Clusteranalyse

Aufgabe 1:

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

Land	Lebenserwartung	Kinder pro Frau
Deutschland	81.1	1.29
Indonesien	79.6	2.02
Japan	85.0	1.33
Israel	81.0	2.90

Die entsprechende Kovarianzmatrix ist

$$\mathbf{X} = \begin{pmatrix} 5.38 & -0.8 \\ -0.8 & 0.57 \end{pmatrix}.$$

Ziel der Aufgabe ist es, ein hierarchisches Clustering für diese 4 Länder mit der euklidischen Distanz durchzuführen.

a) Warum ist es sinnvoll, die Variablen zu normalisieren (zentrieren und skalieren), bevor man die euklidische Distanzmatrix berechnet? Erstellen Sie eine Tabelle mit den normalisierten Variablen.

b) Nach einer Normalisierung der Variablen ergibt sich die unvollständige euklidische Distanzmatrix:

Land	Deutschland	Indonesien	Japan	Israel
Deutschland	0			
Indonesien	1.16	0		
Japan	1.68	2.50	0	
Israel	2.13	??	2.70	??

(i) Vervollständigen Sie die euklidische Distanzmatrix.

(ii) Führen Sie das Complete-Linkage Verfahren durch.

c) Erstellen Sie das entsprechende Dendrogramm.

Aufgabe 2:

Lösen Sie nun folgende Aufgaben in `R`. Verwenden Sie dabei die Tabelle aus Aufgabe 1.

a) Erstellen Sie einen Datensatz mit `Land`, `Lebenserwartung` und `KindproFrau`.

b) Berechnen Sie die euklidische Distanzmatrix für die normalisierten und nicht normalisierten Variablen.

c) Laden Sie das Paket `cluster` und machen Sie sich mit den Funktionen `agnes()` und `diana()` vertraut.

d) Führen Sie ein agglomeratives Clustering mit `agnes()` durch. Verwenden Sie hierbei die Manhattan-Metrik und die euklidische Metrik für das Complete Linkage-Verfahren und vergleichen Sie die Dendrogramme für beide Metriken.
Hinweis: Verwenden Sie den normalisierten Datensatz!

e) Führen Sie ein divisives Clustering mit `diana()` durch. Verwenden Sie hierbei die Manhattan-Metrik und die euklidische Metrik und vergleichen Sie beide Dendrogramme.

f) Führen Sie ein Clustering mit dem Kmeans-Verfahren durch.