Aufgabe 16:

Für die folgende Datenmatrix

Person	1	2	3	4	5
Alter	23	46	51	60	34
Einkommen (in Tsd)	25	55	100	84	40

wurde zum Teil bereits die Matrix der Distanzen zwischen den Objekten berechnet:

$$\boldsymbol{D} = \begin{bmatrix} 0 & 37.8 & 80.06 & 69.64 & 18.6 \\ & 0 & \dots & \dots & \dots \\ & & 0 & \dots & \dots \\ & & & 0 & \dots \\ & & & & 0 \end{bmatrix}.$$

(a) Welches Distanzmaß wurde bei der Berechnung verwendet? Begründen Sie Ihre Antwort anhand eines der schon berechneten Elemente. Berechnen Sie danach die noch fehlenden Elemente in $\boldsymbol{D}$.

(b) Bestimmen Sie ausgehend von fünf einelementigen Klassen der Anfangspartition

$$C^0 = \{C_1^0, \dots, C_5^0\} = \{\{1\}, \dots \{5\}\}$$

eine Hierarchie nach den folgenden Verfahren und zeichnen Sie die zugehörigen Dendrogramme:

i. Single Linkage

ii. Complete Linkage

Aufgabe 17:

Die folgende Tabelle enthält für fünf Länder die Lebenserwartung von Frauen und die Anzahl der Kinder pro Frau.

Objekt	Lebenserwartung	Kinder pro Frau
Deutschland	81.1	1.29
Indien	64.9	2.97
Indonesien	79.6	2.02
Israel	81.0	2.70
Japan	85.0	1.33

Die emprirische Kovarianzmatrix der beiden Variablen lautet

$$\boldsymbol{S} = \begin{bmatrix} 60.30 & -4.40 \\ -4.40 & 0.59 \end{bmatrix}.$$

Im Folgenden ist das Ziel, für diese 5 Länder ein hierarchisches Clustering mit der euklidischen Distanz durchzuführen.

(a) Warum ist es sinnvoll, vor der Berechnung der euklidischen Distanzmatrix die Variablen zu skalieren?

(b) Nach Skalierung der Variablen ergibt sich als euklidische Distanzmatrix

$$\boldsymbol{D} = \begin{bmatrix} 0 & 3.0 & 1.0 & 1.8 & 0.5 \\ & 0 & 2.3 & 2.1 & 3.4 \\ & & 0 & 0.9 & 1.1 \\ & & & 0 & 1.8 \\ & & & & 0 \end{bmatrix}$$

Führen Sie eine Clusteranalyse mit Complete Linkage durch.

(c) Zeichnen Sie das zugehörige Dendrogramm.

(d) Nennen Sie ein anderes häufiges Linkage-Verfahren und erklären Sie kurz, was die beiden Verfahren unterscheidet.