Hijerarhijska klasterizacija

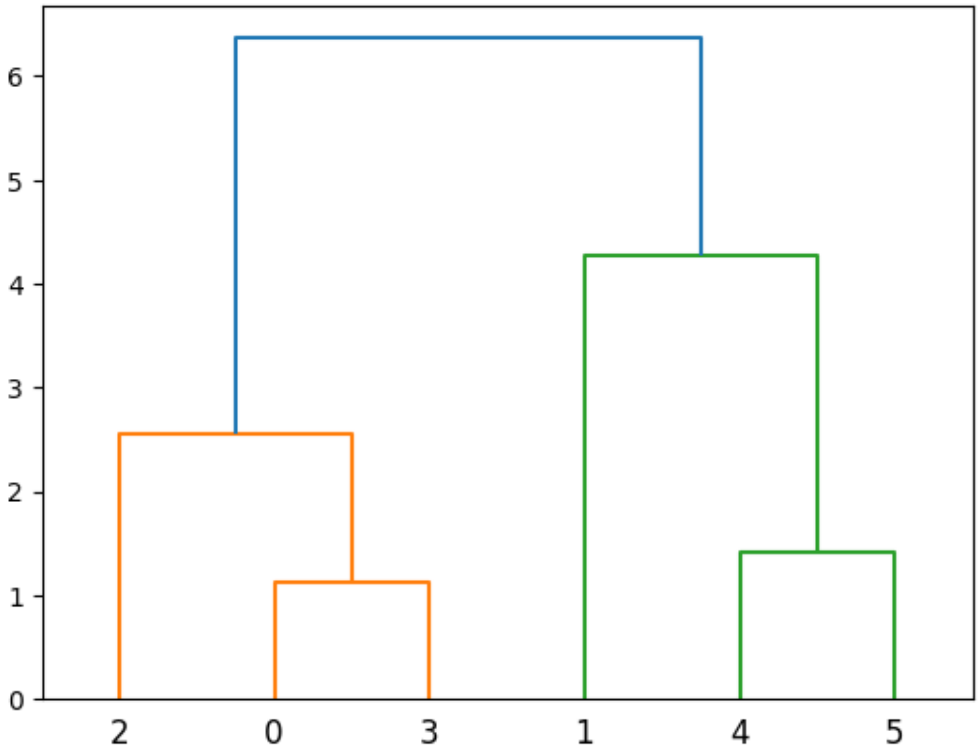
Bojan Tomić

Hijerarhijska klasterizacija

- Metoda klasterizacije u kojoj se pravi hijerarhija klastera
  - Veći klasteri sadrže sve manje i manje klastere dok se ne dođe do klastera koji sadrže samo po jednu instancu
- Dva pristupa
  - Bottom-up, aglomeracija (eng. Agglomerative) - češći
    - U početku, svaka instanca je poseban klaster
    - Redom se najbliži klasteri grupišu u veće klastere i td. dok se ne dobije jedan klaster
  - Top-down, deljenje (eng. Divisive) - ređi

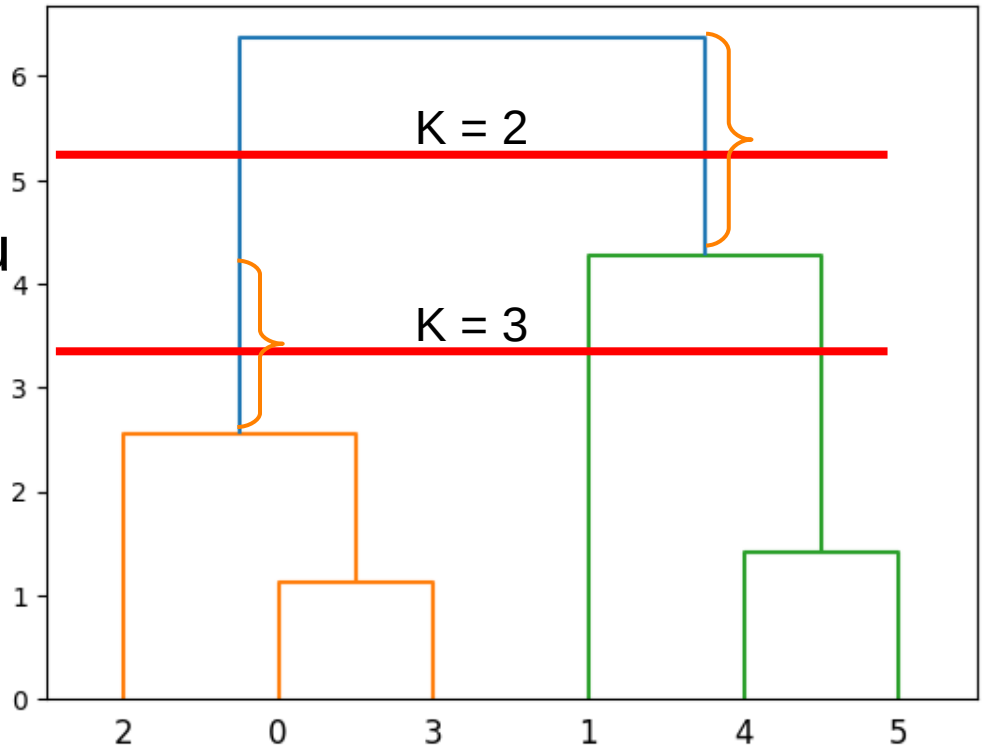
Hijerarhijska klasterizacija

- Dendrogram (hijerarhija klastera)
- X osa – instance klastera
- Y osa - „razmak“ između klastera
- Linije – grupisanje po klasterima
- Bottom-up, top-down



Hijerarhijska klasterizacija

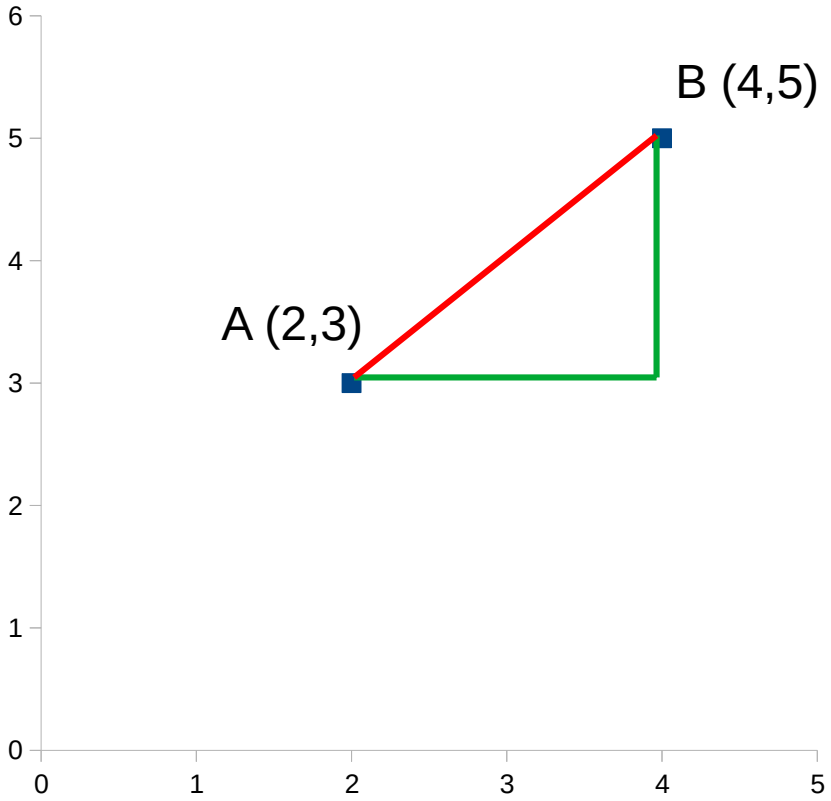
- „Dobar“ broj klastera
  - Presek horizontalnom linijom gde je najveći razmak po Y osi između horizontalnih linija (vitičaste zagrade)
  - Ovde je moguće 2 ili 3 klastera, sa tim da se 2 klastera čine kao bolje rešenje (veći razmak)
  - Boje linija (dva klastera)
  - Subjektivna procena



Hijerarhijska klasterizacija

- Dva ulazna parametra
- Metoda za izračunavanje udaljenosti
  - Euklidska, Menhetn (city block), maksimalna...
- Metoda za povezivanje klastera (u veće klastere) tzv. „linkage“ metode
  - Single linkage (najmanja udaljenost između kl.)
  - Complete linkage (najveća udaljenost)
  - Average linkage (prosečna udaljenost)
  - Ward metoda (min within cluster sum of squares)...

Hijerarhijska klasterizacija



Tačke A i B

- Menhetn udaljenost (city block)
  $d_m = |x_A-x_B| + |y_A-y_B| = |2-4| + |3-5| = 4$

- Euklidska udaljenost
  $d_e^2 = (x_A-x_B)^2 + (y_A-y_B)^2 = (2-4)^2+(3-5)^2 = 8$
  $d_e = 2,8284$

- Maksimalna udaljenost
  $d_{max} = \max(|x_A-x_B|, |y_A-y_B|) = \max(2,2)$
  $d_{max} = 2$

Hijerarhijska klasterizacija

- Povezivanje klastera
- Dva klastera koja su najbliža međusobno se spajaju u veći klaster
- Šta znači najbliži?
- Sa najmanjom udaljenošću instanci jednog prema instancama drugog klastera
  - Single linkage
  - Complete linkage
  - Average linkage

Hijerarhijska klasterizacija

- Single linkeage
- Razdaljina između dva klastera = razdaljina između dve najbliže instance tih klastera

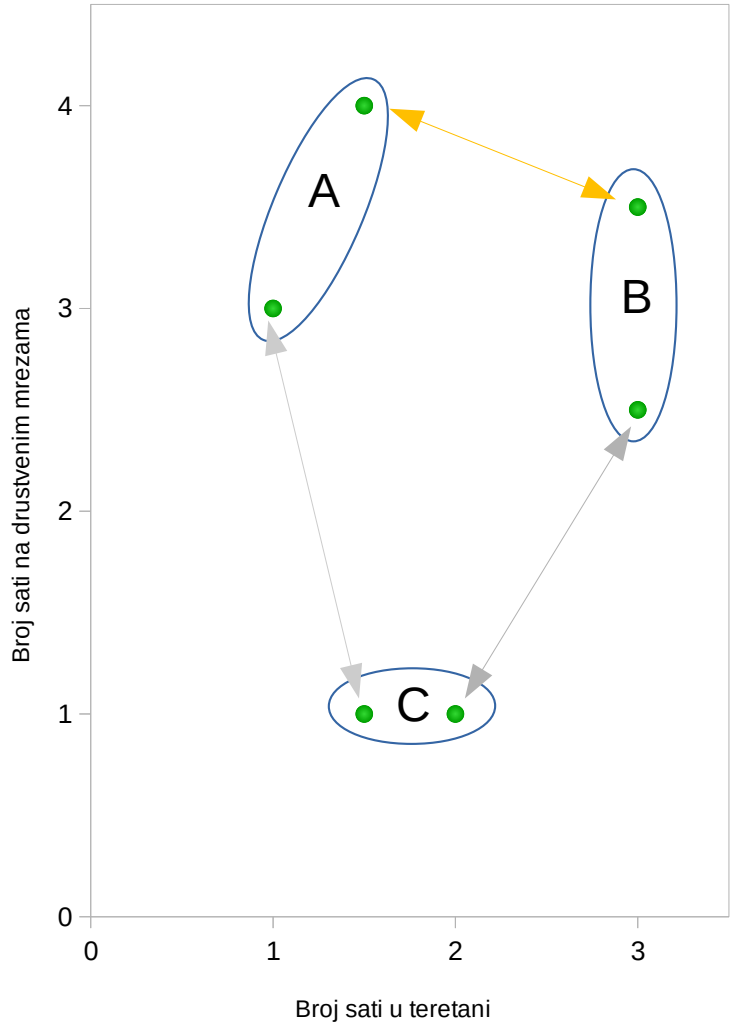
Minimalna udaljenost najbližih instanci

d_{AB} = d ((1.5, 4), (3, 3.5)) = 1.58

d_{AC} = d ((1, 3), (1.5, 1)) = 2.06

d_{BC} = d ((3, 2.5),(2, 1)) = 1.80

Zaključak: klasteri A i B su najbliži i treba ih spojiti u jedan klaster



Hijerarhijska klasterizacija

- Complete linkeage
- Razdaljina između dva klastera = razdaljina između dve najdalje instance tih klastera

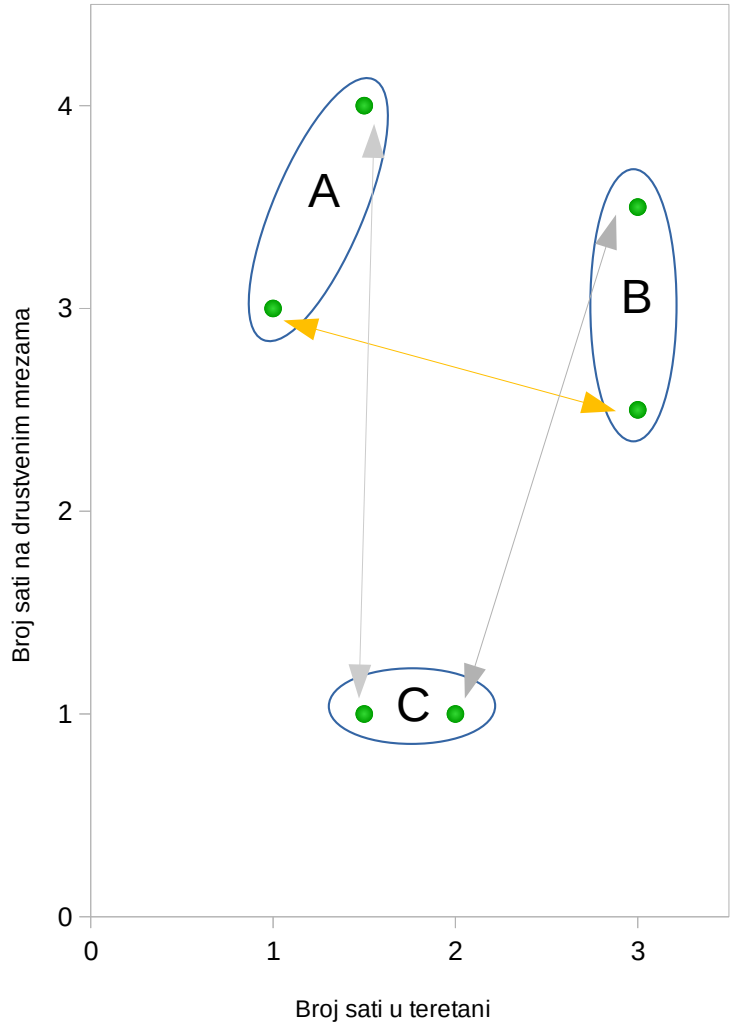
Minimalna udaljenost najdaljih instanci

d_{AB} = d ((1, 3), (3, 2.5)) = 2.06

d_{AC} = d ((1.5, 4), (1.5, 1)) = 3

d_{BC} = d ((3, 3.5),(2, 1)) = 2.69

Zaključak: klasteri A i B su najbliži i treba ih spojiti u jedan klaster



Hijerarhijska klasterizacija

- Average linkeage
- Razdaljina između dva klastera = prosečna razdaljina između svih kombinacija parova instanci tih klastera

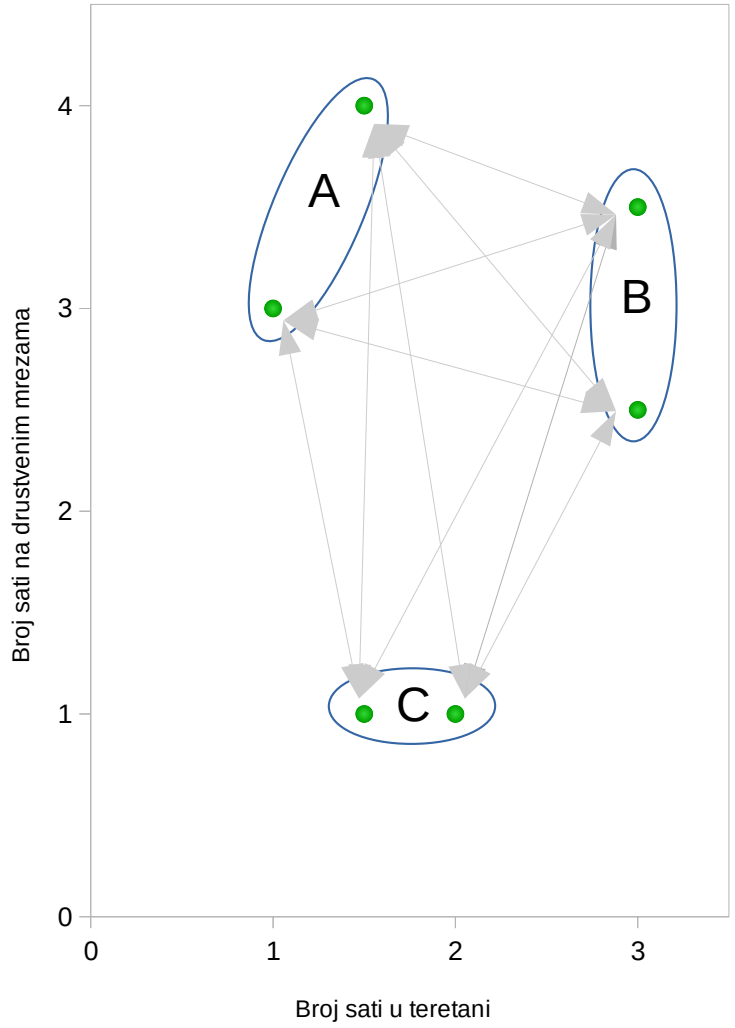
Minimalna prosečna udaljenost svih parova instanci

d_{AB} = (d((1.5, 4),(3, 3.5)) + d((1.5, 4),(3, 2.5)) + d((1, 3),(3, 3.5)) + d((1, 3),(3, 2.5))) / 4

d_{AC} = (d((1.5, 4),(1.5, 1)) + d((1.5, 4),(2, 1)) + d((1, 3),(1.5, 1)) + d((1, 3),(2, 1))) / 4

d_{BC} = (d((3, 3.5),(1.5, 1)) + d((3, 3.5), (2, 1)) + d(((3, 2.5),(1.5, 1)) + d((3, 2.5), (2, 1))) / 4

Zaključak: klasteri A i B su najbliži i treba ih spojiti u jedan klaster

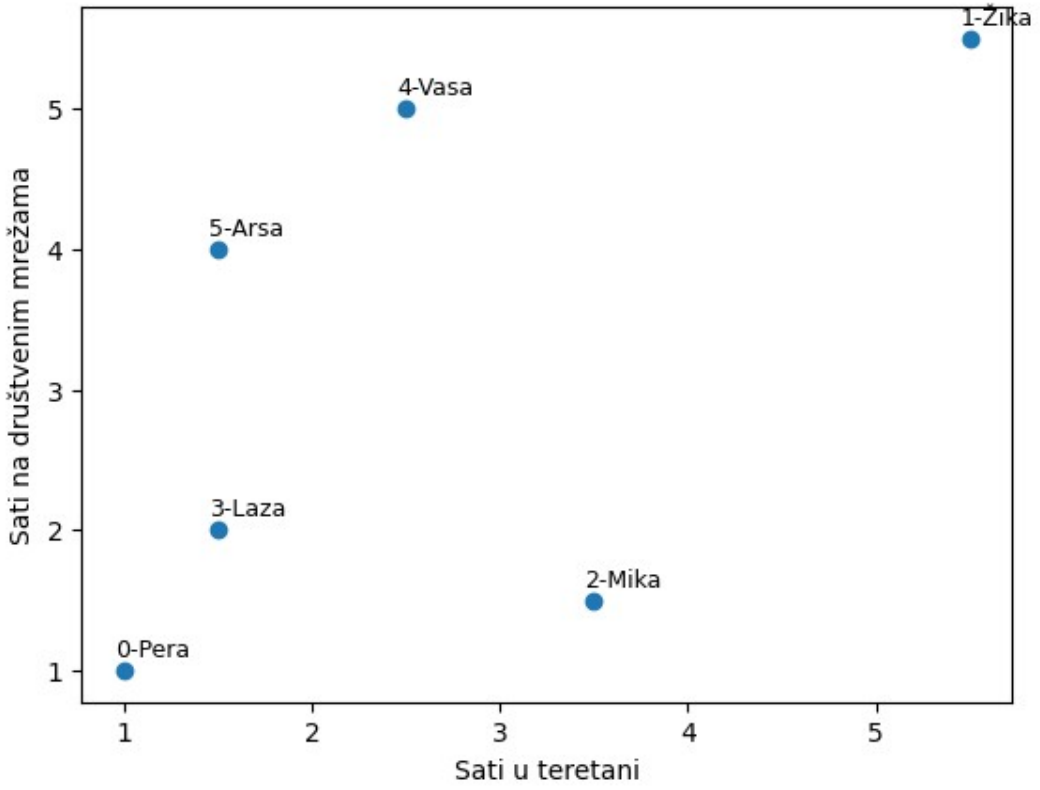


Hijerarhijska klasterizacija – algoritam (bottom-up)

1. Početak - Svaka instanca je poseban klaster
2. Izračunava se udaljenost između svaka dva klastera (matrica udaljenosti – distance matrix)
3. Povezuju se (spajaju) najbliža dva klastera u jedan
4. Ponavljaju se koraci 2 i 3 dok se svi klasteri ne spoje u jedan koji sadrži sve instance

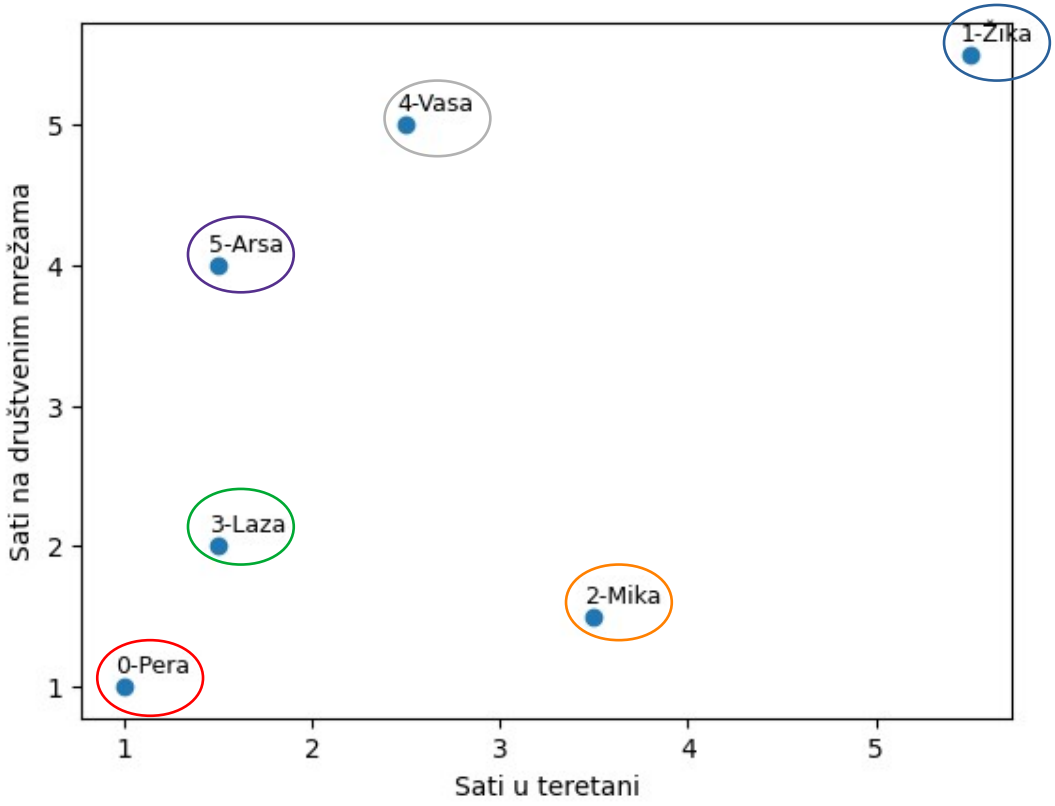
Primer 1

- Podaci za šest osoba
- Prosečan nedeljni broj sati u teretani
- Prosečan dnevni broj sati na društv. mrežama
- Parametri algoritma
  - Euklidska udaljenost
  - Average linkage



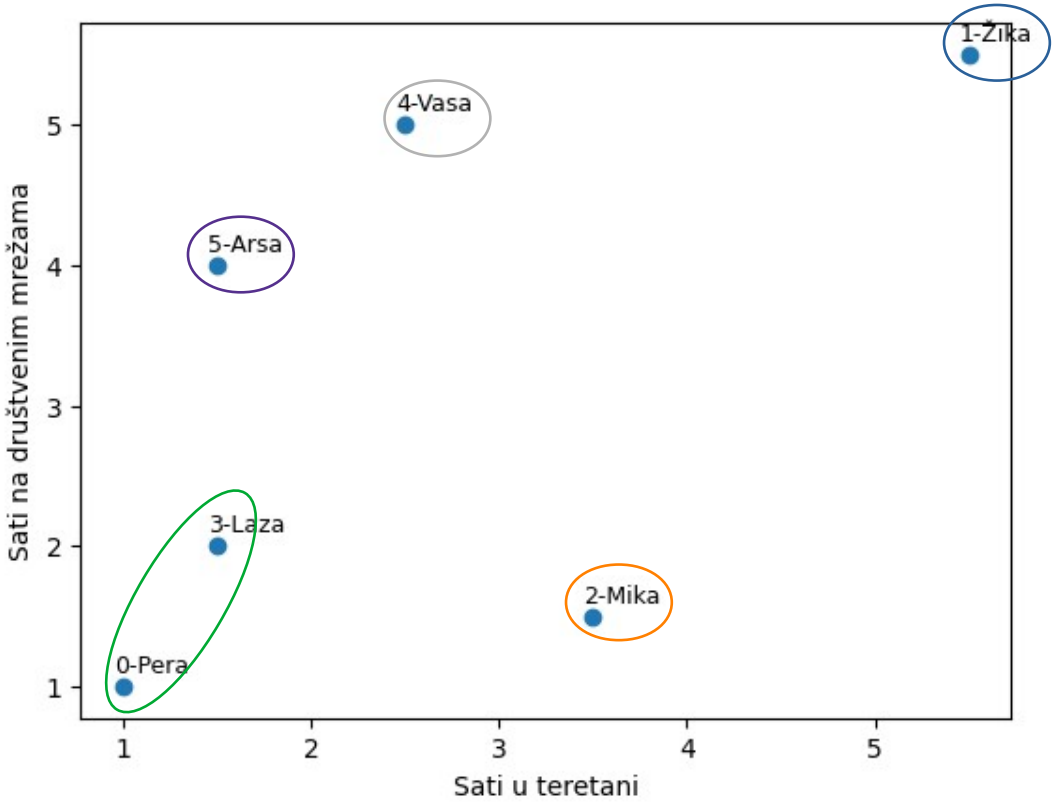
Primer 1

- Početak
  - Svaka instanca je poseban klaster



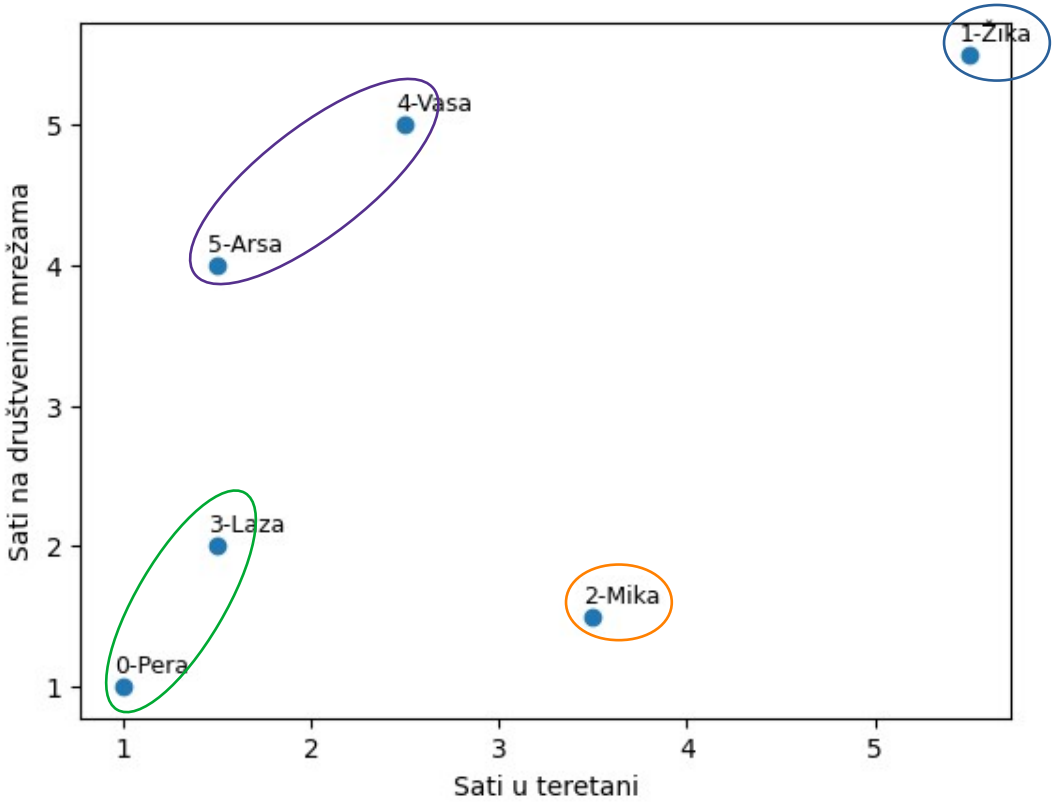
Primer 1

- Iteracija 1
  - Izračunava se matrica udaljenosti (između svaka dva klastera tj. Instance)
  - Spajaju se najbliža dva klastera u jedan (0-Pera i 3-Laza)



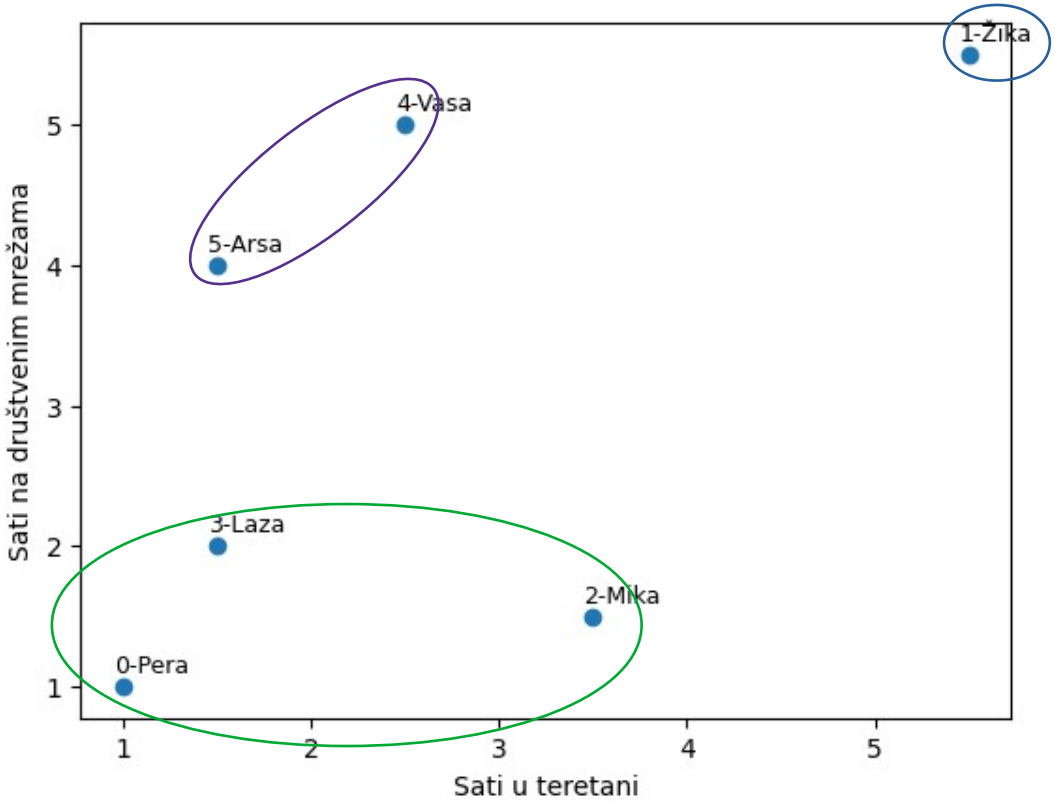
Primer 1

- Iteracija 2
  - Izračunava se matrica udaljenosti (između svaka dva klastera tj. Instance)
  - Spajaju se najbliža dva klastera u jedan (4-Vasa i 5-Arsa)



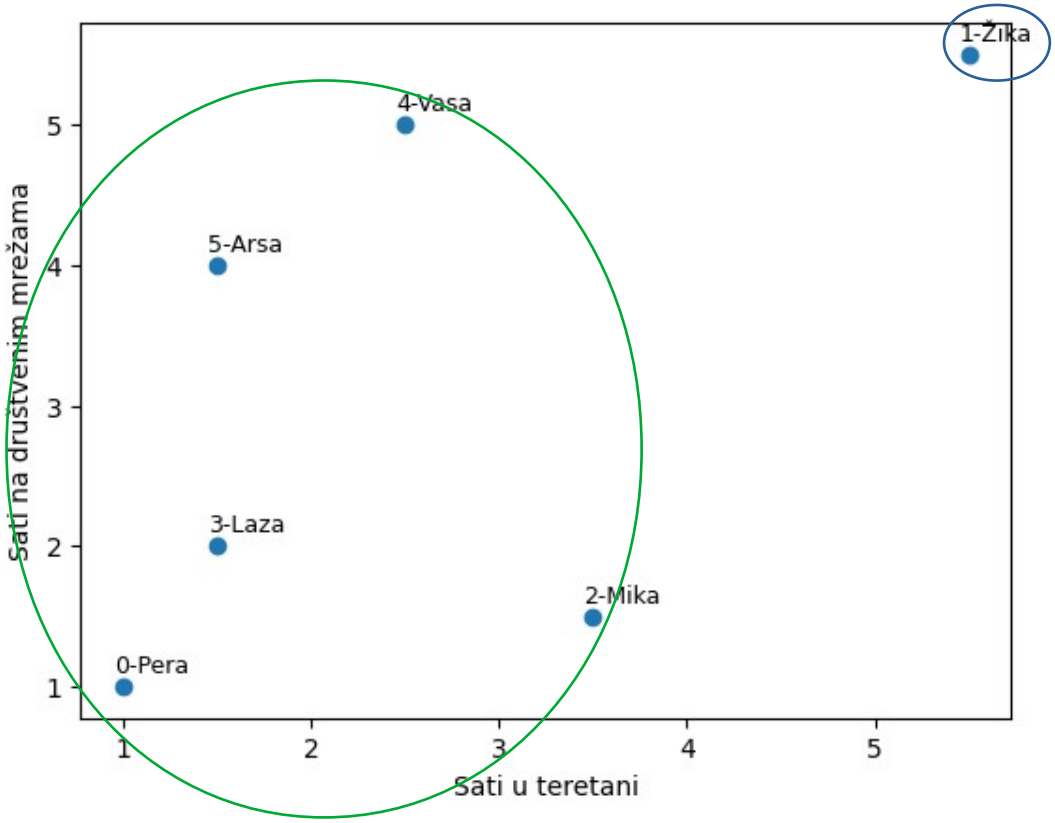
Primer 1

- Iteracija 3
  - Izračunava se matrica udaljenosti (između svaka dva klastera tj. Instance)
  - Spajaju se najbliža dva klastera u jedan ((0-Pera, 3-Laza) sa 2-Mika)



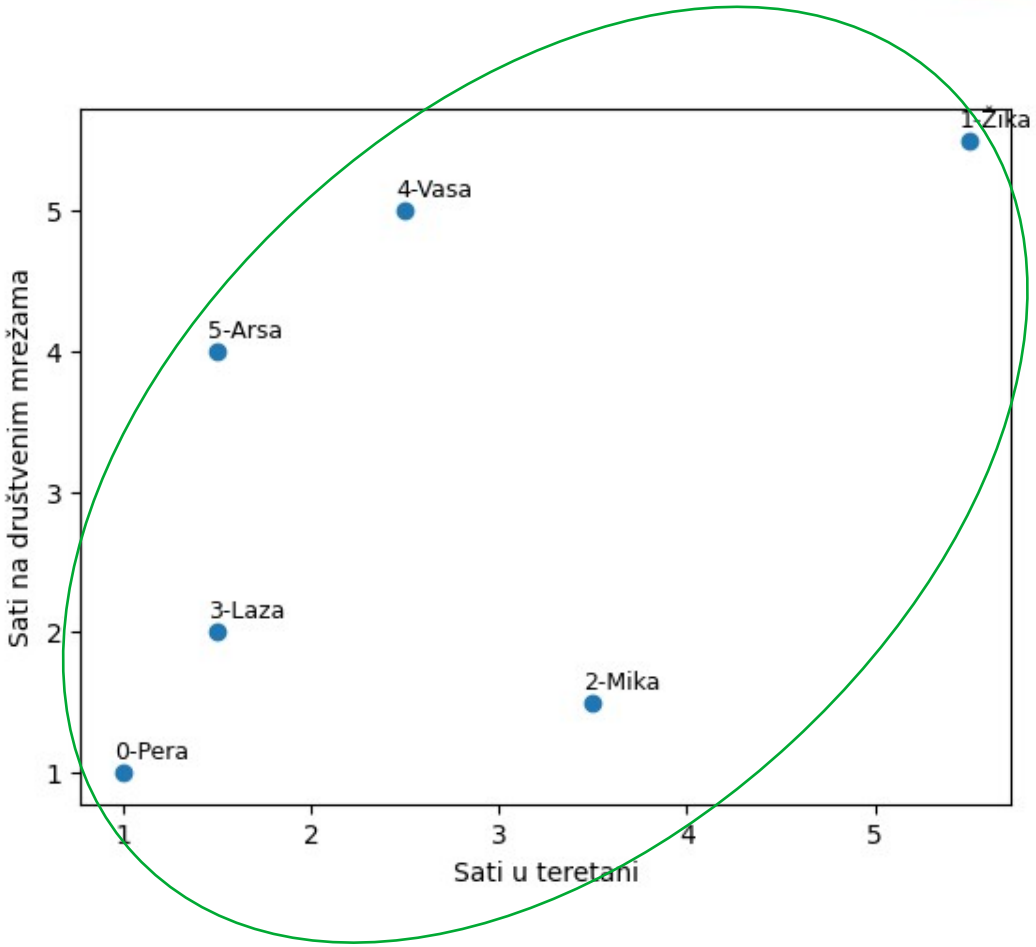
Primer 1

- Iteracija 4
  - Izračunava se matrica udaljenosti (između svaka dva klastera tj. Instance)
  - Spajaju se najbliža dva klastera u jedan ((0-Pera, 3-Laza, 2-Mika) sa (4-Vasa, 5-Arsa))



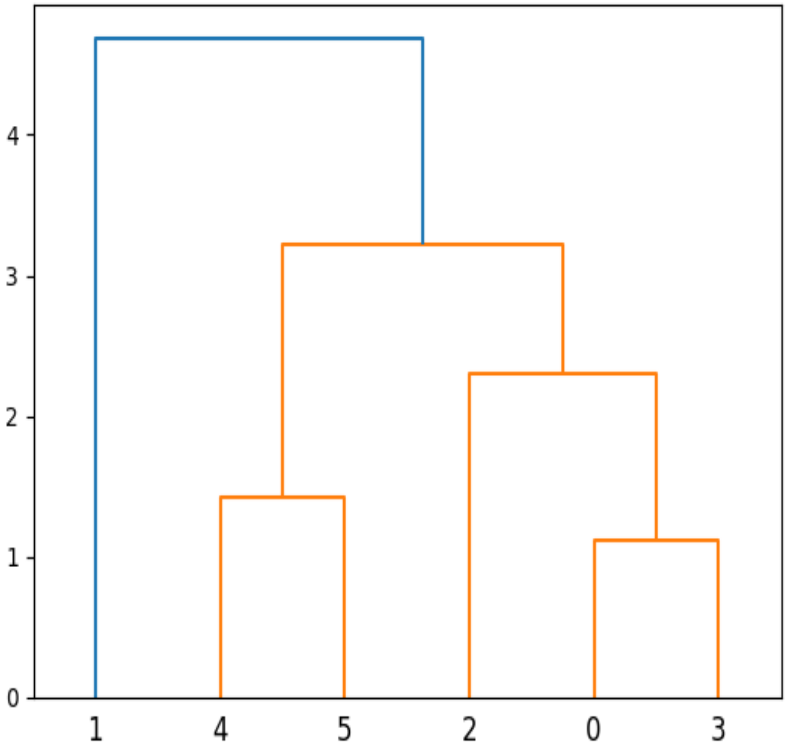
Primer 1

- Iteracija 5
  - Izračunava se matrica udaljenosti (između preostala dva klastera tj. Instance)
  - Spajaju se preostala dva klastera u jedan (koji sadrži sve instance)

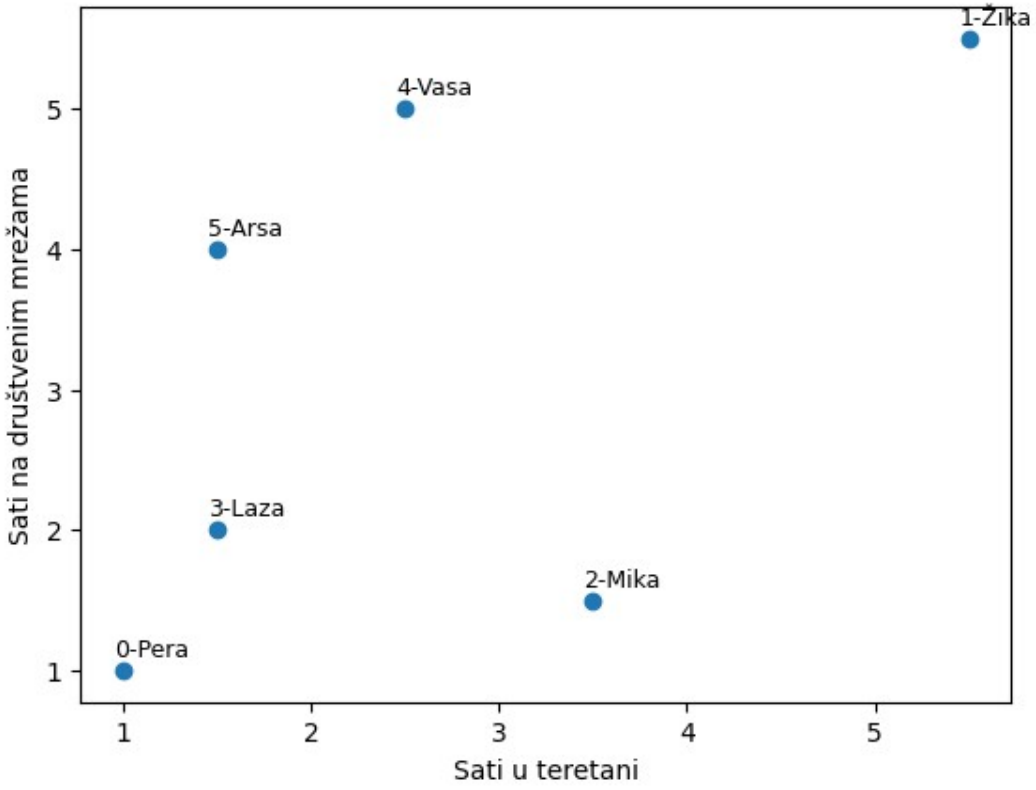


Primer 1

- Kraj (dendrogram)

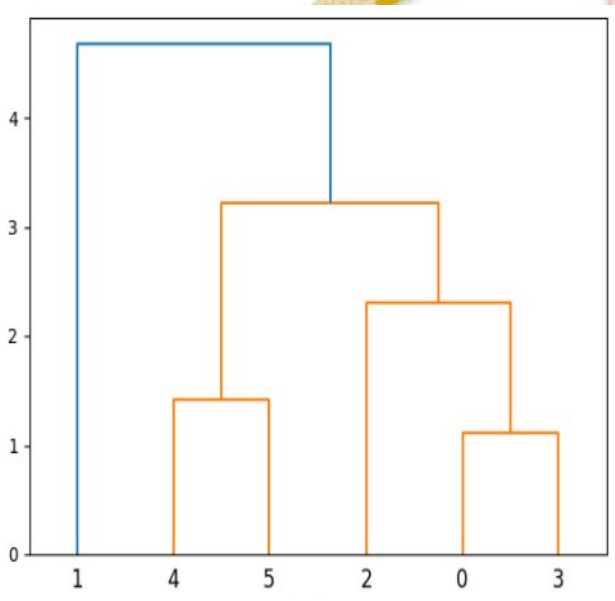


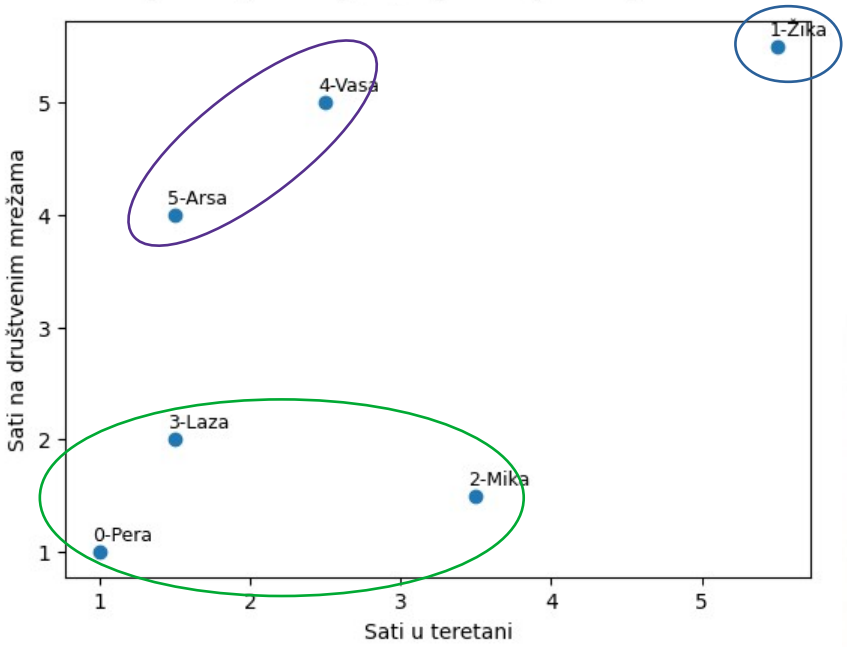
0 Pera
1 Žika
2 Mika
3 Laza
4 Vasa
5 Arsa
Name: Imena, dtype: object



Primer 1

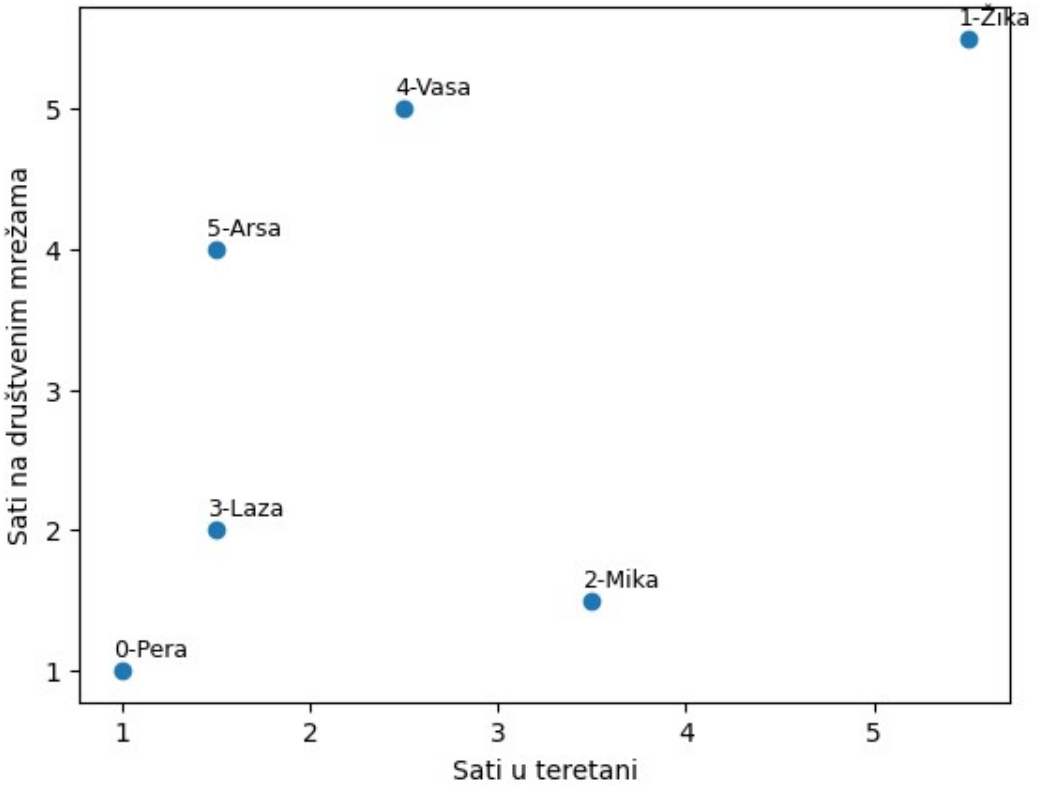
- Tumačenja dendrograma
  - Optimano 2 klastera
  - Visina je najveća(boje linija)
  - Nema mnogo smisla, prvi klaster ima samo jednu instancu (1-Žika) a drugi sve ostale
- Možda ima više smisla sa tri klastera (subjektivno)
  - „zainteresovani za teretanu“ (0,2,3)
  - „zainteresovani za društvene mreže“ (4,5)
  - „fitness influenser“ (1)





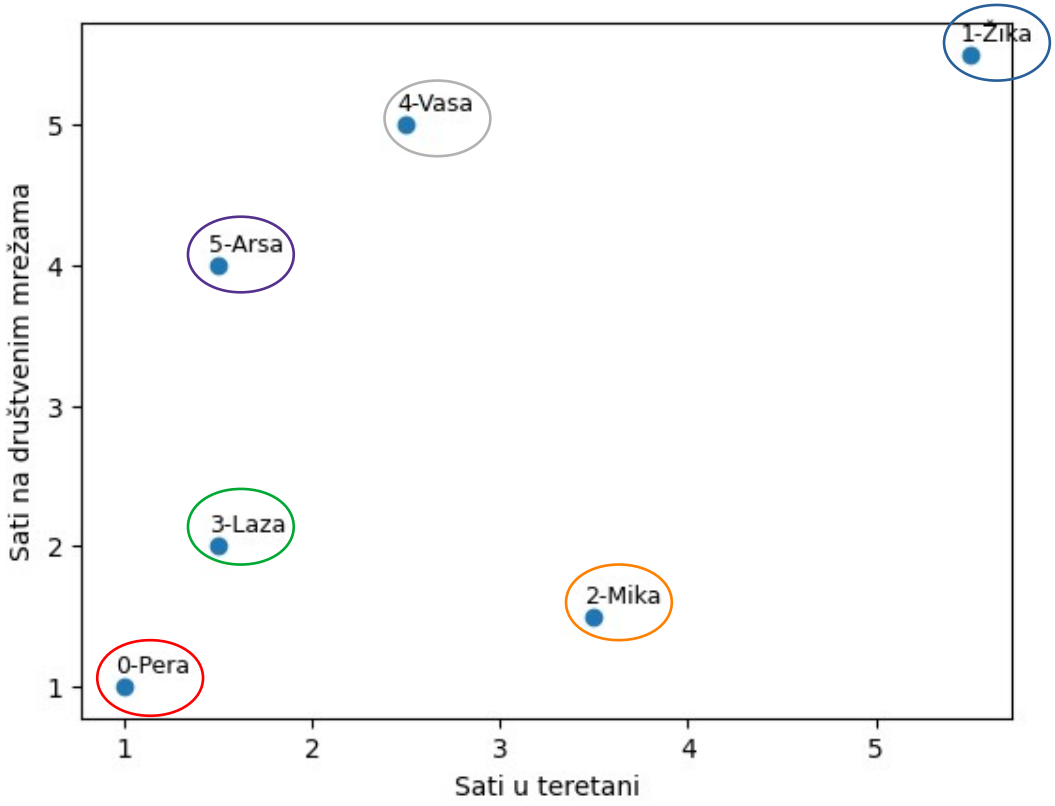
Primer 2

- Isti podaci za šest osoba
- Drugačiji parametri za algoritam
  - Euklidska udaljenost (i dalje)
  - Complete linkage



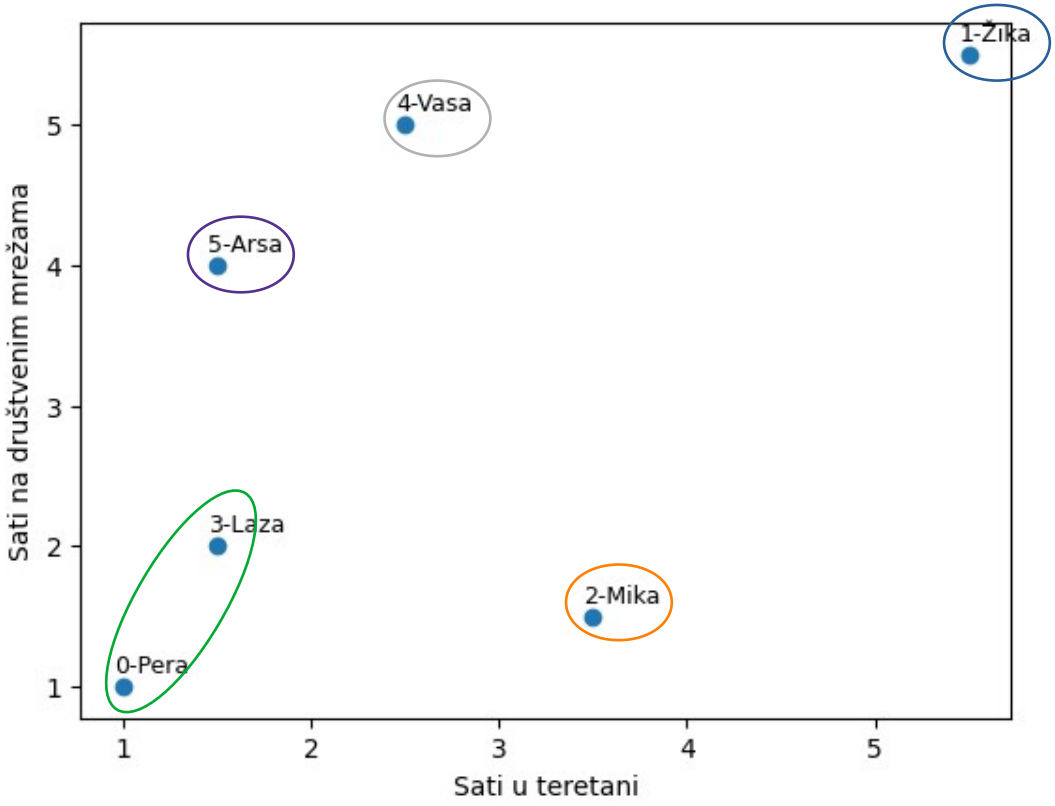
Primer 2

- Početak
  - Svaka instanca je poseban klaster



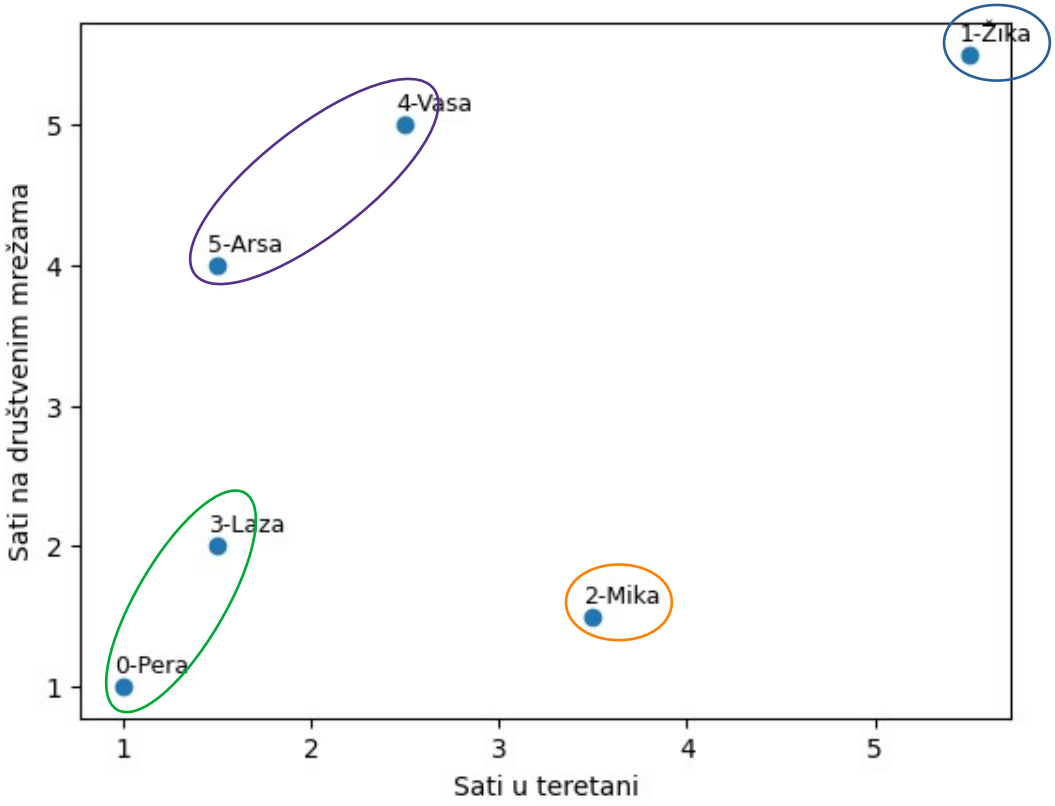
Primer 2

- Iteracija 1
  - Izračunava se matrica udaljenosti (između svaka dva klastera tj. Instance)
  - Spajaju se najbliža dva klastera u jedan (0-Pera i 3-Laza)



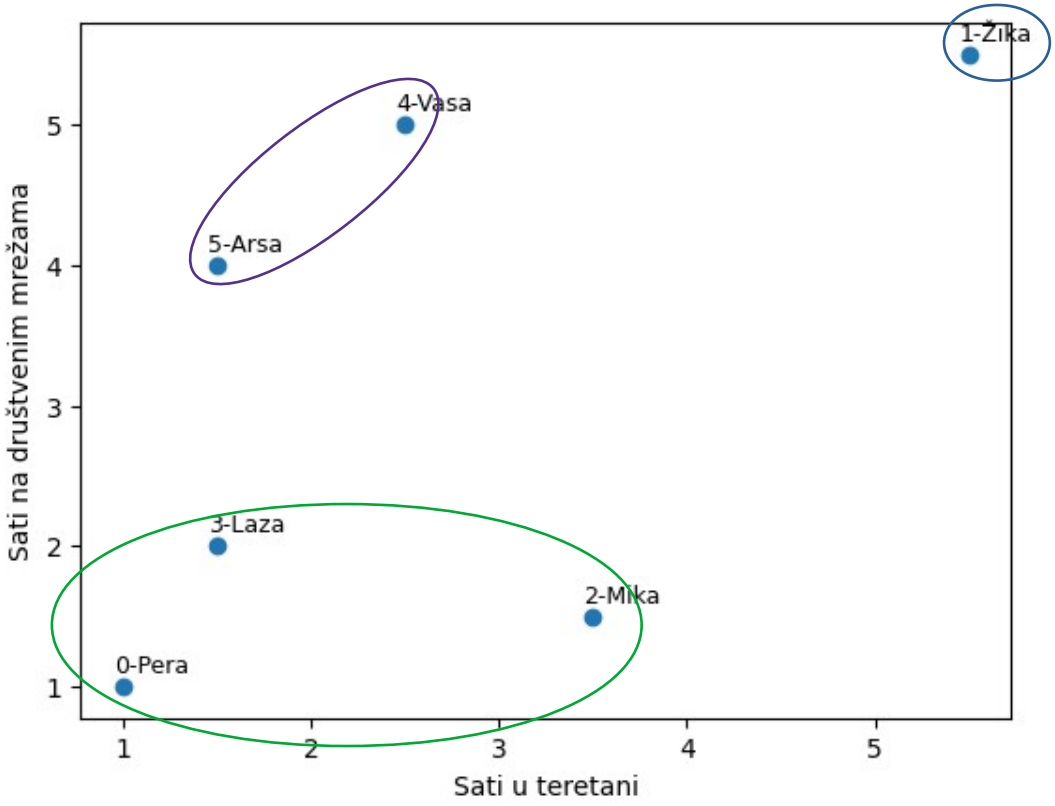
Primer 2

- Iteracija 2
  - Izračunava se matrica udaljenosti (između svaka dva klastera tj. Instance)
  - Spajaju se najbliža dva klastera u jedan (4-Vasa i 5-Arsa)



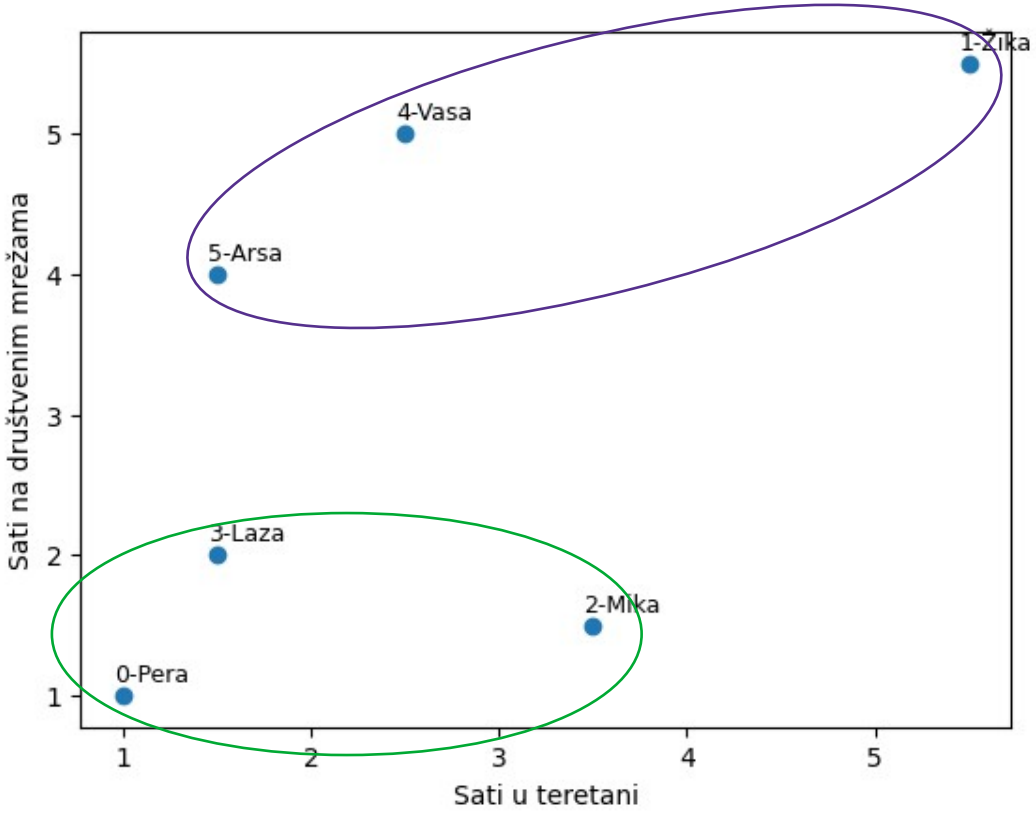
Primer 2

- Iteracija 3
  - Izračunava se matrica udaljenosti (između svaka dva klastera tj. Instance)
  - Spajaju se najbliža dva klastera u jedan ((0-Pera, 3-Laza) sa 2-Mika)



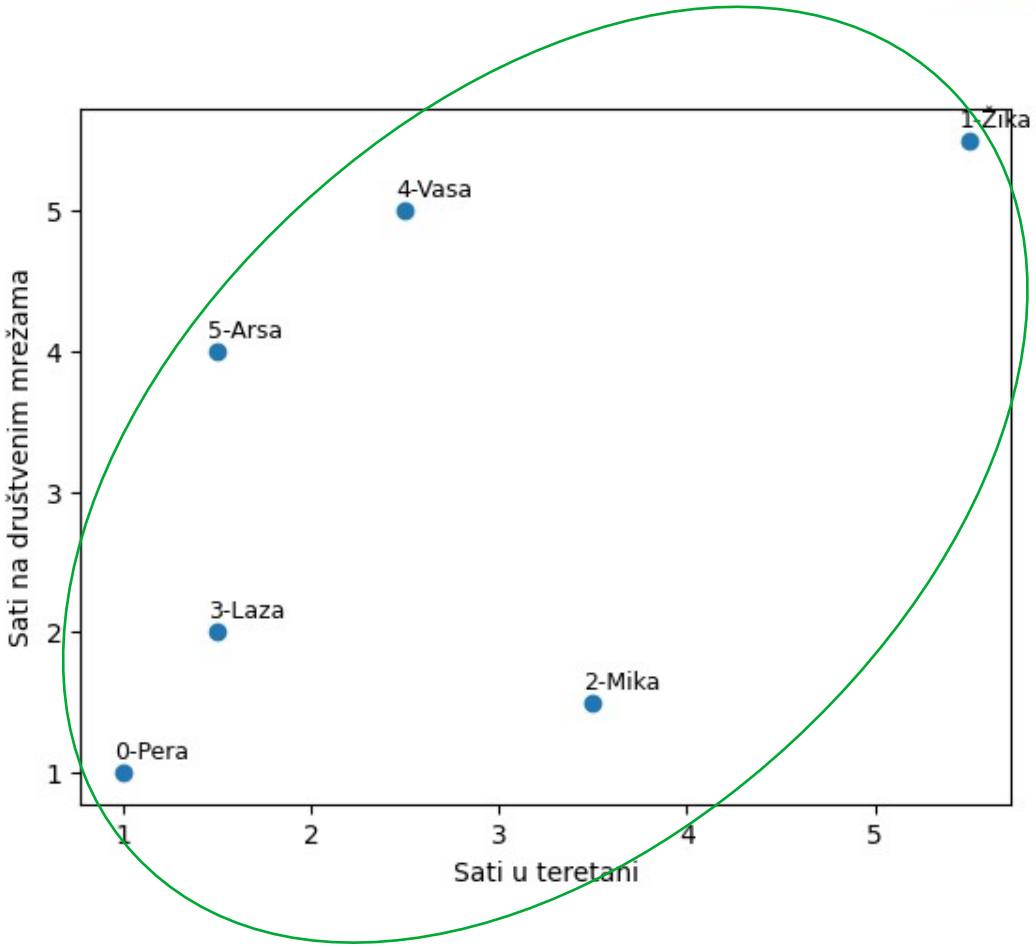
Primer 2

- Iteracija 4
  - Izračunava se matrica udaljenosti (između svaka dva klastera tj. Instance)
  - Spajaju se najbliža dva klastera u jedan ((1-Žika) sa (4-Vasa, 5-Arsa))



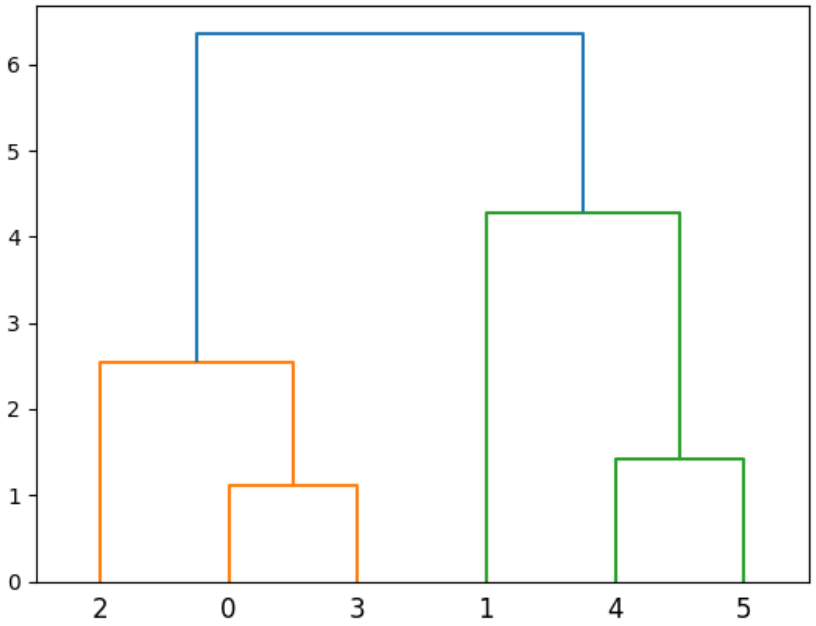
Primer 2

- Iteracija 5
  - Izračunava se matrica udaljenosti (između preostala dva klastera tj. Instance)
  - Spajaju se preostala dva klastera u jedan (koji sadrži sve instance)

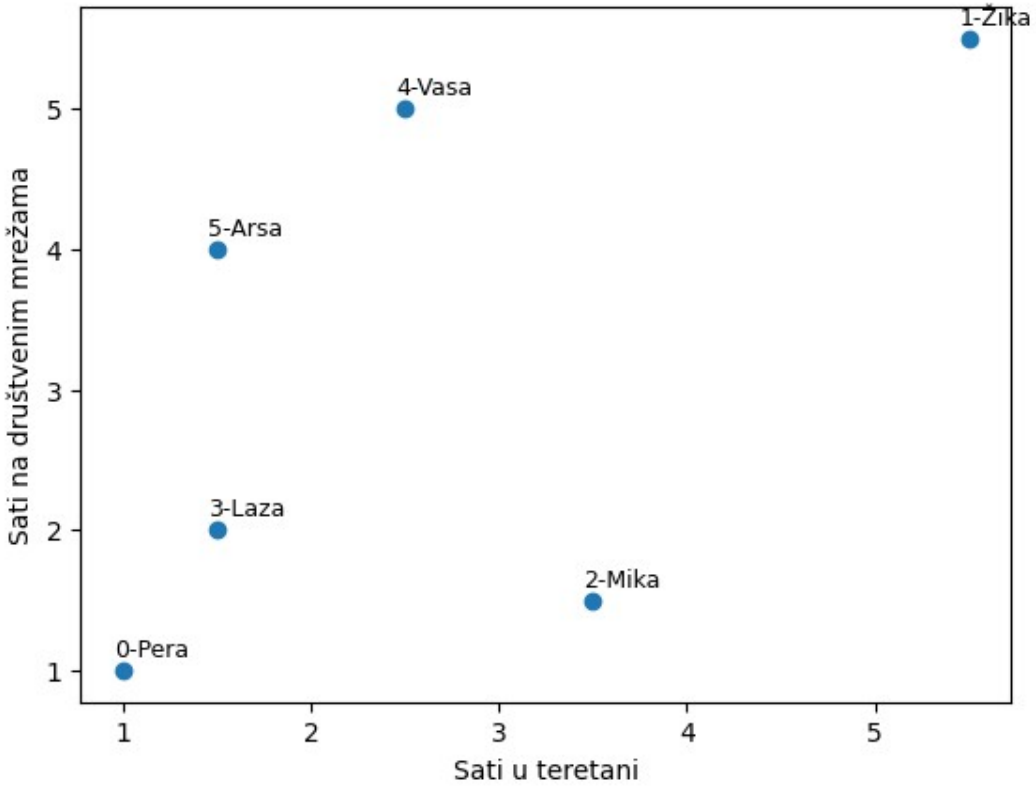


Primer 2

- Kraj (dendrogram)

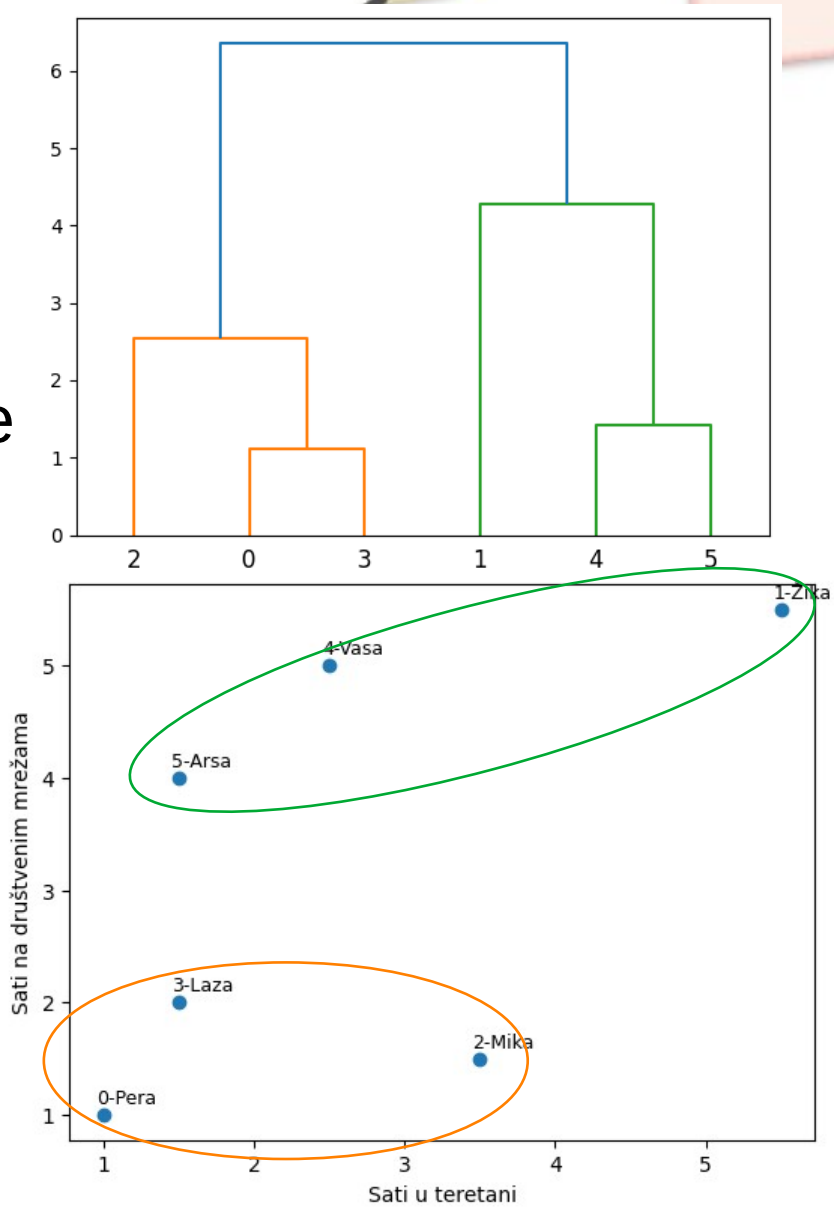


0 Pera
1 Žika
2 Mika
3 Laza
4 Vasa
5 Arsa
Name: Imena, dtype: object



Primer 2

- Tumačenja dendrograma
  - Optimano 2 klastera
  - Visina je najveća(boje linija)
  - Donekle ima smisla, ali opet se nekako čini da je Žika dosta daleko i da je klaster „razvučen“
- Tumačenje (subjektivno)
  - „zainteresovani za teretanu“ (0,2,3)
  - „zainteresovani za društvene mreže“ (1,4,5)

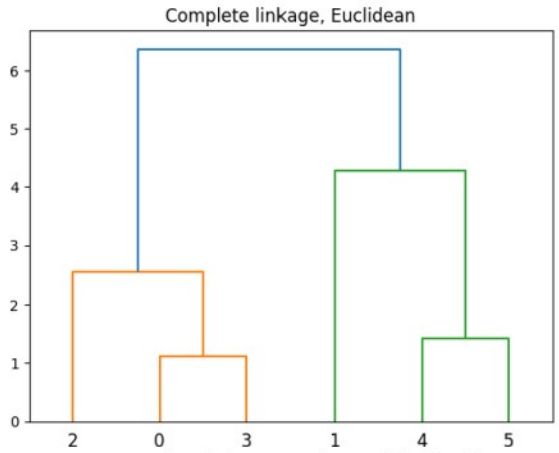


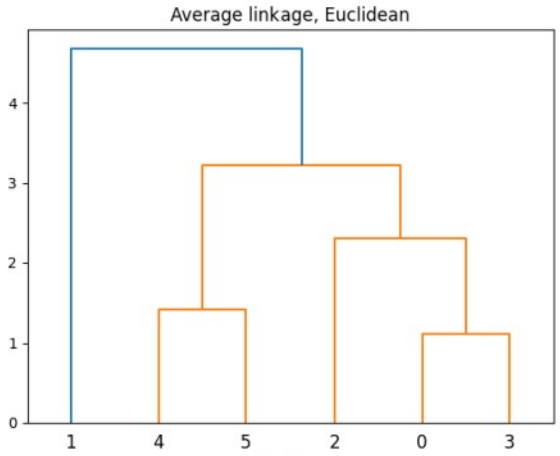
Primeri 1 i 2

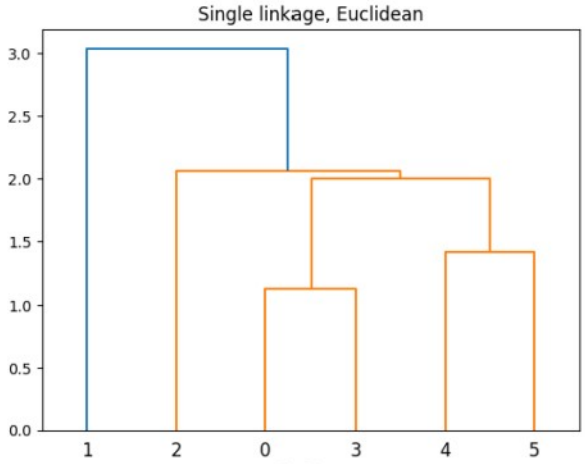
- Nisu baš realni
  - Samo dve promenljive
  - Obe imaju isti raspon od 1 do 5 (nema potrebe za normalizacijom)
  - Samo 6 instanci (redova, merenja, opažanja...)
  - Nema nedostajućih podataka
  - Nema outlier-a (1-Žika nije outlier, iako se čini tako)
- Bez obzira na sve to, rezultati klasterizacije variraju u odnosu na izabrane metode

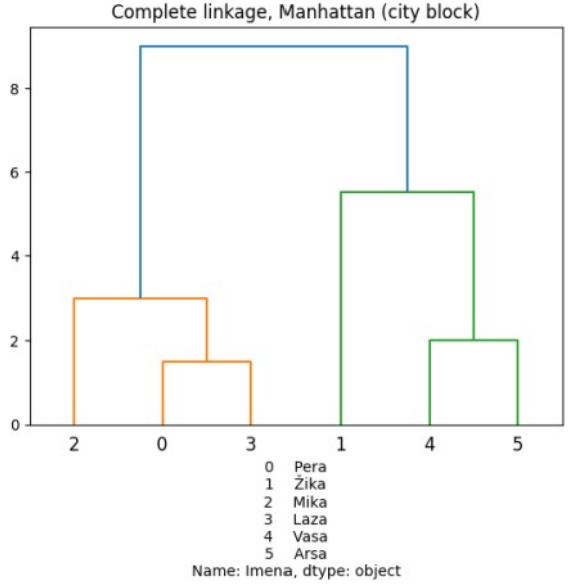
Primer 3

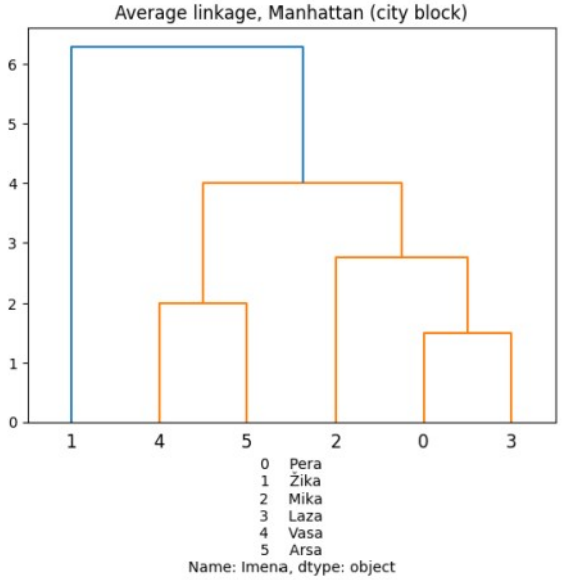
- Isti podaci, druge met. udaljenosti i povezivanja

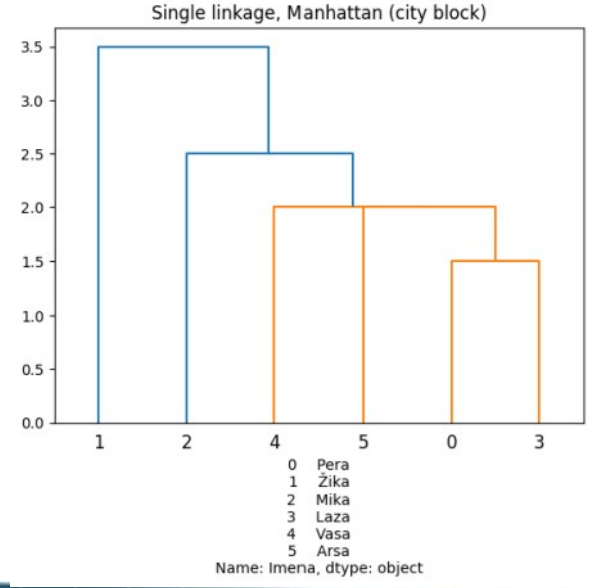












Primer 3

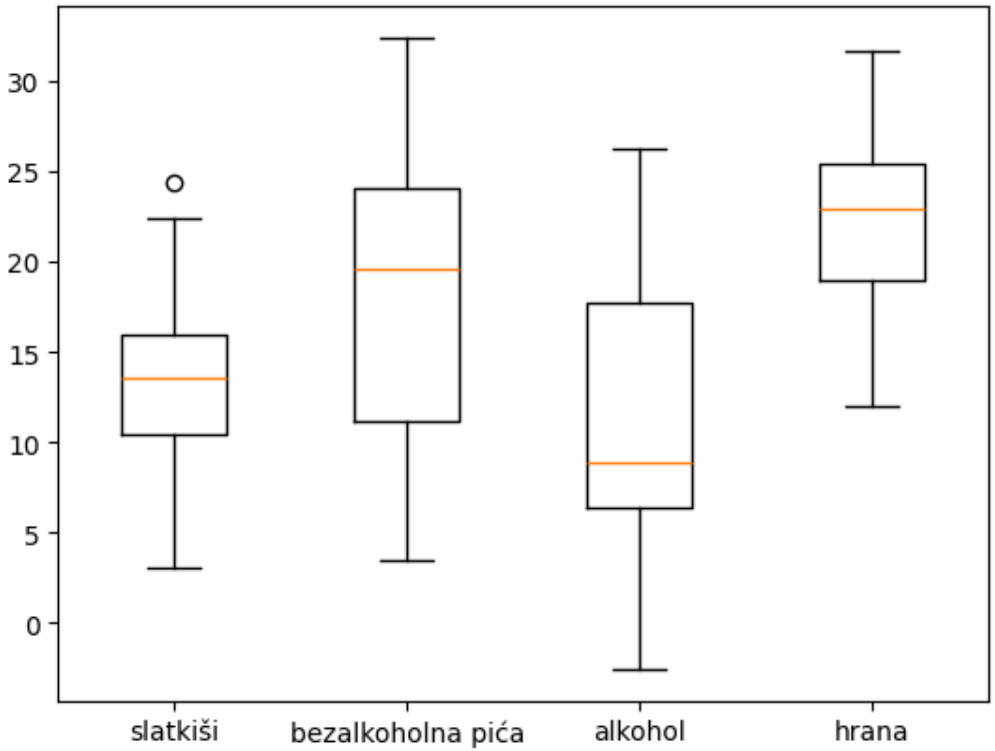
- Isti podaci, druge metode računanja udaljenosti i povezivanja
  - Često utiču na rezultate klasterovanja
  - Dobijaju se drugačiji klasteri (druge instance)
  - Dobija se drugi optimalan broj klastera
  - Drugačije visine na dendogramu ako se promeni metrika udaljenosti
- Često više utiče izbor metode povezivanja nego metode za računanje udaljenosti (iako jedna zavisi od druge)

Primer 3

- Tumačenje rezultata (dendrograma)
- Objektivna procena (objektivne mere)
  - Veličina klastera
  - „Raštrkanost“ klastera – disperzija
  - Udaljenost klastera od drugih klastera
- Subjektivna procena
  - Da li ti klasteri imaju smisla?
  - Domensko znanje i iskustvo pri tumačenju
- Subjektivna procena je često važnija od objektivne

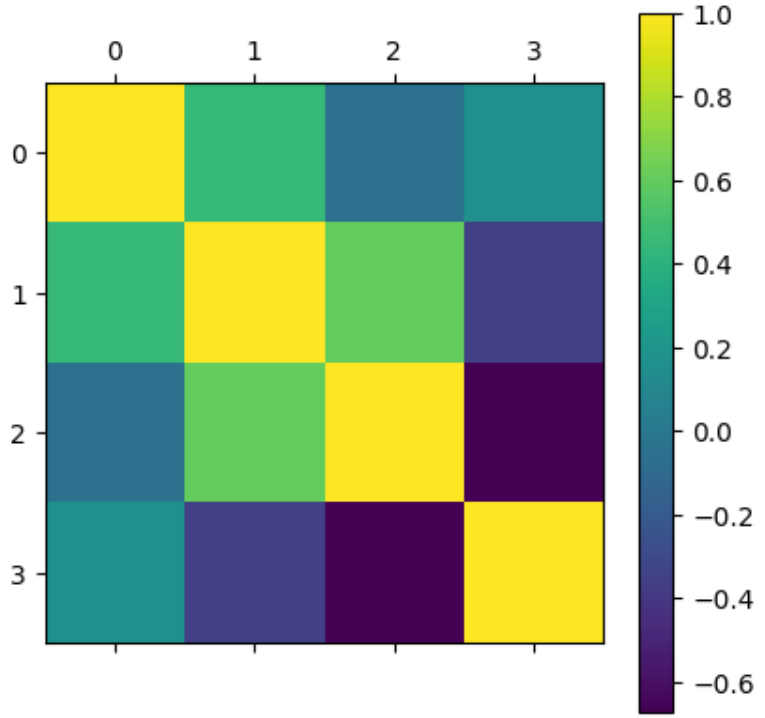
Primer 4

- Podaci:
  - Potrošnja novca na namirnice za više porodica
  - Slatkiši, bezalkoholna pića, alkohol, hrana
- Višedimenzionalni podaci
- Još realniji primer



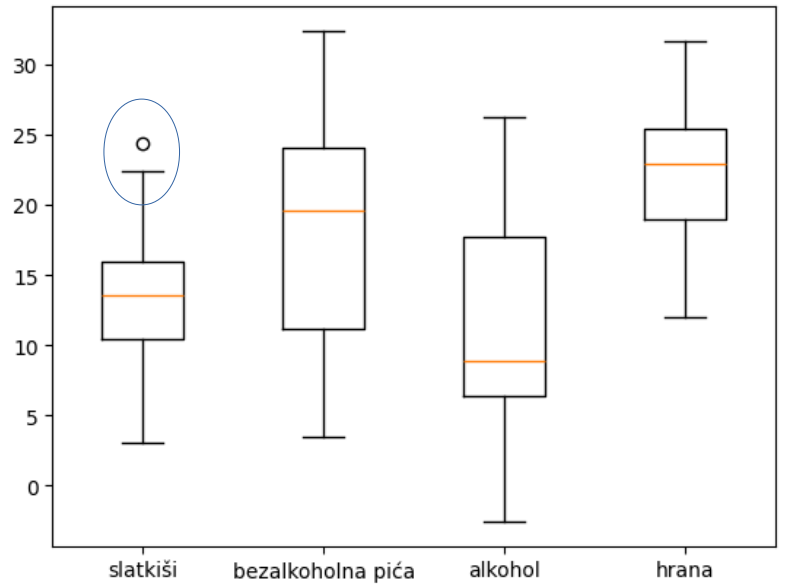
Primer 4

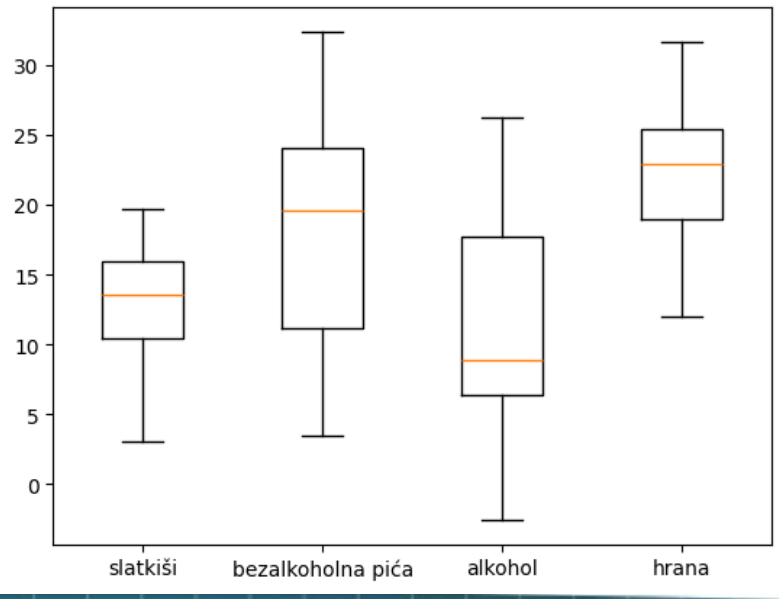
- Probamo korelaciju da vidimo da li su neke promenljive u vezi
- Nažalost, ne daje dobre rezultate
- Velika (negativna) korelacija je samo između hrane (2) i alkohola (3)



Primer 4

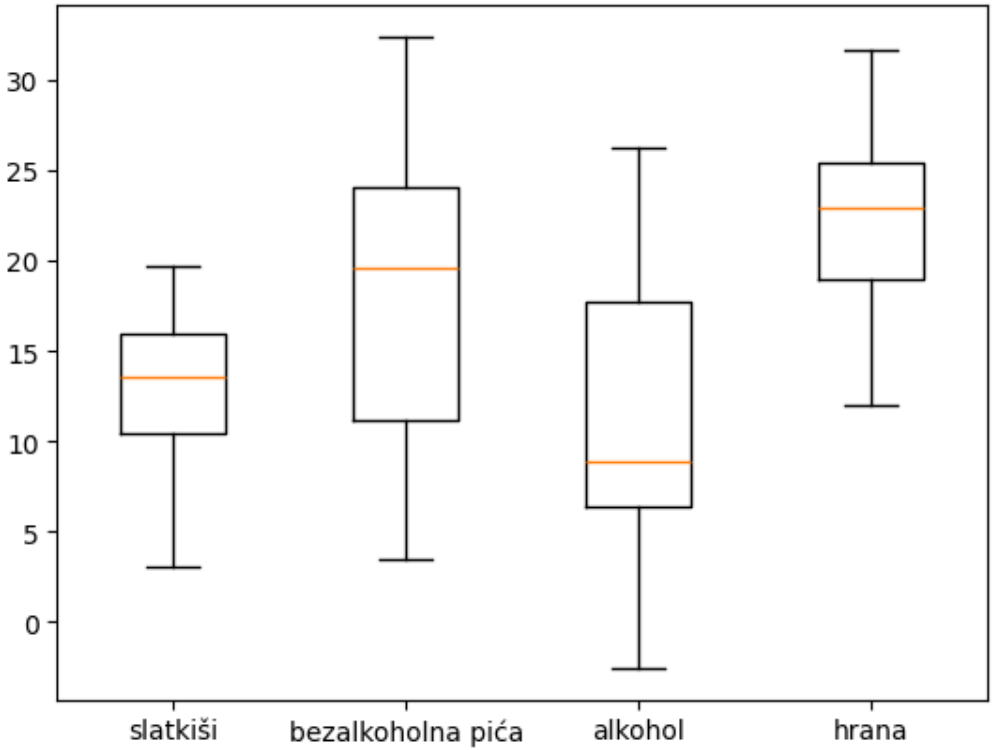
- Imamo outlier-e kod slatkiša
- Samo kod gornje granice
- Winsorize (samo preko 98 percentila)





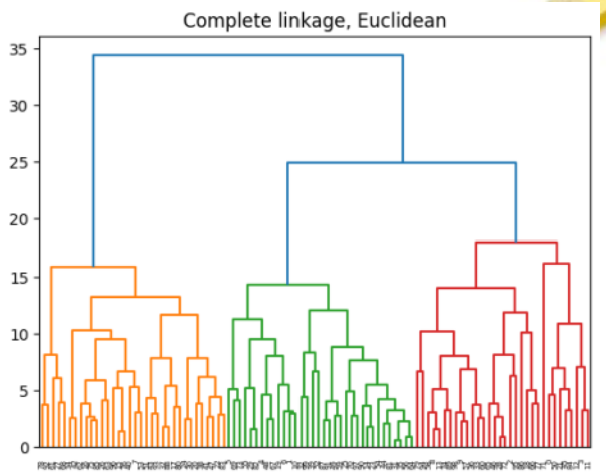
Primer 4

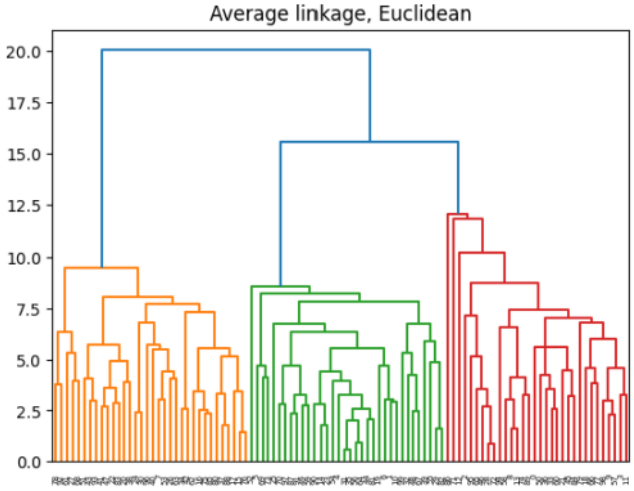
- Probamo hijerarhijsku klasterizaciju
- Klasteri se ne mogu uočiti okom
- Koliko klastera je optimalno?



Primer 4

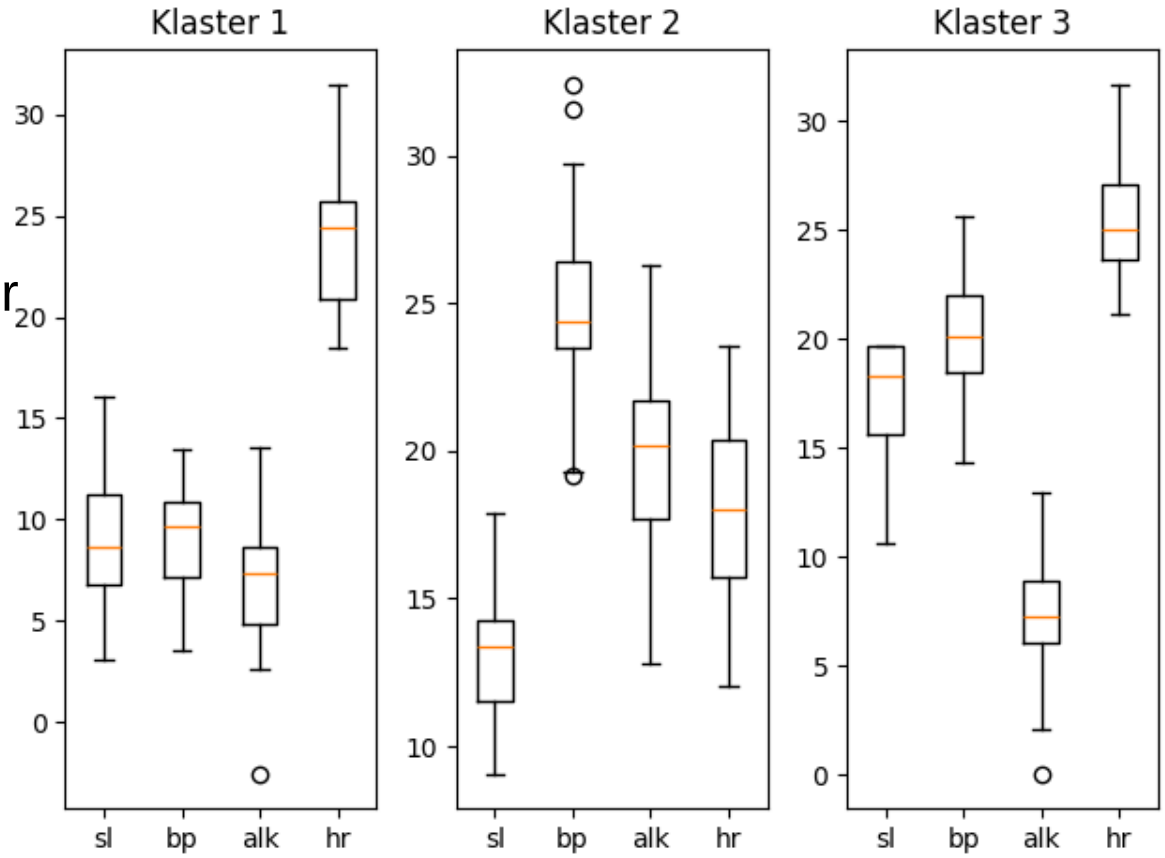
- Dendrogrami sa različitim metodama povezivanja
  - Čini se da je najbolje rešenje sa 3 klastera
  - Ward (2 ili 3)
  - Slična veličina klastera (širina na dendrogramu)





Primer 4

- Tri klastera
  1. Porodice koje se zdravo hrane (klaster 1)
  2. Porodice više vole da piju nego da jedu (klaster 2)
  3. Porodice koje vole puno da jedu i piju, samo ne alkohol (klaster 3)
- Prikaz
  - Opet nije moguć jedan scatterplot
  - Opet više uporednih boxplot-ova



Hijerarhijska klasterizacija

- Prednosti
  - Nije potrebno pretpostaviti broj klastera (K)
  - Dendrogrami su intuitivni i lako se interpretiraju
- Mane
  - Procesorski i memorijski intenzivna (matrica udaljenosti)
  - Osetljiva na izbor metrike udaljenosti
  - Osetljiva na izbor metrike za povezivanje klastera (linkage)
  - Osetljiva na šum i outlier-e