Analýza dat moderně a rozumně

Konference ČES 2023

červen 2023

Proč a jak

Proč jste tu dnes vy a co si chcete odnést??

Co vás na práci s daty štve?

Cíle aneb co si dnes odnesete

Proč - Co - *Jak*

- obrázek o stavu oboru analýzy dat
- principy použitelné s jakýmkoli problémem a nástrojem
- kde vzít veřejná data a jak si s nimi poradit
- praktické provedení některých konceptů z první části
- tipy a triky pro práci s Excelem
- tipy, kam dál

Co není cíl

× naučit se do hloubky statistiku

× naučit se R, Python a git

× pouštět se do nestandardních typů dat

Jak na to půjdeme

principy - postupy - techniky

Jak budeme dnes pracovat 🙏

Bezpečné prostředí

Všichni se učíme

Je OK nevědět

Analytická práce dnes

Důvěra v analýzy díky

1. porozumění uživateli
2. integritě postupu
3. komunikaci výsledků

Práce v kódu, ale zároveň integrace analýzy a interpretace

Datové produkty: nejen PDF reporty (web, interaktivita)

Agilní přístup k práci s daty

Analytik 🧠 a uživatel 👀

Rychle reagovat

Moct navázat na už udělané

Dobře komunikovat: v průběhu i výsledek

Sám/sama datům dobře rozumět

Získat si důvěru: proces, kontrola kvality, transparentnost

Kde data vzít

Otevřená data

ČSÚ <=> otevřená data

ČSÚ <=> data a metadata

ČSÚ <=> Eurostat (někdy lepší)

Data o životním prostředí

Geodata a číselníky

“Ostatní data”

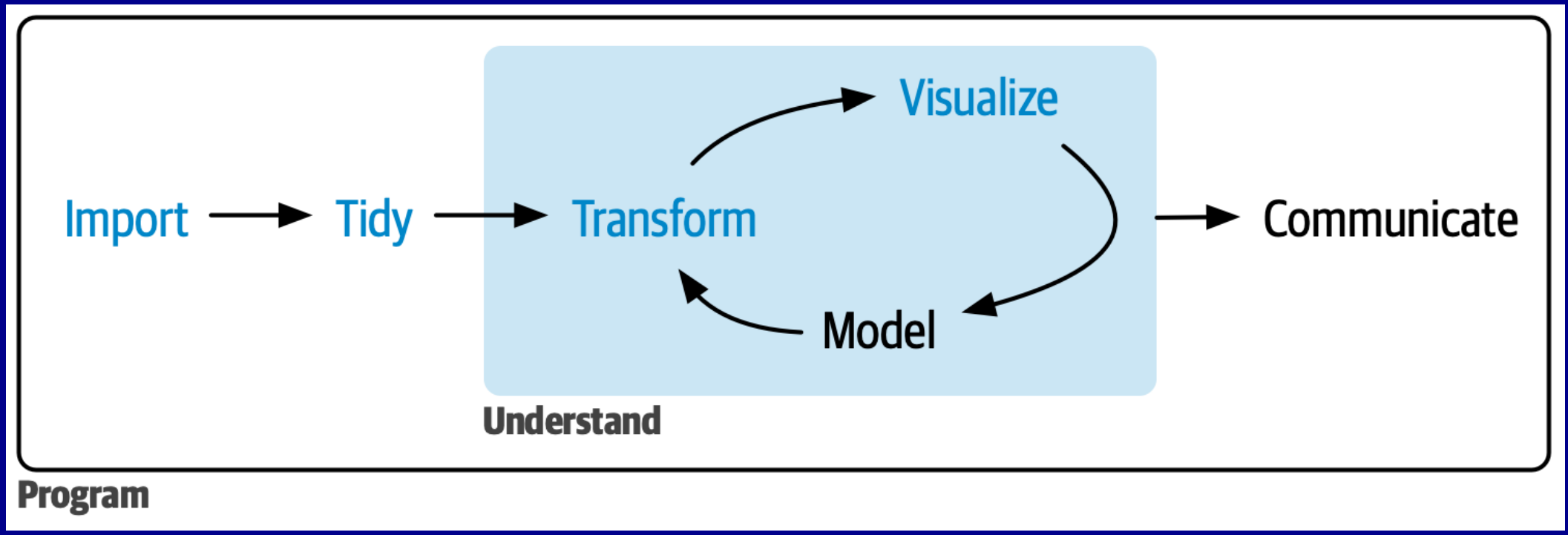
Vaše tipy?

=> **Co to je za data? Kde se vzala? Kde je dokumentace?**

Krok stranou: ČSÚ, katalogy, číselníky

Co s daty dělat

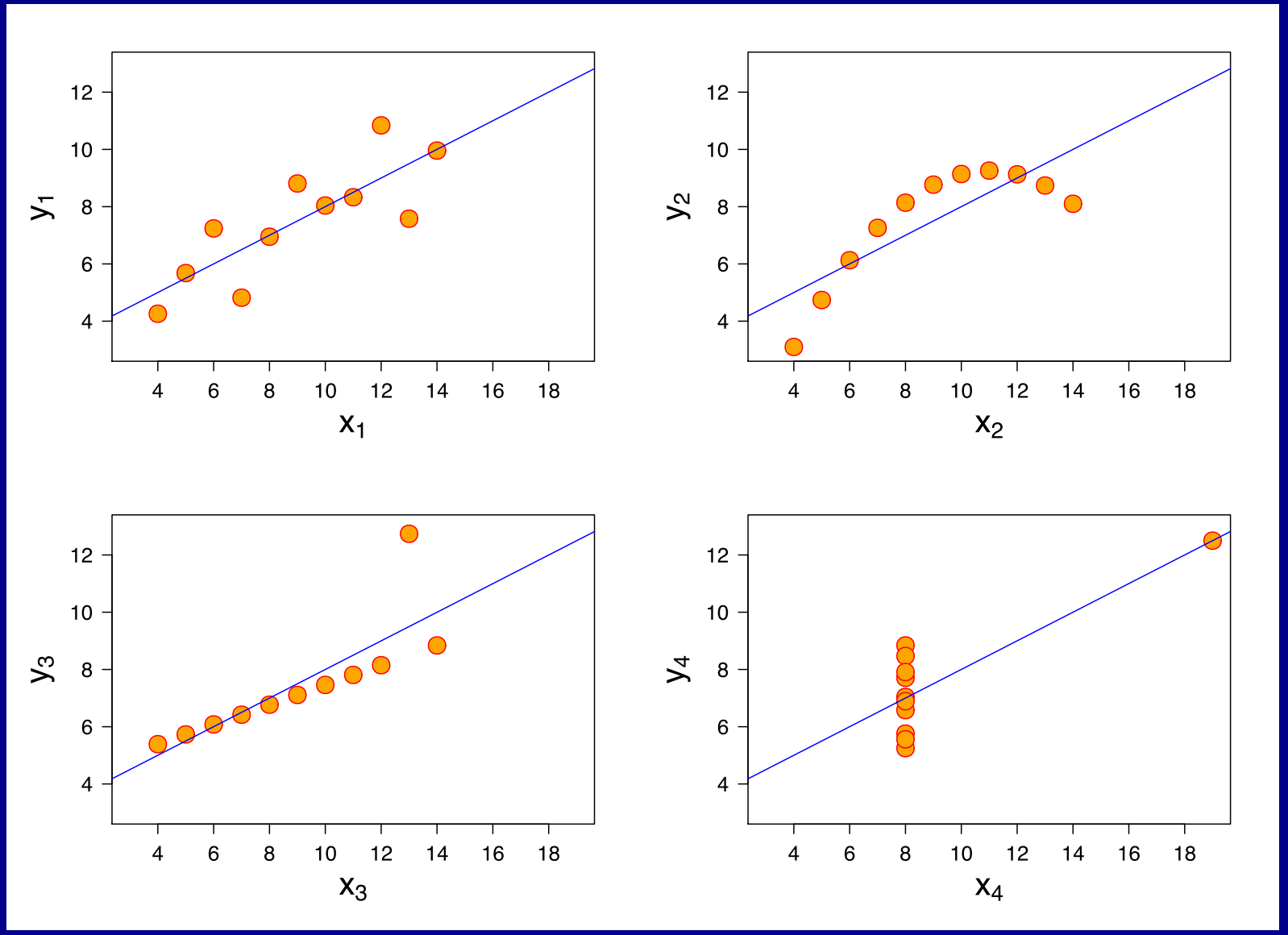
Mentální model



Teze pro postupy

- **Čištění a zpracování dat** jsou součástí analýzy
- **Iterace**
- Vyplatí se investovat do rozumného **postupu a dokumentace**
- **Vizualizace** nejen jako výstup, ale jako nástroj analýzy

Vizualizace



Neboli

Není to lineární proces

Spolupracujete se svým budoucím já.
Vaše budoucí já vám poděkuje
(kolega*yně taky)

Žádná analýza není jednorázovka

Postup + dokumentace =
zkontrolovatelnost
opakovatelnost
automatizovatelnost

Na postupu a organizaci záleží

petrbouchal.xyz/czecheval2023

Co vás štve, když otevřete něčí starý Excel nebo složku s analýzou?

Jak tedy na to

📋 Dokumentovat data, postup, soubory

📥 Oddělit vstupní data od analýzy

📂 Organizace projektu na disku

(👩‍💻 Pracovat v kódu ideálně s open source nástroji)

🔌 (Vytvářet kompletní opakovatelné workflow)

🔢 Verzovat postup (kód) a pokud možno i výstupy

✅ Využívat strojovou validaci vstupů a výstupů

(📕 Používat nástroje na propojení textu a kódu: Jupyter, Rmarkdown, Quarto)

(💿 Zaznamenat prostředí, kde kód běží)

Jak to provést

Organizace projektu


data


analyza


vystupy


README.docx

Ideálně konzistentní napříč projekty

Názvy souborů

> “Naming things is hard”

❌ `hotovo-final1-fin2.1led21.xlsx`

✅ `report01-vypocty_20220926.xlsx`

- popisný název - čitelný pro lidi
- standardní formát datumu (2022-10-23)
- funguje abecední řazení
- lze strojově zpracovat - čitelný pro stroj

Názvy proměnných, sloupců aj.

- konzistentně (CamelCase, snake_case, cesky, Česky)
- strojová vs. lidská čitelnost

Organizace dat: tvary

Data mají různé tvary: dlouhá, široká, něco mezi

Různé tvary pro různé účely

- “dlouhá” data často lepší pro analýzy
- široká pro čtení/srovnání očima

Dostat data do správného tvaru je úkol sám o sobě

Co identifukuje jednu řádku?

Organizace dat: tvary

Základ:

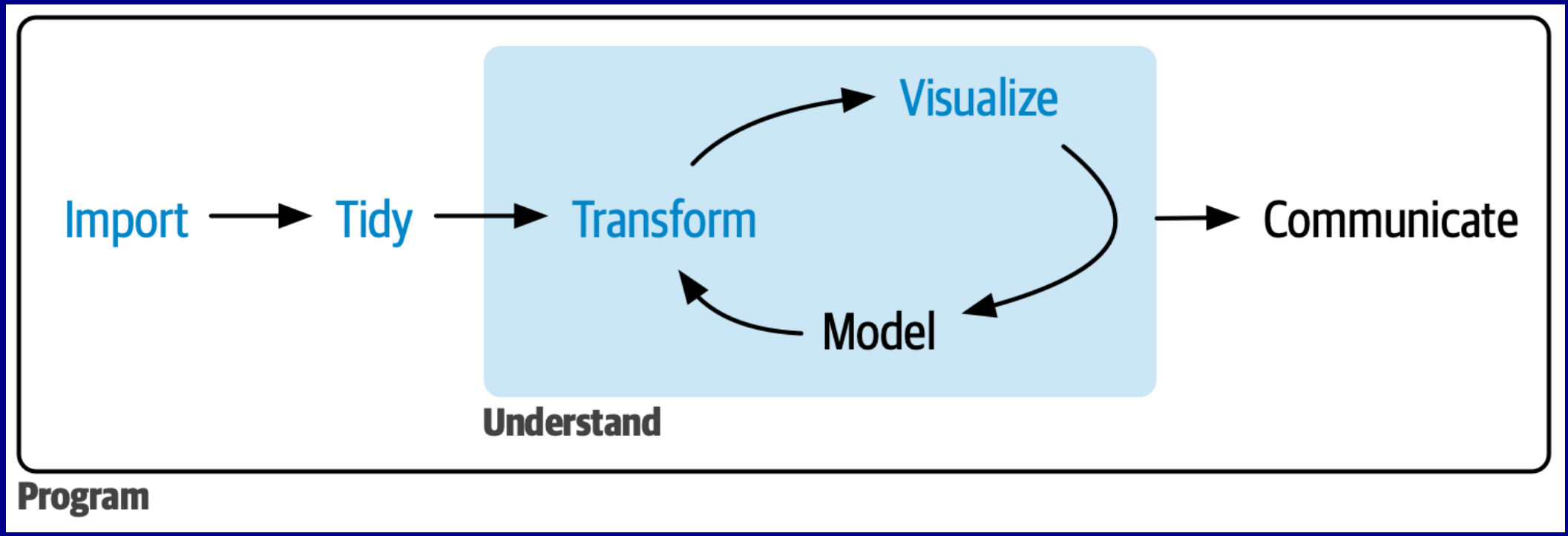
- co informace, to buňka/sloupec
- co řádek, to pozorování
- co datová sada, to tabulka / list / objekt
- ale: co je pozorování??

`pivot` - `unpivot`

`databázová normalizace`

Organizace dat: dobré praxe

- co informace, to buňka/sloupec
- Excel: co list, to tabulka
- Excel: data = text, ne formát
- Excel: přímá napojení na zdroje
- Excel: použít funkci tabulky
- dokumentace (metadata / codebook) blízko dat
- data ukládat nefiltrovaná, kurzor na začátku atd.
- udržovat informaci o původu dat (jasné ID zdroje; URL)

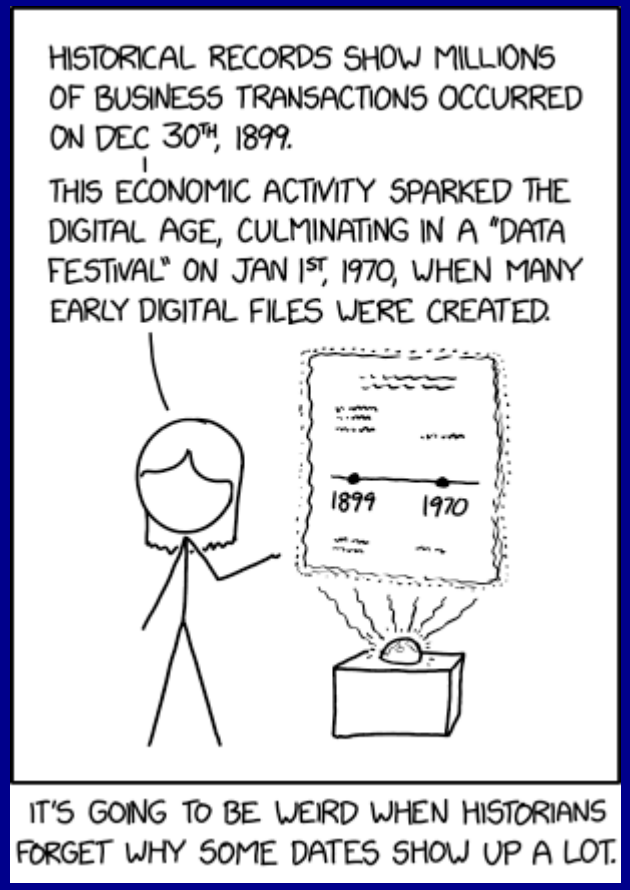

Import
Tidy
Transform
Visualize
Model
Communicate
Understand
Program

Krok stranou: náš datový úkol

“Hledá se obec s největším podílem neobydlených bytů v každém ORP ČR.”

Načtení dat

- CSV / TSV / ; / ,
- pozor na formáty (desetinné čárky, mezery)
- datumy!
- znaková sada
- Excel: PowerQuery >> standard Excel
- NULL, –, -99, “_”

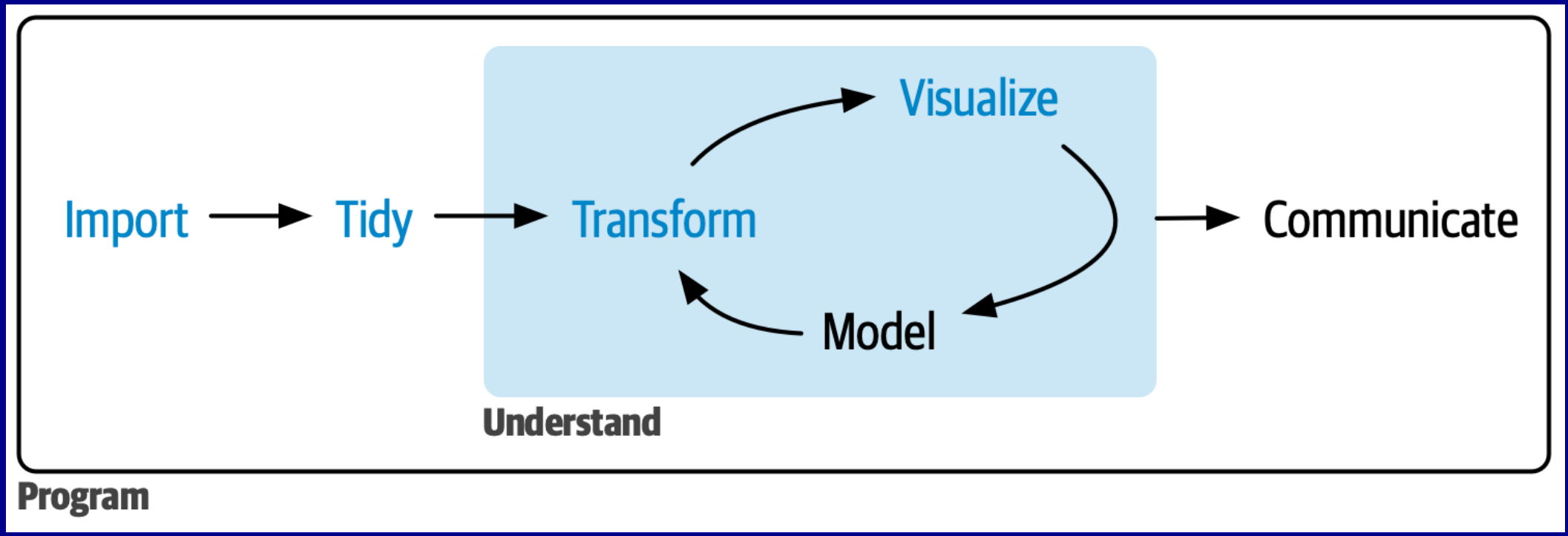


Krok stranou: nástroj PowerQuery

Co

Proč

Jak


Import
Tidy
Transform
Visualize
Model
Communicate
Understand
Program

Čištění dat

Cíl: data vhodná pro analýzu

Formáty

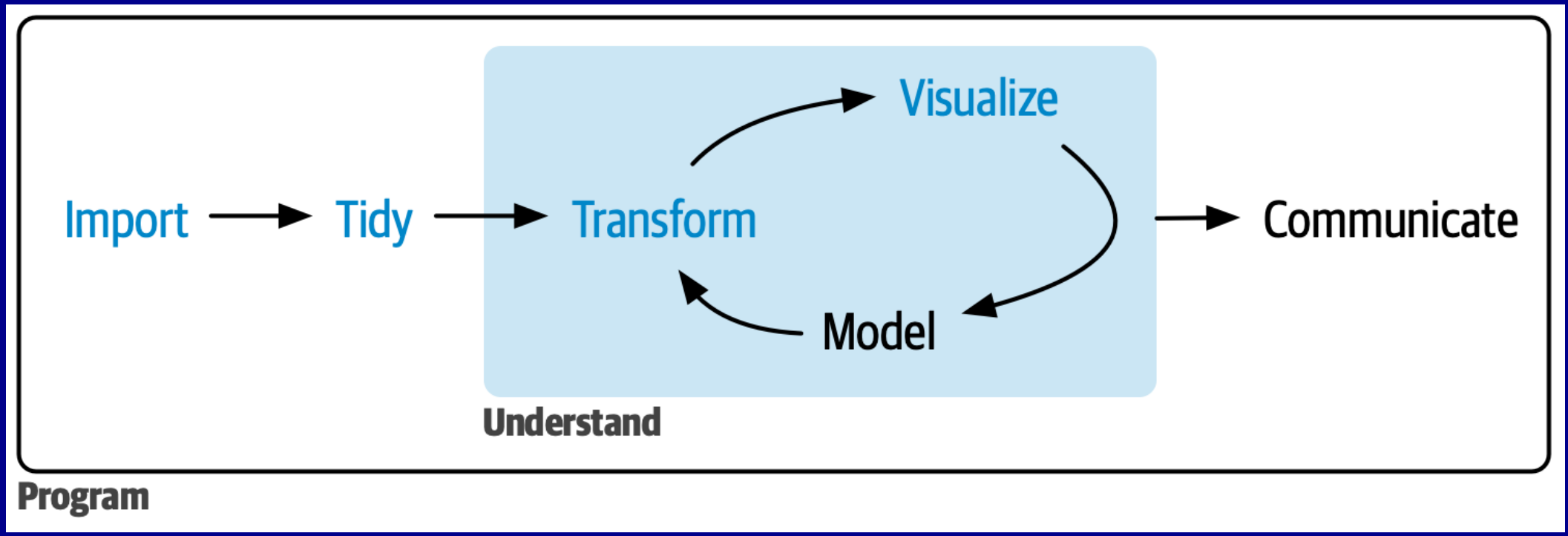
Nepřesnosti?

Duplicity?

Chybějící data?

Nesprávné hodnoty?

Extrémní hodnoty?


Import
Tidy
Transform
Visualize
Model
Communicate
Understand
Program

Transformace dat

Široká

uzemi_kod	uzemi_txt		obvykle neobydlen	obvykle obydlen
500011	Želechovice nad Dřevnicí	859	190	669
500020	Petrov nad Desnou	552	83	469
500046	Libhošť	764	137	627
500062	Krhová	910	128	782
500071	Poličná	771	119	652
500101	Bražec	101	12	89
500127	Doupovské Hradiště	70	6	64
500135	Kozlov	153	32	121
500151	Luboměř pod Strážnou	53	3	50
500160	Město Libavá	267	46	221
500194	Polná na Šumavě	72	2	70

Dlouhá

sldb_rok	uzemi_cis	ukaz_txt	obydlenost_txt	uzemi_txt	uzemi_kod	hodnota
2021	43	Počet bytů		Želechovice nad Dřevnicí	500011	859
2021	43	Počet bytů	obvykle obydlen	Želechovice nad Dřevnicí	500011	669
2021	43	Počet bytů	obvykle neobydlen	Želechovice nad Dřevnicí	500011	190
2021	43	Počet bytů		Petrov nad Desnou	500020	552
2021	43	Počet bytů	obvykle obydlen	Petrov nad Desnou	500020	469
2021	43	Počet bytů	obvykle neobydlen	Petrov nad Desnou	500020	83
2021	43	Počet bytů		Libhošť	500046	764
2021	43	Počet bytů	obvykle obydlen	Libhošť	500046	627
2021	43	Počet bytů	obvykle neobydlen	Libhošť	500046	137
2021	43	Počet bytů		Krhová	500062	910
2021	43	Počet bytů	obvykle obydlen	Krhová	500062	782
2021	43	Počet bytů	obvykle neobydlen	Krhová	500062	128
2021	43	Počet bytů		Poličná	500071	771
2021	43	Počet bytů	obvykle obydlen	Poličná	500071	652
2021	43	Počet bytů	obvykle neobydlen	Poličná	500071	119
2021	43	Počet bytů		Bražec	500101	101
2021	43	Počet bytů	obvykle obydlen	Bražec	500101	89
2021	43	Počet bytů	obvykle neobydlen	Bražec	500101	12
2021	43	Počet bytů		Doupovské Hradiště	500127	70
2021	43	Počet bytů	obvykle obydlen	Doupovské Hradiště	500127	64
2021	43	Počet bytů	obvykle neobydlen	Doupovské Hradiště	500127	6

Dlouhá

ukaz_txt	stav_zeny_txt	vzdelani_zeny_txt	vek_zeny_txt	uzemi_txt	hodnota
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	15 a více let	Praha	4621
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	15 – 19 let	Praha	80
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	20 – 24 let	Praha	440
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	25 – 29 let	Praha	844
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	30 – 34 let	Praha	898
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	35 – 39 let	Praha	747
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	40 – 44 let	Praha	648
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	45 – 49 let	Praha	468
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	50 – 54 let	Praha	167
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	55 – 59 let	Praha	98
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	60 – 64 let	Praha	66
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	65 – 69 let	Praha	70
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	70 – 74 let	Praha	38
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	75 – 79 let	Praha	44
Počet živě narozených dětí	Svobodná	Základní vč. neukončeného	80 a více let	Praha	13
Počet živě narozených dětí	Vdaná	Základní vč. neukončeného	15 a více let	Praha	16219
Počet živě narozených dětí	Vdaná	Základní vč. neukončeného	15 – 19 let	Praha	3
Počet živě narozených dětí	Vdaná	Základní vč. neukončeného	20 – 24 let	Praha	101

Široká

ukaz_txt	stav_zeny_txt	vek_zeny_txt	uzemi_txt	Bez vzdělání	Nezjištěno	Střední vč. vyučení (bez maturity)	Úplné střední (s maturitou), vč. nástavbového a pomaturitního	Vysokoškolské	Vyšší odborné, konzervatoř	Základní vč. neukončeného
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Hlavní město Praha	1	2	1	0	0	0	2
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Jihočeský kraj	0	0	0	0	0	0	0
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Jihomoravský kraj	0	0	0	0	0	0	0
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Jihovýchod	0	0	0	0	0	0	1
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Jihozápad	0	0	0	0	0	0	0
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Karlovarský kraj	0	0	0	0	0	0	0
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Kraj Vysočina	0	0	0	0	0	0	1
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Královéhradecký kraj	0	0	0	0	0	0	0
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Liberecký kraj	0	0	0	0	0	0	1
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Moravskoslezsko	0	0	0	0	0	0	0
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Moravskoslezský kraj	0	0	0	0	0	0	0
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Olomoucký kraj	0	2	0	0	0	0	0
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Pardubický kraj	0	0	0	2	0	0	1
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Plzeňský kraj	0	0	0	0	0	0	0
Počet živě narozených dětí	Nezjištěno	15 – 19 let	Praha	1	2	1	0	0	0	2

Široká

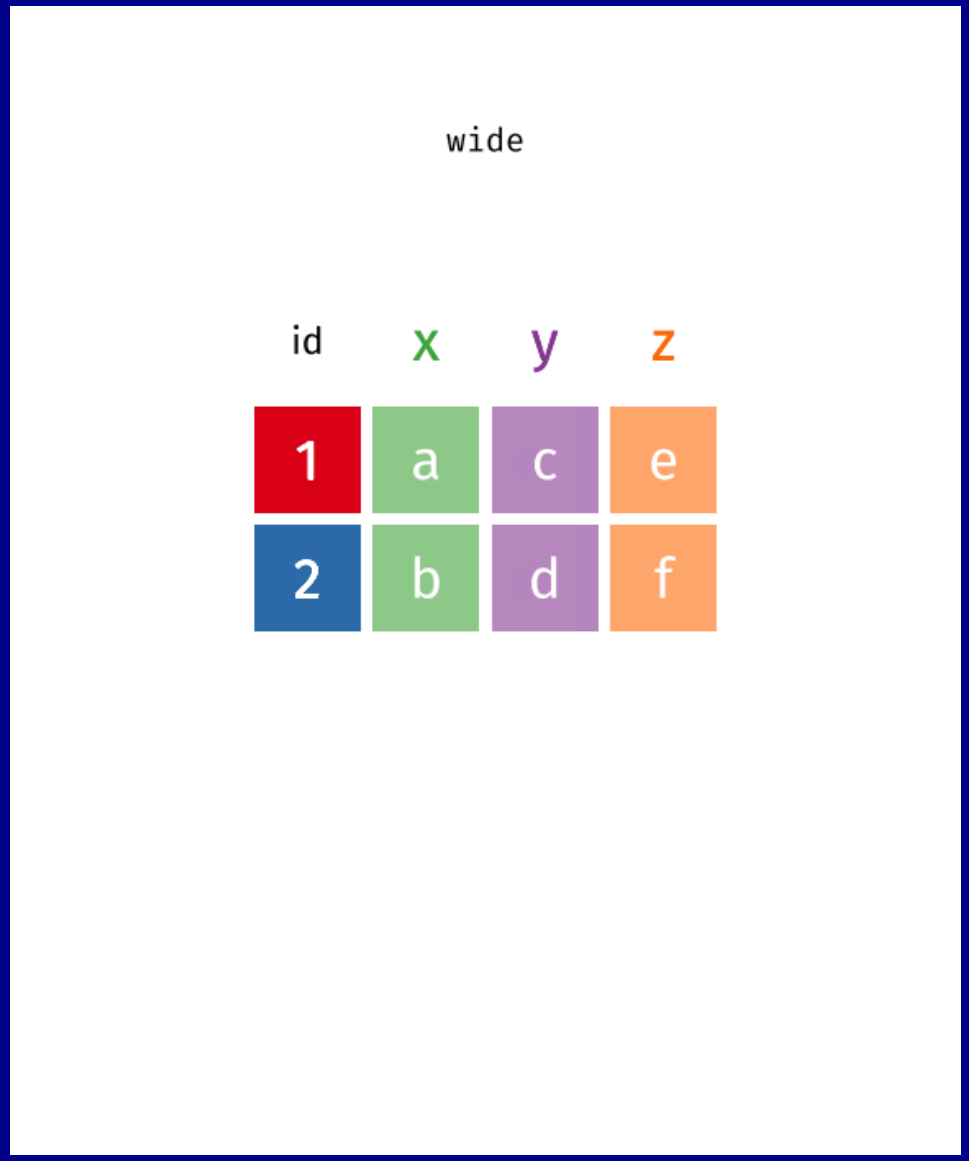
stav_zeny_txt	vzdelani_zeny_txt	vek_zeny_txt	uzemi_txt	Počet živě narozených dětí	Podíl počtu živě narozených dětí na 1000 žen	Podíl počtu živě narozených dětí na 1000 žen s aspoň 1 dítětem
Nezjištěno	Bez vzdělání	15 – 19 let	Hlavní město Praha	1	250.0000	1000.000
Nezjištěno	Bez vzdělání	15 – 19 let	Jihočeský kraj	0	0.0000	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Jihomoravský kraj	0	0.0000	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Jihovýchod	0	0.0000	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Jihozápad	0	0.0000	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Karlovarský kraj	0	*NA*	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Kraj Vysočina	0	*NA*	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Královéhradecký kraj	0	*NA*	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Liberecký kraj	0	*NA*	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Moravskoslezsko	0	*NA*	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Moravskoslezský kraj	0	*NA*	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Olomoucký kraj	0	*NA*	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Pardubický kraj	0	0.0000	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Plzeňský kraj	0	0.0000	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Praha	1	250.0000	1000.000
Nezjištěno	Bez vzdělání	15 – 19 let	Severovýchod	0	0.0000	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Severozápad	0	*NA*	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Střední Čechy	0	0.0000	*NA*
Nezjištěno	Bez vzdělání	15 – 19 let	Střední Morava	0	*NA*	*NA*

Široká

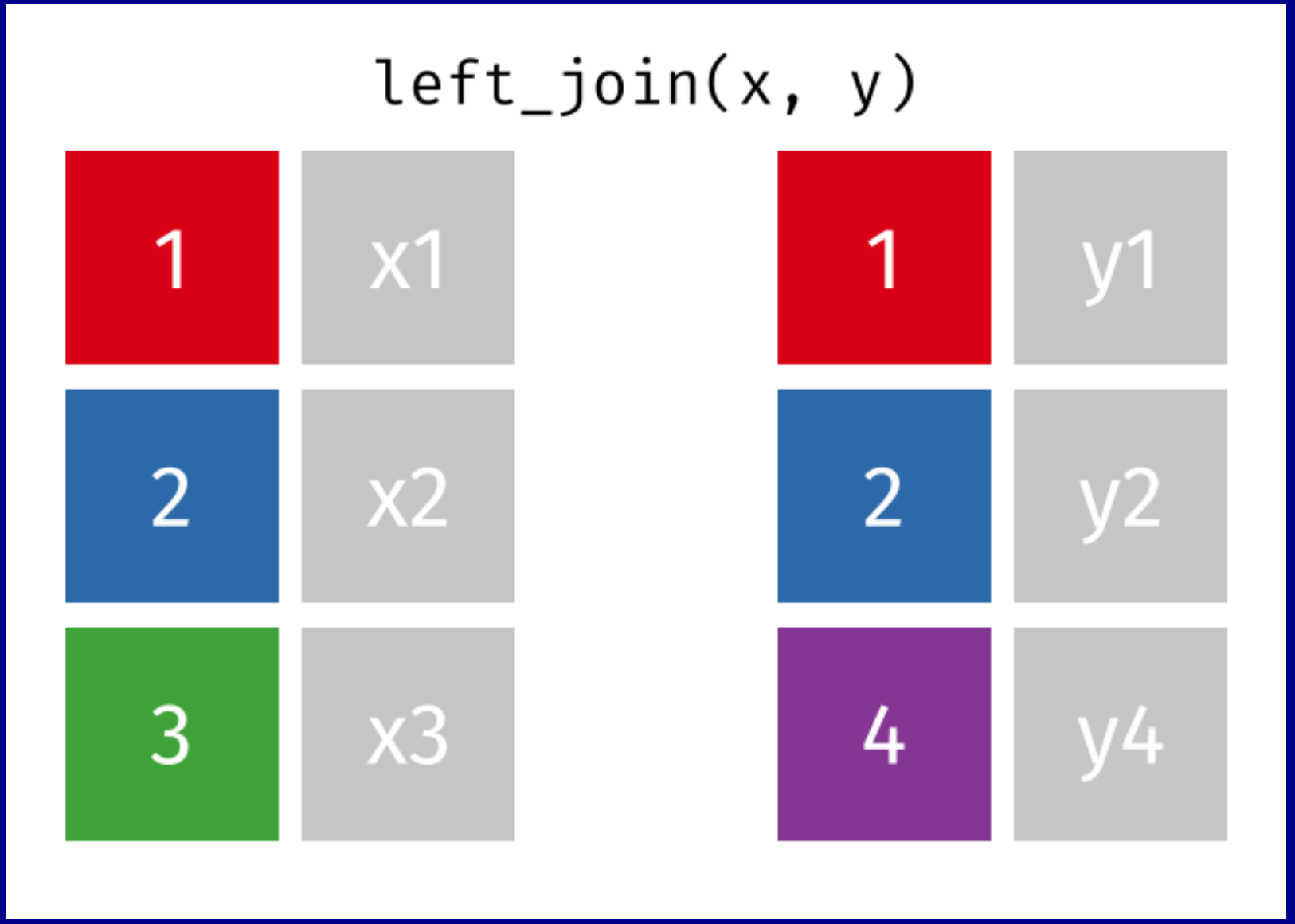
ukaz_txt	stav_zeny_txt	vzdelani_zeny_txt	vek_zeny_txt	Hlavní město Praha	Jihočeský kraj	Jihomoravský kraj
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	15 – 19 let	1	0	0
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	15 a více let	16	1	5
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	20 – 24 let	0	0	1
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	25 – 29 let	2	1	0
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	30 – 34 let	0	0	0
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	35 – 39 let	2	0	0
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	40 – 44 let	2	0	0
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	45 – 49 let	2	0	0
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	50 – 54 let	3	0	0
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	55 – 59 let	2	0	2
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	60 – 64 let	0	0	2
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	65 – 69 let	0	0	0
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	70 – 74 let	0	0	0
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	t	2	0	0
Počet živě narozených dětí	Nezjištěno	Bez vzdělání	80 a více let	0	0	0
Počet živě narozených dětí	Nezjištěno	Nezjištěno	15 – 19 let	2	0	0
Počet živě narozených dětí	Nezjištěno	Nezjištěno	15 a více let	79	57	50

Bez vzdělání

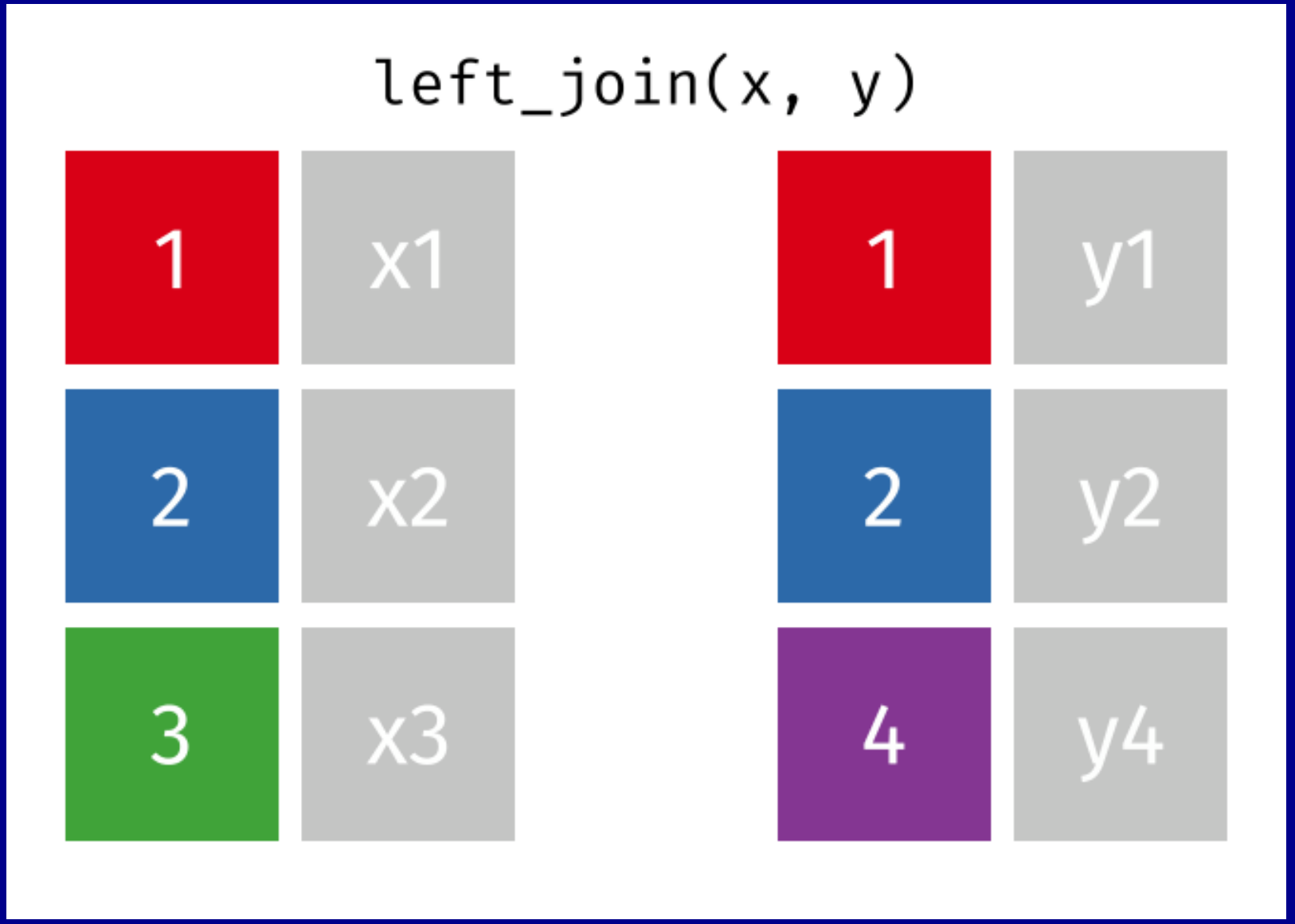
Transformace



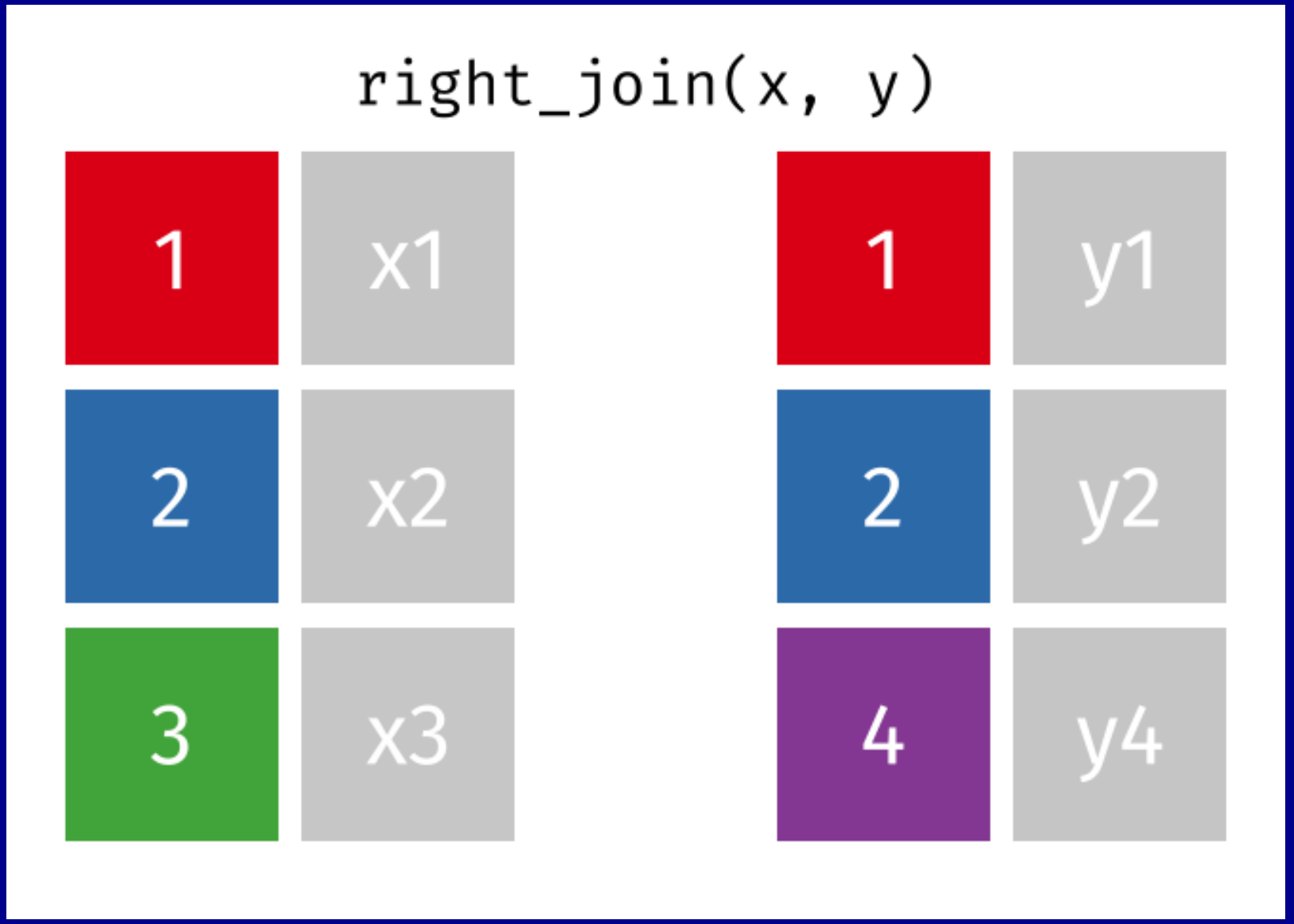
Propojování



Propojování



Propojování



Proč se propojování nedaří

- mezery na začátku a konci
- velká a malá písmena
- dvojité mezery

Číselníky, klasifikace atd.

obec_kod	ukaz_txt	obec_nazev	obvykle obydlen	obvykle neobydlen
554979	Počet bytů	Abertamy	376	201
581291	Počet bytů	Adamov	2042	152
535826	Počet bytů	Adamov	349	42
531367	Počet bytů	Adamov	57	25
547786	Počet bytů	Adršpach	203	95
598925	Počet bytů	Albrechtice	1508	182
547981	Počet bytů	Albrechtice	170	42
576077	Počet bytů	Albrechtice nad Orlicí	388	61
549258	Počet bytů	Albrechtice nad Vltavou	379	230
563528	Počet bytů	Albrechtice v Jizerských horách	124	178
568741	Počet bytů	Albrechtičky	247	67
506761	Počet bytů	Alojzov	93	30
551929	Počet bytů	Andělská Hora	145	42
538001	Počet bytů	Andělská Hora	122	22

Číselníky, klasifikace atd.

https://apl.czso.cz/iSMS/

https://www.cuzk.cz/ruian/Poskytovani-udaju-ISUI-RUIAN-VDP/Ciselniky-ISUI.aspx

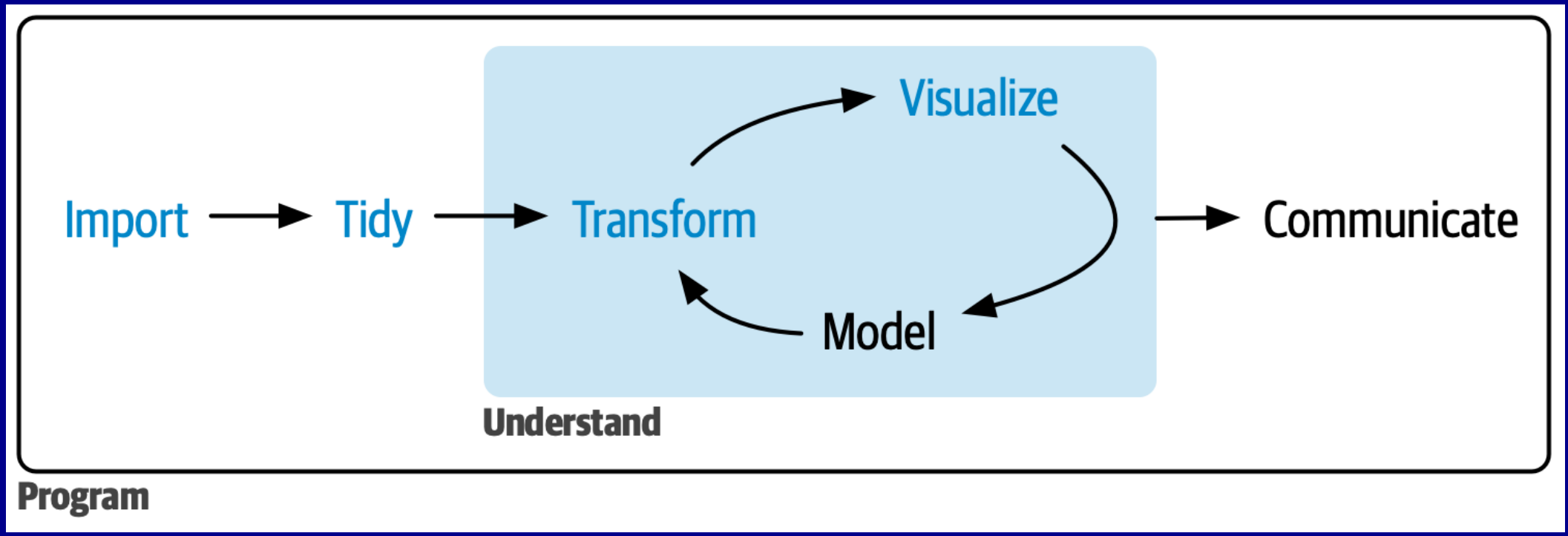
obec_kod	obec_nazev	kraj_kod	kraj_nazev
554979	Abertamy	CZ041	Karlovarský kraj
531367	Adamov	CZ020	Středočeský kraj
535826	Adamov	CZ031	Jihočeský kraj
581291	Adamov	CZ064	Jihomoravský kraj
547786	Adršpach	CZ052	Královéhradecký kraj
547981	Albrechtice	CZ053	Pardubický kraj
598925	Albrechtice	CZ080	Moravskoslezský kraj
576077	Albrechtice nad Orlicí	CZ052	Královéhradecký kraj
549258	Albrechtice nad Vltavou	CZ031	Jihočeský kraj
563528	Albrechtice v Jizerských horách	CZ051	Liberecký kraj

Proč číselník nesedí

- jiný číselník
- jiná verze číselníků
- neshody mezi tvůrcem dat a tvůrcem číselníku
  - je Praha okres?
  - jsou městské části Prahy ORP?

Číselníky, klasifikace atd.

- pozor na platnost (ke kterému datu?)
- pozor na verze (kraje, NUTS kraje, staré NUTS kraje)
- každý číselník správně má jednoho správce
- ale jsou tu překryvy, správci různých dat občas zveřejňují duplicitní číselníky (např. MPSV k datům o nezaměstnanosti)


Import
Tidy
Transform
Visualize
Model
Communicate
Understand
Program

Výpočty a modelování

Exploratorní datová analýza

Explorace a poznávání dat

Proč

Poznat data

Najít problémy

Poznat variabilitu

Detekovat pravidelnosti

Jak

Koukat na data

Hledat podivnosti

Vizualizovat

Sčítat a porovnávat

Zkoumat výseky dat

Co lze v Excelu

Sařadit si data

Souhrnné statistiky

Kontingenční tabulka

Histogram

Box plot

Náhled rozdělení v PowerQuery

Heatmapy (podmíněné formátování)

Základní mapy (kraje)

Korelace, regrese, t-testy

Kam dál

Další nástroje

- Extrakce dat: OpenRefine, Tabula
- Vizualizace: Datawrapper, RawGraphs, Flourish
- Interaktivní: Google Data Studio, PowerBI
- Regulérní výrazy

Práce s kódem

SQL a databáze

R nebo Python?

=> automatizace

Literate programming

ObservableJS

Jupyter Notebooks

R Markdown / Quarto

Verzování (git, Github)

Reflexe

Díky!

pbouchal@gmail.com

petrbouchal.xyz