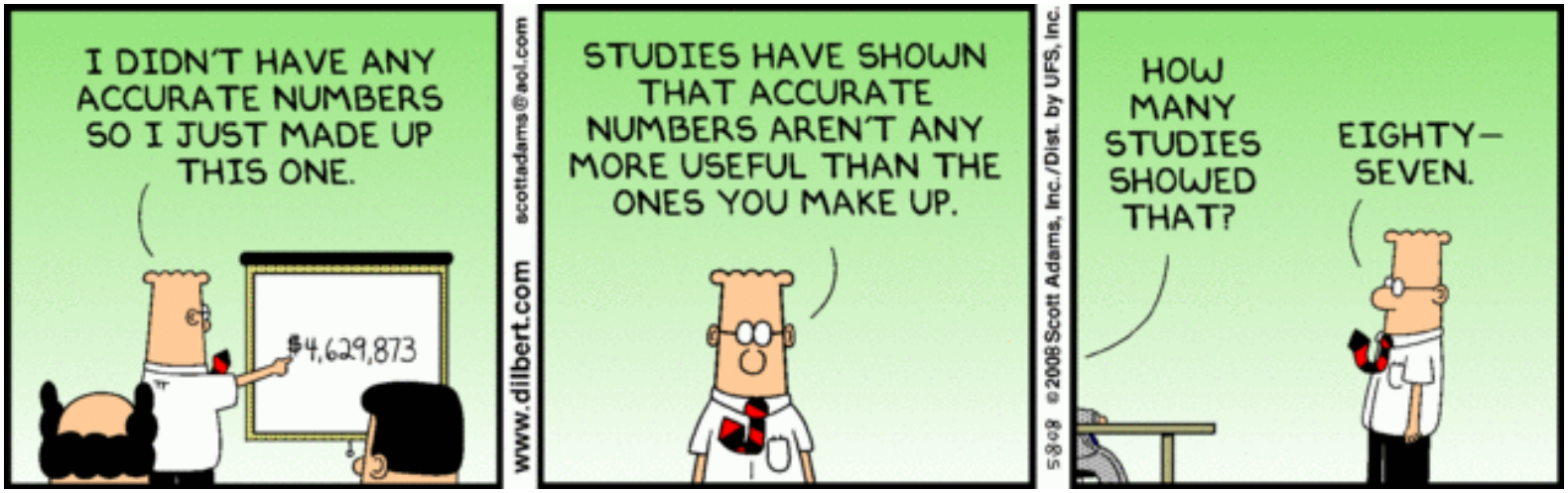
ANALÝZA DAT

Základy analýzy kvantitativních dat

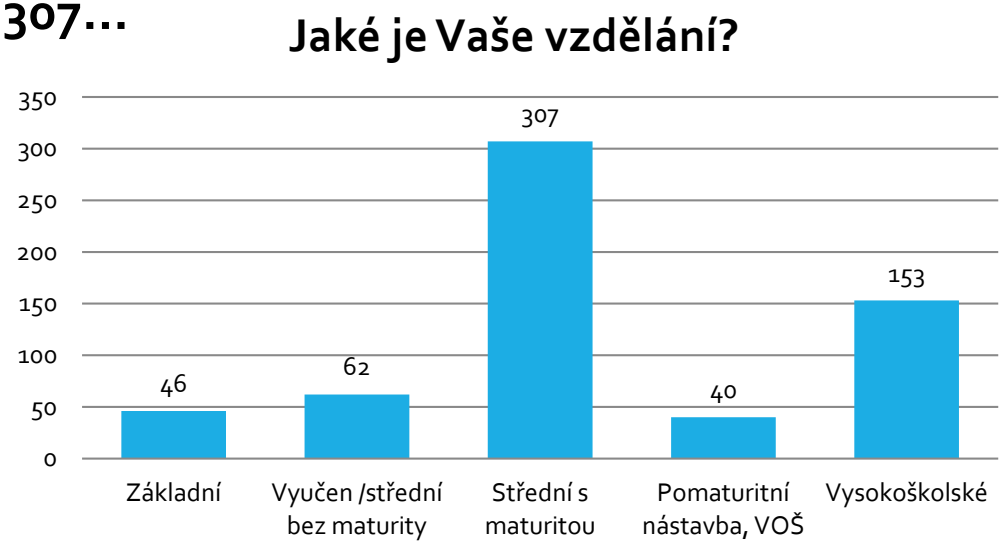
Metodologie pro ISK – podzim 2015


I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
www.dilbert.com scottadams@aol.com
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
5-8-08 ©2008 Scott Adams, Inc./Dist. by UFS, Inc.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.

Začínáme s analýzou (v Excelu)

Nejprve pár termínů:

- **Proměnná (znak) - vzdělání**
- **Hodnota proměnné – ZŠ, SŠ, VŠ …**
- **Četnost – 46, 62, 307…**



Proměnné

- Nominální
  - nabývají nečíselných hodnot a nelze je uspořádat hierarchicky či podle velikosti (nemůžeme určit, která hodnota proměnné je vyšší než jiná. Speciálním případem nominálních hodnot jsou **dichotomické proměnné** (muž/žena, ano/ne). Nominální proměnnou může být např. stav, bydliště, oblíbená barva apod.
- Ordinální
  - nabývají hodnot, u kterých můžeme s jistotou tvrdit, že jedna je vyšší než druhá, nemůžeme však s jistotou tvrdit, o kolik je vyšší. Ordinální proměnnou je například vzdělání, volně formulované frekvence činností.
- Kardinální
  - nabývají skutečných měřitelných číselných hodnot – kardinální proměnnou je například věk, počet dětí, výše platu. Speciálním případem kardinálních proměnných jsou intervalové proměnné (např. výše platu měřená intervaly 0-10000, 10001-20000, 20001-30000…)

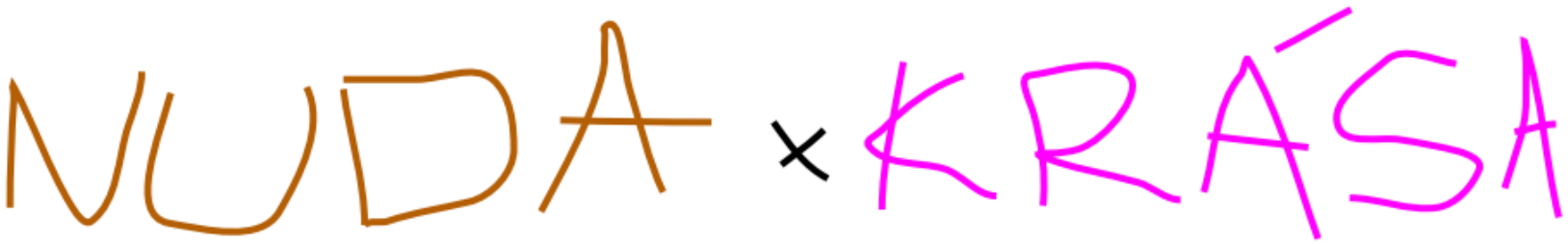
Statistická analýza

- Deskriptivní
  - zabývá se sběrem, sumarizací a prezentací souborů dat. Je to ta „lehčí" statistika, která je dostupná pomocí běžných nástrojů
    - *Jaká je průměrná délka života žen?*
    - *Jaká je mediánová hodnota platu knihovníků v ČR?*
    - *Jaký je minimální a maximální počet knih, který průměrně za rok přečte student KISKu?*
- Induktivní
  - Zabývá se zobecňováním výsledků výzkumu na vzorku na populaci

Zdroje dat

Připravená data
Standardní formáty
Málo práce
Weby institucí

Data neexistují
Existují, ale jsou tajná
Spousta práce
(I programování!)



Obrázek CC: Jan Boček

Toužíte-li po kráse → ***Vizualizace dat*** *(Boček, Marek, Málek, Pospíšil),* ***Datová analytika*** *(Mayer) + celá datová větev*

Zdroje dat

- **Český statistický úřad**
  - *otevřená data z výsledků voleb*
- **Databáze Eurostatu**
  - *http://ec.europa.eu/eurostat/data/database*
- **ČSDA - Český sociálněvědní datový archiv**
  - *ČSDA poskytuje přístupk vybraným českým datovým souborům reprezentativních výzkumů. Bez registrace je možné procházet stránky Webu a informace o archivovaných datech.V archivu najdete například datové soubory z realizovaných měsíčních šetření Centra pro výzkum veřejného mínění (CVVM).*
- **Instituce a jejich repozitáře**
  - *Datacite*
  - ***www.otevrenadata.cz***

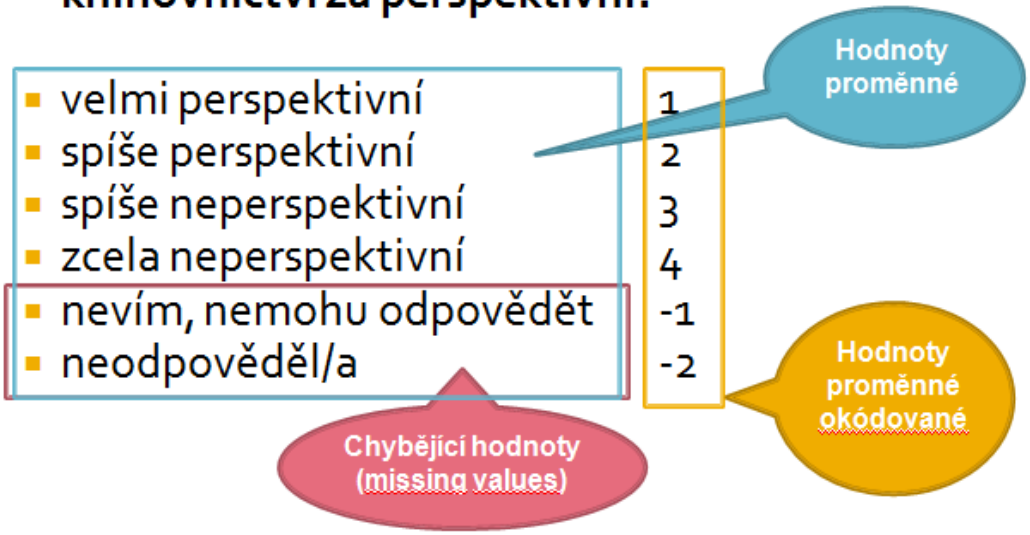
Datové matice

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
	Ot5	Ot6	Ot71	Ot72	Ot73	Ot74	Ot75	Ot76	Ot77	Ot78	Ot8	Ot91	Ot92	Ot93	Ot94	Ot95	Ot96	Ot97	Ot98
ní	ano		0	0	0	0	0	0	0	,	ano, absol	spíše nesc	spíše nesc	vůbec nes	naprosto s	spíše nesc	spíše nesc	spíše souh	vůbec nes
ní	ano	jednou tý	0	0	0	0	0	0	0	,	ne, čerpár	vůbec nes	vůbec nes	vůbec nes	naprosto s	vůbec nes	vůbec nes	naprosto s	vůbec nes
ní	ano	méně čast	5.	1.	3.	4.	6.	7.	2.	8,	ano, absol	spíše nesc	spíše nesc	vůbec nes	spíše souh	nevím, ne	spíše nesc	spíše nesc	spíše nesc
ní	ano	méně čast	1.	7.	2.	4.	3.	5.	6.	,	ano, absol	vůbec nes	spíše souh	spíše nesc	spíše souh	spíše nesc	nevím, ne	naprosto s	spíše nesc
ní	ne	několikrát	1.	3.	4.	8.	5.	7.	6.	2, prezenc	ne, nikdo	spíše nesc	spíše nesc	spíše nesc	nevím, ne	spíše nesc	spíše nesc	spíše souh	spíše nesc
ní	ano	jednou tý	0	0	0	0	0	0	0	,	ne, nikdo	nevím, ne	nevím, ne	nevím, ne	nevím, ne	nevím, ne	nevím, ne	nevím, ne	nevím, ne
ní	ne	jednou m	4.	2.	3.	5.	1.	6.	7.	,	ne, nikdo	naprosto s	spíše souh	nevím, ne	spíše souh	nevím, ne	spíše souh	spíše nesc	nevím, ne
ní	ne	méně čast	1.	7.	6.	5.	3.	2.	4.	8, jiný	ne, čerpár	vůbec nes	nevím, ne	spíše nesc	naprosto s	vůbec nes	vůbec nes	nevím, ne	vůbec nes
ní	ano	jednou tý	0	0	0	0	0	0	0	,	ne, čerpár	vůbec nes	vůbec nes	vůbec nes	naprosto s	vůbec nes	vůbec nes	naprosto s	vůbec nes
ní	ano	méně čast	2.	6.	1.	3.	4.	5.	7.	,	ne, nikdo	spíše souh	spíše nesc	vůbec nes	vůbec nes	spíše souh	nevím, ne	spíše souh	nevím, ne
ní	ano	jednou za	0	0	0	0	0	0	0	,	ano, absol	spíše nesc	spíše nesc	spíše nesc	spíše souh	spíše nesc	spíše nesc	spíše souh	spíše nesc
ní	ano	jednou za	0	0	0	0	0	0	0	,	ne, čerpár	spíše souh	vůbec nes	nevím, ne	naprosto s	nevím, ne	spíše nesc	spíše souh	spíše nesc
ní	ne	jednou za	1.	2.	3.	5.	4.	6.	7.	,	ne, čerpár	spíše nesc	nevím, ne	vůbec nes	naprosto s	spíše nesc	vůbec nes	naprosto s	vůbec nes
ní	ne	méně čast	2.	4.	1.	7.	5.	3.	6.	8,	ne, nikdo	spíše nesc	nevím, ne	spíše nesc	spíše souh	spíše nesc	nevím, ne	spíše souh	spíše nesc
ov	ano	méně čast	0	0	0	0	0	0	0	,	vím jak fu	vůbec nes	vůbec nes	spíše nesc	spíše souh	vůbec nes	vůbec nes	naprosto s	vůbec nes
ní	ano	jednou tý	0	0	0	0	0	0	0	,	ano, mám	spíše nesc	nevím, ne	spíše nesc	nevím, ne	spíše souh	vůbec nes	naprosto s	spíše nesc
ní	ano	jednou m	0	0	0	0	0	0	0	,	ne, čerpár	spíše nesc	spíše souh	vůbec nes	nevím, ne	spíše nesc	spíše souh	spíše souh	spíše nesc
ní	ne	jednou m	1.	3.	2.	4.	5.	7.	6.	,	ano, absol	vůbec nes	vůbec nes	vůbec nes	naprosto s	vůbec nes	vůbec nes	naprosto s	vůbec nes
ní	ne	několikrát	1.	5.	2.	7.	4.	3.	6.	,	ne, čerpár	vůbec nes	spíše nesc	spíše nesc	naprosto s	spíše nesc	spíše nesc	spíše souh	spíše nesc
ní	ano	jednou tý	5.	7.	1.	4.	2.	6.	3.	,	ne, nikdo	nevím, ne	nevím, ne	vůbec nes	nevím, ne	spíše souh	nevím, ne	nevím, ne	spíše souh
ní	ano	méně čast	1.	5.	2.	4.	3.	6.	7.	,	ne, čerpár	nevím, ne	spíše nesc	vůbec nes	naprosto s	spíše nesc	spíše nesc	naprosto s	vůbec nes
ní	ano	jednou m	3.	6.	7.	4.	1.	5.	2.	,	ne, čerpár	vůbec nes	spíše nesc	vůbec nes	naprosto s	vůbec nes	vůbec nes	naprosto s	vůbec nes
ní	ne	jednou m	1.	2.	3.	5.	6.	7.	4.	8,	ne, čerpár	vůbec nes	spíše nesc	spíše nesc	naprosto s	vůbec nes	vůbec nes	naprosto s	vůbec nes
ní	ne	jednou tý	3.	4.	1.	7.	2.	5.	6.	8,	ne, čerpár	vůbec nes	spíše nesc	spíše souh	naprosto s	vůbec nes	vůbec nes	naprosto s	vůbec nes

Datové matice

D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	
)t4	Ot5	Ot6	Ot71	Ot72	Ot73	Ot74	Ot75	Ot76	Ot77	Ot78	Ot8	Ot91	Ot92	Ot93	Ot94	Ot95	Ot96	Ot97	Ot98	Ot99	Ot91
.	1		0	0	0	0	0	0	0	,	2	4	4	5	1	4	4	2	5	5	
.	1	2	0	0	0	0	0	0	0	,	3	5	5	5	1	5	5	1	5	5	
.	1	5	5	1	3	4	6	7	2	8,	2	4	4	5	2	3	4	4	4	4	
.	1	5	1	7	2	4	3	5	6	,	2	5	2	4	2	4	3	1	4	4	
.	2	1	1	3	4	8	5	7	6	2, prezenc	4	4	4	4	3	4	4	2	4	4	
.	1	2	0	0	0	0	0	0	0	,	4	3	3	3	3	3	3	3	3	3	
.	2	4	4	2	3	5	1	6	7	,	4	1	2	3	2	3	2	4	3	3	
.	2	5	1	7	6	5	3	2	4	8, jiný	3	5	3	4	1	5	5	3	5	5	
.	1	2	0	0	0	0	0	0	0	,	3	5	5	5	1	5	5	1	5	5	
.	1	5	2	6	1	3	4	5	7	,	4	2	4	5	5	2	3	2	3	3	
.	1	3	0	0	0	0	0	0	0	,	2	4	4	4	2	4	4	2	4	4	
.	1	3	0	0	0	0	0	0	0	,	3	2	5	3	1	3	4	2	4	4	
.	2	3	1	2	3	5	4	6	7	,	3	4	3	5	1	4	5	1	5	5	
.	2	5	2	4	1	7	5	3	6	8,	4	4	3	4	2	4	3	2	4	4	
!	1	5	0	0	0	0	0	0	0	,	vím jak fu	5	5	4	2	5	5	1	5	5	
.	1	2	0	0	0	0	0	0	0	,	1	4	3	4	3	2	5	1	4	5	
.	1	4	0	0	0	0	0	0	0	,	3	4	2	5	3	4	2	2	4	4	
.	2	4	1	3	2	4	5	7	6	,	2	5	5	5	1	5	5	1	5	4	
.	2	1	1	5	2	7	4	3	6	,	3	5	4	4	1	4	4	2	4	5	
.	1	2	5	7	1	4	2	6	3	,	4	3	3	5	3	2	3	3	2	4	
.	1	5	1	5	2	4	3	6	7	,	3	3	4	5	1	4	4	1	5	5	
.	1	4	3	6	7	4	1	5	2	,	3	5	4	5	1	5	5	1	5	5	
.	2	4	1	2	3	5	6	7	4	8,	3	5	4	4	1	5	5	1	5	5	
	2	2	3	4	1	7	2	5	6	8,	3	5	4	2	1	5	5	1	5	5	

- **2. Považujete obor Informační studia a knihovnictví za perspektivní?**



Spokojenost s nabídkou kurzů						
	Velmi souhlasím	Spíše souhlasím	Ani souhlasím, ani nesouhlasím	Spíše nesouhlasím	Vůbec nesouhlasím	Nevím / nemohu odpovědět
Povinné (A) kurzy mají logickou časovou posloupnost.	1	2	3	4	5	-1
Obsahy jednotlivých povinných (A) kurzů se nepřekrývají.	1	2	3	4	5	-1
Jsem spokojen/a s tematickou šíří nabídky povinně volitelných (B) kurzů.	1	2	3	4	5	-1
Jsem spokojen/a s počtem nabízených povinně volitelných (B) kurzů.	1	2	3	4	5	-1

Validní a chybějící hodnoty

- **Validní hodnoty** jsou ty hodnoty, které započítáváme do analýzy. Jsou to všechny varianty odpovědí, které pro nás mají vysokou informační hodnotu.
- **Chybějící hodnoty** jsou ty hodnoty, kdy respondent zvolí odpověď typu „nevím / nemohu se rozhodnout / nemohu odpovědět" nebo otázku přeskočí a odpověď vůbec neposkytne. I tyto druhy odpovědí pro nás mohou mít informační hodnotu (např. pokud existuje na některou otázku vysoký počet odpovědí „nevím" nebo neodpovědí, měli bychom se zamyslet nad tím, zda respondenti otázce rozumí).
- **Nevalidní hodnoty** – chybné hodnoty (outliers, chyby)

Zásady pro práci s daty

1. **Zálohovat!**
2. **Zálohovat!**
3. **Zálohovat!**

4. **Kontrolovat!**
5. **Popisovat!**
6. **GIGO!**

GIGO!

- Gabrage in → garbage out ☹
- Slučování a rozdělování sloupců
- Hledání a nahrazování textu (CTRL+H)
- Odebrání duplicitních řádků (DATA – Odebrat duplicity)
- Příkazy ZLEVA, ZPRAVA, DOSADIT a další
- Malá/velká písmena: MALÁ, VELKÁ, VELKÁ2
- Odebrání mezer a netisknutelných znaků z textu ( ): PROČISTIT, VYČISTIT
- Úpravy formátování čísel (datum, čas, procenta)
- Transpozice
- Více info – nápověda Excelu
- Tip: projděte si základní operátory v Excelu

První pohled na data

- Různé datové formáty:
  - **XLS, XLSX**
  - CSV, TSV, TXT
  - XML
  - **SAV**
  - JSON, GEOJSON

- https://www.czso.cz/csu/czso/otevrena_data_pro_vysledky_scitani_lidu_domu_a_bytu_2011_-sldb_2011-
  - CSV
  - popisy dat
  - čištění
  - tabulka
  - první pohled na data
  - Kraje – okresy – obce
  - Podmíněné formátování

Cvičení

- Otevřete si seznam studentů Metodologie
- Rozdělte jméno a příjmení do samostatných sloupců
- Vytvořte nové sloupce ze sloupečku Studium:
  - Studijní obor
  - Forma studia
  - Semestr

Cvičení

- Otevřete si seznam studentů Metodologie
- Rozdělte jméno a příjmení do samostatných sloupců
- Vytvořte nové sloupce ze sloupečku Studium:
  - Studijní obor
  - Forma studia
  - Semestr

MÁTE?

- Jaké je nejnižší a nejvyšší UČO?
- Kolik je v souboru kombinovaných a prezenčních studentů?
- V jakém semestru se nacházejí studenti?
  - Aritmetický průměr
  - Medián
  - Modus

Základní statistické operace

- COUNTIF(oblast;"hodnota")

A	B
Prodejce	**Faktura**
Novák	15 000
Novák	9 000
Horák	8 000
Horák	20 000
Novák	5 000
Veselý	22 500
Vzorec	**Popis (výsledek)**
=COUNTIF(A2:A7;"Novák")	Počet faktur od Nováka (3)
=COUNTIF(A2:A7;A4)	Počet faktur od Horáka (2)
=COUNTIF(B2:B7,"< 20000")	Počet faktur s hodnotou nižší než 20 000 (4)
=COUNTIF(B2:B7,">="&B5)	Počet faktur s hodnotou vyšší nebo rovnou 20 000 (2)

Základní statistické operace

- MEDIAN(oblast)
- MODE(oblast)
- PRŮMĚR(oblast)

- MIN(oblast)
- MAX(oblast)