ALDA Fall 2025
HW5
Due: 11/13/2025

TEAM ID : H24 MEMBER : Yujin kim, ykim68

HW5 contains 6 questions. Please read and follow the instructions.

- **DUE DATE FOR SUBMISSION: 11/13/2025 11:45 PM**
- **TOTAL NUMBER OF POINTS: 100**
- **NO PARTIAL CREDIT** will be given so provide concise answers.
- **Submissions and updates should be handled by the same person.**
- **You MUST manually add ALL team members in the submission portal when you submit through Gradescope.**
- Make sure you clearly list **your homework team ID, all team members' names, and Unity IDs, for those who have contributed to the homework contribution** at the top of your submission.
- **[GradeScope and Github]:** Submit a PDF on GradeScope. **No code is required for this homework.**
- The materials on this course website are only for use of students enrolled in this course and **MUST NOT** be retained or disseminated to others.
- By uploading your submission, you agree that you have not violated any university policies related to the student code of conduct (`https://policies.ncsu.edu/policy/pol-11-35-01/`), and you are signing the Pack Pledge: **"I have neither given nor received unauthorized aid on this test or assignment"**.

1. (10 points) [**K-means Clustering**] [**Graded By: Ian Holmes**] Using K-means clustering and Euclidean distance, cluster the 11 data points in Figure 1 into *three* clusters. We assume that the initial seeds are at the indicated points. Answer the following questions:

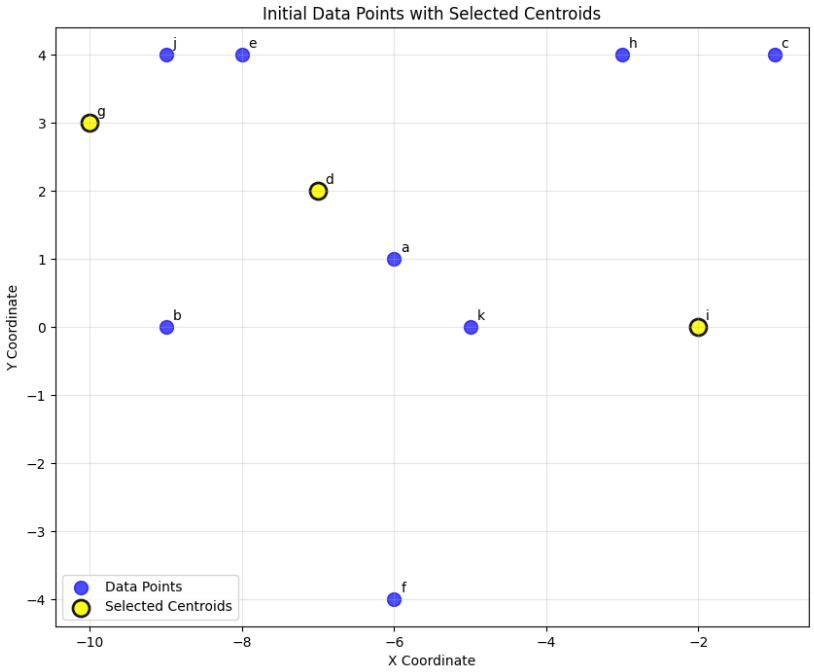


Figure 1: K-means Clustering (a)

Point	X Coordinate	Y Coordinate
a	-6	1
b	-9	0
c	-1	4
d	-7	2
e	-8	4
f	-6	-4
g	-10	3
h	-3	4
i	-2	0
j	-9	4
k	-5	0

Table 1: Coordinates of labeled data points.

(a) $\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - u_i\|^2$

Point	X Coordinate	Y Coordinate
a	-6	1
b	-9	0
c	-1	4
d	-7	2
e	-8	4
f	-6	-4
g	-10	3
h	-3	4
i	-2	0
j	-9	4
k	-5	0

Table 1: Coordinates of labeled data points.

① Euclidian Distance

$\|x_j - u_i\| = \sqrt{(x_j - u_{ix})^2 + (y_j - u_{iy})^2}$

- initial centroids : d(-7, 2), g(-10, 3), i(-2, 0)

a(-6, 1) $d(a,d) = \sqrt{(-6-(-7))^2 + (1-2)^2} = \sqrt{1+1} = \sqrt{2}$ ✓

$d(a,g) = \sqrt{(-6-(-10))^2 + (1-3)^2} = \sqrt{4^2+2^2} = \sqrt{20}$

$d(a,i) = \sqrt{(-6-(-2))^2 + (1-0)^2} = \sqrt{4^2+1^2} = \sqrt{17}$

---

b(-9,0) $d(b,d) = \sqrt{(-9-(-7))^2 + (0-2)^2} = \sqrt{4+4} = \sqrt{8}$ ✓

$d(b,g) = \sqrt{(-9-(-10))^2 + (0-3)^2} = \sqrt{1+9} = \sqrt{10}$

$d(b,i) = \sqrt{(-9-(-2))^2 + (0-0)^2} = \sqrt{49}$

---

c(-1,4) $d(c,d) = \sqrt{(-1-(-7))^2 + (4-2)^2} = \sqrt{36+4} = \sqrt{40}$

$d(c,g) = \sqrt{(-1-(-10))^2 + (4-3)^2} = \sqrt{81+1} = \sqrt{82}$

$d(c,i) = \sqrt{(-1-(-2))^2 + (4-0)^2} = \sqrt{1+16} = \sqrt{17}$ ✓

---

e(-8.4) $d(e,d) = \sqrt{(-8-(-7))^2 + (4-2)^2} = \sqrt{1+4} = \sqrt{5}$

$d(e,g) = \sqrt{(-8-(-10))^2 + (4-3)^2} = \sqrt{4+1} = \sqrt{5}$ ✓ left side

$d(e,i) = \sqrt{(-8-(-2))^2 + (4-0)^2} = \sqrt{36+16} = \sqrt{52}$

$f(-6,-4)$ $d(f,d) = \sqrt{(-6-(-7))^2+(-4-2)^2} = \sqrt{1+36} = \sqrt{37}$

$d(f,g) = \sqrt{(-6-(-10))^2+(-4-3)^2} = \sqrt{16+49} = \sqrt{65}$

$d(f,i) = \sqrt{(-6-(-2))^2+(-4-0)^2} = \sqrt{16+16} = \sqrt{32}$ ✓

$h(-3,4)$ $d(h,d) = \sqrt{(-3-(-7))^2+(4-2)^2} = \sqrt{16+4} = \sqrt{20}$

$d(h,g) = \sqrt{(-3-(-10))^2+(4-3)^2} = \sqrt{49+1} = \sqrt{50}$

$d(h,i) = \sqrt{(-3-(-2))^2+(4-0)^2} = \sqrt{1+16} = \sqrt{17}$ ✓

$j(-9,4)$ $d(j,d) = \sqrt{(-9-(-7))^2+(4-2)^2} = \sqrt{4+4} = \sqrt{8}$

$d(j,g) = \sqrt{(-9-(-10))^2+(4-3)^2} = \sqrt{1+1} = \sqrt{2}$ ✓

$d(j,i) = \sqrt{(-9-(-2))^2+(4-0)^2} = \sqrt{49+16} = \sqrt{65}$

$k(-5,0)$ $d(k,d) = \sqrt{(-5-(-7))^2+(0-2)^2} = \sqrt{4+4} = \sqrt{8}$ ✓

$d(k,g) = \sqrt{(-5-(-10))^2+(0-3)^2} = \sqrt{25+9} = \sqrt{34}$

$d(k,i) = \sqrt{(-5-(-2))^2+(0-0)^2} = \sqrt{9}$

$C_d : \{a,b,d,k\}$ $C_g : \{e,g,j\}$ $C_i : \{c,f,h,i\}$

$\mu_d^{(1)} = \left(\frac{-6-9-7-5}{4}, \frac{1+0+2+0}{4}\right) = \left(\frac{-27}{4}, \frac{3}{4}\right) = (-6.75, 0.75)$

$\mu_g^{(1)} = \left(\frac{-8-10-9}{3}, \frac{4+3+4}{3}\right) = \left(\frac{-27}{3}, \frac{11}{3}\right) = (-9, 3.67)$

$\mu_i^{(1)} = \left(\frac{-1-6-3-2}{4}, \frac{4-4+4+0}{4} = \left(\frac{-12}{4}, \frac{4}{4}\right) = (-3, 1)\right.$

(b) Round 2 centroid

$\mu_d^{(1)} : (-6.75, 0.75)$, $\mu_g^{(1)} : (-9, 3.67)$, $\mu_i^{(1)} : (-3, 1)$

a(-6, 1) $d(a, \mu_d^{(1)}) = \sqrt{(-6-(-6.75))^2 + (1-0.75)^2} = \sqrt{0.625}$ ✓d

$d(a, \mu_g^{(1)}) = \sqrt{(-6-(-9))^2 + (1-3.67)^2} = \sqrt{16.1289}$

$d(a, \mu_i^{(1)}) = \sqrt{(-6-(-3))^2 + (1-1)^2} = \sqrt{9}$

---

b(-9, 0) $d(b, \mu_d^{(1)}) = \sqrt{(-9-(-6.75))^2 + (0-0.75)^2} = \sqrt{5.625}$ ✓d

$d(b, \mu_g^{(1)}) = \sqrt{(-9-(-9))^2 + (0-3.67)^2} = \sqrt{13.4689}$

$d(b, \mu_i^{(1)}) = \sqrt{(-9-(-3))^2 + (0-1)^2} = \sqrt{37}$

---

c(-1, 4) $d(c, \mu_d^{(1)}) = \sqrt{(-1-(-6.75))^2 + (4-0.75)^2} = \sqrt{43.625}$

$d(c, \mu_g^{(1)}) = \sqrt{(-1-(-9))^2 + (4-3.67)^2} = \sqrt{64.1089}$

$d(c, \mu_i^{(1)}) = \sqrt{(-1-(-3))^2 + (4-1)^2} = \sqrt{13}$ ✓i

---

d(-7, 2) $d(d, \mu_d^{(1)}) = \sqrt{(-7-(-6.75))^2 + (2-0.75)^2} = \sqrt{1.625}$ ✓d

$d(d, \mu_g^{(1)}) = \sqrt{(-7-(-9))^2 + (2-3.67)^2} = \sqrt{6.7889}$

$d(d, \mu_i^{(1)}) = \sqrt{(-7-(-3))^2 + (2-1)^2} = \sqrt{17}$

---

e(-8, 4) $d(e, \mu_d^{(1)}) = \sqrt{(-8-(-6.75))^2 + (4-0.75)^2} = \sqrt{12.125}$

$d(e, \mu_g^{(1)}) = \sqrt{(-8-(-9))^2 + (4-3.67)^2} = \sqrt{1.1089}$ ✓g

$d(e, \mu_i^{(1)}) = \sqrt{(-8-(-3))^2 + (4-1)^2} = \sqrt{34}$

f(-6,-4) $d(f,\mu_d^{(1)}) = \sqrt{(-6-(-6.75))^2+(-4-0.75)^2} = \sqrt{23.125}$ ✓d

$d(f,\mu_g^{(1)}) = \sqrt{(-6-(-9))^2+(-4-3.67)^2} = \sqrt{67.8289}$

$d(f,\mu_i^{(1)}) = \sqrt{(-6-(-3))^2+(-4-1)^2} = \sqrt{34}$

---

g(-10,3) $d(g,\mu_d^{(1)}) = \sqrt{(-10-(-6.75))^2+(3-0.75)^2} = \sqrt{15.625}$

$d(g,\mu_g^{(1)}) = \sqrt{(-10-(-9))^2+(3-3.67)^2} = \sqrt{1.4489}$ ✓g

$d(g,\mu_i^{(1)}) = \sqrt{(-10-(-3))^2+(3-1)^2} = \sqrt{53}$

---

h(-3,4) $d(h,\mu_d^{(1)}) = \sqrt{(-3-(-6.75))^2+(4-0.75)^2} = \sqrt{24.625}$

$d(h,\mu_g^{(1)}) = \sqrt{(-3-(-9))^2+(4-3.67)^2} = \sqrt{36.1089}$

$d(h,\mu_i^{(1)}) = \sqrt{(-3-(-3))^2+(4-1)^2} = \sqrt{9}$ ✓i

---

i(-2,0) $d(i,\mu_d^{(1)}) = \sqrt{(-2-(-6.75))^2+(0-0.75)^2} = \sqrt{23.125}$

$d(i,\mu_g^{(1)}) = \sqrt{(-2-(-9))^2+(0-3.67)^2} = \sqrt{62.4689}$

$d(i,\mu_i^{(1)}) = \sqrt{(-2-(-3))^2+(0-1)^2} = \sqrt{2}$ ✓i

---

j(-9,4) $d(j,\mu_d^{(1)}) = \sqrt{(-9-(-6.75))^2+(4-0.75)^2} = \sqrt{15.625}$

$d(j,\mu_g^{(1)}) = \sqrt{(-9-(-9))^2+(4-3.67)^2} = \sqrt{0.1089}$ ✓g

$d(j,\mu_i^{(1)}) = \sqrt{(-9-(-3))^2+(4-1)^2} = \sqrt{45}$

---

k(-5,0) $d(k,\mu_d^{(1)}) = \sqrt{(-5-(-6.75))^2+(0-0.75)^2} = \sqrt{3.625}$ ✓d

$d(k,\mu_g^{(1)}) = \sqrt{(-5-(-9))^2+(0-3.67)^2} = \sqrt{29.4689}$

$d(k,\mu_i^{(1)}) = \sqrt{(-5-(-3))^2+(0-1)^2} = \sqrt{5}$

C_d: {a, b, d, f, k} C_g: {e, g, j} C_i: {c, h, i}

$\mu_d^{(2)} = \left(\frac{-6-9-7-6-5}{5}, \frac{1+0+2-4+0}{5}\right) = \left(\frac{-33}{5}, \frac{-1}{5}\right) = (-6.6, -0.2)$

$\mu_g^{(2)} = \left(\frac{-8-10-9}{3}, \frac{4+3+4}{3}\right) = \left(-\frac{27}{3}, \frac{11}{3}\right) = (-9, 3.67)$ converged

$\mu_i^{(2)} = \left(\frac{-1-3-2}{3}, \frac{4+4+0}{3}\right) = \left(\frac{-6}{3}, \frac{8}{3}\right) = (-2, 2.67)$

Round 3

0.36

a(-6, 1) $d(a, \mu_d^{(1)}) = \sqrt{(-6-(-6.6))^2 + (1-(-0.2))^2} = \sqrt{1.8}$ ✓ d

$d(a, \mu_g^{(1)}) = \sqrt{(-6-(-9))^2 + (1-3.67)^2} = \sqrt{16.1289}$

$d(a, \mu_i^{(1)}) = \sqrt{(-6-(-2))^2 + (1-2.67)^2} = \sqrt{18.7889}$

---

b(-9, 0) $d(b, \mu_d^{(1)}) = \sqrt{(-9-(-6.6))^2 + (0-(-0.2))^2} = \sqrt{5.8}$ ✓ d

$d(b, \mu_g^{(1)}) = \sqrt{(-9-(-9))^2 + (0-3.67)^2} = \sqrt{13.4689}$

$d(b, \mu_i^{(1)}) = \sqrt{(-9-(-2))^2 + (0-2.67)^2} = \sqrt{56.1289}$

---

c(-1, 4) $d(c, \mu_d^{(1)}) = \sqrt{(-1-(-6.6))^2 + (4-(-0.2))^2} = \sqrt{49}$

$d(c, \mu_g^{(1)}) = \sqrt{(-1-(-9))^2 + (4-3.67)^2} = \sqrt{64.1089}$

$d(c, \mu_i^{(1)}) = \sqrt{(-1-(-2))^2 + (4-2.67)^2} = \sqrt{2.7689}$ ✓ i

$d(-7,2)$ $d(d,\mu_d^{(1)}) = \sqrt{(-7-(-6.6))^2+(2-(-0.2))^2} = \sqrt{5}$ ✓ d

$d(d,\mu_g^{(1)}) = \sqrt{(-7-(-9))^2+(2-3.67)^2} = \sqrt{6.7889}$

$d(d,\mu_i^{(1)}) = \sqrt{(-7-(-2))^2+(2-2.67)^2} = \sqrt{25.4489}$

$e(-8.4)$ $d(e,\mu_d^{(1)}) = \sqrt{(-8-(-6.6))^2+(4-(-0.2))^2} = \sqrt{19.6}$

$d(e,\mu_g^{(1)}) = \sqrt{(-8-(-9))^2+(4-3.67)^2} = \sqrt{1.1089}$ ✓ g

$d(e,\mu_i^{(1)}) = \sqrt{(-8-(-2))^2+(4-2.67)^2} = \sqrt{37.7689}$

$f(-6,-4)$ $d(f,\mu_d^{(1)}) = \sqrt{(-6-(-6.6))^2+(-4-(-0.2))^2} = \sqrt{14.8}$ ✓ d

$d(f,\mu_g^{(1)}) = \sqrt{(-6-(-9))^2+(-4-3.67)^2} = \sqrt{57.8289}$

$d(f,\mu_i^{(1)}) = \sqrt{(-6-(-2))^2+(-4-2.67)^2} = \sqrt{17.7689}$

$g(-10,3)$ $d(g,\mu_d^{(1)}) = \sqrt{(-10-(-6.6))^2+(3-(-0.2))^2} = \sqrt{21.8}$

$d(g,\mu_g^{(1)}) = \sqrt{(-10-(-9))^2+(3-3.67)^2} = \sqrt{1.4489}$ ✓ g

$d(g,\mu_i^{(1)}) = \sqrt{(-10-(-2))^2+(3-2.67)^2} = \sqrt{64.1089}$

$h(-3,4)$ $d(h,\mu_d^{(1)}) = \sqrt{(-3-(-6.6))^2+(4-(-0.2))^2} = \sqrt{30.6}$

$d(h,\mu_g^{(1)}) = \sqrt{(-3-(-9))^2+(4-3.67)^2} = \sqrt{36.1089}$

$d(h,\mu_i^{(1)}) = \sqrt{(-3-(-2))^2+(4-2.67)^2} = \sqrt{2.7689}$ ✓ i

---

$i(-2,0)$ $d(i,\mu_d^{(1)}) = \sqrt{(-2-(-6.6))^2+(0-(-0.2))^2} = \sqrt{21.2}$

$d(i,\mu_g^{(1)}) = \sqrt{(-2-(-9))^2+(0-3.67)^2} = \sqrt{62.4689}$

$d(i,\mu_i^{(1)}) = \sqrt{(-2-(-2))^2+(0-2.67)^2} = \sqrt{7.1289}$ ✓ i

$j(-9,4)$ $d(j,\mu_d^{(1)}) = \sqrt{(-9-(-6.6))^2+(4-(-0.2))^2} = \sqrt{23.4}$

$d(j,\mu_g^{(1)}) = \sqrt{(-9-(-9))^2+(4-3.67)^2} = \sqrt{0.1089}$ ✓g

$d(j,\mu_i^{(1)}) = \sqrt{(-9-(-2))^2+(4-2.67)^2} = \sqrt{50.7689}$

$k(-5,0)$ $d(k,\mu_d^{(1)}) = \sqrt{(-5-(-6.6))^2+(0-(-0.2))^2} = \sqrt{2.6}$ ✓d

$d(k,\mu_g^{(1)}) = \sqrt{(-5-(-9))^2+(0-3.67)^2} = \sqrt{29.4689}$

$d(k,\mu_i^{(1)}) = \sqrt{(-5-(-2))^2+(0-2.67)^2} = \sqrt{16.1289}$

$C_d = \{a,b,d,f,k\}$ $C_g = \{e,g,j\}$ $C_i = \{c,h,i\}$

converged in round 2

converged in round3

centroid (final)

$\mu_d^{(3)} : (-6.6,-0.2)$ $\mu_g^{(3)} : (-9, 3.67)$ $\mu_i^{(3)} : (-2, 2.67)$

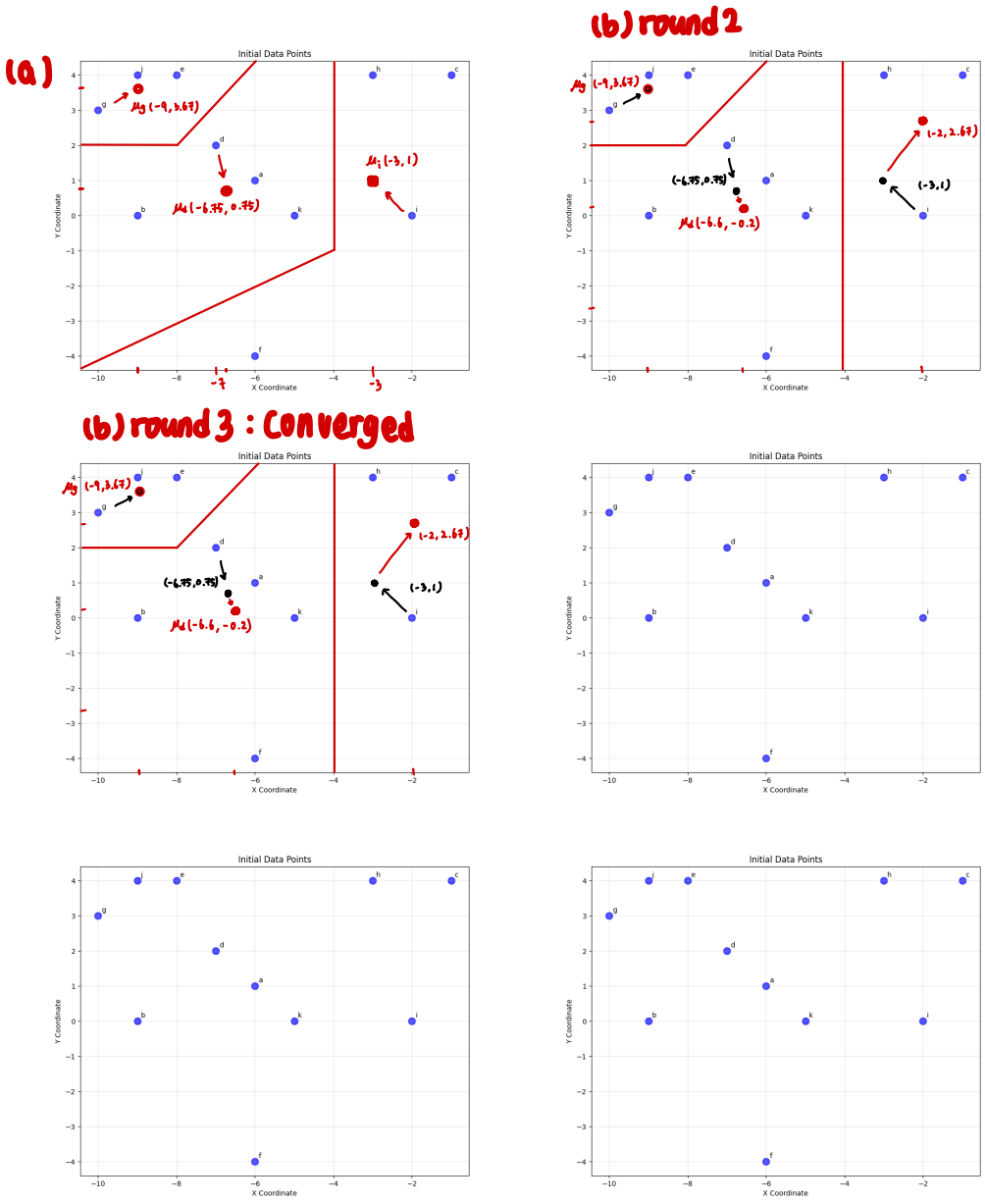


Figure 2: K-means Clustering (b)

(a) (4 points) Run the K-means algorithm for the first iteration 1. Calculate the coordinates of the new centroids. What are the new clusters? Show your work in the first subgraph in Figure 2.

(b) (6 points) On which iteration the K-means clustering algorithm would converge? Draw the resulting clusters and new centroid at the end of each round (including the first round) in the Figure 2. Indicate the coordinates along side corresponding centroids. **Add new graphics if needed; Stop when the algorithm converges and clearly label on the graph where the algorithm converges.** NO PARTIAL CREDIT.

2. (15 points) [**Hierarchical Clustering**] [**Graded by Ian Holmes**] We will use the same dataset as in Question 1 for the following problem. The *Euclidean Distance* matrix between each pair of the data points is listed in the figure below:

Table 2: Euclidean distance matrix for points a–k.

	a	b	c	d	e	f	g	h	i	j	k
a	0.00	3.16	5.83	1.41	3.61	5.00	4.47	4.24	4.12	4.24	1.41
b	3.16	0.00	8.94	2.83	4.12	5.00	3.16	7.21	7.00	4.00	4.00
c	5.83	8.94	0.00	6.32	7.00	9.43	9.06	2.00	4.12	8.00	5.66
d	1.41	2.83	6.32	0.00	2.24	6.08	3.16	4.47	5.39	2.83	2.83
e	3.61	4.12	7.00	2.24	0.00	8.25	2.24	5.00	7.21	1.00	5.00
f	5.00	5.00	9.43	6.08	8.25	0.00	8.06	8.54	5.66	8.54	4.12
g	4.47	3.16	9.06	3.16	2.24	8.06	0.00	7.07	8.54	1.41	5.83
h	4.24	7.21	2.00	4.47	5.00	8.54	7.07	0.00	4.12	6.00	4.47
i	4.12	7.00	4.12	5.39	7.21	5.66	8.54	4.12	0.00	8.06	3.00
j	4.24	4.00	8.00	2.83	1.00	8.54	1.41	6.00	8.06	0.00	5.66
k	1.41	4.00	5.66	2.83	5.00	4.12	5.83	4.47	3.00	5.66	0.00

(a) (8 points) Perform *single* and *complete* link hierarchical clustering. Show your results by drawing corresponding dendrogram. The dendrogram should clearly show the order and the height in which the clusters are merged. **In case of a tie please resolve in alphabetical order of the points' labels**. NO PARTIAL CREDIT.

(b) (4 points) Using Sum of Squared Error (SSE) and assuming there are three clusters, which of the *single link* and *complete link* hierarchical clustering will yield better results? Justify your answer.

(c) (3 points) Compare the clusters from 2(b) with the clusters found using K-means in Question 1 by calculating their corresponding Sum of Squared Errors (SSE)s. According to their SSE results, which is better: K-means or hierarchical clustering?

(a) Single-link

	a	b	c	d	e	f	g	h	i	j	k
a	0.00	3.16	5.83	1.41	3.61	5.00	4.47	4.24	4.12	4.24	1.41
b	3.16	0.00	8.94	2.83	4.12	5.00	3.16	7.21	7.00	4.00	4.00
c	5.83	8.94	0.00	6.32	7.00	9.43	9.06	2.00	4.12	8.00	5.66
d	1.41	2.83	6.32	0.00	2.24	6.08	3.16	4.47	5.39	2.83	2.83
e	3.61	4.12	7.00	2.24	0.00	8.25	2.24	5.00	7.21	1.00	5.00
f	5.00	5.00	9.43	6.08	8.25	0.00	8.06	8.54	5.66	8.54	4.12
g	4.47	3.16	9.06	3.16	2.24	8.06	0.00	7.07	8.54	1.41	5.83
h	4.24	7.21	2.00	4.47	5.00	8.54	7.07	0.00	4.12	6.00	4.47
i	4.12	7.00	4.12	5.39	7.21	5.66	8.54	4.12	0.00	8.06	3.00
j	4.24	4.00	8.00	2.83	1.00	8.54	1.41	6.00	8.06	0.00	5.66
k	1.41	4.00	5.66	2.83	5.00	4.12	5.83	4.47	3.00	5.66	0.00

a-d or k 1.41
b-d 2.83
c-h 2.00 ⑤
d-a 1.41 ②
e-j 1.00 ①
f-k 4.12
g-j 1.41
h-c 2.00
i-k 3.00
j-e 1.00 ④
k-a 1.41

closet point
:e-j

2nd point
:a-d

distance of e,j min
a 3.61 4.24 → 3.61
b 4.12 4.00 → 4.00
c 7.00 8.00 → 7.00
d 2.24 2.83 → 2.24
f 8.25 8.54 → 8.25
g 2.24 1.41 → 1.41 ③
h 5.00 6.00 → 5.00
i 7.21 8.06 → 7.21
k 5.00 5.66 → 5.00
* e,j → g ③

distance of e,j,g
a 3.61 4.47 → 4.47
b 4.00 3.16 → 3.16
c 7.00 9.06 → 7.00
d 2.24 3.16 → 2.24
f 8.25 8.06 → 8.06
h 5.00 7.07 → 5.00
i 7.21 8.54 → 7.21
k 5.00 5.83 → 5.00

distance of a,d
b 3.16 2.83 → 2.83
c 5.83 6.32 → 5.83
e 3.61 2.24 → 2.24
f 5.00 6.08 → 5.00
g 4.47 3.16 → 3.16
h 4.24 4.47 → 4.24
i 4.12 5.39 → 4.12
j 4.24 2.83 → 2.83
k 1.41 2.83 → 1.41
* a,d → k ④

→ distance of a,d,k
b 2.83 4.00 → 2.83
c 5.83 5.66 → 5.66
e 2.24 5.00 → 2.24
f 5.00 4.12 → 4.12
g 3.16 5.83 → 3.16
h 4.24 4.47 → 4.24
j 2.83 5.66 → 2.83
i 4.12 3.00 → 3.00
* a,d,k → e(j,g) ⑥

{a,d,k} – {e,j,g} 2.24 ⑥

	a	b	c	d	e	f	g	h	i	j	k
a	0.00	3.16	5.83	1.41	3.61	5.00	4.47	4.24	4.12	4.24	1.41
b	3.16	0.00	8.94	2.83	4.12	5.00	3.16	7.21	7.00	4.00	4.00
c	5.83	8.94	0.00	6.32	7.00	9.43	9.06	2.00	4.12	8.00	5.66
d	1.41	2.83	6.32	0.00	2.24	6.08	3.16	4.47	5.39	2.83	2.83
e	3.61	4.12	7.00	2.24	0.00	8.25	2.24	5.00	7.21	1.00	5.00
f	5.00	5.00	9.43	6.08	8.25	0.00	8.06	8.54	5.66	8.54	4.12
g	4.47	3.16	9.06	3.16	2.24	8.06	0.00	7.07	8.54	1.41	5.83
h	4.24	7.21	2.00	4.47	5.00	8.54	7.07	0.00	4.12	6.00	4.47
i	4.12	7.00	4.12	5.39	7.21	5.66	8.54	4.12	0.00	8.06	3.00
j	4.24	4.00	8.00	2.83	1.00	8.54	1.41	6.00	8.06	0.00	5.66
k	1.41	4.00	5.66	2.83	5.00	4.12	5.83	4.47	3.00	5.66	0.00

distance of a,d,k,e,j,g
b 2.83 3.16 → 2.83
c 5.66 7.00 → 5.66
f 4.12 8.06 → 4.12
h 4.24 5.00 → 4.24
i 3.00 7.21 → 3.00
* adkejg → b ⑦

distance of c,h
a 5.83 4.24 → 4.24
b 8.94 7.21 → 7.21
d 6.32 4.47 → 4.47
e 7.00 5.00 → 5.00
f 9.43 8.54 → 8.54
g 9.06 7.07 → 7.07
i 4.12 4.12 → 4.12
j 8.00 6.00 → 6.00
k 5.66 4.47 → 4.47

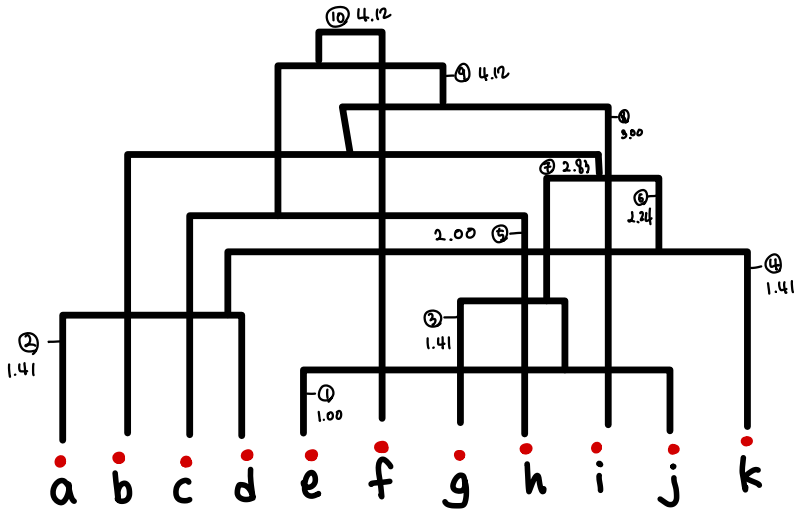
⑦

distance of a,d,k,e,j,g,b
c 5.66 8.94 → 5.66
f 4.12 5.00 → 4.12
h 4.24 7.21 → 4.24
i 3.00 7.00 → 3.00
* a,d,k,e,j,g,b → i ⑧

∴ a,d,k,e,j,g,b,i,c,h
→ f 4.12

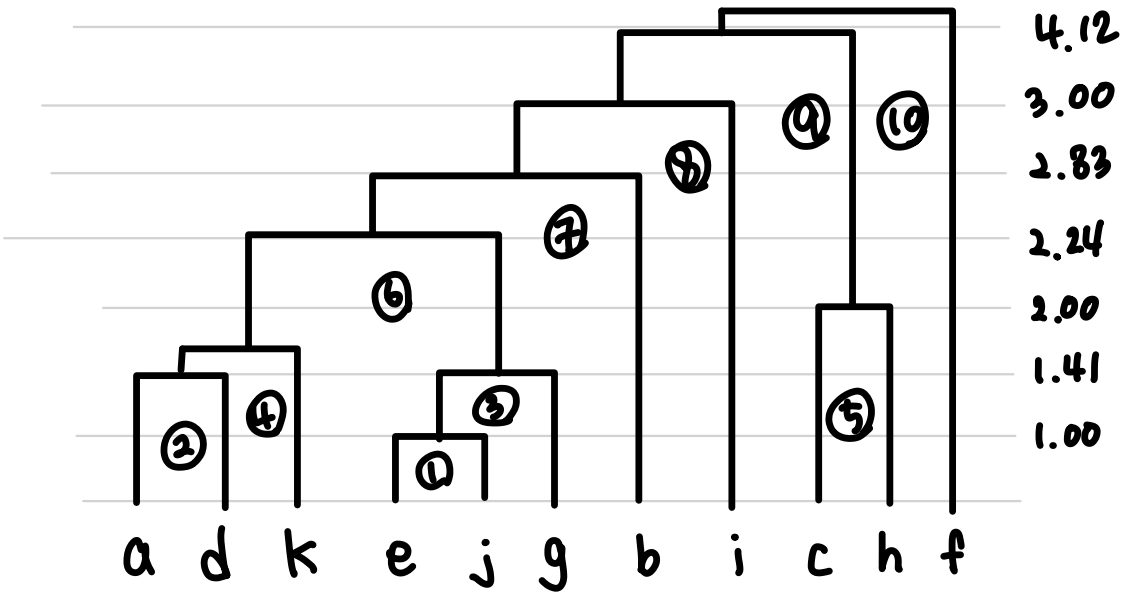
distance of a,d,k,e,j,g,b,i
c 5.66 4.12 → 4.12
f 4.12 5.66 → 4.12
h 4.24 4.12 → 4.12
* a,d,k,e,j,g,b,i → {c,h} 4.12 ⑨

	a	b	c	d	e	f	g	h	i	j	k
a	0.00	3.16	5.83	1.41	3.61	5.00	4.47	4.24	4.12	4.24	1.41
b	3.16	0.00	8.94	2.83	4.12	5.00	3.16	7.21	7.00	4.00	4.00
c	5.83	8.94	0.00	6.32	7.00	9.43	9.06	2.00	4.12	8.00	5.66
d	1.41	2.83	6.32	0.00	2.24	6.08	3.16	4.47	5.39	2.83	2.83
e	3.61	4.12	7.00	2.24	0.00	8.25	2.24	5.00	7.21	1.00	5.00
f	5.00	5.00	9.43	6.08	8.25	0.00	8.06	8.54	5.66	8.54	4.12
g	4.47	3.16	9.06	3.16	2.24	8.06	0.00	7.07	8.54	1.41	5.83
h	4.24	7.21	2.00	4.47	5.00	8.54	7.07	0.00	4.12	6.00	4.47
i	4.12	7.00	4.12	5.39	7.21	5.66	8.54	4.12	0.00	8.06	3.00
j	4.24	4.00	8.00	2.83	1.00	8.54	1.41	6.00	8.06	0.00	5.66
k	1.41	4.00	5.66	2.83	5.00	4.12	5.83	4.47	3.00	5.66	0.00



Single Link

① e-j 1.00
② a-d 1.41
③ {e,j}-g 1.41
④ {a,d}-k 1.41
⑤ c-h 2.00
⑥ {a,d,k}-{e,j,g} 2.24
⑦ {a,d,k,e,j,g}-b 2.83
⑧ {a,d,k,e,j,g,b}-i 3.00
⑨ {a,d,k,e,j,g,b,i}-{c,h} 4.12
⑩ {a,d,k,e,j,g,b,i,c,h}-f 4.12



complete - link

	a	b	c	d	e	f	g	h	i	j	k
a	0.00	3.16	5.83	1.41	3.61	5.00	4.47	4.24	4.12	4.24	1.41
b	3.16	0.00	8.94	2.83	4.12	5.00	3.16	7.21	7.00	4.00	4.00
c	5.83	8.94	0.00	6.32	7.00	9.43	9.06	2.00	4.12	8.00	5.66
d	1.41	2.83	6.32	0.00	2.24	6.08	3.16	4.47	5.39	2.83	2.83
e	3.61	4.12	7.00	2.24	0.00	8.25	2.24	5.00	7.21	1.00	5.00
f	5.00	5.00	9.43	6.08	8.25	0.00	8.06	8.54	5.66	8.54	4.12
g	4.47	3.16	9.06	3.16	2.24	8.06	0.00	7.07	8.54	1.41	5.83
h	4.24	7.21	2.00	4.47	5.00	8.54	7.07	0.00	4.12	6.00	4.47
i	4.12	7.00	4.12	5.39	7.21	5.66	8.54	4.12	0.00	8.06	3.00
j	4.24	4.00	8.00	2.83	1.00	8.54	1.41	6.00	8.06	0.00	5.66
k	1.41	4.00	5.66	2.83	5.00	4.12	5.83	4.47	3.00	5.66	0.00

a-d or k 1.41
b-d 2.83
c-h 2.00
d-a 1.41
e-j 1.00
f-k 4.12
g-j 1.41
h-c 2.00
i-k 3.00
j-e 1.00
k-a 1.41

closet point
① :e-j 1.00
second
② :a-d 1.41
3-rd
③ :c-h 2.00

distance of e,j
a 3.61 4.24 → 4.24
b 4.12 4.00 → 4.12
c 7.00 8.00 → 8.00
d 2.24 2.83 → 2.83
f 8.25 8.54 → 8.54
g 2.24 1.41 → 2.24 ④
h 5.00 6.00 → 6.00
i 7.21 8.06 → 8.06
k 5.00 5.66 → 5.66
* e,j → g

distance of a,d
b 3.16 2.83 → 3.16
c 5.83 6.32 → 6.32
e 3.61 2.24 → 3.61
f 5.00 6.08 → 6.08
g 4.47 3.16 → 4.47
h 4.24 4.47 → 4.47
i 4.12 5.39 → 5.39
j 4.24 2.83 → 4.24
k 1.41 2.83 → 2.83 ⑤
* a,d → k

distance of c,h
* c,h → i
a 5.83 4.24 → 5.83
b 8.94 7.21 → 8.94
d 6.32 4.47 → 6.32
e 7.00 5.00 → 7.00
f 9.43 8.54 → 9.43
g 9.06 7.07 → 9.06
i 4.12 4.12 → 4.12 ⑦
j 8.00 6.00 → 8.00
l 5.66 4.47 → 5.66

distance of a,d,k
b 3.16 4.00 → 4.00 ⑥
c 6.32 5.66 → 6.32
e 3.61 5.00 → 5.00
f 6.08 4.12 → 6.08
g 4.47 5.83 → 5.83
h 4.47 4.47 → 4.47
i 5.39 3.00 → 5.39
j 4.24 5.66 → 5.66
* a,d,k → b

	a	b	c	d	e	f	g	h	i	j	k
a	0.00	3.16	5.83	1.41	3.61	5.00	4.47	4.24	4.12	4.24	1.41
b	3.16	0.00	8.94	2.83	4.12	5.00	3.16	7.21	7.00	4.00	4.00
c	5.83	8.94	0.00	6.32	7.00	9.43	9.06	2.00	4.12	8.00	5.66
d	1.41	2.83	6.32	0.00	2.24	6.08	3.16	4.47	5.39	2.83	2.83
e	3.61	4.12	7.00	2.24	0.00	8.25	2.24	5.00	7.21	1.00	5.00
f	5.00	5.00	9.43	6.08	8.25	0.00	8.06	8.54	5.66	8.54	4.12
g	4.47	3.16	9.06	3.16	2.24	8.06	0.00	7.07	8.54	1.41	5.83
h	4.24	7.21	2.00	4.47	5.00	8.54	7.07	0.00	4.12	6.00	4.47
i	4.12	7.00	4.12	5.39	7.21	5.66	8.54	4.12	0.00	8.06	3.00
j	4.24	4.00	8.00	2.83	1.00	8.54	1.41	6.00	8.06	0.00	5.66
k	1.41	4.00	5.66	2.83	5.00	4.12	5.83	4.47	3.00	5.66	0.00

→ {a,d,k,b} - {e,j,g} : 5.83

	a	b	c	d	e	f	g	h	i	j	k
a	0.00	3.16	5.83	1.41	3.61	5.00	4.47	4.24	4.12	4.24	1.41
b	3.16	0.00	8.94	2.83	4.12	5.00	3.16	7.21	7.00	4.00	4.00
c	5.83	8.94	0.00	6.32	7.00	9.43	9.06	2.00	4.12	8.00	5.66
d	1.41	2.83	6.32	0.00	2.24	6.08	3.16	4.47	5.39	2.83	2.83
e	3.61	4.12	7.00	2.24	0.00	8.25	2.24	5.00	7.21	1.00	5.00
f	5.00	5.00	9.43	6.08	8.25	0.00	8.06	8.54	5.66	8.54	4.12
g	4.47	3.16	9.06	3.16	2.24	8.06	0.00	7.07	8.54	1.41	5.83
h	4.24	7.21	2.00	4.47	5.00	8.54	7.07	0.00	4.12	6.00	4.47
i	4.12	7.00	4.12	5.39	7.21	5.66	8.54	4.12	0.00	8.06	3.00
j	4.24	4.00	8.00	2.83	1.00	8.54	1.41	6.00	8.06	0.00	5.66
k	1.41	4.00	5.66	2.83	5.00	4.12	5.83	4.47	3.00	5.66	0.00

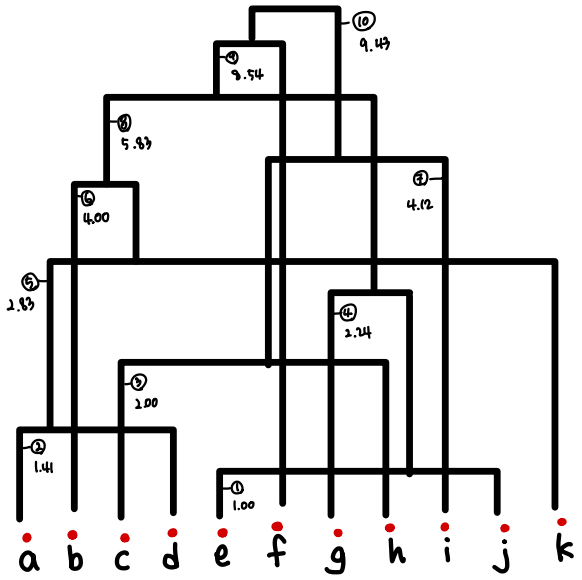
→ {a,d,k,b,e,j,g,f 4 - {c,h,i} : 9.43

distance of a,d,k,b
c 6.32 8.94 → 8.94
e 5.00 4.12 → 5.00
f 6.08 5.00 → 6.08
g 5.83 3.16 → 5.83
h 4.47 7.21 → 7.21
i 5.39 7.00 → 7.00
j 5.66 4.00 → 5.66

distance of e,j,g
a 4.24 4.47 → 4.47
b 4.12 3.16 → 4.12
c 8.00 9.06 → 9.06
d 2.83 3.16 → 3.16
f 8.54 8.06 → 8.54
h 6.00 7.07 → 7.07
i 8.06 8.54 → 8.54
k 5.66 5.83 → 5.83

* a,d,k,b,e,j,g → f : 8.54

distance of c,h,i
a 5.83 4.12 → 5.83
b 8.94 7.00 → 8.94
d 6.32 5.39 → 6.32
e 7.00 7.21 → 7.21
f 9.43 5.66 → 9.43
g 9.06 8.54 → 9.06
j 8.00 8.06 → 8.06
k 5.66 3.00 → 5.66



Complete link

① e - j 1.00

② a - d 1.41

③ c - h 2.00

④ {e, j} - g 2.24

⑤ {a, d} - k 2.83

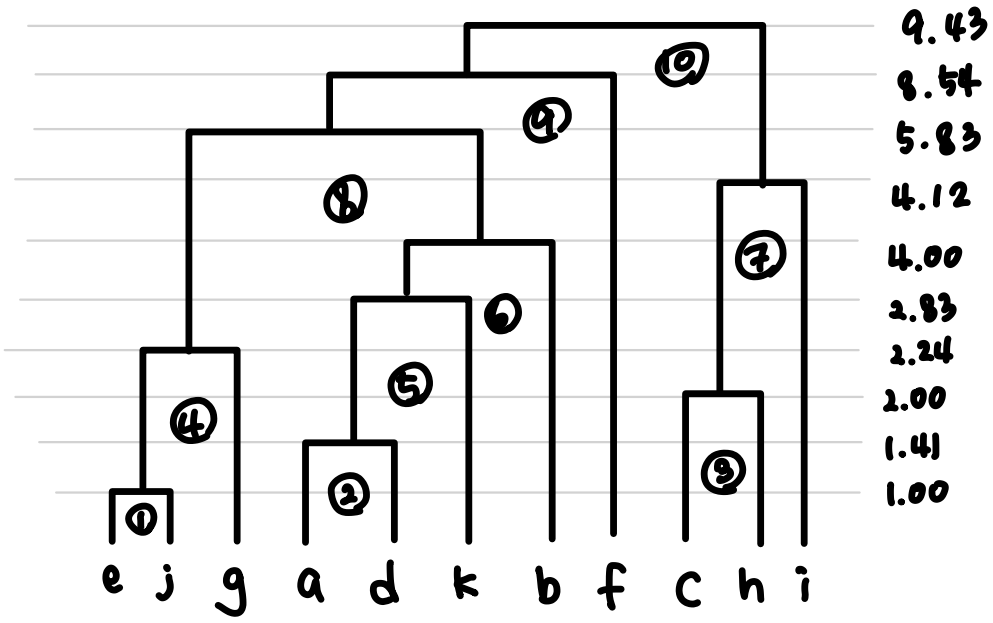
⑥ {a, d, k} - b 4.00

⑦ {c, h} - i 4.12

⑧ {a, d, k, b} - {e, j, g} 5.83

⑨ {a, d, k, b, e, j, g} - f 8.54

⑩ {a, d, k, b, e, j, g, f} - {c, h, i} 9.43



(b) Single link

$C_1 = \{a, d, k, e, j, g, b, i\}$

$C_2 = \{c, h\}$

$C_3 = \{f\}$

Complete link

$C_1 = \{a, d, k, b, e, j, g\}$

$C_2 = \{c, h, i\}$

$C_3 = \{f\}$

$SSE(C_1) = \frac{1}{8} \times 555.88 = 69.48$

	a	b	c	d	e	f	g	h	i	j	k
a	0.00	3.16	5.83	1.41	3.61	5.00	4.47	4.24	4.12	4.24	1.41
b	3.16	0.00	8.94	2.83	4.12	5.00	3.16	7.21	7.00	4.00	4.00
c	5.83	8.94	0.00	6.32	7.00	9.43	9.06	2.00	4.12	8.00	5.66
d	1.41	2.83	6.32	0.00	2.24	6.08	3.16	4.47	5.39	2.83	2.83
e	3.61	4.12	7.00	2.24	0.00	8.25	2.24	5.00	7.21	1.00	5.00
f	5.00	5.00	9.43	6.08	8.25	0.00	8.06	8.54	5.66	8.54	4.12
g	4.47	3.16	9.06	3.16	2.24	8.06	0.00	7.07	8.54	1.41	5.83
h	4.24	7.21	2.00	4.47	5.00	8.54	7.07	0.00	4.12	6.00	4.47
i	4.12	7.00	4.12	5.39	7.21	5.66	8.54	4.12	0.00	8.06	3.00
j	4.24	4.00	8.00	2.83	1.00	8.54	1.41	6.00	8.06	0.00	5.66
k	1.41	4.00	5.66	2.83	5.00	4.12	5.83	4.47	3.00	5.66	0.00

$\rightarrow 3.16^2 + 1.41^2 + 3.61^2 + 4.47^2 + 4.12^2 + 4.24^2 + 1.41^2 = 81.927$

$\rightarrow 2.83^2 + 4.12^2 + 3.16^2 + 7^2 + 4^2 + 4^2 = 115.969$

$\rightarrow 2.24^2 + 3.16^2 + 5.39^2 + 2.83^2 + 2.83^2 = 60.074$

$\rightarrow 2.24^2 + 7.21^2 + 1.00^2 + 5^2 = 83.002$

$\rightarrow 8.54^2 + 1.41^2 + 5.83^2 = 109.909$

$\rightarrow 8.06^2 + 3^2 = 73.964$

$\rightarrow 5.66^2 = 32.036$

$SSE(C_2) = \frac{1}{2} \times 4 = 2$

$SSE(C_3) = 0$

$SSE_{single} = 0 + 2 + 69.48 = 71.48$

	a	b	c	d	e	f	g	h	i	j	k
a	0.00	3.16	5.83	1.41	3.61	5.00	4.47	4.24	4.12	4.24	1.41
b	3.16	0.00	8.94	2.83	4.12	5.00	3.16	7.21	7.00	4.00	4.00
c	5.83	8.94	0.00	6.32	7.00	9.43	9.06	2.00	4.12	8.00	5.66
d	1.41	2.83	6.32	0.00	2.24	6.08	3.16	4.47	5.39	2.83	2.83
e	3.61	4.12	7.00	2.24	0.00	8.25	2.24	5.00	7.21	1.00	5.00
f	5.00	5.00	9.43	6.08	8.25	0.00	8.06	8.54	5.66	8.54	4.12
g	4.47	3.16	9.06	3.16	2.24	8.06	0.00	7.07	8.54	1.41	5.83
h	4.24	7.21	2.00	4.47	5.00	8.54	7.07	0.00	4.12	6.00	4.47
i	4.12	7.00	4.12	5.39	7.21	5.66	8.54	4.12	0.00	8.06	3.00
j	4.24	4.00	8.00	2.83	1.00	8.54	1.41	6.00	8.06	0.00	5.66
k	1.41	4.00	5.66	2.83	5.00	4.12	5.83	4.47	3.00	5.66	0.00

$\rightarrow 3.16^2 + 1.41^2 + 3.61^2 + 4.47^2 + 4.24^2 + 1.41^2 = 64.9524$

$\rightarrow 2.83^2 + 4.12^2 + 3.16^2 + 4^2 + 4^2 = 66.9689$

$\rightarrow 2.24^2 + 3.16^2 + 2.83^2 + 2.83^2 = 31.021$

$\rightarrow 2.24^2 + 1^2 + 5^2 = 31.0176$

$\rightarrow 1.41^2 + 5.83^2 = 35.977$

$\rightarrow 5.66^2 = 32.036$

$SSE(C_1) = \frac{1}{7} \times 261.9729 = 37.425$

$SSE(C_2) = \frac{1}{3} \times 37.9488 = 12.65$

$SSE(C_3) = 0$

$SSE_{complete} = 0 + 12.65 + 37.425 = 50.07$

$SSE_{single} = 71.48 > SSE_{complete} = 50.07$

Using the SSE criterion with 3-cluster, complete-link provides the better clustering result.

(c) - Calculate SSE_{kmeans} of Q1.

$C_1 = \{a, b, d, f, k\}$ $\mu_1 = (-6.6, -0.2)$

$C_2 = \{e, g, j\}$ $\mu_2 = (-9, 3.67)$

$C_3 = \{c, h, i\}$ $\mu_3 = (-2, 2.67)$

(1) $a(-6,1)$, $b(-9,0)$, $d(-7,2)$, $f(-6,-4)$, $k=(-5,0)$

- $(6-(-6.6))^2 + (1-(-0.2))^2 = 1.6$
- $(-9-(-6.6))^2 + (0-(-0.2))^2 = 5.8$
- $(-7-(-6.6))^2 + (2-(-0.2))^2 = 5$
- $(-6-(-6.6))^2 + (-4-(-0.2))^2 = 14.8$
- $(-5-(-6.6))^2 + (0-(-0.2))^2 = 2.6$

$SSE(C_1) = 1.8 + 5.8 + 5 + 14.8 + 2.6 = 30$

(2) $e(-8,4)$, $g(-10,3)$, $j=(-9,4)$

- $(-8-(-9))^2 + (4-3.67)^2 = 1.11$
- $(-10-(-9))^2 + (3-3.67)^2 = 1.44$
- $(-9-(-9))^2 + (4-3.67)^2 = 0.11$

$SSE(C_2) = 1.11 + 1.44 + 0.11 = 2.67$

(3) $c(-1,4)$, $h(-3,4)$, $i=(-2,0)$

- $(-1-(-2))^2 + (4-2.67)^2 = 2.78$
- $(-3-(-2))^2 + (4-2.67)^2 = 2.78$
- $(-2-(-2))^2 + (0-2.67)^2 = 7.11$

$SSE(C_3) = 2.78 + 2.78 + 7.11 = 12.67$

$SSE_{k\text{-means}} = 30 + 2.67 + 12.67 = 45.34$

$SSE_{k\text{-means}}(45.34) < SSE_{complete}(50.07) < SSE_{single}(71.48)$

k-means provides better clustering according to the SSE criterion, as it groups the data more compactly.

3. (8 points) [**Frequent Itemset**] [**Graded By: Tural Mehtiyev**] For the transaction Table 6 given below, please answer the following questions:

TID	Items Bought
T1	{C, D, A, G}
T2	{F, B, G, H}
T3	{G, C, H, B, F}
T4	{C, G, B, H}
T5	{E, G, F, H}
T6	{D, E, F, G, C}
T7	{F, C, H, G, D}
T8	{D, H, G, C, B}
T9	{F, D, H, B}
T10	{B, G, H}
T11	{C, E, H, A}
T12	{H, B, G, D}
T13	{E, F, H, A}
T14	{G, F}
T15	{C, H, D, B}

Table 3: Transactions Data

A:3 D:7 E:4
C:8 C-G:6 C-F:3 C-B:4 C-H:6
G:11 G-F:6 G-B:6 G-H:8
F:8 F-B:3 F-H:5
B:8 B-H:8
H:12

(a) (1 point) Explain what is frequent itemset and give an example of 2-itemset that is frequent itemset with support count = 8.

(b) (3 points) Explain what is closed frequent itemset and list ALL of them with support count = 8. No partial credit.

(c) (3 points) Explain what is maximal frequent and list ALL of maximal itemset with support count = 8. No partial credit.

(d) (1 point) Compute the support and confidence for association rule $\{G, H\} \rightarrow \{B\}$.

(a) A frequent itemset is an itemset whose support count is greater than or equal to given minimum support threshold.
2-itemset {B,H} is a frequent itemset with support count 8 (appears in T2, T3, T4, T8, T9, T10, T12, T15)
{G,H} (appears in T2, T3, T4, T5, T7, T8, T10, T12)

(b) Support count = 8
1- itemset : {B}, {C}, {F}
1- itemset : {B,H}, {G,H}
Closed Closed

{C} support < 8 → closed
{F} support < 8 → closed
{B} → X closed

{G,H} support < 8 → closed
{B,H} support < 8 → closed

A closed frequent item set is a frequent itemset for which no strict superset has the same support count.
For this dataset, the closed frequent itemsets with support count 8 are {C}, {F}, {B,H}, {G,H}.

TID	Items Bought
T1	{C, D, A, G}
T2	{F, B, G, H}
T3	{G, C, H, B, F}
T4	{C, G, B, H}
T5	{E, G, F, H}
T6	{D, E, F, G, C}
T7	{F, C, H, G, D}
T8	{D, H, G, C, B}
T9	{F, D, H, B}
T10	{B, G, H}
T11	{C, E, H, A}
T12	{H, B, G, D}
T13	{E, F, H, A}
T14	{G, F}
T15	{C, H, D, B}

(c) frequent ≥ 8

- 1 itemset
  - {B}: 8 → x maximal : {B,H}: 8
  - {C}: 8 → maximal : supersets {C,G}, {C,F}, {C,B}, {C,H} < 8
  - {F}: 8 → maximal : supersets {G,F}, {F,B}, {F,H} < 8
  - {G}: 11 → x maximal : {G,H} = 8
  - {H}: 12 → x maximal : {B,H}, {G,H} = 8
- 2 itemset
  - {B,H}: 8 → maximal
  - {G,H}: 8 → maximal

{B, H, G}: 6

A maximal frequent itemset is a frequent itemset for which no strict superset is also frequent.

For this dataset, the maximal frequent itemsets with minimum support count 8 are {C}, {F}, {B,H}, {G,H}

(d) association rule {G,H} → {B}

- frequency of {G,H,B}: 6 (T2, T3, T4, T8, T10, T12)

  support = $\frac{6}{15} = 0.4$

- frequency of {G,H}: 8 (T2, T3, T4, T5, T7, T8, T10, T12)

  {G,H,B}: 6

  confidence = $\frac{6}{8} = 0.75$

4. (13 points) [**Association Analysis**] [**GRADER: Ian Holmes**] Consider the following school supply transactions shown in the Table 4 below.

Transaction ID	Items ordered
1	{Backpack, Eraser, Marker} 3
2	{Calculator, Marker, Scissors} 3
3	{Marker, Pencil} 2
4	{Eraser, Pen, Scissors} 3
5	{Eraser, Glue, Marker, Notebook} 4
6	{Notebook, Pen, Scissors} 3
7	{Eraser, Glue, Pen, Pencil} 4
8	{Backpack, Calculator, Glue, Marker, Pen} 5
9	{Glue, Marker, Pen} 3
10	{Backpack, Notebook, Scissors} 3

Table 4: School Supply Transactions Data

For each of the following question, briefly explain your answers in 2-3 sentences. NO PARTIAL CREDIT.

(a) (2 points) How many unique items are in this data set? What is the maximum size of itemsets that can be extracted from this data set (only including itemsets that have ≥ 1 support count)?

(b) (2 points) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

(c) (2 points) What is the *maximum number* of 3-itemsets that can be derived from this data set (including those have zero support)?

(d) (3 points) Find an itemset (of size 2 or larger) that has the largest support.

(e) (4 points) Find one pair of items, a and b, such that the rules $\{a\} \rightarrow \{b\}$ and $\{b\} \rightarrow \{a\}$ have the same confidence, and the support for each rule is greater than or equal to 0.2.

(a) Unique item : Backpack, Eraser, Marker, Calculator, Scissors, Pencil, Pen, Glue, Notebook
(9) (3) (4) (6) (2) (4) (2) (5) (4) (3)

maximum size : 5 (T8)

(b) association rule item : 9

Max rules = $3^d - 2^{d+1} + 1$

$= 3^9 - 2^{10} + 1 = 19683 - 1024 + 1 = 18660$

(c) item d = 9

$\frac{9!}{3!(9-3)!} = \frac{9!}{3!6!} = \frac{9 \cdot 8 \cdot 7}{3 \cdot 2 \cdot 1} = 84$

maximum number of 3-itemsets = 84

(d) ▲ frequency : Marker > Pen > Glue = Scissors = Eraser > Backpack = Notebook > Calculator = Pencil

Support : 3 → {Glue, Marker} : 3

{Glue, Pen} : 3

Support : 2 → {Marker, Pen} : 2, {Backpack, Marker} : 2, {Pen, Scissors} : 2, {Notebook, Scissors} : 2, {Eraser, Marker} : 2, {Eraser, Glue} : 2

(e) {a} → {b} confidence = {b} → {a} confidence

$conf(a \rightarrow b) = \frac{support(a,b)}{support(a)}$

$conf(b \rightarrow a) = \frac{support(a,b)}{support(b)}$

∴ support(a) = support(b) ≥ 0.2

support(a, b) ≥ 2

{Glue, Eraser} : 2 - T5, T7

Conf(Glue → Eraser) = $\frac{2}{4} = 0.5$

Conf(Eraser → Glue) = $\frac{2}{4} = 0.5$

5. (24 points) [**Apriori algorithm**] [**GRADED By Rajesh Debnath**] Consider the data set shown in Table 5 and answer the following questions using Apriori algorithm.

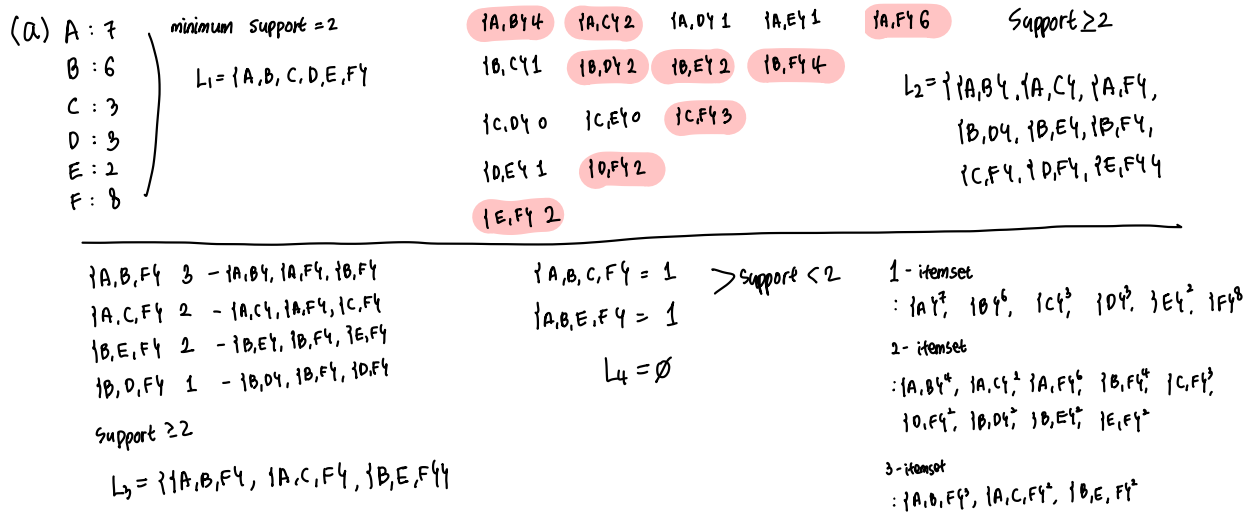
TID	Items
t_1	A, B, C, F
t_2	A, D, F
t_3	A, B
t_4	A, B, F
t_5	B, D
t_6	A, B, E, F
t_7	A, C, F
t_8	A, F
t_9	B, D, E, F
t_{10}	C, F

Table 5: Apriori algorithm

(a) (10 points) Show (compute) each step of the frequent itemset generation process using the Apriori algorithm, with a minimum support count of 2.

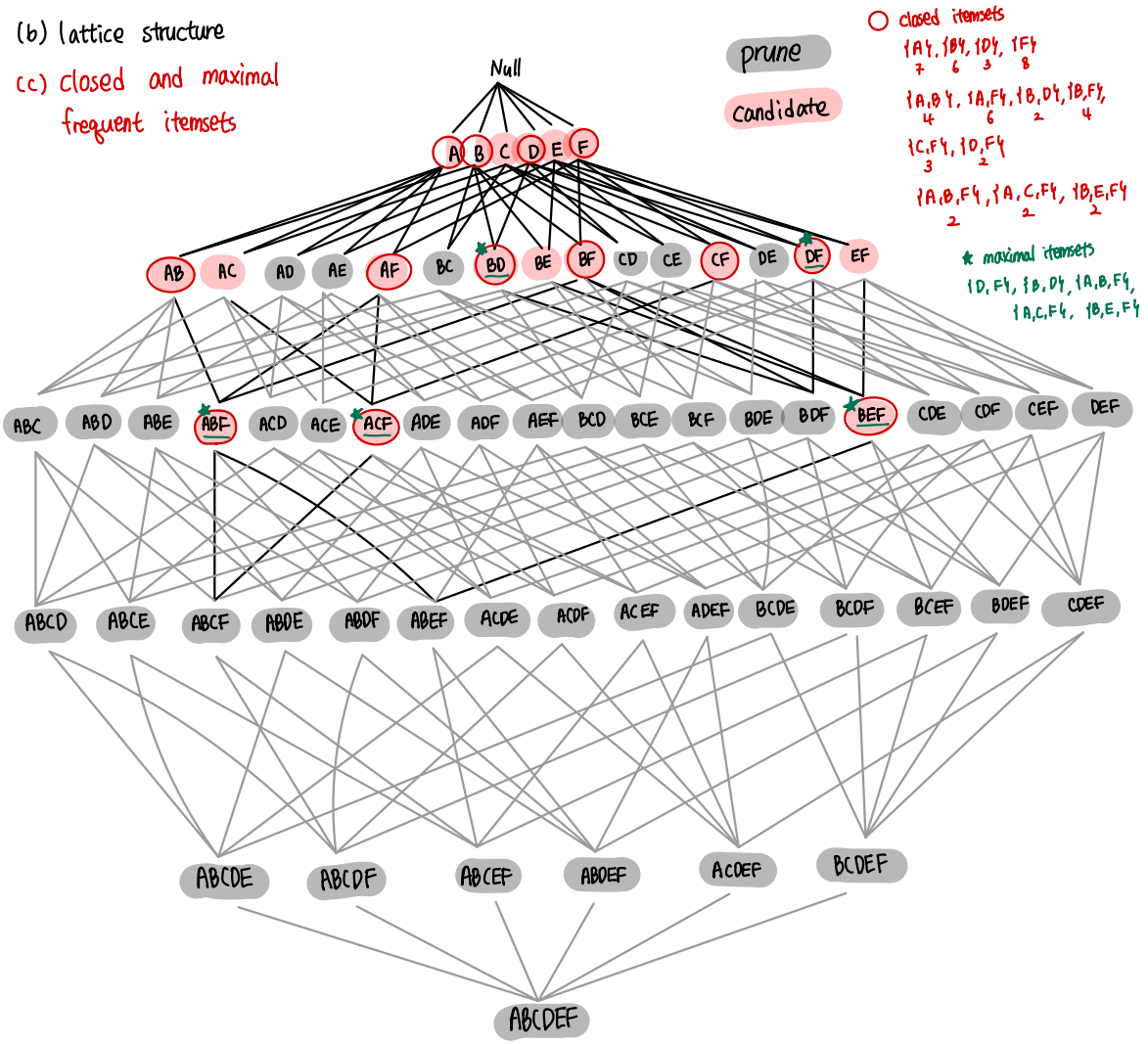
(b) (10 points) Show the lattice structure for the data given in the table above and mark the pruned branches if any. (Scanned hand-drawing is acceptable as long as it is clear.)

(c) (4 points) Mark closed and maximal frequent itemsets on the lattice structure from (b) if there are any.



(b) lattice structure

(c) closed and maximal frequent itemsets



6. (30 points) [**Frequent Pattern Tree**] [**GRADED by Tural Mehtiyev**] Consider the following data set shown in Table 6 and answer the following questions using FP-Tree.

TID	Items Bought
T1	{A, B, D, G}
T2	{B, D, E, H}
T3	{A, B, D, H}
T4	{B, D, F, G}
T5	{A, B, C, D}
T6	{A, B, C, G}
T7	{A, B, C, D, E, F}
T8	{B, C, E, F, G}
T9	{A, C, D, H}
T10	{C, D}
T11	{D, E, F, H}
T12	{A, D, E, G, H}
T13	{A, B, E, F, H}
T14	{A, C, F, G}
T15	{A, B, C}

A: 10
B: 10
C: 8
D: 10
E: 6
F: 6
G: 6
H: 6

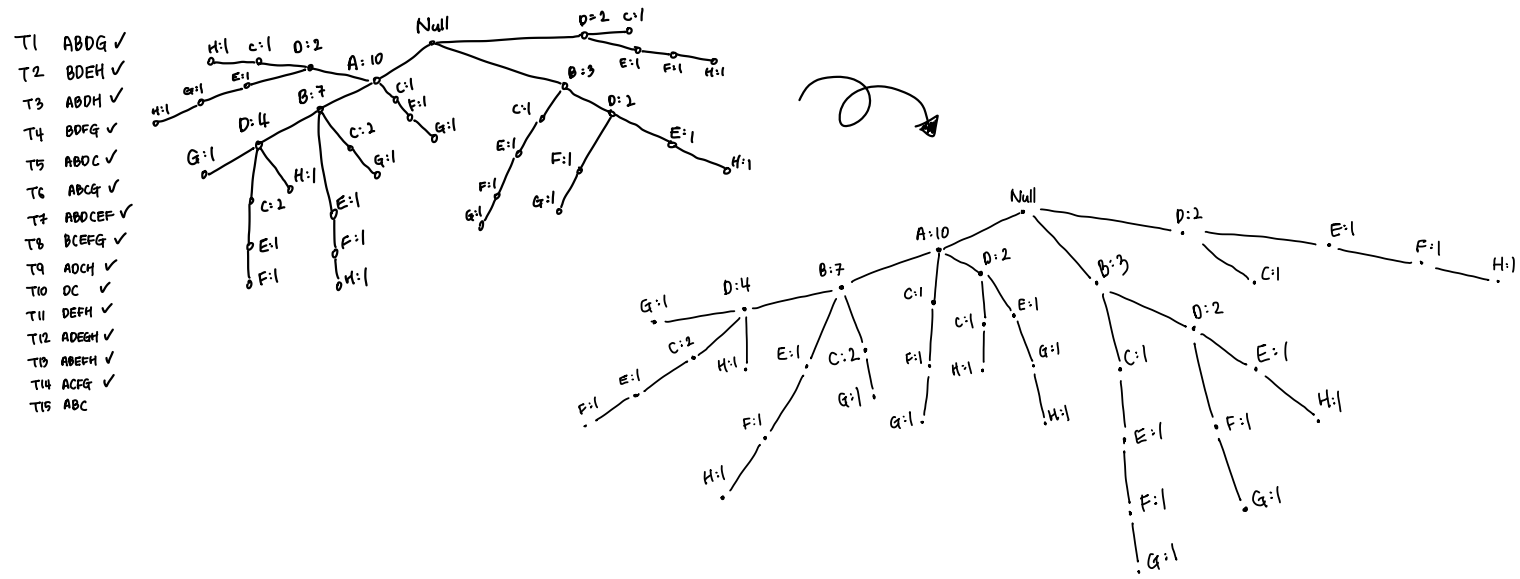
Table 6: Transactions Data

(a) (15 points) Construct an FP-Tree for the set of transactions in the table below as the first step towards identifying the itemsets with minimum support count of 2 (at least 2 occurrences). Hint: Do not forget to include the header table that locates the starts of the corresponding linked item lists through the FP-Tree. NO PARTIAL CREDIT.

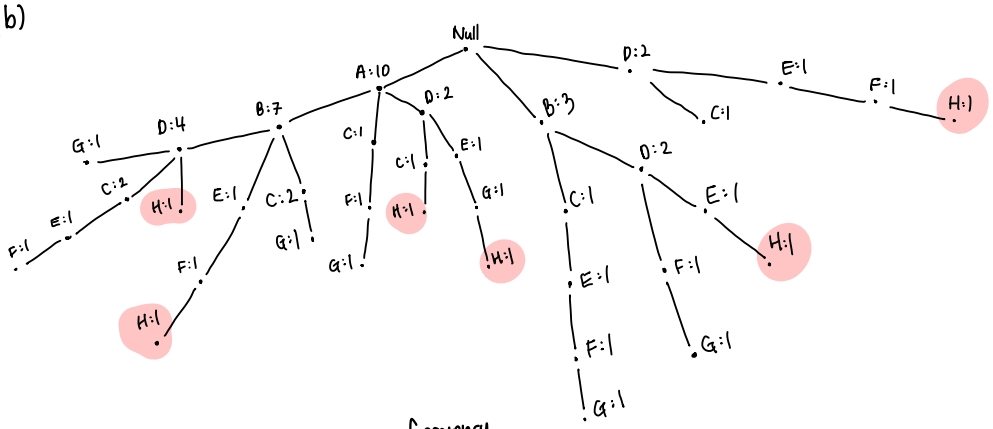
(b) (15 points) Using the FP-Tree constructed and support count = 2, generate all the frequent patterns with the base of item H step by step. List the frequent itemsets in alphabetical order. NO PARTIAL CREDIT.

(a) A=B=D > C > E=F=G=H ≥2
10 8 6

T1 ABDG ✓
T2 BDEH ✓
T3 ABDH ✓
T4 BDFG ✓
T5 ABDC ✓
T6 ABCG ✓
T7 ABDCEF ✓
T8 BCEFG ✓
T9 ADCH ✓
T10 DC ✓
T11 DEFH ✓
T12 ADEGH ✓
T13 ABEFH ✓
T14 ACFG ✓
T15 ABC



(b)



A→B→D→H : prefix {A,B,D}
A→B→E→F→H : prefix {A,B,E,F}
A→D→C→H : prefix {A,D,C}
A→D→E→G→H : prefix {A,D,E,G}
B→D→E→H : prefix {B,D,E}
D→E→F→H : prefix {D,E,F}

frequency
A : 4
B : 3
C : 1 < 2 X
D : 5
E : 4
F : 2
G : 1 < 2 X

{A,B,D} → {D,A,B}
{A,B,E,F} → {A,E,B,F}
{A,D,C} → {D,A}
{A,D,E,G} → {D,A,E}
{B,D,E} → {D,E,B}
{D,E,F} → {D,E,F}

D > A = E > B > F

{D,A,B}
{A,E,B,F}
{D,A}
{D,A,E}
{D,E,B}
{D,E,F}

A : 4
B : 3
D : 5
E : 4
F : 2
(H : 6)

AB : 2 AD : 3 AE : 2 ~~AF : 1~~
BD : 2 BE : 2 ~~BF : 1~~
DE : 3 ~~DF : 1~~ EF : 2
→ AB, AD, AE, BD, BE, DE, EF

ABD : 1 ABE : 1 ADE : 1
AEF : 1 BDE : 1 BEF : 1
DEF : 1
} < 2

* frequent items *

{H} ①
{A,H}, {B,H}, {D,H}, {E,H}, {F,H} ⑤
{A,B,H}, {A,D,H}, {A,E,H}, {B,D,H}, {B,E,H}, {D,E,H}, {E,F,H} ⑦
} total : 13