Shluková analýza návyků studentů a jejich vztah k akademickému výkonu

1. Deskripce datového souboru

Tato statistická zpráva se věnuje identifikaci behaviorálních profilů studentů na základě jejich každodenních návyků. K analýze byl využit dataset Student Habits & Performance, který byl uměle vytvořen tak, aby realisticky kopíroval skutečná data. Tento dataset obsahuje 1000 záznamů, z nichž každý řádek reprezentuje jednoho studenta. Mezi každodenní návyky studentů autoři řadí: dobu strávenou studiem, spánek, používání sociálních médií, čas strávený na platformě Netflix, kvalitu stravy, hodnocení duševního zdraví a další.

2. Cíl analýzy

Pro analýzu dat byla zvolena metoda vícenásobného shlukování (K-means clustering), jelikož umožňuje identifikovat vnitřní strukturu v souboru 1000 respondentů. Na rozdíl od jednoduchých korelačních metod, které sledují vztah pouze dvou proměnných, shluková analýza umožňuje analyzovat celou kombinaci osmi sledovaných návyků současně. To umožňuje vytvořit komplexní typologii studentů a následně zjistit, jak se tyto přirozeně vzniklé skupiny liší ve svém akademickém výkonu.

Cílem shlukové analýzy tedy bylo seskupit studenty do homogenních shluků pomocí metody K-means na základě těchto specifických proměnných: počet hodin studia, využívání sociálních sítí, sledování Netflixu, školní docházka, délka spánku, kvalita stravy, hodnocení duševního zdraví a zapojení do mimoškolních aktivit. Následně jsme si za cíl zvolili zhodnotit, zda a jak behaviorální vzorce studentů korelují s jejich reálným úspěchem u zkoušek.

3. Postup při shlukové analýze

Před samotným provedením analýzy K-means byla provedena analýza optimálního počtu shluků. Na základě metody lokte a následného posouzení teoretické interpretovatelnosti jednotlivých profilů bylo rozhodnuto o rozdělení souboru do tří shluků (k=3). Toto nastavení se ukázalo jako ideální kompromis mezi statistickou přesností a schopností dat jasně popsat odlišné typy studentského chování bez nadbytečné fragmentace souboru.

Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=736.

Vzhledem k odlišným jednotkám měření byly všechny vstupní proměnné pro účely analýzy standardizovány na Z-skóry, což zajistilo jejich rovnocennou váhu při výpočtu vzdáleností mezi shluky. Pro vlastní seskupování byla zvolena metoda K-means. Na základě explorace dat byl stanoven počet shluků k=3, což se jevilo jako optimální řešení pro jasnou diferenciaci typických profilů studentů. Veškerá statistická analýza probíhala v programech MS Excel a Statistica.

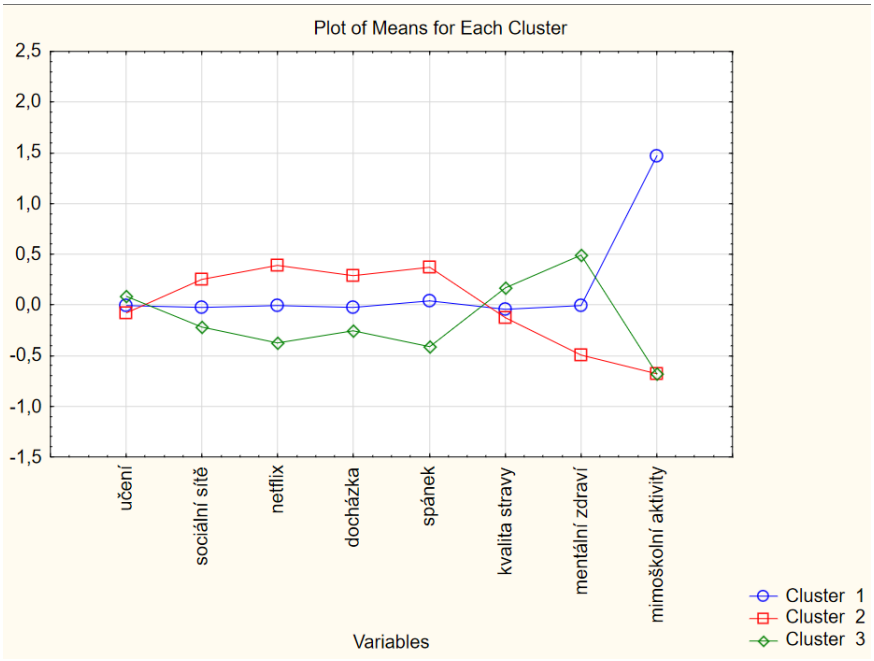
4. Interpretace výsledků

V této části analyzujeme charakteristiky tří identifikovaných shluků studentů. Profily jsou definovány průměrnými hodnotami Z-skórů v jednotlivých kategoriích návyků.

4.1. Charakteristika shluků

Na základě výsledku shlukové analýzy jsme definovali 3 shluky studentů dle jejich každodenních návyků. Tyto shluky jsou znázorněny v níže přiloženém profilovém grafu.

Graf 1: Profilový graf průměrů shluků (Z-skóry)



Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=736.

Profilový graf zobrazuje průměrné hodnoty sledovaných proměnných ve třech identifikovaných shlucích, přičemž hodnoty jsou vyjádřeny ve formě z-skórů (M = 0, SD = 1). Z grafu je patrné, že shluk 1 se vyznačuje převážně průměrnými hodnotami ve všech sledovaných oblastech, s výraznou odchylkou v podobě nadprůměrné míry mimoškolních aktivit, což naznačuje skupinu studentů aktivních zejména mimo školní prostředí. Shluk 2 dosahuje nadprůměrných hodnot v oblasti využívání sociálních sítí, sledování Netflixu, docházky a spánku, avšak současně vykazuje podprůměrné hodnoty v kvalitě stravy, mentálním zdraví a mimoškolních aktivitách. Naproti tomu shluk 3 se vyznačuje podprůměrným využíváním digitálních médií jako jsou sociální sítě a Netflix a současně nadprůměrnými hodnotami v oblasti kvality stravy a mentálního zdraví. Rozdíly mezi shluky tak ukazují na existenci odlišných behaviorálních profilů studentů, které se liší zejména v oblasti trávení volného času, životosprávy a mentálního zdraví. Pro přesnější přehled níže přikládáme tabulku s konkrétními průměry shluků neboli centroidy.

Tabulka 1: Centroidy shluků (v jednotkách Z-skóre)

Proměnná	**Shluk 1**	**Shluk 2**	**Shluk 3**
Čas strávený učením [h]	-0,005	-0,081	**+0,084**
Čas strávený na sociálních sítích [h]	-0,027	**+0,249**	-0,220
Čas strávený na Netflixu [h]	-0,008	**+0,388**	-0,376
Školní docházka	-0,030	**+0,287**	-0,260
Čas strávený spánkem [h]	+0,041	**+0,374**	-0,407
Kvalita stravy	-0,045	-0,128	**+0,170**
Posouzení svého mentálního zdraví	-0,007	-0,492	**+0,492**

Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=736.

Mimoškolní aktivity	**+1,464**	-0,683	-0,683

4.2. Vzdálenosti mezi jednotlivými shluky

Tabulka euklidovských vzdáleností mezi průměry shluků potvrzuje separaci identifikovaných skupin. Nejvyšší míru odlišnosti vykazují shluky 1 a 3 (vzdálenost 0,817), což značí výrazně diferencované typy návyků těchto skupin studentů. Naopak nejnižší vzdálenost byla naměřena mezi shluky 2 a 3 (0,591). Tento výsledek naznačuje, že ačkoliv se tyto dvě skupiny rozcházejí v klíčové proměnné, tzn. školní docházce, v ostatních sledovaných návycích vykazují vyšší míru podobnosti než ve vztahu ke shluku 1.

Tabulka 2: Matice euklidovských vzdáleností mezi shluky

	Shluk 1	**Shluk 2**	**Shluk 3**
Shluk 1	0,000	0,814	0,817
Shluk 2	0,814	0,000	0,591
Shluk 3	0,817	0,591	0,000

4.3. Analýza rozptylu

Analýza rozptylu (viz Tabulka 3) byla využita k posouzení míry diferenciace jednotlivých shluků v rámci sledovaných proměnných. Výsledky ukázaly, že nejvýraznějším diferenciačním znakem mezi shluky je školní docházka ($p < 0,001$), zatímco ostatní proměnné přispívají k odlišení shluků v menší míře.

Tabulka 3: Analýza rozptylu (ANOVA)

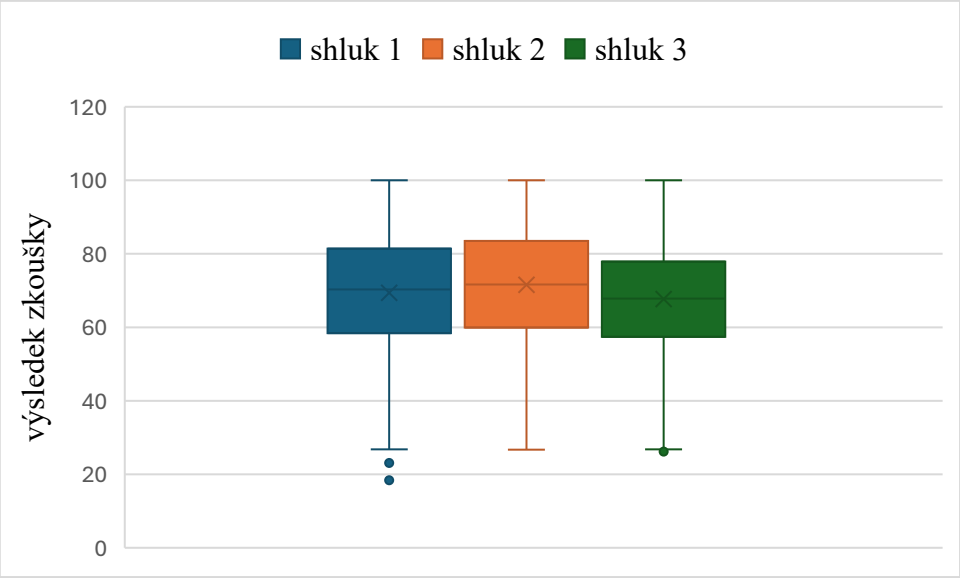
Proměnná	**F hodnota**	**Df**	**p-hodnota**
Čas strávený učením [h]	0,258	2;997	0,772
Čas strávený na sociálních sítích [h]	1,407	2;997	0,245

Čas strávený na Netflixu [h]	0,171	2;997	0,843
Školní docházka	**2588,943**	2;997	**<0,001**
Čas strávený spánkem [h]	1,175	2;997	0,309
Kvalita stravy	1,453	2;997	0,234
Posouzení svého mentálního zdraví	0,414	2;997	0,661
Mimoškolní aktivity	2,479	2;997	0,084

4.4.Vztah identifikovaných shluků k úspěšnosti u zkoušky

Studenti mohli u zkoušky dosáhnout maximálního počtu 100 bodů a minimálně 0 bodů. Průměrné skóre celého souboru se rovnalo 69,6 bodům. Následující analýza se zaměřuje na to, jak se tyto výsledky lišily v závislosti na příslušnosti k jednotlivým shlukům. Rozložení dosažených bodů napříč identifikovanými skupinami zachycuje níže uvedený krabicový graf.

Graf 2: Rozložení dosažených bodů u zkoušek napříč jednotlivými shluky



Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=736.

Analýza ukázala, že nejvyšší průměrnou úspěšnost vykazuje shluk 2 (M=71,65). Při pohledu na profil tohoto shluku je pravděpodobné, že důležitým faktorem úspěchu u zkoušky může být vysoká míra školní docházky (+0,287) a dostatek spánku (+0,374).

První shluk, charakteristický výrazným zapojením se do mimoškolních aktivit, průměrně dosahuje pouze o 1,35 bodu méně než ve druhém shluku (M=70,3). Navzdory vysoké mimoškolní aktivitě dosahují srovnatelných výsledků.

Třetí shluk vykazuje v porovnání s ostatními nižší úspěšnost (M=67,8). Ve srovnání se shlukem 2 vykazují studenti ve třetím shluku o 3,85 bodu méně. Ačkoliv se tito studenti nejméně rozptylují sociálními sítěmi a vykazují nejlepší duševní zdraví, jejich výkon může souviset s podprůměrnou docházkou (-0,26) a deficitem spánku (-0,407). Data tedy potvrzují, že zejména míra absence je důležitým faktorem hrajícím roli v úspěšnosti u zkoušek.

Vliv příslušnosti ke shluku na výsledné skóre u zkoušky byl testován pomocí jednofaktorové analýzy rozptylu (viz Tabulka 4). Výsledky potvrdily, že rozdíly mezi průměry shluků jsou statisticky významné ($F(2, 998) = 21,6$; $p < 0,001$). Velikost účinku vyjádřená hodnotou $\eta^2 = 0,042$ naznačuje, že zvolené shlukování vysvětluje přibližně 4,2 % celkové variability studijních výsledků. Ačkoliv se jedná o spíše malý účinek, statistická významnost potvrzuje, že rozdíly v návycích mají na akademický úspěch mírný dopad ($\eta^2 = 0,042$).

Tabulka 4: Výsledky ANOVA

Průměry			**Počet studentů (N)**			**F**	**p-hodnota**	**η^2**
Shluk 1	Shluk 2	Shluk 3	Shluk 1	Shluk 2	Shluk 3	21,6	<0,001	0,042
70,3	71,65	67,8	467	290	243			

Následné post-hoc testování pomocí Tukeyho HSD testu potvrdilo, že všechny tři shluky se od sebe ve výsledném skóre statisticky významně liší ($p < 0,05$). To znamená, že každá z identifikovaných skupin studentů představuje odlišnou úroveň akademické úspěšnosti.

Tabulka 5: Tukeyův post hoc test

	Shluk 1	**Shluk 2**	**Shluk 3**
Shluk 1	0,000	0,003	0,004
Shluk 2	0,003	0,000	<0,001
Shluk 3	0,004	<0,001	0,000

Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=736.

5. Diskuse

Z provedené analýzy vyplývá, že ačkoliv se studenti liší v celé škále návyků, nejvýrazněji odlišovala jednotlivé skupiny studentů školní docházka. Výsledky dále naznačují, že vyšší školní docházka může souviset s vyšší úspěšností u zkoušky ($p > 0,05$), nicméně vzhledem k povaze analýzy nelze vyvozovat kauzální vztahy. Vliv příslušnosti ke shluku na studijní výkon byl sice statisticky významný, avšak velikost efektu byla relativně nízká ($\eta^2 = 0,042$), což naznačuje, že sledované návyky vysvětlují pouze omezenou část variability studijních výsledků. Mezi ostatními proměnnými nebyly nalezeny statisticky významné rozdíly mezi shluky, a jejich vztah k výkonu tak nelze na základě této analýzy jednoznačně potvrdit.