Tidy data

Organizing data

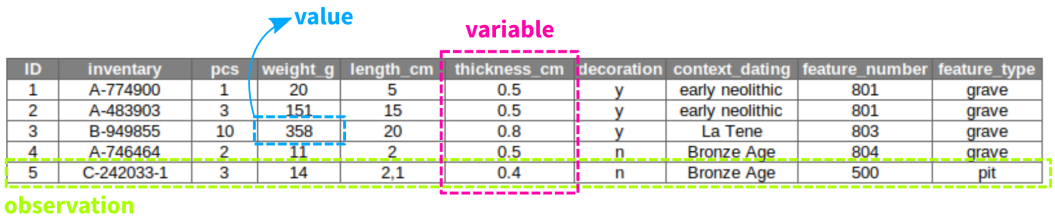
What is a database?

(Brainstorming)

Relational database - basic terms

ID	inventary	pcs	weight_g	length_cm	thickness_cm	decoration	context_dating	feature_number	feature_type
1	A-774900	1	20	5	0.5	y	early neolithic	801	grave
2	A-483903	3	151	15	0.5	y	early neolithic	801	grave
3	B-949855	10	358	20	0.8	y	La Tene	803	grave
4	A-746464	2	11	2	0.5	n	Bronze Age	804	grave
5	C-242033-1	3	14	2,1	0.4	n	Bronze Age	500	pit

Relational database - basic terms

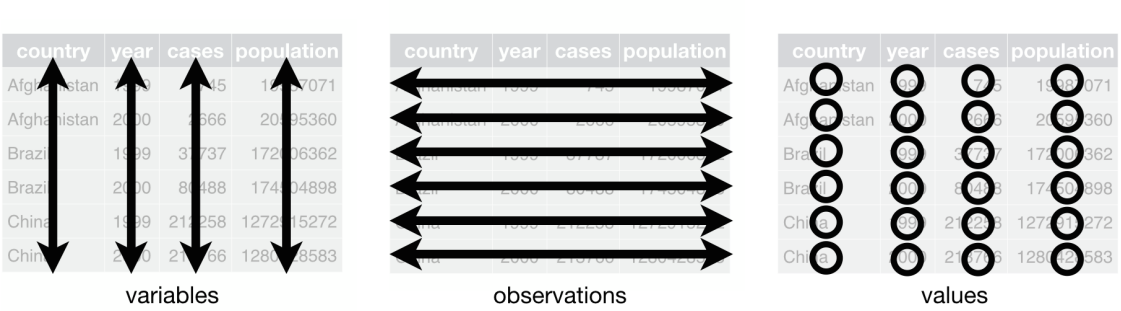


ID	inventary	pcs	weight_g	length_cm	thickness_cm	decoration	context_dating	feature_number	feature_type
1	A-774900	1	20	5	0.5	y	early neolithic	801	grave
2	A-483903	3	151	15	0.5	y	early neolithic	801	grave
3	B-949855	10	358	20	0.8	y	La Tene	803	grave
4	A-746464	2	11	2	0.5	n	Bronze Age	804	grave
5	C-242033-1	3	14	2,1	0.4	n	Bronze Age	500	pit

- **variable** (*proměnná* / *atribut*) - napr. hmotnosť keramického fragmentu
- **observation** / **object** (*entita* / *záznam*) - napr. konkrétny keramický fragment
- **value** (*hodnota*)

- **primary key / unique ID** (*primární klíč*)

Tidy data



Types of variables

Categorical

- **dichotomies** (*dichotomická*) - prítomnosť alebo neprítomnosť nejakého javu ("Y/N")
- **nominal** (*nominální*) - archeologické datovanie ("neolit"), číslo objektu, keramický typ
- **ordinal / rank** (*ordinální, pořadová*) - tlupa / kmeň / náčelníctvo / štát, alebo: komponenta / sídelný areál / nadkomunitný areál

Numeric

- **discrete** (*diskrétní*) - môžu byť len *celé* čísla - napr. počet lokalít, počet bronzových spôn
- **continuous** (*spojité / metrické*) - môže byť akékoľvek *reálne* číslo - hmotnosť ker. fragmentu, dĺžka železného meča
- **interval** (*intervalové*) - numericky vyjadrujú vzdialenosť na vyjadrenie hierarchického vzťahu, nemajú ale zmysluplný nultý bod, umožňujú relatívne porovnávanie ale nie kalkulácie (napr. BC/AD - rozdiel medzi 400AD a 800AD je 400 rokov, 400AD ale nieje 2 krát staršie než 800AD)
- **ratio** (*poměrové*) - na rozdiel od intervalových premenných umožňujú kalkulácie (napr. vek - objekt starý 1000 rokov je 2krát starší než objekt starý 500; dĺžka meča aj.)

Tidy data

What's wrong?

ID	inventary	pcs	weight	length_cm	thickness_cm	decoration	context_dating	feature_number	feature_type	site
1	A-774900	one	20	5	0.5	y	early neolithic	801	grave	Vedrovice
ID_2	A-483903	3	151 g	Length 15, width 10	0,5	y	early neolithic	č. 801	grave	
III	B-949855	5+5	0,3 kg	20	0.8	y	La Tene	803	burial	
4ID	A-746464	2	11	2	0.5-0.8	missing	bronz	804	grave	Pohansko
5	C-242033-1	3	14	2,1	0.4	-	Bronze Age	500	pit	

sex:	male
	female

Basic tidy data principles

- One variable in one column.
- One observation in one row.
- One value in one cell.
- Do not use color codes.
- Backup your data!
- Be consistent!

Assignments

- Read Karl Broman's guide on how to organize data in spreadsheets. As an article: Broman, K. W. and Woo, K. H. 2017: Data Organization in Spreadsheets. *The American Statistician* 72(1): 2–10, DOI: https://doi.org/gdz6cm.
- Read chapter **Data** in *Quantitative analysis in archaeology* book by VanPool and Leonard (2011).