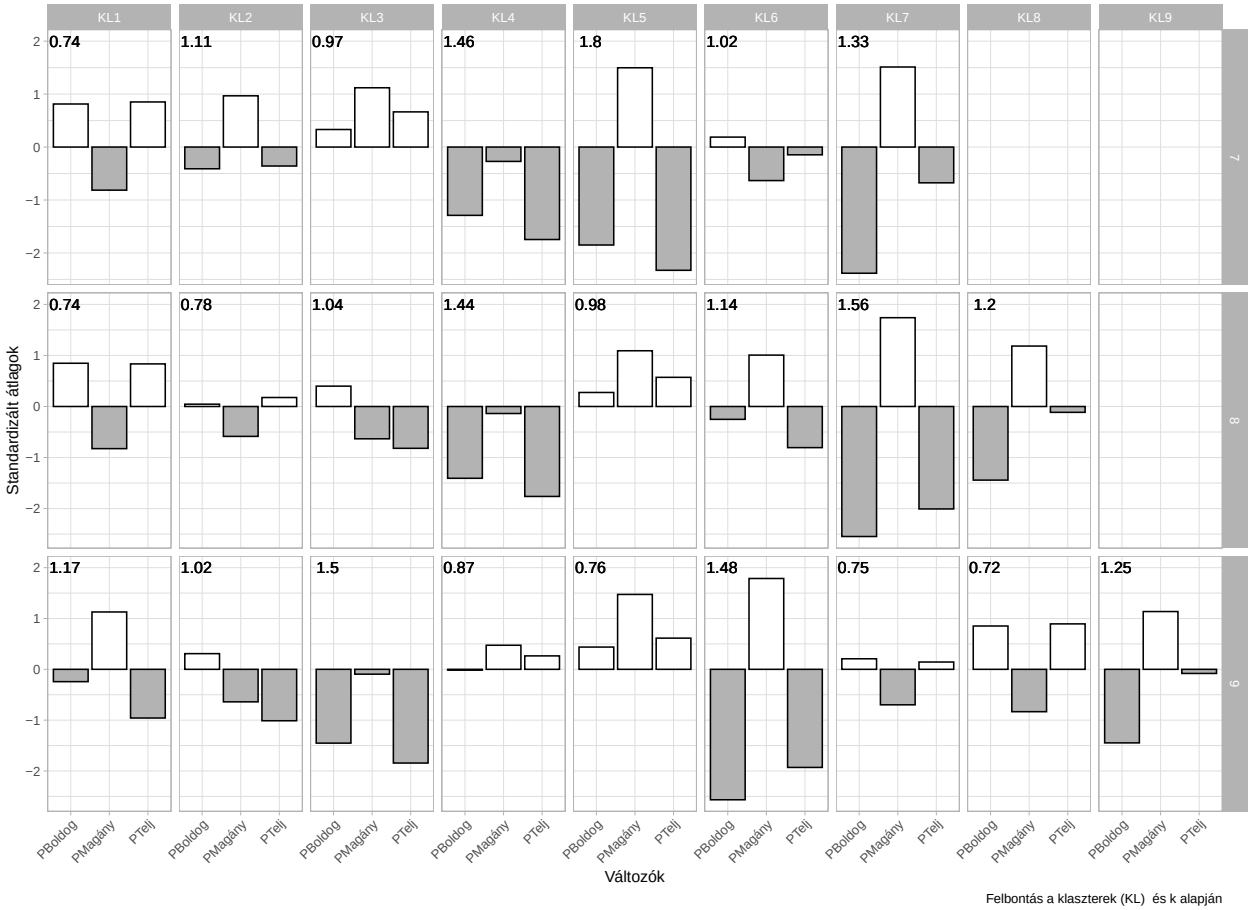
Házi feladatok megoldása 9.

k-középpontú klaszteranalízis R-ben

Smahajcsik-Szabó Tamás, M9IJYM

1. Végezz k-közép elemzést R-ben a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel, standardizálással k = 7 és 9 között!

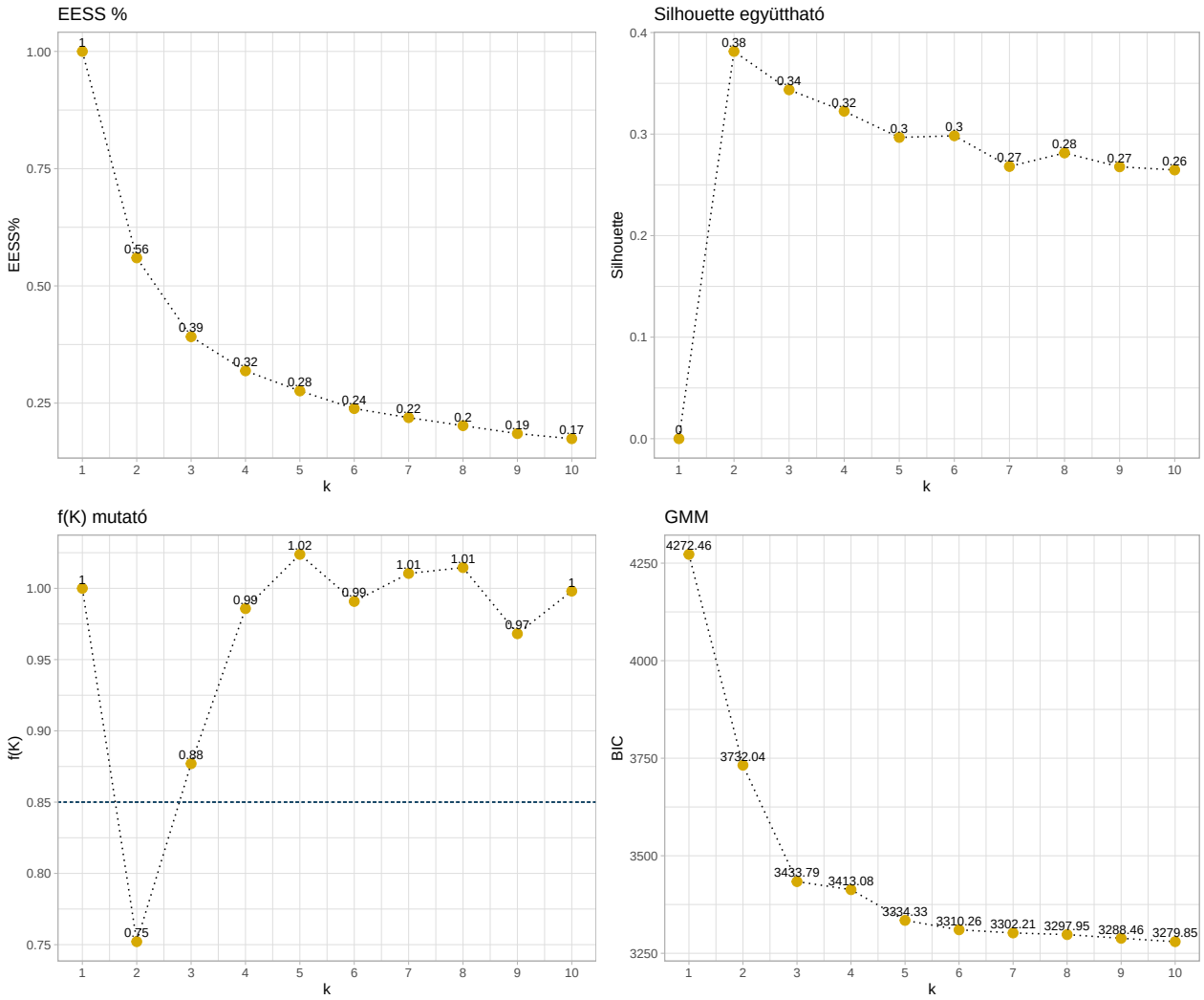
A k-közép elemzéseket 5 kezdeti centroid struktúrával, maximális 20 iterációval végeztem (MacQueen-féle algoritmussal). Az eredményekről az alábbi áttekintő ábra tájékoztat. A képződött klaszterek standard átlagait, és a homogenitási együtthatókat is feltüntettem.



1.ábra A klaszterstruktúra áttekintése

2. Hány klaszteres megoldás tűnik a legjobbnak az 1. feladat változói esetében a 6.4.1. alpontban leírt R-beli módszerek alapján (vö. 6.6-6.10. ábrák)?

A fenti ábra kedvezőtlenebb homogenitási indexeivel összhangban, a k=7 és k=9 közötti megoldás nem optimális a segítő ábrák alapján sem.



2. ábra: Segítő ábrák

Az EESS% könyökábrája (bal felül), illetve a Silhouette együttható (jobb felül) egy k=2 megoldás fölényét erősíti a k = 7, 8 illetve 9 megoldásokkal szemben. Az *f(K)* mutató (bal alul) a k=2 megoldást emeli ki, k=2-nél ereszkedik f(k) értéke a 0.85 küszöb alá. A k=7 és k=9 megoldások közül a k=9 esetén kedvezőbb kissé a mutató, de mindegyikre nézve suboptimális a jelzés. GMM függvénnyel tesztelve az adatokat a BIC értéke egyaránt alacsony k=7 és k=9 között, de egyrészt nem optimális a BIC ezen struktúrák mellett, másrészt minden más segítő ábra a k=7, 8 vagy 9 megoldások nem kielégítő voltát erősíti.

3. Mentsd el az 1. feladat klaszterváltozóit k = 7 és 9 között, tedd át ROPstatba és számítsd ki a Validálás modullal a főbb QC mutatókat! Melyik klaszterszám megoldása tűnik a legjobbnak?

4. Végezz k-medoid elemzést R-ben a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel, standardizálással k = 7 és 9 között!

5. Mentsd el a 4. feladat klaszterváltozóit k = 7 és 9 között, tedd át ROPstatba és számítsd ki a Validálás modullal a főbb QC mutatókat! Melyik klaszterszám megoldása tűnik a legjobbnak?