# Paskaita

## Apie klasterinę analizę

doc. dr. Rūta Simanavičienė

# Turinys

- Sąvokos;
- Klasterinės analizės tikslai ir etapai;
- Objektų panašumo matai;
- Klasterinės analizės metodų klasifikacija;
- Hierarchiniai metodai – ***jungimo metodai***;
- Nehierarchiniai metodai – ***k-vidurkių metodas***.

# Sąvokos

- **Klasterių (klasterinė) analizė**, arba **klasterizacija** yra objektų suskirstymas į klasterius, pagal jų panašumą.
- **Klasterizacija** yra pagrindinė duomenų išgavimo užduotis ir pagrindinė technika atliekant statistinę duomenų analizę, taip pat yra vartojama daugybėje kitų sričių: mašininiame mokyme, atpažinimo teorijoje, vaizdų analizėje, informacijos paieškoje, bioinformatikoje, duomenų suspaudime, ir kompiuterinėje grafikoje.
- Terminą „**klasteris**" 1939 metais pirmasis pavartojo R. Trajonas (R. Tryon). Klasterinė analizė kartais dar vadinama *taksonomine analize*.
- **Klasifikavimas** (angl. *classification*) yra objektų priskyrimas tam tikroms tikslinėms grupėms (angl. *target groups*), dar kitaip vadinamoms ***klasėmis***.

# Klasterinės analizės tikslai

- Nereiktų tapatinti terminų: grupė ir klasteris. **Klasteris** – panašių objektų grupė (*apie objektų panašumą vėliau*).
- **Klasterinės analizės tikslas** – suskirstyti objektus taip, kad skirtumai klasterių viduje būtų kuo mažesni, o tarp klasterių – kuo didesni.
- Skirstydami objektus į klasterius nežinome kiek klasterių egzistuoja tiriamojoje populiacijoje, vadinasi ***klasterinė analizė*** yra ***egzistuojančių struktūrų paieška.***
- Klasterinės analizės metodo parinkimas ir rezultatų interpretacija **priklauso tik nuo tyrėjo**.

# Klasterinės analizės etapai

**Klasterizuodami turime pereiti 5 etapus:**

- Pasirinkti klasterizuojamus objektus.
- Nuspręsti, pagal kokius požymius klasterizuosime.
- Pasirinkti **kiekybinį matą**, kuriuo matuosime objektų panašumą.
- Vienu ar **kitu metodu** suskirstyti objektus į klasterius.
- Peržiūrėti gautus rezultatus.

# 1 – 2 klasterinės analizės etapai

- Klasterizuojamų objektų ir klasterizavimo požymių parinkimą diktuoja konkretaus tyrimo tikslai bei uždaviniai. Tai - ne statistiko reikalas.
- Pvz.:
  - **Ekonomistas** siekia išskirstyti valstybes pagal jų ekonominius bei demografinius rodiklius;
  - **Psichologas** visus tiriamuosius nori suskirstyti į klasterius pagal jų intelektą, savitvardą bei temperamentą;
  - **Medikas** tiriamus ligonius klasifikuoja pagal įvairius organizmo funkcinius parametrus ir pan.
- Visais atvejais skirstymas į klasterius prasideda tada, kai jau turime *objektų aibę* ir kiekvieną objektą aprašančių *skaitinių* rodiklių aibę.

# 3 – 5 klasterinės analizės etapai

- Kiekybinio **panašumo matų** yra ne vienas.
- Nuo pasirinkto mato priklauso klasterizacijos rezultatai. Tiems patiems duomenims taikydami skirtingus klasterinės analizės metodus, galime gauti skirtingus rezultatus.
- Turėdami kiekybinį panašumo matą, galime pasakyti, kurios objektų poros panašesnės.
- **Klasterizacijos metodas** leidžia nustatyti principus, pagal kuriuos sudaromi klasteriai, ir atsakyti į klausimą, ką reiškia klasterių panašumas.
- Suskirstę objektus į klasterius, dar turime patikrinti, ar gauti rezultatai neprieštarauja sveikam protui.

# Klasterinės analizės elementai

- Tiriamieji objektai žymimi: **X, Y, ..., Z**
- Atlikus objektų matavimus pagal ***m*** požymių, gauta:
- $(x_1, x_2, \dots, x_m)$, $(y_1, y_2, \dots, y_m)$, ..., $(z_1, z_2, \dots, z_m)$ - atitinkamų objektų ***m*** požymių reikšmių vektoriai.

# Kiekybiniai objektų panašumo matai

- **Panašumas** - subjektyvus dalykas. Statistikoje kiek lengviau, nes klasterizuojame atsižvelgdami į **skaičius** (panašumo matų reikšmes).
- Turėdami **kiekybinį panašumo matą**, galime pasakyti, kurios objektų poros **panašesnės**.
- Tačiau ir čia labai daug priklauso nuo ***matuojamų požymių tipo*** (ar požymis tolydus, ar diskretus, ar dvireikšmis), nuo matavimų skalės ir nuo pasirinkto panašumo mato.
- Dažniausiai naudojami panašumo matai:

  1) **metriniai** atstumo matai,

  *2) koreliacijos koeficientai,*

  *3) asociatyvumo koeficientai.*

# Objektų panašumo matai: Metriniai atstumo matai

**Metriniai atstumo matai** naudojami tada, kai objektus charakterizuojantys požymiai matuojami pagal *intervalų* arba *santykių skalę*.

- Šiuos matus tiksliau būtų vadinti skirtingumo matais - kuo didesnė reikšmė, tuo objektai mažiau panašūs.
- Daugelis metrinių atstumų matų yra ***metrikos***. Kas yra metrika ir kodėl svarbu, kad metrinis atstumo matas būtų metrika?

**Metrika** - tai skaitinė neneigiama dviejų objektų **X** ir ***Y*** funkcija ***d(X, Y),*** tenkinanti **sąlygas:**

1) Simetriškumo: $d(X,Y) = d(Y,X)$;
2) Trikampio nelygybės: $d(X,Y) \leq d(X,Z) + d(Y,Z)$;
3) Netapačių objektų atskiriamumo: jei $X \neq Y$, tai $d(X,Y) \neq 0$;
4) Tapačių objektų neatskiriamumo: jei $d(X,Y) = 0$, tai X ir Y identiški.

# Metrikos pavyzdys

**Pavyzdys: Euklido atstumo kvadratas**

Paimkime Euklido atstumą, kuris objektams X ir Y apibrėžiamas taip:

$$d(X,Y) = \|X-Y\| = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \cdots + (x_m-y_m)^2}$$

**Tarkim X, Y ir Z yra trys studentai, kurių amžius ir ūgis atitinkamai yra:**
**(21, 180), (22, 178) ir (23, 179).**

1) $\|X-Y\|^2 = 5$, $\|X-Z\|^2 = 5$ ir $\|Y-Z\|^2 = 2$

– atstumai tarp objektų nelygūs nuliui (3 savybė);

2) $\|X-Y\|^2 = \|Y-X\|^2 = 5$ – (1 savybė);

3) $\|X-Y\|^2 = 5 \leq \|X-Z\|^2 + \|Y-Z\|^2 = 7$ – (2 savybė);

4) Jeigu būtų du studentai, kurių amžius ir ūgis vienodi, turėtume tapačių objektų neatskiriamumą (4 savybė), tada $d(X,Y) = 0$.

# Kiekybinių duomenų atstumo matai

| Atstumas | $d(X, Y)$ formulė |
|---|---|
| Euklido | $\lVert X - Y \rVert = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$ |
| Euklido atstumo kvadratas | $\lVert X - Y \rVert^2 = \sum_{i=1}^{m} (x_i - y_i)^2$ |
| Minkovskio | $\left( \sum_{i=1}^{m} \lvert x_i - y_i \rvert^l \right)^{1/l}, \quad l > 0$ |
| Manheteno (blokinis) | $\sum_{i=1}^{m} \lvert x_i - y_i \rvert$ |
| Čebyšovo | $\max_i \lvert x_i - y_i \rvert$ |
| Vektorių kampo kosinusas | $\sum_{i=1}^{m} (x_i y_i) \left( \sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i^2 \right)^{-1/2}$ |
| Mahalanobio atstumo kvadratas | $(x - y)' V^{-1} (x - y)$, $V$ – požymių reikšmių vektorių kovariacinė matrica |

Tarkim turime taškus **A** ir **B**.

**Euklido** atstumas tarp jų

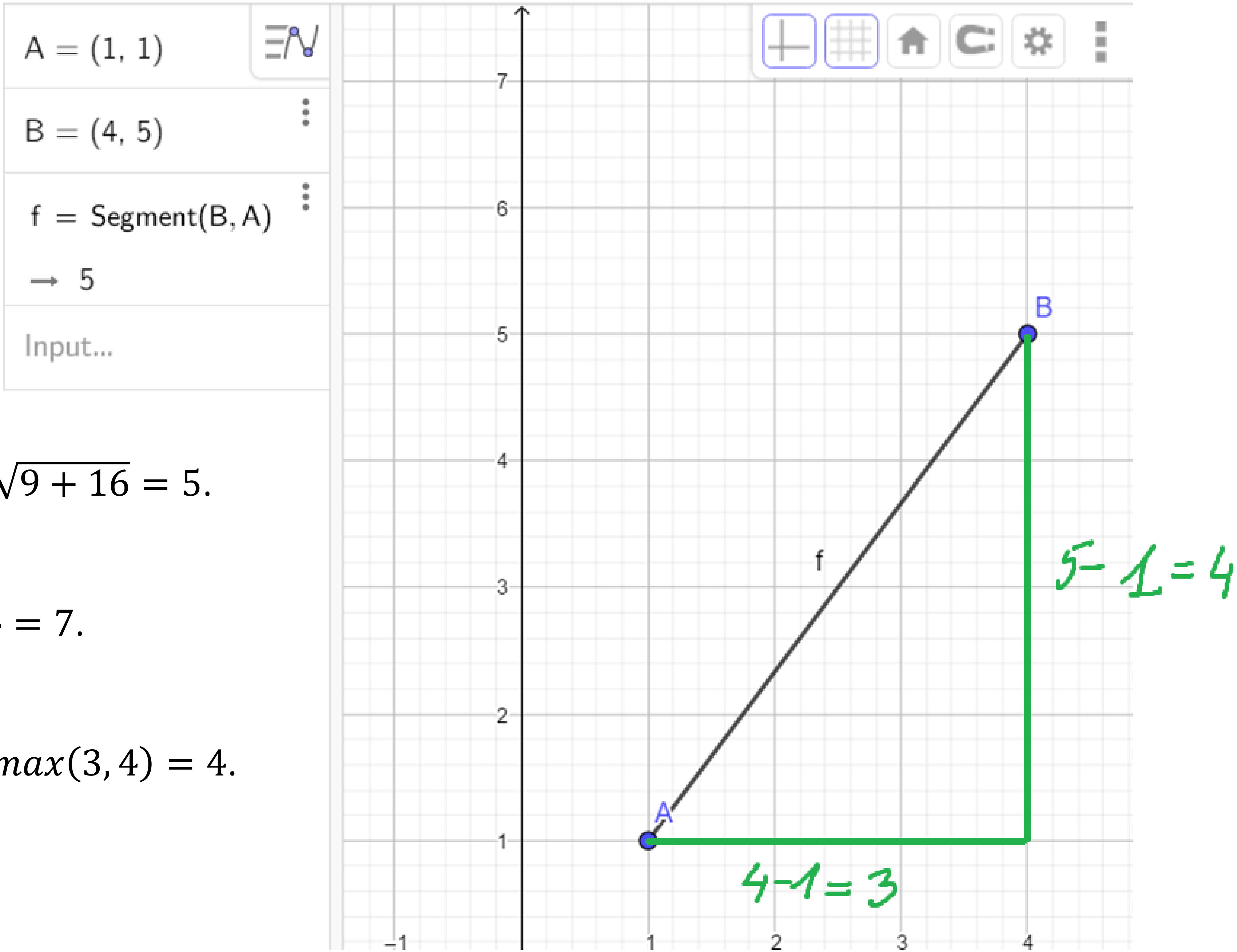
$$d(A, B) = \sqrt{(4-1)^2 + (5-1)^2} = \sqrt{9+16} = 5.$$

**Manheteno** atstumas

$$d(A, B) = |4-1| + |5-1| = 3 + 4 = 7.$$

**Čebyšovo** atstumas

$$d(A, B) = max(|4-1|, |5-1|) = max(3, 4) = 4.$$



# Metrinių atstumo matų trūkumas

- Vienas iš metrinių atstumo matų trūkumų - nevienoda skirtingai matuojamų požymių įtaka. Kintamieji, kurių sklaidos charakteristikos įgyja dideles reikšmes, gali nustelbti mažai įvairuojančių kintamųjų įtaką.
- Tarkime, turime du vektorius **A(1; 0)** ir **B(0; 6)**.

Euklido atstumas tarp jų yra $\sqrt{1^2 + 6^2} = 6{,}083$. Atstumą faktiškai lemia antroji vektorių koordinatė.

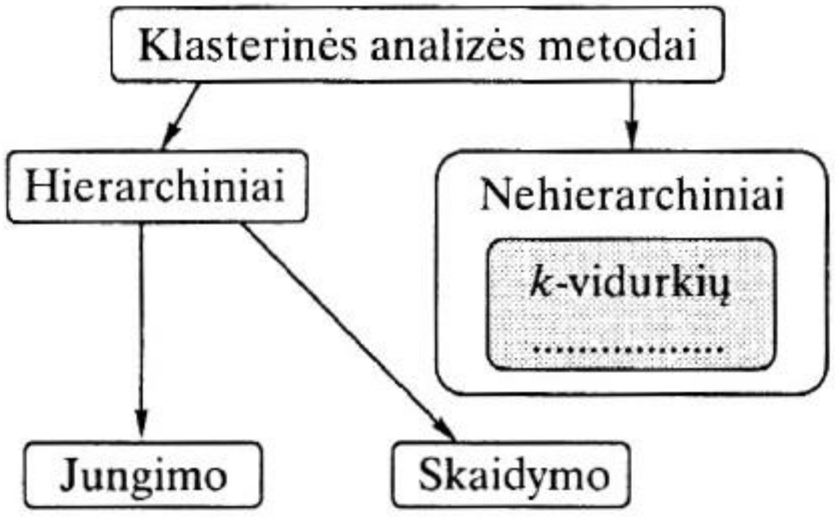
- Vienas iš būdų išvengti šio trūkumo – užuot naudojus kintamųjų reikšmes, imti jų standartizuotąsias reikšmes (*z-reikšmes*), t.y.

$$z_i = \frac{x_i - \bar{x}}{s},$$

Čia $x_i$ - *i-oji* kintamojo **X** reikšmė, $\bar{x}$ - kintamojo **X** reikšmių vidurkis, $s$ - kintamojo **X** reikšmių standartinis nuokrypis. $Z \sim N(0{,}1)$.

# Klasterinės analizės metodų klasifikacija

- ***Klasterinės analizės metodas leidžia nustatyti principus***, pagal kuriuos sudaromi klasteriai, ir atsakyti į klausimą, ką reiškia klasterių panašumas.
- Klasterinės analizės metodų yra daug. Jie skiriami pagal tai, kaip parenkami **panašumo matai**, atstumo tarp klasterių nustatymo **kriterijai** bei kokia skirstymo į klasterius **strategija**.
- Pagrindinės klasterinės analizės metodų klasės pavaizduotos schemoje:



# Hierarchiniai ir nehierarchiniai klasterizavimo metodai

Klasterinės analizės metodai skirstomi į dvi klases – hierarchiniai ir nehierarchiniai metodai.

- **Hierarchinių metodų** rezultatai nusako klasterių tarpusavio hierarchiją, t. y. visi objektai laikomi vienu dideliu klasteriu, kurį sudaro mažesni klasteriai, šiuos savo ruožtu dar mažesni ir t.t.
  - Taikydami hierarchinius metodus, nustatome bendrą visų klasterių tarpusavio priklausomybių struktūrą ir tik po to sprendžiame, koks klasterių skaičius optimalus.
  - Hierarchiniai metodai skirstomi į ***jungimo*** ir ***skaidymo*** metodus.
  - ***Jungimo*** (angl. *agglomerative)* ***metodai*** smulkius klasterius jungia vis į stambesnius, kol galų gale lieka vienas.
  - ***Skaidymo*** (angl. *divisive)* ***metodai*** yra loginė jungimo metodų priešingybė. Vienintelis klasteris nuosekliai skaidomas į dalis.
- **Nehierarchiniai metodai** paprastai taikomi tada, kai iš anksto žinomas (pasirenkamas) klasterių skaičius ir norima tiriamus objektus klasterizuoti.
- Šiame kurse aptarsime vieną hierarchinių metodų klasę - ***jungimo metodus*** bei vieną dažniausiai naudojamų nehierarchinių metodų - ***k-vidurkių metodą***.

# Hierarchinių jungimo metodų strategija

**Bendroji klasterizavimo schema:**

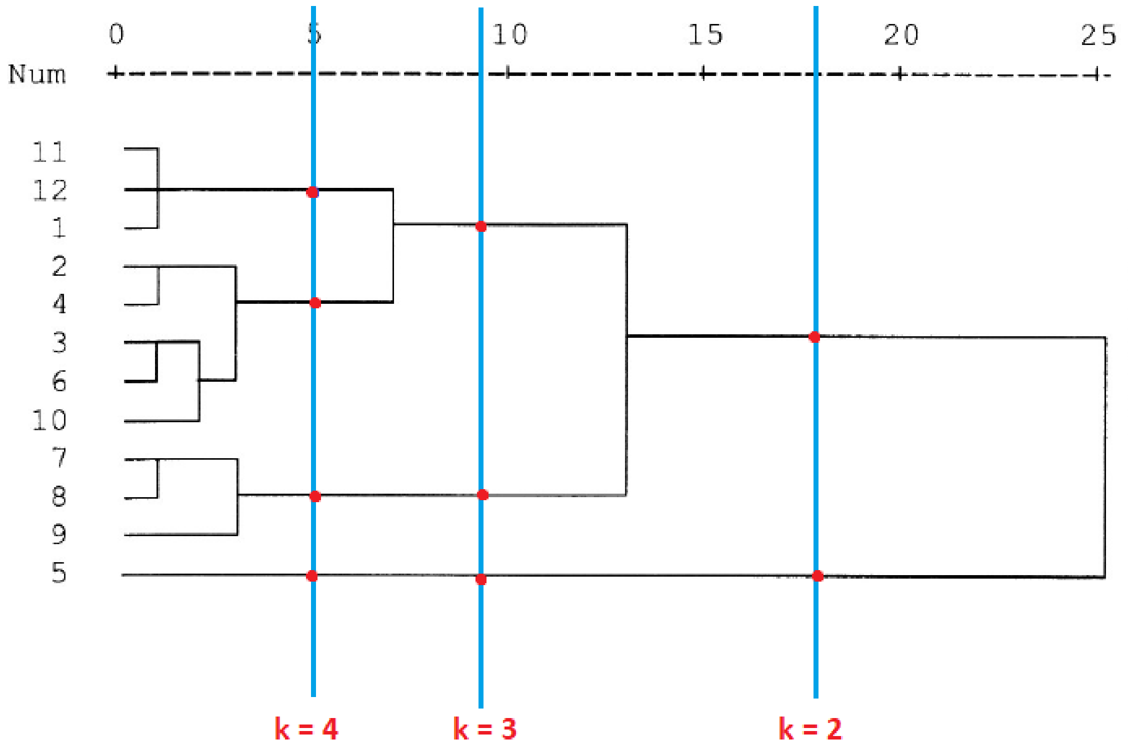
1. Turime N klasterių po 1 objektą ir NxN simetrinę atstumų matricą $(d_{ij}), i, j = \overline{1, N}$.
2. Pagal atstumų matricą nustatome du klasterius, tarp kurių atstumas yra **mažiausias** (kurie yra panašiausi). Tarkime, kad tai klasteriai *U* ir *V.*
3. Sujungiame klasterius *U* ir *V.* Naują klasterį pavadiname (UV). Tada atstumų matricą pakeičiame taip:
   a) išbraukiame stulpelius ir eilutes, atitinkančius klasterius *U* ir *V;*
   b) pridedame eilutę ir stulpelį su **atstumais** tarp *(UV)* ir likusiųjų klasterių.
4. Kartojame 2 ir 3 žingsnius (N — 1) kartų. Procesą baigiame, kai visi objektai yra viename klasteryje.

Šio proceso schema vaizduojama grafiku, vadinamu ***dendrograma*** *(*angl*. dendrogram).*

**Tyrėjas** pats sprendžia (dažniausiai žiūrėdamas į dendrogramą bei klasterizavimo eigos schemą), kuriuo etapu objektų paskirstymas į klasterius yra optimalus.

# Dendrograma

- ***Dendrograma*** yra medžio pavidalo diagrama. Ji yra patogus įrankis norint pavaizduoti klasterių išsidėstymą taikant hierarchinio klasterizavimo metodus.
- Iš dendrogramos matyti kurie klasteriai yra panašiausi ir kada jie buvo jungiami. Vienoje dendrogramos ašyje yra pateikiamos klasės, kitoje – atstumai. Laužtė, jungianti objektus, rodo, koks atstumas tarp klasterių ir kada šie klasteriai buvo sujungti.



# *Klasterių panašumo matai ir jungimo metodai*

- Tarkime, turime du klasterius *U* ir *V*. **Atstumas** $d(X_i, Y_j)$ tarp objektų $X_i \in U \ ir \ Y_j \in V$ matuojamas **vienu** iš kiekybiniams duomenims apibrėžtų **matų**.
- Dažniausiai naudojami **atstumai** $\boldsymbol{d}(\boldsymbol{U}, V)$ tarp dviejų klasterių ***U* ir *V*** pateikti lentelėje.

| Atstumas | $d(U, V)$ formulė |
|---|---|
| Vienetinės jungties (artimiausio kaimyno) | $d(U, V) = \min\limits_{X_i \in U, Y_j \in V} d(X_i, Y_j)$, <br> $X_i$ – $i$-asis $U$ objektas, $Y_j$ – $j$-asis $V$ objektas |
| Pilnosios jungties (tolimiausio kaimyno) | $d(U, V) = \max\limits_{X_i \in U, Y_j \in V} d(X_i, Y_j)$, |
| Vidutinės jungties | $d(U, V) = \sum_{X_i \in U} \sum_{Y_j \in V} d(X_i, Y_j)/(n_U n_V)$, <br> $n_U$, $n_V$ – klasterių objektų skaičius |
| Centrų | $d(U, V) = d(\overline{U}, \overline{V})$, <br> $\overline{U}$, $\overline{V}$ – klasterius sudarančių <br> objektų požymių vektorių vidurkiai |
| Vordo | $d(U, V) = \lVert \overline{U} - \overline{V} \rVert^2/(1/n_U + 1/n_V)$ |

# *Klasterių panašumo matai ir jungimo metodai (angl.)*

| Lietuviškai | Angliškai | Pastabos |
|---|---|---|
| Vienetinės jungties<br>(Artimiausio kaimyno) | Single Linkage<br>(Nearest Neighbor) | $d(A,B) = min\{d(x_i, y_j), x_i \in A \;\&\; y_j \in B\}$ |
| Pilnosios jungties<br>(Tolimiausio kaimyno) | Complete Linkage<br>(Farthest Neighbor) | $d(A,B) = max\{d(x_i, y_j), x_i \in A \;\&\; y_j \in B\}$ |
| Vidutinės jungties | Average Linkage | $d(A,B) = \frac{1}{n_A n_B}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d(x_i, y_j)$ |
| Centrų | Centroid | $d(A,B) = d(\bar{x}_A, \bar{y}_B)$ |
| Medianų | Median | $m_{AB} = \frac{1}{2}(\bar{x}_A + \bar{y}_B)$ |
| Vordo | Ward'method<br>(method = **"ward.D2"**)<br>*Jeigu naudojate Euklido metriką, ne jos kvadratą.*<br>(Murtagh & Legendre, 2014) | $d(A,B) = \frac{\lVert\bar{x}_A - \bar{y}_B\rVert^2}{\frac{1}{n_A} + \frac{1}{n_B}}$ |

# *Jungimo metodų taikymo pavyzdys (1)*

- Kaip pavyzdį paimsime duomenų lentelę iš vadovėlio (Čekanavičius ir Murauskas, 2002).
- Turime automobilių galingumo ir benzino sunaudojimo(100-ui km) duomenis.
- Suskirstykite automobilius į klasterius pagal turimus duomenis.

| | Automobilis | Galingumas | Degalai |
|---|---|---|---|
| 1 | Vreno | 95 | 8 |
| 2 | Saudi | 92 | 8 |
| 3 | Ituzu | 95 | 10 |
| 4 | Delicija | 94 | 6 |
| 5 | Mopel | 93 | 5 |

# *Jungimo metodų taikymo pavyzdys (2)*

- Atstumams tarp automobilių matuoti naudosime **Euklido atstumo kvadratą** – gausime atstumų matricą $(d_{ij}), i, j = \overline{1,5}$.

Tarkime $A_1(95, 8)$; $A_2(92, 8)$, vadinasi $\|A_1 - A_2\|^2 = (95 - 92)^2 + (8 - 8)^2 = 3^2 + 0 = 9$.

- Atstumus tarp klasterių skaičiuosime taikydami **vienetinės jungties metodą.** Kiekvienu klasterizavimo etapu jungiami panašiausia klasteriai, t.y. tie, tarp kurių atstumas mažiausias.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 9 | 4 | 5 | 13 |
| 2 | 9 | 0 | 13 | 8 | 10 |
| 3 | 4 | 13 | 0 | 17 | 29 |
| 4 | 5 | 8 | 17 | 0 | 2 |
| 5 | 13 | 10 | 29 | 2 | 0 |

- Kadangi $\min(d_{ik}) = d_{45} = 2$, tai objektus 4 ir 5 sujungiame į klasterį (45) ir turime sudaryti naują atstumų matricą, taikydami ***vienetinės jungties metodą***, t.y.:

$$d_{(45)1} = \min(d_{41}, d_{51}) = \min(5, 13) = 5 \text{ ir t.t.}$$

# *Jungimo metodų taikymo pavyzdys (3)*

- Kartojant bendrąją klasterizavimo schemą, galiausiai gauname du klasterius, kuriuos galime apjungti į vieną.

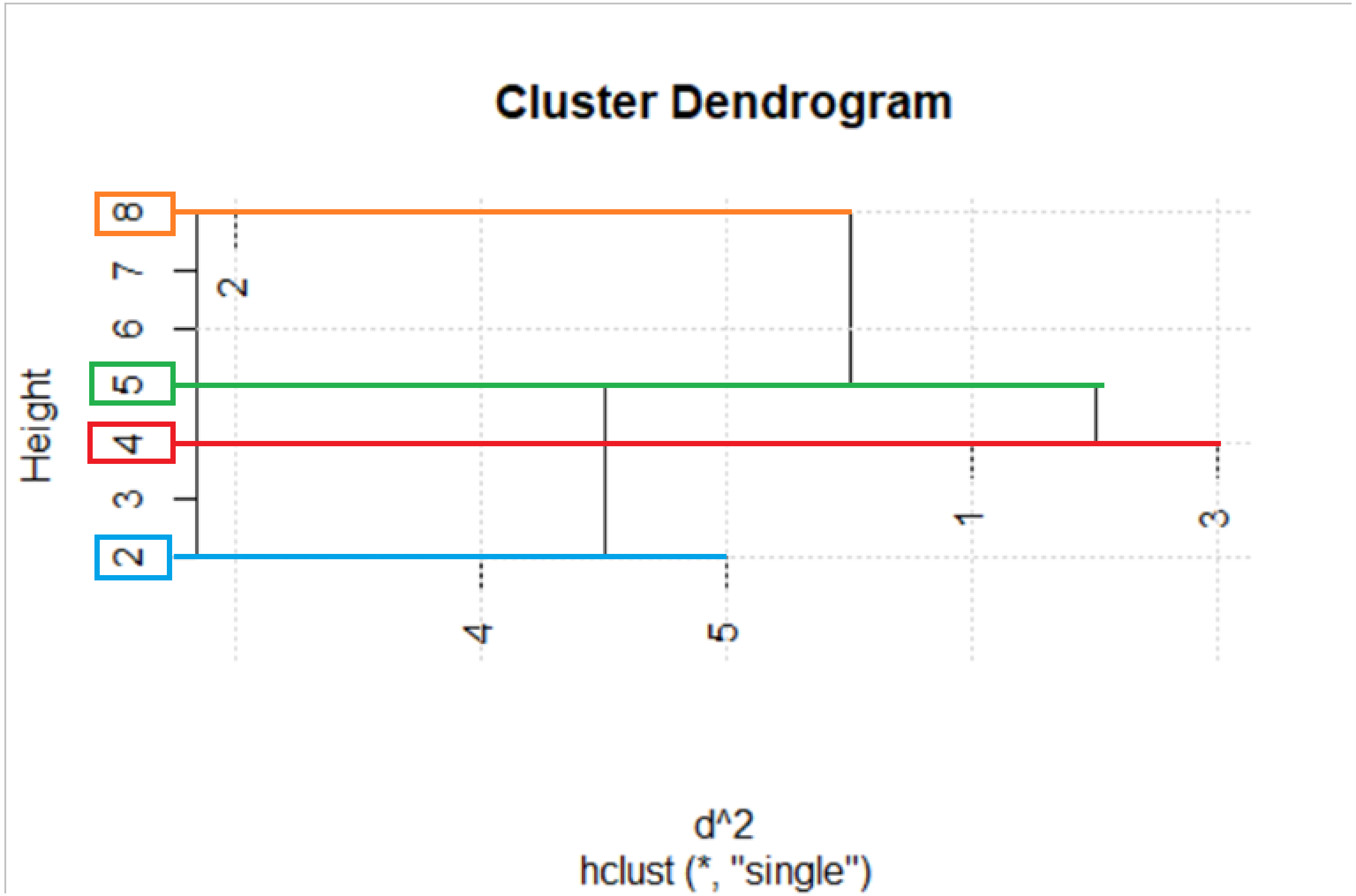
| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 9 | 4 | 5 | 13 |
| 2 | 9 | 0 | 13 | 8 | 10 |
| 3 | 4 | 13 | 0 | 17 | 29 |
| 4 | 5 | 8 | 17 | 0 | 2 |
| 5 | 13 | 10 | 29 | 2 | 0 |

| | 1 | 2 | 3 | (45) |
|---|---|---|---|---|
| 1 | 0 | 9 | 4 | 5 |
| 2 | 9 | 0 | 13 | 8 |
| 3 | 4 | 13 | 0 | 17 |
| (45) | 5 | 8 | 17 | 0 |

| | (13) | 2 | (45) |
|---|---|---|---|
| (13) | 0 | 9 | 5 |
| 2 | 9 | 0 | 8 |
| (45) | 5 | 8 | 0 |

| | (1345) | 2 |
|---|---|---|
| (1345) | 0 | 8 |
| 2 | 8 | 0 |

# Klasterizavimo dendrograma



# Hierarchinių klasterizavimo metodų trūkumai

Skaičiavimams naudojama atstumų matrica. Jeigu turime > 300 objektų, tuomet gausime atstumų matricą iš daugiau nei 90000 elementų.

Hierarchinis klasterizavimas neatsako į klausimą kiek yra klasterių.

Jeigu žinome kiek klasterių turime gauti, tuomet geriau taikyti nehierarchinius klasterizavimo metodus.

Nagrinėsime nehierarchinį klasterizavimo metodą – *k*-vidurkių metodą.

# *k*-vidurkių metodas

***k*-vidurkių** klasterizavimo procedūrą sudaro 3 žingsniai:

1. Objektai suskirstomi į ***k*** pradinių klasterių;
2. Paeiliui apskaičiuojamas kiekvieno objekto **atstumas** iki klasterių **centrų** (atstumas paprastai skaičiuojamas naudojantis Euklido metrika arba jos kvadratu). Objektas skiriamas į artimiausią klasterį. Klasterių centrai perskaičiuojami.
3. 2 žingsnis kartojamas tol, kol perskirstymų daugiau nėra.

**Vienas iš *k-vidurkių* metodų trūkumų** – klasterių skaičių reikia nustatyti iš anksto. Yra keletas argumentų, prieštaraujančių išankstiniam klasterių skaičiaus nustatymui:

- Pasirinktieji klasterių centrai yra iš vieno klasterio ir gauti klasteriai mažai skiriasi.
- Net jei iš tiesų žinoma, kad objektų populiacijoje yra ***k*** klasterių, tiriamoje objektų imtyje gali nepasitaikyti atstovų iš ***k**-ojo* klasterio.
- Objektas, kurio požymių reikšmių vektorius yra iš išskirčių, gali sudaryti atskirą klasterį.

**Klasterinės analizės tikslas** - egzistuojančių struktūrų paieška, tačiau, nurodant pradinį klasterių skaičių, struktūra yra primetama.

# *k*-vidurkių metodo taikymo pavyzdys (1)

- Atsitiktinai parinkti 4 piliečiai įvertino savo materialinę padėtį ir šalies ekonominę situaciją skalėje nuo -5 (labai blogai) iki +5 (labai gerai).
- Suskirstykite 4 respondentus į 2 klasterius: *pesimistų* ir *optimistų*.

| Respondentas | Materialinė padėtis ($X$) | Šalies situacija ($Y$) |
|---|---|---|
| $A$ | 5 | 3 |
| $B$ | −1 | 1 |
| $C$ | 0 | −3 |
| $D$ | −2 | −1 |

**Sprendimas**:

- Savo nuožiūra objektus suskirstome į 2 klasterius (AB) ir (CD) ir apskaičiuojame šių klasterių centrus:

| Klasteris | Centras $X$ | Centras $Y$ |
|---|---|---|
| $(AB)$ | $(5+(-1))/2=2$ | $(3+1)/2=2$ |
| $(CD)$ | $(0+(-2))/2=-1$ | $((-3)+(-1))/2=-2$ |

# *k*-vidurkių metodo taikymo pavyzdys (2)

***Pirmoji iteracija (žingsnis).*** Apskaičiuojame respondentų atsakymų vektorių atstumus nuo klasterių centrų ir priskiriame respondentus artimiausiam klasteriui. Naudosime **Euklido atstumo kvadratą:**

$\|A - \overline{AB}\|^2 = (5-2)^2 + (3-2)^2 = 10;$

$\|B - \overline{AB}\|^2 = (-1-2)^2 + (1-2)^2 = 10;$

$\|C - \overline{AB}\|^2 = 41;$

$\|D - \overline{AB}\|^2 = 25;$

$\|A - \overline{CD}\|^2 = 61;$

$\|\boldsymbol{B} - \overline{\boldsymbol{CD}}\|^{\mathbf{2}} = \mathbf{9};$

$\|C - \overline{CD}\|^2 = 2;$

$\|D - \overline{CD}\|^2 = 2.$

***B*** respondentas yra artimesnis klasteriui *(CD)* nei *(AB),* todėl jį priskiriame **klasteriui *(CD).***

# $k$-vidurkių metodo taikymo pavyzdys (3)

Sudaromi nauji klasteriai ir apskaičiuojami jų centrai:

$\overline{BCD}_X = \frac{-1+0-2}{3} = -1;$

$\overline{BCD}_Y = \frac{1-3-1}{3} = -1.$

| Klasteris | Centras $X$ | Centras $Y$ |
|---|---|---|
| $(A)$ | 5 | 3 |
| $(BCD)$ | $-1$ | $-1$ |

***Antroji iteracija (žingsnis).*** Apskaičiuojame respondentų atsakymų vektorių atstumus nuo klasterių centrų ir priskiriame respondentus artimiausiam klasteriui. Naudosime Euklido atstumo kvadratą:

$\|A - \bar{A}\|^2 = (5-5)^2 + (3-3)^2 = \mathbf{0;}$

$\|B - \bar{A}\|^2 = (-1-5)^2 + (1-3)^2 = 40;$

$\|C - \bar{A}\|^2 = 61;$

$\|D - \bar{A}\|^2 = 65;$

$\|A - \overline{BCD}\|^2 = \left(5-(-1)\right)^2 + (3-(-1))^2 = 52;$

$\|B - \overline{BCD}\|^2 = \left(-1-(-1)\right)^2 + (1-(-1))^2 = 0;$

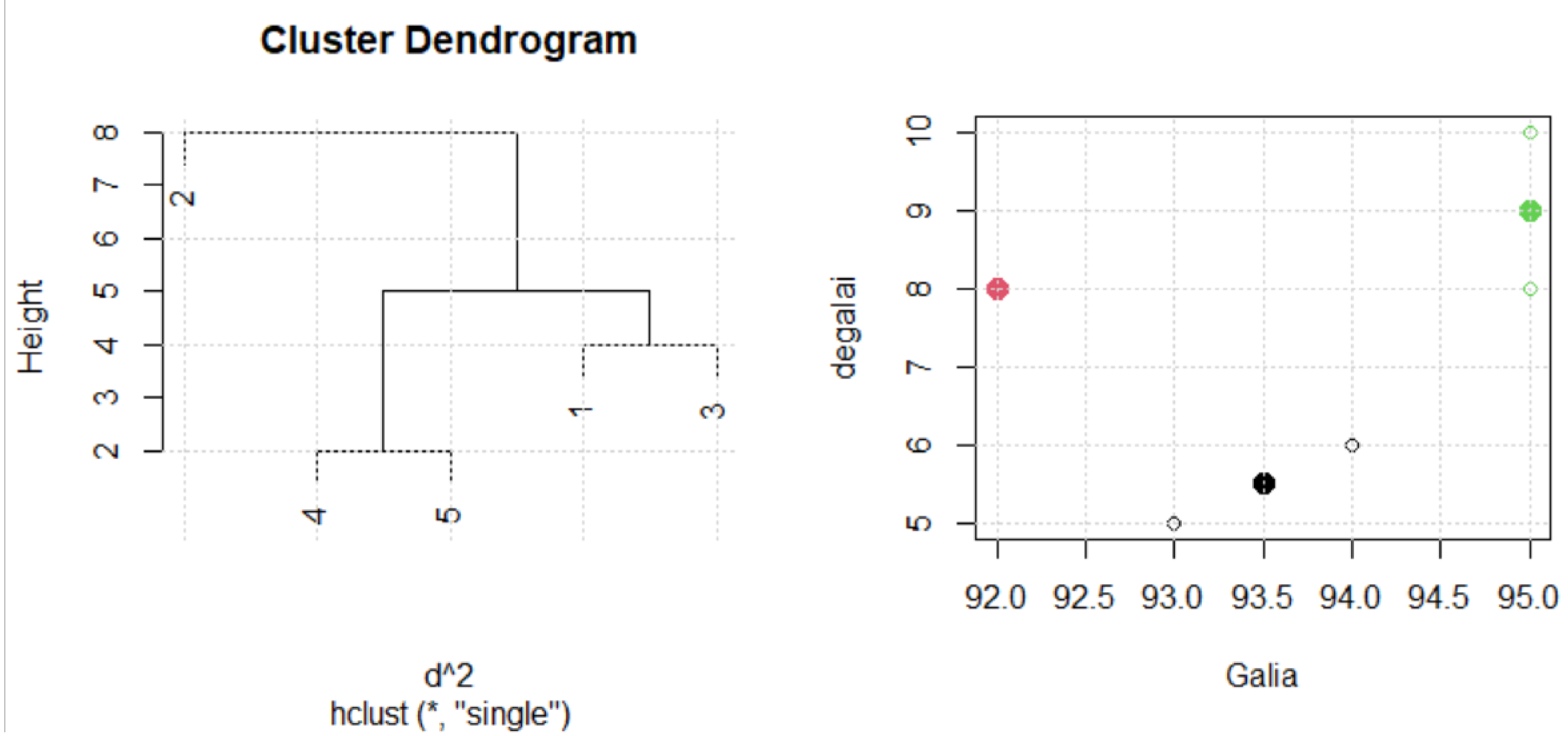
$\|C - \overline{BCD}\|^2 = 5;$

$\|D - \overline{BCD}\|^2 = 1.$

Kiekvieno objekto atstumas iki “savo”klasterio centro yra mažiausias. Turime 1 **optimistą** ir 3 **pesimistus**.

| | Automobilis | Galingumas | Degalai |
|---|---|---|---|
| 1 | Vreno | 95 | 8 |
| 2 | Saudi | 92 | 8 |
| 3 | Ituzu | 95 | 10 |
| 4 | Delicija | 94 | 6 |
| 5 | Mopel | 93 | 5 |

# Heirarchinio ir nehierarchinio klasterizavimo rezultatų vizualizacija



# Ką reikia turėti omenyje taikant klasterinės analizės metodus

1) Klasterinėje analizėje yra daug euristinių, neturinčių teorinio pagrindimo, metodų. Viena iš problemų - dažnai nėra aišku, ar klasterizuojamų objektų aibė yra populiacija, ar populiacijos dalis (imtis). Taigi kyla sunkumų vertinant imties reprezentatyvumą, rezultatų statistinį reikšmingumą ir pan.

2) Klasterinės analizės metodai buvo konstruojami įvairioms sritims, todėl juose yra nemažai specifiškumų.

3) Tiems patiems duomenims taikydami **skirtingus klasterinės analizės metodus**, galime **gauti skirtingus rezultatus**.

4) Neturint išankstinės informacijos apie nagrinėjamų duomenų struktūras, gautus rezultatus lyginti sunku. Objektų klasterizavimui *rekomenduojama* taikyti keletą klasterizavimo metodų, tuomet galime tikėtis patikimesnių rezultatų.

# R komandos klasterizavimui

| Komandos | Aprašymai |
| --- | --- |
| dist(X, method = "euclidean",...) | Sukuriama atstumų matrica, kurios elementai nurodo atstumus tarp matricos X eilučių, taikant atstumo skaičiavimo metodą:<br><br>"euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski". |
| which.max(x) | Nurodo vektoriaus x didžiausio elemento indeksą. |
| which(sąlyga(X), ....) | Išveda masyvo X elemento, kuris tenkina masyvo X elementams apibrėžtą sąlygą, indeksus. |
| hclust(d, method = "complete") | Objektus suskirsto į klasterius naudojant objektų atstumų matricą d, klasterių jungimo metodai: **ward.D2**; **single**; **complete**; **average**; **centroid**. |
| plot(x) | Brėžiama hierarchinės klasterinės analizės dendrograma, kai x – rezultatas gautas įvykdžius ***hclust()*** |
| kmeans(X,k) | Atliekamas objektų, kurių duomenys pateikti matricoje X suskirstymas į k klasterių. |

# Papildomos R komandos

| Komandos | Aprašymai |
|---|---|
| cutree(Hcl,3) | Nurodo, kiekvieno imties objekto, klasterio numerį, atlikus hierarchinį klasterizavimą. |
| rect.hclust(fit, k=3, border="red") | Ant dendrogramos nubrėžiami stačiakampiai žymintys klasterius. |
| subset(dat,weight>=200&agegp=="young") | Išveda objektus priskirtus poaibiui, kurio elementai tenkina duotas sąlygas. |
| plot(Hcl, cex = 0.6, hang = -1); grid() | Dendrogramos braižymas |
| **sub_grp <- cutree(Hcl, k = 3)**<br>**table(sub_grp)** | Suskaičiuojama kiek objektų yra kiekviename klasteryje. |
| | |

# Naudota literatūra


- Čekanavičius, V. & Murauskas. G. (2002). *Statistika ir jos taikymai, II dalis. Vilnius: TEV*.
- Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, **31**(3), 274-295.