividing data into bins:

Since the data is sorted in asceding order, we put group of 3 in each bin accordingly.

D_1:

① partition into equal-depth bins

Bin 1: 13, 15, 16
Bin 2: 16, 19, 20
Bin 3: 20, 21, 22
Bin 4: 22, 25, 25
Bin 5: 25, 25, 30
Bin 6: 33, 33, 35
Bin 7: 35, 35, 35
Bin 8: 36, 40, 45
Bin 9: 46, 52, 70

→

② Smoothing by bin means:

Bin 1: 15, 15, 15
Bin 2: 18, 18, 18
Bin 3: 21, 21, 21
Bin 4: 24, 24, 24
Bin 5: 27, 27, 27
Bin 6: 34, 34, 34
Bin 7: 35, 35, 35
Bin 8: 40, 40, 40
Bin 9: 56, 56, 56

D_2:

① Bin 1: 5, 10, 11
Bin 2: 13, 15, 35
Bin 3: 50, 55, 72
Bin 4: 92, 204, 215

smoothing →

②
Bin 1: 9, 9, 9
Bin 2: 21, 21, 21
Bin 3: 59, 59, 59
Bin 4: 170, 170, 170

(b) In this method each element in a bin is replaced by mean value of the bin and therefore help remove noise from data. It can help provide more accurate results. This allows important patterns to more clearly stand out. This helps to predict trends.

(c)

i. Equal - frequency partitioning.

D_1: $27 \div 3 = 9$ numbers/bin

Bin 1: 13. 15. 16. 16. 19. 20. 20. 21. 22
Bin 2: 22. 25. 25. 25. 25. 30. 33. 33. 35
Bin 3: 35. 35. 35. 36. 40. 45. 46. 52. 70

D_2: $12 \div 3 = 4$ numbers/bin

Bin 1: 5. 10. 11. 13
Bin 2: 15. 35. 50. 55
Bin 3: 72. 92. 204. 215

we ~~divide~~ divide numbers into 3 bins, each has same number of data.

By using this meathod, we have better data scaling because interval has same number of samples.

ii. Equal - width partitioning

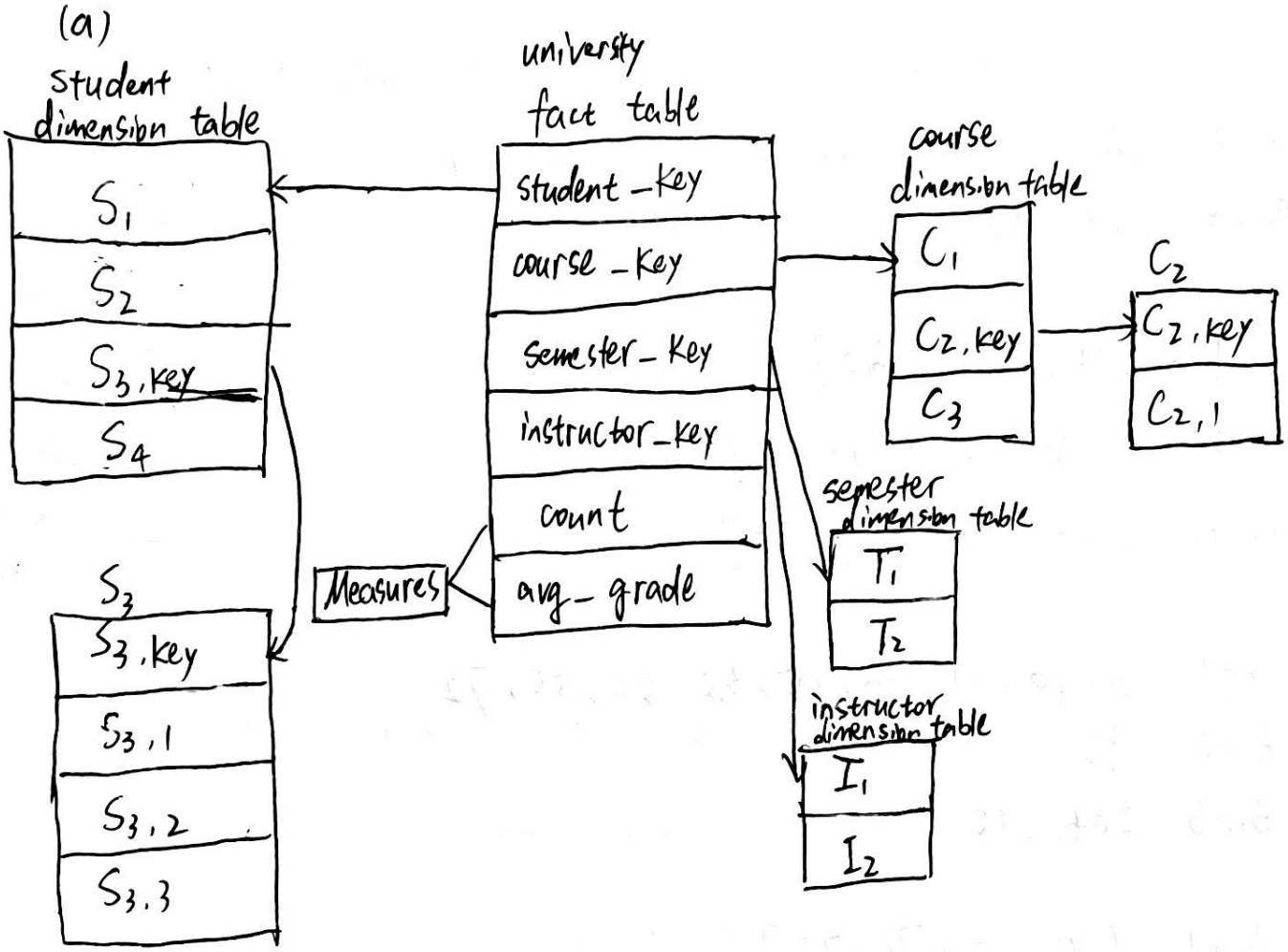
Divides the range into 3 intervals of equal size: uniform grid

the width of D_1 is: $\frac{70 - 13}{3} = 19$

the width of D_2 is: $\frac{215 - 5}{3} = 70$

2.

(a)



(b) We should perform the following OLAP operations:

1. Roll-up on course from course_key to department
2. Roll-up on semester from semester_key to all
3. Slice for course = "Computer Science"

(c) We have 4 dimensions and each has 5 levels

$\Rightarrow$ cube will contain $5^4 = 625$ cuboids

$$= \prod_{i=1}^{4} (L_i + 1) = 5 \times 5 \times 5 \times 5 = 625$$

including the base and apex cuboids.

D_1:

Bin 1: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30

Bin 2: 33, 33, 35, 35, 35, 35, 36, 40, 45, 46

Bin 3: 52, 70

Reason:

Bin 1: [13, 13+19=32]

Bin 2: [min, 13+2×19=51]

Bin 3: [min, 13+3×19=70]

D_2:

Bin 1: 5, 10, 11, 13, 15, 35, 50, 55, 72

Bin 2: 92

Bin 3: 204, 215

Reason:

Bin 1: [5, 5+70=75]

Bin 2: [min, 5+2×70=145]

Bin 3: [min, 5+3×70=215]

This meathod is the most straightforward, but outliers may dominate presentation. Besides, skewed data is not handled well using this meathod.

3.

closed pattern: a pattern X is closed if X is frequent, and there exists no super-pattern $Y \supset X$, with same support of X.

max pattern: a pattern X is a maximal pattern if X is frequent and there exists no frequent super-pattern $Y \supset X$

Since closed pattern is a lossless compression of frequent patterns, we can find max pattern from closed patterns.

(a) minsup = 1, find pattern with occurrences ≥ 1

4 closed patterns:

P_1: "$\{a_1, a_2, \cdots, a_{12}\} : 3$"

P_2: "$\{a_{10}, a_{11}, a_{20}\} : 3$"

P_3: "$\{a_1, a_2, \cdots a_{20}\} : 2$"

P_4: "$\{a_1, a_2, \cdots a_{30}\} : 1$"

1 maximum patten:

P: "$\{a_1, a_2, \cdots a_{30}\} : 1$"

30-itemsets, the largest among closed patterns.

(b) minsup = 2, find pattern with occurrence ≥ 2

3 closed patterns:

P_1: "$\{a_1, a_2 \cdots a_{12}\} : 3$"

P_2: "$\{a_{10}, a_{11}, a_{20}\} : 3$"

P_3: "$\{a_1, a_2, \cdots a_{20}\} : 2$"

1 maximal pattern

P: "$\{a_1, a_2, \cdots a_{20}\} : 2$"

20-itemsets, largest among closed patterns.

(c) minsup = 4, find pattern with occurrence ≥ 4

there are 4 transactions in total, the pattern must happen in every transaction.

$T_1 \cap T_2 \cap T_3 \cap T_4 = \{a_{10}, a_{11}\}$

$\Rightarrow$ 1 closed pattern P = "$\{a_{10}, a_{11}\} : 4$"

1 maximal pattern P: "$\{a_{10}, a_{11}\} : 4$"

4.

(a)

there is a total of 4 transactions contains subset $\{A, B\}$ out of 11 transactions in total

$$\Rightarrow S(A, B) = \frac{4}{11}$$

there is a total of 8 transactions contains subset $\{A\}$ out of 11 transactions in total

$$\Rightarrow S(A) = \frac{8}{11}$$

$\Rightarrow$ confidence c of $A \Rightarrow B$:

$$C = \frac{S(A,B)}{S(A)} = \frac{\frac{4}{11}}{\frac{8}{11}} = \frac{1}{2} = 50\%$$

(b)

All frequent itemsets from Apriori Algorithm:

$\{A\}: 8$ $\{B\}: 7$ $\{C\}: 7$ $\{D\}: 5$

$\{A, B\}: 4$ $\{A, C\}: 5$ $\{A, D\}: 4$ $\{B, C\}: 5$ $\{B, D\}: 3$

$\{A, B, C\}: 3$

(b)

TID	Items
T_1	A, B, C
T_2	A, D, E
T_3	B, D
T_4	A, B, D
T_5	A, C
T_6	B, D
T_7	A, C
T_8	A, B, C, D, E
T_9	B, C
T_{10}	A, D
T_{11}	A, B, C

C_1 — First scan →

Itemset	Sup
{A}	8
{B}	7
{C}	7
{D}	5
{E}	2

F_1 — filter, minsup=3 →

Itemset	Sup
{A}	8
{B}	7
{C}	7
{D}	5

↓

Itemset
{A, B}
{A, C}
{A, D}
{B, C}
{B, D}
{C, D}

C_2 — 2nd scan →

Itemset	Sup
{A, B}	4
{A, C}	5
{A, D}	4
{B, C}	5
{B, D}	3
{C, D}	1

C_2 →

Itemset	Sup
{A, B}	4
{A, C}	5
{A, D}	4
{B, C}	5
{B, D}	3

↓ C_3

Itemset
{A, B, C}
{A, B, D}
{A, C, D}
{B, C, D}

C_3 — 3rd scan →

Itemset	sup
{A, B, C}	3
{A, B, D}	2
{A, C, D}	1
{B, C, D}	1

F_3 — filter →

Itemset	sup
{A, B, C}	3

(c)

1. Scan DB once, find all single item frequency pattern:

   A: 8 B: 7 C: 7 D: 5 E: 2

   minsup = 3

2. Sort frequency items in descendy order

   A-list = A - B - C - D

TID	Items	Ordered itemlist
T_1	A, B, C	A, B, C
T_2	A, D, E	A, D
T_3	B, D	B, D
T_4	A, B, D	A, B, D
T_5	A, C	A, C
T_6	B, C	B, C
T_7	A, C	A, C
T_8	A, B, C, D, E	A, B, C, D
T_9	B, C	B, C
T_{10}	A, D	A, D
T_{11}	A, B, C	A, B, C

minsup = 3, exclude E

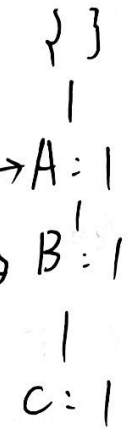
3. Scan DB again and find the ordered frequent itemlist for each transaction

4. Insert the ordered frequent itemlist into an FP-tree for each transaction, with shared sub-branches merged, counts accumulated.

i). After inserting T_1.

Header Table

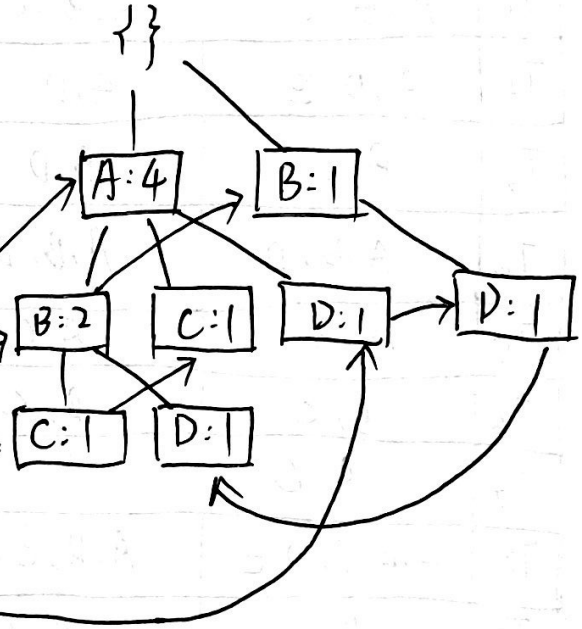
Itemset	frequency	header
{A}	8	
{B}	7	
{C}	7	
{D}	5	



ii) Inserting T_5:

Header Table

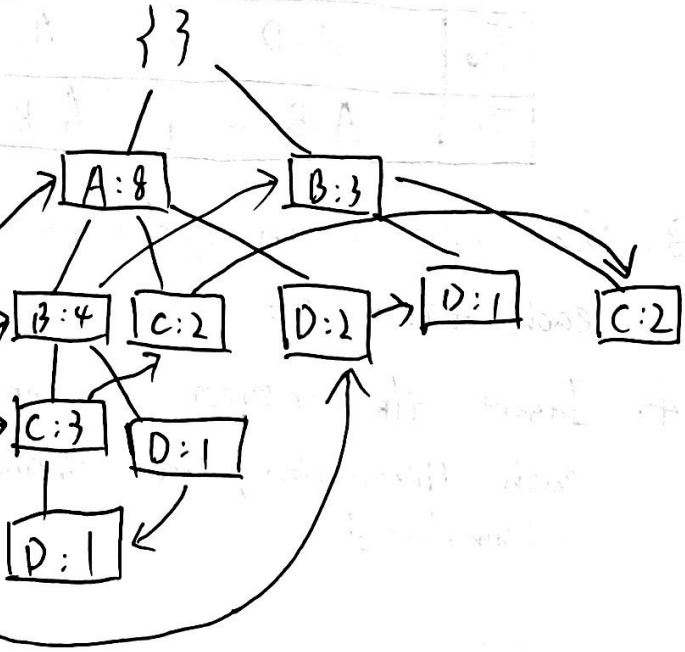
Itemset	frequency	header
{A}	8	
{B}	7	
{C}	7	
{D}	5	



iii) Inserting T_{11}:

Header Table

Itemset	Frequency	Header
{A}	8	
{B}	7	
{C}	7	
{D}	5	



(d)

To calculate the conditional database, we start from bottom to top.

As we know, conditional database is the database under the condition that D exists.

D's conditional database, that pattens containing D as follows:

A:2 ABC:1 AB:1 B:1

C's conditional database:

AB:3 A:2 B:2

B's conditional database:

A:4

Since A is the most frequent item, we don't need its conditional DB.

Item	Conditional database	Conditional FP-tree	Frequent Pattens Generated
D	{{A:2}, {ABC:1}, {AB:1}, {B:1}}	<A:4, B:3>, <B:1>	{A,D:4}, {B,D:3}
C	{{AB:3}, {A:2}, {B:2}}	<AB:5>, <B:2>	{A,C:5}, {B,C:5}, {A,B,C:3}
B	{{A:4}}	<A:4>	{A,B:4}

⇒ All itemsets using FP-growth algorithm:

{A}:8 {B}:7 {C}:7 {D}:5

{A,B}:4 {A,C}:5 {A,D}:4 {B,C}:5 {B,D}:3 {A,B,C}:3

5.

(a)

$$Kulc(A,B) = \frac{1}{2}\left(P(A|B) + P(B|A)\right)$$

$$= \frac{1}{2}\left(\frac{S(A\cup B)}{S(A)} + \frac{S(A\cup B)}{S(B)}\right)$$

$$= \frac{1}{2}\left(\frac{\frac{a}{a+b+c+d}}{\frac{a+c}{a+b+c+d}} + \frac{\frac{a}{a+b+c+d}}{\frac{a+b}{a+b+c+d}}\right)$$

$$= \frac{1}{2}\left(\frac{a}{a+c} + \frac{a}{a+b}\right)$$

Since the final furmula $\frac{1}{2}\left(\frac{a}{a+c} + \frac{a}{a+b}\right)$ has nothing to do with d, we know that $Kulc(A,B)$ is null invariant. $Kulc(A,B)$ only related to the supports of A, B and $A\cup B$, but not related to total number.

(b) $$Lift(A,B) = \frac{S(A\cup B)}{S(A)\times S(B)} = \frac{P(A\cup B)}{P(A)\cdot P(B)} = \frac{\frac{a}{a+b+c+d}}{\frac{a+c}{a+b+c+d}\times\frac{a+b}{a+b+c+d}}$$

$$= \frac{a(a+b+c+d)}{(a+c)(a+b)}$$

As we can see, the result is related to d, so it is not null invariance.

Based on Lift(A,B), A,B will be considered independent if $P(A \cup B) = P(A)P(B)$

$$\Rightarrow \frac{a}{a+b+c+d} = \frac{a+c}{a+b+c+d} \times \frac{a+b}{a+b+c+d}$$

$$\Rightarrow ad = bc$$

$\Rightarrow$ A, B will be considered independent if $ad = bc$ in this contingency table.

(c) $$\text{Lift}(A,B) = \frac{P(A \cup B)}{P(A) \cdot P(B)} = \frac{S(A \cup B)}{S(A) \times S(B)}$$

$$\text{cosine}(A,B) = \frac{P(A \cup B)}{\sqrt{P(A) \times P(B)}} = \frac{S(A \cup B)}{\sqrt{S(A) \times S(B)}} = \sqrt{P(A|B) \times P(B|A)}$$

For cosine, the square root is taken on the product of the probabilities of A and B. The cosine measure can be viewed as a harmonized lift measure.

By taking the square root, the cosine value is only influenced by the supports of A, B and $A \cup B$, and not by the total number of transactions. Therefore, such difference make Cosine(A,B) null invariant.