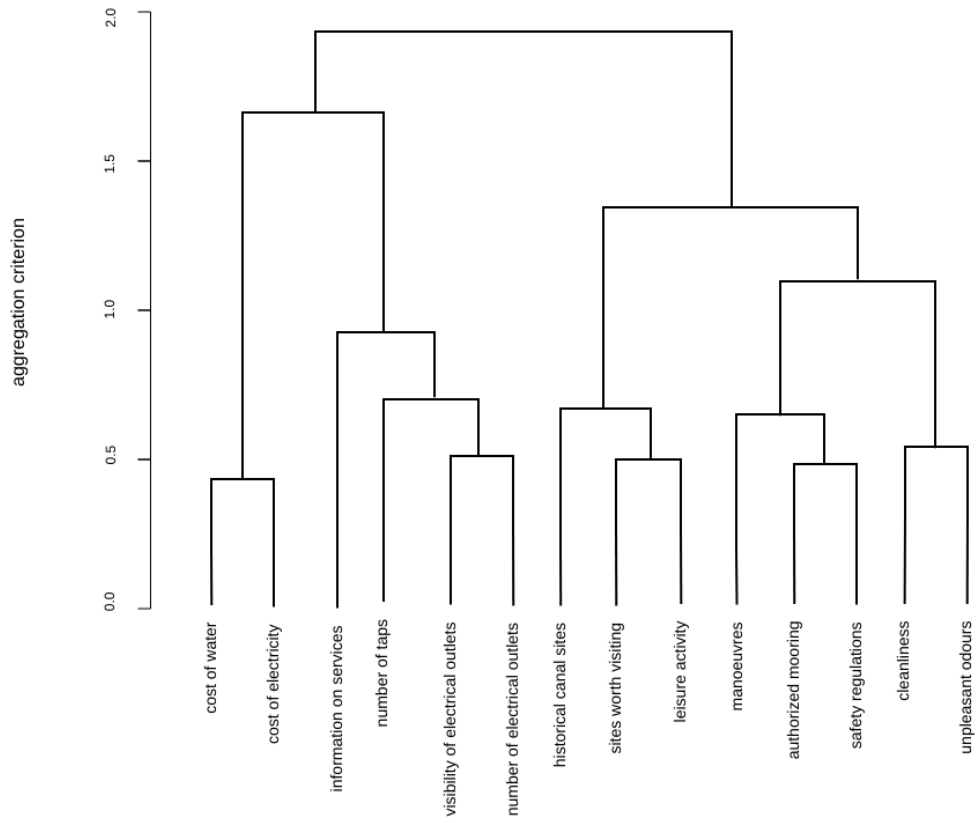
Studijų dalyko **„Duomenų analizė“** klausimai ir uždaviniai egzaminui

Teorijos klausimai iš temos KLASTERIZAVIMAS (Klasterinės analizės teorija.pdf)

1. Kuo skiriasi klasterizacija nuo klasifikavimo?
2. Koks yra klasterinės analizės tikslas?
3. Išvardykite 5 klasterinės analizės etapus;
4. Pagrindiniai 3 klasterinės analizės elementai;
5. Kokios matavimų skalės duomenims skaičiuojami metriniai atstumo matai?
6. Kada objektai panašesni: kai atstumo mato reikšmė didesnė, ar kai mažesnė?
7. Kokias sąlygas turi tenkinti skaitinė neneigiama funkcija $d(X,Y)$, kad ją būtų galima vadinti ***metrika***?
8. Kuo skiriasi hierarchiniai ir nehierarchiniai klasterizavimo metodai?
9. Užrašykite hierarchinių jungimo metodų klasterizavimo schemos 4 žingsnius.
10. Užrašykite nehierarchinio ***k-vidurkių*** metodo 3 pagrindinius klasterizavimo schemos žingsnius.
11. Pagal pateiktą dendrogramą galima stebėti, kaip objektai suskirstomis į ***tris*** klasterius. Surašykite kurie objektai priklauso kiekvienam klasteriui.
    Pvz.:



Teorijos klausimai iš temos STATISTINĖS HIPOTEZĖS (Statistinės hipotezės (2024-04-17)

1. Kas yra statistinė hipotezė? Pateikti apibrėžimą.
2. Ką nurodo nulinė ir alternatyvioji hipotezės? – parašyti paaiškinimą.
3. Kas yra statistinis kriterijus? Parašyti apibrėžimą.
4. Paaiškinti kada padaroma I rūšies klaida, kada II rūšies klaida.
5. Pagal pateiktą uždavinio struktūrą nurodyti, kuris statistinis kriterijus yra taikomas.
6. Ką parodo reikšmingumo lygmuo?
7. Paaiškinkite *p-reikšmės* naudojimą hipotezių išvadų formulavime.
8. Bus pateikti statistinių hipotezių pavyzdžiai, kai taikomas vienas iš statistinių kriterijų (*t* kriterijus, *Chi_kvadrato* kriterijus, *Vilkoksono* kriterijus, *Kruskalo-Voliso* kriterijus, *ANOVA*, *Frydmano* kriterijus) Jus reiks pagal nurodytą p-reikšmę atsakyti kokia yra hipotezės išvada.

Teorijos klausimai iš temos FAKTORINĖ ANALIZĖ (Faktorinė_paskaita.pdf)

1. Apibrėžkite faktorinės analizės uždavinį;
2. Apibrėžkite faktorinės analizės tikslą;
3. Išvardykite faktorinės analizės pagrindinius 4 etapus;
4. Paaiškinkite kokie elementai sudaro faktorinės analizės modelį:

$$X_1 = \lambda_{11}F_1 + \lambda_{12}F_2 + \ldots + \lambda_{1m}F_m + \varepsilon_1,$$
$$X_2 = \lambda_{21}F_1 + \lambda_{22}F_2 + \ldots + \lambda_{2m}F_m + \varepsilon_2,$$
$$\vdots$$
$$X_k = \lambda_{k1}F_1 + \lambda_{k2}F_2 + \ldots + \lambda_{km}F_m + \varepsilon_k.$$

5. Išvardykite 4 faktorinės analizės modelio prielaidas;
6. Nurodykite tris reikalavimus, kuriuos turi tenkinti duomenys, kad jie tiktų faktorinei analizei;
7. Kokią išvadą galima padaryti atlikus ***Bartlett'o sferiškumo*** testą ir gavus žemiau pateiktus rezultatus?

```
> cortest.bartlett(cor(D),n = nrow(D))
$chisq
[1] 207.1133

$p.value
[1] 7.591189e-36

$df
[1] 15
```

8. Kokias išvadas galima padaryti gavus ***Kaiserio-Meyerio-Olkino mato ir MSA_i*** reikšmes pateiktas žemiau:

```
> KMO(D)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = D)
Overall MSA =  0.83
MSA for each item =
 mpg  cyl disp   hp drat   wt
0.85 0.84 0.82 0.83 0.87 0.79
```

9. Užrašykite: a) tiesines daugdaras, kurios vadinamos pagrindinėmis komponentėmis; b) sąlygas, kurias turi tenkinti šios tiesinės daugdaros.

10. Tarkim, pirmoji pagrindinė komponentė yra $Y_1 = \alpha_{11}X_1 + \cdots + \alpha_{1k}X_k$. Kaip apskaičiuojami koeficientai $\alpha_{11}, \alpha_{12}, \ \ldots, \alpha_{1k}$?

11. Kiek procentų bendrosios dispersijos paaiškina pirmoji pagrindinė komponentė?
12. Kiek procentų bendrosios dispersijos paaiškina **dvi pirmosios** pagrindinės komponentės, pagal žemiau pateiktus rezultatus?

```
> pk <- prcomp(X)
> summary(pk)
Importance of components:
```

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.0463	1.0715	0.57737	0.39289	0.3533	0.22799
Proportion of Variance	0.6979	0.1913	0.05556	0.02573	0.0208	0.00866
Cumulative Proportion	0.6979	0.8892	0.94481	0.97054	0.9913	1.00000

13. Taikydami pagrindinių komponenčių metodą skaičiuojame koreliacijų matricą. Kaip koreliacijų matricos tikrinės reikšmės ir tikriniai vektoriai susiję su pagrindinių komponenčių metodu?

Uždaviniai sprendžiami su *Python*

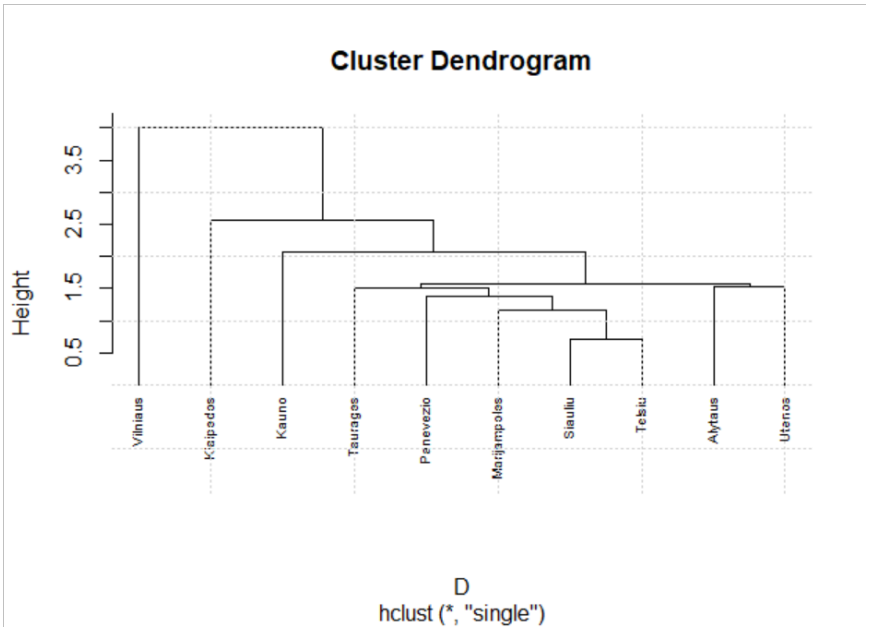
1. Socialiniai bei ekonominiai 10-ties apskričių 1999 metų rodikliai pateikti lentelėje (žemiau) ir duomenų faile ***Duomenys_e.txt***. Kokius apskričių klasterius galima sudaryti pagal šiuos rodiklius? Čia: a_1 – gyventojų aprūpinimas gyvenamuoju plotu (kiek vienam gyventojui vidutiniškai tenka naudingo ploto (m^2)); a_2 – bendrasis vidaus produktas (*BVP*) vienam gyventojui (*tūkst. Lt*); a_3 – materialinės investicijos (*tūkst. Lt*); a_4 – nusikalstamumas (kiek užregistruota nusikaltimų, tenkančių 10000 gyventojų); a_5 – gyventojų aprūpinimas telefonais butuose (100 gyventojų); a_6 – tiesioginės užsienio investicijos (tūkst. Lt, sausio 1 d. duomenys).

Pagal visus požymius klasterizuokite apskritis:

a) Pradžioje standartizuokite kintamuosius;
b) Atstumui skaičiuoti naudokite Euklido atstumo matą.
c) Klasterių jungimui taikykite **vienetinės** jungties (**pilnosios**, **vidutinės**, **centrų**) metodą;
d) Nubrėžkite dendrogramą.
e) Atsakykite į klausimą su kokiomis apskritimis Alytus patenka į vieną klasterį, jeigu klasterių skaičius yra 5?

	Apskritis	a1	a2	a3	a4	a5	a6
1	Alytaus	22.5	9.3	260433	122	24.5	171585
2	Kauno	22.9	11.1	1052822	209	27.3	792675
3	Klaipedos	19.6	12.5	690000	247	28.7	850725
4	Marijampoles	19.8	7.7	146914	164	24.1	20806
5	Panevezio	22.9	9.8	396002	210	23.9	334743
6	Siauliu	21.1	8.8	296607	189	24.3	147408
7	Taurages	21.6	7.0	35921	156	22.4	17674
8	Telsiu	21.1	10.0	503425	171	24.4	115108
9	Utenos	24.5	10.3	268631	140	23.5	91212
10	Vilniaus	20.6	15.8	2502666	236	26.9	3959258

Atsakymai:
Alytaus apskritis patenka į vieną klasterį su Utenos apskritimi



2. Naudodami duomenų rinkinį ***Data1.txt*** patikrinkite ar kintamųjų {*DriversKilled, drivers, front, rear, kms, PetrolPrice, VanKilled*} duomenys tinka faktorinei analizei. Išvadas padarykite pagal:
   a. Bartleto sferiškumo kriterijų;
   b. Kaizerio-Mejerio-Olkino (KMO) matą;
   c. Kiekvieno kintamojo tinkamumo MSA_i matą.
   d. Rasti kovariacijų matricos tikrinius vektorius ir tikrines reikšmes;
   e. Apskaičiuokite pagrindinių komponenčių dispersijas;
   f. Kiek procentų bendrosios dispersijos paaiškina pirmosios dvi pagrindinės komponentės?

Atsakymai:

a) Kadangi $p - reikšmė = 7{,}591189 \times 10^{-36} < 0{,}05$, tai galima teigti, kad stebimi kintamieji tarpusavyje koreliuoja. Tolimesnei analizei reikia pašalinti kintamąjį *„rear“*, nes jis nekoreliuoja su kintamaisiais: *„PetrolPrice“* ir *„VanKilled“*.

b) Pašalinus kintamąjį *„rear“*, KMO = 0,76 – vadinasi stebimi duomenys faktorinei analizei tinka patenkinamai;

c) Kiekvieno kintamojo tinkamumo mato MSA_i reikšmės:

DriversKilled	drivers	front	kms	PetrolPrice	VanKilled
0.73	0.70	0.81	0.72	0.83	0.85

Vadinasi visi stebimi kintamieji tinka faktorinei analizei, nes visos $MSA_i > 0{,}5$.

d) Kovariacijų matricos tikrinės reikšmės yra:

3.54993267 0.90979465 0.71836521 0.47856852 0.25927503 0.08406391

Kovariacijų matricos tikrinių vektorių koordinatės:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	-0.4474766	0.39646082	-0.21921027	-0.2183988	0.49459109	-0.549738094
[2,]	-0.4879876	0.26723031	-0.15935054	-0.1935005	0.06573816	0.789493171
[3,]	-0.4621808	0.24910202	0.05455618	0.1782174	-0.79210255	-0.249344814
[4,]	0.3266051	0.68412420	-0.03160174	0.6291622	0.13017983	0.107296955
[5,]	0.3422249	0.08455782	-0.85691203	-0.2500165	-0.27953304	-0.028051707
[6,]	-0.3524669	-0.48392046	-0.43391357	0.6517244	0.16895615	0.004023482

e) Pagrindinių komponenčių dispersijos yra:

3.54993267 0.90979465 0.71836521 0.47856852 0.25927503 0.08406391

f) Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.8841	0.9538	0.8476	0.69179	0.50919	0.28994
Proportion of Variance	0.5917	0.1516	0.1197	0.07976	0.04321	0.01401
Cumulative Proportion	0.5917	0.7433	0.8630	0.94278	0.98599	1.00000

Pirmosios dvi pagrindinės komponentės paaiškina 74,33 % bendrosios kintamųjų dispersijos.

3.. Tyrimui naudokite duomenų rinkinio ***Data1.txt*** duomenis.

Duomenų rinkinio aprašymas:

`UKDriverDeaths` is a time series giving the monthly totals of car drivers in Great Britain killed or seriously injured Jan 1969 to Dec 1984. Compulsory wearing of seat belts was introduced on 31 Jan 1983.

`DriversKilled` – car drivers killed.
`Drivers` – same as `UKDriverDeaths`.
`Front` – front-seat passengers killed or seriously injured.
`Rear` – rear-seat passengers killed or seriously injured.
`Kms` – distance driven.
`PetrolPrice` – petrol price.
`VanKilled` – number of van ('light goods vehicle') drivers.
`Law` – 0/1: was the law in effect that month?

Naudodami duomenų rinkinio Data1.txt duomenis, patikrinkite žemiau pateiktas hipotezes.

1) Apskaičiuokite koreliacijos koeficientą tarp kintamųjų: `DriversKilled ir drivers`. Ar apskaičiuota koreliac ija statistiškai reikšminga? Atsakymą pagrįskite.

2) Ar galima teigti, jog kintamasis `DriversKilled` turi normalųjį pasiskirstą? Atsakymą pagrįskite.

3) Ar galima teigti, kad kintamųjų `Front` ir `Rear` skirstiniai skirasi? Atsakymą pagrįskite.