Statistinės duomenų analizės praktinės užduotys

2017

12. **Klasterinė analizė.**

(a) (ČM II.8.4) Socialiniai bei ekonominiai 10-ties apskričių 1999 metų rodikliai pateikti lentelėje.

Apskritis	a_1	a_2	a_3	a_4	a_5	a_6	a_7
Alytaus	22.5	9.3	260433	122	24.5	171585	-1.3
Kauno	22.9	11.1	1052822	209	27.3	792675	-0.6
Klaipėdos	19.6	12.5	690000	247	28.7	850725	-0.2
Marijampolės	19.8	7.7	146914	164	24.1	20806	-0.6
Panevėžio	22.9	9.8	396002	210	23.9	334743	-1.8
Šiaulių	21.1	8.8	296607	189	24.3	147408	-1.1
Tauragės	21.6	7.0	35921	156	22.4	17674	-1.3
Telšių	21.1	10.0	503425	171	24.4	115108	-0.6
Utenos	24.5	10.3	268631	140	23.5	91212	-3.9
Vilniaus	20.6	15.8	2502666	236	26.9	3959258	-1.0

Čia a_1 - gyventojų aprūpinimas gyvenamuoju plotu (kiek vienam gyventojui vidutiniškai tenka naudingo ploto (m^2); a_2 - bendrasis vidaus produktas (BVP) vienam gyventojui (tūkst. litų); a_3 - materialinės investicijos (t8kst. litų); a_4 - nusikalstamumas (kiek užregistruota nusikaltimų, tenkančių 10000 gyventojų); a_5 - gyventojų aprūpinimas telefonais butuose (100 gyventojų); a_6 - tiesioginės užsienio investicijos (tūkst. litų, sausio 1 d. duomenys); a_7 - natūralus gyventojų prieauglis (1000 gyventojų).

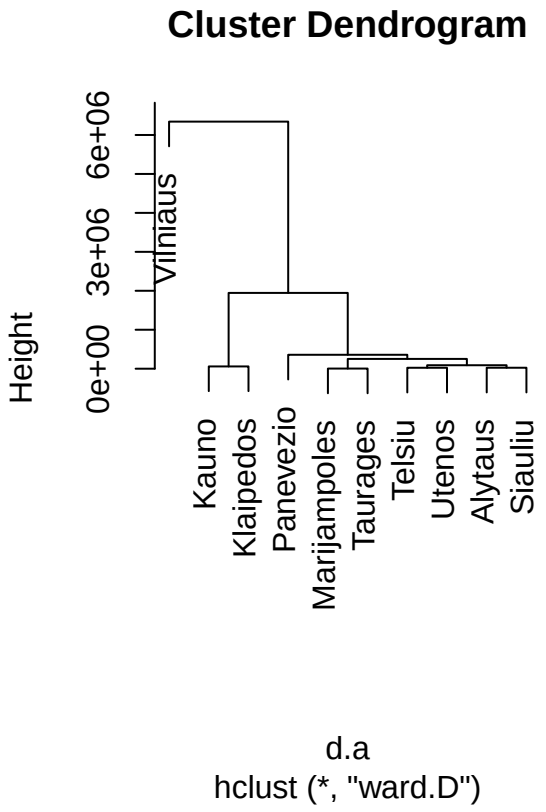
```
dat.a <- matrix(c(22.5, 9.3, 260433, 122, 24.5, 171585, -1.3,
                  22.9, 11.1, 1052822, 209, 27.3, 792675, -0.6,
                  19.6, 12.5, 690000, 247, 28.7, 850725, -0.2,
                  19.8, 7.7, 146914, 164, 24.1, 20806, -0.6,
                  22.9, 9.8, 396002, 210, 23.9, 334743, -1.8,
                  21.1, 8.8, 296607, 189, 24.3, 147408, -1.1,
                  21.6, 7.0, 35921, 156, 22.4, 17674, -1.3,
                  21.1, 10.0, 503425, 171, 24.4, 115108, -0.8,
                  24.5, 10.3, 268631, 140, 23.5, 91212, -3.9,
                  20.6, 15.8, 2502666, 236, 26.9, 3959258, -1.0),
                nrow = 10, ncol = 7,
                byrow = TRUE,
                dimnames = list(c("Alytaus", "Kauno", "Klaipedos", "Marijampoles",
                                  "Panevezio", "Siauliu", "Taurages", "Telsiu",
                                  "Utenos", "Vilniaus"),
                                c("a1", "a2", "a3", "a4", "a5", "a6", "a7")))

dat.a <- as.data.frame(dat.a)
```

Kokius apskričių klasterius galima sudaryti pagal šiuos rodiklius? Vordo ir pilnosios jungties metodais suklasterizuokite apskritis pagal požymius a_1, a_2, a_4 , a_5 ir a_6.

```
d.a <- dist(dat.a[,-c(3, 7)], method = "euclidean")
par(mfrow=c(1,2))
cluster.a.ward <- hclust(d.a, method="ward")
```

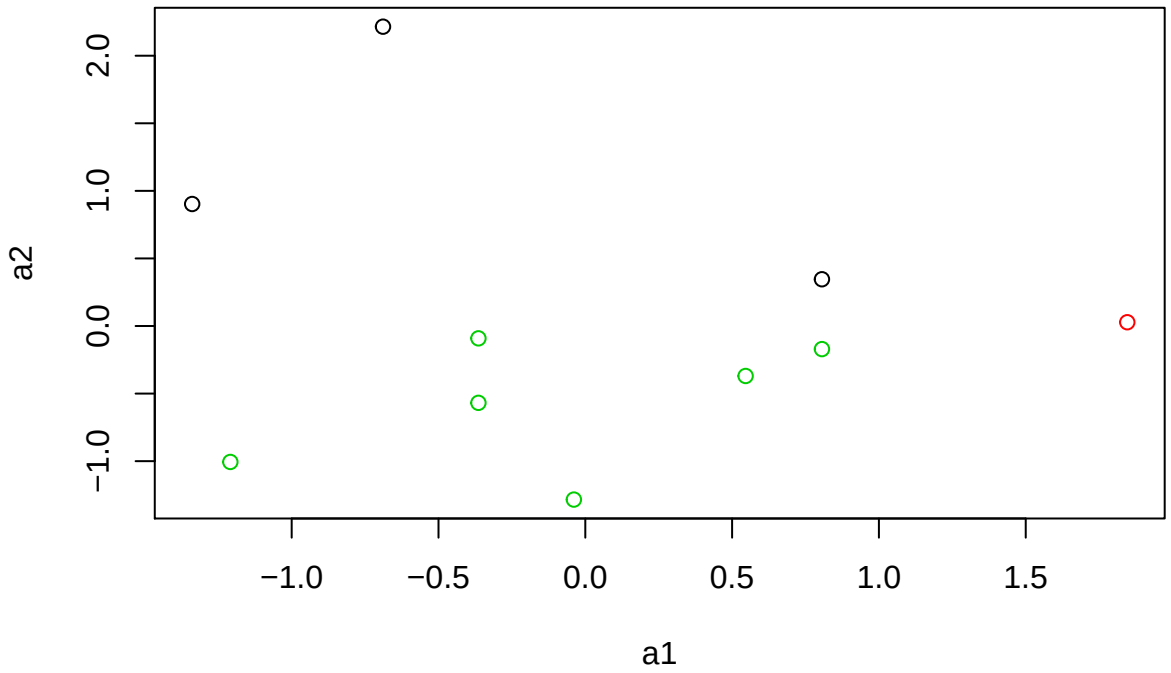
```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
plot(cluster.a.ward)
cluster.a.complete <- hclust(d.a, method="complete")
plot(cluster.a.complete)
```



Ar rezultatai skiriasi? Standartizavę požymius, k-vidurkių metodu suskirstykite apskritis į tris klasterius. Ar rezultatai skiriasi nuo gautų kitais metodais?

```
dat.a.scaled <- scale(dat.a, center = TRUE, scale = TRUE)
cluster.a.kmeans <- kmeans(dat.a.scaled, centers = 3)
plot(dat.a.scaled, col=cluster.a.kmeans$cluster)
```

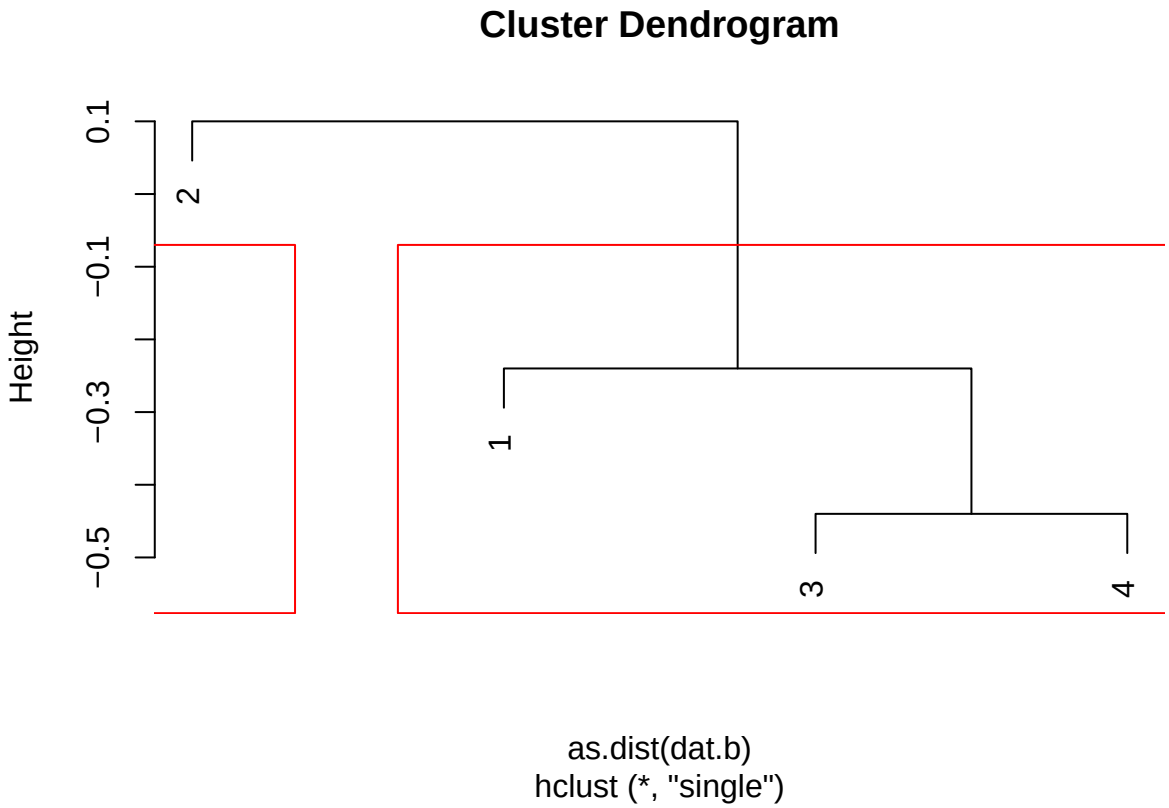


(b) (ČM II.8.5) Turime 4 alaus bendrovių akcijų kainų koreliacijos matricą:

	(1)	(2)	(3)	(4)
(1)	1	0.51	0.39	-0.24
(2)	0.51	1	0.72	0.10
(3)	0.39	0.72	1	-0.44
(4)	-0.24	0.10	-0.44	1

Suskirstykite bendroves į du klasterius.

```
dat.b <- matrix(c(1, 0.51, 0.39, -0.24,
                  0.51, 1, 0.72, 0.10,
                  0.39, 0.72, 1, -0.44,
                  -0.24, 0.10, -0.44, 1),
                nrow = 4, ncol = 4)

cluster.b <- hclust(as.dist(dat.b), method="single")
plot(cluster.b)
rect.hclust(cluster.b, k=2, border="red")
```

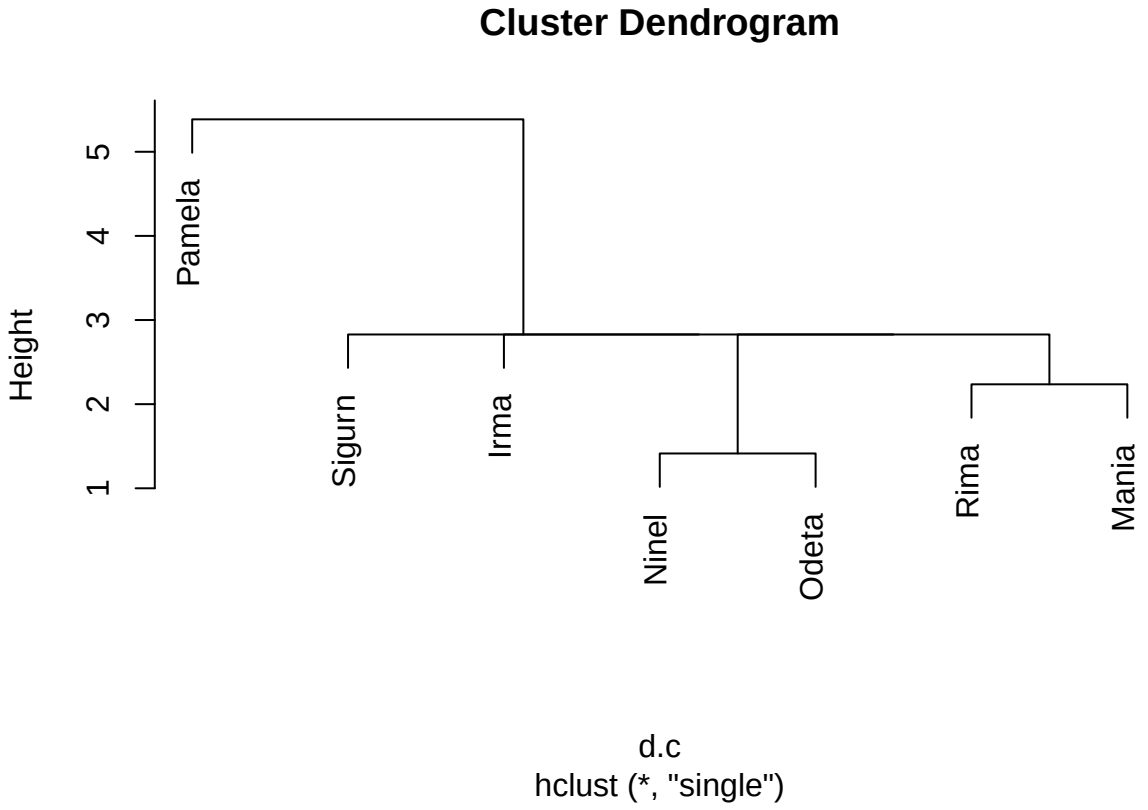


(c) (ČM II.8.6) Kazanova visas damas klasterizuoja pagal krūtinės ir klubų matmenis.

Dama	Krūtinė	Klubai
Rima	100	106
Irma	96	106
Mania	99	108
Ninel	98	104
Odeta	97	103
Pamela	94	111
Sigurn	100	102

Vienetinės jungties metodu pakartokite Kazanovos klasterizavimą.

```
dat.c <- matrix(c(100, 106,  96, 106,  99, 108,  98, 104,
                  97, 103,  94, 111,  100, 102),
                nrow = 7, ncol = 2,
                byrow = TRUE,
                dimnames = list(c("Rima", "Irma", "Mania", "Ninel",
                                  "Odeta", "Pamela", "Sigurn"),
                                c("Krutine", "Klubai")))

dat.c <- as.data.frame(dat.c)

d.c <- dist(dat.c, method = "euclidean")
cluster.c <- hclust(d.c, method="single")
plot(cluster.c)
```



Padaryta su R version 3.4.2 (2017-09-28), x86_64-pc-linux-gnu.