Open Data Lab

Workshop #3:

Študijný materiál

*Toto podujatie bolo podporené z dotačnej schémy Ministerstva školstva, vedy, výskumu a športu SR Programy pre mládež na roky 2014 – 2021, ktorú administruje IUVENTA – Slovenský inštitút mládeže.*



Autor: Lukáš Jankovič  
[lukas.jankovic@alvaria.sk](mailto:lukas.jankovic@alvaria.sk)

Obsah

[Spracovanie otvorených dát Štatistického úradu SR 3](#_Toc87880612)

[Spracovanie otvorených dát mesta Prešov 5](#_Toc87880613)

[Ako urobiť projekt nad otvorenými dátami čo najlepší 10](#_Toc87880614)

**Poznámka:**

Ak je v texte spomenuté stlačenie tlačidla myši, myslí sa tým ľavé tlačidlo. V prípade ak ide o pravé tlačidlo, je to v texte uvedené.

**Pojmy**

Prvok – graf, tabuľka, mapa

Vizualizácia – celý súbor (môže obsahovať viac prvkov a viac strán)

# Spracovanie otvorených dát Štatistického úradu SR

Štatistický úrad SR zverejňuje otvorené dáta pomocou týchto kanálov:

1. [Profil na centrálnom portáli otvorených dát data.gov.sk](https://data.gov.sk/organization/f4787c6f-9fa3-406c-b8d5-d374f1e1f2d3)
2. [DataCube](http://datacube.statistics.sk/)
3. [API](https://slovak.statistics.sk/wps/portal/ext/Databases/Open_data/!ut/p/z1/jY_BCsIwGIMfqandbD12wrpCFX9Hu9mL9CQFnR7E51eGV-tyC3xJCItsZHFKr3xJz3yf0vXjT3F9Du6gmoZrKB4kbNtTRX7HDdVsmIGt0V0lHaCcqWF1548bEgJasLgkjx_SWJYvALFcP7A4I5qIehcCTFi1sIIb7L0HWvkFShf_jTxu3o_I9g2cryaW/dz/d5/L2dJQSEvUUt3QS80TmxFL1o2X1ZMUDhCQjFBMDBVTDkwQVJEMzE4SjYyQ1Mx/)

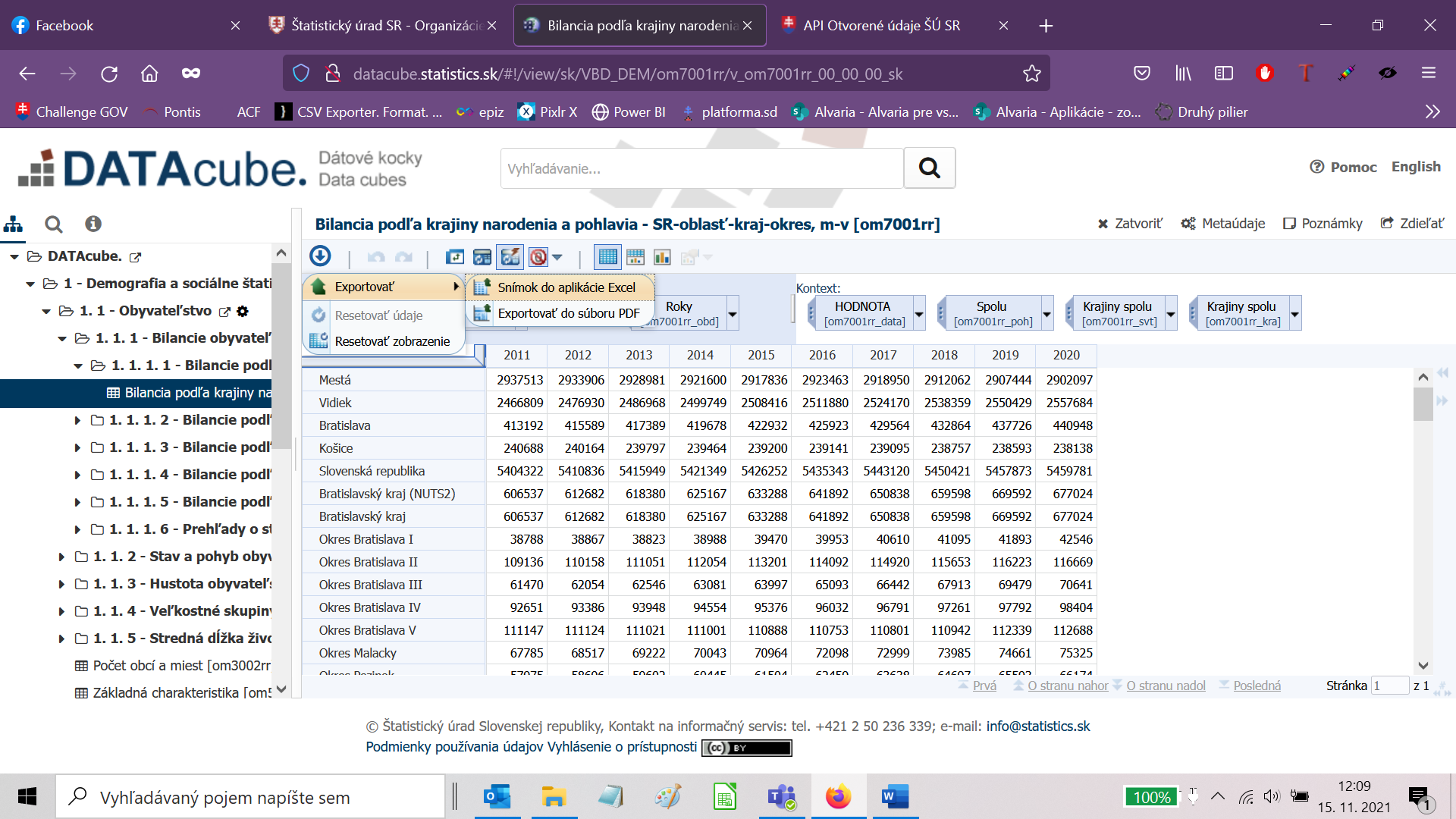
Obsah zverejňovaný pomocou týchto kanálov **nie je rovnaký**, nie všetky datasety sú zverejňované cez všetky kanály, preto odporúčame pri hľadaní datasetu **skontrolovať všetky kanály**.

**Data.gov.sk**

Na centrálnom portáli otvorených dát si používateľ môže stiahnuť priamo datasety v rôznych formátoch vrátane CSV a XLSX.

**DataCube**

Pomocou DataCube je možné stiahnuť datasety vo formáte XLSX. Dataset je potrebné nájsť alebo nalistovať v ľavom menu, otvoriť ho a údaje je možné stiahnuť cez voľby **Exportovať ­­> Snímok do aplikácie Excel**. Je potrebné si dať pozor či sa exportuje len aktuálna snímka alebo všetky.



**API**

API Štatistického úradu **umožňuje vygenerovať si z datasetov a rôznych dimenzií odkaz** (napr. rok, obec, kraj a ďalšie), pomocou ktorého je možné stiahnuť **stále aktuálne** dáta. Používateľ musí zvoliť dataset a ďalej vyklikať, o ktoré údaje má záujem. Následne klikne na tlačidlo **Generovať**, ktoré **vygeneruje odkaz**. Po kliknutí na odkaz sa stiahne dataset podľa zvolených údajov a formáte. Podporované formáty sú JSON, CSV, XML, XLSX a ODS.

Obrázok, na ktorom je text, snímka obrazovky, monitor

Automaticky generovaný popis

Stiahnutý dataset **nie je však jednoducho použiteľný (napr. v nástrojoch na vizualizáciu dát)**, lebo obsahuje rôzne stránky a dáta označené pomocou kódov (dáta sú uložené vo formáte json-stat čo je špeciálny formát pre viacdimenzionálne štatistické dáta).

Pre **zjednodušenie spracovania** je možné dataset **prekódovať pomocou služby** [**https://json-stat.org/format/exporter/**](https://json-stat.org/format/exporter/) . Najskôr je nutné stiahnuť dataset vo formáte json a ten je následne exportérom uložený ako tabuľkové csv ,s ktorým sa lepšie pracuje.

Obrázok, na ktorom je stôl

Automaticky generovaný popis

# Spracovanie otvorených dát mesta Prešov

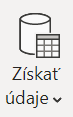
Mesto Prešov zverejňuje otvorené dáta [na svojom webe](https://egov.presov.sk/Default.aspx?NavigationState=1100:0:). Dáta sú zverejnené vo formátoch:

* XML, JSON, XML.net
* SHP, KML, GEO JSON (geo dáta)

Dáta je možné použiť aj v **nástrojoch na spracovanie otvorených dát**, napr. Microsoft PowerBI.

**Bežné datasety**

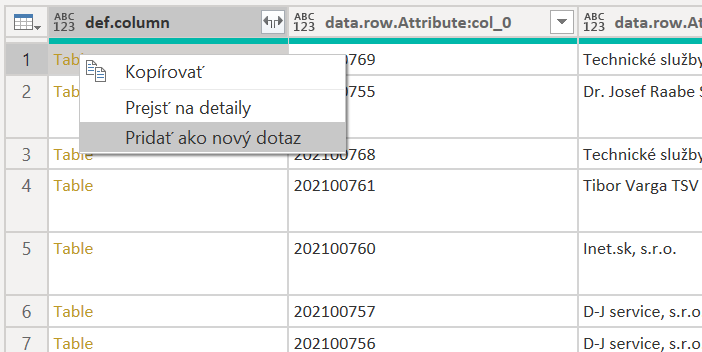
Pre prácu v PowerBI odporúčame stiahnuť formát XML. Ten je potom možné v PowerQuery ďalej spracovať.

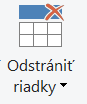
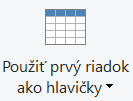
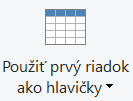
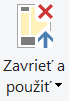
1. Klikneme na tlačidlo **Získať údaje** .
2. Nalistujeme voľbu XML.
3. Klikneme na **Pripojiť**.
4. Nalistujeme dataset.
5. Vyberieme zdroj - „rs“.
6. Klikneme na **Načítať**.
7. V pravom paneli stĺcov klikneme pravým tlačidlom na názov datasetu a zvolíme **Upraviť dotaz**.
8. V PowerQuery rozbalíme stĺpce **okrem** **def.column** kliknutím na ikonu .

Obrázok, na ktorom je stôl

Automaticky generovaný popis

1. Klikneme pravým tlačidlom myši na názov **názov Table sĺpca def.column** a zvolíme **Pridať ako nový dotaz**.



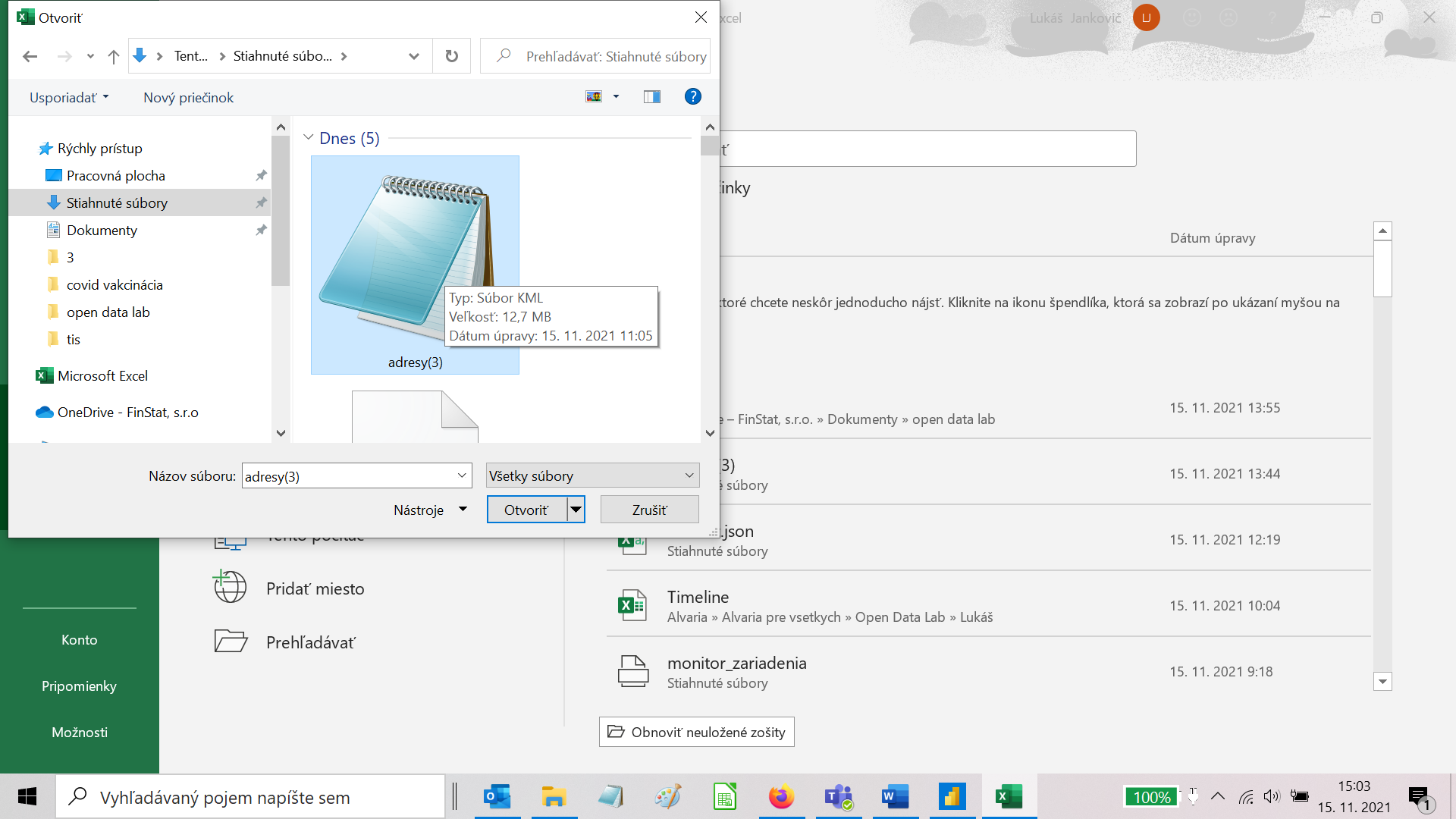
1. V novom dotaze, **pri stĺpci Attribute:col\_id** **nahradíme text „col“ za „Attribute:col“**
2. Transponujeme nový dotaz voľbou **Transformovať > Transponovať **.
3. Voľbou **Odstrániť riadky > Odstrániť spodné riadky ** odstránime **4** spodné riadky.
4. Zoradíme riadky vzostupne tlačidlom .
5. Použijeme voľbu **Transformovať > Použiť prvý riadok ako hlavičky **.
6. Vrátime sa pomocou menu vľavo na pôvodný dataset a vyberieme pomocou tlačidla v ľavom hornom rohu  **Pripojiť dotazy** novovytvorený dataset.
7. Odstránime stĺpec **def**.
8. Zvolíme **Transformovať > Obrátiť poradie riadkov **.
9. Použijeme voľbu **Transformovať > Použiť prvý riadok ako hlavičky **.
10. Odstránime prebytočné stĺpce.
11. Klikneme na tlačidlo **Zavrieť a použiť** .

**Geo dáta**

Datasety **obsahujú geografickú** informáciu buď vo forme **bodov** (adresy, školy, atď.) alebo **kriviek** (cesty, plochy, atď.).

**Goegrafické informácie o bodoch je možné spracovať v PowerBi**, musíme však predtým dáta **previesť do formátu, s ktorým vie pracovať PowerBI** (**XLSX alebo CSV**).

1. **V Microsoft Excel (alebo obdobnom software) zvolíme Otvoriť a Prehľadávať**, zvolíme všetky súbory a nájdeme KML súbor s geografickými informáciami a klikneme na tlačidlo **Otvoriť**.



1. Potvrdíme všetky okná kliknutím na tlačidlo OK.
2. **Klikneme na roh tabuľky**, vyberie sa nám celá tabuľka.
3. **Klikneme pravým tlačidlo myši na ľubovoľné číslo riadku a zvolíme Výška riadku**, kde vyplníme napr. hodnotu **12**, klikneme na tlačidlo **OK**.
4. Vytvoríme si nový hárok.
5. Vyberieme si v stĺpci **name** pomocou filtra názov a zvolíme tlačidlo OK.

Obrázok, na ktorom je stôl

Automaticky generovaný popis

1. Skopírujeme stĺpec **ns1:value** tak, že na **jeho názov klikneme myšou a stlačíme CTRL+C.**
2. **Vložíme stĺpec do nového hárku** tak, že na písmeno prázdneho stĺpca klikneme pravým tlačidlom myši a zvolíme **Hodnoty.**

Obrázok, na ktorom je stôl

Automaticky generovaný popis

1. **Pomenujeme stĺpec** tým, že prepíšeme prvú hodnotu v ňom.

Obrázok, na ktorom je stôl

Automaticky generovaný popis

1. Rovnakým postupom nakopírujeme požadované stĺpce do toho istého nového hárku.
2. Ak chceme pridať GPS súradnice, vyberieme si ľubovoľný názov v kroku č. 6. Následne v kroku č. 7 použijeme stĺpec **ns1:coordinates.** Vložíme stĺpec do toho istého nového hárku (krok č. 8). Následne nový stĺpec pomenujeme. Vyberieme stĺpec s GPS súradnicami (klikneme na písmeno v jeho záhlaví) a zvolíme **Údaje > Text na stĺpce** Obrázok, na ktorom je stôl

   Automaticky generovaný popis. **Vyberieme Oddelené a ako oddeľovač zvolíme čiarku**. Klikneme na tlačidlo **Dokončiť**. Pomenujeme stĺpce.
3. **Uložíme údaje ako XLSX súbor** a importujeme do PowerBI. Pri importe nezabudnime zvoliť novovytvorený hárok.

Obrázok, na ktorom je stôl

Automaticky generovaný popis

Obrázok, na ktorom je stôl

Automaticky generovaný popis

Obrázok, na ktorom je stôl

Automaticky generovaný popis

# Ako urobiť projekt nad otvorenými dátami čo najlepší

Tieto rady Vám dúfam pomôžu pri dokončovaní projektu. Je to zoznam dôležitých vecí, na ktoré nezabudnite, keď budete projekt dokončovať a ktoré môžu Váš projekt posunúť smerom ku kvalitnejšej vizualizácii otvorených dát.

**Prehľadnosť**

Vizualizácia má byť v prvom rade **prehľadná**. Znamená to dobrú zrozumiteľnosť a pochopiteľnosť. **Prvkov (grafy, mapy, filtre) by** **nemalo byť priveľa (môžeme ich rozdeliť na stránky)**, mali by byť **rovnomerne rozmiestnené** a mali by mať **medzi sebou súvis**. Takisto sa vyhnite chybám, ktoré by mohli skresliť grafy, napr. chybná mierka alebo nedodržiavanie konvencií (viac častých chýb nájdete [v tomto článku](https://thenextweb.com/news/7-most-common-data-visualization-mistakes)). Taktiež **formátovanie by malo byť zjednotené**, mali by sa používať rovnaké farby, písma, veľkosti písma. Na vizualizáciu by malo byť príjemné sa pozerať.

**Pomocné texty slúžia k dovysvetleniu pojmov**, prípadne toho, čo graf alebo mapa hovorí. **Nebojte sa ich používať. Prvky by mali byť nadpisy a grafy názvy ôs** (nie je nutné ak je prítomná legenda alebo ide o roky**), ak je na stránke viac prvkov mala by tiež mať nadpis.**

**Na úvodnej stránke uveďte zdroj údajov spolu s odkazom, meno autora, kontakt na autora.** Tieto údaje sú užitočné aby Vás mohol používateľ kontaktovať alebo Vám mohol položiť prípadné otázky k vizualizácii. V prípade ak chcete aby bola Vaša vizualizácia čo najviac využívaná, je vhodné uviesť licenciu, napríklad z rodiny Creative Commons, ktorá používateľom povie, čo môžu s vizualizáciou robiť.

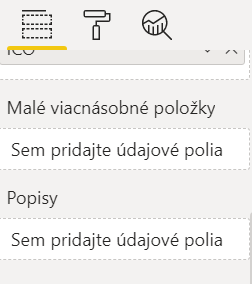
**Interaktivita**

Vizualizácia v nástrojoch nám umožňuje pridať filtre alebo iné možnosti interaktivity. Preto **je žiadúce, aby si používateľ mohol vizualizáciu prispôsobiť podľa seba**. Na jednej stránke môže byť aj viacero filtrov, ktoré môžu byť rôzneho typu (napr. vybratie hodnôt alebo rozsah). Vždy však platí, že daný filter by mal byť **pre používateľa užitočný** a **mal by** **mať logické opodstatnenie**.

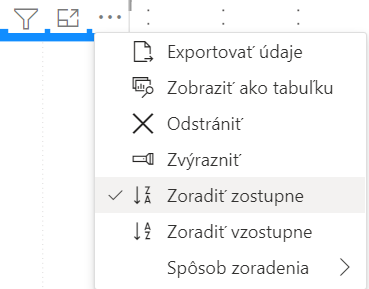
Takisto **kliknutím na časť grafu je možné filtrovať údaje**. Toto je možné využívať napríklad pridaním tabuľky a grafu, pričom v závislosti od kliknutia na graf sa môže meniť aj tabuľka.

Nie **všetky filtre musia interagovať so všetkými prvkami na stránke**, toto je možné nastaviť. Vždy by však malo byť zrejmé napr. podľa rozmiestnenia, ktorý filter interaguje s ktorými prvkami.

Vhodným doplnkom ku údajom zobrazeným v prvku je **textový popis – bublina**, ktorá sa **zobrazí pri ukázaní na časť grafu alebo bod na mape**. Tu je vhodné umiestniť podrobnejšie údaje. V PowerBI sú tieto údaje v poli **Popisy.**



Nezabudnite tiež na **zoradenie údajov**, buď podľa najvyššej hodnoty po najnižšiu alebo podľa abecedy.



**Spracovanie údajov**

Spracovanie údajov je to prvé, čím by sme sa mali pri tvorbe vizualizácie zaoberať. **Je dôležité skontrolovať správnosť údajov, napríklad tým že si porovnáme údaje priamo v datasete s tým čo vidíme vo vizualizácii. Vhodné je skontrolovať aj vlastné vypočítané stĺpce.**

Pokiaľ niektoré údaje v datasete sú zjavne chybné, **môžeme ich vylúčiť z vizualizácie, pokiaľ ich vieme identifikovať. Pokiaľ je dataset príliš rozsiahly, môžeme informáciu pridať do popisu aplikácie** (údaje neboli čistené), aby bolo zrejmé že chyby nevznikli v našej vizualizácii alebo boli prítomné v datasete.

**Dajte si tiež pozor na správne formáty.** Napríklad PSČ a IČO by mali byť zobrazené ako text (kvôli nulám na začiatku), sumy by mali obsahovať oddeľovač tisícov a menu.

Takisto pokiaľ **boli prepájané** 2 alebo viac datasetov, **je vhodné si námatkovo skontrolovať správnosť údajov**.

Pokiaľ sme **údaje čistili**, **je to vhodné uviesť do popisu vizualizácie**.