Hazi feladatok megoldasa 2.

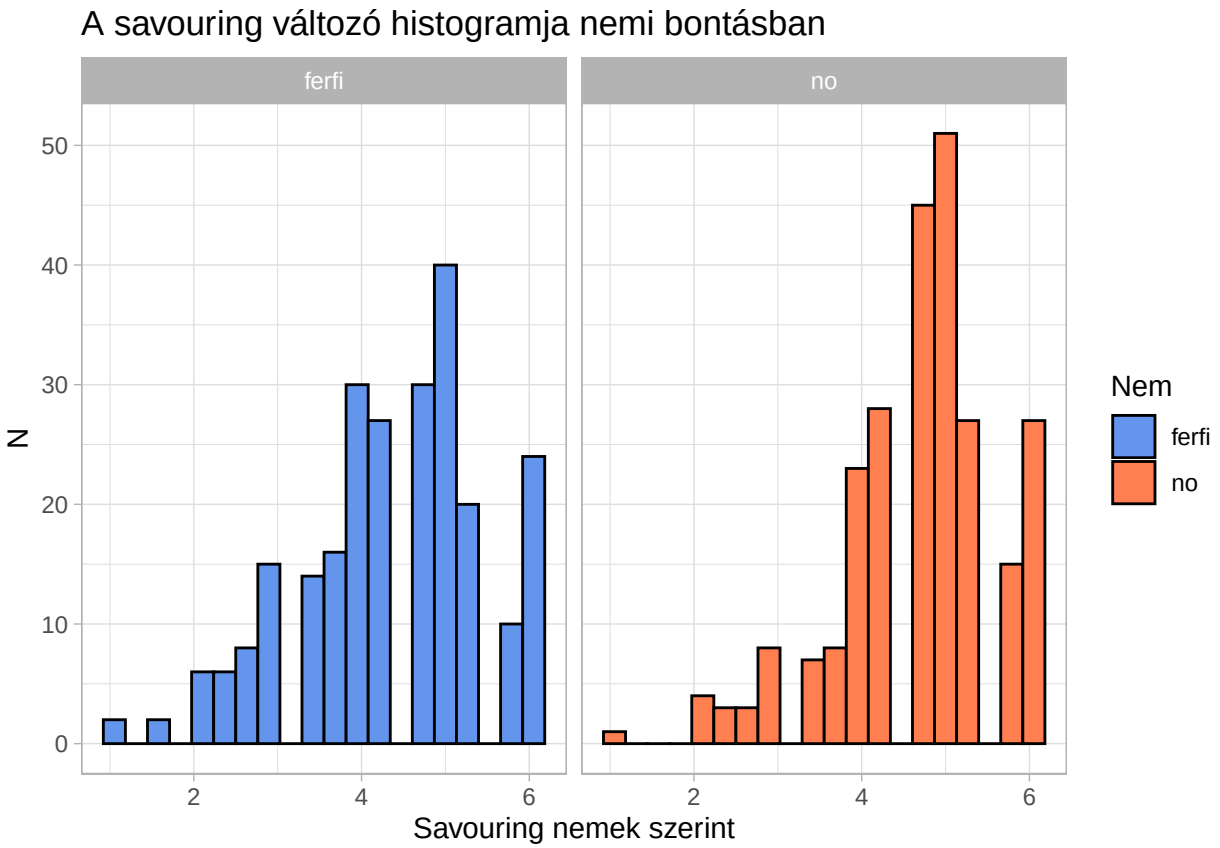
Smahajcsik-Szabó Tamás, M9IJYM

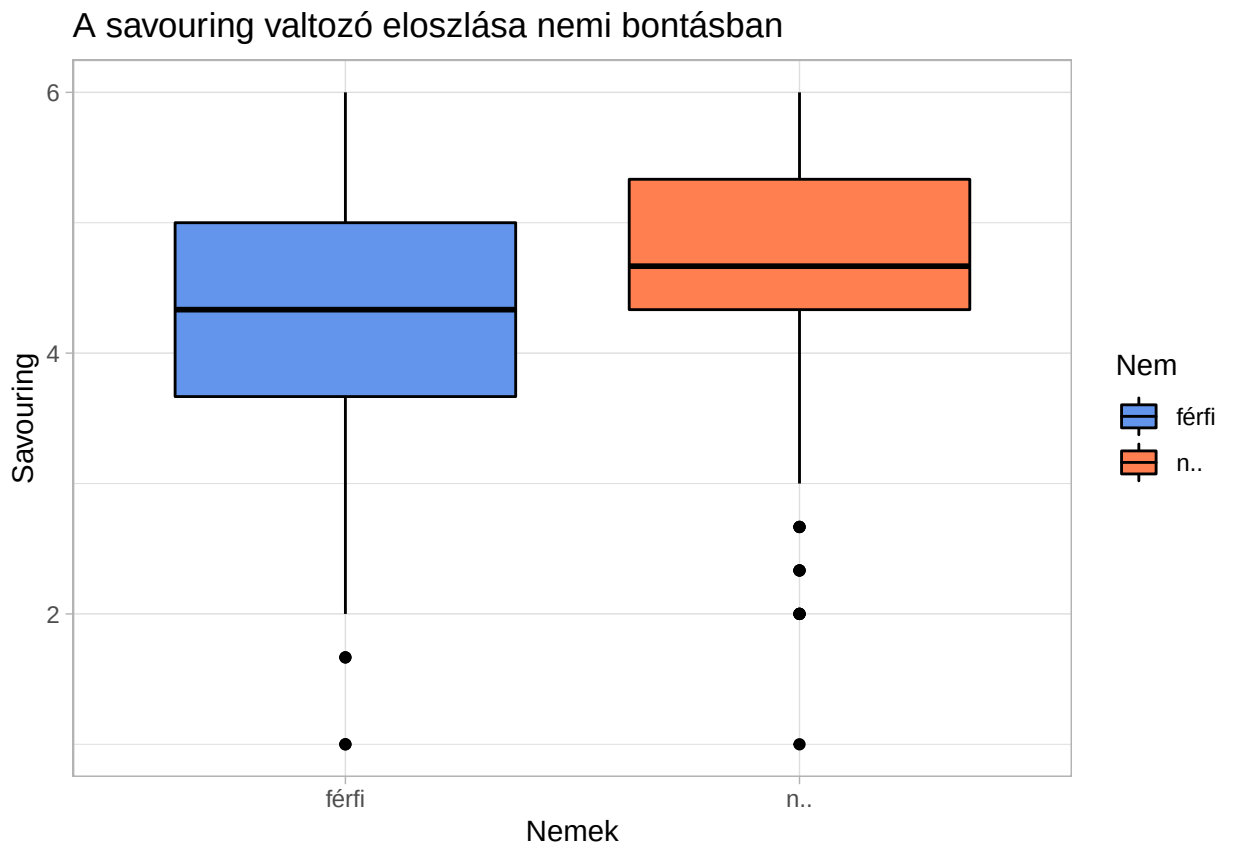
1. A boldog.msw fájl Pboldog és Pegészs változója esetében milyen gyakran fordul elő, hogy fordított kapcsolat van a két változó között? (Pl. beteg ember boldog vagy jó egészségben lévő boldogtalan)

A két változó esetében 24.4% esetben fordul elő diszkordáns pár. A Knight-féle algoritmussal számolva az érték kissé magasabb, 27.7%, mivel itt az X es Y valtozókra vontakoztatott egyező (tied) értékeket is figyelembe vesszük.

2. A nők Savoring szintje magasabb, mint a férfiaké? Cohen d mekkora? Milyen gyakran fordul elő a globális nemi trenddel ellentétes viselkedés? Mi ennek a szakmai relevanciája?

A normalitás és a homoscedasticitás sérülése miatt Mann-Whitney-féle U-tesztet végeztem a "savouring" valtozóra nézve nemi bontásban. A teszt eredménye W=2.5604×10^4, szignifikancia szintje pedig p=4.31×10^{-4}, melyek nemi különbséget valószínűsítenek a női résztvevők javára. A hatásméret azonban alacsonyabb (Cohen's d = -0.33).





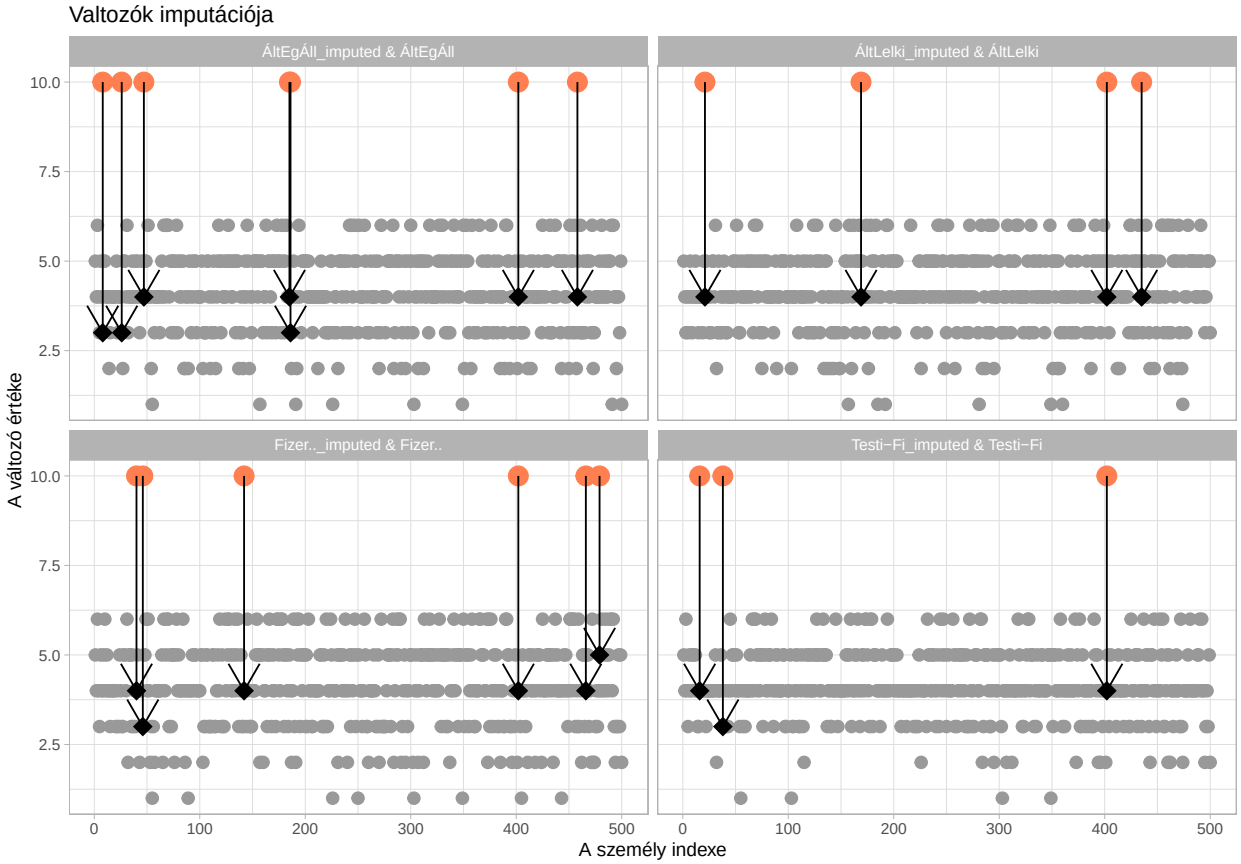
3. Elemezd a Testi_fi, Áltlelki, Áltegáll, Fizerö változókat a hiányzó adatok alapján! Melyiküknél a legkisebb/legnagyobb a hiányzó értékek száma?

A hiányzó adatok változónkénti számáról és százalékértékéről az alábbi táblázat tájékoztat.

Változó	Hiányzó adat	% hiányzó
Testi-Fi	3	0.6%
ÁltLelki	4	0.8%
Fizerő	6	1.2%
ÁltEgÁll	7	1.4%

4. A 3. feladat változóinak hiányzó értékeit próbáld meg pótolni. Melyik módszer a legjobb erre? A pótlás után mentsd el az msw fájlt boldog_imp.msw néven.

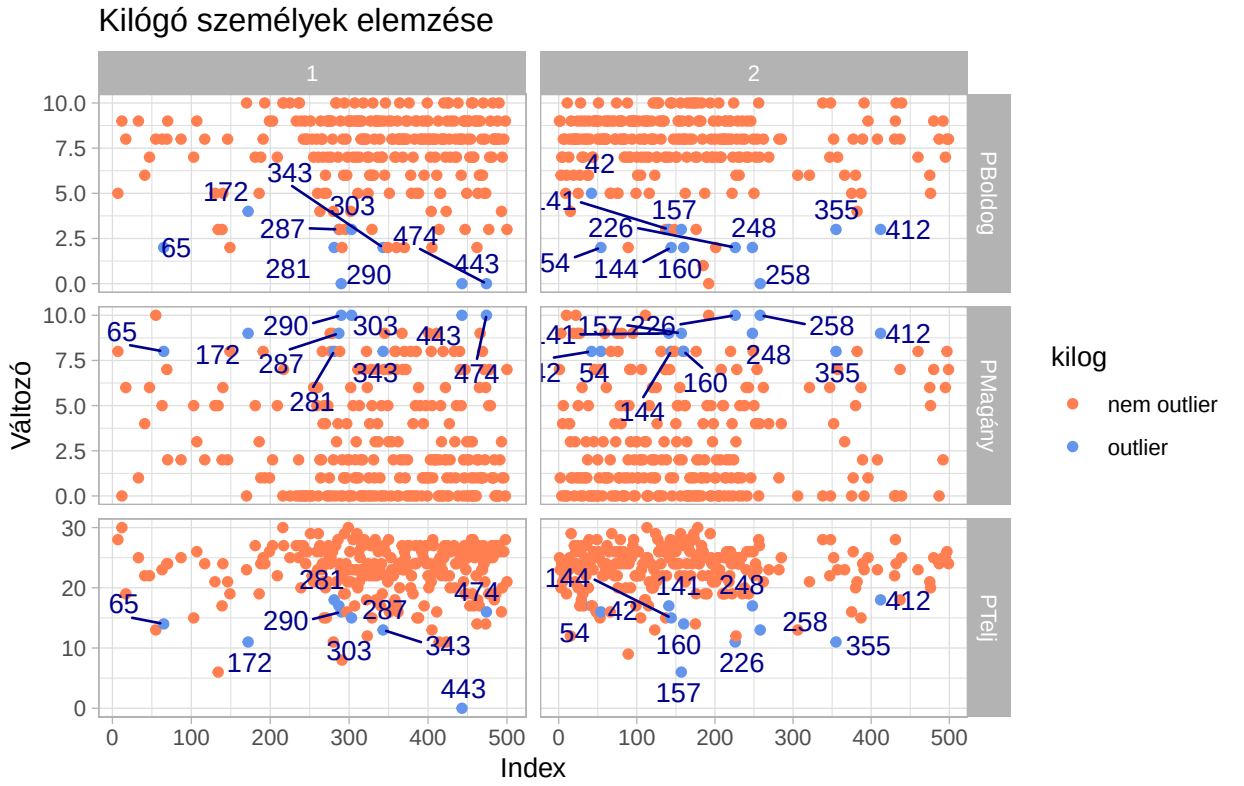
A hiányzó adatokat a k legközelebbi szomszéd módszerével (iker) pótoltam, mely során az R nyelven elérhető *caret* csomag adta megoldással éltem: Minden teljes adatsort mutató esetet felhasználtam egy tréning adatbázishoz, ahol prediktor változóként az indexen kívül minden egyéb változó beépült a modellbe; kivélt a pótlandó változók képeztek, melynek meglévő adatai a tréning során használt kimeneti változó szerepét töltötték be. Az ily módon kiképzett *kNN* modell szolgált alapul arra, hogy segítségével bejósoljam a hiányzó adatokat. A predikcióhoz használt prediktorváltozókat a hiányzó változóértékekkel rendelkező személyek meglévő, ép adatai adták. Az eredményeket az alábbi ábra foglalja össze.



Az eleve adott adatpontokat szürkével jelöltem, a hiányzó adatok a láthatóság és az ábra kedvéért egy az adatterjedelmen kívül elhelyezkedő értéket (esetünkben 10) kaptak. A nyilak ezen, a valóságban hiányzó értékek imputált, új értékeire mutatnak.

5. Van-e outlier a PTELJ, Pboldog, Pmagány változók értékhármasainál? Ha igen, hányas sorszámú személy a legkilógóbb és milyen értékekkel? Férfi ez vagy nő? Tudsz valamilyen magyarázatot, hogy mitől outlier ez a személy, ha megnézed a többi változójának értékét is? Lehet, hogy ezek kamu adatok?

Az outlierek azonosításához a medián abszolút szórás (median abolsute deviation, MAD) mérőszámát választottam, mert az átlag helyett egy robusztusabb centrális mutatót, a mediánt veszi alapul. A 2 * MAD / 0.6745 képlettel számoltam mindhárom kérdéses változó tekintetében és azonosítottam azokat, akik értékei a fenti kritériánál alacsonyabbak vagy nagyobbak.



Az ábrán nemi és változónkénti bontásban szerepel az adateloszlés. Kékkel jelöltem a feltételezhetően outlier személyeket, index változójukat pedig azonosításukhoz hasznosítottam. Nemi bontás tekintetében nem mutatkozik nagy egyensúly-borulás valamelyik nem oldalára. A változók tekintetében látható mintázat alapján magányosabb, kevésbé boldog, csökkenten teljesítő személyek - nem valószínűsítem, hogy tetszőlegesen megadott, vagy valótlan adatok lennének. A 443 sorszámú személy csak minimum - maximum értékeket adott meg, esetében feltételezhető, hogy nem valós az adat.

6. Van-e outlier a 8 Diener-tétel együttesénél? Ha igen, hányas sorszámú a két legkilógóbb személy és milyen értékekkel? Nemük? Tudsz valamilyen magyarázatot, hogy mitől outlierek, ha megnézed a többi változójuk értékét is? Lehet, hogy ezek kamu adatok?

Az előző feladathoz hasonló módszertannal elemeztem az adatokat a kilógó esetek tekintetében és noha nemi egyensúly mutatkozik az outlierek tekintetében, megállapítható, hogy az alacsony értékek tartományaiban vannak visszatérő személyek, akik következetesen hamis adatokat közöltek. Férfiaknál a 291, 303, 443, 462, 471 és 493 sorszámú résztvevőket említeném, hölgyeknél a 67, 89, 157, 192 vagy a 353 számú is alacsony pontú, kilógó személy.

Kilógó személyek elemzése