Zpráva o stavu štěstí ve světě - shluková analýza[1]

Úvod

Žijeme ve světě dat. Nejen statistikové, ale i datoví analytici a vůbec běžní lidé v různých oborech občas stojí před problémem, jak zpracovat, vyhodnotit a prezentovat určitá data, často velmi rozsáhlá. Vezměme si jako jednoduchý příklad následující tabulku, která obsahuje údaje o 40 lidech, u kterých jsme zjišťovali jejich pohlaví, vzdělání a výšku:

```
  pohlavi vzdelani vyska
1       M       ZA   181
2       M       VS   173
3       M       ZA   181
4       M       ZA   182
5       M       ZA   172
...
36      F       VS   167
37      F       SS   183
38      F       SS   168
39      F       SS   175
40      F       ZA   168
```

Mohli bychom pak například naše data vyhodnotit jako celek s ohledem na údaj “výška”, který bychom reprezentovali pomocí střední hodnoty (průměru) a zjistili bychom, že nejnižší účastník našeho výzkumu měří 161 cm, nejvyšší 187 cm a průměrná výška je 173,7 cm.

```
Min.    1st Qu. Median  Mean     3rd Qu. Max.
161.0   168.0   173.0   173.7   178.8   187.0
```

Přestože sumární hodnoty mohou být užitečné v určitých případech, zajímavější by mohlo být podívat se na naše data z pohledu atributů pohlaví nebo vzdělání:

```
  Pohlavi Prum(cm)
1       F 169.75
2       M 177.60
```

```
  Vzdelani Prum(cm)
1       SS 173.08
2       VS 174.43
3       ZA 173.25
```

[1] *Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=156*

Přestože podobné popisné statistiky, ať už v podobě sumární nebo napříč skupinami, mají v datové analýze své místo, často stojíme před mnohem komplikovanějšími problémy. V našem jednoduchém příkladu, jsme pracovali pouze s jednou proměnnou - výška. Naše data bychom ale mohli rozšířit o množství dalších tělesných ukazatelů např. hmotnost, objem pasu, objem hrudníku, množství podkožního tuku, ale také například o zdánlivě nesouvisející údaje jako je měsíční příjem, počet automobilů v domácnosti nebo rychlost internetového připojení. V určitých situacích nehledáme cíleně ani tak korelace mezi jednotlivými ukazateli, jako spíše určitou příslušnost ke skupině nebo **shluku**. Tento typ práce s daty spadající do kategorie vícerozměrných statistických metod se nazývá **shluková analýza** (angl. cluster analysis). Jejím cílem je nalézt v datech pozorování, která jsou si podobnější než ostatní a zároveň zajistit, aby jednotlivé shluky byly dostatečně odlišné.

Existuje velké množství algoritmů a postupů, které lze pro shlukovou analýzu použít. Volba správného postupu souvisí s účelem analýzy a také s povahou a kontextem konkrétních dat. Mezi základní metody patří hierarchické a nehierarchické shlukování. V našem příkladu se věnujeme pouze nehierarchickému shlukování, konkrétně algoritmu k-means.

Data

Pro naši analýzu jsme použili **World Happiness Data**[2] report za rok 2022. Jedná se o každoroční mezinárodní výzkum probíhající od roku 2012. Výzkum probíhá v jednotlivých zemích a data jsou také podle států prezentována. Základním ukazatelem je skór štěstí, který je výsledkem self-reportingu účastníků výzkumu. Účastníci jsou vyzváni, aby ohodnotili svůj život na škále 0-10, přičemž 0 odpovídá nejhoršímu možnému životu a 10 nejlepšímu. Pro naši analýzu nás nicméně nezajímá tento skór, ale odvozené faktory[3]:

- Skutečný HDP na hlavu
- Sociální podpora
- Očekávaná doba dožití ve zdraví
- Svoboda činnit životní rozhodnutí
- Štědrost
- Vnímání korupce

World Happiness report, jeho metodika, tvorba jednotlivých odvozených faktorů je předmětem mnoha polemik a také kritiky. Účelem naší práce není tyto aspekty hodnotit ani dále rozebírat. Důležitá jsou data a možnost jejich zpracování a prezentace.

[2] https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2022

[3] Faktory se skutečně odvíjejí z celkového skóru, nikoli naopak. Výzkum zjišťuje pouze odpověď na otázku ohledně vnímaného celkového štěstí, či spokojenosti. Šest faktorů je poté modelováno na základě tohoto skóru. Sedmý faktor je tzv. Dystopia, který slouží jako regresní faktor a zároveň proměnná, která zahrnuje skór, který nelze vysvětlit ostatními šesti.

Následující tabulky ukazují prvních a posledních pět zemí dle skóru štěstí. Česká republika se v roce 2022 umístila na příjemném 18. místě za Velkou Británií a před Belgií.

	Country	Happiness score	Dystopia (1.83) + residual	Explained by: GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption
1	Finland	7.821	2.518	1.892	1.258	0.775	0.736	0.109	0.534
2	Denmark	7.636	2.226	1.953	1.243	0.777	0.719	0.188	0.532
3	Iceland	7.557	2.320	1.936	1.320	0.803	0.718	0.270	0.191
4	Switzerland	7.512	2.153	2.026	1.226	0.822	0.677	0.147	0.461
5	Netherlands	7.415	2.137	1.945	1.206	0.787	0.651	0.271	0.419

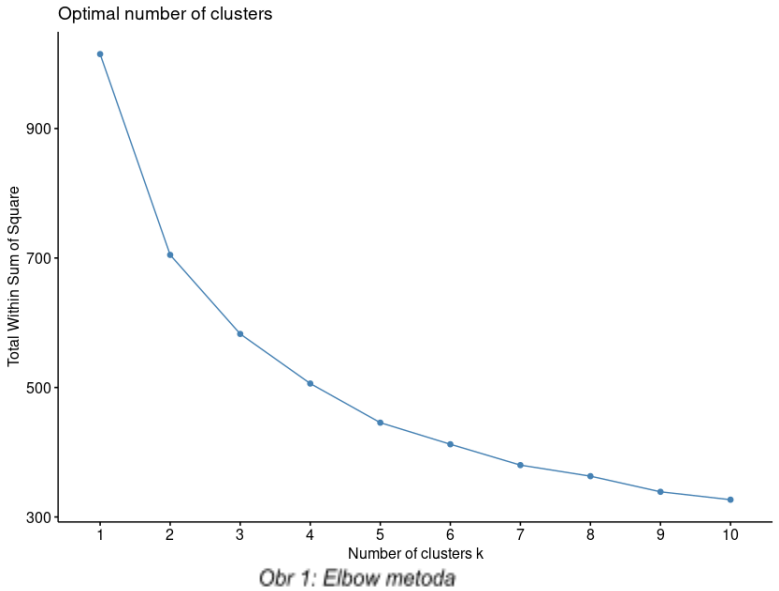
…

142	Botswana*	3.471	0.187	1.503	0.815	0.280	0.571	0.012	0.102
143	Rwanda*	3.268	0.536	0.785	0.133	0.462	0.621	0.187	0.544
144	Zimbabwe	2.995	0.548	0.947	0.690	0.270	0.329	0.106	0.105
145	Lebanon	2.955	0.216	1.392	0.498	0.631	0.103	0.082	0.034
146	Afghanistan	2.404	1.263	0.758	0.000	0.289	0.000	0.089	0.005

Přestože bychom si v případě těchto dat mohli vystačit s popisnými statistikami a vyhodnocovat například maxima, minima či střední hodnoty v rámci jednotlivých zemí, shluková analýza nabízí možnost seskupit data do shluků. V případě k-means algoritmu hledáme taková pozorování (země), která jsou si nejbližší v rámci euklidovského prostoru tzn. sdílejí vzdálenost k určitému bodu v geometrickém prostoru. Tyto body, které tvoří pomyslné středy shluků, se nazývají **centroidy**.

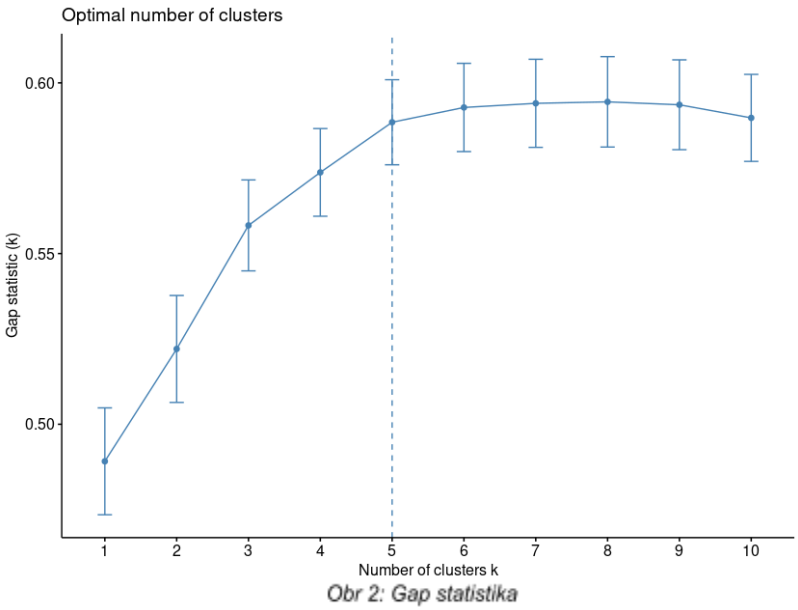
Počet shluků

Počet shluků je klíčový ukazatel, který musíme v případě nehierarchického shlukování s použitím k-means **definovat předem**. Tento fakt vyžaduje od výzkumníka jistou úroveň obeznámení s daty, intuici, případně zakotvení v nějaké teorii. Pokud bychom chtěli uvést příklad z oblasti psychologie, pak například v případě aplikace shlukové analýzy na tisíce pozorování z dotazníku osobnosti “Velké pětky”, můžeme předpokládat, že shluků hledáme pět. Pro zjištění předpokládaného počtu shluků můžeme využít také některé statistické metody. Jako první demonstrujeme použití tzv. **elbow method**. Principem této metody je vizuálně nalézt bod, ve kterém křivka přestane



Obr 1: Elbow metoda

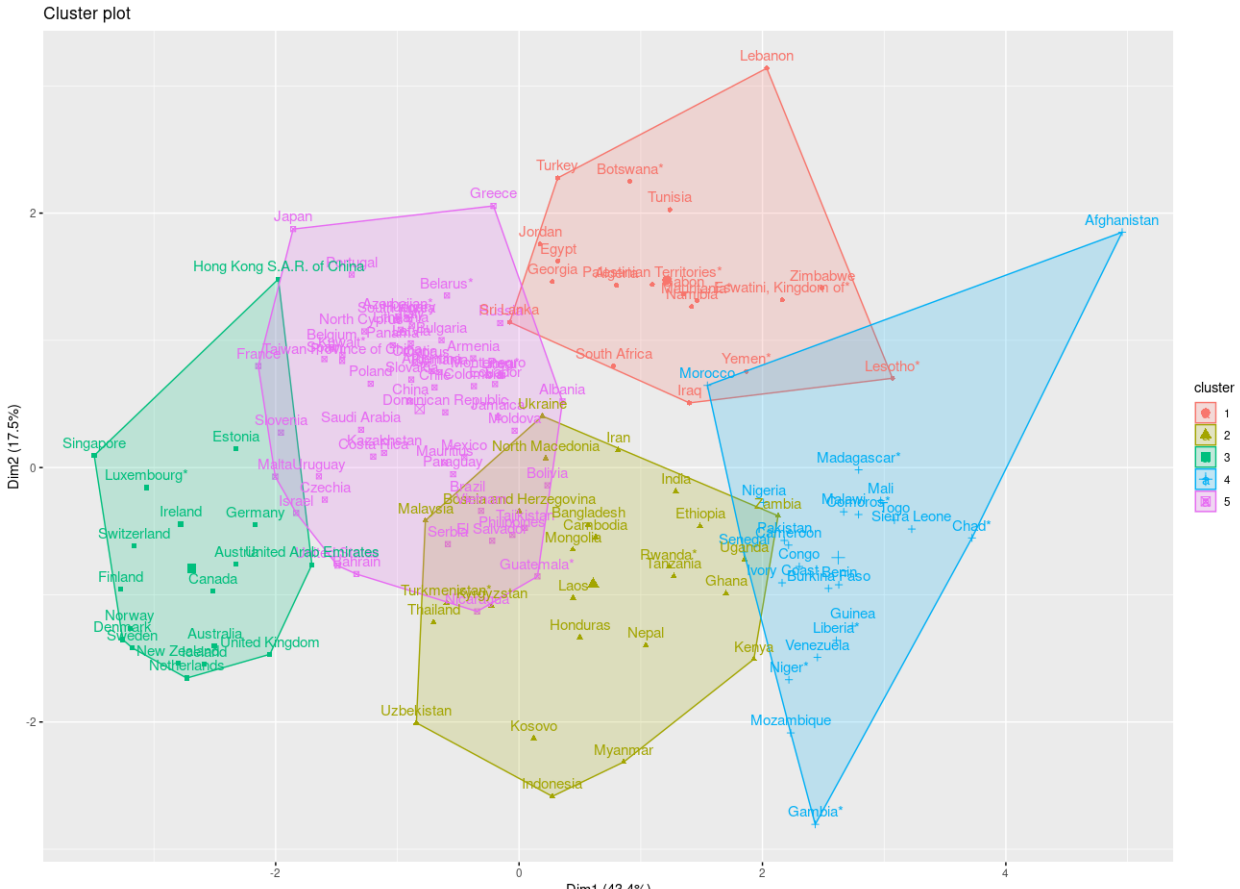
“padat” prudce dolů (jako například mezi prvním a druhým bodem v našem grafu viz obr. 1). Nevýhodou je (jako v našem případě) jistá neurčitost - v některých případech nemusí být tento bod zcela zřejmý. V elbow grafu pozorujeme tento zlom už někde kolem 4 shluků, ale mohli bychom uvažovat také o 5. Pro přesnější určení počtu shluků existují pokročilejší metody. Jednou z nich je tzv. **gap statistika**. Pokud ji aplikujeme na naše data, vidiíme, že optimální počet shluků je skutečně 5, viz obr. 2.



Obr 2: Gap statistika

Shluková analýza

Nyní můžeme na naše data aplikovat k-means algoritmus s předpokládaným počtem 5 shluků a data vizualizovat.



Obr 3: Shluky států

Tento příklad vizualizace je jen jedním z mnoha možných. Jak je z obrázku zřejmé, už při relativně nízkém počtu pozorování (v našem případě n=146) se popisy jednotlivých pozorování (názvy států) překrývají a čitelnost je problematická. Pro rychlou orientaci v datech a představu o jednotlivých shlucích a příslušnosti pozorování k nim nicméně může být tento typ vizualizace užitečný.