

Klasifikace konzumentů YTB videí podle vybraných kritérií¹

YouTube je jednou z největších a nejpobulárnějších online platform pro sdílení videí na internetu. Založena byla v roce 2005 a od roku 2006 je jejím vlastníkem společnost Google (Wikipedia, 2023). Za 18 let existence YouTube se tato platforma stala pro mnoho tvůrců, které dnes označujeme jako youtubery, zdrojem finančního příjmu (Sedláček, 2016; Statista, nedat.). Postojům vůči YouTube a tím, jaký obsah je na YouTube nejpobulárnější se již zabývala celá řada výzkumů (Brodersen et al., 2012; Chatzopoulou et al., 2010; Welbourne & Grant, 2016).

Profesor Jeff Galak z Carnegie Mellon University pro svůj YouTube kanál Data Demystified vytvořil dotazník², ve kterém zjišťoval zvyky respondentů při sledování YouTube videí a za pomoci služby Amazon mechanical turk získal odpovědi od 1000 respondentů (Galak, nedat.).

V dotazník zjišťoval například kolik minut denně respondenti trávili sledováním YouTube videí v posledních 7 dnech, kolik různých kanálů odebírají, na jakém zařízení sledují YouTube nejčastěji nebo jak důležité jsou pro ně různé vlastnosti videí, na které se koukají.

V naší zprávě se pokusíme rozdělit tyto respondenty do skupin podle jejich hodnocení YTB jako platformy, množství konzumovaného obsahu v posledních 7 dnech a toho, na kterých zařízeních YTB videa sledují³.

Hodnocení platformy bylo získáno 5 položkami hodnocenými na 7 bodových škálách Likertova typu v dimenzích:

- příšerný – úžasný
- špatný – dobrý
- nudný – zábavný
- nekvalitní – kvalitní
- neužitečný – užitečný

¹ Data a další informace o této zprávě jsou dostupné na adrese

<https://dostal.vyzkum-psychologie.cz/stat4?i=125>

² Původní data z dotazníku jsou dostupná zde: [Data Demystified YouTube Survey.sav - Disk Google](#)

³ Data byla vyextrahována a převedena z původního souboru .sav do formátu .xlsx s využitím programu RStudio, balíčku „foreign“, a následně analyzována v programu TIBCO Statistica.

V otázkách konzumace jsme využili odpovědi na položky týkající se:

- počtu odebíraných kanálů
- počet kanálů, jejichž obsah respondent sledoval v posledních 7 dnech
- počet zhlédnutých videí

Dále použijeme odpovědi na položku, kde respondenti uváděli, na kterých zařízeních sledují YTB videa nejčastěji v procentech.

Shluková analýza

Pro rozdělení respondentů do skupin využijeme metod shlukové analýzy (cluster analysis). Termín shluková analýza můžeme chápat jako nadřazený pojem celé rodině metod, ze kterých pro naši analýzu využijeme konkrétně metodu hierarchické shlukové analýzy a metodu K-Means.

Princip metody K-Means spočívá v tom, že se vybere počet K shluků, do kterých se budou data rozdělovat. Následně se náhodně vybere K centroidů, což jsou bodové reprezentace každého shluku. Tyto centroidy jsou inicializovány buď náhodně, nebo pomocí nějakého algoritmu. Poté se data přiřadí ke shlukům na základě nejmenší vzdálenosti k centroidu každého shluku.

Po přiřazení všech dat ke shlukům se vypočítá nový centroid každého shluku jako průměr všech dat, která jsou k němu přiřazena. Následně se proces opakuje, přičemž data jsou přiřazována ke shlukům na základě nových centroidů. Tento proces se opakuje, dokud se pozice centroidů neustálí.

Metoda K-Means nám však nenabízí odpověď na to, do kolika skupin by bylo vhodné naše data roztrždit. K odhadu tohoto parametru tedy využijeme druhou ze zmíněných metod, tedy hierarchickou shlukovou analýzu.

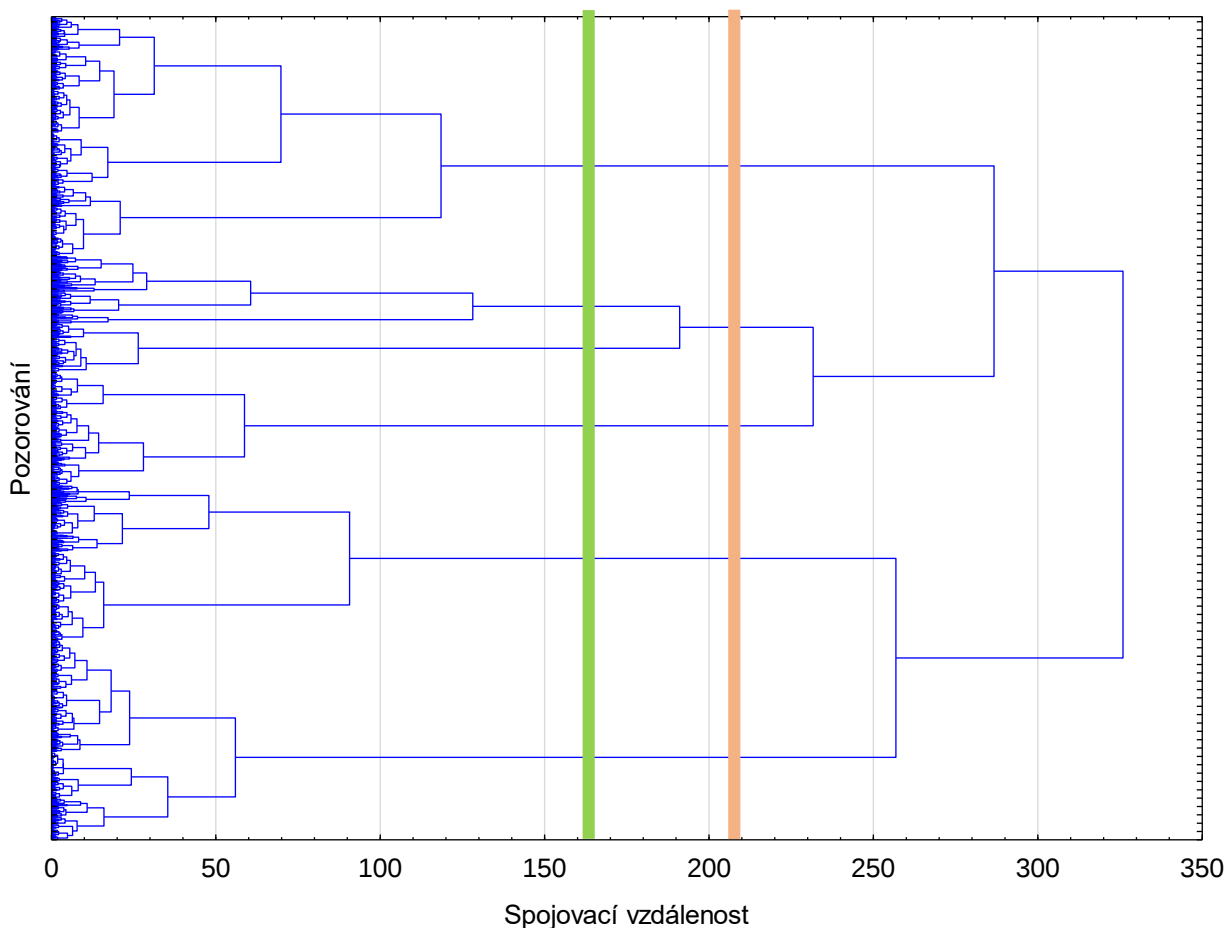
Hierarchická shluková analýza

Před analýzou jsme z našeho souboru vyřadili 15 respondentů pro velké množství chybějících odpovědí. Z důvodu různých měřítek škál u jednotlivých položek jsme před další analýzou převedli odpovědi na t-skóry. Na data zbylých 985 respondentů jsme aplikovali postupy hierarchické shlukové analýzy. Princip této metody spočívá v tom, že se vypočítávají vzdálenosti mezi jednotlivými datovými body a na základě těchto vzdáleností se vytváří dendrogram. Dendrogram zobrazuje, jak jsou jednotlivé datové body spojeny v různých

úrovních shluků. Na začátku je každý datový bod sám o sobě jedním shlukem. Následně se dva nejbližší shluky spojí do jednoho shluku, který se stává novým uzlem dendrogramu. Tento proces se opakuje a shluky jsou postupně spojovány do vyšších úrovní dendrogramu, až se vytvoří jediný shluk obsahující všechny data.

Kritérium pro výběr dvou nejbližších shluků může být různé. My jsme se rozhodli využít Wardovy metody, která je založená na minimalizaci součtu kvadratických odchylek v každém kroku spojování shluků. Výhodou této metody je, že se snaží minimalizovat rozdíly mezi jednotlivými shluky a může vést k tvorbě poměrně homogenních shluků s co nejmenší možnou vnitřní variabilitou.

Obr. 1 – Stromový graf hierarchické shlukové analýzy



Na ose Y našeho dendrogramu (obr. 1) vidíme jednotlivé pozorování jako samostatné shluky, které se postupně slučují do větších a větších klastrů. Na ose X lze vidět spojovací vzdálenost mezi shluky spočítanou pomocí Wardovy metody. Dle dendrogramu se ukazuje jako nejschůdnější rozdělení respondentů do pěti skupin („rozříznutí“ reprezentováno oranžovou vertikální čarou), případně rozdělení do šesti skupin (zelená vertikální čára). Pro další analýzu

jsme zvolili řešení s rozdělením do 5 skupin, které ačkoli se zdá o něco méně robustním než rozdělení do 6 skupin, nabídne pro naše účely přehlednější a snáze interpretovatelné výsledky.

Výsledky – metoda K-Means

Po zvolení vhodného počtu klastrů jsme využili k rozdělení respondentů do skupin metodu K-Means. S její pomocí jsme rozdělili našich 985 respondentů do pěti klastrů (15 respondentů z původních 1000 bylo pro chybějící data z analýzy vyřazeno), které budeme dále označovat jako klastř A (n = 300), B (n = 275), C (n = 253), D (n = 96) a klastř E (n = 61). V tabulce 1 lze vidět výsledky analýzy rozptylu našich klastrů, dle kterých se odpovědi našich skupin signifikantně liší ve všech dotazovaných položkách.

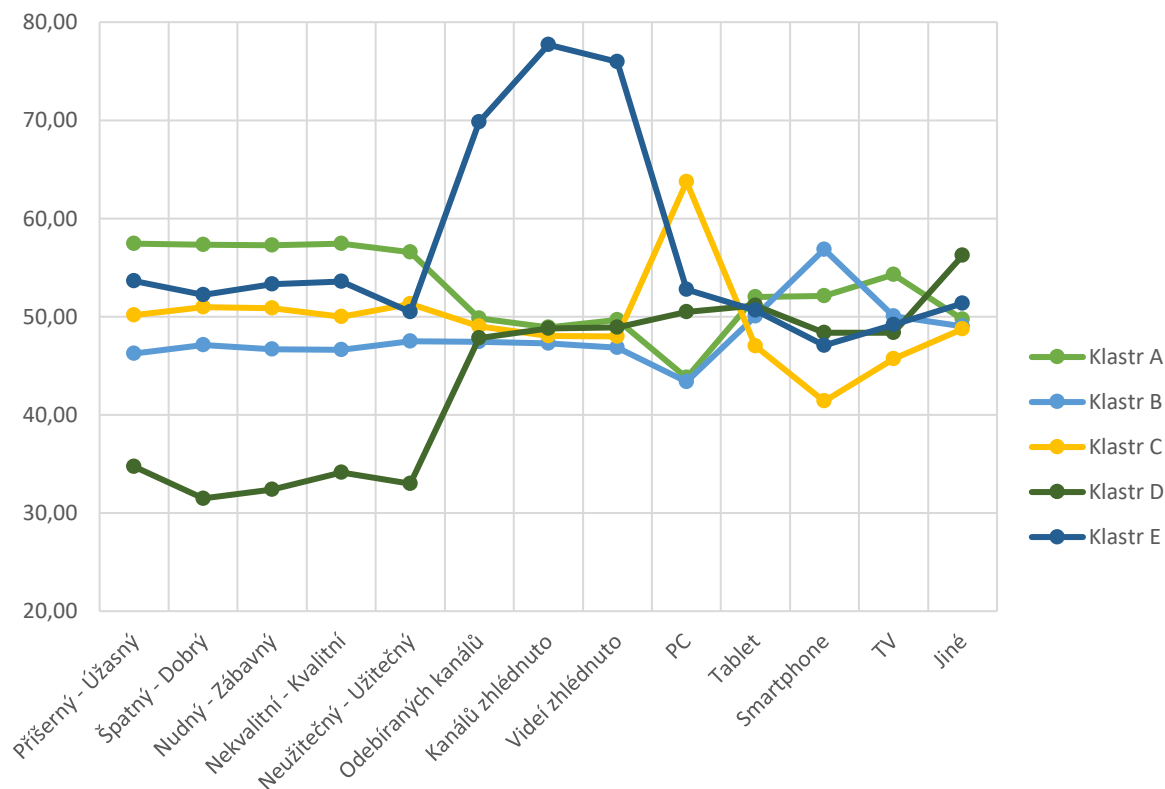
Tab. 1 – Analýza rozptylu klastrů získaných metodou K-Means

Proměnná	Between	df	Within	df2	F	signif.
Příšerný – Úžasný	435,230	4	548,770	980	194,309	0,000
Špatný – Dobrý	518,143	4	465,857	980	272,497	0,000
Nudný – Zábavný	494,597	4	489,403	980	247,600	0,000
Nekvalitní – Kvalitní	446,231	4	537,769	980	203,296	0,000
Neužitečný – Užitečný	428,907	4	555,093	980	189,306	0,000
Odebíraných kanálů	264,737	4	719,263	980	90,177	0,000
Kanálů zhlédnuto	503,439	4	480,562	980	256,663	0,000
Videí zhlédnuto	450,174	4	533,826	980	206,608	0,000
PC	718,258	4	265,742	980	662,194	0,000
Tablet	36,189	4	947,811	980	9,355	0,000
Smartphone	338,230	4	645,770	980	128,322	0,000
TV	104,471	4	879,529	980	29,101	0,000
Jiné	45,142	4	938,858	980	11,780	0,000

Na obrázku 2 lze vidět graf průměrných odpovědí respondentů všech klastrů pro jednotlivé položky dotazníku. Naš největší klastř A dle grafu hodnotí YTB nejpozitivněji ze všech klastrů a sleduje videa především na smartphonech, TV nebo tabletech. Klastř B se pak odlišuje

především tím, že sleduje YTB převážně jen na smartphonech. Do klastru C nám pak spadli především ti respondenti, kteří oproti ostatním preferují sledování videí na PC a nejméně využívají YTB na svých telefonech. Nejvýraznější data v grafu však patří našim dvěma nejmenším klastrům, kdy klaster D hodnotí YouTube celkově negativně, ačkoli se v konzumaci obsahu příliš neliší od zbylých skupin. Do posledního z klastrů pak připadly osoby, které ve srovnání s ostatními konzumují největší množství YTB obsahu.

Obr. 2 – Graf průměrných odpovědí jednotlivých klastrů v t-skórech



Metoda zároveň nabízí i možnost ohodnotit jednotlivé respondenty podle toho, jak blízko jsou jejich odpovědi centroidu příslušného klastru, tedy jak moc se jejich odpovědi podobají „průměrnému členu klastru“. Tento údaj dále označujeme jako „vzdálenost“. Menší vzdálenost pak znamená, že je respondent lepším reprezentantem svého klastru. Pro ilustraci jsme do tabulky 2 zanesli odpovědi nejreprezentativnějšího a nejméně reprezentativního člena každého z klastrů, který je však stále dle použité metody lepším reprezentantem jednoho klastru nežli klastru jiného.

Tab. 2 – Odpovědi vybraných reprezentantů jednotlivých klastrů a jejich vzdálenost od centroidu

Číslo respondenta	920	811	204	230	760	782	164	201	3	191
Příšerný – Úžasný	6	7	5	5	5	5	4	4	6	5
Špatný – Dobrý	7	7	6	6	6	5	4	7	5	5
Nudný – Zábavný	7	7	6	5	6	5	4	4	6	5
Nekvalitní – Kvalitní	7	7	5	6	6	6	4	5	7	5
Neužitečný – Užitečný	7	7	6	6	6	7	4	7	6	5
Odebíraných kanálů	24	19	5	0	25	300	1	0	159	0
Kanálů zhlédnuto	7	4	2	5	8	14	10	0	157	150
Videí zhlédnuto	67	33	15	32	74	20	19	0	605	500
PC	20 %	0 %	10 %	0 %	98 %	100 %	70 %	0 %	100 %	0 %
Tablet	5 %	0 %	10 %	0 %	2 %	0 %	0 %	0 %	0 %	0 %
Smartphone	45 %	0 %	70 %	25 %	0 %	0 %	10 %	0 %	0 %	0 %
TV	30 %	0 %	10 %	0 %	0 %	0 %	20 %	0 %	0 %	0 %
Jiné	0 %	100 %	0 %	75 %	0 %	0 %	0 %	100 %	0 %	100 %
Klastr	A	A	B	B	C	C	D	D	E	E
Vzdálenost	0,25	2,65	0,23	2,01	0,18	1,42	0,46	2,69	0,67	2,74

Zdroje

- Brodersen, A., Scellato, S., & Wattenhofer, M. (2012). *Youtube around the world: Geographic popularity of videos*. 241–250.
- Galak, J. (nedat.). *Data Demystified—YouTube*. Získáno 10. duben 2023, z <https://www.youtube.com/@DataDemystified>
- Chatzopoulou, G., Sheng, C., & Faloutsos, M. (2010). *A first step towards understanding popularity in YouTube*. 1–6.
- Sedláček, J. (2016). *The Rise of the YouTube Celebrity: The Migration of Young Audiences from TV to Independent Content Creators* [Diplomová práce, Univerzita Karlova]. <https://dspace.cuni.cz/bitstream/handle/20.500.11956/82956/120240856.pdf?sequence=1&isAllowed=y>
- Statista. (nedat.). *Biggest social media platforms 2023*. Získáno 10. duben 2023, z <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>
- Welbourne, D. J., & Grant, W. J. (2016). Science communication on YouTube: Factors that affect channel and video popularity. *Public understanding of science*, 25(6), 706–718.
- Wikipedia. (2023). *YouTube*. https://en.wikipedia.org/w/index.php?title=YouTube&oldid=1149050950#cite_note-2