# Práce s (kvantitativními) daty

## Hlavní teze z bloku vzdělávacího kurzu pro analytiky veřejné správy

Všechny podklady k této části kurzu na webu.

### Jak vypadá datová analýza dnes

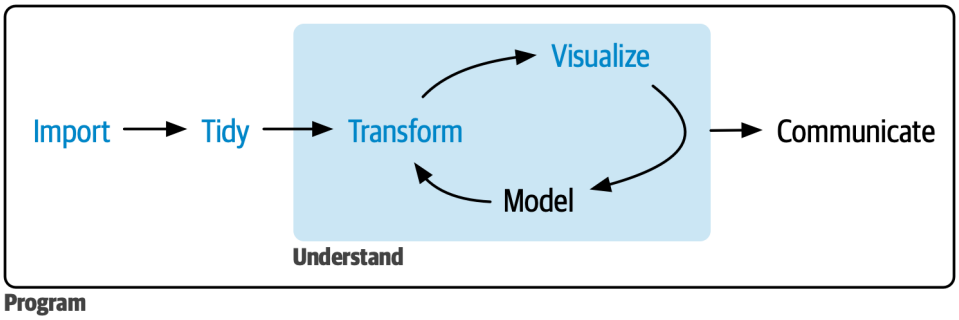
Klade se důraz na uživatele a porozumění jeho potřebám. Důvěru uživatelů výstupů analytik získává svou schopností reagovat na jejich potřeby, transparentností a integritou postupu a komunikací výsledků.

Pracuje se v kódu, vytváří se tzv. reprodukovatelné workflow.

Pracuje se iterativně, práce řídí agilně, vytváří se spíš produkty než projekty.

Častěji vznikají webové a interaktivní produkty, zároveň slábne předěl mezi analytickým postupem a výstupem.

### 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 a dokumentace projektu, dat, postupu (readme, krycí list datového souboru)
- organizace a názvy souborů v projektové složce
- názvy proměnných/sloupců
- 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; mnohá data ČSÚ jsou prezentována i v katalogu Eurostatu, 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 a Eurostatu či Národní katalog otevřených dat. Data o životním prostředí a zdrav(otnictv)í hledejte u (CENIA resp. ÚZIS).

Číselníky a jiná metadata typicky spravuje ČSÚ (Databáze metainformací); používejte ty správné, aktuální a od zdroje. Prostorová data hledejte na ČÚZK; velká část už je dostupná v otevřené formě.

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

Pokud sbíráte data nebo vytváříte nové datové sady spojováním a transformací jiných, stáváte se de facto správci dat (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.

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

PowerQuery 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
- osvojit si postupy užitečné při přechodu k práci v programovacích a databázových nástrojích
- provést některé operace, které v Excelu nejdou
- naladit se na logiku práce s daty v PowerBI
- využít datové zdroje ve formátech, které z Excelu nejsou dostupné
- více se soustředit na to, co data říkají, než jak s nimi operovat

Na složitější věci budete potřebovat jiné nástroje: na vizualizace zkuste RawGraphs, Datawrapper, 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.