1.

(a) Smoothing of D_1

Bin1 : 13 15 16

mean of Bin1 : $\frac{13+15+16}{3} \approx 15$

Smoothing by bin mean :

Bin 1 : 15 15 15

Iterating the steps above for the next Bins 2~9

B2 : 18 18 18 B6 : 34 34 34

B3 : 21 21 21 B7 : 35 35 35

B4 : 24 24 24 B8 : 40 40 40

B5 : 27 27 27 B9 : 56 56 56

Smoothing of D2

B1 : 9 9 9 B4 : 170 170 170

B2 : 21 21 21

B3 : 59 59 59

(b)

Comment-
Smoothing by bin means has a higher performance of approximation in D2, because according to the calculations, the variances of the bin in D1 and D2 are relatively small and large, this technique is better to use in a dataset with small variance, otherwise this will cause huge errors in the analysis of processed data.

Mean of variances of all bins in D_1 =

$$\frac{1}{N}\sum\sigma^2 = \frac{1}{9}\sum_{j=1}^{9}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} = \frac{1}{9}\times(1.56+2.89+0.67+2+5.56+0.89+0+13.56+104) = 14.57$$

Means of variances of all bins in D_2 =

$$\frac{1}{N}\sum\sigma^2 = \frac{1}{4}(6.89+98.67+88.67+3088.22) \approx 820.61$$

(c) Equal-frequency partitioning

In D_1

B_1 : 13 15 16 16 19 20 20 21 22

B_2 : 22 25 25 25 25 30 33 33 35

B_3 : 35 35 35 36 40 45 46 52 70

In D_2

B_1 : 5 10 11 13

B_2 : 15 35 50 55

B_3 : 72 92 204 215

Mean of variances of all bins in D_1 =

$$\frac{1}{N}\sum\sigma^2 = \frac{1}{3}\times(8.44+19.43+118.62) = 48.83$$

Means of variances of all bins in D_2 =

$$\frac{1}{N}\sum\sigma^2 = \frac{1}{3}\times(8.69+242.19+4129.19) = 1460.02$$

Equal-width partitioning

In D_1

width : $\frac{70-13}{3} = 19$

B_1 : (13 ~ 32)
13 15 16 16 19 20 21 22 22
25 25 25 25 30

B2 : (33 ~ 51)
33 33 35 35 35 35 36 40 45
46

B3 : 70 52

In D_2 :
width : $\frac{215-5}{3} = 70$

B_1 : (5 ~ 74) 5 10 11 13 15 35 50 55 72
(75~144) 92
(145~215) 204 215

Mean of variances of all bins in D_1 =

$$\frac{1}{N}\sum\sigma^2 = \frac{1}{3}\times(21.57+20.21+81) \approx 40.93$$

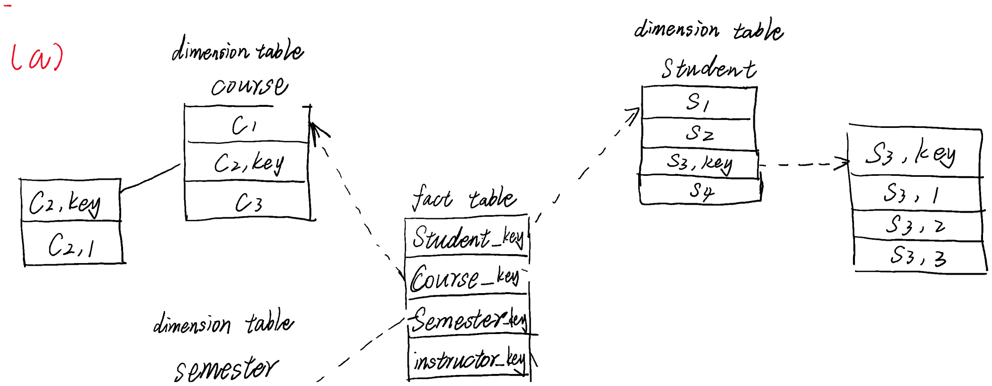
Means of variances of all bins in D_2 =

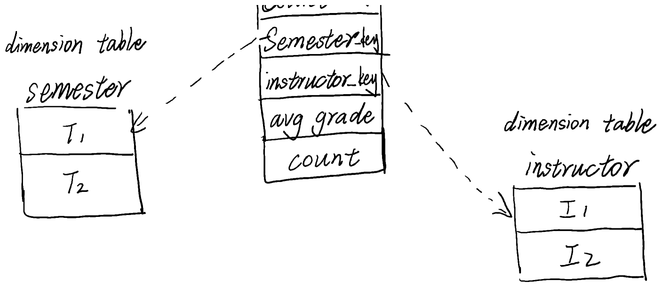
$$\frac{1}{N}\sum\sigma^2 = \frac{1}{2}\times(523.58+30.25) = 276.915$$

Comment--
No matter by Equal-frequency partitioning or Equal-width partitioning, mean of variances become larger in D1, which is a DS with relatively small variance originally. However, in a DB with large variance like D2, equal-width partitioning can obviously decrease the average variance.

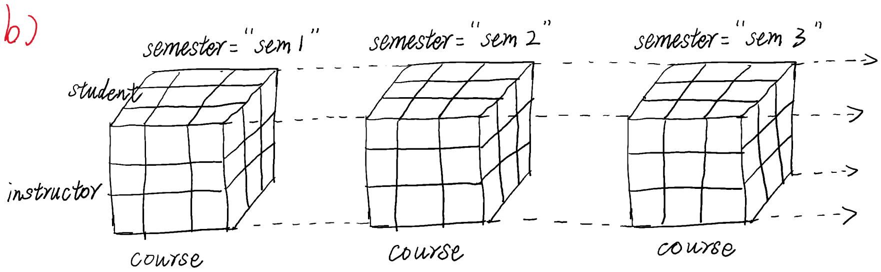
2.

(a)





(b)



(c)

OLAP Operations to list the average grade of Computer Science courses for each student:
1.Roll-up on course (from to department)
1.Roll-up on student (from to university)
2.Dice for course, student with department="CS" and university="Big University"
4.Drill down on student from university to student

Data cube can be view as a lattice of cuboids, if there is not other level in a dimension, then the cuboid contains the attribute of dimension itself and "none", so the total number of cuboids would be 2^n.(n - the number of dimensions) However, if there are some other levels, for example in student, the cuboid contains student, major, status, university, totally 4. Then the total number of cuboids should be computed by: (L - the number of levels)

$$\text{Total cubiods} = \prod_{i=1}^{n} (L_i + 1) = 5^4$$

So the cuboids will be contained is $5^4 = 625$.

3. (a)

Closed pattern: A pattern (itemset) X is closed if X is frequent, and there exists no super-pattern Y ⊃ X, with the same support as X.

closed patterns: Four

P_1: "$\{a_1, a_2, \cdots a_{11}\}$", P_2: "$\{a_1, a_2, \ldots a_{12}\}$"

P_3: "$\{a_1, a_2, \cdots a_{20}\}$", P_4: "$\{a_1, a_2, \cdots a_{30}\}$"

Maximal pattern: A pattern X is a maximal pattern if X is frequent and there exists no frequent super-pattern Y ⊃ X.

Maximal patterns: One

P_1: "$\{a_1, a_2, \ldots, a_{30}\}$"

(b) The closed patterns must contain in (a), except for those itemsets which don't satisfy the minimum support.

Closed patterns: Three

P_1: "$\{a_1, a_2, \cdots a_{11}\}$", P_2: "$\{a_1, a_2, \cdots a_{12}\}$,

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

The maximal patterns could be more as it allows the support of Y to be 1, but we need to notice that they must meet the minimum support.

Maximal Patterns: One

P_1: "$\{a_1, a_2, \ldots a_{20}\}$"

(c) Similarly

Closed Patterns: One

P_1: "$\{a_1, a_2, \ldots a_{11}\}$"

Maximal Patterns: One

P_1: "$\{a_1, a_2, \ldots a_{11}\}$"

4.

(a)

Support of $A \cup B$

$s\{A,B\}$ = occurences of an itemset $\{A,B\}$

$= 4$

$c = sup\{A,B\} / sup\{A\}$

$= 4/8 = 0.5$

(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,C
T_7	A,C
T_8	A,B,C,D,E
T_9	B,C
T_{10}	A,D
T_{11}	A,B,C

1st Scan →

Item set	support
A	8
B	7
C	7
D	5
E	2

1st filter →

Item set	support
A	8
B	7
C	7
D	5

→

Itemset
AB
AC
AD
BC
BD
CD

2nd Scan →

Itemset	Support
AB	4
AC	5
AD	4
BC	5
BD	3
CD	1

2nd fliter

Itemset	Support
AB	4
AC	5
AD	4
BC	5
BD	3

←

Itemset
ABC
ABD
ACD
BCD

3rd Scan

Itemset	Support
ABC	3
ABD	2
ACD	1
BCD	1

3rd fliter

Itemset	support
ABC	3

(c)

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,C
T_7	A,C
T_8	A,B,C,D,E
T_9	B,C
T_{10}	A,D
T_{11}	A,B,C

1. Scan DB once, find single item frequent pattern

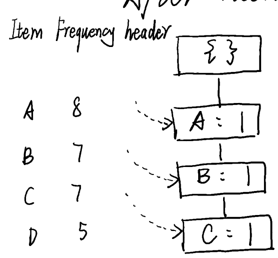
A:8 B:7 C:7 D:5

2. F-list = A-B-C-D

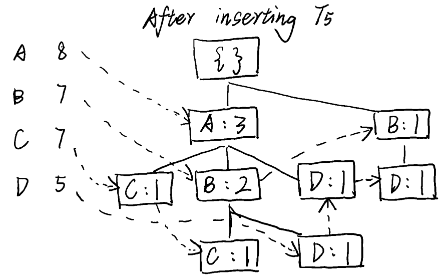
3.

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

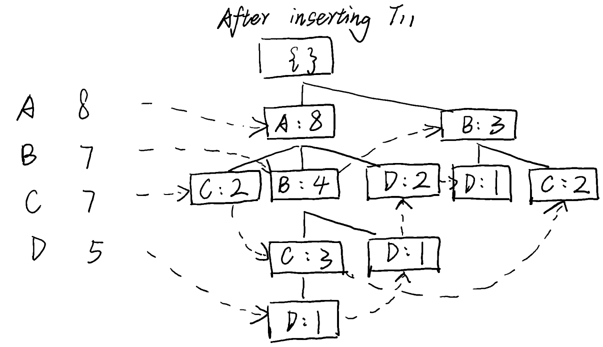
After inserting T_1



After inserting T_5



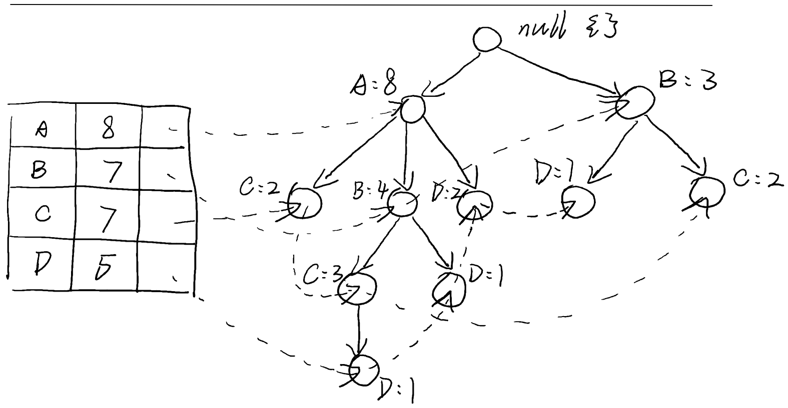
After inserting T_{11}



(d)

Item	Conditional Pattern Base	Conditional FP-tree	Frequent Patterns Generated
D	{{A,B,C:1}, {A,B:1}, {A:2}, {B:1}}	<A:4, B:2>, <B:1>	{A, D:4}, {B, D:3},
C	{{A:2}, {A,B:3}, {B:2}}	<A:5, B:3>,	{A,B,C:3}, {A,C:5}, {B,C:3}
B	{{A:4}}	<A:4>	{A,B:4}

	{B:2}}		{A,C:5}, {B,C:3}}
B	{{A:4}}	<A:4>	{A,B:4}



5.

(a)

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

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

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

thus $kulc(A,B) \leq 1$, we can obtain that $Kulc(A,B)$ is null invariant.

(b)

$$Lift(A,B) = \frac{s(A\cup B)}{s(A)\times s(B)}$$

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

Suppose A, B are independent, $Lift(A,B) = 1$

i.e., $\frac{a(a+b+c+d)}{(a+c)(a+b)} = 1$

$$a^2+ab+ac+ad = a^2+ab+ac+bc$$

$$ad = bc$$

(c) The only difference is that Cosine has a square root in the denominator. When we take the square root of the denominator, we can cancel out "a+b+c+d" from the fuction entirely. then it becomes $\frac{a}{(a+c)\times(a+b)}$ in the case above, which making the measure null-invariant.