Práce s (kvantitativními) daty

Hlavní teze z bloku vzdělávacího programu pro analytiky veřejné správy

[Všechny podklady k této části programu na webu](https://petrbouchal.xyz/analytici/).

## Jak vypadá datová analýza dnes

Klade se důraz na uživatele a porozumění jeho potřebám. Důvěru uživatelů výstupů analytik získává svou schopností reagovat na jejich potřeby, [transparentností a integritou postupu](https://best-practice-and-impact.github.io/qa-of-code-guidance/intro.html) a komunikací výsledků.

Pracuje se v kódu, vytváří se tzv. [reprodukovatelné workflow](https://analysisfunction.civilservice.gov.uk/support/reproducible-analytical-pipelines/).

Pracuje se iterativně, práce řídí agilně, vytváří se spíš [produkty](https://locallyoptimistic.com/post/data-product-manager/) než projekty.

Častěji vznikají webové a interaktivní produkty, zároveň [slábne předěl mezi analytickým postupem a výstupem](https://r4ds.hadley.nz/quarto.html).

## Jak o tom přemýšlet: mentální model postupu práce s daty



Zdroj: <https://r4ds.hadley.nz/whole-game.html>

Práce s daty má své fáze, ale je to také iterativní proces. Čištění a transformace dat jsou součástí analýzy. Vizualizace dat slouží nejen ke komunikaci, ale i k analýze – porozumění dat.

K analýzám se často musíme vracet, proto je třeba dokumentovat data a postup.

Spolupracujete buď z kolegy, nebo s vaším budoucím já: jedni i druzí vám budou vděční, pokud budete dobře dokumentovat, kde se vzala jaká data, co jsme s nimi udělali a jak a proč.

## Jak to uřídit a vyznat se v tom

Když si ustálíte základní postupy a standardy – pro sebe nebo v týmu –, zůstane vám víc mentální kapacity na obsahová/analytická rozhodnutí namísto těch triviálních. Také si tím snížíte bariéru k tomu tyto praktické principy dodržovat. Rozhodněte si předem a dodržujte:

* předem daný způsob [organizace](https://rstats.wtf/projects.html#work-in-a-project) a dokumentace projektu, [dat](https://doi.org/10.1080/00031305.2017.1375989), postupu (readme, krycí list datového souboru)
* organizace a [názvy souborů](https://github.com/jennybc/how-to-name-files/raw/main/how-to-name-files.pdf) v projektové složce
* [názvy proměnných/sloupců](http://www2.stat.duke.edu/~rcs46/lectures_2015/01-markdown-git/slides/naming-slides/naming-slides.pdf)
* způsob, jak okamžitě zachycovat poznámky z postupu práce; cokoli byste později zapomněli, hned zapište.
* postup pro uzavření/předání analýzy (dokument, schůzka, …): přiměje vás vše zachytit a zdokumentovat

## Kde vzít data

ČSÚ některá data [poskytuje i jako otevřená data](https://www.czso.cz/csu/czso/katalog-produktu?filtr=true&vlastnosti=22); mnohá data ČSÚ jsou prezentována i v [katalogu Eurostatu](https://ec.europa.eu/eurostat/web/products-catalogues), často v analyticky přívětivější formě nebo v alternativních agregacích, které se vám mohou hodit.

Pro analytickou práci využívejte data ve standardizovaných formátech: u ČSÚ a Eurostatu otevřená data. Využijte katalogy [ČSÚ](https://www.czso.cz/csu/czso/katalog-produktu?filtr=true&vlastnosti=22) a [Eurostatu](https://ec.europa.eu/eurostat/web/products-catalogues) či [Národní katalog otevřených dat](https://data.gov.cz/datov%C3%A9-sady). Data o životním prostředí a zdrav(otnictv)í hledejte u ([CENIA](https://www.cenia.cz/data/) resp. [ÚZIS](https://www.uzis.cz/index.php?pg=vystupy)).

Číselníky a jiná metadata typicky spravuje ČSÚ ([Databáze metainformací](https://www.czso.cz/csu/czso/databaze_metainformaci)); používejte ty správné, aktuální a od zdroje. Prostorová data hledejte na [ČÚZK](https://geoportal.cuzk.cz/%28S%283ulvz4m5hf2m3zsnuh0s0qxq%29%29/Default.aspx?mode=TextMeta&side=data_uvod&text=data_sady&head_tab=sekce-05-gp&menu=52&news=yes); velká část už je dostupná v otevřené formě.

Spolu s daty stahujte (a čtětě!) dokumentaci a metadata, abyste rozuměli, jak data vznikla, co v nich je a není, na co se v nich dá a nedá spolehnout.

Pokud pracujete s daty o území, může vám pomoct:

* Databáze [KROK](https://www.czso.cz/csu/czso/databaze-krok-otevrena-data)/[MOS](https://www.czso.cz/csu/czso/databaze-mos-otevrena-data): krajská/obecní data zkompilovaná do jedné velké sady ve formě otevřených dat
* [Struktura území](https://www.czso.cz/csu/czso/uzemi-sidelni-struktura): v excelové i strojově čitelné formě (otevřená data)

Pokud sbíráte data nebo vytváříte nové datové sady spojováním a transformací jiných, stáváte se [de facto správci dat](https://www.gov.uk/guidance/creating-and-sharing-spreadsheets) (v praktickém smyslu, ne právním). Dokumentujte, jak data vznikla, odkud pocházejí, co v nich je, jak je použít. Tak, abyste je mohli snadno a bez velkého vysvětlování někomu předat.

Pokud pro vás analytickou práci dělá někdo jiný - dodavatel, resortní výzkumák, jiný tým - je rozumné po něm chtít dokumentaci a předávku dat a projektu na podobné úrovni.

## Proč na to zkusit vzít PowerQuery (a kdy vzít do ruky něco jiného)

[PowerQuery](https://support.microsoft.com/en-us/office/about-power-query-in-excel-7104fbee-9e62-4cb9-a02e-5bfb1a6c536a) v Excelu vám – oproti standardnímu Excelu – pomůže

* uvést do praxe principy moderní práce s daty: dokumentace, transparentnost postupu, reprodukovatelnost a zkontrolovatelnost, automatizace, [nedestruktivní přístup ke zdrojovým datům](https://support.microsoft.com/en-us/office/import-or-export-text-txt-or-csv-files-5250ac4c-663c-47ce-937b-339e391393ba#ID0EBBN=Newer_versions)
* [osvojit si postupy užitečné při přechodu k práci v programovacích a databázových nástrojích](https://taraskaduk.com/posts/2018-03-29-power-query/)
* provést některé [operace, které v Excelu nejdou](https://exceloffthegrid.com/power-query-unpivot-data/)
* naladit se na [logiku práce s daty v PowerBI](https://learn.microsoft.com/en-us/power-bi/transform-model/desktop-query-overview)
* využít [datové zdroje ve formátech, které z Excelu nejsou dostupné](https://learn.microsoft.com/en-us/power-query/connectors/)
* více se soustředit na to, co data říkají, než jak s nimi operovat

Na složitější věci budete potřebovat jiné nástroje: na vizualizace zkuste [RawGraphs](https://www.rawgraphs.io/), [Datawrapper](https://www.datawrapper.de/), PowerBI či Tableau; do složitější statistiky, větších nebo komplexnějších dat a větších projektů se pusťte spíš v R/Pythonu nebo Stata a o kód se starejte v gitu. To jsou také technologie, které má smysl se učit, pokud vám Excel nestačí.