



# Klasterizacija

Bojan Tomić  
Jelena Jovanović

# Klasterizacija kao zadatak VI



- Osnove Python-a
  - Numpy, Pandas, Scikit-learn...
- Utvrđivanje zavisnosti i predviđanje
- Klasifikacija
- Klasterizacija
- Pretraživanje

# Klasterizacija kao zadatak VI



- Utvrđivanje zavisnosti i predviđanje
  - Korelacije, (linearna) regresija
- Klasifikacija
  - Stabla odlučivanja, KNN, neuronske mreže
- Klasterizacija
  - K-means, hijerarhijska klasterizacija
- Pretraživanje (search)
  - Breadth-first, depth-first, best-first

A yellow pencil and a pink eraser are positioned in the top right corner of the white paper, suggesting a classroom or study environment.

Šta je klasterizacija?

# Šta je klasterizacija?

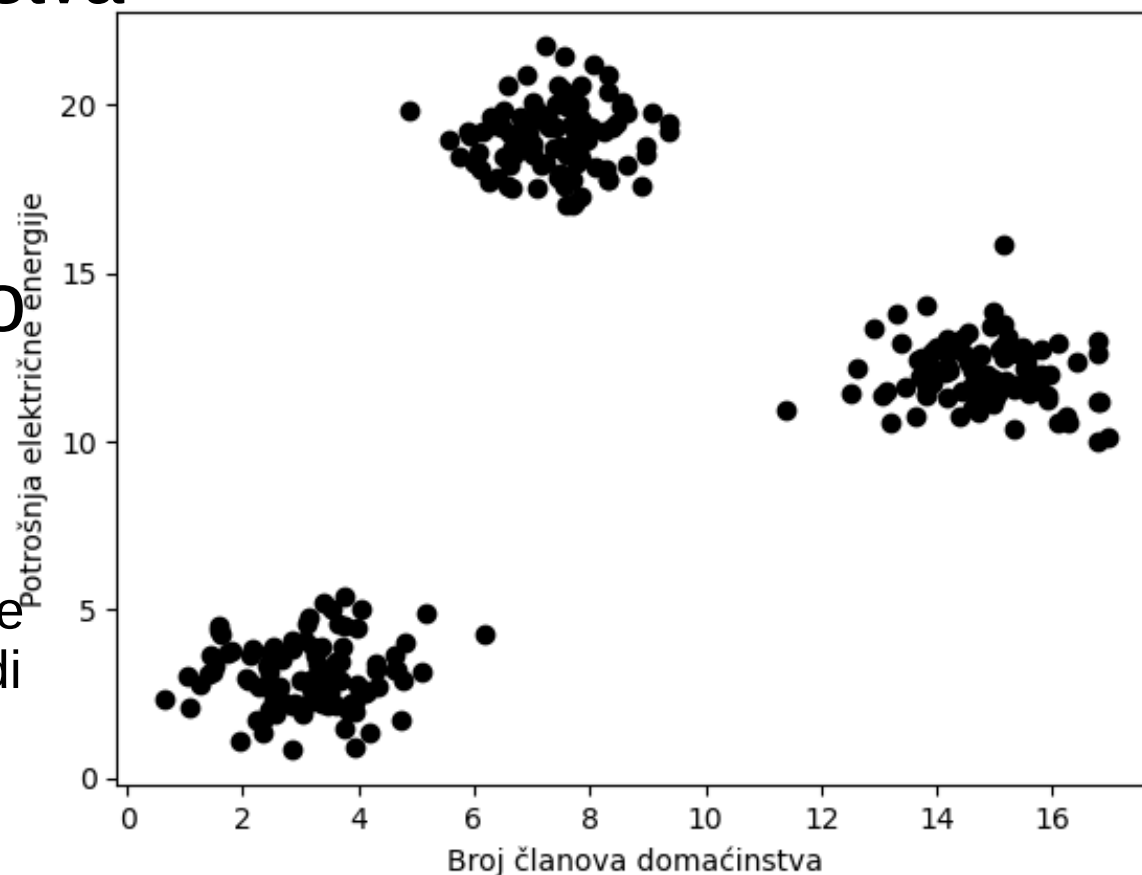
- Klasterizacija je zadatak grupisanja instanci, tako da za svaku instancu važi da je sličnija (bliža) instancama iz svoje grupe (klastera), nego instancama iz drugih grupa (klastera)
- Ciljevi
  - Identifikacija potencijalnih klastera sličnih instanci
  - Svrstavanje novih instanci u identifikovane klastere
- Nije klasifikacija – jer nisu poznate klase
- Nije regresija – jer se ne traži zavisnost, samo sličnost.



# Primer 1

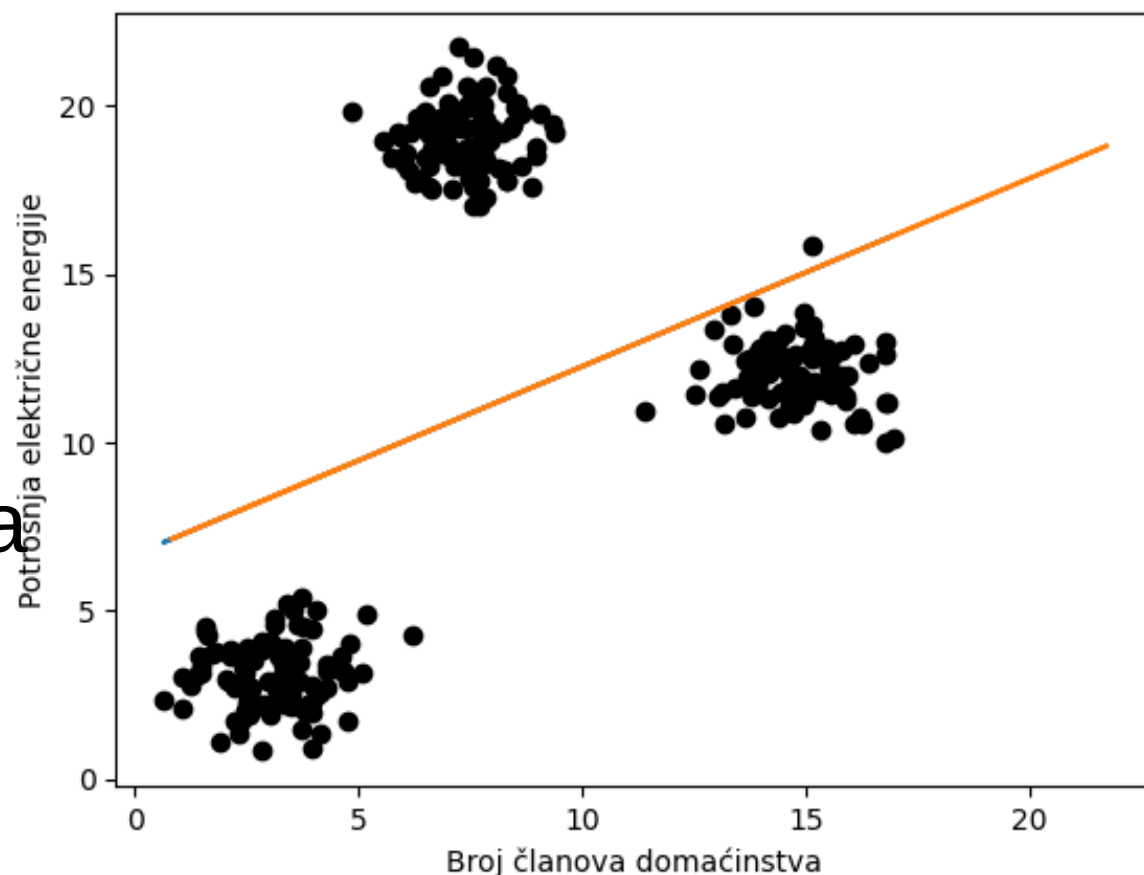
- Podaci:
  - Potrošnja električne energije domaćinstva
  - Broj članova domaćinstva\*
- Jedna tačka-jedno domaćinstvo

\* broj članova domaćinstva je uvek ceo broj, ali je ovde radi stvaranja „idealnog“ primera stavljeno drugačije



# Primer 1

- Ovo nije problem koji se rešava linearnom regresijom
- Nema (linearne) veze među promenljivima
- U tom smislu, „problem“ je grupa domaćinstava sa 5-10 članova

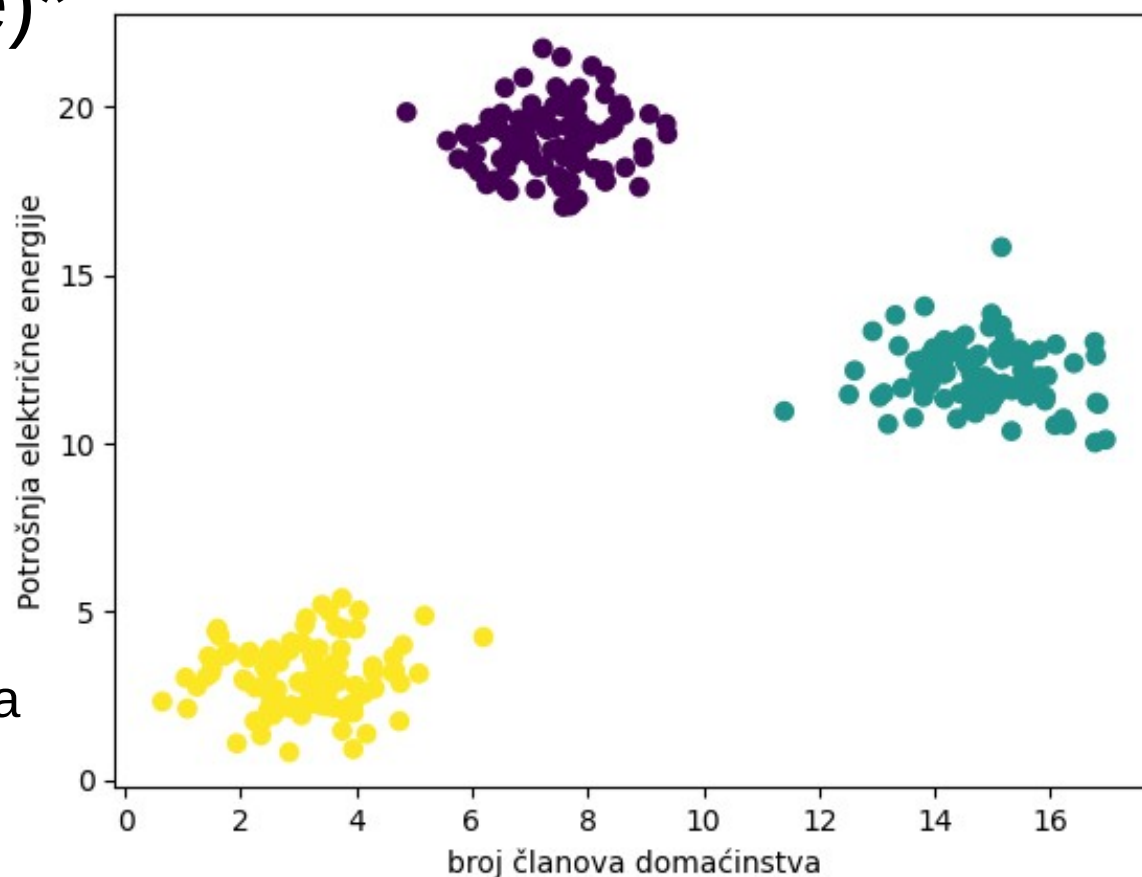


# Primer 1

- Uradimo klasterizaciju
- Tri klastera (grupe)\*

- 1) Mala domaćinstva sa malom potrošnjom - žuta (stanovi?)
- 2) Srednja domaćinstva sa visokom potrošnjom - ljubičasta (gradske porodične kuće?)
- 3) Velika domaćinstva sa prosečnom potrošnjom - zelena (seoska porodična imanja?)

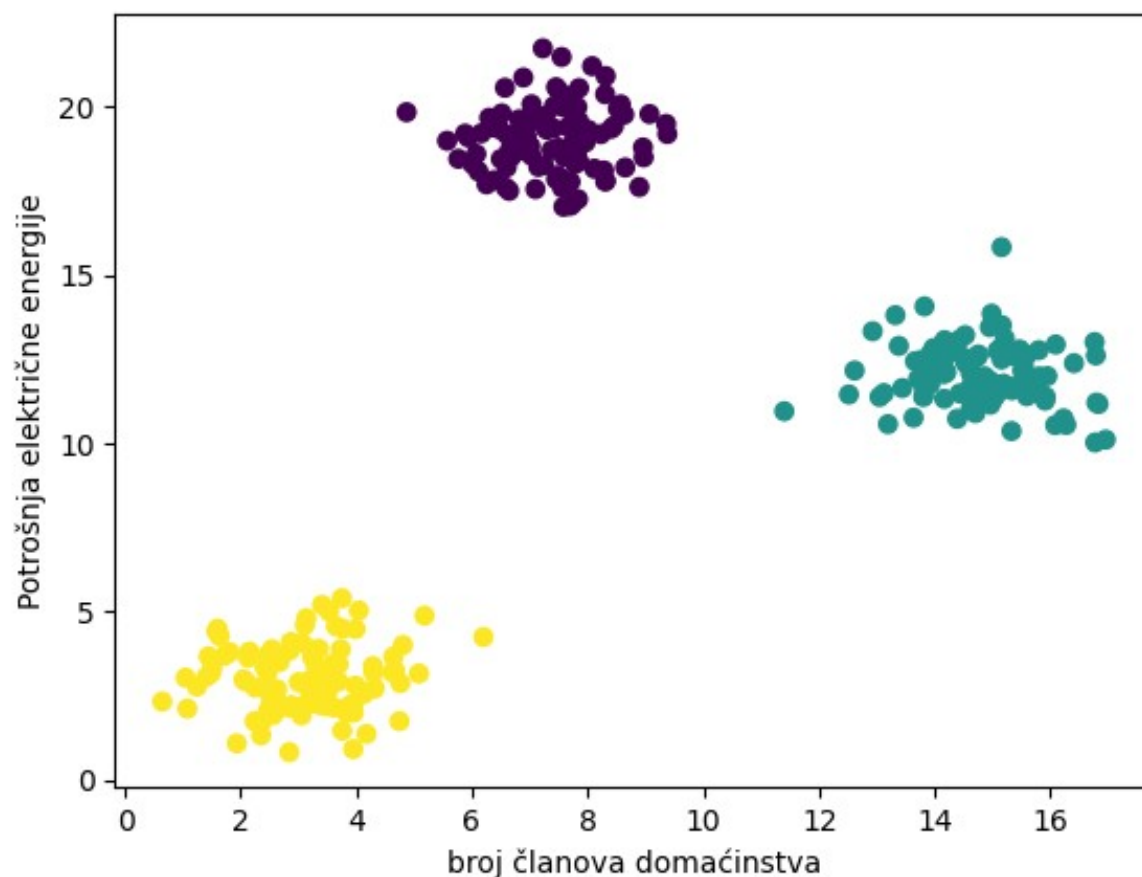
\* *Subjektivno tumačenje*





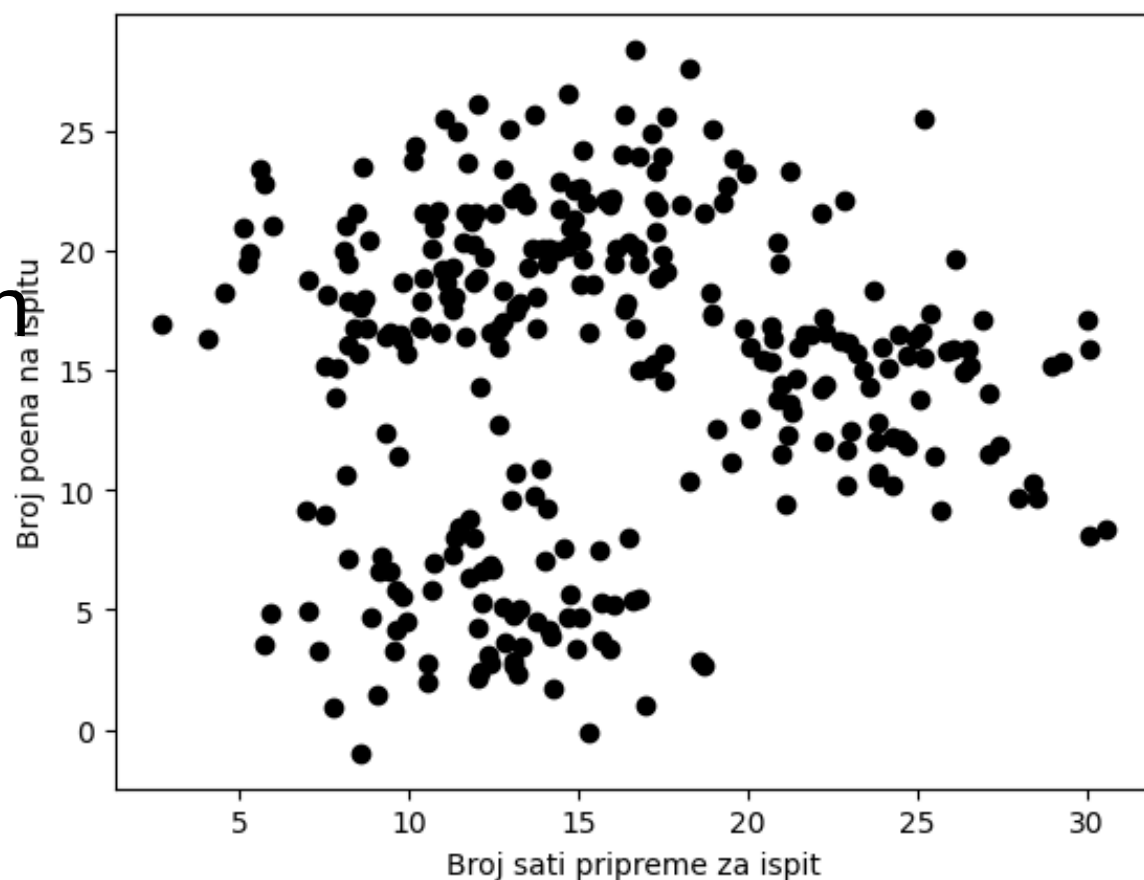
# Primer 1

- Samo jedno rešenje (3 klastera)
- Dve promenljive
- Idealni klasteri
  - Gusto zbijeni
  - Međusobno razdvojeni
  - Iste veličine
- Idealan, nerealan primer (jednorog)



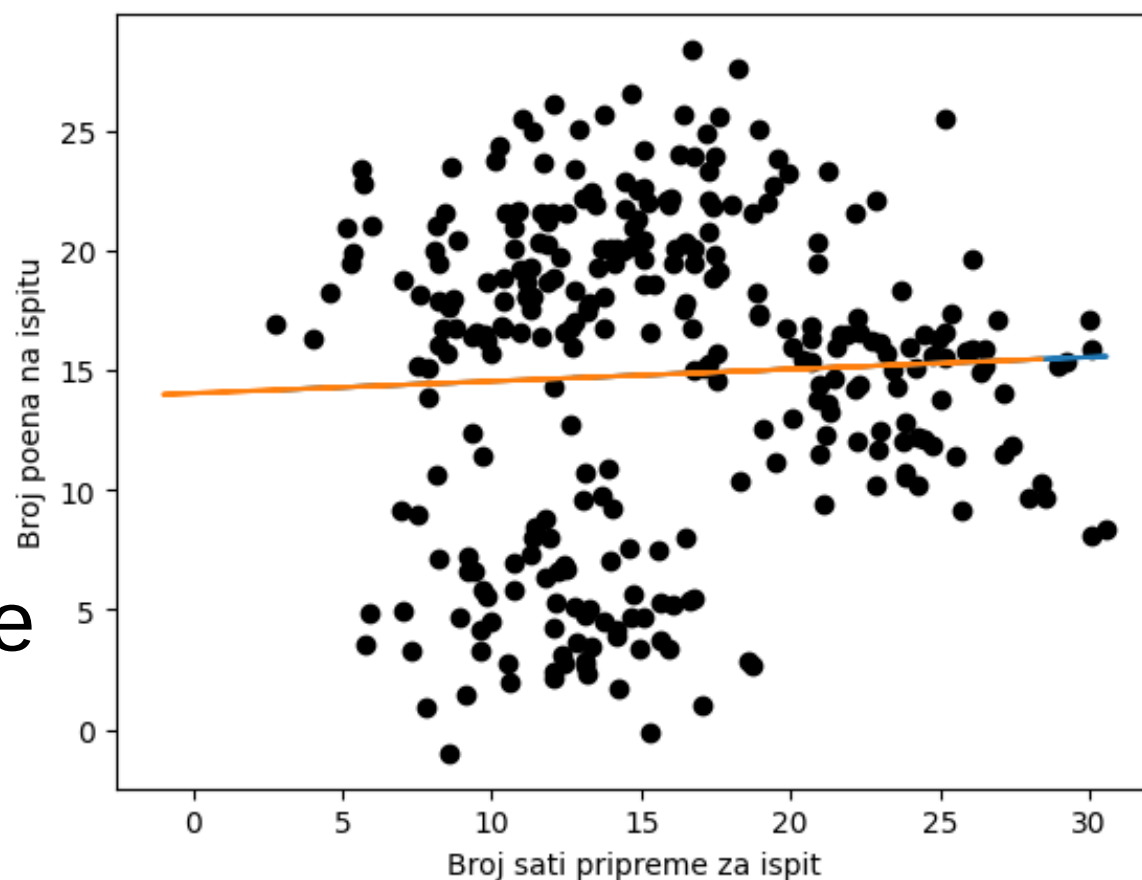
# Primer 2

- Podaci:
  - Broj sati pripreme za ispit
  - Broj ostvarenih poena na ispitu
- Jedna tačka-jedan student
- Malo realniji primer



# Primer 2

- Ni ovo nije problem koji se rešava regresijom
- Nema (linearne) veze među promenljivima
- Prevelika je „raštrkanost“ podataka oko linije regresije



# Primer 2

- Uradimo klasterizaciju

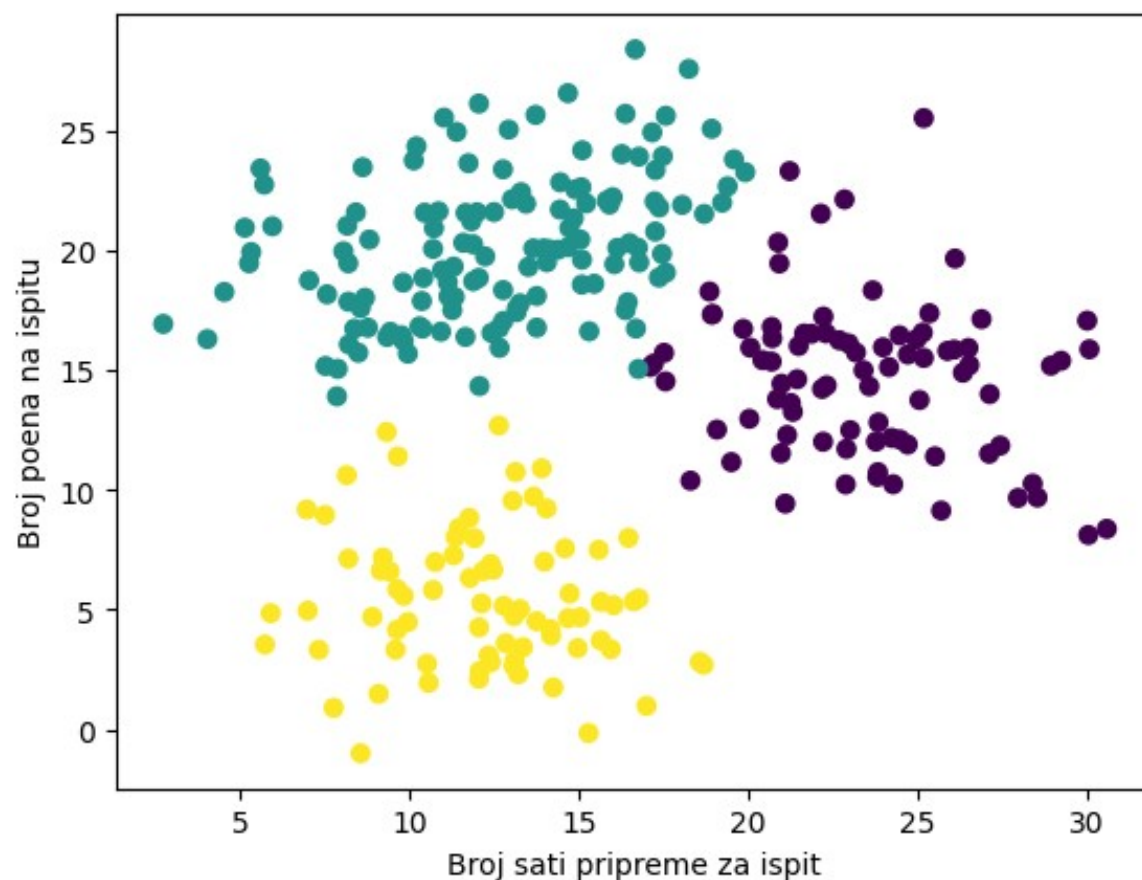
- Tri klastera (možda)?\*

1) Malo učili, slabo uradili (žuta)

2) Malo ili srednje učili, dobro uradili (zelena)

3) Puno učili, osrednje uradili (ljubičasta)

\* *Subjektivno tumačenje*



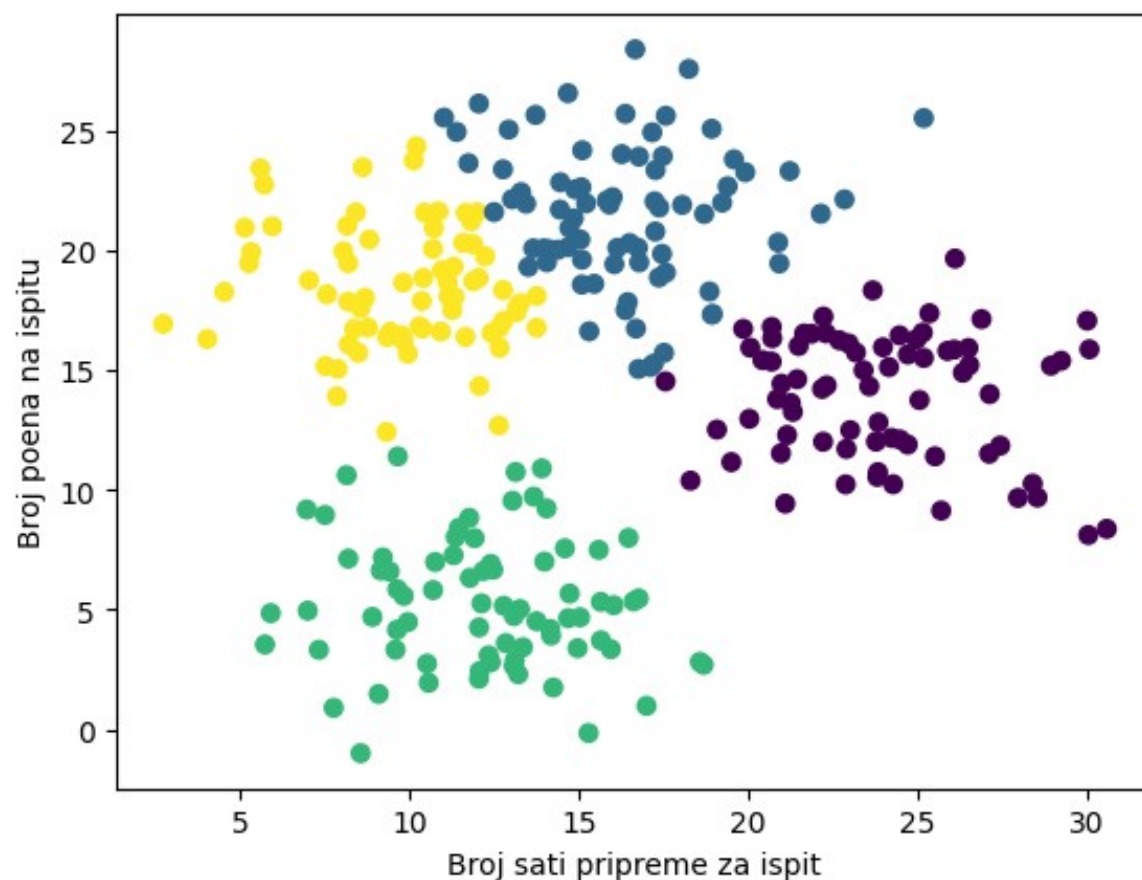


# Primer 2

- Četiri klastera (možda)?\*

- 1) Malo učili, slabo uradili (zelena)
- 2) Malu učili, dobro uradili (žuta)
- 3) Srednje učili, dobro uradili (plava)
- 4) Puno učili, osrednje uradili (ljubičasta)

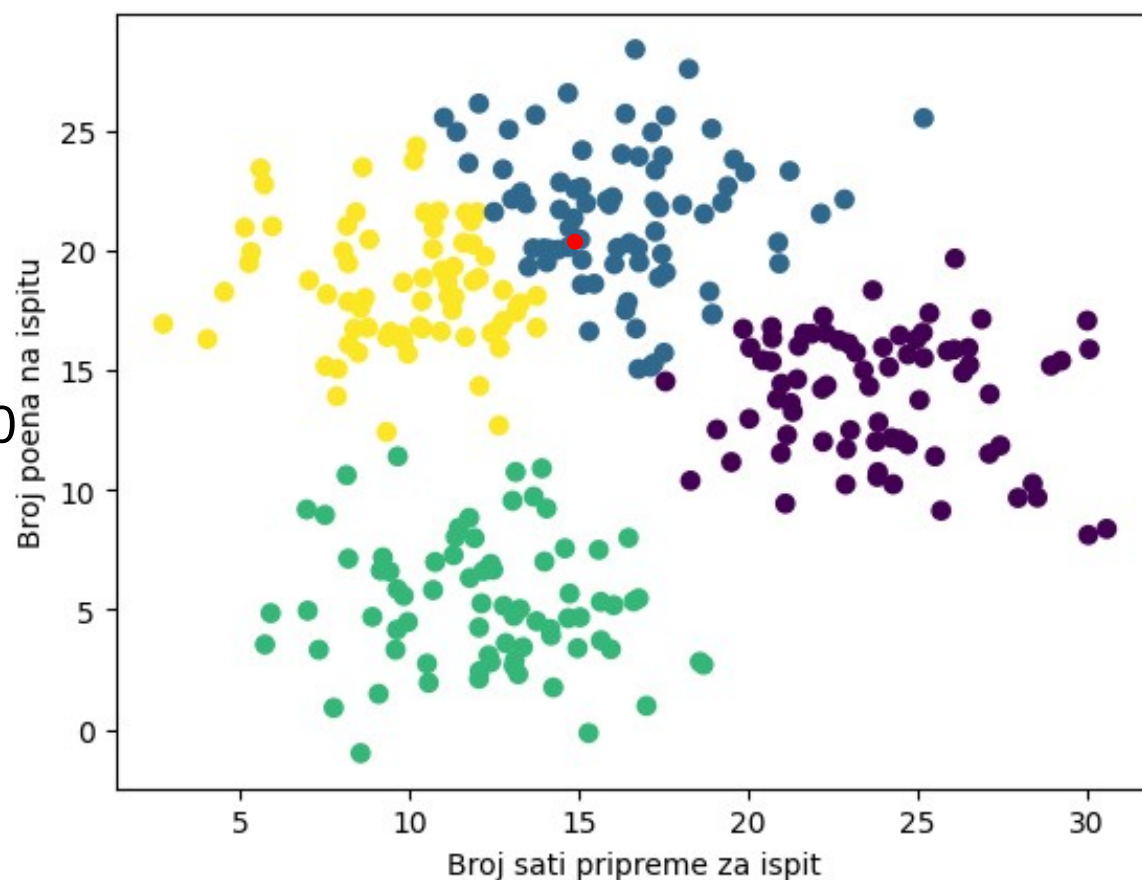
\* *Subjektivno tumačenje*





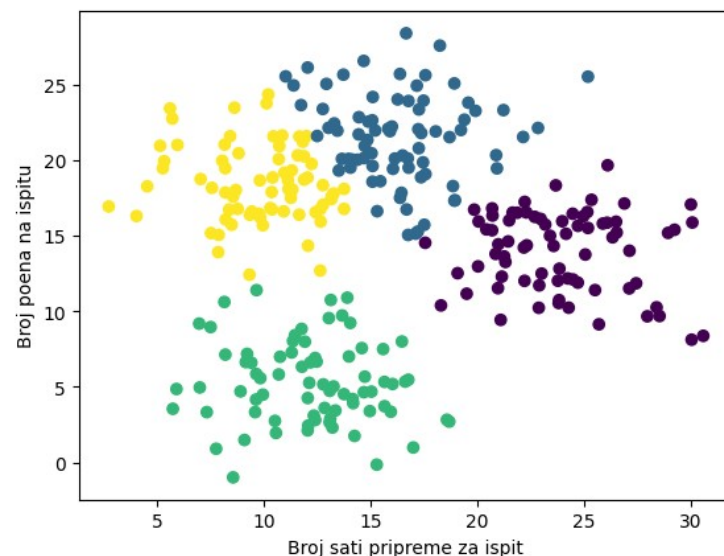
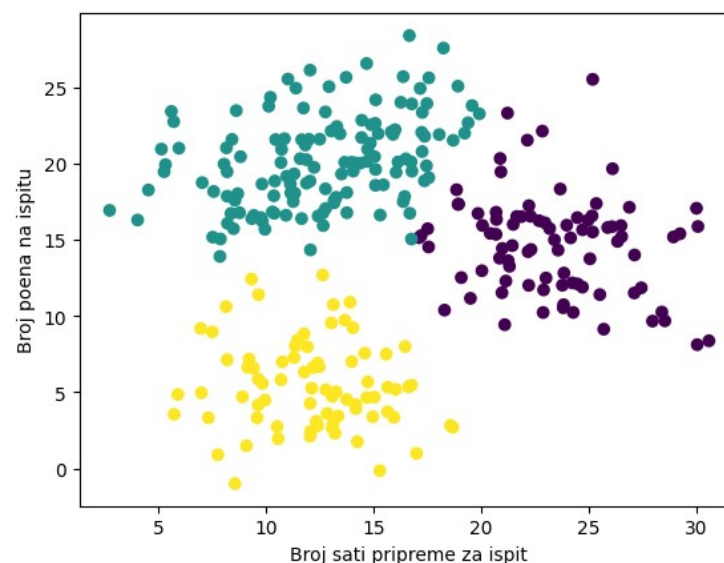
# Primer 2

- Dodatni zadatak klasterizacije
- Svrstavanje novih instanci u nađene klasterere:
  - Kom klasteru pripada student koji je 15 sati pripremao ispit i dobio 20 poena (crvena tačka)?
  - Prema modelu sa 4 klastera, pripada plavom klasteru „Srednje učili, dobro uradili“.



# Primer 2

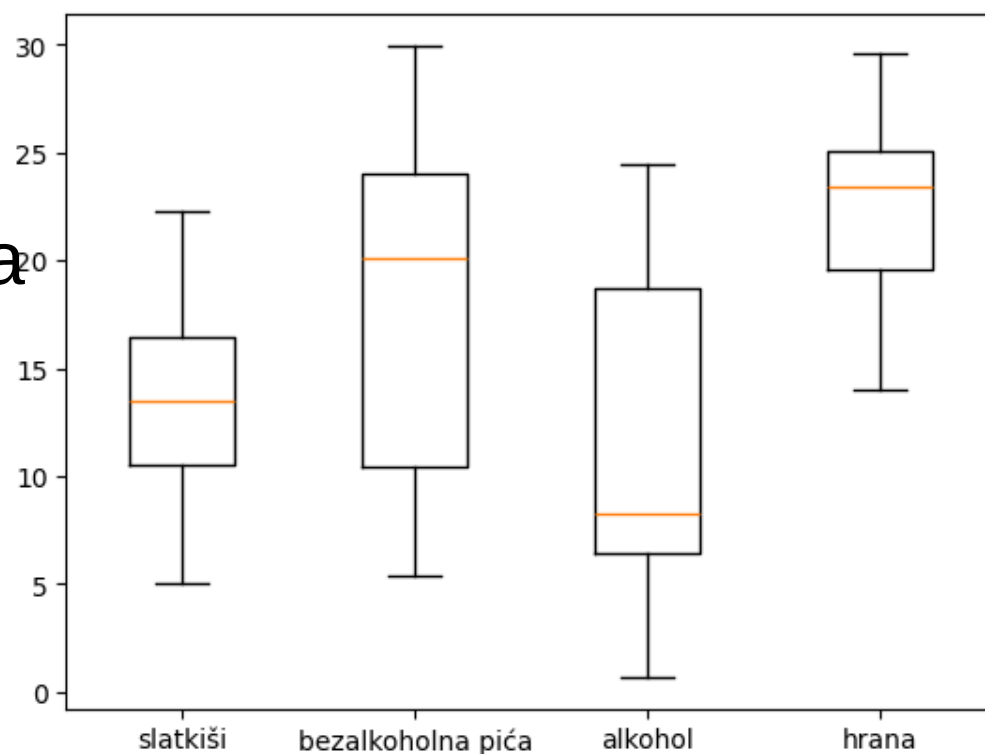
- Dva rešenja koja mogu imati smisla (3 ili 4 klastera)
- Dve promenljive
- Klasteri nisu potpuno odvojeni
- Malo realniji primer



# Primer 3

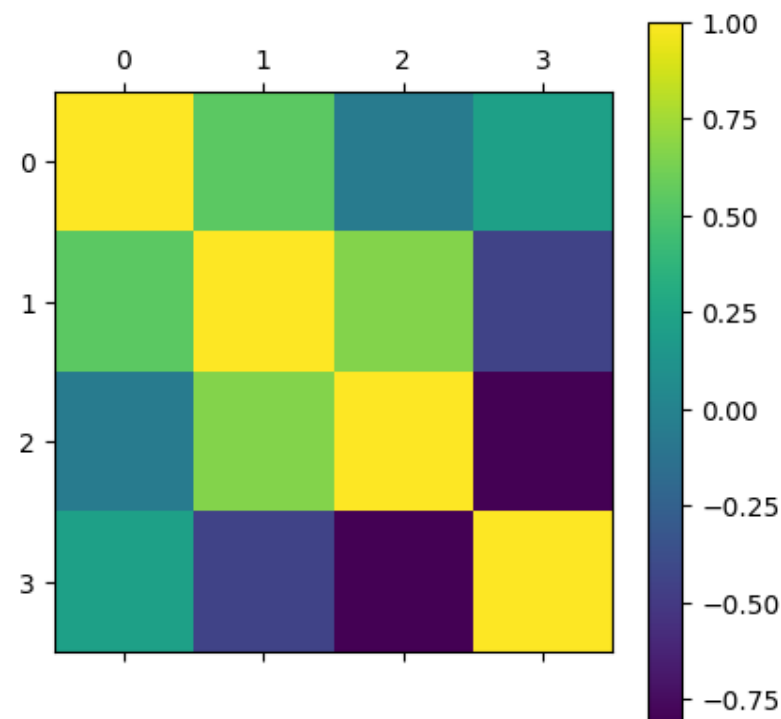


- Podaci:
  - Potrošnja novca na namirnice za više porodica
  - Slatkiši, bezalkoholna pića, alkohol, hrana
- Višedimenzionalni podaci
- Još realniji primer



# Primer 3

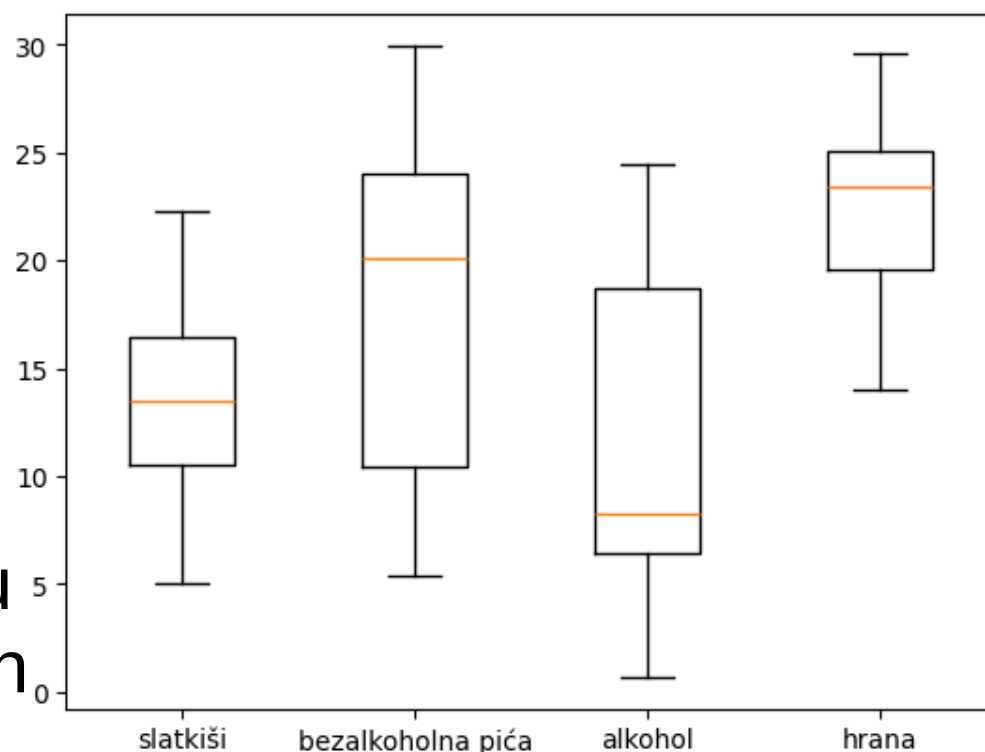
- Probamo korelaciju da vidimo da li su neke promenljive u vezi
- Nažalost, ne daje dobre rezultate
- Velika (negativna) korelacija je samo između hrane (2) i alkohola (3)



# Primer 3



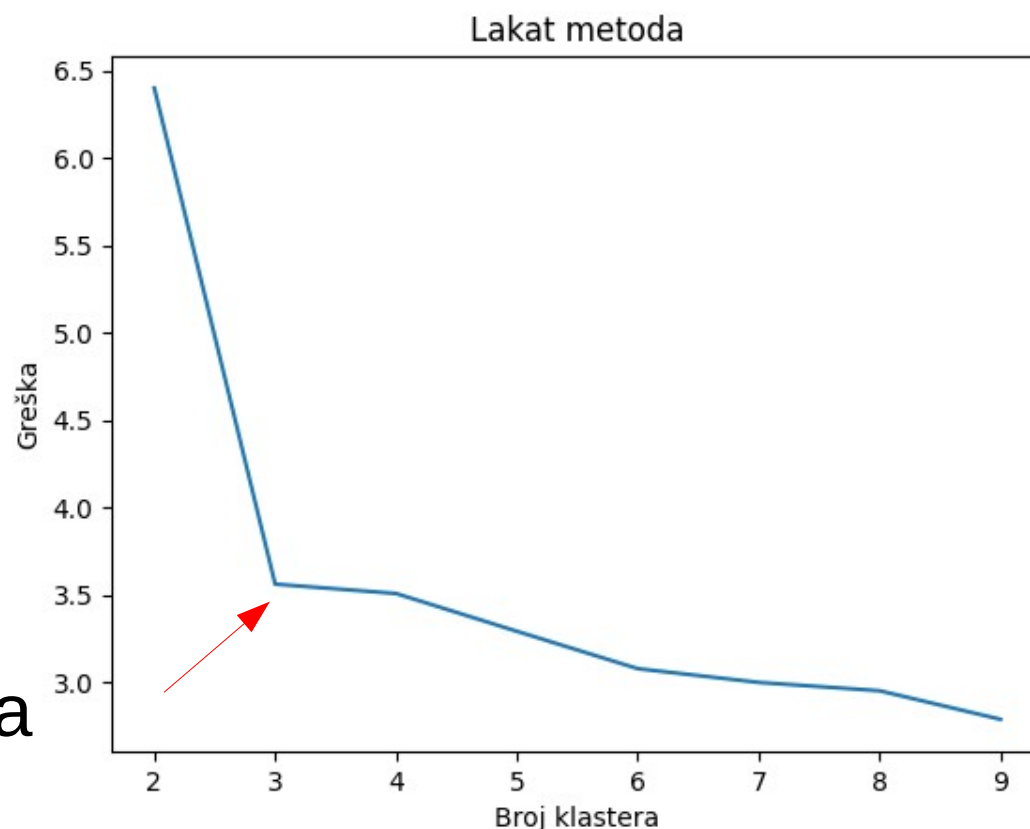
- Uradimo klasterizaciju
- Problem: klasteri se ne mogu uočiti okom
  - Nije moguće nacrtati jedan scatterplot
  - Koristi se više boxplot-ova za uvid u vrednosti promenljivih
- Koliko klastera je optimalno?





# Primer 3

- Proba se na više načina, sa npr. 2 do 9 klastera
- Za svako od rešenja se izračuna greška pri klasterizaciji
- Grafik sa greškom klasterizacije
  - „Lakat“(elbow)metoda
  - Optimalno 3 klastera



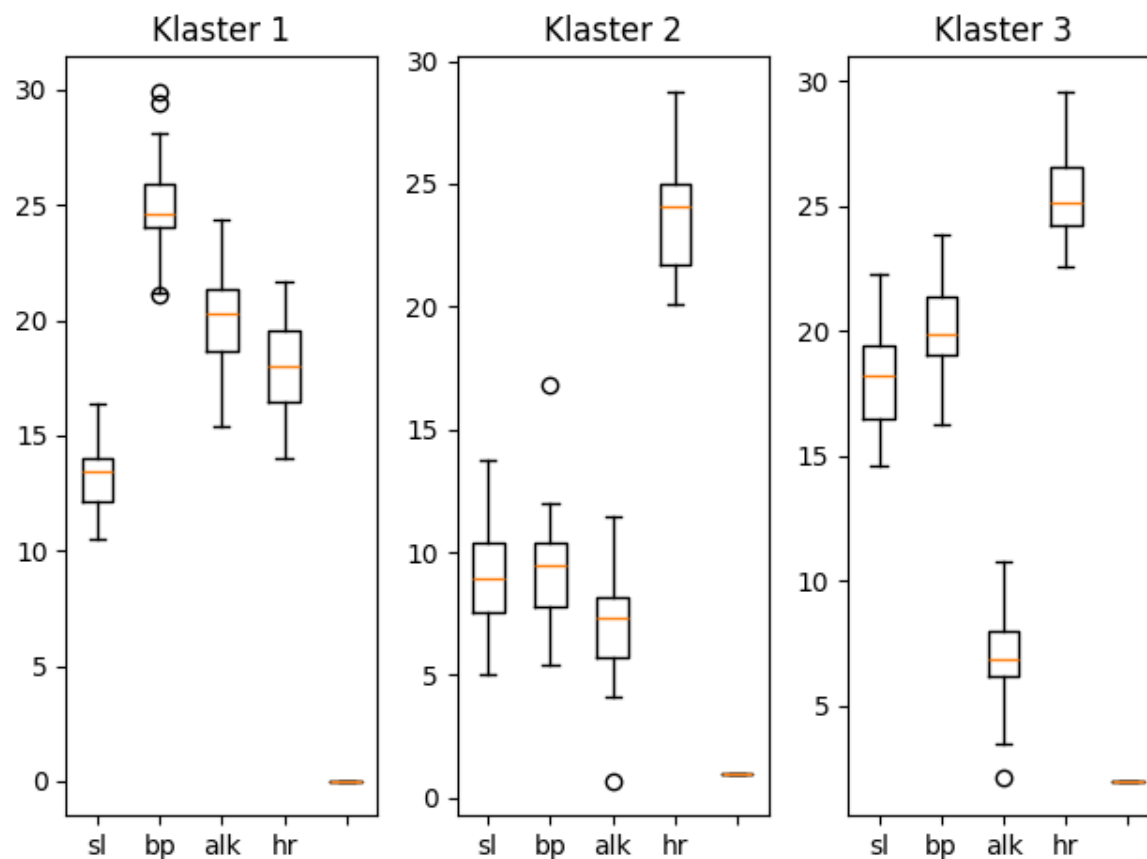
# Primer 3

- Tri klastera

1. Porodice koje puno piju (klaster 1)
2. Porodice koje se zdravo hrane (klaster 2)
3. Porodice koje vole puno da jedu i piju (klaster 3)

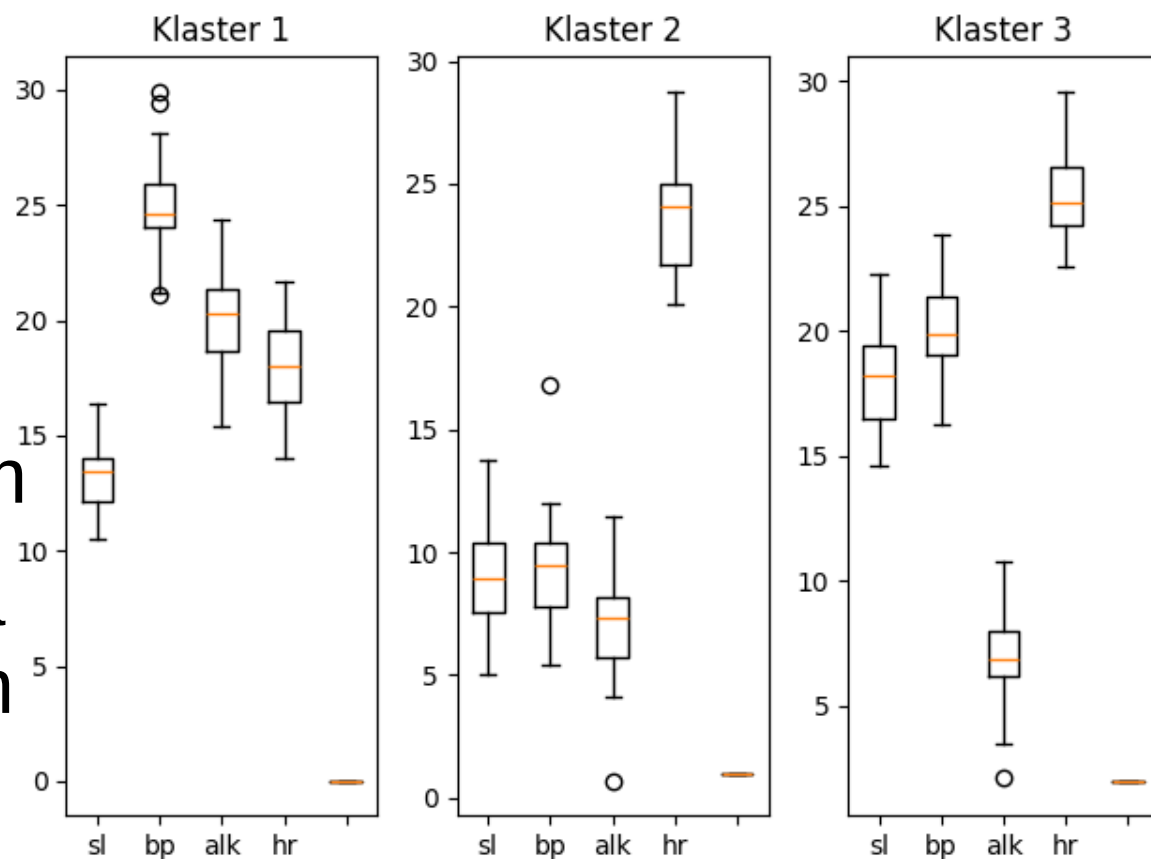
- Prikaz

- Nije moguć jedan scatterplot
- više uporednih boxplot-ova




# Primer 3

- Više promenljivih
- Više rešenja
- Klasteri nisu potpuno odvojeni
- Klasterne nije moguće uočiti okom
- Neophodna analiza klastera algoritmom
- Prilično realan primer




# Klasterizacija - karakteristike



- Eksploratorna analiza podataka
  - Tek se utvrđuje da li postoje neke grupe ili ne
  - Obično se koristi da se neko upozna sa podacima
- Nenadgledano mašinsko učenje (unsupervised)
  - Ne postoji „tačno“ ili „uzorno“ rešenje
  - Ne postoji trening set i set za validaciju

# Klasterizacija - ograničenja



- Klasterizaciju ima smisla primeniti kada su:
  - Numerički podaci u pitanju
  - Višedimenzijski podaci (bar dve promenljive)
  - Podaci neistraženi (nisu poznate klase, zakonitosti)
- Često postoje i neke pretpostavke ili iskustveni predosećaji koje želimo proveriti klasterizacijom
  - Npr. pretpostavljamo da postoje tri različite grupe gostiju u nekom restoranu



# Oblasti primene



- Segmentacija tržišta
- Uočavanje grupa u društvenim mrežama
- Identifikacija korisnika koje karakterišu slični oblici interakcije sa sadržajima nekog Web sajta/aplikacije
- Grupisanje objekata (npr., slika/dokumenata) radi lakše i efektivnije pretrage
- ...

A yellow pencil and a pink eraser are positioned in the top right corner of the white paper, suggesting a drawing or writing activity.

Kako klasterizacija funkcioniše?

# Kako klasterizacija funkcionira?

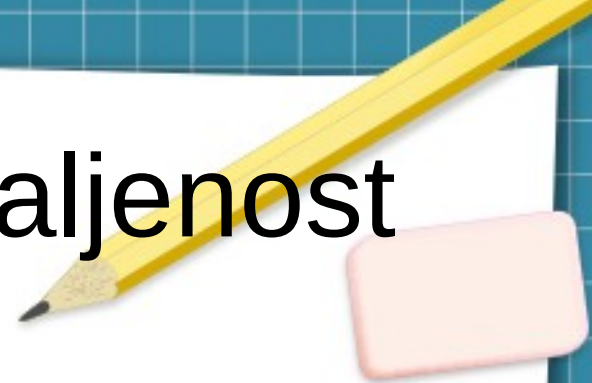


- „Slične“ instance – šta to znači?
- Pojam udaljenosti
  - Euklidska, Menhetn (city block), kosinusna...
- Metode klasterizacije
  - KMeans
  - Hijerarhijska klasterizacija
  - ...

# Šta to znači „slične“ instance?



- Sličnost se izračunava korišćenjem neke mere udaljenosti ili sličnosti:
  - Udaljenost dve instance (Euklidska ili Manhattan)
  - Sličnost dve instance (kosinusna sličnost ili koeficijent korelacije)



# aljenost

- $$d = \sqrt{\sum_{j=1}^n (x_{sj} - x_{tj})^2}$$

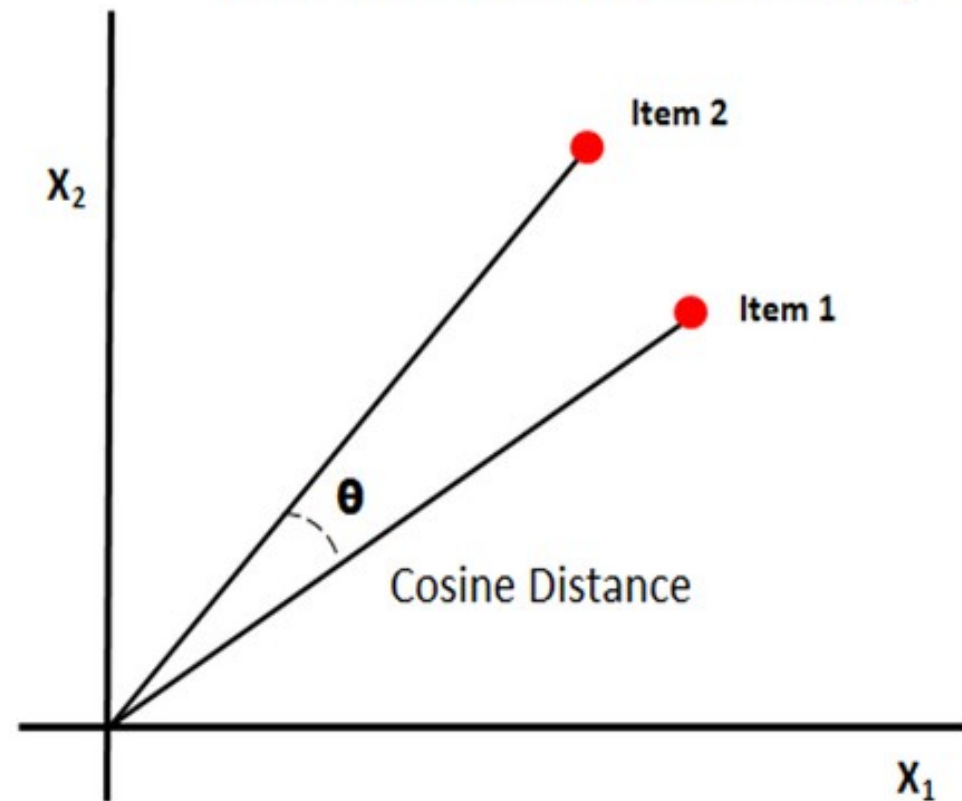
- $$d = \sum_{j=1}^n \|x_{sj} - x_{tj}\|$$

A map of Manhattan, New York, showing a route. A red line starts at a red pin in the Upper West Side, near Morningside Park, and runs south along the Hudson River. A blue line branches off from the red line near Central Park, runs east along the top of Central Park, and then runs south along the Hudson River. The route ends at a green pin in the Theater District, near Times Square. Key landmarks and streets labeled include: Hudson River, Joe DiMaggio Hwy, Broadway, Lincoln Square, Central Park, The Lake, Morningside Park, Upper West Side, Upper East Side, Central Park West, Madison Ave, Park Ave, E 96th St, E 92nd St, E 88th St, E 84th St, E 80th St, E 76th St, E 72nd St, E 68th St, E 64th St, E 60th St, E 56th St, E 52nd St, E 48th St, E 44th St, E 40th St, E 36th St, E 32nd St, E 28th St, E 24th St, E 20th St, E 16th St, E 12th St, E 8th St, E 4th St, E 1st St, W 14th St, W 18th St, W 22nd St, W 26th St, W 30th St, W 34th St, W 38th St, W 42nd St, W 46th St, W 50th St, W 54th St, W 58th St, W 62nd St, W 66th St, W 70th St, W 74th St, W 78th St, W 82nd St, W 86th St, W 90th St, W 94th St, W 98th St, W 100th St, W 104th St, W 108th St, W 112th St, W 116th St, W 120th St, W 124th St, W 128th St, W 132nd St, W 136th St, W 140th St, W 144th St, W 148th St, W 152nd St, W 156th St, W 160th St, W 164th St, W 168th St, W 172nd St, W 176th St, W 180th St, W 184th St, W 188th St, W 192nd St, W 196th St, W 200th St, W 204th St, W 208th St, W 212th St, W 216th St, W 220th St, W 224th St, W 228th St, W 232nd St, W 236th St, W 240th St, W 244th St, W 248th St, W 252nd St, W 256th St, W 260th St, W 264th St, W 268th St, W 272nd St, W 276th St, W 280th St, W 284th St, W 288th St, W 292nd St, W 296th St, W 300th St, W 304th St, W 308th St, W 312th St, W 316th St, W 320th St, W 324th St, W 328th St, W 332nd St, W 336th St, W 340th St, W 344th St, W 348th St, W 352nd St, W 356th St, W 360th St, W 364th St, W 368th St, W 372nd St, W 376th St, W 380th St, W 384th St, W 388th St, W 392nd St, W 396th St, W 400th St, W 404th St, W 408th St, W 412th St, W 416th St, W 420th St, W 424th St, W 428th St, W 432nd St, W 436th St, W 440th St, W 444th St, W 448th St, W 452nd St, W 456th St, W 460th St, W 464th St, W 468th St, W 472nd St, W 476th St, W 480th St, W 484th St, W 488th St, W 492nd St, W 496th St, W 500th St, W 504th St, W 508th St, W 512th St, W 516th St, W 520th St, W 524th St, W 528th St, W 532nd St, W 536th St, W 540th St, W 544th St, W 548th St, W 552nd St, W 556th St, W 560th St, W 564th St, W 568th St, W 572nd St, W 576th St, W 580th St, W 584th St, W 588th St, W 592nd St, W 596th St, W 600th St, W 604th St, W 608th St, W 612th St, W 616th St, W 620th St, W 624th St, W 628th St, W 632nd St, W 636th St, W 640th St, W 644th St, W 648th St, W 652nd St, W 656th St, W 660th St, W 664th St, W 668th St, W 672nd St, W 676th St, W 680th St, W 684th St, W 688th St, W 692nd St, W 696th St, W 700th St, W 704th St, W 708th St, W 712th St, W 716th St, W 720th St, W 724th St, W 728th St, W 732nd St, W 736th St, W 740th St, W 744th St, W 748th St, W 752nd St, W 756th St, W 760th St, W 764th St, W 768th St, W 772nd St, W 776th St, W 780th St, W 784th St, W 788th St, W 792nd St, W 796th St, W 800th St, W 804th St, W 808th St, W 812th St, W 816th St, W 820th St, W 824th St, W 828th St, W 832nd St, W 836th St, W 840th St, W 844th St, W 848th St, W 852nd St, W 856th St, W 860th St, W 864th St, W 868th St, W 872nd St, W 876th St, W 880th St, W 884th St, W 888th St, W 892nd St, W 896th St, W 900th St, W 904th St, W 908th St, W 912th St, W 916th St, W 920th St, W 924th St, W 928th St, W 932nd St, W 936th St, W 940th St, W 944th St, W 948th St, W 952nd St, W 956th St, W 960th St, W 964th St, W 968th St, W 972nd St, W 976th St, W 980th St, W 984th St, W 988th St, W 992nd St, W 996th St, W 1000th St, W 1004th St, W 1008th St, W 1012th St, W 1016th St, W 1020th St, W 1024th St, W 1028th St, W 1032nd St, W 1036th St, W 1040th St, W 1044th St, W 1048th St, W 1052nd St, W 1056th St, W 1060th St, W 1064th St, W 1068th St, W 1072nd St, W 1076th St, W 1080th St, W 1084th St, W 1088th St, W 1092nd St, W 1096th St, W 1100th St, W 1104th St, W 1108th St, W 1112th St, W 1116th St, W 1120th St, W 1124th St, W 1128th St, W 1132nd St, W 1136th St, W 1140th St, W 1144th St, W 1148th St, W 1152nd St, W 1156th St, W 1160th St, W 1164th St, W 1168th St, W 1172nd St, W 1176th St, W 1180th St, W 1184th St, W 1188th St, W 1192nd St, W 1196th St, W 1200th St, W 1204th St, W 1208th St, W 1212th St, W 1216th St, W 1220th St, W 1224th St, W 1228th St, W 1232nd St, W 1236th St, W 1240th St, W 1244th St, W 1248th St, W 1252nd St, W 1256th St, W 1260th St, W 1264th St, W 1268th St, W 1272nd St, W 1276th St, W 1280th St, W 1284th St, W 1288th St, W 1292nd St, W 1296th St, W 1300th St, W 1304th St, W 1308th St, W 1312th St, W 1316th St, W 1320th St, W 1324th St, W 1328th St, W 1332nd St, W 1336th St, W 1340th St, W 1344th St, W 1348th St, W 1352nd St, W 1356th St, W 1360th St, W 1364th St, W 1368th St, W 1372nd St, W 1376th St, W 1380th St, W 1384th St, W 1388th St, W 1392nd St, W 1396th St, W 1400th St, W 1404th St, W 1408th St, W 1412th St, W 1416th St, W 1420th St, W 1424th St, W 1428th St, W 1432nd St, W 1436th St, W 1440th St, W 1444th St, W 1448th St, W 1452nd St, W 1456th St, W 1460th St, W 1464th St, W 1468th St, W 1472nd St, W 1476th St, W 1480th St, W 1484th St, W 1488th St, W 1492nd St, W 1496th St, W 1500th St, W 1504th St, W 1508th St, W 1512th St, W 1516th St, W 1520th St, W 1524th St, W 1528th St, W 1532nd St, W 1536th St, W 1540th St, W 1544th St, W 1548th St, W 1552nd St, W 1556th St, W 1560th St, W 1564th St, W 1568th St, W 1572nd St, W 1576th St, W 1580th St, W 1584th St, W 1588th St, W 1592nd St, W 1596th St, W 1600th St, W 1604th St, W 1608th St, W 1612th St, W 1616th St, W 1620th St, W 1624th St, W 1628th St, W 1632nd St, W 1636th St, W 1640th St, W 1644th St, W 1648th St, W 1652nd St, W 1656th St, W 1660th St, W 1664th St, W 1668th St, W 1672nd St, W 1676th St, W 1680th St, W 1684th St, W 1688th St, W 1692nd St, W 1696th St, W 1700th St, W 1704th St, W 1708th St, W 1712th St, W 1716th St, W 1720th St, W 1724th St, W 1728th St, W 1732nd St, W 1736th St, W 1740th St, W 1744th St, W 1748th St, W 1752nd St, W 1756th St, W 1760th St, W 1764th St, W 1768th St, W 1772nd St, W 1776th St, W 1780th St, W 1784th St, W 1788th St, W 1792nd St, W 1796th St, W 1800th St, W 1



# Kosinusna sličnost

- Kosinus ugla između vektora povučениh od koordinatnog početka do dve instance (tačke)
- Kosinusna sličnost
- Kosinusna udaljenost (1-kosinusna slič.)



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

# Procena rezultata klasterizacije



- Ocena uspešnosti modela je dosta teža nego kod nadgledanog mašinskog učenja
- Ovde nemamo precizne metrike koje nedvosmisleno ukazuju na to koliko je model “dobar”

# Procena rezultata klasterizacije



- Pod “dobrim” rešenjem se podrazumeva model koji:
  - Dobro deli instance u nepreklapajuće grupe (klasterne) (objektivna procena)
  - Koristan je za dati zadatak / problem zbog koga se klasterovanje i radi (subjektivna procena)

# Procena rezultata klasterizacije



- Neki od objektivnih kriterijuma za procenu kvaliteta klastera:
  - Međusobna udaljenost težišta
    - što su težišta dalje jedno od drugog, to je stepen preklapanja klastera manji, i njihov kvalitet viši
  - Max udaljenost instanci u okviru istog klastera
  - Min udaljenost instanci iz različitih klastera
  - Suma kvadrata unutar klastera
    - suma kvadrata odstupanja instanci u okviru klastera od težišta klastera
  - *Veličina svakog klastera (broj instanci)?*

# Procena rezultata klasterizacije



- Problem: ne postoje metrike koje ukazuju na to koliko je neko rešenje sveukupno dobro, odnosno korisno za dati zadatak
- Subjektivna procena korisnosti klastera za dati domen i zadatak je značajnija od opisanih objektivnih metrika
- Domensko znanje presudno za evaluaciju, tj. izbor optimalnog skupa klastera



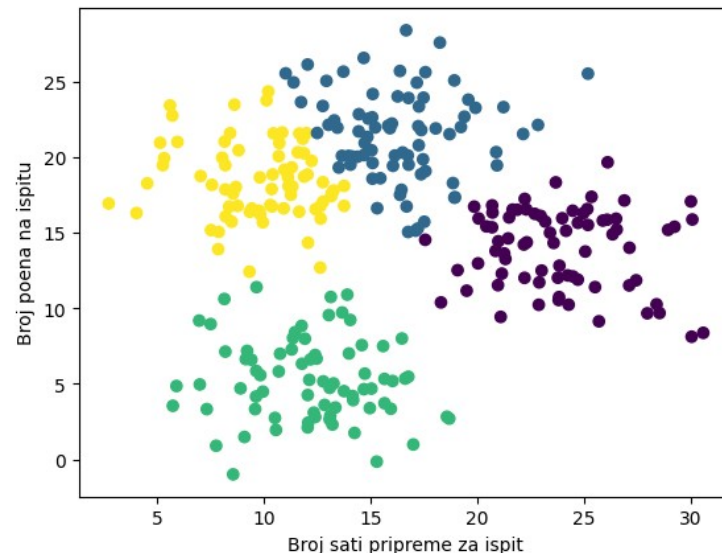
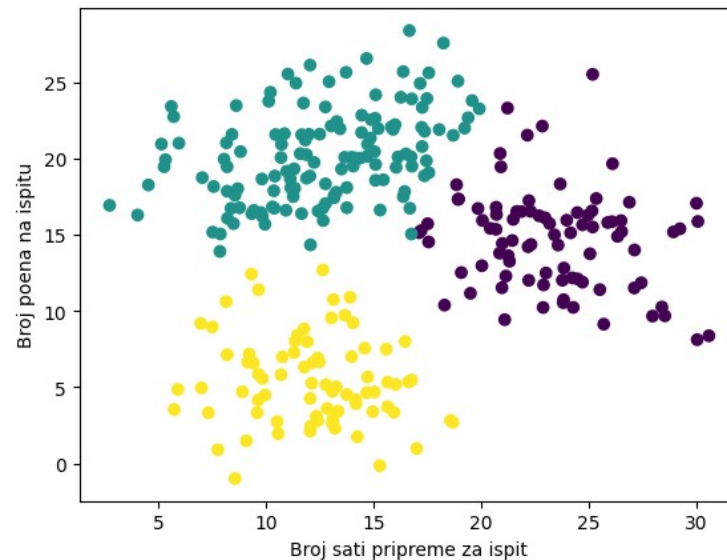
# Procena rezultata klasterizacije

- Rešenje sa tri klastera ili

- 1) Malo učili, slabo uradili (žuta)
- 2) Malo ili srednje učili, dobro uradili (zelena)
- 3) Puno učili, osrednje uradili (ljubičasta)

- Rešenje sa četiri klastera?

- 1) Malo učili, slabo uradili (zelena)
- 2) Malo učili, dobro uradili (žuta)
- 3) Srednje učili, dobro uradili (plava)
- 4) Puno učili, osrednje uradili (ljubičasta)



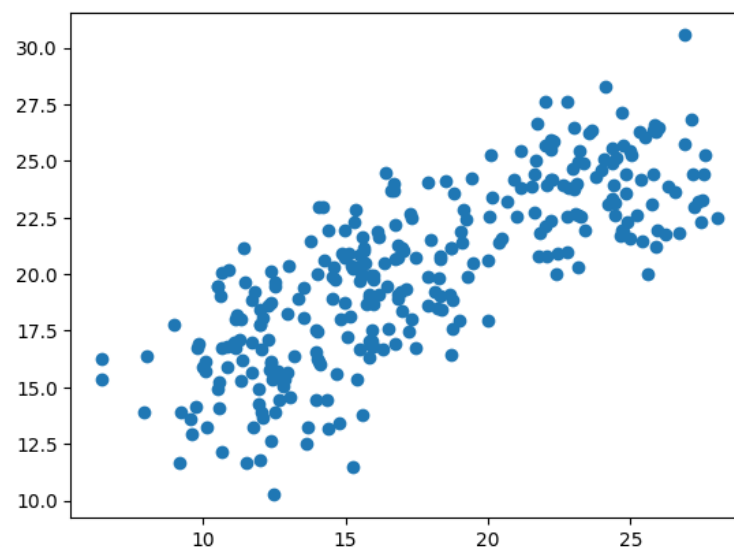
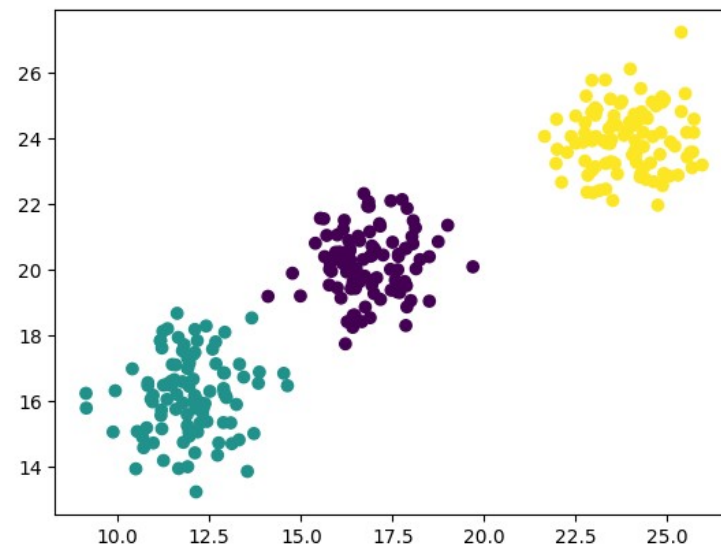
# Klasterizacija - problemi



- Nedostajući podaci (NaN)
  - Zbog lošeg merenja, dizajna istraživanja, više sile...
  - Nije moguće odrediti udaljenost instance od drugih
- Moguća rešenja (prednosti i mane):
  - Rad sa parcijalnim skupom podataka
    - Izbacivanje celih instanci (redova) sa NaN
    - Izbacivanje promenljivih koje imaju mnogo NaN
  - Ubacivanje vrednosti umesto NaN
    - Prosečna vrednost (mean) ili medijana umesto NaN
    - Imputacija vrednosti

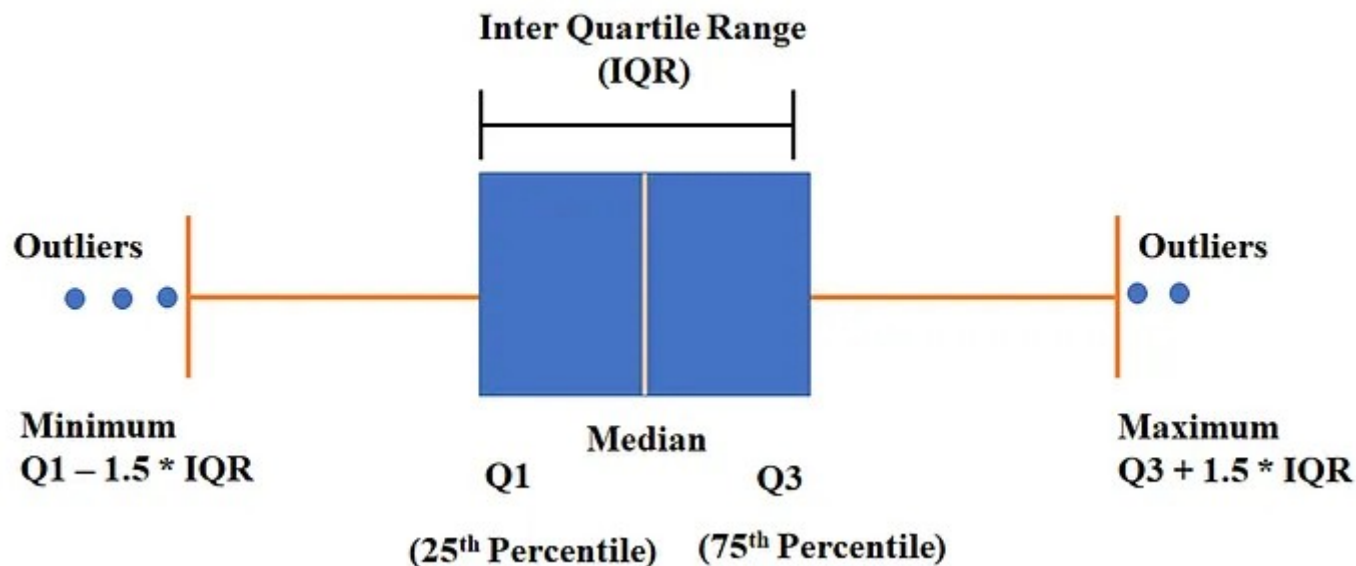
# Klasterizacija - problemi

- Korelacija promenljivih
  - Ako je par promenljivih visoko korelisan, to utiče na udaljenost (može da bude problem)
  - U ekstremnom slučaju, dobija se samo jedan klaster
- Moguće rešenje:
  - Izbacivanje jedne visoko korelisane promenljive (iz svakog para) iz analize



# Klasterizacija - problemi

- Netipične/ekstremne vrednosti (eng. „outliers“)
  - Ekstremna vrednost neke promenljive u nekoj instanci može da „povuče“ ceo klaster na neku stranu





# Klasterizacija - problemi



- Moguće rešenje:
  - Izbacivanje promenljive (ako ima npr.  $> 10\%$  outlier-a)
  - Zamena ekstremnih vrednosti nekim drugim (winsorize/winsorization metoda)
    - Obično se biraju percentili kao donje i gornje granice (npr. 5% i 95% ako ima outliera na obe strane)
    - Pronađu se vrednosti iz skupa podataka koje odgovaraju tim percentilima.
    - Outlier-i koji su iznad se zamene vrednošću 95% percentila
    - Outlier-i koji su ispod se zamene vrednošću 5% percentila
    - Ponovo se proveriti da li ima outlier-a i, ako ih ima, ponovi se ceo proces sa drugim percentilima



# Klasterizacija - problemi



- Različite skale promenljivih
  - Ako je jedna promenljiva u rasponu od 1 do 100 a druga od 0 do 1, vrednost prve promenljive će dominantno uticati na ukupnu udaljenost.
- Moguće rešenje:
  - Normalizacija (svođenje na skalu 0 do 1).

# Klasterizacija - postupak



1) Učitavanje podataka

2) Inicijalni izbor promenljivih (objektivan i subjektivan)

3) Priprema podataka

1) Provera nedostajućih vrednosti (NaN)

- Ako ih ima, izbacivanje celih instanci ili zamena nedostajućih vrednosti (više načina)

2) Provera korelacije promenljivih

- Ako ima korelacije, izbacivanje po jedne promenljive iz svakog para

3) Provera ekstremnih vrednosti promenljivih (outliers)

- Ako ih ima, izbacivanje celih instanci ili zamena ekstremnih vrednosti (više načina)

4) Provera raspona vrednosti promenljivih

- Ako su rasponi različiti (ili u svakom slučaju) uraditi normalizaciju

# Klasterizacija - postupak



4) Izbor metode klasterizacije i parametara

5) Izvršavanje izabrane metode klasterizacije

6) Procena rezultata klasterizacije

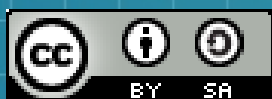
1) Procena prema objektivnim kriterijumima

- Međusobna udaljenost težišta, obično metodom sume kvadrata unutar klastera

2) Procena prema subjektivnim kriterijumima

- Koliko dobijeni klasteri imaju smisla (na osnovu prethodnog znanja i iskustva)

3) Vraćanje na korake 4 i 5 ako rezultati procene nisu dobri (druga metoda i/ili parametri)



This work is licensed under a Creative Commons  
Attribution-ShareAlike 3.0 Unported License.  
It makes use of the works of Mateus Machado Luna.

