למידת מכונה - תרגיל בית מס׳ 4
אשכול היררכי – Hirarchical Clustering

מרצה: ד״ר לואי עבדאללה

מתרגלים: מר. סאלח אבו שאהין

מר. אנדריאס נסייר

תרגיל ללמידה עצמי לא להגשה

תרגיל מספר 1:

בטבלה שלפניך שמונה נקודות במישור, עליך לבצע אשכול היררכי Hierarchical Clustering של נקודות אלו, ולצייר את ה- dendrogram של האשכול כאשר משתמשים ב- :

א) single-linkage
ב) complete-linkage
ג) average-linkage
ד) centroid-linkage

מה תהיה תוצאת האשכול עבור 3 אשכולות בכל אחת מהשיטות הנ"ל?

Sample	X	Y
A_1	2	5
A_2	8	4
A_3	5	8
A_4	7	5
A_5	1	2
A_6	4	9
A_7	2	3
A_8	6	9

Solution:

1.א. טבלת מרחקים – Single Linkage

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0.000	6.083	4.243	5.000	3.162	4.472	2.000	5.657
A2	6.083	0.000	5.000	1.414	7.280	6.403	6.083	5.385
A3	4.243	5.000	0.000	3.606	7.211	1.414	5.831	1.414
A4	5.000	1.414	3.606	0.000	6.708	5.000	5.385	4.123
A5	3.162	7.280	7.211	6.708	0.000	7.616	1.414	8.602
A6	4.472	6.403	1.414	5.000	7.616	0.000	6.325	2.000
A7	2.000	6.083	5.831	5.385	1.414	6.325	0.000	7.211
A8	5.657	5.385	1.414	4.123	8.602	2.000	7.211	0.000

רואים שיש לנו 4 זוגות שונים שהמרחק האוקלידי בניהם הינו הקטן ביותר: 1.414 נבחר באחד מהם, נניח $A_2 \cup A_4$

	A1	$A2 \cup A4$	A3	A5	A6	A7	A8
A1	0.000	5.000	4.243	3.162	4.472	2.000	5.657
$A2 \cup A4$	5.000	0.000	3.606	6.708	5.000	5.385	4.123
A3	4.243	3.606	0.000	7.211	1.414	5.831	1.414
A5	3.162	6.708	7.211	0.000	7.616	1.414	8.602
A6	4.472	5.000	1.414	7.616	0.000	6.325	2.000
A7	2.000	5.385	5.831	1.414	6.325	0.000	7.211
A8	5.657	4.123	1.414	8.602	2.000	7.211	0.000

כנ"ל גם פה נבחר באחד הזוגות נניח: $A_3 \cup A_6$

	A1	$A2 \cup A4$	$A3 \cup A6$	A5	A7	A8
A1	0.000	5.000	4.243	3.162	2.000	5.657
$A2 \cup A4$	5.000	0.000	3.606	6.708	5.385	4.123
$A3 \cup A6$	4.243	3.606	0.000	7.211	5.831	1.414
A5	3.162	6.708	7.211	0.000	1.414	8.602
A7	2.000	5.385	5.831	1.414	0.000	7.211
A8	5.657	4.123	1.414	8.602	7.211	0

$A_3 \cup A_6 \cup A_8$

	A1	$A2 \cup A4$	$A_3 \cup A_6 \cup A_8$	A5	A7
A1	0.000	5.000	4.243	3.162	2.000
$A2 \cup A4$	5.000	0.000	3.606	6.708	5.385
A3 ∪ A6	4.243	3.606	0.000	7.211	5.831
A5	3.162	6.708	7.211	0.000	1.414
A7	2.000	5.385	5.831	1.414	0.000

$A_5 \cup A_7$

	A1	$A2 \cup A4$	$A_3 \cup A_6 \cup A_8$	$A_5 \cup A_7$
A1	0.000	5.000	4.243	2.000
$A2 \cup A4$	5.000	0.000	3.606	5.385
A3 ∪ A6	4.243	3.606	0.000	5.831
$A_5 \cup A_7$	2.000	5.385	5.831	0.000

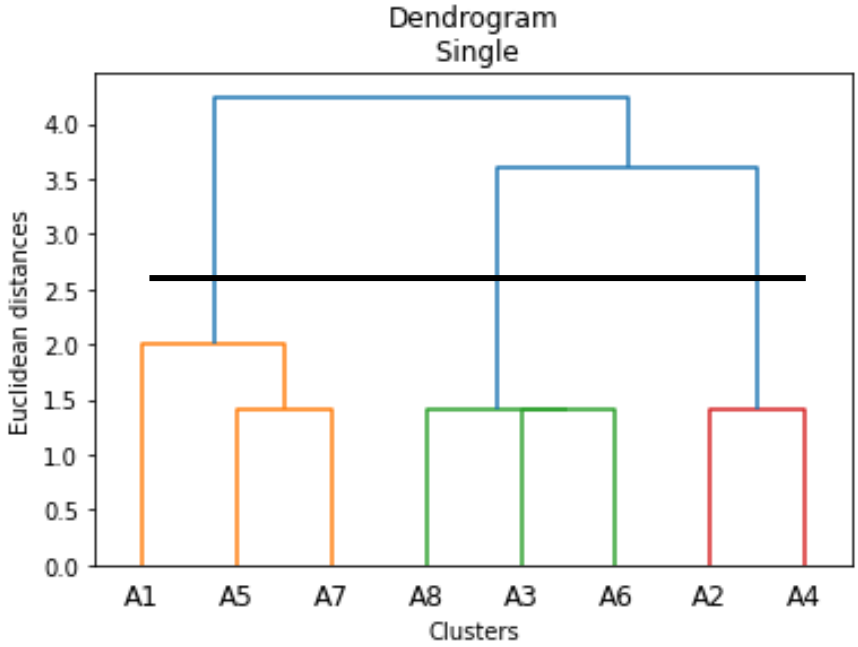
$A_1 \cup A_5 \cup A_7$

	$A_1 \cup A_5 \cup A_7$	$A2 \cup A4$	$A_3 \cup A_6 \cup A_8$
$A_1 \cup A_5 \cup A_7$	0.000	5.000	4.243
$A2 \cup A4$	5.000	0.000	3.606
A3 ∪ A6	4.243	3.606	0.000

$A_3 \cup A_6 \cup A_8 \cup A_2 \cup A_4$

	$A_1 \cup A_5 \cup A_7$	$A_3 \cup A_6 \cup A_8 \cup A_2 \cup A_4$
$A_1 \cup A_5 \cup A_7$	0.000	4.243
$A_3 \cup A_6 \cup A_8 \cup A_2 \cup A_4$	4.243	0.000

נותרו לנו 2 אשכולות כלומר נאחד ביניהם: $\{A1 \cup A5 \cup A_7\} \cup \{A3 \cup A6 \cup A8 \cup A2 \cup A4\}$



ע"מ למצוא את תוצאת האשכול עבור 3 אשכולות – נעביר קו אופקי מקביל לציר X בכך שיחתוך את הדנדרוגרמה ב- 3 נקודות חיתוך (קו שחור בגרף) ונקבל:

$$Cluster1: \{A_1, A_5, A_7\} \quad Cluster2: \{A_8, A_3, A_6\} \quad Cluster3: \{A_2, A_4\}$$

ב.טבלת מרחקים: Complete Linkage

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0.000	6.083	4.243	5.000	3.162	4.472	2.000	5.657
A2	6.083	0.000	5.000	1.414	7.280	6.403	6.083	5.385
A3	4.243	5.000	0.000	3.606	7.211	1.414	5.831	1.414
A4	5.000	1.414	3.606	0.000	6.708	5.000	5.385	4.123
A5	3.162	7.280	7.211	6.708	0.000	7.616	1.414	8.602
A6	4.472	6.403	1.414	5.000	7.616	0.000	6.325	2.000
A7	2.000	6.083	5.831	5.385	1.414	6.325	0.000	7.211
A8	5.657	5.385	1.414	4.123	8.602	2.000	7.211	0.000

$A_2 \cup A_4$

	A1	$A2 \cup A4$	A3	A5	A6	A7	A8
A1	0.000	6.083	4.243	3.162	4.472	2.000	5.657
$A2 \cup A4$	6.083	0.000	5.000	7.280	6.403	6.083	5.385
A3	4.243	5.000	0.000	7.211	1.414	5.831	1.414
A5	3.162	7.280	7.211	0.000	7.616	1.414	8.602
A6	4.472	6.403	1.414	7.616	0.000	6.325	2.000
A7	2.000	6.083	5.831	1.414	6.325	0.000	7.211
A8	5.657	5.385	1.414	8.602	2.000	7.211	0.000

$A_3 \cup A_6$

	A1	$A2 \cup A4$	$A3 \cup A6$	A5	A7	A8
A1	0.000	6.083	4.472	3.162	2.000	5.657
$A2 \cup A4$	6.083	0.000	6.403	7.280	6.083	5.385
$A3 \cup A6$	4.472	6.403	0.000	7.616	6.325	2.000
A5	3.162	7.280	7.616	0.000	1.414	8.602
A7	2.000	6.083	6.325	1.414	0.000	7.211
A8	5.657	5.385	2.000	8.602	7.211	0.000

$A_5 \cup A_7$

	A1	$A2 \cup A4$	$A3 \cup A6$	$A5 \cup A7$	A8
A1	0.000	6.083	4.472	3.162	5.657
$A2 \cup A4$	6.083	0.000	6.403	7.280	5.385
$A3 \cup A6$	4.472	6.403	0.000	7.616	2.000
$A5 \cup A7$	3.162	7.280	7.616	0.000	8.602
A8	5.657	5.385	2.000	8.602	0.000

$A_3 \cup A_6 \cup A_8$

	A1	$A2 \cup A4$	$A3 \cup A6 \cup A8$	$A5 \cup A7$
A1	0.000	6.083	5.657	3.162
$A2 \cup A4$	6.083	0.000	6.403	7.280
$A3 \cup A6 \cup A8$	5.657	6.403	0.000	8.602
$A5 \cup A7$	3.162	7.280	8.602	0.000

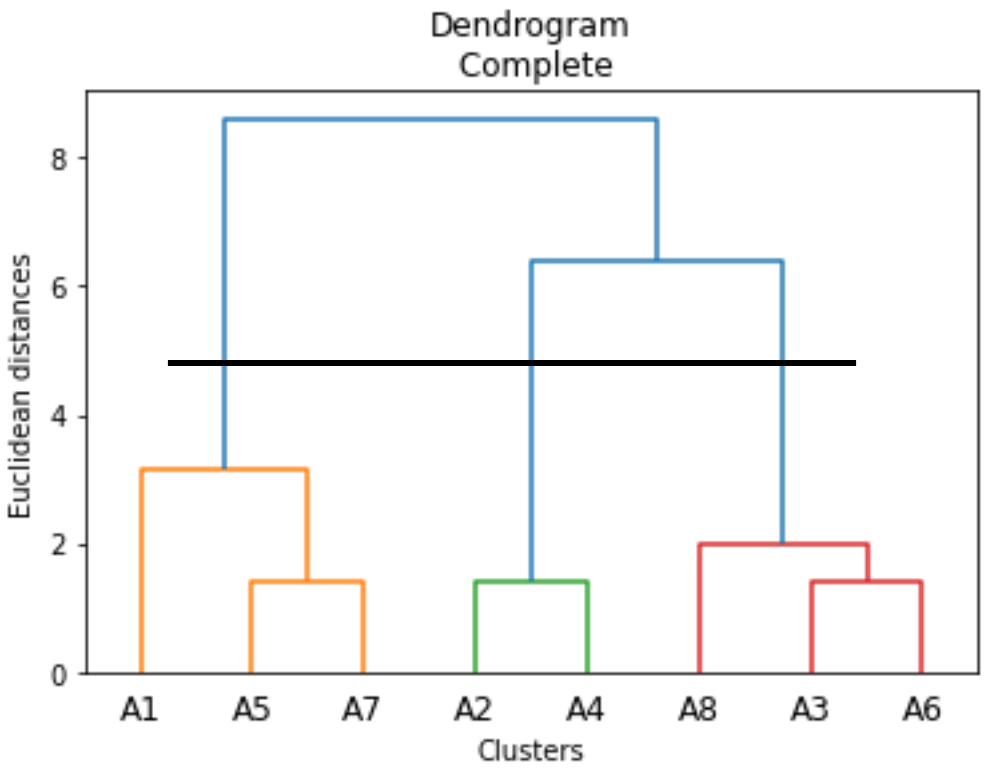
$A_1 \cup A_5 \cup A_7$

	$A_1 \cup A_5 \cup A_7$	$A2 \cup A4$	$A3 \cup A6 \cup A8$
$A_1 \cup A_5 \cup A_7$	0.000	7.280	8.602
$A2 \cup A4$	7.280	0.000	6.403
$A3 \cup A6 \cup A8$	8.602	6.403	0.000

$A_2 \cup A_4 \cup A_3 \cup A_6 \cup A_8$

	$A_1 \cup A_5 \cup A_7$	$A_2 \cup A_4 \cup A_3 \cup A_6 \cup A_8$
$A_1 \cup A_5 \cup A_7$	0.000	8.602
$A_2 \cup A_4 \cup A_3 \cup A_6 \cup A_8$	8.602	0.000

נותרו לנו 2 אשכולות כלומר נאחד ביניהם: $\{A1 \cup A5 \cup A_7\} \cup \{A_2 \cup A_4 \cup A_3 \cup A_6 \cup A_8\}$



ע"מ למצוא את תוצאת האשכול עבור 3 אשכולות – נעביר קו אופקי מקביל לציר X בכך שיחתוך את הדנדרוגרמה ב- 3 נקודות חיתוך (קו שחור בגרף) ונקבל:

$$Cluster1: \{A_1, A_5, A_7\} \quad Cluster2: \{A_2, A_4\} \quad Cluster3: \{A_8, A_3, A_6\}$$

ג. טבלת מרחקים: Average Linkage

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0.000	6.083	4.243	5.000	3.162	4.472	2.000	5.657
A2	6.083	0.000	5.000	1.414	7.280	6.403	6.083	5.385
A3	4.243	5.000	0.000	3.606	7.211	1.414	5.831	1.414
A4	5.000	1.414	3.606	0.000	6.708	5.000	5.385	4.123
A5	3.162	7.280	7.211	6.708	0.000	7.616	1.414	8.602
A6	4.472	6.403	1.414	5.000	7.616	0.000	6.325	2.000
A7	2.000	6.083	5.831	5.385	1.414	6.325	0.000	7.211
A8	5.657	5.385	1.414	4.123	8.602	2.000	7.211	0.000

$$A_3 \cup A_6$$

$$dist_{(\{A_1\},\{A_3,A_6\})} = \frac{\left(dist(A_1,A_3) + dist(A_1,A_6)\right)}{2} = \frac{4.243 + 4.472}{2} = 4.357$$

$$dist_{(\{A_2\},\{A_3,A_6\})} = \frac{\left(dist(A_2,A_3) + dist(A_2,A_6)\right)}{2} = \frac{5 + 6.403}{2} = 5.701$$

$$dist_{(\{A_4\},\{A_3,A_6\})} = \frac{\left(dist(A_4,A_3) + dist(A_4,A_6)\right)}{2} = \frac{3.606 + 5}{2} = 4.303$$

$$dist_{(\{A_5\},\{A_3,A_6\})} = \frac{\left(dist(A_5,A_3) + dist(A_5,A_6)\right)}{2} = \frac{(7.211 + 7.616)}{2} = 7.413$$

$$dist_{(\{A_7\},\{A_3,A_6\})} = \frac{\left(dist(A_7,A_3) + dist(A_7,A_6)\right)}{2} = \frac{(5.831 + 6.325)}{2} = 6.078$$

$$dist_{(\{A_8\},\{A_3,A_6\})} = \frac{\left(dist(A_8,A_3) + dist(A_8,A_6)\right)}{2} = \frac{(1.414 + 2)}{2} = 1.707$$

	A1	A2	$A3 \cup A6$	A4	A5	A7	A8
A1	0.000	6.083	4.357	5.000	3.162	2.000	5.657
A2	6.083	0.000	5.701	1.414	7.280	6.083	5.385
$A3 \cup A6$	4.357	5.701	0.000	4.303	7.413	6.078	1.707
A4	5.000	1.414	4.303	0.000	6.708	5.385	4.123
A5	3.162	7.280	7.413	6.708	0.000	1.414	8.602
A7	2.000	6.083	6.078	5.385	1.414	0.000	7.211
A8	5.657	5.385	1.707	4.123	8.602	7.211	0.000

$A5 \cup A7$

$$dist_{(\{A_1\},\{A_5,A_7\})} = \frac{\left(dist(A_1,A_5) + dist(A_1,A_7)\right)}{2} = \frac{3.162 + 2}{2} = 2.581$$

$$dist_{(\{A_2\},\{A_5,A_7\})} = \frac{\left(dist(A_2,A_5) + dist(A_2,A_7)\right)}{2} = \frac{7.280 + 6.083}{2} = 6.6815$$

$$dist_{(\{A_3,A6\},\{A_5,A_7\})} = \frac{\left(dist(A_3,A_5) + dist(A_3,A_7) + dist(A_6,A_5) + dist(A_6,A_7)\right)}{4}$$
$$= \frac{7.211 + 5.831 + 7.616 + 6.325}{4} = 6.745$$

$$dist_{(\{A_4\},\{A_5,A_7\})} = \frac{\left(dist(A_4,A_5) + dist(A_4,A_7)\right)}{2} = \frac{6.708 + 5.385}{2} = 6.046$$

$$dist_{(\{A_8\},\{A_5,A_7\})} = \frac{\left(dist(A_8,A_5) + dist(A_8,A_7)\right)}{2} = \frac{(8.602 + 7.211)}{2} = 7.906$$

	A1	A2	$A3 \cup A6$	A4	$A5 \cup A7$	A8
A1	0.000	6.083	4.357	5.000	2.581	5.657
A2	6.083	0.000	5.701	1.414	6.681	5.385
$A3 \cup A6$	4.357	5.701	0.000	4.303	6.745	1.707
A4	5.000	1.414	4.303	0.000	6.046	4.123
$A5 \cup A7$	2.581	6.681	6.745	6.046	0.000	7.906
A8	5.657	5.385	1.707	4.123	7.906	0.000

$A2 \cup A4$

$$dist_{(\{A_1\},\{A_2,A_4\})} = \frac{\left(dist(A_1,A_2) + dist(A_1,A_4)\right)}{2} = \frac{6.083 + 5}{2} = 5.541$$

$$dist_{(\{A_3,A6\},\{A_2,A_4\})} = \frac{\left(dist(A_3,A_2) + dist(A_3,A_4) + dist(A_6,A_2) + dist(A_6,A_4)\right)}{4}$$

$$= \frac{5 + 3.606 + 6.403 + 5}{4} = 5.0022$$

$$dist_{(\{A_5,A_7\},\{A_2,A_4\})} = \frac{\left(dist(A_5,A_2) + dist(A_5,A_4) + dist(A_7,A_2) + dist(A_7,A_4)\right)}{2}$$
$$= \frac{7.280 + 6.708 + 6.083 + 5.385}{4} = 6.364$$

$$dist_{(\{A_8\},\{A_2,A_4\})} = \frac{\left(dist(A_8,A_2) + dist(A_8,A_4)\right)}{2} = \frac{(5.385 + 4.123)}{2} = 4.754$$

	A1	$A2 \cup A4$	$A3 \cup A6$	$A5 \cup A7$	A8
A1	0.000	5.541	4.357	2.581	5.657
$A2 \cup A4$	5.541	0.000	5.002	6.364	4.754
$A3 \cup A6$	4.357	5.002	0.000	6.745	1.707
$A5 \cup A7$	2.581	6.364	6.745	0.000	7.906
A8	5.657	4.754	1.707	7.906	0.000

$$A3 \cup A6 \cup A8$$

$$dist_{(\{A_1\},\{A_3,A_6,A_8\})} = \frac{\left(dist(A_1,A_3) + dist(A_1,A_6) + dist(A_1,A_8)\right)}{3}$$
$$= \frac{4.243 + 4.472 + 5.657}{3} = 4.79$$

$$dist_{(\{A_2,A_4\},\{A_3,A_6,A_8\})} = \frac{\begin{pmatrix} dist(A_2,A_3) + dist(A_2,A_6) + dist(A_2,A_8) \\ +dist(A_4,A_3) + dist(A_4,A_6) + dist(A_4,A_8) \end{pmatrix}}{6}$$

$$= \frac{5 + 6.403 + 5.385 + 3.606 + 5 + 4.123}{6} = 4.919$$

$$dist_{(\{A_5,A_7\},\{A_3,A_6,A_8\})} = \frac{\begin{pmatrix} dist(A_5,A_3) + dist(A_5,A_6) + dist(A_5,A_8) \\ +dist(A_7,A_3) + dist(A_7,A_6) + dist(A_7,A_8) \end{pmatrix}}{6}$$

$$= \frac{7.211 + 7.616 + 8.602 + 5.831 + 6.325 + 7.211}{6} = 7.132$$

	A1	$A2 \cup A4$	$A3 \cup A6 \cup A8$	$A5 \cup A7$
A1	0.000	5.541	4.79	2.581
$A2 \cup A4$	5.541	0.000	4.919	6.364
$A3 \cup A6 \cup A8$	4.79	4.919	0.000	7.132
$A5 \cup A7$	2.581	6.364	7.132	0.000

$$A1 \cup A5 \cup A7$$

$$dist_{(\{A_2,A_4\},\{A_1,A_5,A_7\})} = \frac{\begin{pmatrix} dist(A_2,A_1) + dist(A_2,A_5) + dist(A_2,A_7) \\ +dist(A_4,A_1) + dist(A_4,A_5) + dist(A_4,A_7) \end{pmatrix}}{6}$$
$$= \frac{6.083 + 7.28 + 6.083 + 5 + 6.708 + 5.385}{6} = 6.089$$

$$dist_{(\{A_3,A_6,A_8\},\{A_1,A_5,A_7\})} = \frac{\begin{pmatrix} dist(A_3,A_1) + dist(A_3,A_5) + dist(A_3,A_7) \\ +dist(A_6,A_1) + dist(A_6,A_5) + dist(A_6,A_7) \\ +dist(A_8,A_1) + dist(A_8,A_5) + dist(A_8,A_7) \end{pmatrix}}{9}$$

$$= \frac{\begin{matrix} 4.243 + 7.211 + 5.831 \\ +4.472 + 7.616 + 6.325 \\ +5.657 + 8.602 + 7.211 \end{matrix}}{9} = 6.352$$

	$A1 \cup A5 \cup A7$	$A2 \cup A4$	$A3 \cup A6 \cup A8$
$A1 \cup A5 \cup A7$	0.000	6.089	6.352
$A2 \cup A4$	6.089	0.000	4.919
$A3 \cup A6 \cup A8$	6.352	4.919	0.000

$$A2 \cup A4 \cup A3 \cup A6 \cup A8$$

$$dist_{(\{A_1,A_5,A_7\},\{A_2,A_4,A_3,A_6,A_8\})} =$$

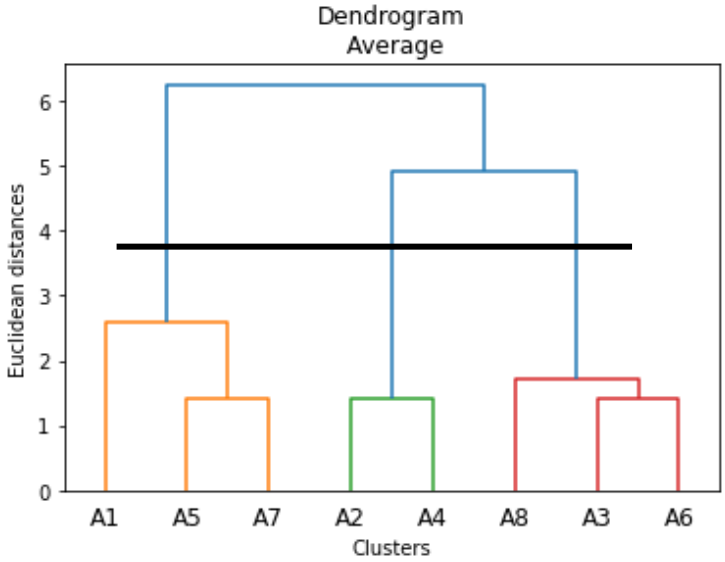
$$\frac{\begin{pmatrix} dist(A_1,A_2)+dist(A_1,A_4)+dist(A_1,A_3)+dist(A_1,A_6)+dist(A_1,A_8)+ \\ dist(A_5,A_2)+dist(A_5,A_4)+dist(A_5,A_3)+dist(A_5,A_6)+dist(A_5,A_8)+ \\ dist(A_7,A_2)+dist(A_7,A_4)+dist(A_7,A_3)+dist(A_7,A_6)+dist(A_7,A_8)+ \end{pmatrix}}{6}$$

$$= \frac{\begin{matrix} 6.083+5+4.243+4.472+5.657 \\ +7.28+6.708+7.211+7.616+8.602 \\ +6.083+5.385+5.831+6.325+7.211 \end{matrix}}{15} = 6.247$$

	$A1 \cup A5 \cup A7$	$A2 \cup A4 \cup A3 \cup A6 \cup A8$
$A1 \cup A5 \cup A7$	0.000	6.247
$A2 \cup A4 \cup A3 \cup A6 \cup A8$	6.247	0.000

נותרו לנו 2 אשכולות אי לכך נמזג אותם ביחד וכך נקבל אשכול אחד:

$$\{A1 \cup A5 \cup A7\} \cup \{A2 \cup A4 \cup A3 \cup A6 \cup A8\}$$



ע"מ למצוא את תוצאת האשכול עבור 3 אשכולות – נעביר קו אופקי מקביל לציר X בכך שיחתוך את הדנדרוגרמה ב- 3 נקודות חיתוך (קו שחור בגרף) ונקבל:

$$Cluster1: \{A_1, A_5, A_7\} \quad Cluster2: \{A_2, A_4\} \quad Cluster3: \{A_8, A_3, A_6\}$$

ד. טבלת מרחקים – Centroid Linkage

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0.000	6.083	4.243	5.000	3.162	4.472	2.000	5.657
A2	6.083	0.000	5.000	1.414	7.280	6.403	6.083	5.385
A3	4.243	5.000	0.000	3.606	7.211	1.414	5.831	1.414
A4	5.000	1.414	3.606	0.000	6.708	5.000	5.385	4.123
A5	3.162	7.280	7.211	6.708	0.000	7.616	1.414	8.602
A6	4.472	6.403	1.414	5.000	7.616	0.000	6.325	2.000
A7	2.000	6.083	5.831	5.385	1.414	6.325	0.000	7.211
A8	5.657	5.385	1.414	4.123	8.602	2.000	7.211	0.000

טבלת הדגימות המקורית:

Sample	X	Y
A_1	2	5
A_2	8	4
A_3	5	8
A_4	7	5
A_5	1	2
A_6	4	9
A_7	2	3
A_8	6	9

רואים שהמרחק הקטן ביותר הינו בין אשכול A_5 ו- A_7 ולכן נבחר לאחר ביניהם:

נחשב את מרכז האשכול $A_5 \cup A_7$:

$$Center\{A_5, A_7\} = \frac{1+2}{2}, \frac{2+3}{2} = (1.5, 2.5)$$

$$dist(Center\ A_1, Center\{A_5, A_7\}) = \sqrt{(2-1.5)^2 + (5-2.5)^2} = 2.549$$

$$dist(Center\ A_2, Center\ \{A_5, A_7\}) = \sqrt{(8-1.5)^2 + (4-2.5)^2} = 6.67$$

$$dist(Center\ A_3, Center\ \{A_5, A_7\}) = \sqrt{(5-1.5)^2 + (8-2.5)^2} = 6.519$$

$$dist(Center\ A_4, Center\ \{A_5, A_7\}) = \sqrt{(7-1.5)^2 + (5-2.5)^2} = 6.041$$

$$dist(Center\ A_6, Center\ \{A_5, A_7\}) = \sqrt{(4-1.5)^2 + (9-2.5)^2} = 6.964$$

$$dist(Center\ A_8, Center\ \{A_5, A_7\}) = \sqrt{(6-1.5)^2 + (9-2.5)^2} = 7.9$$

	A1	A2	A3	A4	$A5 \cup A7$	A6	A8
A1	0.000	6.083	4.243	5.000	2.549	4.472	5.657
A2	6.083	0.000	5.000	1.414	6.67	6.403	5.385
A3	4.243	5.000	0.000	3.606	6.519	1.414	1.414
A4	5.000	1.414	3.606	0.000	6.041	5.000	4.123
$A5 \cup A7$	2.549	6.67	6.519	6.041	0.000	6.964	7.9
A6	4.472	6.403	1.414	5.000	6.964	0.000	2.000
A8	5.657	5.385	1.414	4.123	7.9	2.000	0.000

כעת, נבחר אחד מזוגות האשכולות שיש ביניהם מרחק הקטן ביותר, נבחר A_2 ו- A_4:

טבלת הדגימות המקורית:

Sample	X	Y
A_1	2	5
A_2	8	4
A_3	5	8
A_4	7	5
A_5	1	2
A_6	4	9
A_7	2	3
A_8	6	9

נחשב את מרכז האשכול $A_2 \cup A_4$:

$$Center\{A_2, A_4\} = \frac{8+7}{2}, \frac{4+5}{2} = (7.5, 4.5)$$

$$dist(Center A_1, Center\ \{A_2, A_4\}) = \sqrt{(2-7.5)^2 + (5-4.5)^2} = 5.522$$

$$dist(Center\ A_3, Center\ \{A_2, A_4\}) = \sqrt{(5-7.5)^2 + (8-4.5)^2} = 4.301$$

$$dist(Center\ \{A_5, A_7\}, Center\ \{A_2, A_4\}) = \sqrt{(1.5-7.5)^2 + (2.5-4.5)^2} = 6.324$$

$$dist(Center\ A_6, Center\ \{A_2, A_4\}) = \sqrt{(4-7.5)^2 + (9-4.5)^2} = 5.7$$

$$dist(Center\ A_8, Center\ \{A_2, A_4\}) = \sqrt{(6-7.5)^2 + (9-4.5)^2} = 4.743$$

	A1	$A2 \cup A4$	A3	$A5 \cup A7$	A6	A8
A1	0.000	5.522	4.243	2.549	4.472	5.657
$A2 \cup A4$	5.522	0.000	4.301	6.324	5.7	4.743
A3	4.243	4.301	0.000	6.519	1.414	1.414
$A5 \cup A7$	2.549	6.324	6.519	0.000	6.964	7.9
A6	4.472	5.7	1.414	6.964	0.000	2.000
A8	5.657	4.743	1.414	7.9	2.000	0.000

כעת, נבחר אחד מזוגות האשכולות שיש ביניהם מרחק הקטן ביותר, כלומר או A_3 ו- A_6 או A_3 ו- A_8 נבחר באחד מהם נניח ב- A_3 ו- A_6:

טבלת הדגימות המקורית:

Sample	X	Y
A_1	2	5
A_2	8	4
A_3	5	8
A_4	7	5
A_5	1	2
A_6	4	9
A_7	2	3
A_8	6	9

נחשב את מרכז האשכול $A_3 \cup A_6$:

$$Center\{A_3, A_6\} = \frac{5+4}{2}, \frac{8+9}{2} = (4.5, 8.5)$$

$$dist(Center\ A_1, Center\ \{A_3, A_6\}) = \sqrt{(2-4.5)^2 + (5-8.5)^2} = 4.301$$

$$dist(Center\ \{A_2, A_4\}, Center\{A_3, A_6\}) = \sqrt{(7.5-4.5)^2 + (4.5-8.5)^2} = 5$$

$$dist(Center\ \{A_5, A_7\}, Center\{A_3, A_6\}) = \sqrt{(1.5-4.5)^2 + (2.5-8.5)^2} = 6.708$$

$$dist(Center\ A_8, Center\{A_3, A_6\}) = \sqrt{(6-4.5)^2 + (9-8.5)^2} = 1.581$$

	A1	$A2 \cup A4$	$A_3 \cup A_6$	$A5 \cup A7$	A8
A1	0.000	5.522	4.301	2.549	5.657
$A2 \cup A4$	5.522	0.000	5	6.324	4.743
$A_3 \cup A_6$	4.243	4.301	0.000	6.519	1.414
$A5 \cup A7$	2.549	6.324	6.708	0.000	7.9
A8	5.657	4.743	1.581	7.9	0.000

כעת, רואים שהמרחק הקטן ביותר הינו בין אשכול $\{A_3 \cup A_6\}$ ו- A_8 ולכן נבחר לאחר ביניהם. להלן טבלת הדגימות המקורית:

Sample	X	Y
A_1	2	5
A_2	8	4
A_3	5	8
A_4	7	5
A_5	1	2
A_6	4	9
A_7	2	3
A_8	6	9

$$Center\{A_8, \{A_3 \cup A_6\}\} = \frac{5+4+6}{3}, \frac{8+9+9}{3} = (5{,}8\frac{2}{3})$$

$$dist(CenterA_1, Center\{A_3, A_6, A_8\}) = \sqrt{(2-5)^2 + \left(5 - 8\frac{2}{3}\right)^2} = 4.737$$

$$dist(Center\{A_2, A_4\}, Center\{A_3, A_6, A_8\}) = \sqrt{(7.5-5)^2 + \left(4.5 - 8\frac{2}{3}\right)^2} = 4.86$$

$$dist(Center\{A_5 \cup A_7\}, Center\{A_3, A_6, A_8\}) = \sqrt{(1.5-5)^2 + \left(2.5 - 8\frac{2}{3}\right)^2} = 7.090$$

	A1	$A2 \cup A4$	$A_3 \cup A_6 \cup A_8$	$A5 \cup A7$
A1	0.000	5.522	4.737	2.549
$A2 \cup A4$	5.522	0.000	4.86	6.324
$A_3 \cup A_6 \cup A_8$	4.737	4.86	0.000	7.090
$A5 \cup A7$	2.549	6.324	7.090	0.000

כעת, רואים שהמרחק הקטן ביותר הינו בין אשכול $\{A_5 \cup A_7\}$ ו- A_1 ולכן נבחר לאחר ביניהם. להלן טבלת הדגימות המקורית:

Sample	X	Y
A_1	2	5
A_2	8	4
A_3	5	8
A_4	7	5
A_5	1	2
A_6	4	9
A_7	2	3
A_8	6	9

$$Center\{A_1, \{A_5 \cup A_7\}\} = \frac{2+1+2}{3}, \frac{5+2+3}{3} = (1\frac{2}{3}, 3\frac{1}{3})$$

$$dist(Center\{A_2, A_4\}, Center\{A_1, A_5, A_7\}) = \sqrt{\left(7.5 - 1\frac{2}{3}\right)^2 + \left(4.5 - 3\frac{1}{3}\right)^2} = 5.948$$

$$dist(Center\{A_3, A_6, A_8\}, Center\{A_1, A_5, A_7\}) = \sqrt{\left(5 - 1\frac{2}{3}\right)^2 + \left(8\frac{2}{3} - 3\frac{1}{3}\right)^2} = 6.29$$

	$A1 \cup A5 \cup A7$	$A2 \cup A4$	$A_3 \cup A_6 \cup A_8$
$A1 \cup A5 \cup A7$	0.000	5.948	6.29
$A2 \cup A4$	5.948	0.000	4.86
$A_3 \cup A_6 \cup A_8$	6.29	4.86	0.000

כעת, רואים שהמרחק הקטן ביותר הינו בין אשכול $\{A_3 \cup A_6 \cup A_8\}$ ו- $\{A_2 \cup A_4\}$ ולכן נבחר לאחר ביניהם. להלן טבלת הדגימות המקורית:

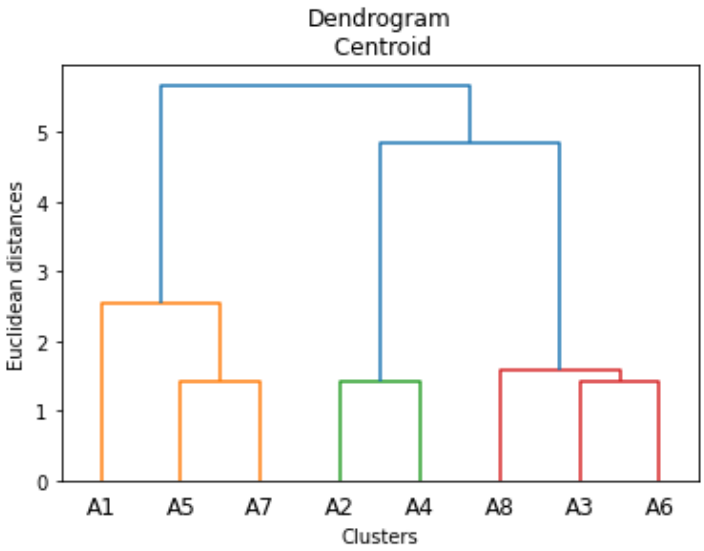
Sample	X	Y
A_1	2	5
A_2	8	4
A_3	5	8
A_4	7	5
A_5	1	2
A_6	4	9
A_7	2	3
A_8	6	9

$$Center\{\{A_2 \cup A_4\},\{A_3,A_6,A_8\}\} = \frac{8+5+7+4+6}{5},\frac{4+8+5+9+9}{5} = (6,7)$$

$$dist(Center\{A_1,A_5,A_7\},Center\{A_2,A_4,A_3,A_6,A_8\}) = \sqrt{\left(1\frac{2}{3}-6\right)^2+\left(3\frac{1}{3}-7\right)^2}$$
$$= 5.676$$

	$A1 \cup A5 \cup A7$	$A2 \cup A4 \cup A_3 \cup A_6 \cup A_8$
$A1 \cup A5 \cup A7$	0.000	5.676
$A2 \cup A4 \cup A_3 \cup A_6 \cup A_8$	5.676	0.000

נקבל בסוף אשכול אחד שהינו: $\{A_1 \cup A_5 \cup A_7\} \cup \{A_2 \cup A_4 \cup A_3 \cup A_6 \cup A_8\}$



ע"מ למצוא את תוצאת האשכול עבור 3 אשכולות – נעביר קו אופקי מקביל לציר X בכך שיחתוך את הדנדרוגרמה ב- 3 נקודות חיתוך (קו שחור בגרף) ונקבל:

$$Cluster1: \{A_1,A_5,A_7\} \quad Cluster2: \{A_8,A_3,A_6\} \quad Cluster3: \{A_2,A_4\}$$

שאלה מס' 2:

לתרגיל בית זה, צורף קובץ נתונים בשם "covid19_stocks.csv" המייצג ערכי המניות בחודשי תקופת הקורונה. להלן פירוט הפיצ'רים:

שם פיצ'ר	הסבר
Time in months	מספר חודשים מאז התפרצות נגיף הקורונה
Stock Value	ערכי מניות בתקופת הקורונה

א. הכרת הנתונים ועיבוד מקדים:

- יש לייבא ולטעון את כל קובץ הנתונים בעזרת הפייתון למשתנה data.
- הדפיסו את מימדי קובץ הנתונים.
- יש להציג את גרפי הפיזור Scatter Plot, עבור 2 הפיצ'רים.
  כאשר הגרף חייב להכיל כותרת ראשית וכותרות לצירים.
  כמה אשכולות אתם רואים בגרף? תשובה: 2 אשכולות

- נרמול:
  יש לנרמל את הנתונים שבמשתנה data באמצעות min/max ולשמור בתוך משתנה בשם min_max_data

ב. הרצת אלגוריתם Hierarchical Clustering:

- להריץ אלגוריתם אשכול היררכי עבור הנתונים שנורמלו לפי min/max כלומר הנתונים השמורים ב- min_max_data לפי 4 שיטות חישוב המרחקים בין אשכולות כלומר: single, complete, average and wards linkages
  כאשר נרצה להריץ אותם עבור 2 אשכולות בכל שיטה.
- להציג את סדר חיבור האשכולות (ה- Dendrogram).

ג. ויזואליזציה והסקת מסקנות:

- עבור כל אחד מגרפי הפיזור Scatter Plot הבאים, יש להציג את תוצאת האשכול בהתאם ל- 4 שיטות חישוב מרחק בין אשכולות, כאשר כל אשכול יהיה בצבע אחר, גם כאן לא לשכוח להוסיף כותרת ראשית וכותרות לצירים.

- איזה שיטת חישוב מרחק בין אשכולות הינה המתאימה ביותר לנתונים אלו? האם לדעתכם יש שיטה יותר טובה מהשנייה? יש להסביר איך זה מתיישב עם התוצאה שקיבלתם.
  תשובה: שיטת חישוב מרחק לפי Single Linkage תיתן את תוצאת האשכול הטובה ביותר לדאטה סט זה, היות ולפי שיטה זו, עדכון המרחקים יהיה לפי המרחק הקטן יותר מבין שני המרחקים, בין אשכול מסויים לבין כל אחד משני האשכולות שחוברו יחדיו, כתוצאה מכך, החיבור יהיה תחילה בין הדגימות בתוך כל אשכול.