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

15. 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

Pro popis adresy školy existuje překvapivě až moc sloupců vzhledem k povaze datasetu, který má především sdělit počet studentů na daných oborech v daných školách.

Sloupec *nazev* vyjadřuje název školy. Zároveň je zde v některých případech zaznamenána i celá adresa školy s číslem popisným. Celkově je zde zaznamenáno 54 škol. Jak již bylo zmíněno, v některých případech je v názvu školy zaznamenána i adresa školy (ať už celá, částečná nebo že se jedná o příspěvkovou organizaci), což vzhledem k nedodržení formátu pro všechny školy je zbytečná informace a postačil by pouhý název školy, protože v dalších sloupcích jsou uvedeny údaje o lokalitě školy.

Sloupec *nazev_okresu* obsahuje informaci o názvu okresu, ve kterém se škola nachází. Celkem se v datasetu nachází 5 okresů.

Následující sloupec *kod_okresu* obsahuje kódy jednotlivých okresů, díky kterým lze identifikovat okres v daném kraji. Formát kódu okresu je čtyřmístné číslo (RÚIÁN kód, viz **odkaz**). Všechny zadané kódy okresů jsou validní a odkazují na správný název okresu.

Další sloupec je *nazev_obce*, ve kterém se definuje název obce, ve které škola sídlí. Celkem se v datasetu nachází 18 obcí.

Stejně jako pro okres se zde nachází sloupec s kódem obce - *kod_obce*. Kód obce je šestimístné číslo (RÚIÁN kód, viz **odkaz**). Všechny zadané kódy obcí jsou validní a odkazují na správný název obce.

Sloupec *nazev_ulice* obsahuje název ulice, ve které se škola nachází.

Sloupec *cislo_domovni* obsahuje domovní číslo školy.

Sloupec *typ_cisla_domovniho* definuje typ domovního čísla (číslo popisné, číslo evidenční). V datasetu se nachází pouze školy, které mají pouze číslo popisné.

Sloupec *cislo_orientacni* obsahuje číslo orientační pro danou školu. Pouze 11 škol má i číslo orientační.

Následující sloupce obsahují celou adresu pracoviště školy. Pracoviště nemusí mít totožnou adresu jako je hlavní adresa. Tyto sloupce není potřeba zmiňovat rozepisovat, protože mají stejný význam jako sloupce *nazev_obce*, *nazev_ulice*, *cislo_domovni*, *typ_cisla_domovniho* a *cislo_orientacni*.

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 ořízntutím 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 devítimístné 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.
Zároveň 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

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

Významné atributy:

- *nazev* - název školy je nutný atribut, je potřeba vědět na jaké škole je daný počet studentů
- *nazev_ okresu / kod_ okresu* - zařazen jako významný, protože některé školy mohou mít totožný název i s jinými školami z jiných okresů (např. **Střední škola služeb, obchodu a gastronomie**
- *nazev_ obce / kod_ obce* - zařazen ze stejného principu jako *nazev_ okresu*
- *obor_ nazev* - povinné
- všechny atributy vyznačující počet studentů v prvním až osmém ročníku *pocet_ studentu_ x_ rocnik* - povinné, počet studentů v x-ročníku
- *pocet_ absolventu_ 2019_ 2020* - velice zajímavá informace, určitě je dobré vědět kolik absolventů prošlo školou za rok 2019-2020 - může ovlivnit názor na studium na dané škole
- *pocet_ prijatych_ studentu* - taktéž zajímavá informace, může reflektovat dnešní zájem studentů o daný obor

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
- *nazev_ ulice* - může se stát, že v rámci jedné obce mohou být dvě školy se stejnými názvy (velice málo pravděpodobné) - proto pouze nice-to-have.
- *cislo_ domovni* - stejné jako u *nazev_ ulice*

- *cislo_orientacni* - stejné jako u *nazev_ulice*
- *obor_kod* - název oboru může být stejný pro více vzdělávacích druhů
- *vzdelavani_delka_roky* - může být zajímá informace pro některé lidi
- *vzdelavani_druh* - může být zajímá informace pro některé lidi
- *vzdelavani_forma* - může být zajímá informace pro některé lidi
- *pocet_studentu_celkem* - nice-to-have informace, lze dopočítat z předchozích atributů *pocet_studentu_x_rocnik*

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
- *typ_cisla_domovniho* - typ čísla domovního bude vždy **číslo popisné**, protože **číslo evidenční** slouží pouze pro stavby určené k rekreaci.
- všechny atributy určující pozici pracoviště - nepoužitelné pro dataset zobrazující počty studentů

4 Závěr

Z analýzy datasetu vyplynulo, že z celkových 39 atributů jich je 16 významných, které slouží pro identifikaci oboru na škole a aktuální počet studujících žáků.

V datasetu je 11 atributů obsahující dodatečné informace o škole a o počtu studentů.

Celkem 12 atributů (což dělá necelou třetinu všech atributů) nemá žádné použitelné informace vzhledem k povaze datasetu - informovat případné nové studenty (nebo i normální obyvatelstvo) o dostupných oborech v okresech ležících v Karlovarském kraji.

Všechny atributy mají homogenní datový typ pro všechny své hodnoty. Atributy *cislo_orientacni* a *cislo_orientacni_pracoviste* obsahovali prázdné hodnoty v +- 70% případech (očekávaná hodnota, většina budov nemá své číslo orientační). Jediné špatně zadané hodnoty byly u sloupce *obor_kod*, kde ve 4 případech byly kódy zapsány s dodatečnou mezerou na konci. Cross-reference na další datasety neexistují.

V datasetu je celkem 129 oborů z různých škol a s ruznými druhy vzdělávání. Celkový počet škol je 54 nacházejících se v 5 okresech a 18 obcích / městech.