ETH Zürich
Statistik II (BIOL, HST) **Lösungen zur Serie 12 - 01**

1. **Aufgabe**
Von 441 kranken Patienten wurden Blutproben entnommen. Das Blut wurde auf 5 Bestandteile im Blut (gleiche Einheiten) untersucht. Es wird vermutet, dass es zwei unterschiedliche Gruppen von Patienten geben könnte. Können Sie diese Vermutung bestätigen?

   Die Daten befinden sich im data frame `dat` im R-data file *ueb*806253.*rda*.

   (a) Laden Sie die Daten. Der Wert in Zeile 323 und Spalte 1 ist 0.76.

   (b) Die Daten in den verschiedenen Spalten wurden in den gleichen Einheiten gemessen und sollen *nicht* skaliert werden. Die maximale Standardabweichung der Variablen ist 4.49.

   (c) Berechnen Sie nun k-means mit euklidischer Distanz auf den Daten um zwei Gruppen zu finden. Verwenden Sie random seed 23 und 10 zufällige Startkonfigurationen im k-means Algorithmus. Das Within-Sum-Of-Squares ist 1463.09.

   (d) Personen 37 und 55 sind im gleichen Cluster.

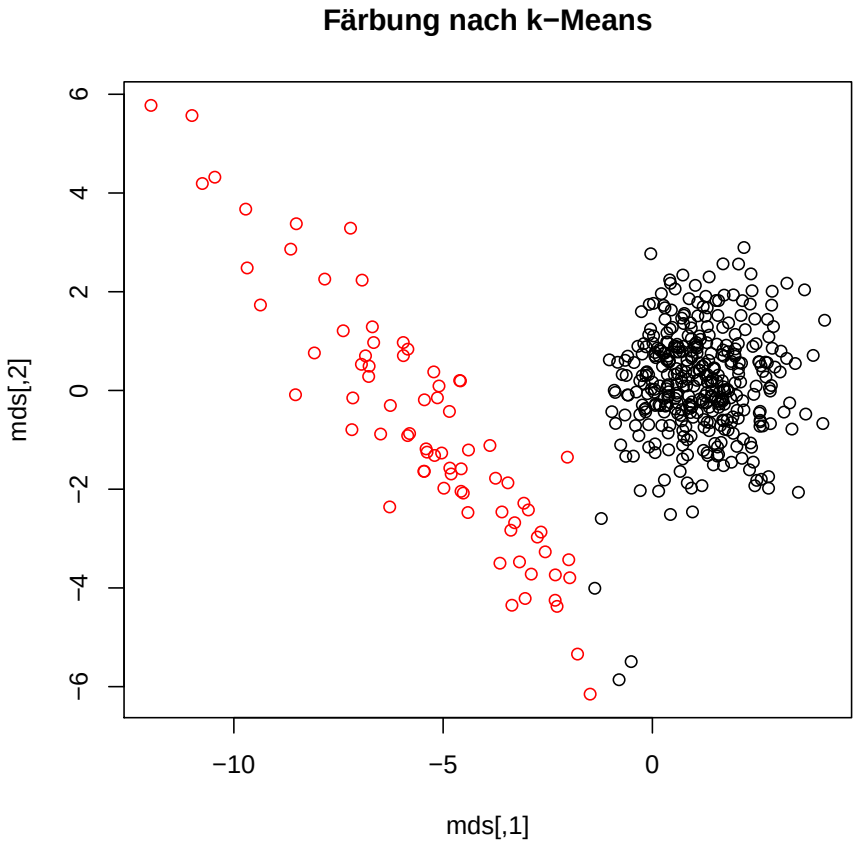
   (e) Die "Average silhouette width" der Cluster ist 0.19.

   (f) Berechnen Sie nun noch ein hierarchisches Clustering mit average linkage. Schneiden Sie das Dendrogramm so ab, dass sich zwei Gruppen ergeben. Vergleichen Sie nun die Gruppierung von k-means und von hierarchischem Clustering. Die beiden Methoden finden in etwa die gleichen Cluster (d.h., die Cluster sind gleich wenn man maximal 10% der Datenpunkte ignorieren darf).

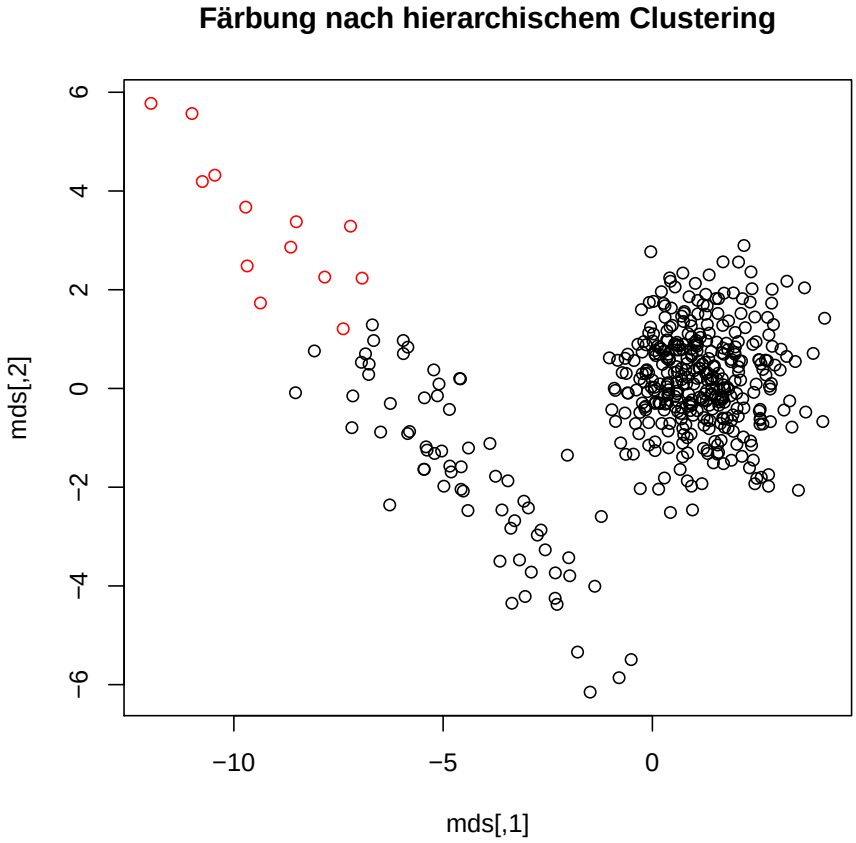
   Lösung

```
> load("ueb806253.rda")
> # a)
> row <- 323
> col <- 1
> round(dat[row, col], 2)

[1] -0.22

> # b)
> round(apply(dat, 2, sd), 2)

   1    2    3    4    5
1.29 1.96 1.36 2.03 1.23

> # c)
> set.seed(23)
> km <- kmeans(dat, centers = 2, nstart = 10)
> round(sum(km$withinss), 2)

[1] 2971.07

> # d)
> km$cluster[37]
```

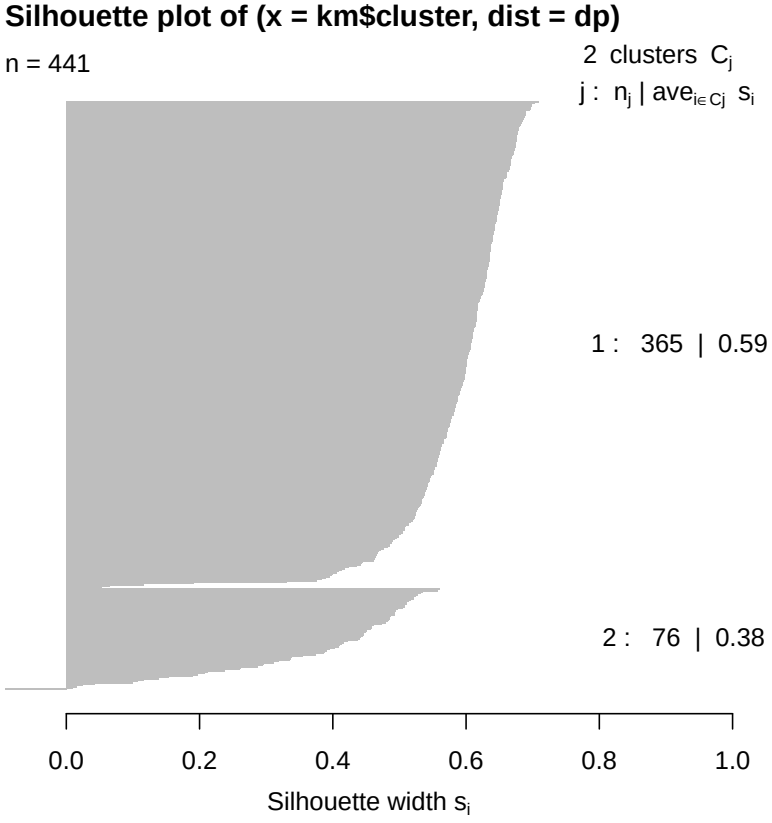
```
[1] 2

> km$cluster[55]

[1] 2

> # e)

> library(cluster)
> dp <- dist(dat)
> sl <- silhouette(km$cluster, dp)
> round(mean(sl[, 3]), 2)

[1] 0.55

> # f)

> cc <- hclust(dp, method = "average")
> grpsHC <- cutree(cc, k = 2)
> table(km$cluster, grpsHC)

   grpsHC
      1   2
  1 365   0
  2  63  13

> # Optionale Plots

> mds <- cmdscale(dp)
> plot(mds, col=km$cluster, main="Färbung nach k-Means")
```




```
> plot(mds, col=grpsHC, main="Färbung nach hierarchischem Clustering")
```


Färbung nach hierarchischem Clustering
mds[,2]
6
4
2
0
−2
−4
−6
−10
−5
0
mds[,1]

```
> plot(sl)
```



(a) **False**. Der korrekte Wert ist -0.22.

(b) **False**. Der korrekte Wert ist 2.03.

(c) **False**. Der korrekte Wert ist 2971.07.

(d) **True**. Person 37 ist im Cluster 2, Person 55 ist im Cluster 2.

(e) **False**. Der korrekte Wert ist 0.55.

(f) **False**. Die beiden Cluster weichen in 63 Datenpunkten voneinander ab.

2. **Aufgabe**
In einer Umfrage wurden 8 Personen nach ihren Präferenzen in 13 verschiedenen Themen befragt. Die Antworten sind entweder numerisch oder kategorisch. Verwende ein geeignetes Verschiedenheitsmass um die Verschiedenheit zwischen Personen zu beschreiben.

Die Daten befinden sich im data frame `dat` im R-data file *ueb*122895.*rda*.

(a) Person 5 ist am ähnlichsten zu Person 4.

Lösung

```
> load("ueb122895.rda")

> library(cluster)
> daisy(dat, metric = "gower")

Dissimilarities :
          1         2         3         4         5         6         7
```

```
2 0.3769834
3 0.5114636 0.4819755
4 0.7371994 0.5538651 0.3757010
5 0.7045230 0.5981118 0.2515174 0.1575205
6 0.4204179 0.1532776 0.4152141 0.4932558 0.5313505
7 0.3106013 0.5863964 0.5117852 0.7685996 0.7392163 0.6111434
8 0.2345396 0.4490002 0.5346612 0.7603970 0.7277207 0.4131677 0.2620744

Metric :  mixed ;  Types = I, I, I, I, N, I, I, I, I, I, N, I, I
Number of objects : 8
```

(a) **True**. Die Gowers-Verschiedenheit zwischen Person 5 und 4 ist 0.16 und kleiner als alle anderen Distanzen von Person 5 zu anderen Personen.

3. **Aufgabe**
Von 423 kranken Patienten wurden Blutproben entnommen. Das Blut wurde auf 5 Bestandteile im Blut (gleiche Einheiten) untersucht. Es wird vermutet, dass es zwei unterschiedliche Gruppen von Patienten geben könnte. Können Sie diese Vermutung bestätigen?

Die Daten befinden sich im data frame `dat` im R-data file *ueb*885287.*rda*.

(a) Laden Sie die Daten. Der Wert in Zeile 400 und Spalte 3 ist 0.23.

(b) Die Daten in den verschiedenen Spalten wurden in den gleichen Einheiten gemessen und sollen *nicht* skaliert werden. Die maximale Standardabweichung der Variablen ist 1.98.

(c) Berechnen Sie nun k-means mit euklidischer Distanz auf den Daten um zwei Gruppen zu finden. Verwenden Sie random seed 23 und 10 zufällige Startkonfigurationen im k-means Algorithmus. Das Within-Sum-Of-Squares ist 1267.01.

(d) Personen 36 und 67 sind im gleichen Cluster.

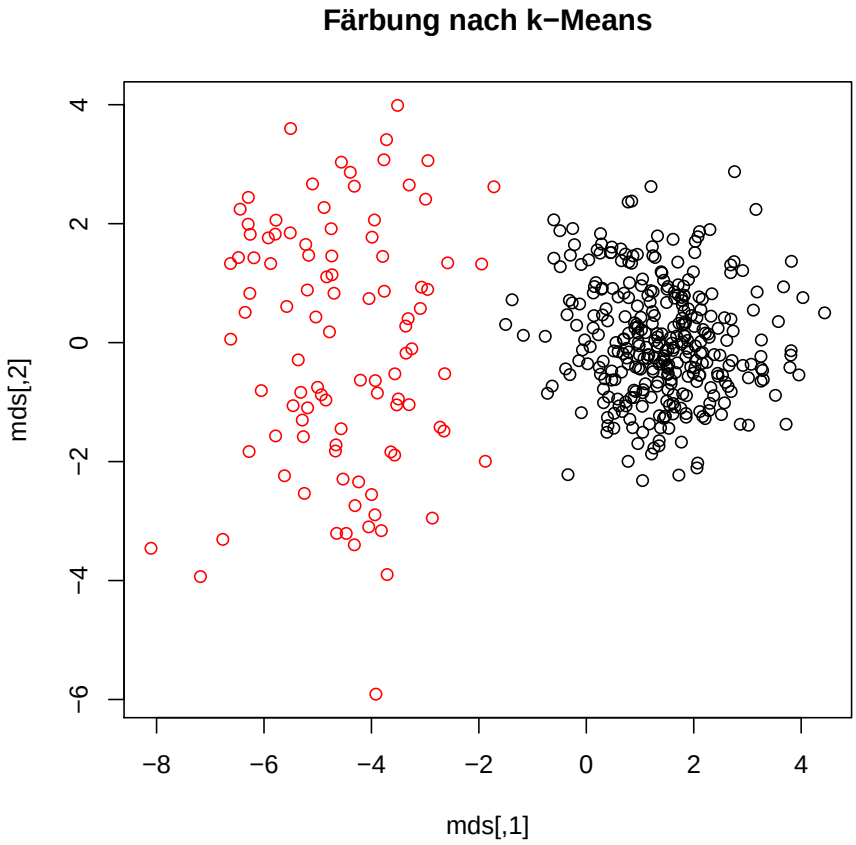
(e) Die “Average silhouette width” der Cluster ist 0.19.

(f) Berechnen Sie nun noch ein hierarchisches Clustering mit average linkage. Schneiden Sie das Dendrogramm so ab, dass sich zwei Gruppen ergeben. Vergleichen Sie nun die Gruppierung von k-means und von hierarchischem Clustering. Die beiden Methoden finden substantiell verschiedene Cluster (d.h., die Cluster unterscheiden sich, auch wenn man 10% der Datenpunkte ignorieren dürfte).

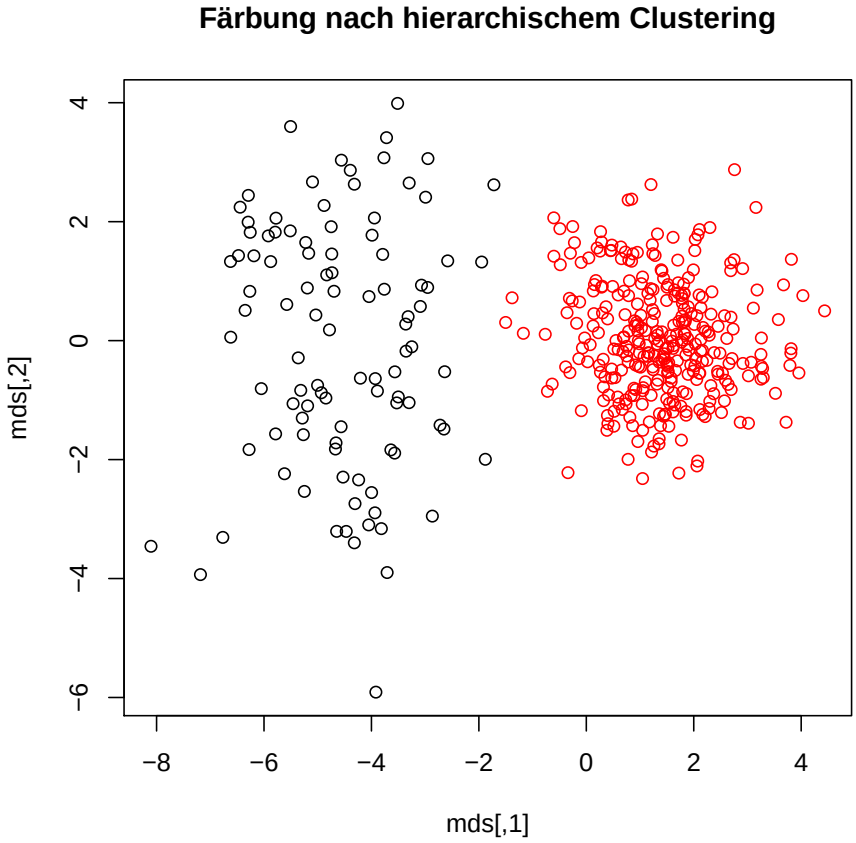
Lösung

```
> load("ueb885287.rda")
> # a)
> row <- 400
> col <- 3
> round(dat[row, col], 2)

[1] 0.23

> # b)
> round(apply(dat, 2, sd), 2)

   1    2    3    4    5
1.28 1.98 1.33 1.63 1.43

> # c)
```

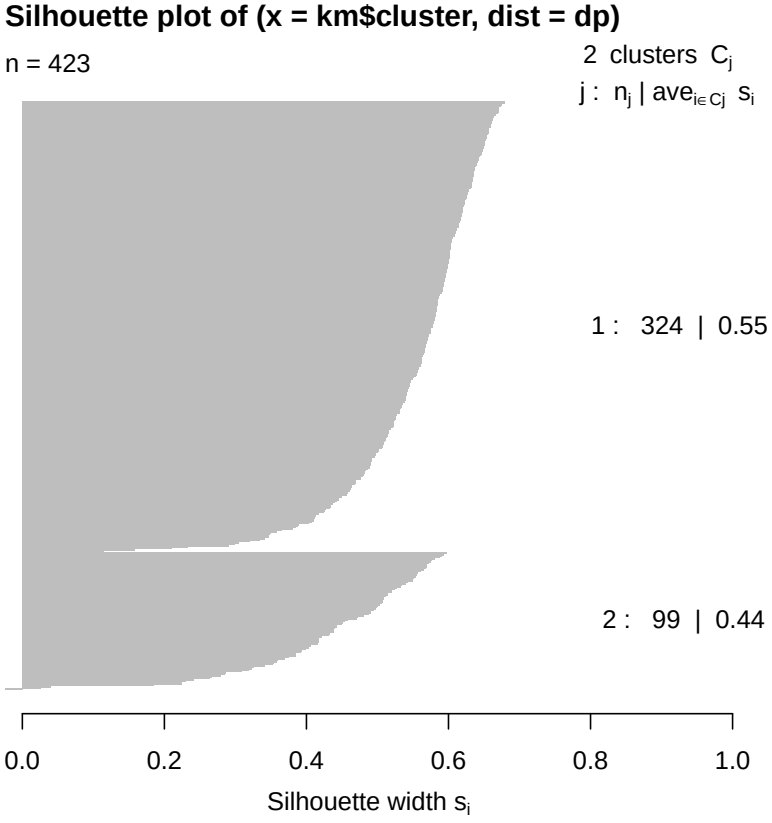
```
> set.seed(23)
> km <- kmeans(dat, centers = 2, nstart = 10)
> round(sum(km$withinss), 2)

[1] 2377.98

> # d)
> km$cluster[36]

[1] 2

> km$cluster[67]

[1] 2

> # e)

> library(cluster)
> dp <- dist(dat)
> sl <- silhouette(km$cluster, dp)
> round(mean(sl[, 3]), 2)

[1] 0.53

> # f)

> cc <- hclust(dp, method = "average")
> grpsHC <- cutree(cc, k = 2)
> table(km$cluster, grpsHC)

   grpsHC
      1   2
  1   0 324
  2  99   0

> # Optionale Plots

> mds <- cmdscale(dp)
> plot(mds, col=km$cluster, main="Färbung nach k-Means")
```




```
> plot(mds, col=grpsHC, main="Färbung nach hierarchischem Clustering")
```


Färbung nach hierarchischem Clustering
mds[,2]
mds[,1]
4
2
0
−2
−4
−6
−8
−6
−4
−2
0
2
4

```
> plot(sl)
```



(a) **True**. Der korrekte Wert ist 0.23.

(b) **True**. Der korrekte Wert ist 1.98.

(c) **False**. Der korrekte Wert ist 2377.98.

(d) **True**. Person 36 ist im Cluster 2, Person 67 ist im Cluster 2.

(e) **False**. Der korrekte Wert ist 0.53.

(f) **False**. Die beiden Cluster weichen in 0 Datenpunkten voneinander ab.