Házi feladatok megoldása 6. Hierarchikus klaszterelemzés

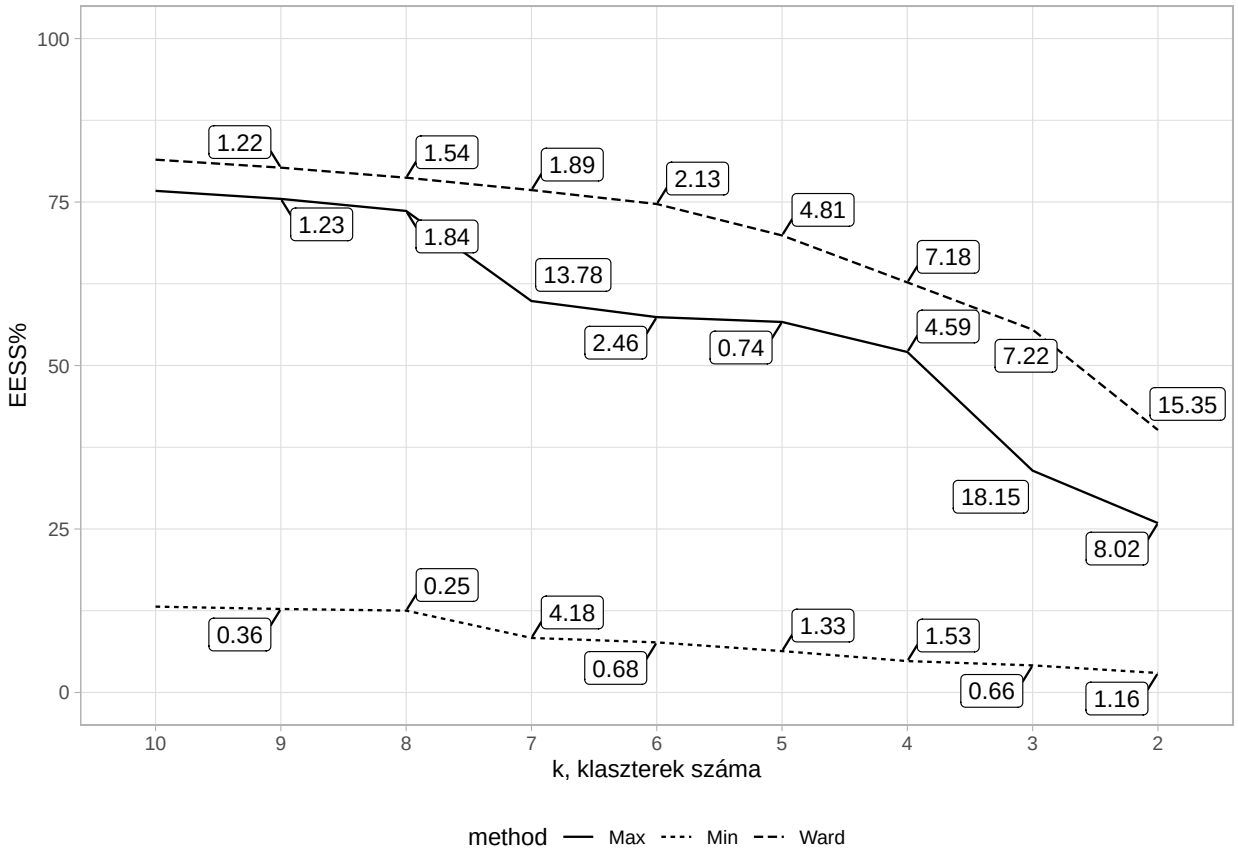
Smahajcsik-Szabó Tamás, M9IJYM

1. Végezz HKA-t ROPstatban a minimális távolság módszerrel a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel, standardizálással! Foglald táblázatba a QC-ket k=2 és k=10 között! Hány klaszteres megoldás tűnik a legjobbnak?

A minimális távolság módszerével végzett hierarchikus klaszeterelemzés eredményét az alábbi táblázat foglalja össze k=2 és k = 10 között.

KL#	Ess_increase	EESS%	Pontbisz	XBmod	Sil.eh.	HC_avg	HCmin-HCmax	Clusters
10	87.268	13.13	0.455	-0.341	0.448	1.738	0,00-1,78	1 (489) 7 (1)
9	5.415	12.77	0.470	-0.344	0.525	1.745	0,00-1,78	1 (490) 263 (1)
8	3.759	12.52	0.493	-0.346	0.645	1.750	0,00-1,78	1 (491) 258 (3)
7	62.567	8.34	0.401	-0.376	0.638	1.833	0,00-1,86	1 (494) 55 (1)
6	10.275	7.66	0.409	-0.381	0.687	1.847	0,00-1,87	1 (495) 134 (1)
5	19.781	6.33	0.380	0.302	0.697	1.873	0,00-1,89	1 (496) 157 (1)
4	23.016	4.80	0.333	0.421	0.714	1.904	0,00-1,92	1 (497) 185 (1)
3	9.850	4.14	0.363	0.835	0.767	1.917	0,00-1,92	1 (498) 192 (1)
2	17.317	2.98	0.381	0.935	0.889	1.940	0,00-1,94	1 (499) 443 (1)

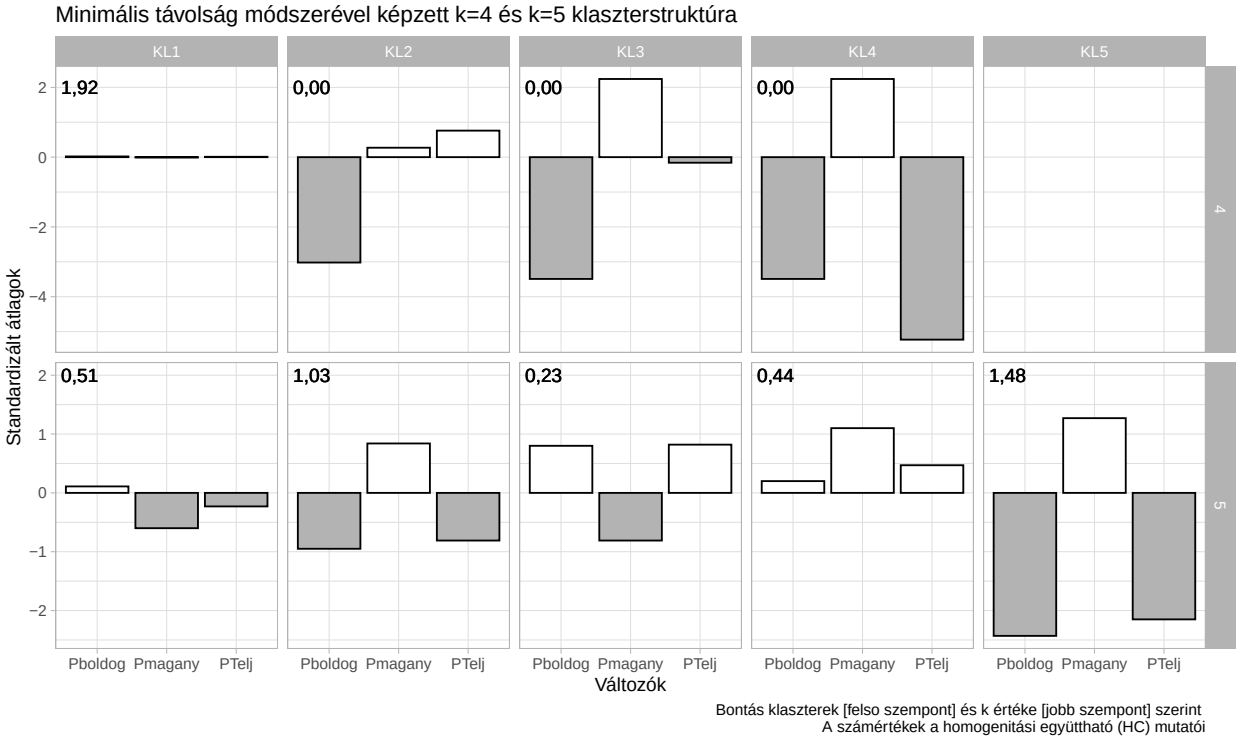
Továbbá a jelen, a 3. és az 5. feladatra vonatkozóan, az alábbi ábra az EESS% összarányában, illetve a minimális, a maximális és a Ward klaszterösszevonási módszerek szerinti lebontásban mutatja, miként változik k értékével az EESS% értéke is.



A minimális távolság módszere, mint klaszterképzési alap, nem kielégítő módszer, a 2 és 10 közötti klaszterszámban a megmagyarázott variancia következetesen rossz, így optimális megoldást ebben a k-tartományban, ezzel a módszerrel nem tudunk képezni. Az egyéb adekvációs mutatók, így a Xien-beni módosított együtthatója, a pontbiszeriális együttható és a Silhouette együttható mintázata alapján a k=4 és k=5 megoldás tűnik jobbnak, az egységesen rossz megoldások küzül.

2. Készíts ábrát az előző feladat legjobb megoldásának centroidjai alapján a sima és a standardizált átlagokra!

Az alábbi ábra a klaszterek (felső szempont) és k értékei (jobb szempont, esetünkben 2 elemzés) mentén mutatja be a standardizált átlagok szintjeit a három változó mentén.



Az értelmezhetőség alapján az 5 klaszteres megoldás tűnik megfelelőbbnek, de fontos tudni, hogy alacsony a megmagyarázott variancia a struktúra hátterében. A k=4 elemzés során (felső ábrasor), keletkezett egy átlagos csoport (KL1), amely minden értékben az átlagra illeszkedik (0), noha HC szerint kevésbé homogén klaszter. Adott egy KL2 kalszter mely boldogtalan, de egyébként enyébben magányos, és jobban teljesítő személyeket sűrít. A KL3 klaszterben magányos, boldogtalan és rosszul teljesítők állnak, míg a KL4 az előbbi klaszterhez hasonló, de kifejezetten teljesíteni képtelen boldogtalan, magányosokat foglal magában. KL3 és KL4 mintázata hasonló, a struktrúra szakmailag kérdéseket vet fel. A k=5 esetében már érdekesebb a kép. Adott egy KL2 és KL3 klaszter, melyek mintázata egymás ellentéte: KL2-ben magányos, boldogtalan és alulteljesítők állnak, KL3-ban pedig jól teljesítő, nem magányos, boldogabb személyek. Adott továbbá egy KL4, ahol magányosságuk ellenére jobban teljesítő és boldogabb személyek állnak. Szakmailag érdekes klaszter. A KL1-ben kevésbé magányos, viszonylag boldog, de (talán egyéb okból?) rosszabbul teljesítők állnak, míg a KL5 a korábbi k=4 struktuára jellegzetes depresszívjeit fogja közre.

3. Végezz HKA-t ROPstatban a maximális távolság módszerrel a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel, standardizálással! Foglald táblázatba a QC-ket k=2 és k=10 között! Hány klaszteres megoldás tűnik a legjobbnak?

Az alábbi táblázat összegzi a maximális távolság módszerével képzett kalszterestruktúrákat.

KL#	Ess_increase	EESS%	Pontbisz	XBmod	Sil.eh.	HC_avg	HCmin-HCmax	Clusters
10	5.629	76.71	0.461	0.654	0.630	0.475	0,00-0,79	3 (24) 42 (13)
9	18.419	75.48	0.462	0.662	0.624	0.499	0,00-0,79	15 (10) 54 (18)
8	27.602	73.64	0.461	0.637	0.615	0.536	0,00-1,21	1 (87) 2 (258)
7	206.197	59.86	0.513	0.447	0.535	0.811	0,00-1,21	9 (26) 17 (49)
6	36.923	57.40	0.518	0.518	0.524	0.860	0,00-1,21	15 (28) 443 (1)
5	11.059	56.66	0.517	0.510	0.525	0.875	0,78-1,43	3 (37) 9 (75)
4	68.680	52.07	0.522	0.553	0.564	0.965	0,79-1,43	1 (345) 3 (112)
3	271.695	33.92	0.658	0.547	0.697	1.326	0,79-1,43	1 (457) 7 (14)
2	120.086	25.90	0.633	0.844	0.774	1.485	1,43-1,49	1 (471) 15 (29)

A kedvező adekvációs mutatók (például Pontbiszeriális együttható, Xien-Beni és Silhouette együttható értékei mind 0.5 felettiek) mellett a k=4 struktúrát támogatja az, hogy k=3-ra lépés során egy jelentősebb 18.15% pontnyi EESS% esés következik be. Ennek alapján a négyklaszteres megoldást erősíti a maximális távolság módszere.

4. Készíts ábrát az előző feladat legjobb megoldásának centroidjai alapján a sima és a standardizált átlagokra!

A leghomogénebb KL1 kalszter boldog, kevébé magányos, jól teljesítő személyeket foglal magába. A KL2 klaszter a magas magányossága ellenére boldog és jól teljesítők klasztere, míg a KL3-4 hasonló mintázatú, csak fokozatában eltérő heterogén klaszterek, melyekbe a magányos, szomorú és rosszul teljesítők tartoznak. KL4 akár klinikai depresszív csoport is lehet, noha heterogenitása > 1 szinten áll.

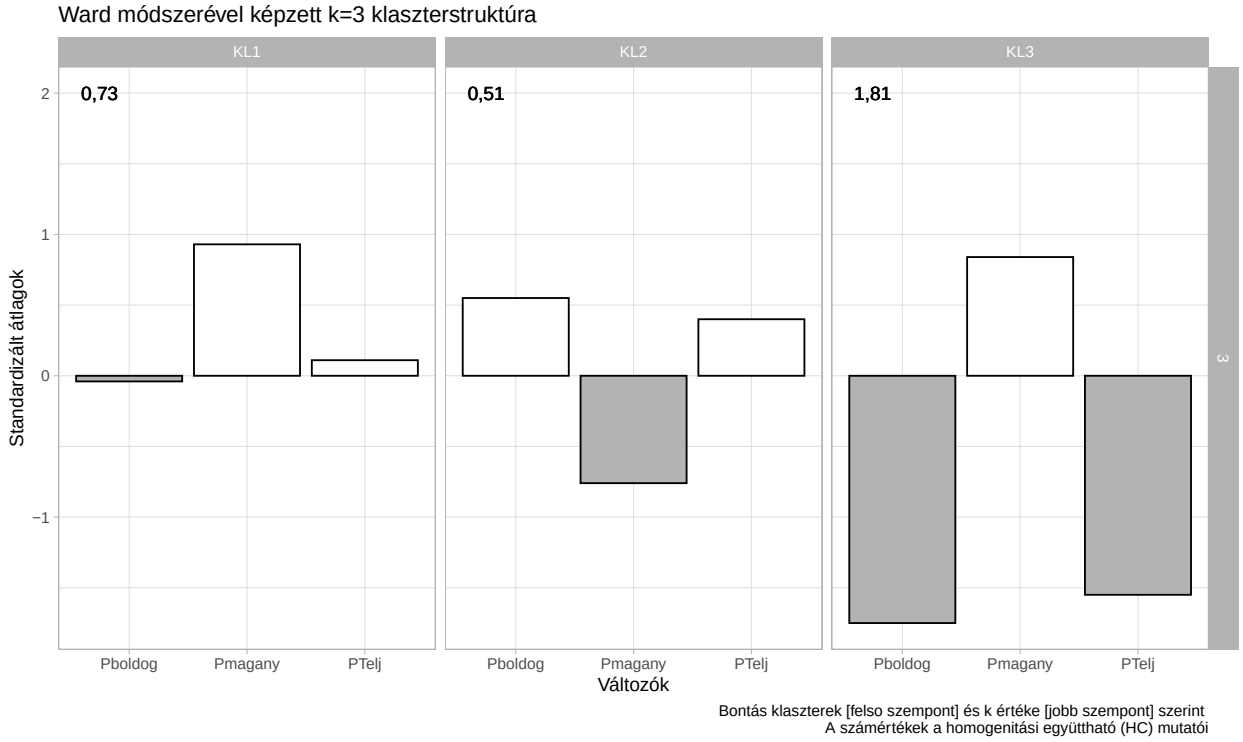
5. Végezz HKA-t ROPstatban a Ward módszerrel a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel, standardizálással! Foglald táblázatba a QC-ket k=2 és k=10 között! Hány klaszteres megoldás tűnik a legjobbnak?

A lenti táblázat eredményei és az 1. feladatban közölt könyök-grafikák alapján a k=3, azaz háromklaszteres megoldást fogadom el a Ward-módszerrel végrehajtott hierarchikus klaszterelemzésre. Ezt egyrészt alátámasztja, hogy k = 2 esetén több mint 15 százalékpontnyi esés következne be a magyarázott varianciában, a k=4 struktúrát pedig a Xien-Beni, a Pontbiszeriális együtthatóban mutatott értékei kedvezőtlenebb struktúrának írják le, mint a választott k=3 megoldást.

KL#	Ess_increase	EESS%	Pontbisz	XBmod	Sil.eh.	HC_avg	HCmin-HCmax	Clusters
10	17.603	81.48	0.342	0.473	0.592	0.379	0,18-1,03	1 (53) 14 (26)
9	18.362	80.26	0.351	0.496	0.590	0.404	0,24-1,03	3 (25) 6 (40)
8	23.059	78.72	0.355	0.456	0.598	0.434	0,24-1,03	4 (106) 29 (27)
7	28.290	76.83	0.370	0.534	0.610	0.471	0,24-1,03	54 (25) 148 (13)

KL#	Ess_increase	EESS%	Pontbisz	XBmod	Sil.eh.	HC_avg	HCmin-HCmax	Clusters
6	31.904	74.70	0.369	0.491	0.614	0.513	0,24-1,44	3 (65) 15 (25)
5	71.989	69.89	0.361	0.394	0.595	0.609	0,24-1,44	3 (90) 54 (38)
4	107.419	62.71	0.330	0.250	0.605	0.750	0,24-1,86	2 (160) 4 (133)
3	108.133	55.49	0.465	0.639	0.674	0.894	0,47-1,86	1 (79) 2 (293)
2	229.741	40.14	0.514	0.716	0.694	1.200	0,98-1,86	1 (372) 3 (128)

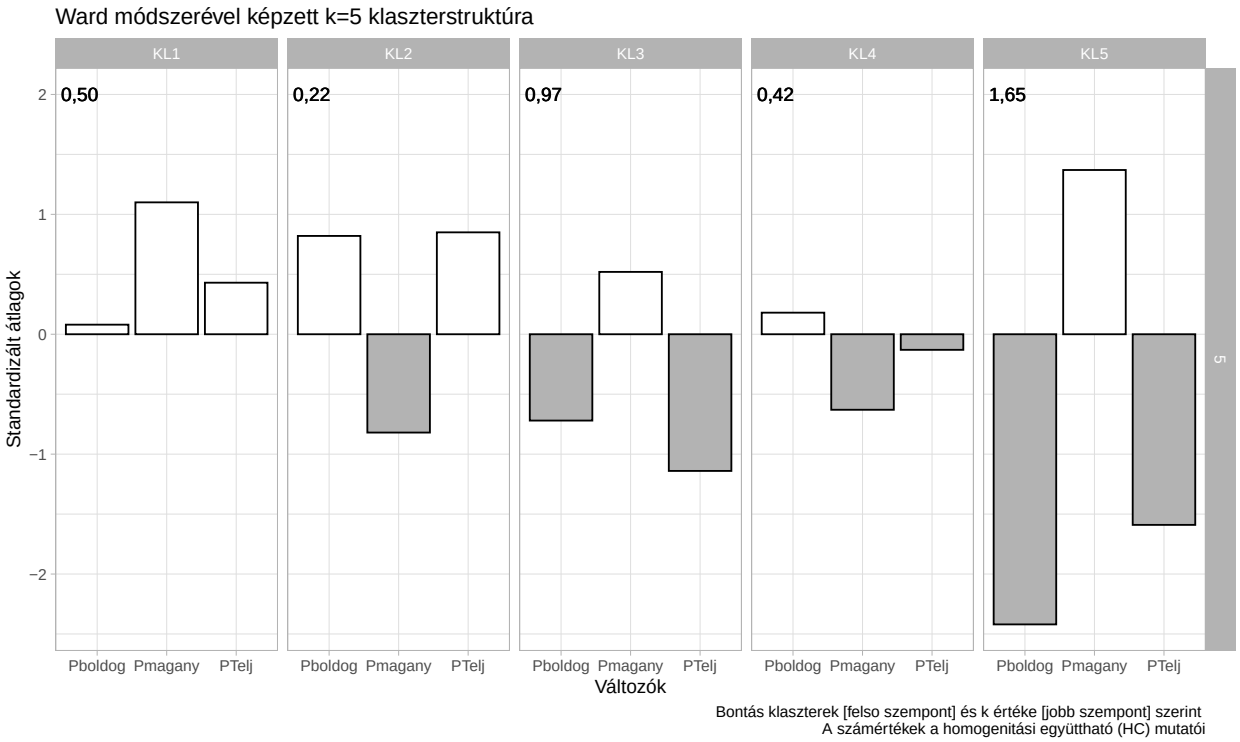
6. Készíts ábrát az előző feladat legjobb megoldásának centroidjai alapján a sima és a standardizált átlagokra!



A kapott struktúra alapján elhatárolódik három, karakteres csoport. KL1-ben a magányos, de jól teljesítő, az átlagnál kicsit boldogtalanabb "reziliens" személyeket értelmezem. Adott a KL2-es klaszter a kevésbé magányos, boldog és jól teljesítők köre. KL3 pedig a depresszív alcsoport, magányos, rosszul teljesít és boldogtalan, noha a klaszter heterogenitása már nem kívánatos szintű.

7. Az 5. feladat beállításaival állj meg az 5-klaszteres megoldásnál! Melyik itt a leghomogénebb és a legheterogénebb klaszter? Mit gondolsz, miért pont ezek? EESS% mekkora ennél a megoldásnál?

A Ward módszerével futtatott klaszterösszevonás k=5 mellett egy olyan klaszterstruktúrát eredményez (EESS% = 69.89%), ahol a leghomogénebb klaszter a KL2 (HC = 0.22), a legheterogénebb pedig a KL5 (HC = 1.65, immár a nem kívánt > 1 tartományban). Mivel feltételezhető, hogy a teljesítmény és a boldogság tételekre adott válaszok pozitívan korrelálnak, míg a magányosság tétel ezekkel negatív kapcsolatban áll, ezért a KL2 alacsony homogenitása abban áll, hogy e három tétel mentén az elvárt mintázatot mutató személyek állnak, közepes standardizált átlagokkal. Ellenben a KL5 klaszter, noha mintázata elvárt (magas magányosság, alacsony teljesítmény és boldogság), jó láthatóan magas standardizált átlagai azt valószínűsítik, a személyek átlagos távolsága egymástól nagyobb, a klaszter sokszínűbb eseteket fog át.


Ward módszerével képzett k=5 klaszterstruktúra
KL1
KL2
KL3
KL4
KL5
0,50
0,22
0,97
0,42
1,65
5
Standardizált átlagok
2
1
0
−1
−2
Pboldog
Pmagany
PTelj
Változók
Bontás klaszterek [felso szempont] és k értéke [jobb szempont] szerint
A számértékek a homogenitási együttható (HC) mutatói