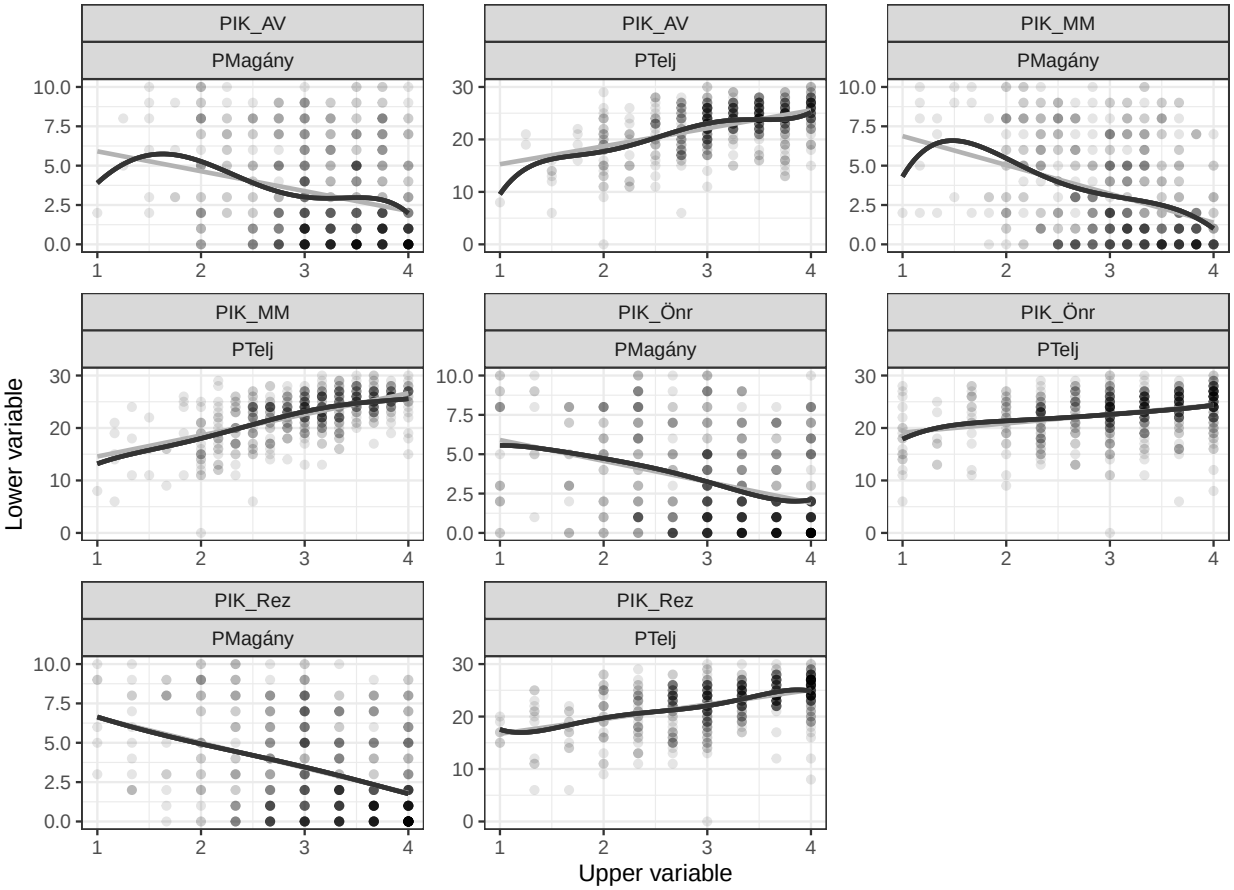
Házi feladatok megoldása 5. Klaszterelemzés

Smahajcsik-Szabó Tamás, M9IJYM

1. Keress nemlineáris kapcsolatot a PIK 4 skálája és a PERMA Magányosság (Pmagány) és Teljesítmény (Ptelj) skálája között!

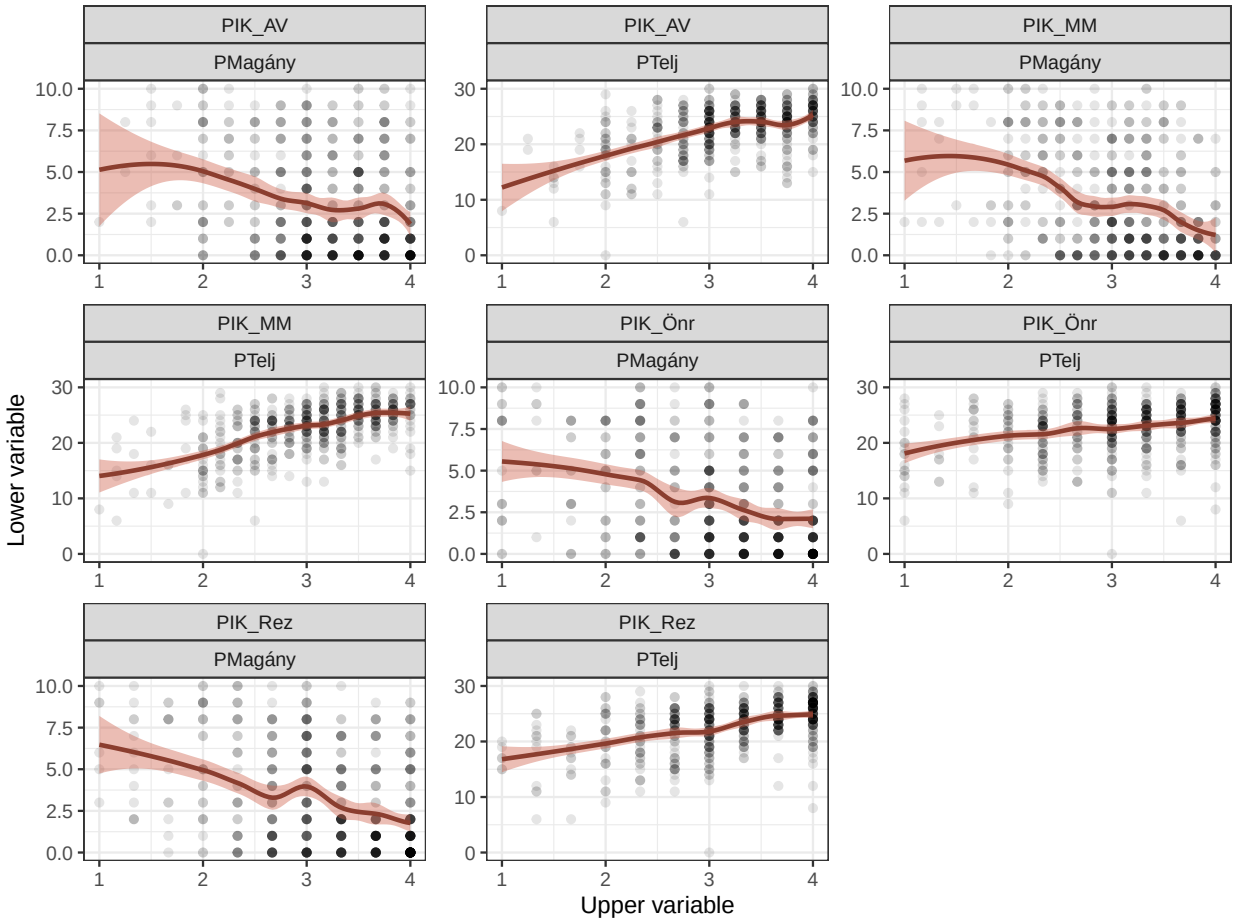
Az alábbi ábrán egy 1-4 fok közötti polinomiális trendvonalat illesztettem az adatokra. A legvilágosabb vonal lineáris trendvonal, míg a 2-4 fokok közti vonalak jelentős átfedéssel a sötét trendvonalak. További elemzések során, lokálisan súlyozott regressziós elemzést is futtattam (LOWESS), mely hasonlóképpen, elsősorban a Magányosság tételelinél rajzolt ki nem lineáris kapcsolatokat.



Mint a lentebbi táblázatból látható, a megmagyarázott variancia (adjusted R squared, ARS) értéke páronkénti polinomiális regressziók esetében maximum értékét nagyobb mint 1 fok esetén éri el több változópár esetében is. A fok értékét 5-ben maximalizáltam, és a legtöbb esetben a harmadfokú polinomiális komponens eredményezi a legmagasabb megmagyarázott variancia arányt. Magasabb értékeknél (maximális fok 8) elltűnik a PIK-AV és a Teljesítmény esetén a lineáris kapcsolat, és egy nem lineáris trendvonal jobb magyarázó erővel bír.

x	y	ARS	degree
PIK_MM	PTelj	0.3365856	3
PIK_MM	PMagány	0.1472861	3
PIK_AV	PTelj	0.2396300	1
PIK_AV	PMagány	0.0655383	3
PIK_Önr	PTelj	0.1026863	3
PIK_Önr	PMagány	0.1237704	5
PIK_Rez	PTelj	0.2363600	3
PIK_Rez	PMagány	0.1679531	5

Alább a lokálisan súlyozott smoother trendvonalakat közlöm konfidencia régiókkal.



Összességében a Magányosság tétel PIK-tételekkel való kapcsolataiban mutatkozik nem lineáris kapcsolat mind polinomiális, mint LOWESS trendvonalak tekintetében.

2. Számítsd ki Vargha (2021) alapján a 4.2. ábra B és C személyének távolságát a 4.1. táblázat összes távolságtípusára!

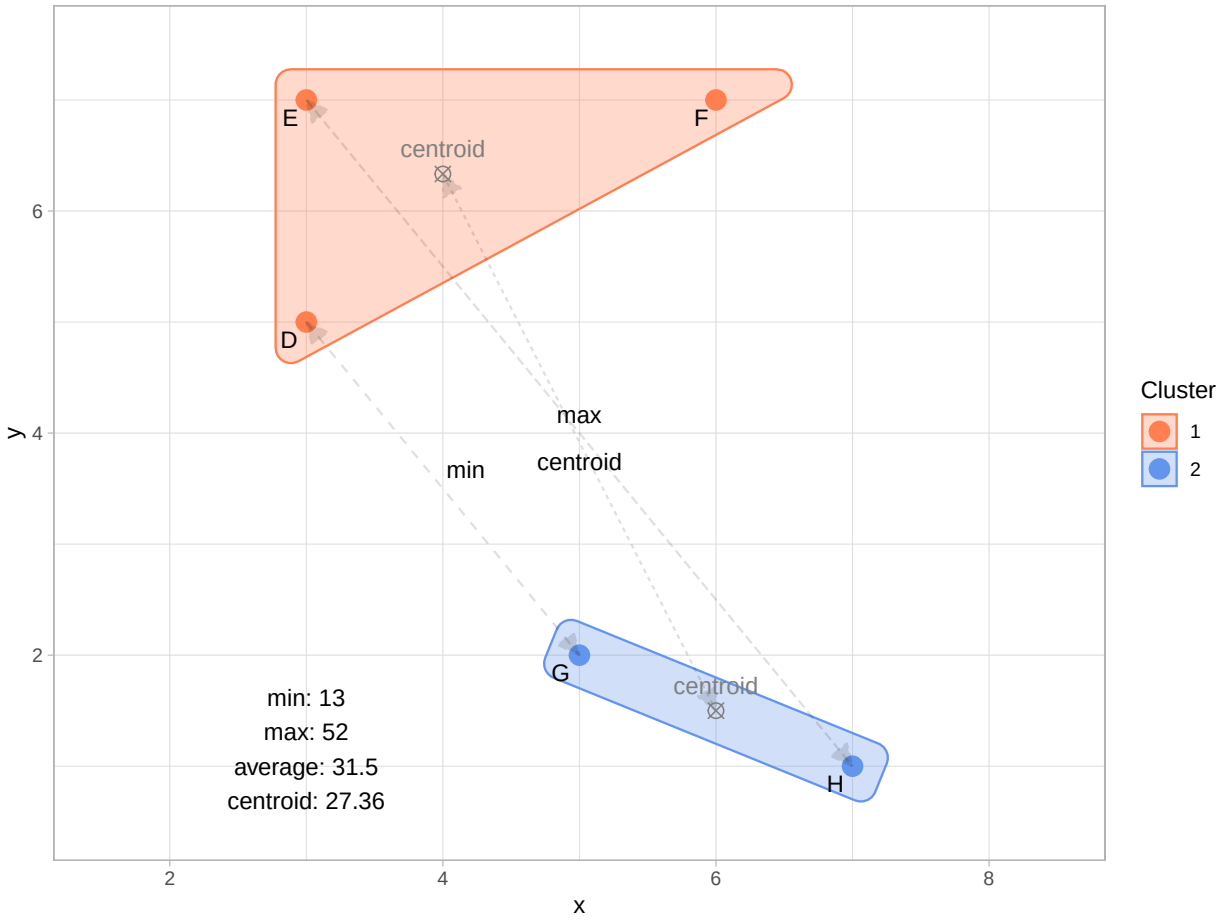
Az alábbi táblázatban összegeztem a távolságértékeket típus szerint.

Távolság	Típus	Képlet
2.500	ASED	((3 - 5)^2 + (3 - 2)^2 + (3 - 2)^2 + (2 - 4)^2)/4 = 2.5
10.000	SED	(3 - 5)^2 + (3 - 2)^2 + (3 - 2)^2 + (2 - 4)^2 = 10
3.162	ED	SQRT((3 - 5)^2 + (3 - 2)^2 + (3 - 2)^2 + (2 - 4)^2) = 3.162
6.000	Manhattan	(|3 - 5| + |3 - 2| + |3 - 2| + |2 - 4|) = 6
2.000	Csebisev	Max(|3 - 5|, |3 - 2|, |3 - 2|, |2 - 4|) = 2
1.333	Pearson	1 - (-0.333)

3. Számítsd ki a 4.4. ábrán látható KL2 és KL3 klaszter távolságát a 4.2. táblázat 1., 2., 3. és 5. távolságtípusára, ha a személytávolságra a SED távolságot használjuk!

Alább először táblázatosan közlöm, majd grafikonon ábrázolom a SED távolságokat a KL2 és KL3 klaszterek között.

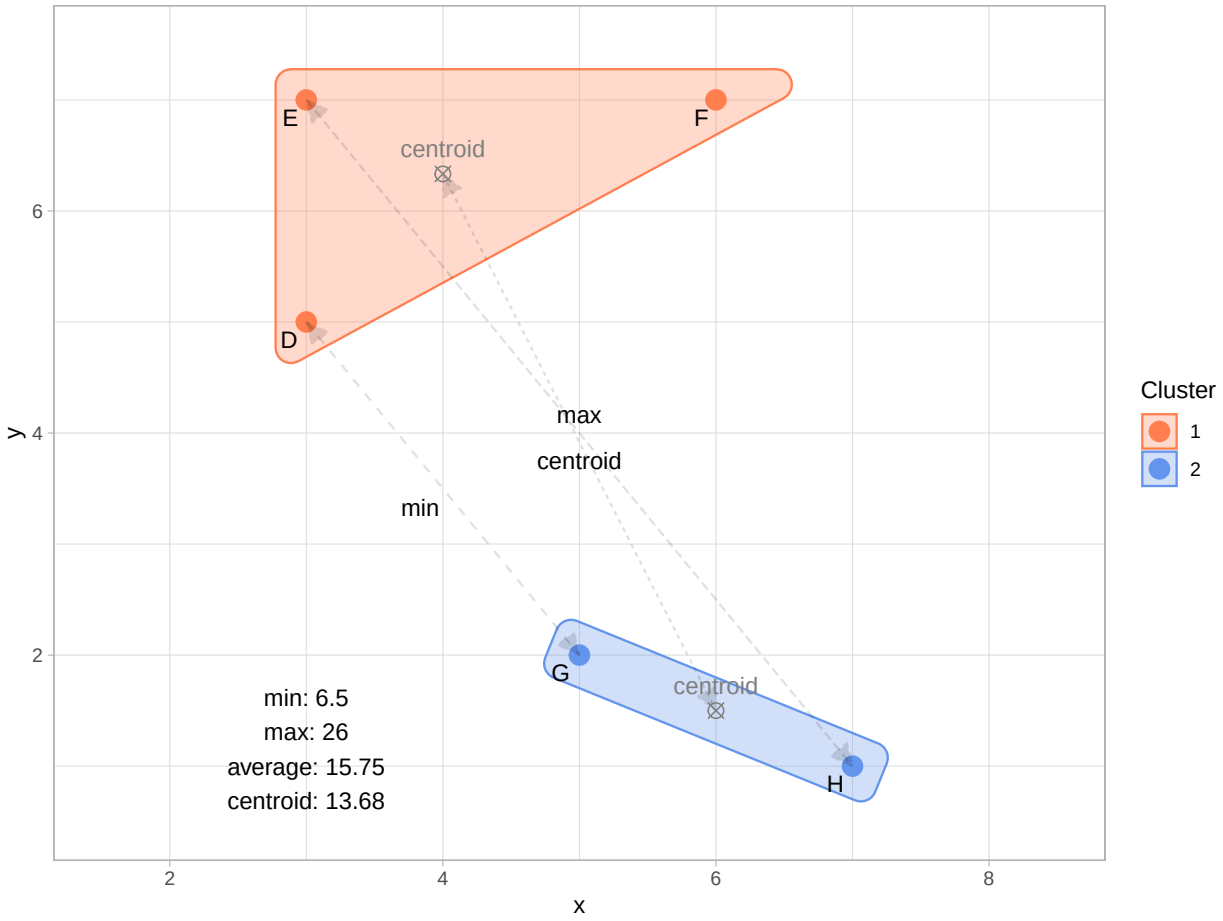
method	d
min	13.00
max	52.00
average	31.50
centroid	27.36



4. Számítsd ki a 4.4. ábrán látható KL2 és KL3 klaszter távolságát a 4.2. táblázat 1., 2., 3. és 5. távolságtípusára, ha a személytávolságra az ASED távolságot használjuk!

Alább először táblázatosan közlöm, majd grafikonon ábrázolom az ASED távolságokat a KL2 és KL3 klaszterek között.

method	d
min	6.50
max	26.00
average	15.75
centroid	13.68



5. A ROPstat Relokáció modulja segítségével készíts 3, 4 és 5 klaszteres megoldást a Ptelj és Pmagány változóra standardizálással és hasonlítsd össze e megoldásokat a 4.4. táblázat adekvációs mutatói segítségével! Melyik klaszterstruktúra tűnik a legjobbnak?

A standardizálást követően képzett klaszterek adekvációs mutatóit az alábbi táblázat foglalja össze. A megmagyarázott varianciaarány (EESS%) alapján minél több klasztert képzünk a k-központú elemzés során, az arány emelkedik, hiszen természetszerűleg közelítunk az adatok eleve adott változatosságához. A pontbiszeriális mutatók tekintetében azonban a k=3 struktúra képvisel egy kicsivel jobb, homogénebb konfigurációt. Ezt árnyalja a CLdelta érték, mely k=4 esetén jelez nagyobb fokú szeparációt a különböző klaszterbe tartozók távolságának arányával az azonosba tartozókhoz képest. Hasonló mintázatot tükröz a Silhouette mutató is, 4 klaszternél a személyek átlagosan közelebb vannak a saját klaszter centrumához mint a legközelebbi idegen klaszteréhez; de erre utal hasonló mintázatával a módosított Xien-Beni index is. A k=4 scenario fölényét

támasztja alá a GDI24 is, mely az ideális > 1,5 értékét ekkor veszi fel. A HCÁtlag szerint is k=4 egy homogén struktrát eredményez. Noha k=5 esetében a homogenitás kedvezőbb, az összes egyéb mutató (különösen a GDI24) szerint alulmarad az öt klaszteres megoldás a néggyel szemben.

k	EESS%	PB	XBmod	SH	HCÁtlag	CLdelta	GDI24	HCmin-max
3	64,96	0,51	0,792	0,769	0,705	1,041	1,004	0,43-1,67
4	74,59	0,497	0,845	0,782	0,512	1,047	1,572	0,33-1,04
5	78,94	0,459	0,743	0,759	0,426	1,012	0,783	0,30-1,04

Amolyan függelékként alább táblázatban összegzek további adekvációs mutatókat a klaszterek száma szerint oszlopokba rendezve.