Clusterová analýza modelu Big Five

Osobnostní model Big Five od Golberga je v dnešní době naprostý staple psychologického výzkumu. Teoreticky model tvrdí, že v jádru osobnosti je pět nezávislých faktorů: Otevřenost (O), Svědomitost (S), Extraverze (E), Přívětivost (P), Neuroticismus (N). Bohužel se tyto faktory v praxi neukazují zcela nezávislé. Různé studie uvádějí různě velké korelace, například metaanalýza Van Der Linden et al. (2010) obsahující 144 117 respondentů našla tyto vztahy (uvádím jen r > 0,3):

Otevřenost	Extraverze	r = 0,31
Svědomitost	Přívětivost	r = 0,31
Svědomitost	Neuroticismus	r = -0,32

Tabulka 1: Korelace jednotlivých dimenzí podle Van Der Linden et al. (2010)

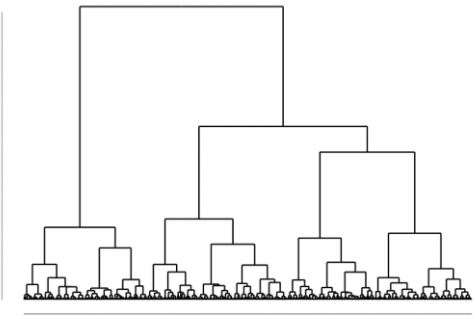
Proto jsem se rozhodl model Big Five prozkoumat pomocí custerové analýzy, abych odhalil, jestli se jednotlivá vyplnění shlukují do typických vyplnění. Dle mého názoru jedna z nevýhod tohoto modelu je, že je špatně srozumitelný pro neodborníky, jelikož je velmi komplexní. Oproti MBTI, který je veřejností oblíbený (I Tinder nabízí možnost si přidat svůj typ z MBTI), je celkový možný počet kombinací prakticky neomezený, jelikož jednotlivé faktory nejsou dichotomické, ale spojité, což je pro interpretaci náročnější. Pokud by se podařilo najít pomocí clustrové analýzy typická vyplnění, tak by se interpretace zjednodušila na konečný počet typických profilů (které by potom byly dobře srozumitelné) a atypické vyplnění, která by se bohužel stále interpretovala stejně.

Data

Použitá data jsou z volně dostupné databáze (https://github.com/automoto/big-five-data), poté jsem jednotlivé faktory převedl do Z skórů, ty budou označovány jako AgrZ (Přívětivost); ExtZ (Extraverze); OpeZ (Otevřenost); ConZ (Svědomitost); NeuZ (Neuroticismus). Chtěl jsem si zkusit práci s datasetem, který má přes 300 000 vyplnění, ale při práci v programu JASP 0.18.2.0 mi při analýzách vyskočila chyba, že na takto velký dataset nemám dostatek RAM, proto jsem zcela drakonicky dataset zmenšil na pouze prvních 1500 vyplnění.

Shluková analýza

Jak jsem již zmiňoval, tak k analýze byl použit program JASP 0.18.2.0, zde jsem použil metodu hierarchického clusterování. Jako metodu shlukování jsem použil Wardovu metodu s Euklidovou metrikou. Z první analýzy vychází tento dendrogram:



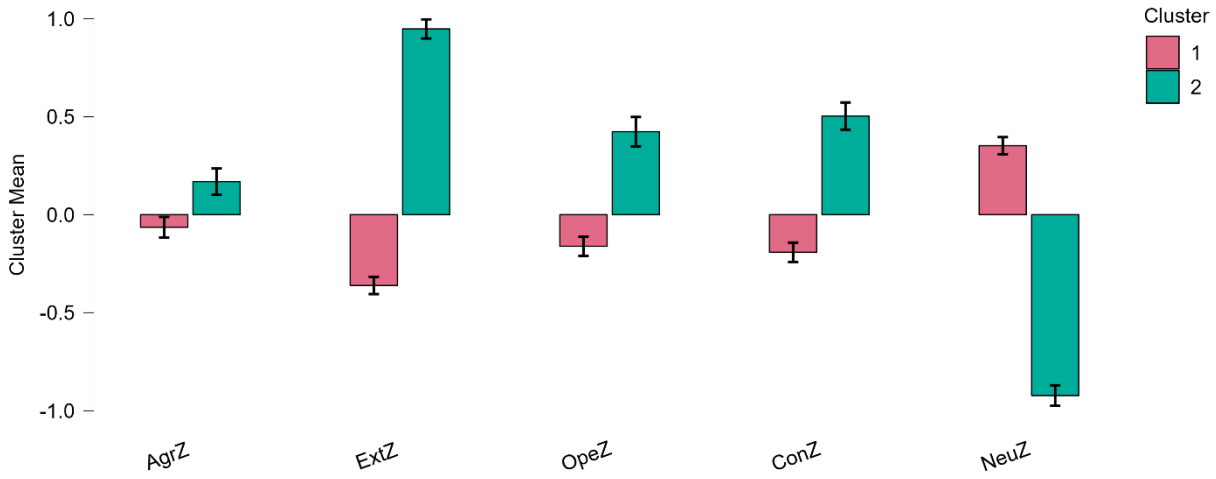
Obrázek 1: Dendrogram

Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=353.

Z tohoto obrázku se nabízí jako řešení buď 2, nebo 4 clustery. Tak tedy prozkoumáme jednotlivá řešení.

Řešení s dvěma clustery

Toto řešení se jeví velmi elegantní a snadné na interpretaci. Čísla v Tabulce 2 ukazují průměry clusterů v jednotlivých dimenzích, jelikož máme proměnné standardizované v Z skóre, tak se zároveň jedná i cohenovo d. Cluster 1 jsou tedy lidé se slabě podprůměrnou extraverzí a neuroticismem a Cluster 2 jsou lidé, kteří mají velmi vysokou extraverzi a vysokou svědomitost a velmi nízký neuroticismus. Mohli bychom si tedy tyto dvě skupiny pojmenovat podle jejich nejsilnějších rozdílů na Introvertě-úzkostný typ a Extravertně-stabilní typ. To už poté můžeme přelabelovat podle Galénovy teorie na Melancholik (Introvertně-úzkostný) a na Sangvinik (Extravertně-stabilní). Jako zajímavost u tohoto modelu bych uvedl, že u obou clusterů se přívětivost „vyprůměrovala“ téměř do nuly (menší než slabý efekt).



Obrázek 2: Průměry clusterů v jednotlivých dimenzích; 2 clustery

	AgrZ	ExtZ	OpeZ	ConZ	NeuZ
Cluster 1	-0.06	-0.36	-0.16	-0.19	0.35
Cluster 2	0.17	0.95	0.42	0.50	-0.92

Tabulka 2: Průměry clusterů v jednotlivých dimenzích; nad 0,2 slabý efekt (oranžově); nad 0,5 střední (červeně); 0,8 silný (sytě červeně)

Jediné dvě metriky z Tabulky 3, které umím interpretovat jsou Silhoutte a Dunn index, které naznačují, že tento model není příliš vhodný. Jako pravidlo palce se uvádí, že by Silhoutte měl být alespoň 0,5 pro kvalitní model. Stejně tak Dunn index blížící se nule naznačuje nízkou kvalitu řešení. Pokud chápu Calinski-Harabasz index správně, tak ten sám nemá vypovídací hodnotu, jen se dá porovnávat mezi jednotlivými modely, který je více vhodný. Interpretaci Pearson's γ jsem plně nepochopil, ale nechávám ji tu pro vzdělanější.

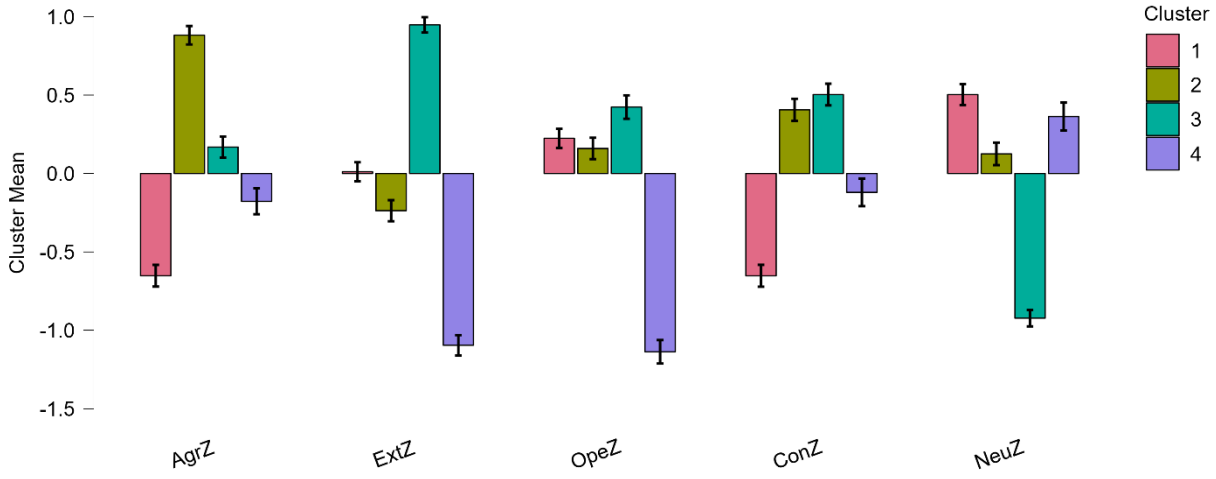
	Value
Pearson's γ	0.220
Dunn index	0.043
Silhouette	0.150
Calinski-Harabasz index	303.8

Tabulka 3: Metriky ukazující sílu modelu; 2 clustery

Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=353.

Řešení s čtyřmi clustery

Už na první pohled na obrázek 3 je toto řešení těžší na interpretaci viz Obrázek 3 a Tabulka 4. Cluster 1 jsou lidé, kteří mají nízkou přívětivost, nízkou svědomitost, vysoký neuroticismus, lehce nadprůměrnou otevřenost. Cluster 2 má velmi vysokou přívětivost, lehce nadprůměrnou svědomitost a lehce podprůměrnou extraverzi. Cluster 3 má velmi vysokou extraverzi, velmi nízký neuroticismus, vysokou svědomitost a lehce nadprůměrnou otevřenost. Cluster 4 má velmi nízkou extraverzi a otevřenost a lehce nadprůměrný neuroticismus. Pro srozumitelnost hned označím Cluster 3 jako Sangvinik, jelikož je totožný s Clusterem 2 z předchozího modelu. Cluster 2 bychom mohli označit jako Přívětivé, jelikož je to jediný cluster, který má vysokou přívětivost. Cluster 4 se poté nejsilněji vyznačuje nízkou otevřeností a introverzí, ten si dovolím kreativně pojmenovat jako Isolovaný typ. Poslední Cluster 1, úzkostní lidé s nízkou svědomitostí a přívětivostí, byl oříšek, ani po sáhodlouhé konzultaci s ChatGPT jsem nepřišel na slovo, které by tuto skupinu dobře popsalo. ChatGPT nabízel slovo *„neúspěšní“*, to by sice asi do jisté míry vystihlo tuto skupinu, ale působí to neprofesionálně. Proto jsem se rozhodl pro termín Úzkostně-nesvědomitý-nepřívětivý typ. Takže naše 4 typy by byly: Sangvinik, Přívětivý, Isolovaný a Úzkostně-nesvědomitý-nepřívětivý. K těmto typům by se pak daly přiřadit historické osoby, aby to bylo více ilustrativní a lidé se s tím více ztotožňovali (stejně jako to dělá www.16personalities.com).



Obrázek 3: Průměry clusterů v jednotlivých dimenzích; 4 clustery

	AgrZ	ExtZ	OpeZ	ConZ	NeuZ
Cluster 1	-0.65	0.01	0.22	-0.65	0.50
Cluster 2	0.88	-0.24	0.16	0.41	0.13
Cluster 3	0.17	0.95	0.42	0.50	-0.92
Cluster 4	-0.18	-1.10	-1.14	-0.12	0.36

Tabulka 4: Průměry clusterů v jednotlivých dimenzích; nad 0,2 slabý efekt (oranžově); nad 0,5 střední (červeně); 0,8 silný (sytě červeně)

Pokud se podíváme na Tabulku 5, opět vidíme, že Silhouette a Dunn index poukazuje na nízkou kvalitu modelu. Calinski-Harabasz index vychází u předchozího modelu vyšší, což naznačuje, že model s dvěma clustery je lepší řešení.

Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=353.

	Value
Pearson's γ	0.305
Dunn index	0.053
Silhouette	0.110
Calinski-Harabasz index	260,7

Tabulka 5: Metriky ukazující sílu modelu; 4 clustery

Závěr

Osobně si nemyslím, že by kterýkoliv z těchto modelů byl užitečný. Model s dvěma clustery má být sice ten lepší, ale je velmi redukcionistický, kdybychom chtěli rozdělit lidi na ose Melancholik-Sangvinik, tak vůbec nepotřebujeme model Big Five, což jde vidět i podle toho, že toto řešení vůbec nepoužívá přívětivost. Redukce jde vidět i v tom, že naši Melancholici jsou mnohem blíže průměru než Sangvinici, ale lingvisticky to už z těchto labelů není na první pohled poznat. Model se čtyřmi clustery je zajímavější, už jsou to nějaké nové typologie, které člověk před tím neviděl. Nicméně podle metrik ukazujících sílu jednotlivých modelů nemůžu ani jeden doporučit.

Popravdě cílem této práce bylo právě naučit se interpretovat sílu modelu v clusterové analýze, jelikož na hodině jsme se učili jen porovnávat modely mezi sebou. Proto jsem si na začátku položil zdánlivě stupidní otázku, abych získal dva nekvalitní modely, aby správně řešení bylo zamítnout oba modely, ne vybrat ten „méně špatný“.

Zdroje:

Van Der Linden, D., Te Nijenhuis, J., & Bakker, A. B. (2010). The General Factor of Personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, *44*(3), 315–327. https://doi.org/10.1016/j.jrp.2010.03.003

https://github.com/automoto/big-five-data