Aufgabenblatt 2

1. Gegeben sei der nachstehende zweidimensionale Datensatz. Führen Sie ein K-means Clustering mit $K = 3$ unter Verwendung der euklidischen Distanz durch. Verwenden Sie die ersten drei Punkte als Anfangszentroiden. Geben Sie bei jeder Algorithmeniteration jeweils die Distanzen zwischen Zentroiden und allen Punkten an und berechnen Sie nach jeder Neuzuordnung der Punkte die veränderten Zentroiden.

	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12
x	2.0	2.0	2.0	2.5	2.5	3.0	4.0	4.0	4.5	4.5	4.5	4.5
y	1.0	1.5	2.0	1.0	2.0	4.0	1.0	2.5	1.0	1.5	2.5	3.0

2. Eine Schule möchte ihre Schüler nach den Leistungen bei zwei Zwischenprüfungen gruppieren. Es wird davon ausgegangen, dass es mindestens 2 Cluster von Schülern gibt. Laden Sie die Datei `clustering-student-mat.csv` ein. Die Datei enthält zu jeder der beiden Prüfungen die Anzahl der erzielten Punktzahl für insgesamt 395 Schüler.
   Führen Sie je ein K-means-Clustering für alle $k \in \{2, 3, \ldots, 8\}$ durch. Stellen Sie die Clusterzuordnungen der Punkte in einem Streudiagramm (Scatter Plot) dar.

3. Ermitteln Sie für das Clustering aus Ausgabe 2 den optimalen Wert für die Anzahl der Cluster K mithilfe des Silhouetten-Koeffizienten. Bewerten Sie das Ergebnis im Hinblick auf die Repräsentativität der Zentroiden bezüglich ihres Clusters.

4. Gegeben sei die nachstehende Distanzmatrix. Führen Sie agglomeratives hierarchisches Clustering mit *single* und *complete* Linkage durch. Stellen Sie das Ergebnis in einem Dendrogramm dar. Das Dendrogramm sollte die Reihenfolge des Zusammenfügens der Punkte darstellen.

	a	b	c	d	e
a	0.00	0.02	0.90	0.36	0.53
b	0.02	0.00	0.65	0.15	0.24
c	0.90	0.65	0.00	0.59	0.45
d	0.36	0.15	0.59	0.90	0.56
e	0.53	0.24	0.45	0.56	0.00

Datensatz für Aufgabe 2:
http://isgwww.cs.uni-magdeburg.de/cv/lehre/VisAnalytics/material/exercise/datasets/clustering-student-mat.csv