

Vizuální gramotnost

VISK2, UISK FF UK, Knihovna AVČR

David Jiří Šlosar

3.4. 2024

Kontaktní údaje

Jméno: Mgr. David Jiří Šlosar

Email: davidjiri.slosar@ff.cuni.cz; slosar@lib.cas.cz



ÚSTAV INFORMAČNÍCH STUDIÍ
A KNIHOVNICTVÍ
Filozofická fakulta
Univerzita Karlova



KNIHOVNA
AKADEMIE VĚD ČR

Kdo jste?

Obsah kurzu

- Co jsou data a jak s nimi pracujeme.
- Statistika
- Grafy
- Techniky a vizualizace

Proč to děláme?

Chceme získat informaci z dat.

Chceme efektivně sdělit informaci...

... nebo vyvolat emoce.

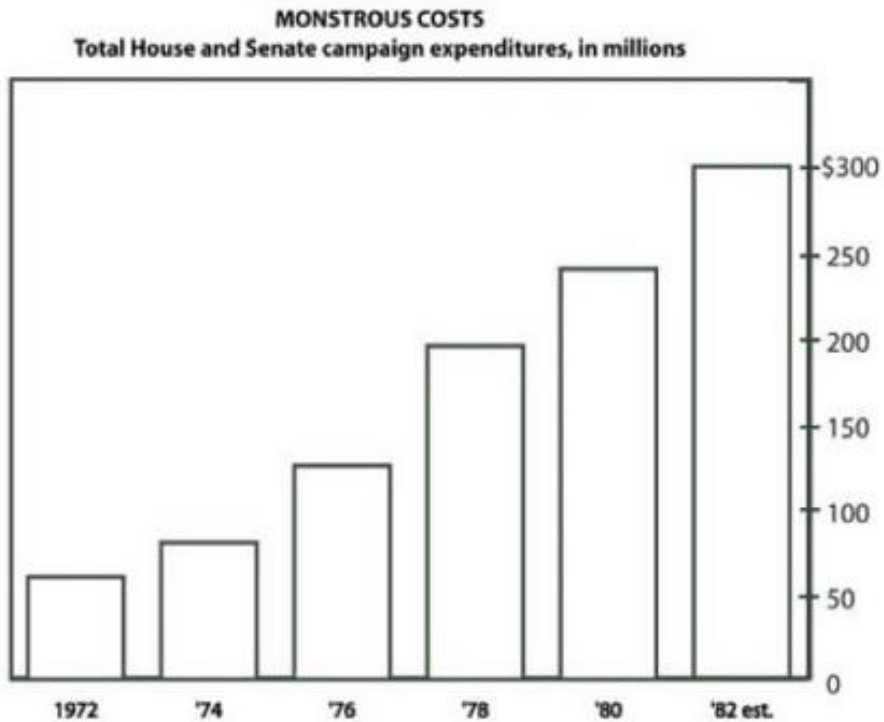
Nechceme manipulovat.



Dvě školy

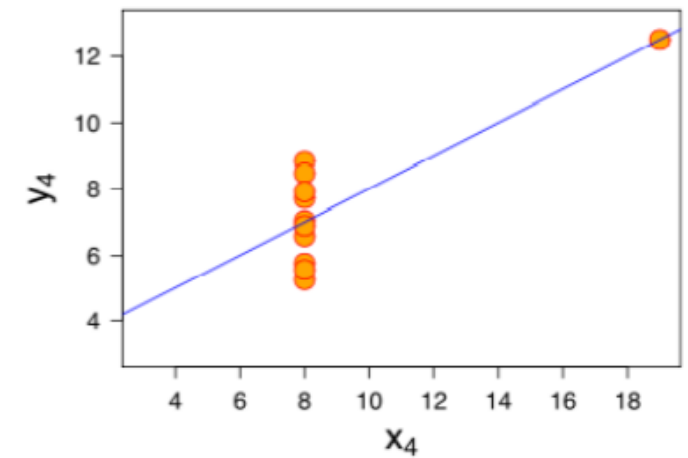
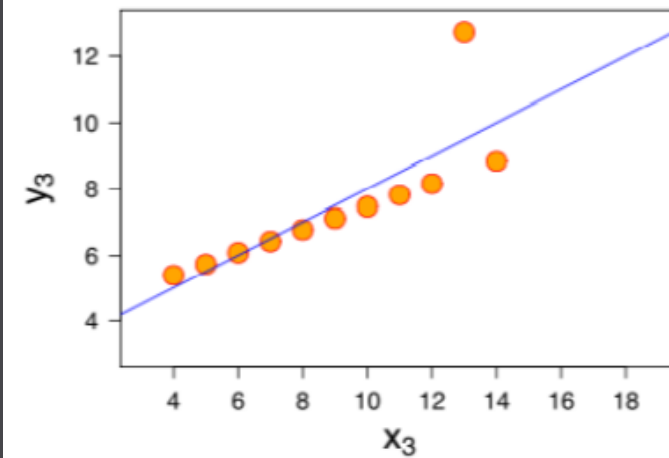
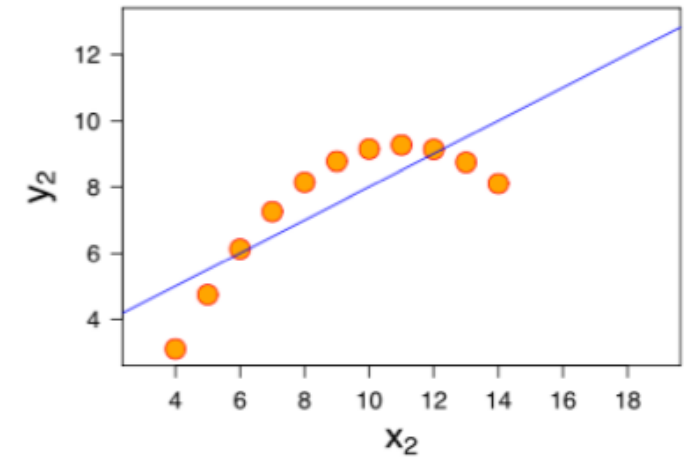
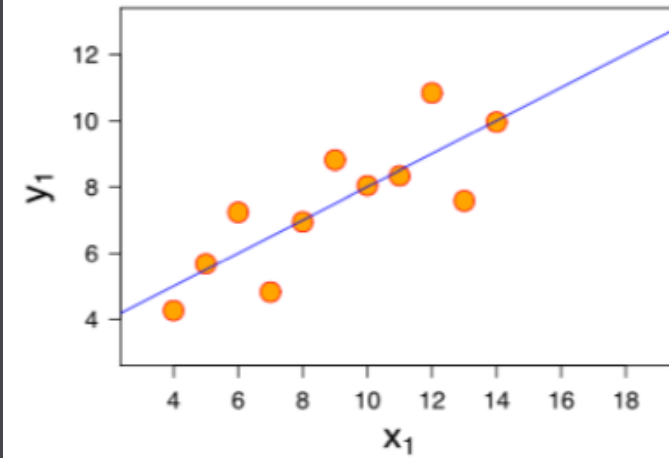
• Ben Shneiderman

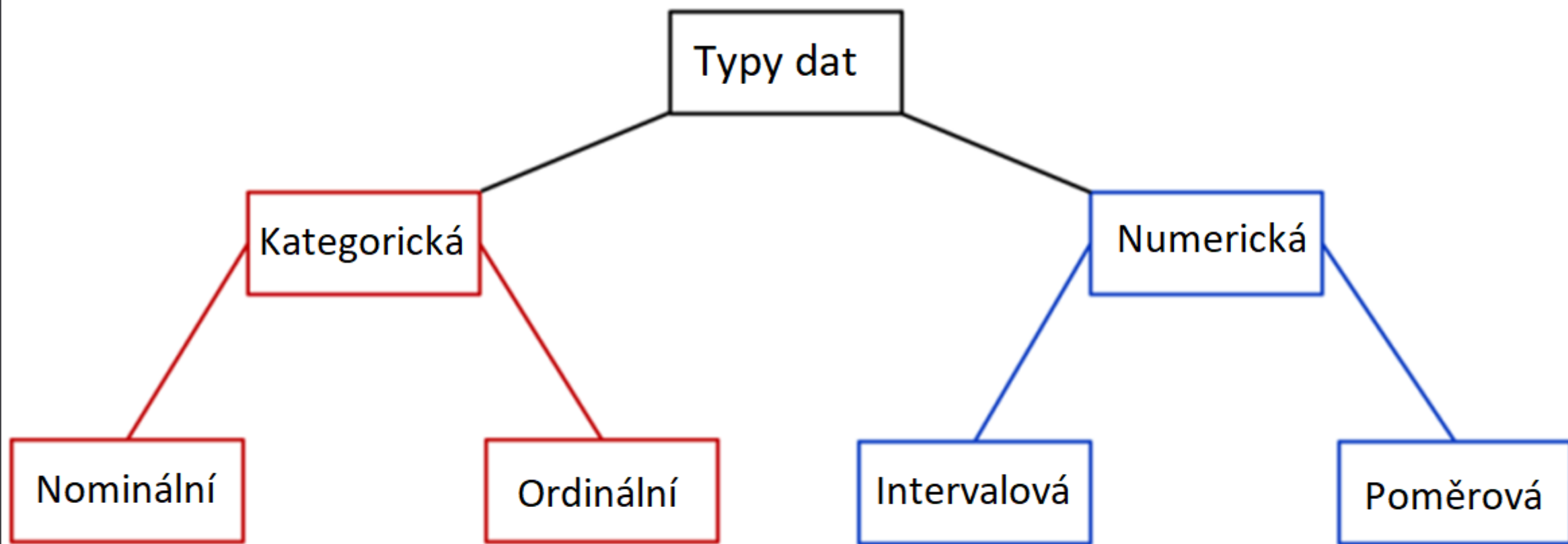
Nigel Holmes



Anscombův kvartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89





Typy dat

- Nominální – jména, vlastnosti.
- Ordinální – mohou být seřazena, ale neznáme velikost rozestupů.
- Intervalová – arbitrární nula, nemohou být násobena.
- Poměrová – bez omezení.
- Dichotomická – ano/ne, 1/0, pravda/nepravda
- <https://youtu.be/hZxznzfnt5v8>

Excel pro zpracování dat

Pro:

- Široce rozšířený.
- Základní obsluhu zvládne téměř každý.
- Umožňuje operace nad jednotlivými záznamy i jejich hromadné zpracování.
- Obsahuje VBA, lze tedy do jisté míry implementovat replikovatelnost operací.

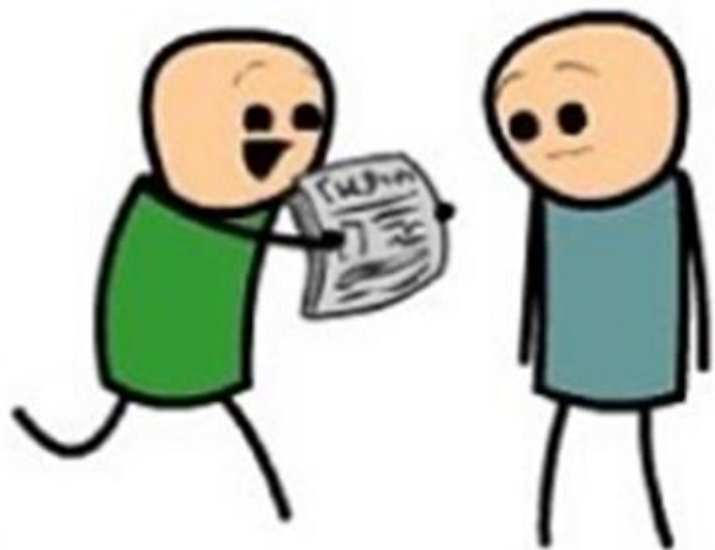
Proti:

- Pracovní postup je obtížně replikovatelný.
- Udává se, že až 30 % datových sad vědeckých studií obsahují alespoň jednu chybu způsobenou Excelem.
- Není sjednocený mezinárodní standard dat.
- Jedná se o komerční SW a nesplňuje tak myšlenku Open Science.
- Neumožňuje práci s velkými objemy dat (více než milion záznamů).

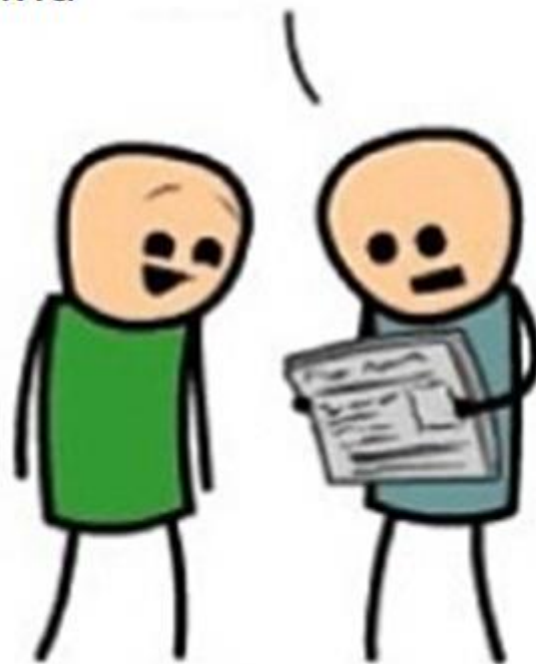
Kvantifikace

- Převod reálného a komplexního světa do jednoho či několika málo pojmů či číselných veličin
- Pouze úzký aspekt reality
- Kontext je stále nezbytný.
- Vstup pro statistiku.

Zatraceně, chlape!
Podívej se na tohle!



"Studie zjistila, že pro 50 %
populace je statistika
nudná"



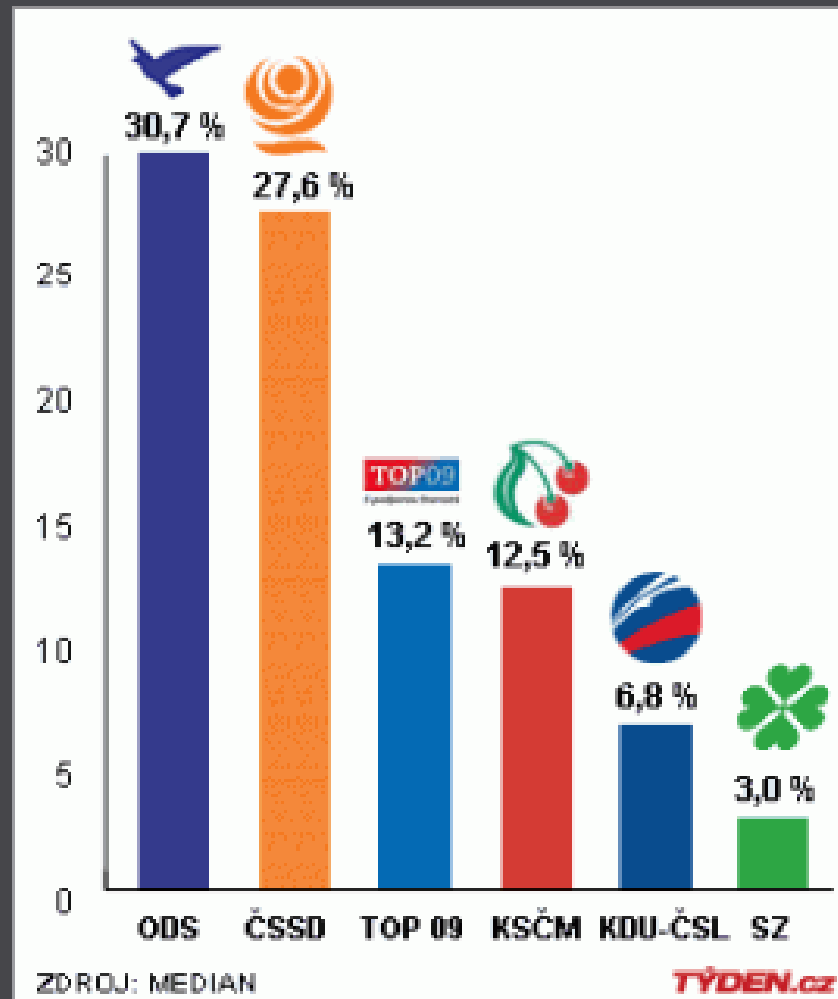
Cyanide and Happiness © Explosm.net

Četnosti

- Absolutní a relativní četnosti
- Osoba A vypila 5 silných piv (0,5l, 6 % obj. alkoholu).
- Osoba B vypila 8 ovocných piv (0,5l, 2,5 % obj. alkoholu).
- Osoba A vypila absolutně nižší množství tekutin, ale s vyšším relativním podílem alkoholu. Osoba A je relativně opilejší.

Procento vs. procentní bod

- Je mnohem snazší a v kontextu správnější uvést, že ODS by měla získat o 3,1 procentního bodu víc než ČSSD.
- ODS získala přibližně o 11,1 % víc hlasů.

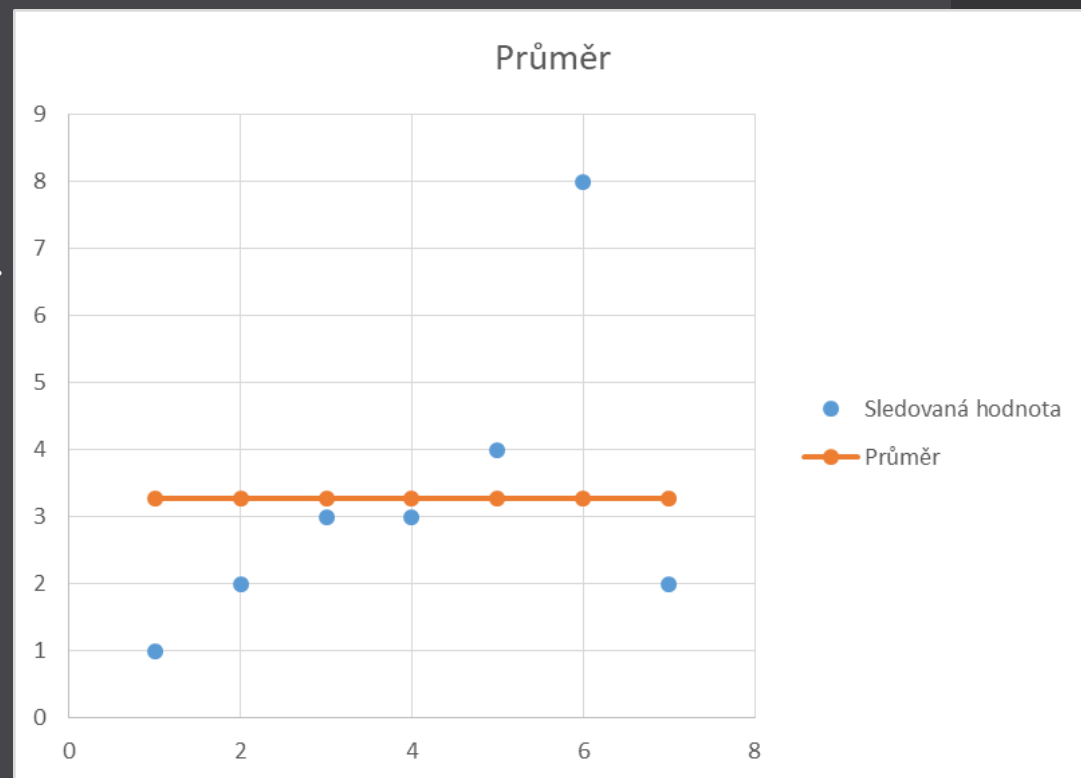


Průměr (mean, average)

- Aritmetický, (vážený), geometrický, harmonický.

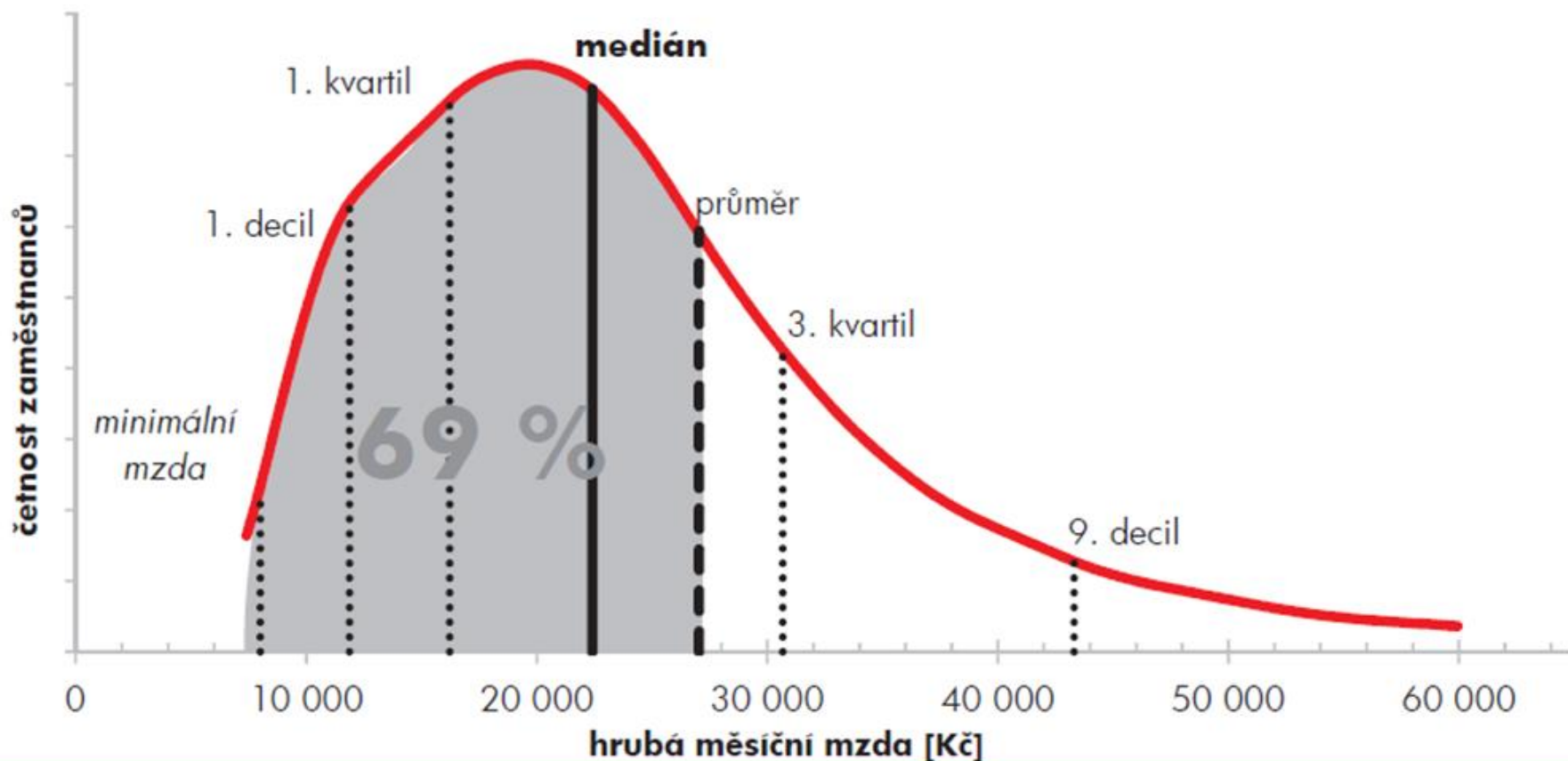
$$\bar{x} = \frac{1}{n} (x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i,$$

- Kde x_1 až x_n jsou zjištěné hodnoty a n je jejich počet.
- Součet odchylek od průměru je nulový.
- Je silně ovlivněn odlehlými hodnotami.



Distribuce mezd v roce 2014

Distribuce mezd



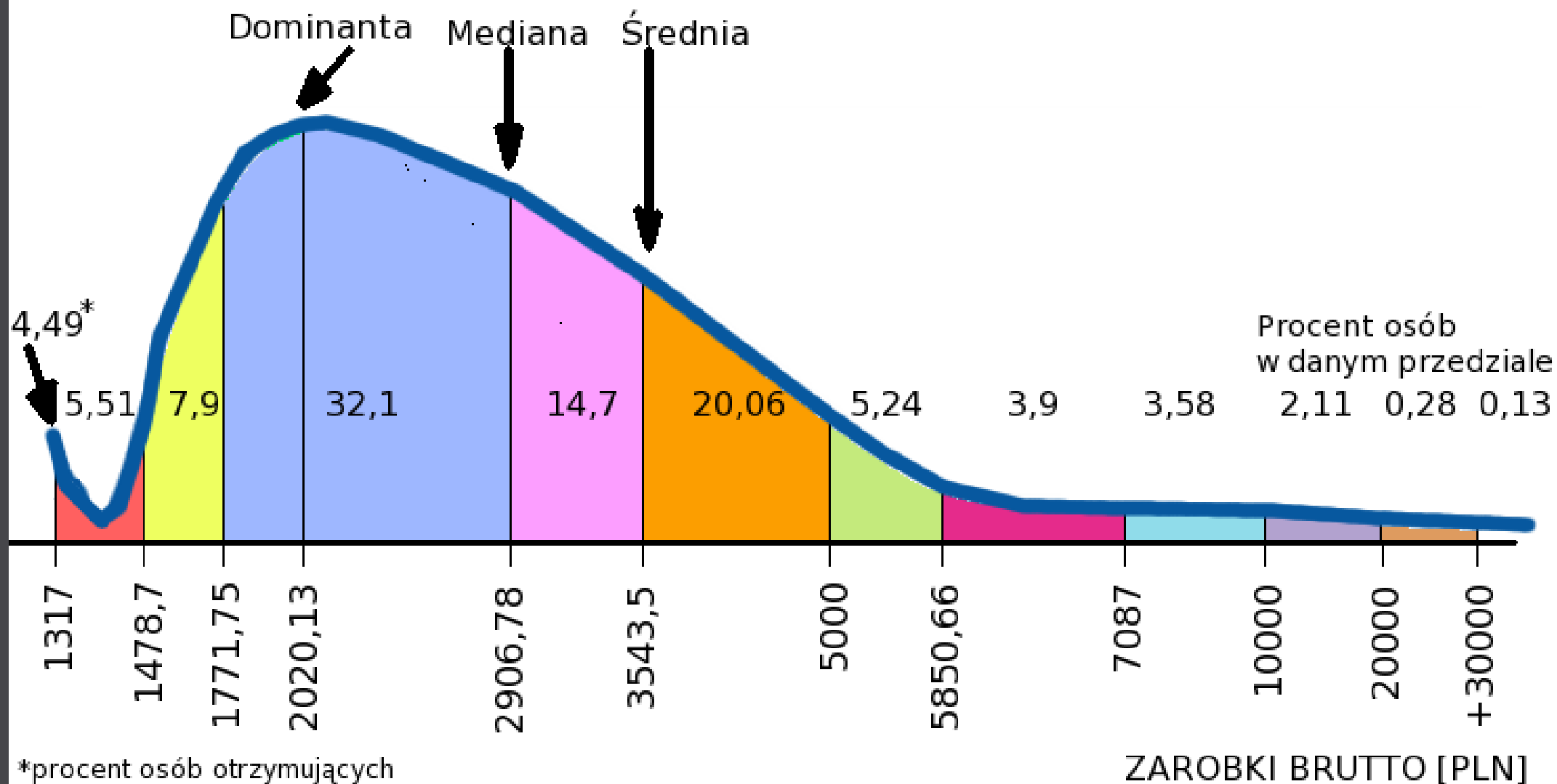
WWW.ISPV.CZ

ČMKOS



Procent osób zarabiających dane kwoty brutto

Na podstawie danych GUS za 2010 rok, like-a-geek.jogger.pl



*procent osób otrzymujących wynagrodzenie minimalne

Průměr průměrů jednotlivin není průměr celku.

Publikace autora 1	Citací	Publikace autora 2	Citací
1. publikace	120	1. publikace	1
2. publikace	6	2. publikace	2
3. publikace	5	3. publikace	1
4. publikace	8	4. publikace	1
5. publikace	3	5. publikace	0
6. publikace	12	6. publikace	0
7. publikace	7		
8. publikace	3		
9. publikace	0		
10. publikace	0		
Průměr:	16,4	Průměr:	0,833333
Suma:	164	Suma:	5
Průměr průměrů:	$(16,4 + 0,833)/2 = 8,616666667$		
Průměr celku:	10,5625		

Medián

Je hodnota prostřední statistické jednotky.

50 % hodnot se nachází nad ním a 50 % pod ním.

Výpočet je ovlivněn sudostí či lichostí statistického souboru.

Musí být vždy seřazen podle velikosti, aby bylo možné jej vypočítat.

Je součástí kvantilů.

Kvantily

Medián (50%), kvartil (25 %, 50%, 75 %), decil (10 %, 20%... 90 %), percentil atd.

Rozdělení statistického souboru na x stejných dílů.

Vhodné použít při silně zešikmených distribucích.

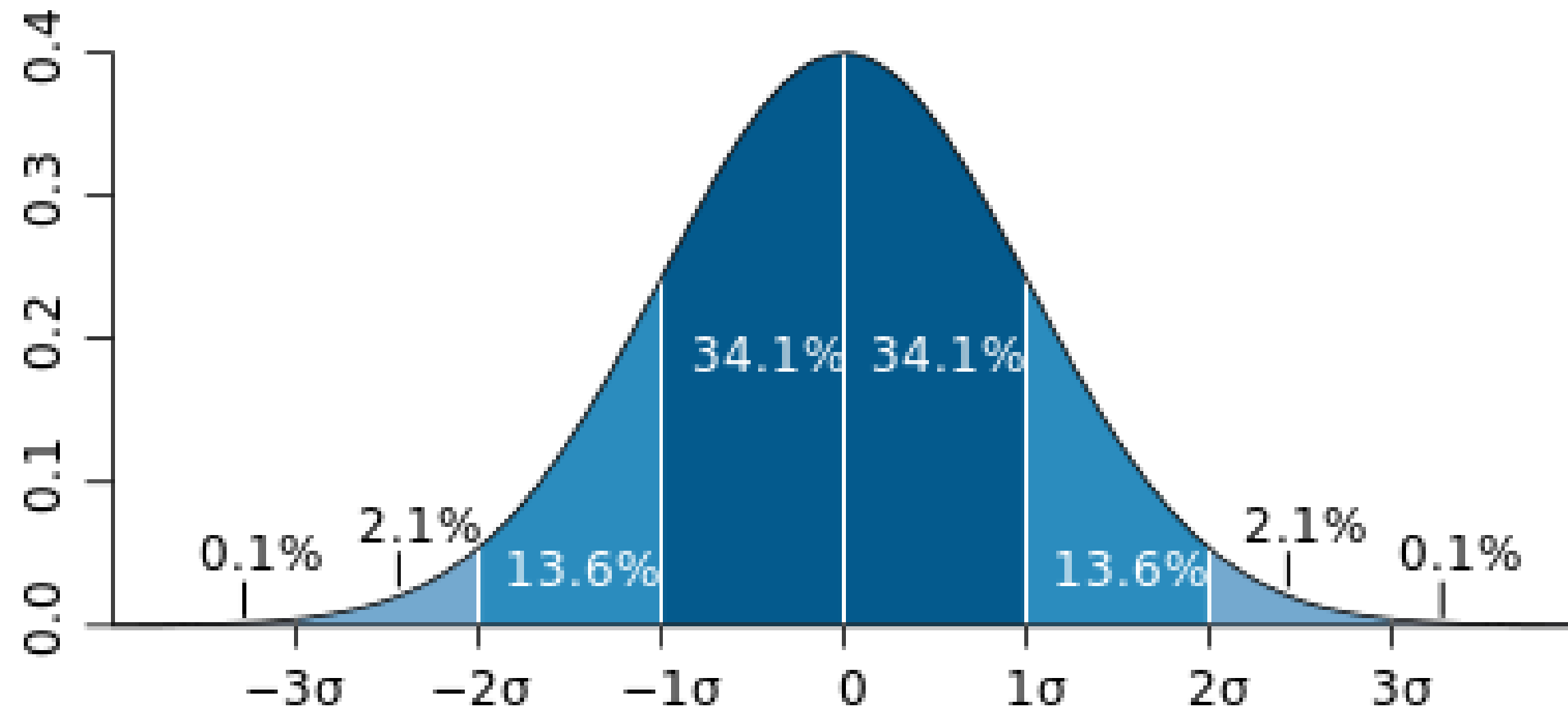
Je možné stanovit nejen rozdělení souboru, ale i polohu konkrétního znaku -
=`percentrank.inc(oblast;hodnota)`,
případně `*.exc`.



Modus

Nejčastěji se vyskytující hodnota.

0, 0, 1, 1, 1, 2, 3, 4, 5, 6, 6, 6, 6, 6, 6, 7, 8, 9 => modus = 6



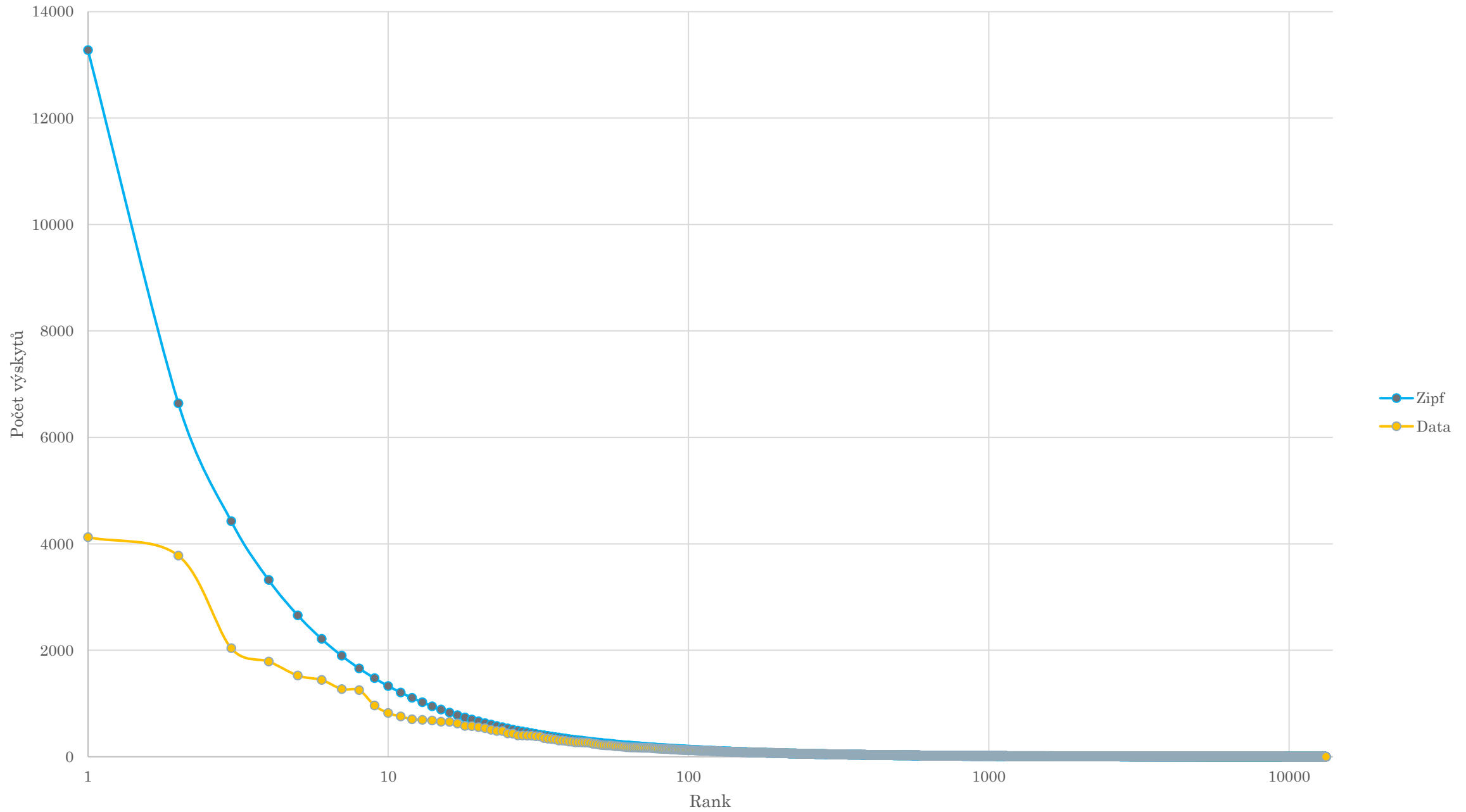
Zipfova transformace

- Kniha: Harry Potter a kámen mudrců
- Znaků: 441 610
- Slov: 74 791
- Unikátních slov: 13 277

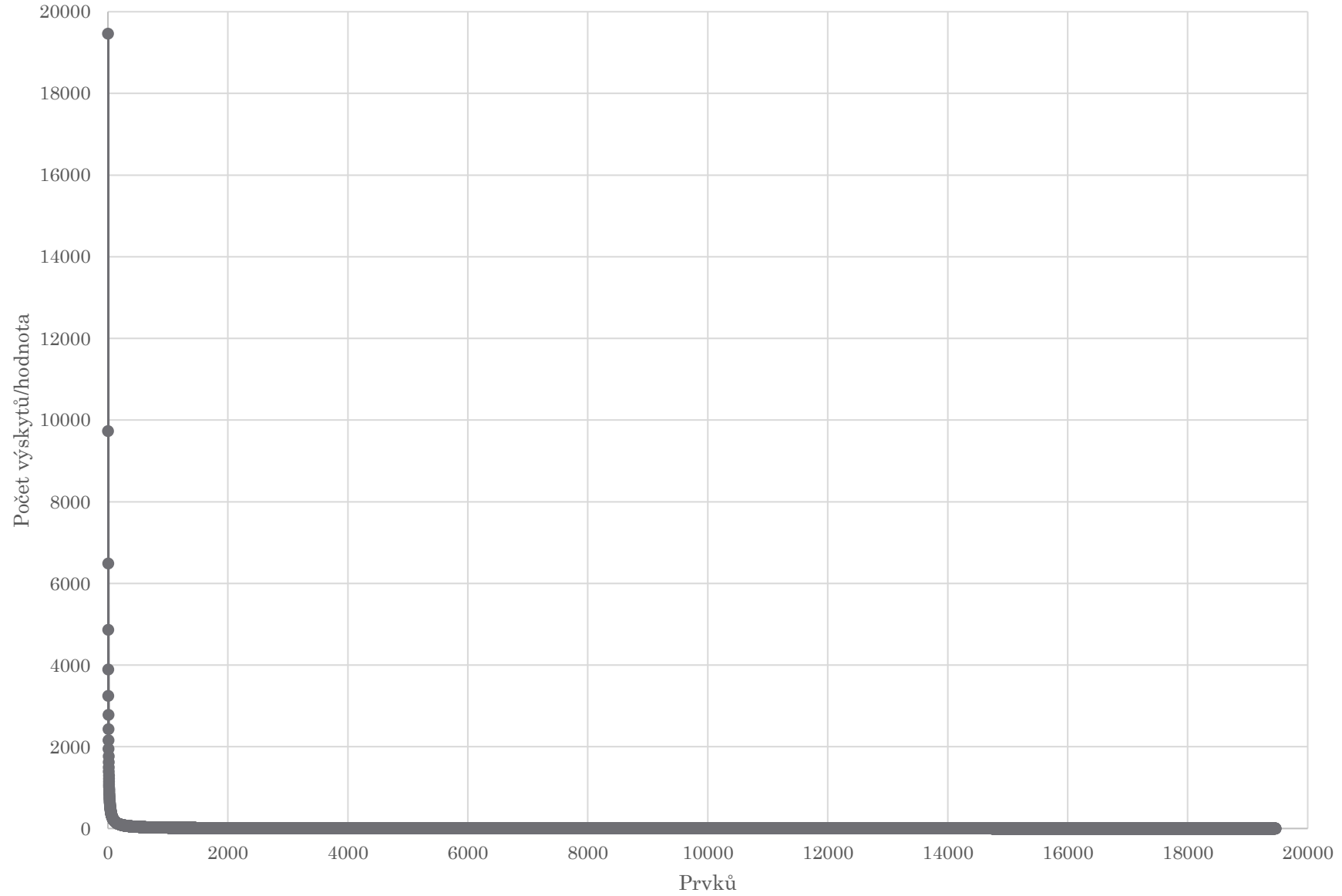
Rank	Slovo	Počet výskytů
1	a	4124
2	se	3778
3	na	2038
4	to	1789
5	že	1524
6	Harry	1442
7	si	1271
8	v	1250
9	je	963
10	do	820
11	s	757
12	z	704
13	ho	692
14	Když	681
15	Co	658
16	už	655
17	mu	622
18	jako	578
19	jak	575
20	ale	554

Rank	Slovo	Počet výskytů
1	Harry	1442
2	řekl	505
3	ještě	434
4	Ron	398
5	Hagrid	393
6	jsem	383
7	aby	381
8	byl	348
9	něco	336
10	měl	284
11	jeho	281
12	Já	258
13	Harryho	240
14	Hermiona	233
15	bylo	219
16	byla	196
17	vůbec	196
18	všichni	188
19	strýc	176
20	Brumbál	167

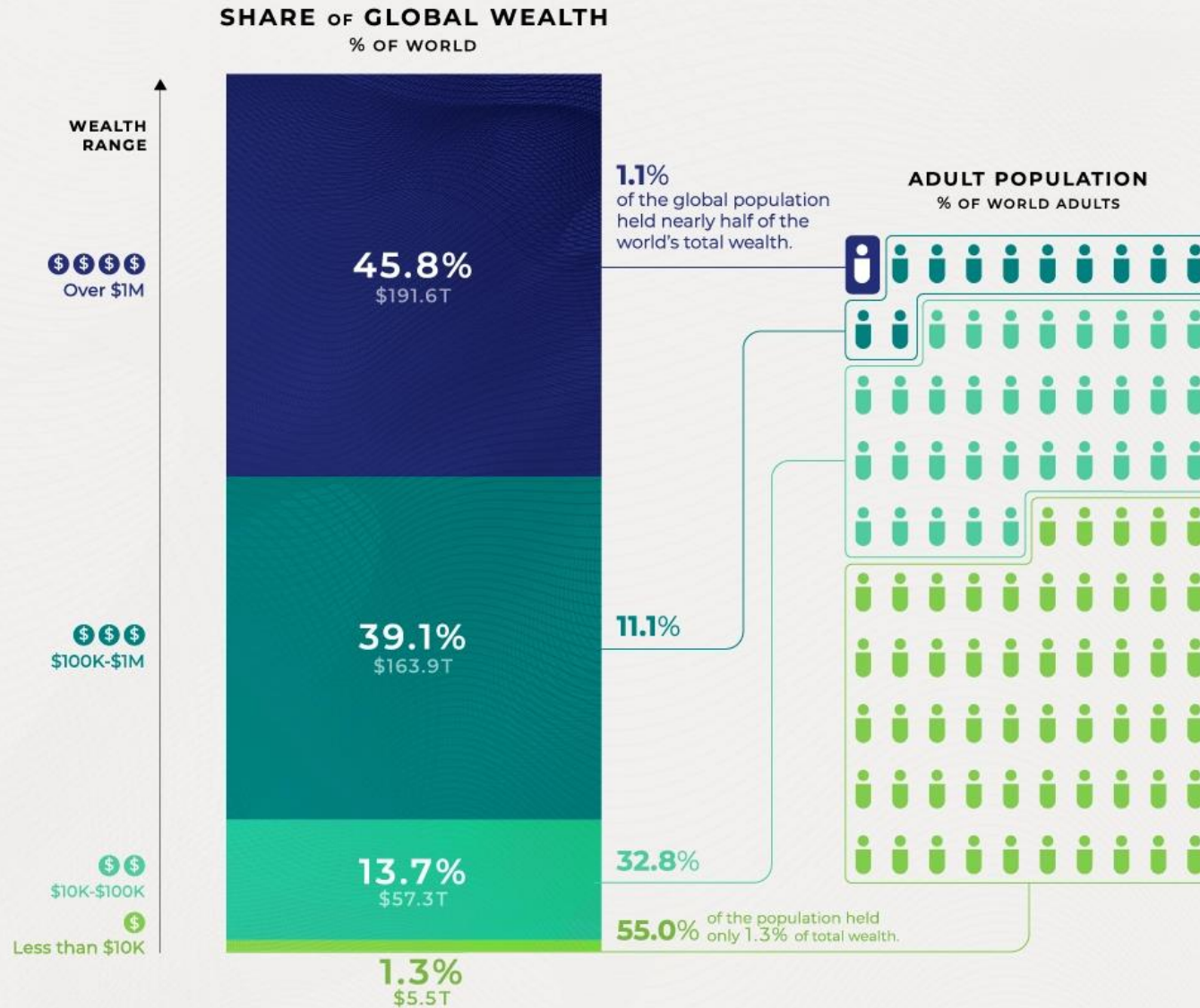
Porovnání Zipfovy transformace knihy s ideálním Zipfovým zákonem

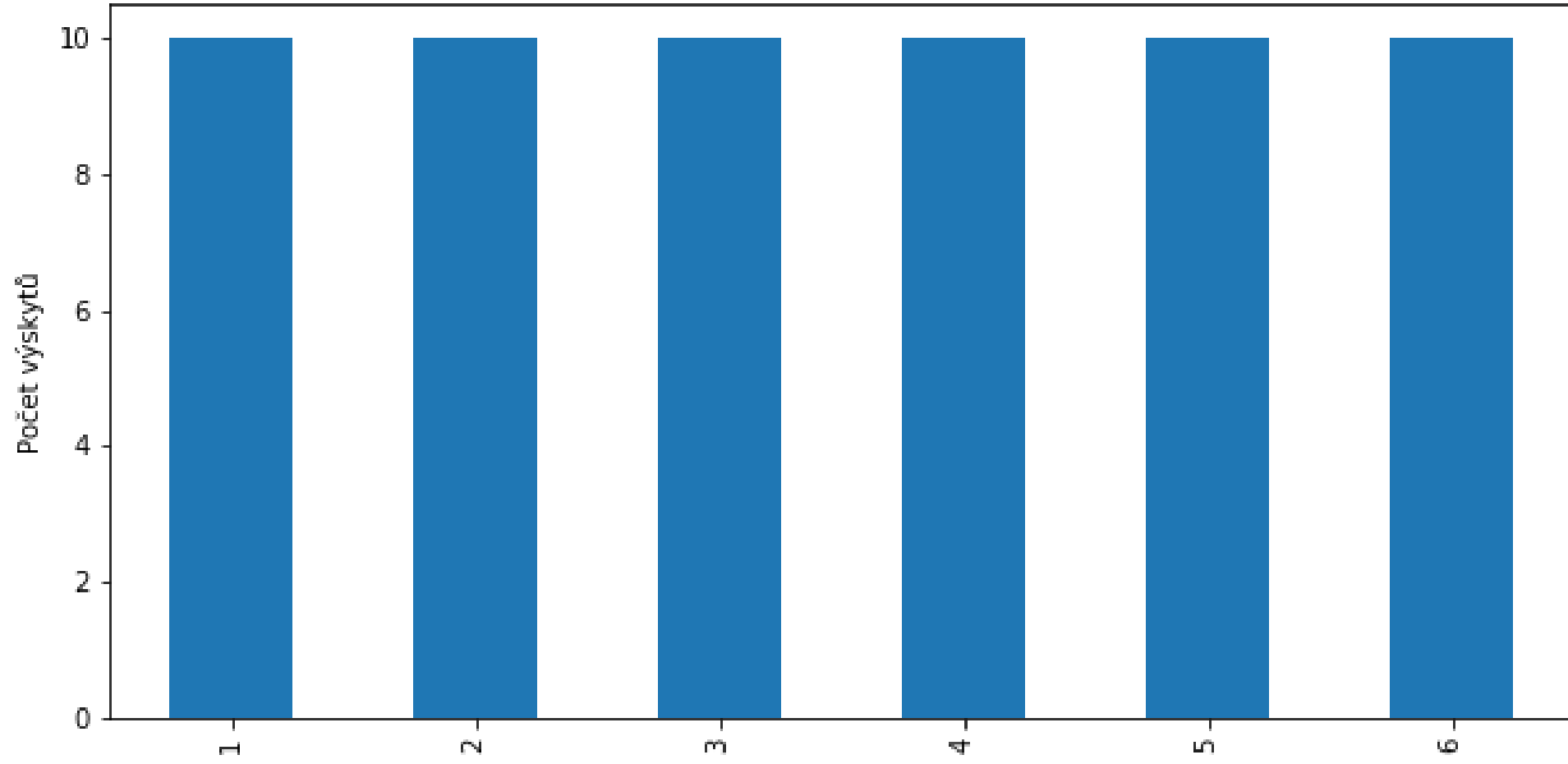


Zipf

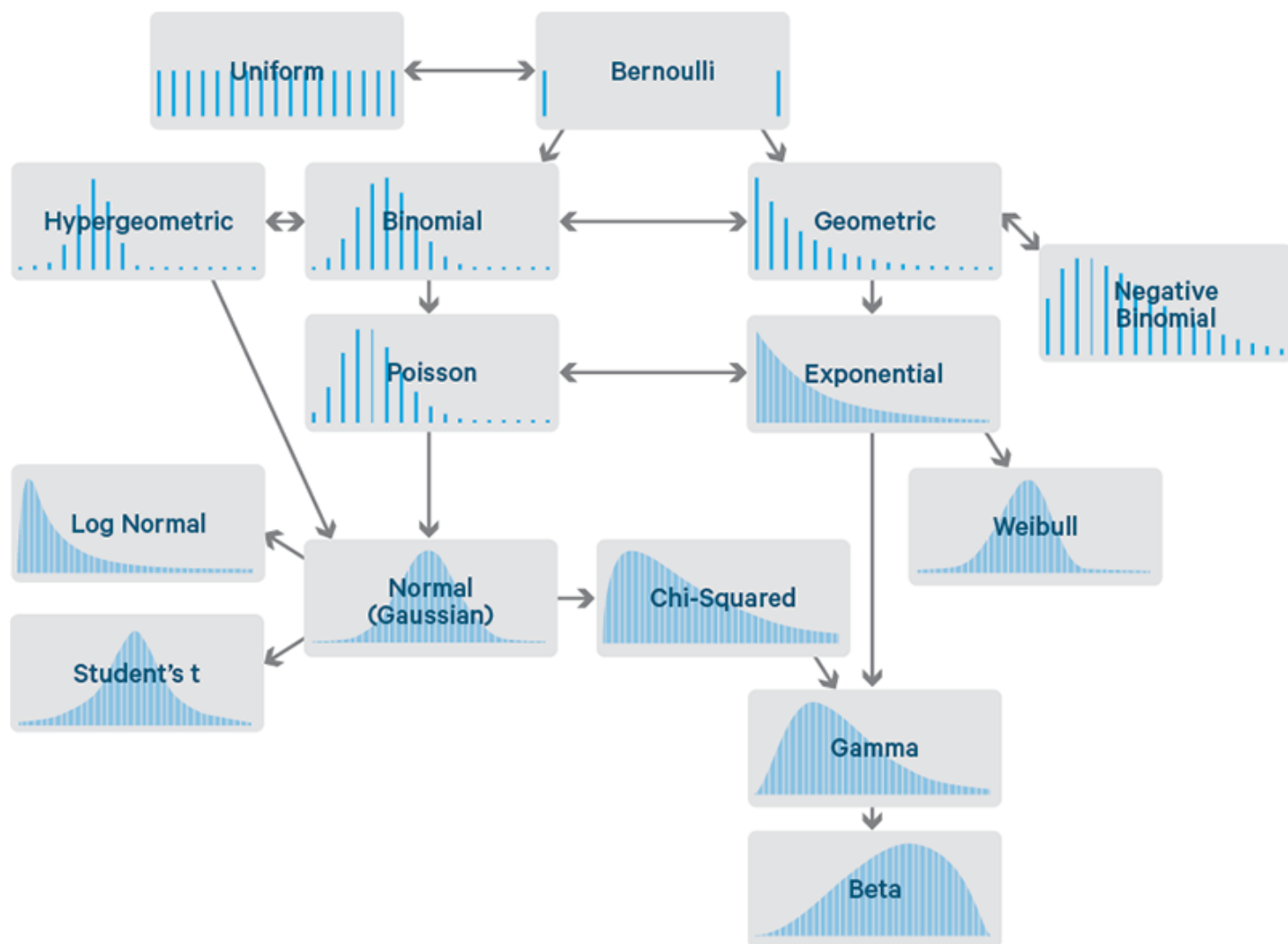


THE GLOBAL WEALTH DISTRIBUTION





Rozložení/distribuce/rozdělení



zdroj: towardsdatascience.com

Korelace

...neimplikuje kauzalitu!

Pokud je obsažena kauzalita, tak vyjadřuje míru vzájemného vztahu.

Pearsonův korelační koeficient.

Nabývá hodnot od -1 do 1.

Samotná korelace nemá matematickou teorii => stejná hodnota korelačního koeficientu má rozdílné významy v jiných kontextech.

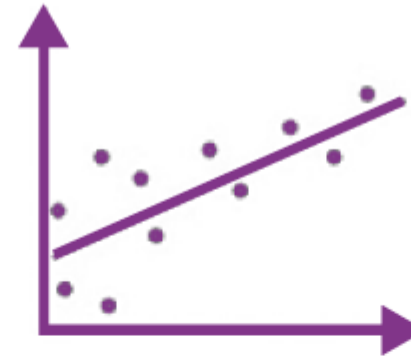
=CORREL()

Obvykle se jedná o Pearsonovu.

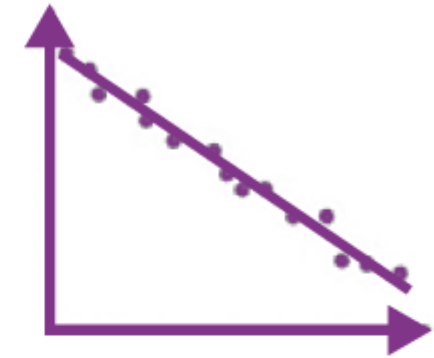
Existuje také Spearmanova pořadová korelace.



Strong positive correlation



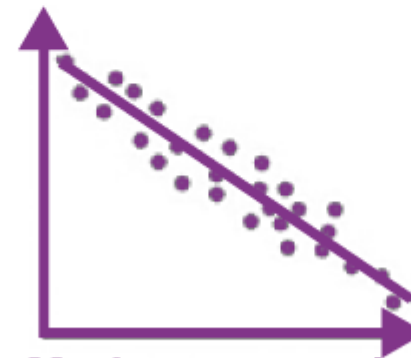
Weak positive correlation



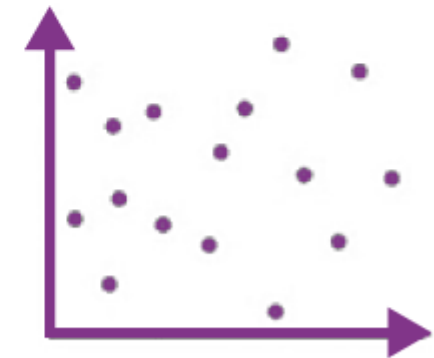
Strong negative correlation



Weak negative correlation



Moderate negative correlation



No correlation

Dříve jsem si myslel, že korelace implikuje kauzalitu.



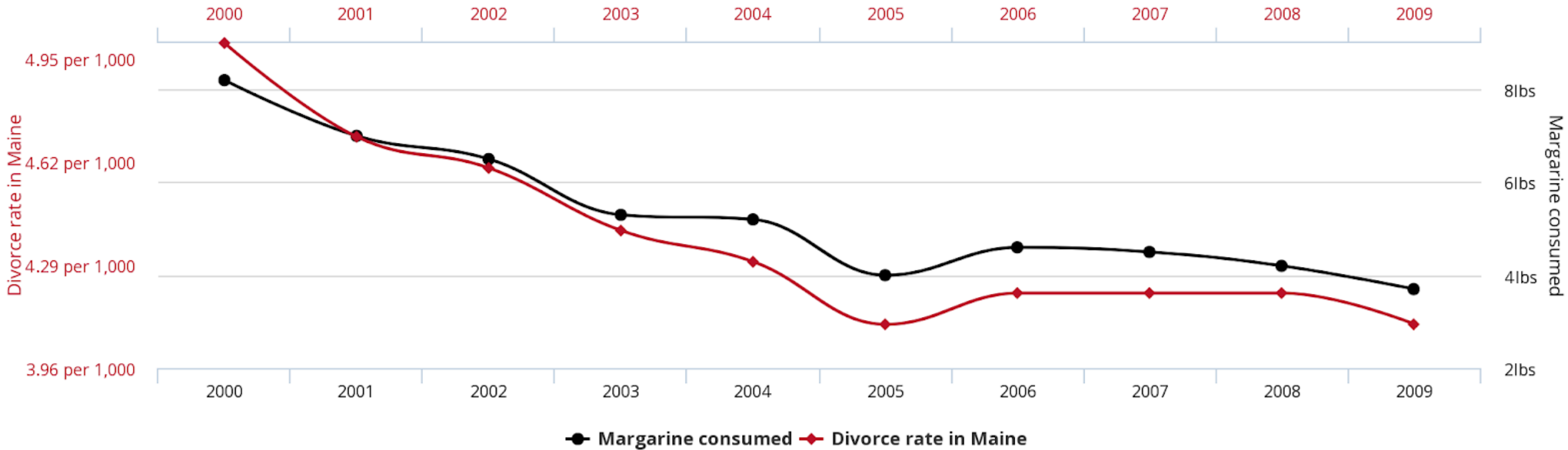
Po absolvování kurzu statistiky si to však nemyslím.



Zdá se, že ti ten kurz pomohl.



Divorce rate in Maine correlates with Per capita consumption of margarine

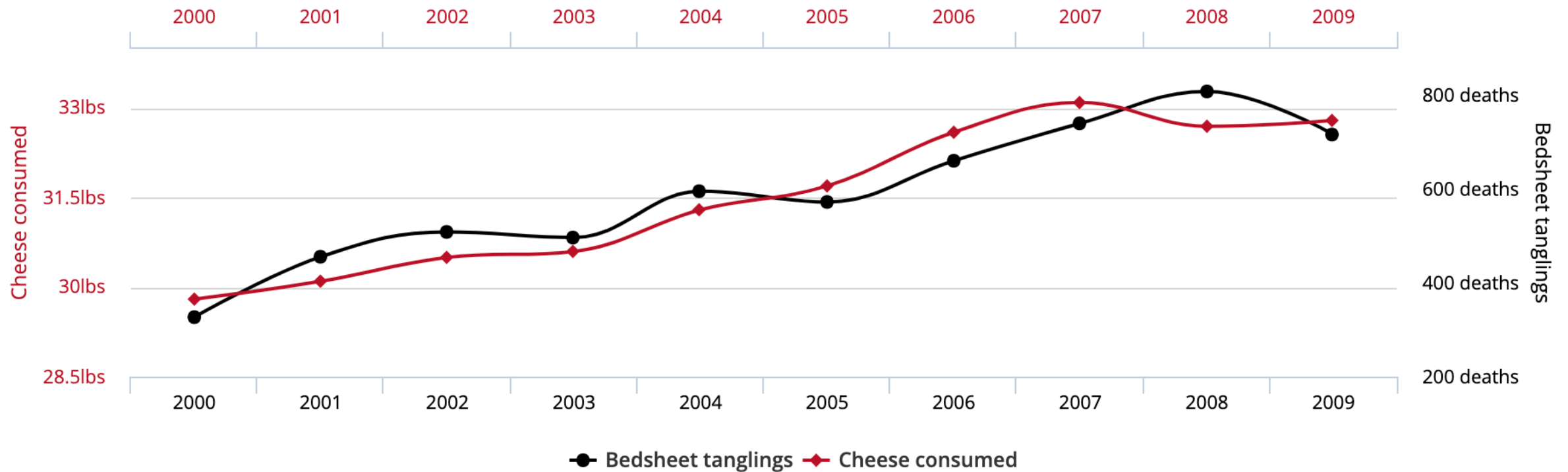


Per capita cheese consumption

correlates with

Number of people who died by becoming tangled in their bedsheets

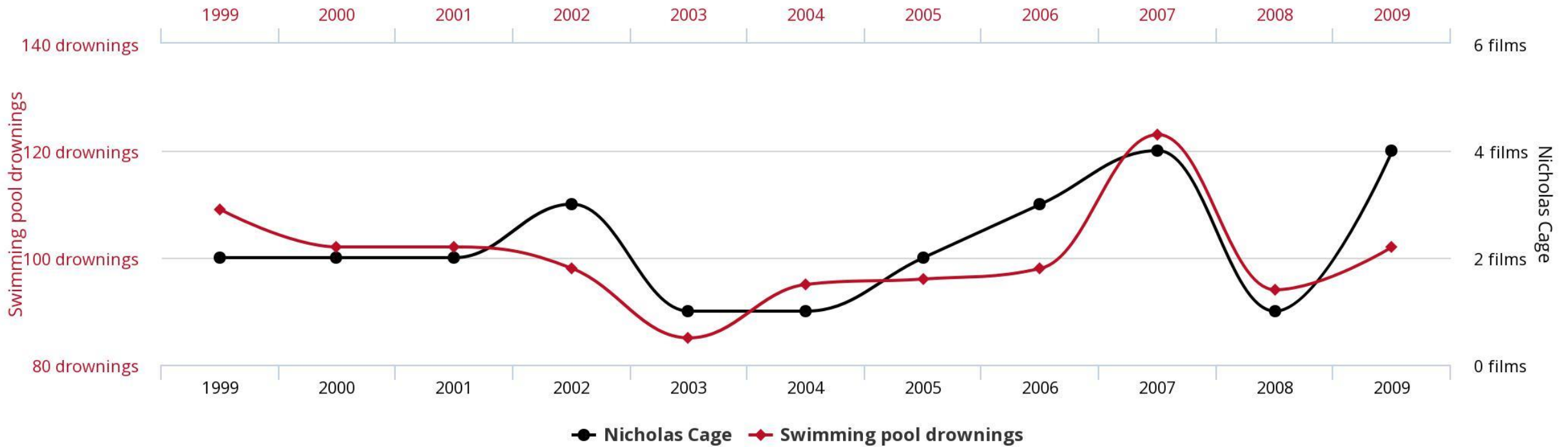
Correlation: 94.71% (r=0.947091)



tylervigen.com

Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in



tylervigen.com

Extrapolace, Interpolace

- Extrapolace je proces, kde z dostupných dat se snažíme zjistit budoucí vývoj nebo minulá stav. Snažíme se jít mimo hranice datasetu.
- Interpolace je proces, kde v rámci dostupných zjišťujeme mezilehlé body. Je zvyšována jemnost zobrazení.

Grafická prezentace

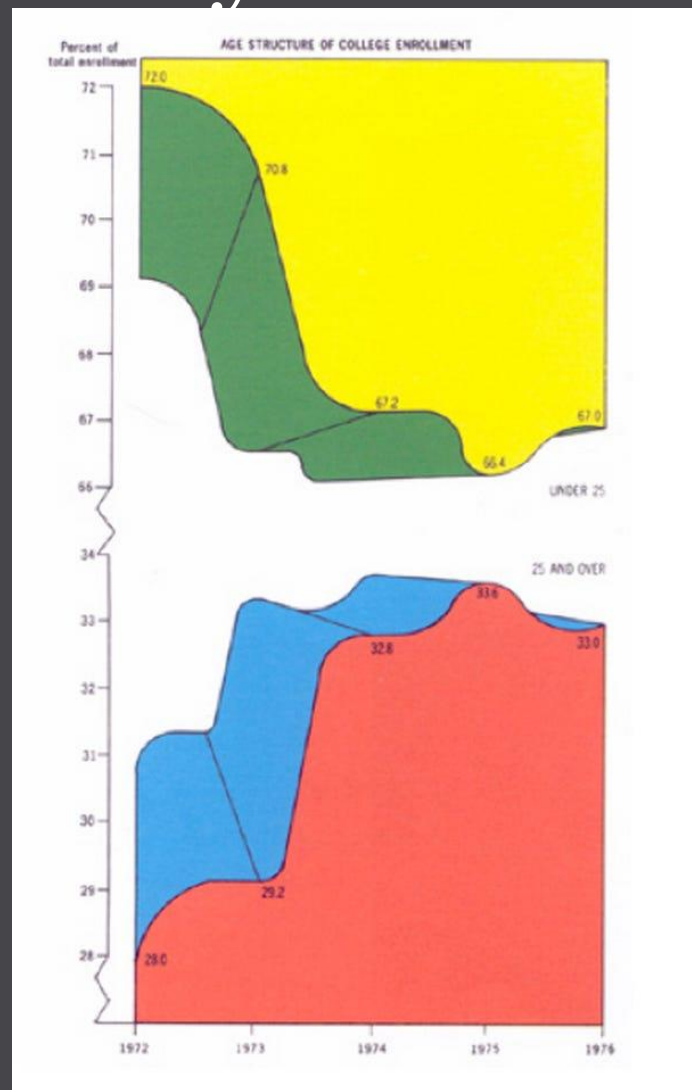
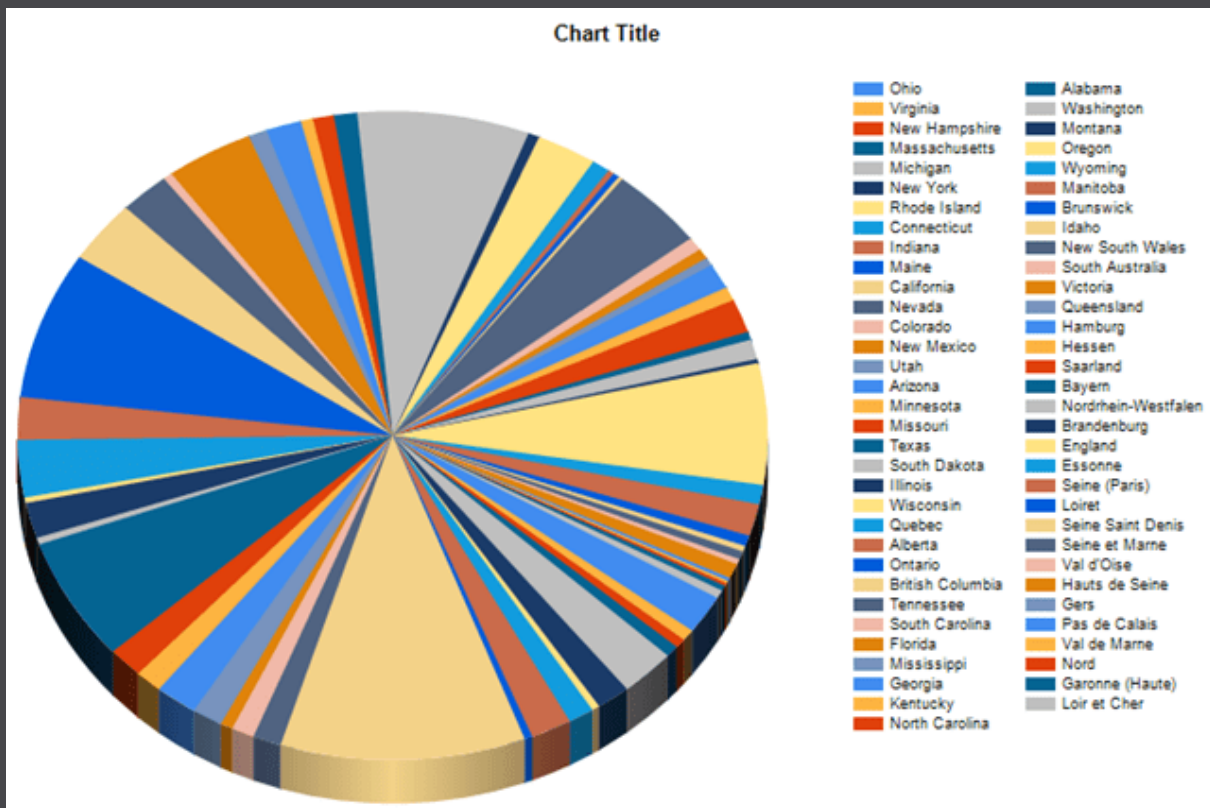
- Zanesením do grafu je možné vizuální reprezentace jinak nepřehledných číselných hodnot.
- Jsou více sugestivní a snáze interpretovatelné než číselné hodnoty.
- Je možné si vybrat z několika typů.

S grafem

- Obvykle rychlejší pochopení
- Větší emoční dopad
- Silnější efekt na paměť

Výsečový/paprskový/povrchový

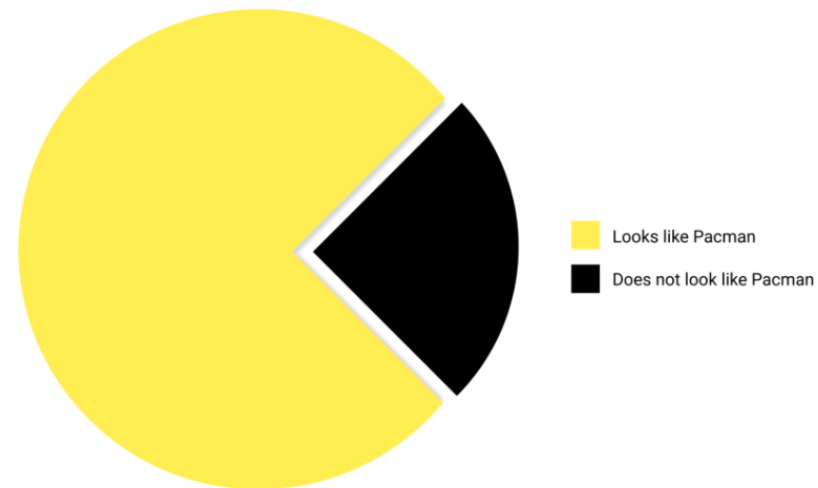
- Nepoužívat, pokud to alespoň trochu jde.



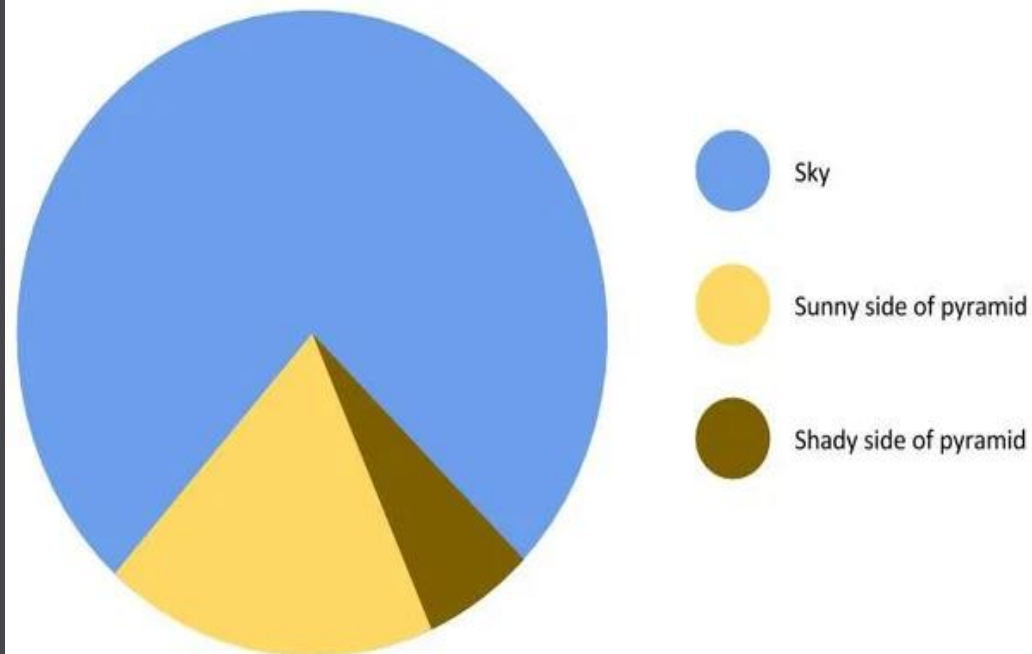
Koláčový graf

- Přítel nenechá svého přítele používat koláčový graf.

One of the few good reasons
to use a Pie Chart



Not all pie charts are bad! Image by the author.

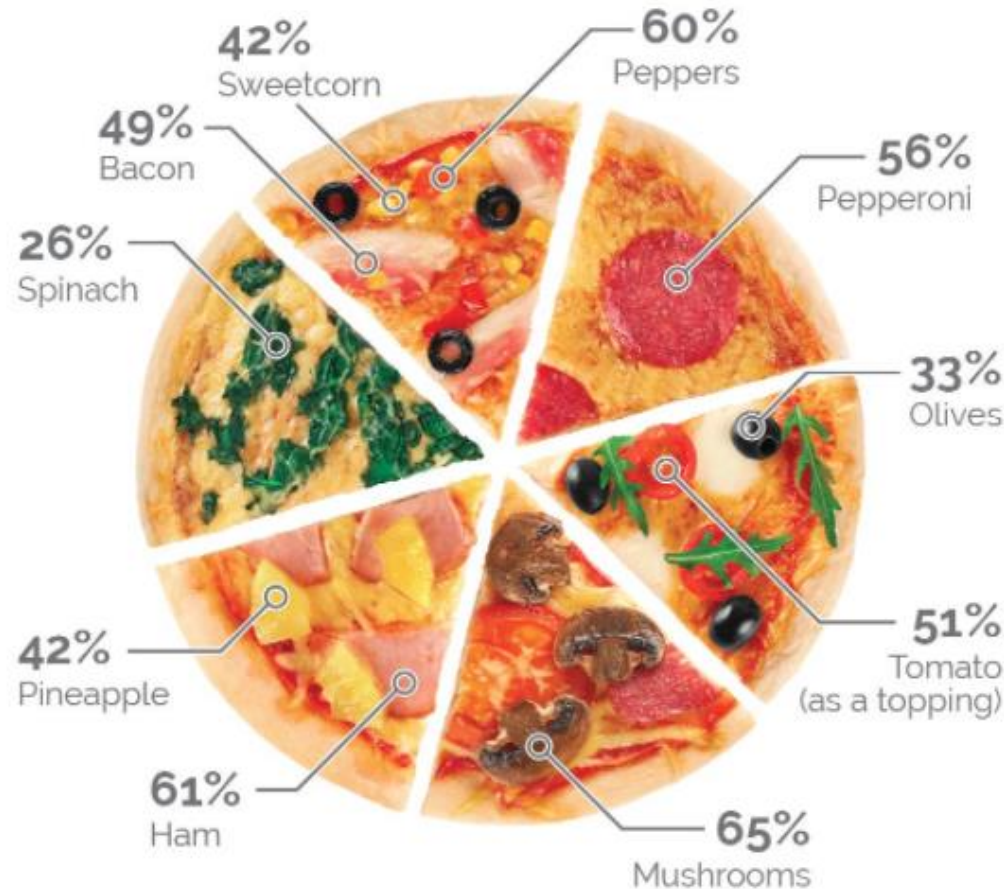




Forget pepperoni - mushroom is Britain's most liked pizza topping (65%), followed by onion (62%) and then ham (61%)

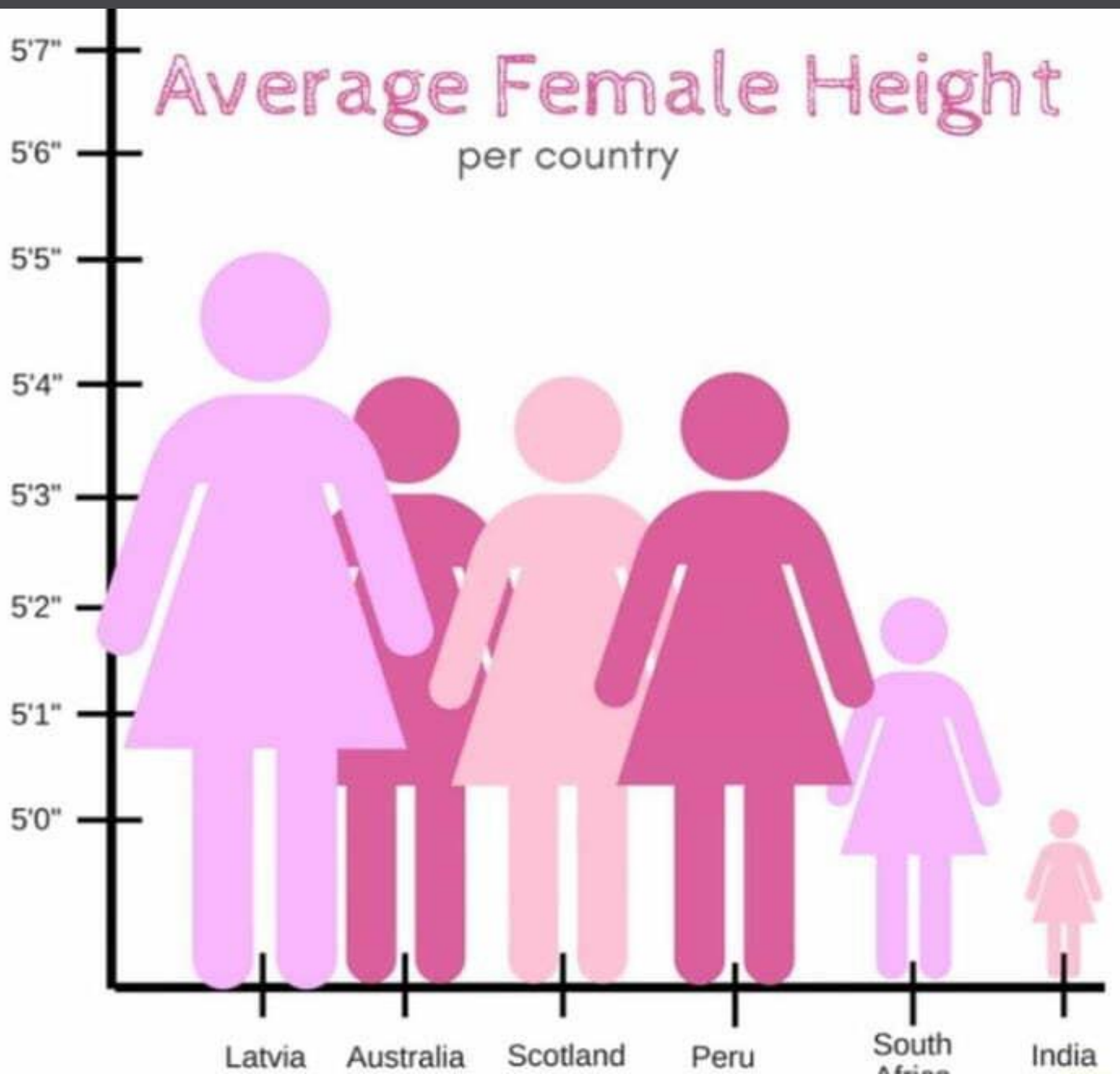
yougov.co.uk/news/2017/03/0 ...

Generally speaking, which of the following toppings do you like on a pizza? Select as many as you like



Other items not depicted include: onions (62%), chicken (56%), beef (36%), chillies (31%), jalapeños (20%), pork (20%), tuna (20%), anchovies (18%), 2% of people say they only like

Average Female Height per country





JOB LOSS BY QUARTER



FOX NEWS
FOX NEWS
.COM

SOURCE: BLS

AMERICA'S
NEWSROOM

N FAIRFAX, VA... BYRD WAS ADMITTED TO THE HC

S&P ▼ 3.08

UNEMPLOYMENT RATE

UNDER PRESIDENT OBAMA

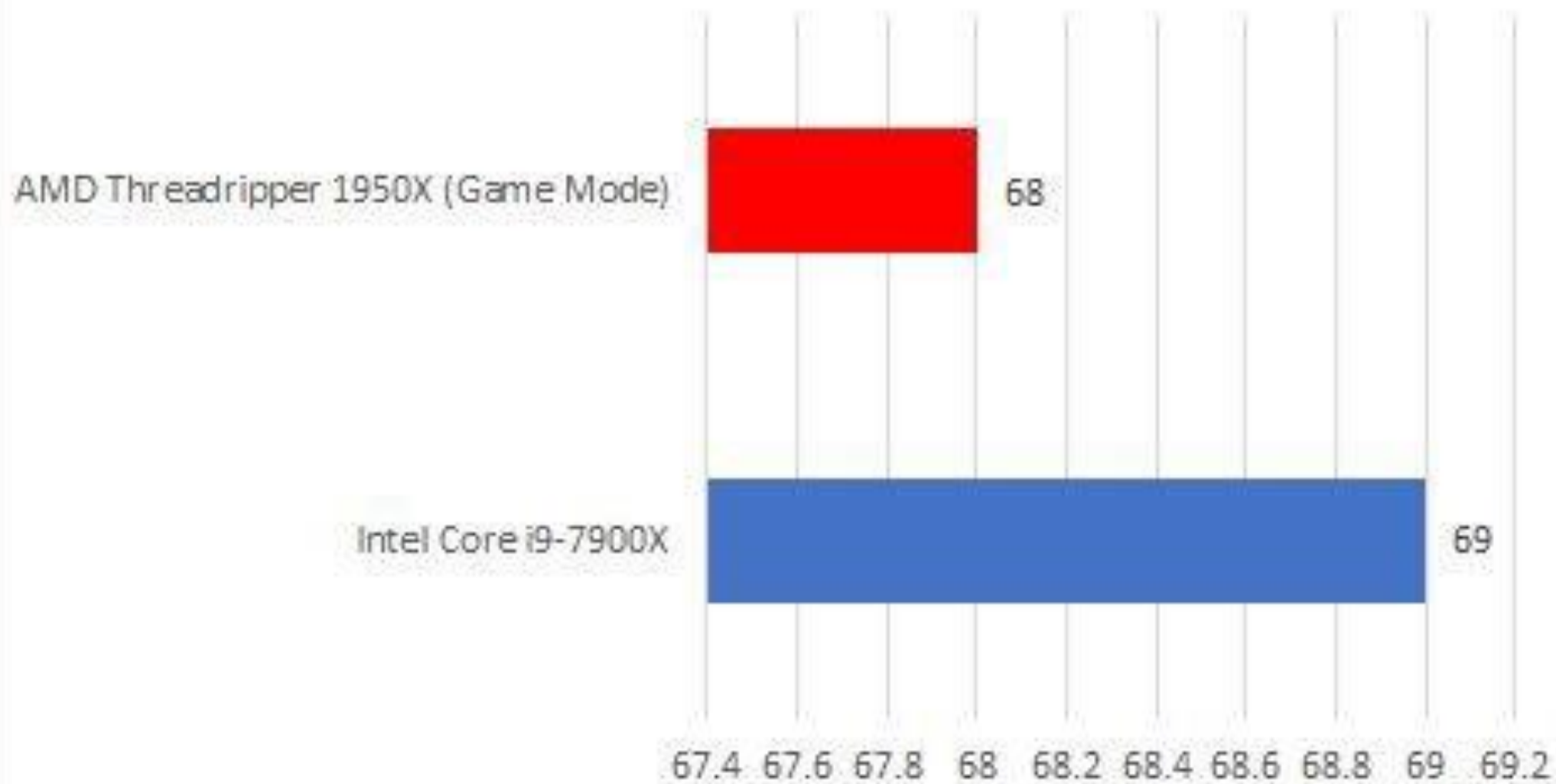


2011

SOURCE: BUREAU OF LABOR STATISTICS

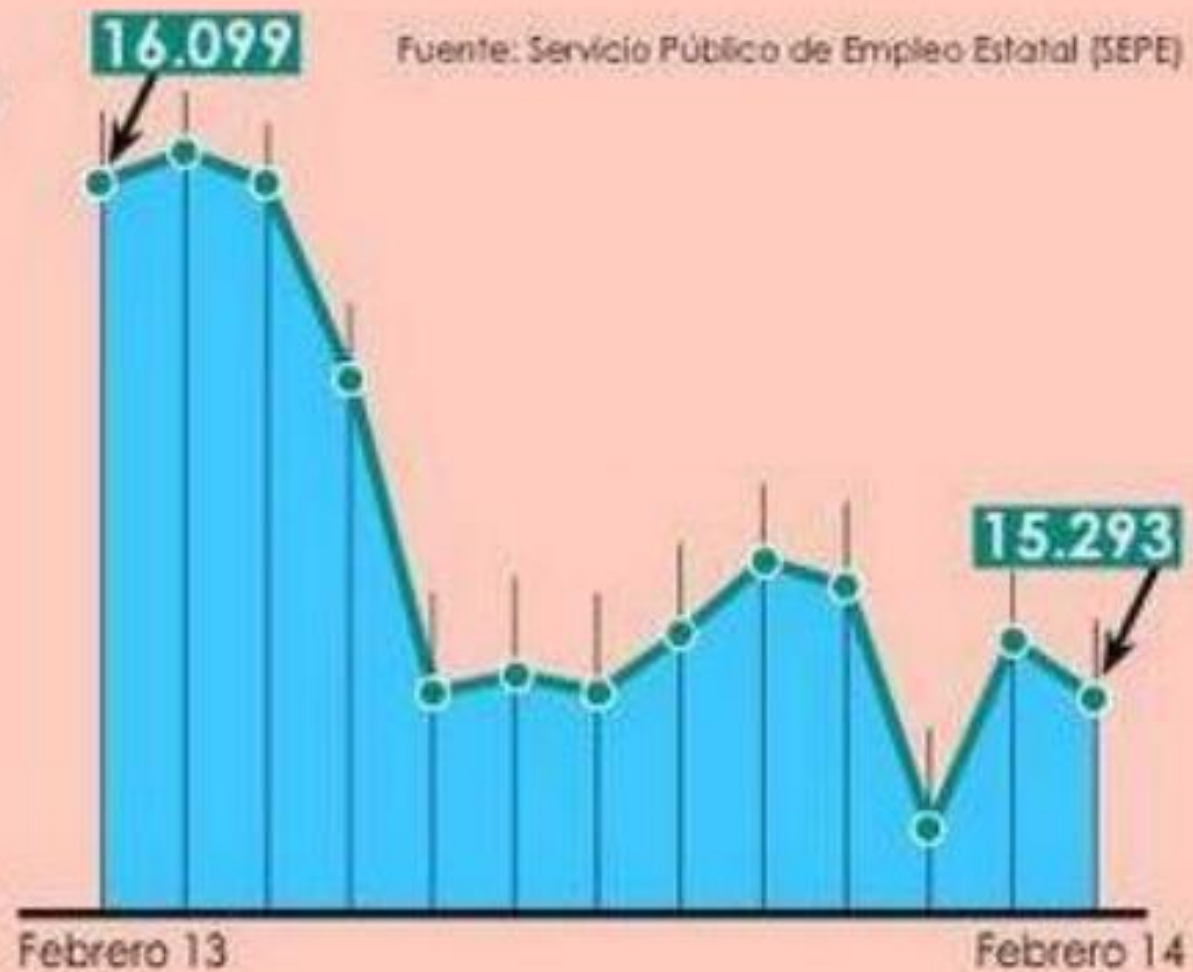
EAR WAR AND A TROOP WITHDRAWAL AT THE E/ NAS FUT 2,292.50

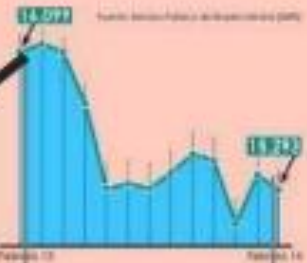
Total War: Warhammer @ 2160p (fps)





Fuente: Servicio Público de Empleo Estatal (SEPE)





AVERAGE ANNUAL ECONOMIC GROWTH

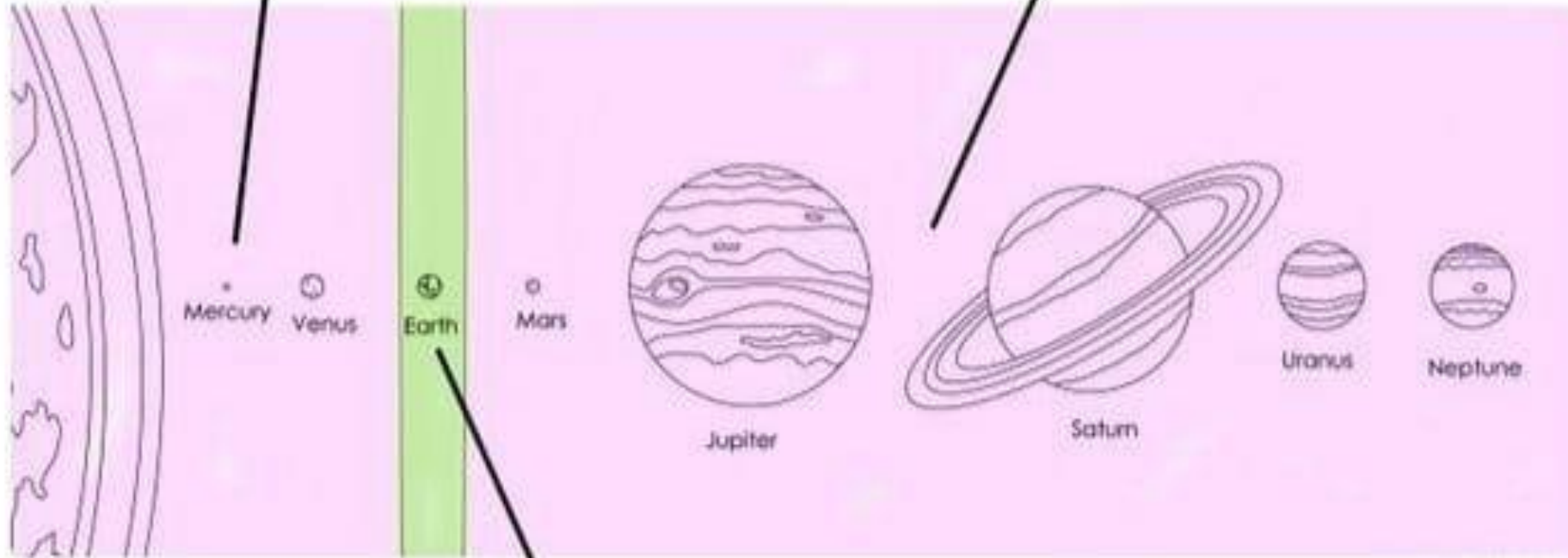


***UP TO JUNE 30, 2016**
SOURCE: BEA

CHART TO HELP DETERMINE RISK OF BEAR ATTACK:

NO RISK OF BEAR ATTACK

ALSO NO RISK OF BEAR ATTACK

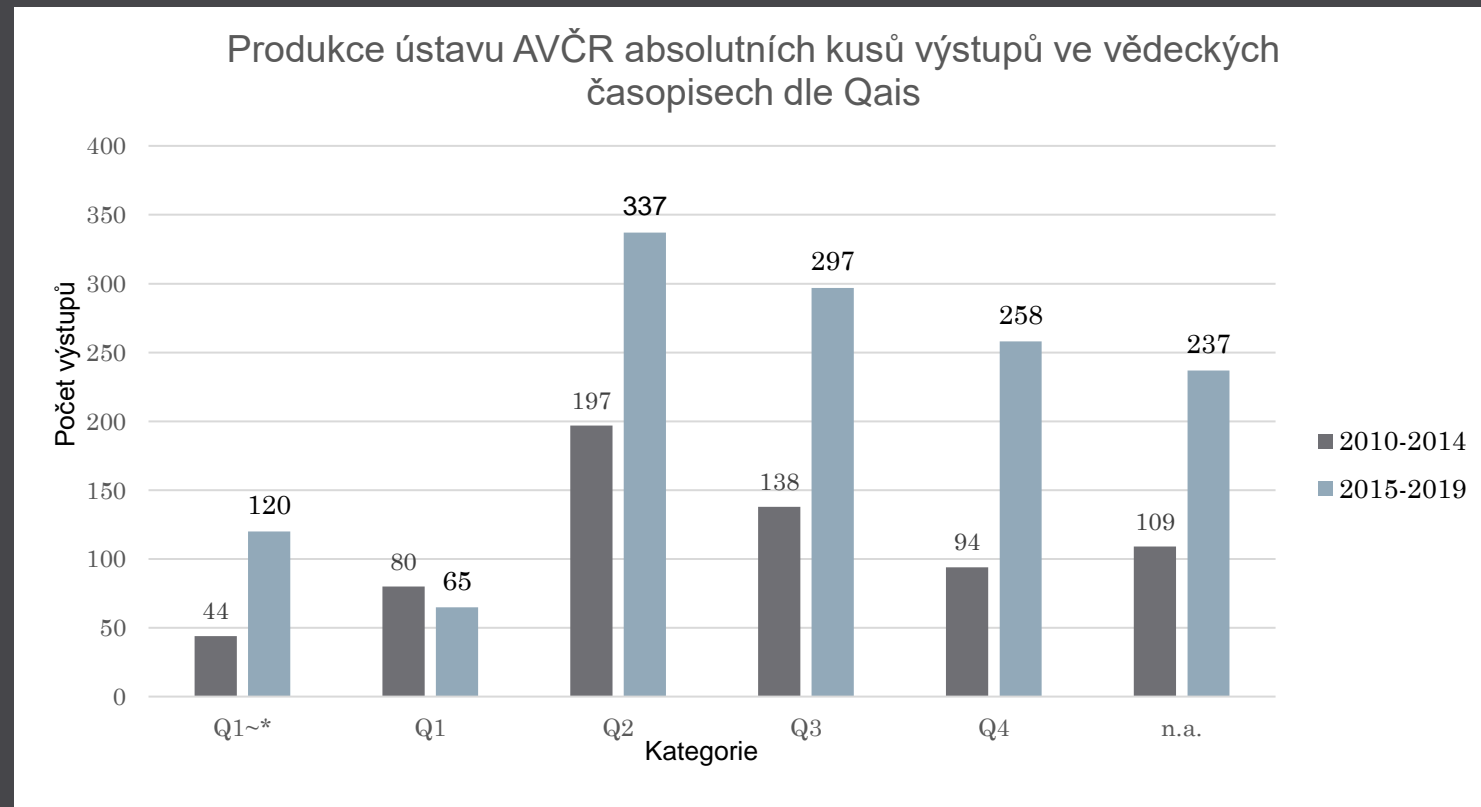


REALLY VERY HIGH RISK OF BEAR ATTACK



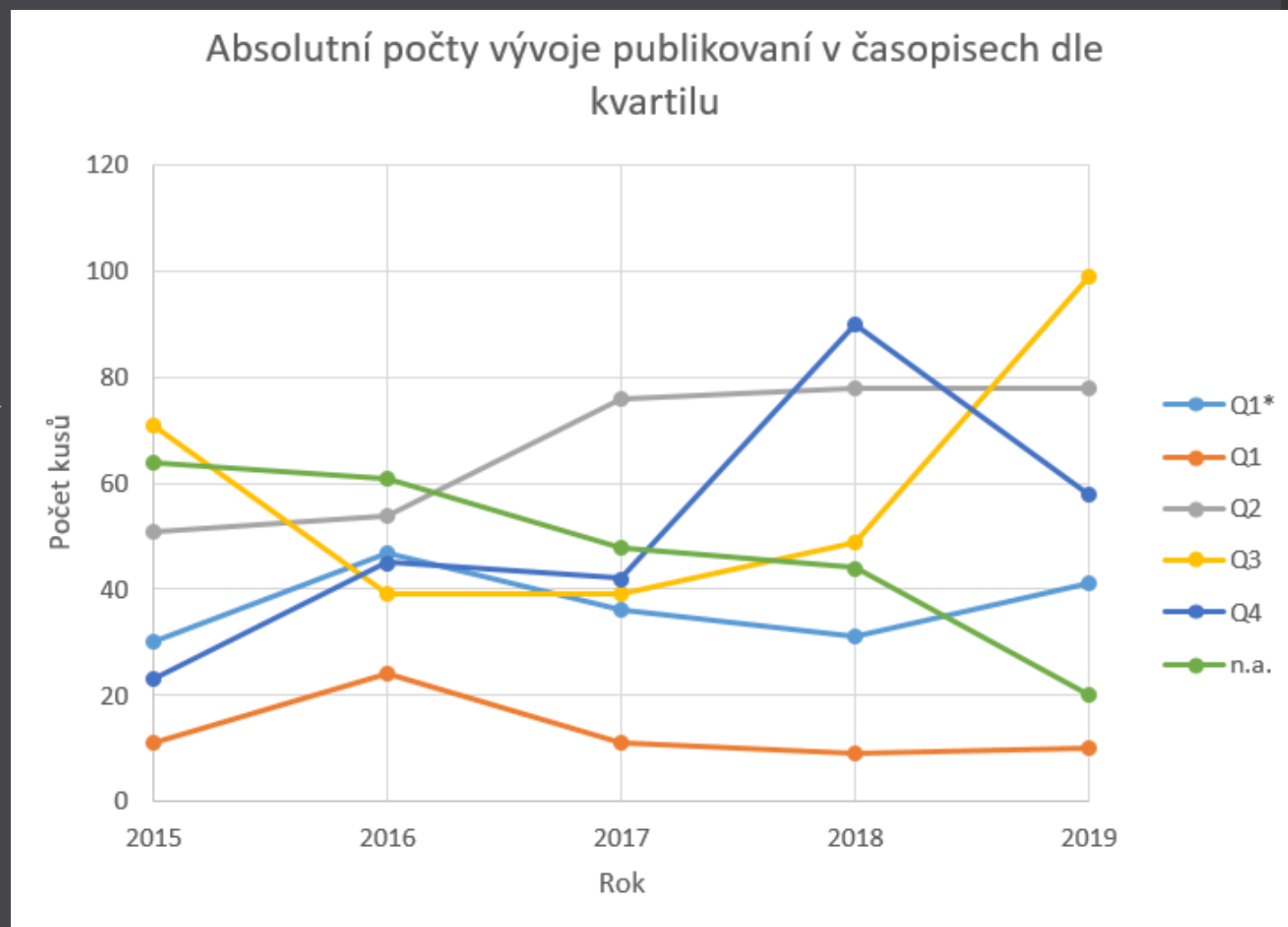
Náležitosti grafu

- Název – stručný, ale výstižný. Pokud by obsahoval příliš mnoho textu, tak je možné jej umístit do popisku.
- Popis os.
- Popis hodnot.
- Legenda.
- Vhodné škálování.



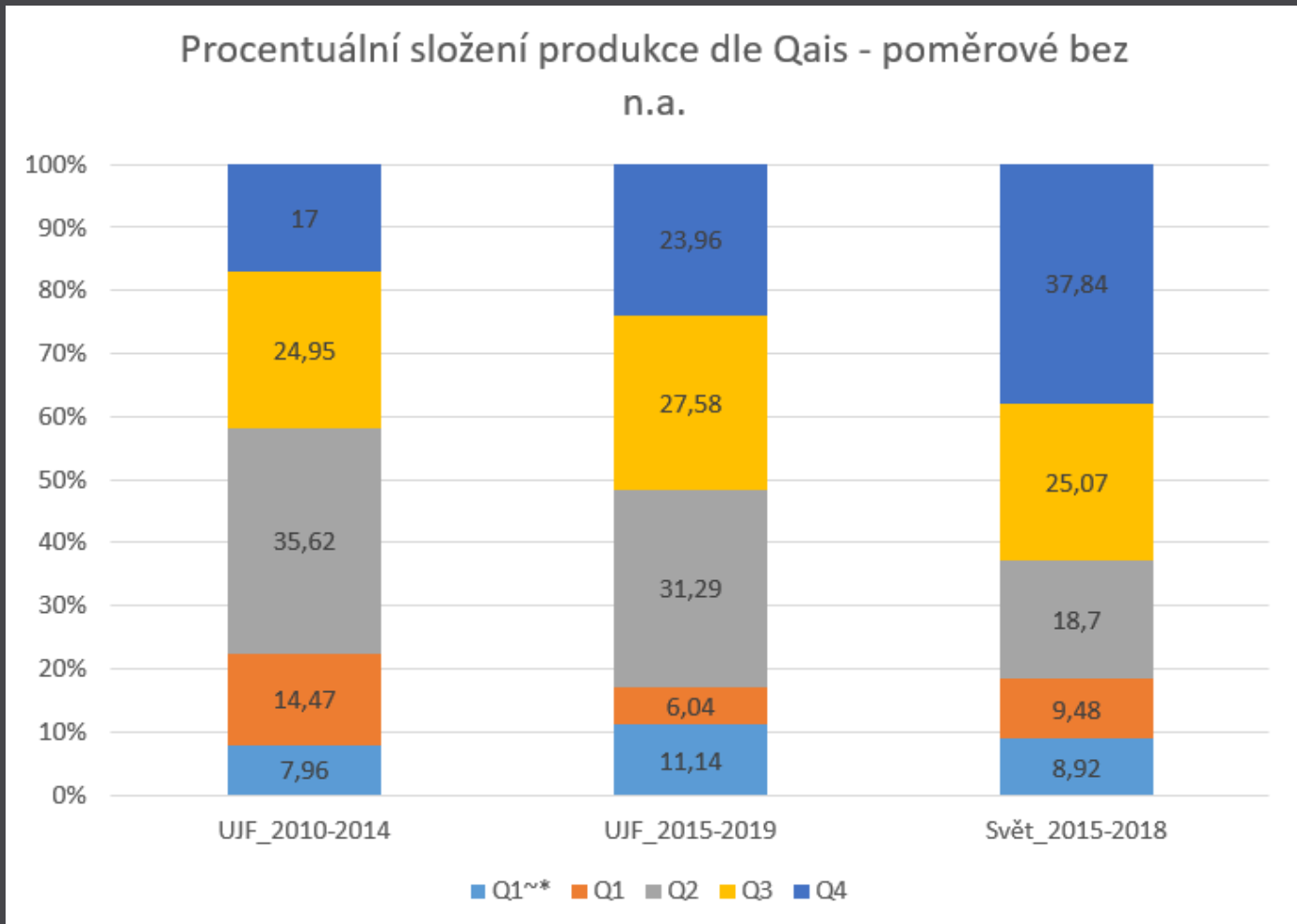
Spojnicový/bodový graf

- Vhodný na prezentaci trendů a rozvoju.
- Je zapotřebí volit zda je vhodné body pospojovat spojnicemi a pokud ano, tak rovnými či vyhlazenými.
- Bodové grafy bez spojnic mohou mít různé velikosti bodů (bublinový graf).



Sloupcový graf

- Sloupcový graf je vhodný například pro zobrazování četností či poměrů srovnatelných sad.
- Někdy nazýván jako histogram.





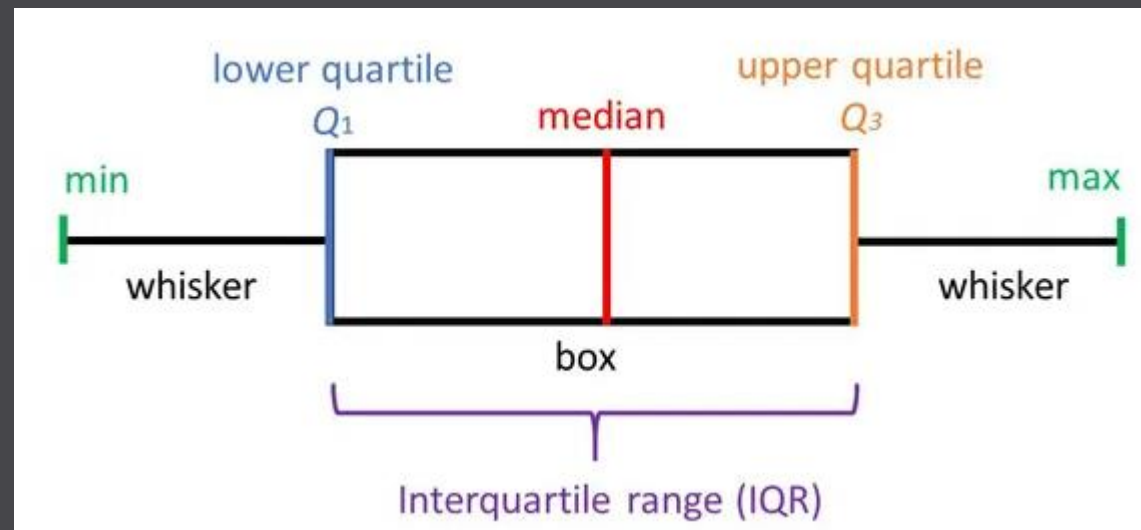
UNDEREXPOSURE

CORRECT EXPOSURE

OVEREXPOSURE

Krabicový graf

- Pokud je k dispozici, tak poskytuje mnoho údajů na relativně malém prostoru.
- Může mít zaznačený i průměr.
- Přehledně porovnává více datových sad.

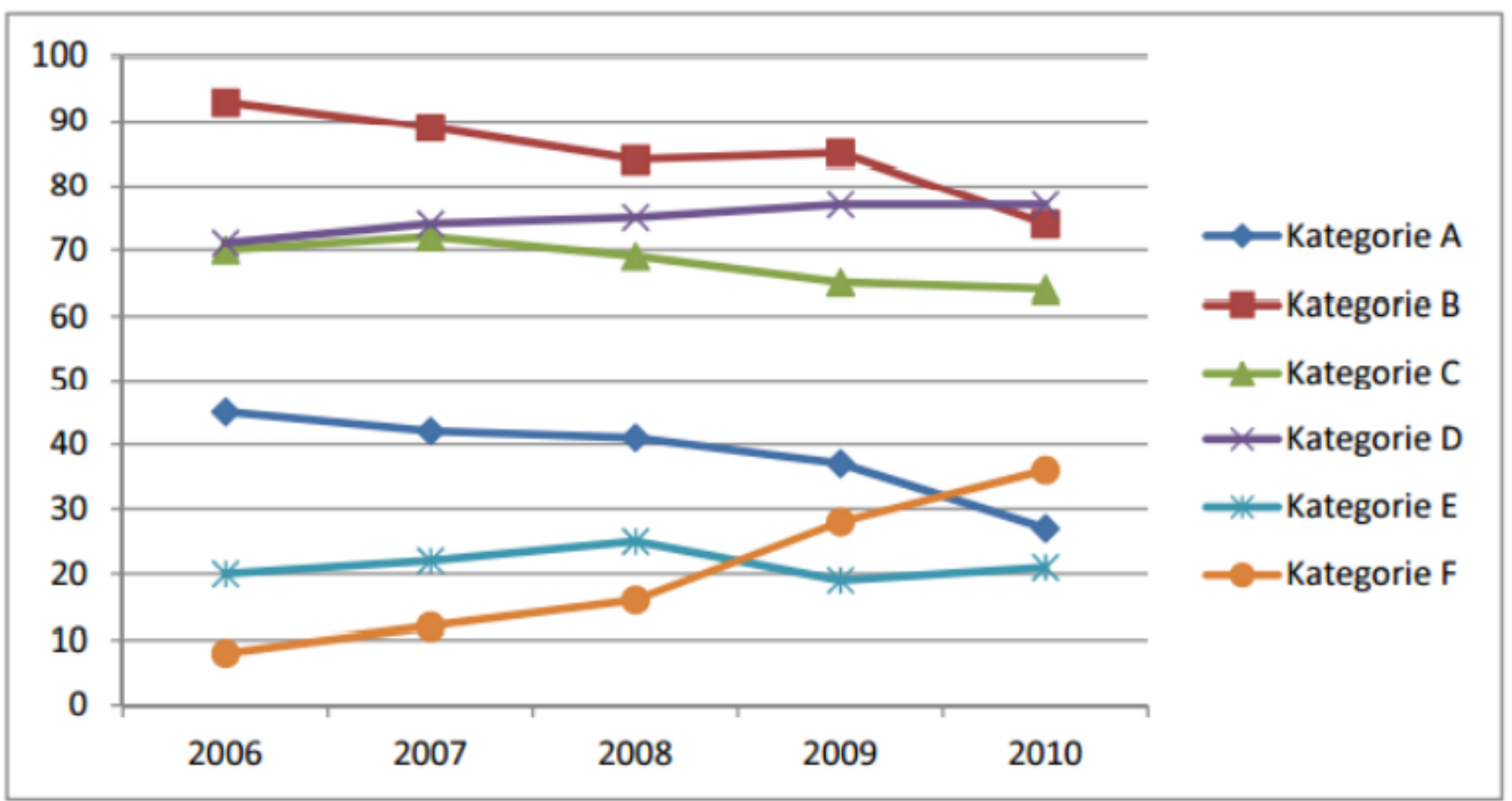


Prvky grafické prezentace

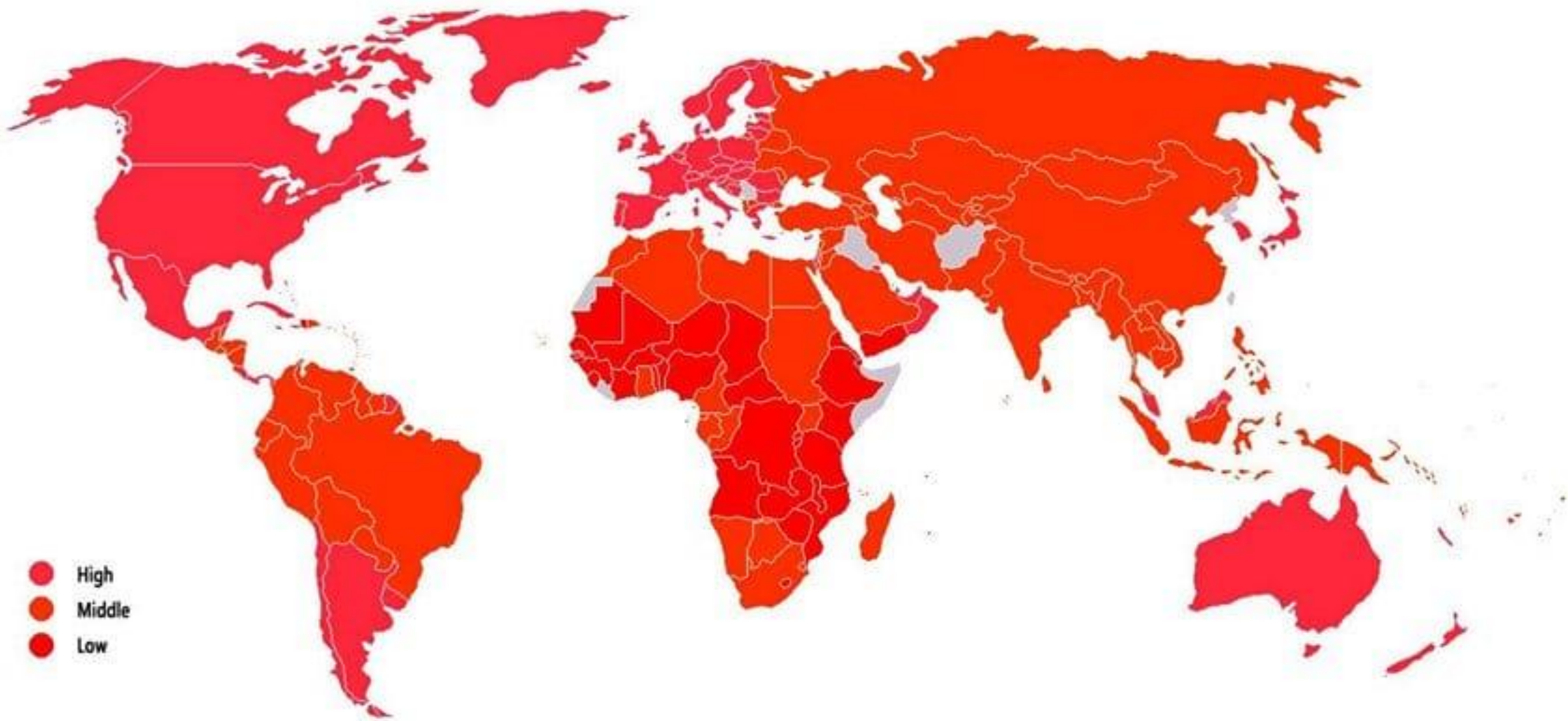
- Co, pro koho a jak dlouho určuje jak. – „If you know the enemy and know yourself, you need not fear the result of a hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat. If you know neither the enemy nor yourself, you will succumb in every battle.“ – Sun Tzu, Art of War.
- U grafů dejte publiku minutu na prozkoumání, pokud mluvíte, tak i dvě.
- Hutnost informací – grafy jsou kompresí balíků dat, je nezbytné vhodně zvolit úroveň agregace.
- Kontext – umožněte pochopení grafů i mimo prezentace, doplňte je komentáři. Graf je smysluplný i bez Vás.
- Rozsah – Je skutečně nezbytně nutné, aby byly prezentovány všechny grafy? Můžeme nějaké smazat?
- Informování přebíjí krásu.
- Pokud vstupují špatná data, tak už se s tím nedá nic moc dělat.

Prvky grafické prezentace

- Jednoduché formátování – Žádný *Comic Sans*, *zbytečné barvy* a smajlíky ☹ Příliš pestrá grafika odpoutává pozornost, příliš fádni může nudit.
- 15463279 => 15,463,279 nebo 15 463 279
- Lower the number of numbers, because a lot of numbers is only for a small number of number lowers.
- Opatrně s opakováním informací. Neměly by být duplicitní. Redundance je nežádoucí. Stejně tak nadbytečnost.
- Žádné 3D a animace.
- Pokud to není nutné, tak nemíchat typy grafů.
- Mind the zero, be data hero.
- Příliš komplikovaný graf lze rozložit do několika dílčích.
- Jsou mezi námi! Barvoslepí lidé... a tiskárny.

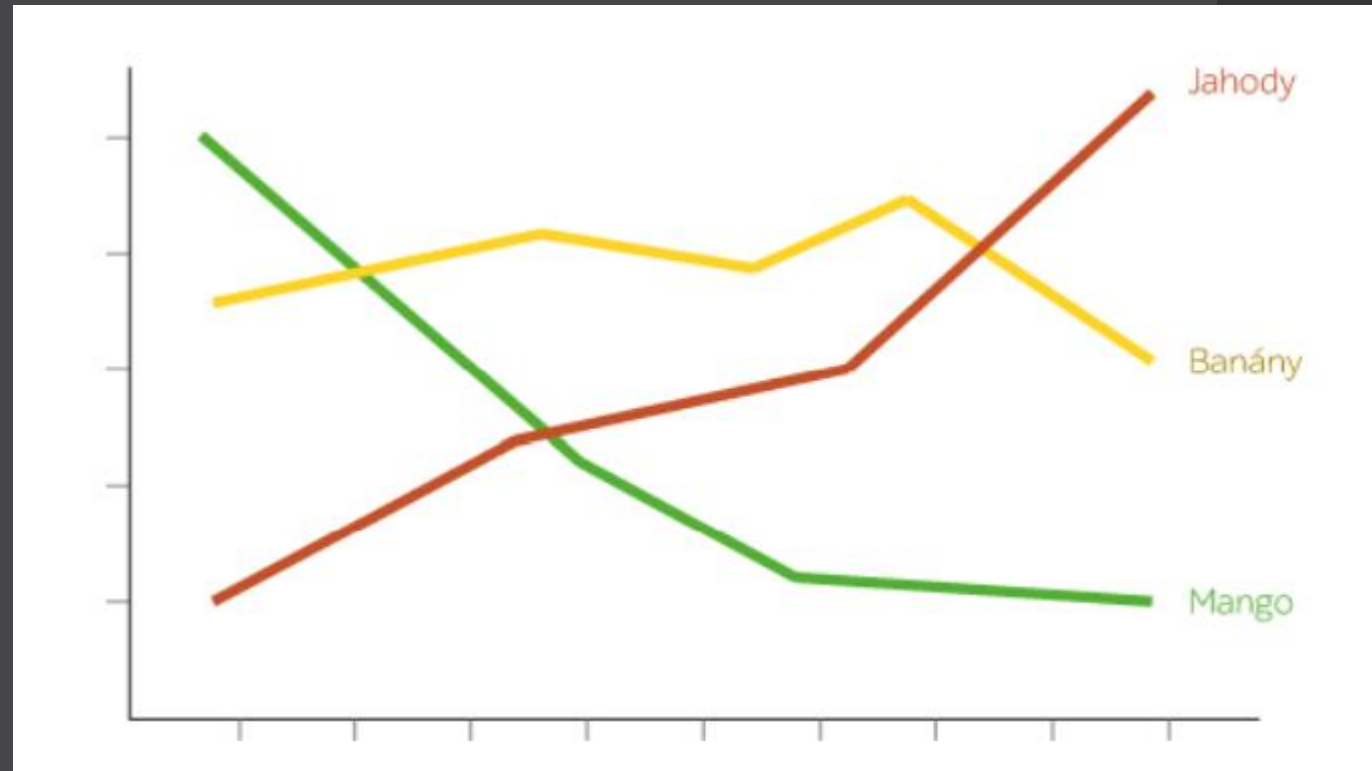


World Incidence of Color Blindness



„manipulační techniky“

- Barva
- Měřítko
- Doba zobrazení
- Vynechání informace
- Fabrikace



Vizualizace

- Předání zprávy, zapůsobení na recipienta: od dat ke grafům.
- Estetika: Dějiny krásy, Dějiny ošklivosti. Kritik umění.
- Fotografie: Zlatý řez, histogram, kontrast.
- Webové stránky: UX, barvy, intuitivní rozložení.
- Architektura/urbanistika: Vizuální smog

Opakování

Literatura

- MAREK, Tomáš. Efektivní vizualizace dat <https://muni.cz/go/5a467f>
- E-learningové kurzy MUNI:
- <https://kisk.phil.muni.cz/kisk4future/datova-gramotnost/osnova-kurzu-datova-gramotnost>
- <https://kisk.phil.muni.cz/kisk4future/vizualizace/osnova-kurzu-grafy-a-vizualizace-dat>
- HINDLS, Richard, Stanislava HRONOVÁ a Jan SEGER. *Statistika pro ekonomy*. 5. vyd. Praha: Professional Publishing, c2004. ISBN 80-864-1959-2.
- LAŠÁK, Pavel. Jak na Excel [online]. 2021 [cit. 2021-10-17]. Dostupné z: <https://office.lasakovi.com/excel/>

Děkuji za pozornost