FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

Západočeská univerzita v Plzni

Databázové systémy a metody zprac.inf.2

kiv/DBM2

Dokumentace semestrální práce

Vojtěch Daniš́ík
A19N0028P
danisik@students.zcu.cz

13. prosince 2021

Obsah

1 Zadání

Již několik let je oblíbeným buzzwordem Open Data, což ve své podstatě jsou data poskytnutá veřejnosti pro libovolné další zpracování. Na našem území je nejčastěji s tímto pojmem spojován portál https://data.gov.cz/, který slouží jako katalog datasetů veřejné správy. Problém je, že momentálně katalog obsahuje 135 510 záznamů a je obtížné udělat si přehled o tom, jaké informace lze vůbec očekávat, že by se zde mohly objevit.

- Vyberte si podmnožinu dat na Portálu otevřených dat, případně jiný zdroj.

- Popište makroskopicky témata, kterých se datasety týkají.

- Vybranou podmnožinu zkoumejte detailně a popište jejich kvalitu* a použitelnost**.

* např. úplnost, smysluplné hodnoty, homogenní datový typ v rámci atributu, použití číselníků místo volného textu (problematické by byla existence výrazů "pozit.", "pozitivní", "kladné"ve stejném kontextu), ...

** např. definice atributů, existence cross-referencí na jiné datasety, použití číselníků (kódová značka pro město místo jeho názvu), ...

Data: https://data.gov.cz/, nebo jiné dle uvážení.

2 Analýza datasetu

Z portálu https://data.gov.cz jsem vybral dataset s názvem: **Počty studentů krajských vyšších odborných škol - Královéhradecký kraj**.

Informace o datasetu:

- velikost souboru: 132 kB
- počet sloupců: 39
- počet řádků: 326
- rok: Školní rok 2019-2020

Sloupce v datasetu by se dali rozdělit do 5 informativních bloků:

- adresa školy (obec, kód obce, ulice, ...)
- právní informace (např. IČO, IZO, REDIZO, ...)
- informace o studiích (název školy, počet studentů, název oboru, ...)
- zeměpisné informace (souřadnice)
- dodatečné informace (id školy v rámci databáze)

2.1 Popis sloupců - adresa školy

todo

2.2 Popis sloupců - právní informace

V datasetu jsou zmíněny právní informace o škole, a to především jejich *IČO*, *IZO* a *REDIZO*.
IČO znamená identifikační číslo osoby, což je osmímísné identifikační číslo právnické osoby. V datasetu se nachází IČO jako čtyřmístné až osmimístné. To může být zapříčiněno oříznutím všech nul nazačátku (může být takto uloženo v databázi nebo bylo upraveno při exportu) .V datasetu se nachází 54 odlišných IČO a nachází se ve všech záznamech.

IZO znamená identifikační znak organizace, který má formu devítimístného čísla a začíná číslicí 0, 1 nebo 2. IZO může být pro každou část školy

různé (pokud jsou pod jednou hlavičkou dohromady střední škola, vyšší odborná škola, ...). V případě otevření datasetu pomocí excelu / openoffice / power BI se může IZO vyskytovat jako např. čtyřmístné číslo. To je zapříčiněno oříznutím všech nul, které se vyskytují nazačátku čísla. V některým případech může IZO být stejné jako IČO, jako v případě našeho datasetu. V datasetu se nachází 54 odlišných hodnot, které jsou buď stejné jako IČO, nebo začínají číslem 0 nebo 1 a mají 9 číslic.

REDIZO je resortní identifikátor právnické osoby, které je taktéž devitimístné číslo, ale zpravidla začínají číslicí 6 a je unikátní pro školu. V datasetu se nachází 54 odlišných hodnot, které splňují formát.

2.3 Popis sloupců - informace o studiích

V rámci tohoto bloku se popíše informace o oborech - název, typ, délka vzdělávání, počet studujících studentů, ...

První sloupec je *obor_nazev*, který značí název oboru na dané škole. Celkem se v datasetu vyskytuje 123 oborů.

Sloupec *obor_kod* označuje identifikační kód oboru, který se skládá z kombinace čísel a písmen a jeho délka je 7 znaků (ve zjednodušeném formátu bez pomlček a lomítka). Pro jeden obor může existovat více identifikačních kódu, které zároveň určují i druh vzdělávání (např. gastronomie s odborným výcvikem a maturitou, gastronomie s vyučením i maturitou - nástavbové studium). Celkem se v datasetu vyskytuje 129 oborových kódů, což značí, že je zde 6 oborů, které mají rozdílný druh vzdělávání. Zároveň jsou zde špatně zadány 4 kódy - obsahují dodatečnou mezeru za posledním znakem.
Formát oborových kódů: https://www.vyberskoly.cz/jak-se-vyznat-v-kodech-oboru

Sloupec *vzdelavani_delka_roky* určuje délku studia oboru v počtu let. V datasetu existují obory, které lze studovat 1 rok (zkrácená studia nebo střední vzdělávání bez výučního listu a maturity), ale i osmiletá gymnázia. Žádná škola neposkytuje obor, který trvá 7 let.

Každý obor má i svůj druh vzdělávání, který je v datasetu reprezentován sloupcem *vzdelavani_druh*. Druh vzdělávání určuje výstup z daného oboru - maturita, výuční list, ... Celkem je zde 6 druhů vzdělávání:

- nástavbové studium - pro absolventy s výučním listem

- střední vzdělávání
- střední vzdělávání s maturitní zkouškou
- střední vzdělávání s výučním listem
- zkrácené studium k získání středního vzdělání s maturitní zkouškou
- zkrácené studium k získání středního vzdělání s výučním listem

Další sloupec je *vzdelavani_forma*, který určuje formu vzdělávání oboru. V datasetu existují pouze 2 formy - denní a dálková. Dálková forma se ve většině případů vyskytuje hlavně u nástavbového studia.

Následující sloupce s názvem *pocet_studentu_1_rocnik* až *pocet_studentu_8_rocnik* vyjadřují počty studentů v ročnících.

Sloupec *pocet_studentu_celkem* vyjadřuje celkový počet studentů ve všech ročnících na daném oboru.

Sloupec *pocet_absolventu_2019_2020* vyjadřuje počet absolventů daného oboru za školní rok 2019-2020.

Poslední sloupec *pocet_prijatych_studentu* vyjadřuje celkový počet přijatých žáků do prvních ročníků daného oboru za tento rok.

2.4 Popis sloupců - zeměpisné informace

V datasetu se nachází sloupce popisující zeměpisné informace školy. Sloupce *wkt*, x a y určují pozici školy na mapě.
Wkt je well-known text formát, který obsahuje souřadnice y,x a určuje pozici objektu na mapě. Tento sloupec nemá žádný informativní význam společně se sloupci x a y (adresa školy je detailně popsána, není potřeba ukládat souřadnice, alespoň ne v datasetu zkoumající počty studentů).

2.5 Popis sloupců - dodatečné informace

V datasetu se nachází sloupec s názvem *OBJECTID*, které obsahuje jednotlivé číselné identifikátory oborů na školách v rámci databáze. Tento sloupec nemá žádný informativní význam.

Zárověň se zde nachází sloupec *dp_ id*, u kterého nelze zjistit jakou nese informaci. Hodnoty typu 'ZOSS1' jsou nic neříkající a je zřejmé, že tento atribut nemá žádný informativní význam vzhledem k účelu tohoto datasetu.

3 Použitelnost datasetu

todo

Atributy datasetu jsou rozděleny do 3 bloků na základě použitelnosti: významné, Nice-to-have a nepoužitelné
Významné atributy:

- 1

Nice-to-have atributy:

- *IČO* - rozhodně není potřeba ukládat do datasetu, ale pro některé lidi to může být zajímavá informace, díky které můžou o škole zjistit více informací - např. právní informace nebo více informací k oborům (za jde obor dobíhající, kapacita oboru, ...)
- *IZO* - stejné jako pro IČO
- *REDIZO* - stejné jako pro IČO

Nepoužitelné atributy:

- *OBJECTID* - id oboru v interní databázi není potřeba ukládat do datasetů
- *wkt* - není potřeba ukládat zeměpisné informace o škole u datasetů zobrazující počty studentů na oborech, navíc v datasetu je uložena celá adresa školy (i jejich pracovišť)
- *x* - stejné jako pro wkt
- *y* - stejné jako pro wkt
- *dp_ id* - neidentifikovaltený atribut, který ale podle mého stejně nebude mít důležitou informaci k počtům studentů

4 Závěr

todo

počet škol, oborů