Vysoké učení technické v Brně
Fakulta informačních technologií

UPA - Ukládání a příprava dat

Projekt: návrh zpracování a uložení dat

Téma: 04: COVID-19 (dr. Rychlý)

Kapoun Petr, Bc. - xkapou04
Nováček Pavel, Bc. - xnovac16
Willaschek Tomáš, Bc. - xwilla00

7. prosince 2020

Obsah

1 Zvolené téma

04: COVID-19 (dr. Rychlý)

2 Řešitelé

- Bc. Kapoun Petr – xkapou04,
- Bc. Nováček Pavel – xnovac16,
- Bc. Willaschek Tomáš – xwilla00.

3 Zvolené dotazy a formulace vlastního dotazu

- **Dotaz A:** vytvořte popisné charakteristiky pro alespoň 4 údaje (např. věk, pohlaví, okres, zdroj nákazy) z datové sady COVID-19: Přehled osob s prokázanou nákazou dle hlášení krajských hygienických stanic (využijte krabicové grafy, histogramy, atd.).
- **Dotaz B:** určete vliv počtu nemocných a jeho změny v čase na sousední okresy (aneb zjistěte jak se šíří nákaza přes hranice okresů).
- **Vlastní dotaz:** určete vliv věku na délku nemoci a úmrtnost.

4 Stručná charakteristika zvolené datové sady

Základním zdrojem dat pro všechny dotazy jsou **otevřené datové sady COVID-19 v ČR**[4] dostupné skrze veřejné API[1] ve formátech JSON a CSV.

Datová sada `COVID-19: Přehled osob s prokázanou nákazou dle hlášení krajských hygienických stanic (v2)` obsahuje následující charakteristiky o nakažených osobách:

- `datum` – datum, kdy byla osoba pozitivně testována,
- `vek` – věk nakažené osoby,
- `pohlavi` – pohlaví nakažené osoby,
- `kraj_nuts_kod` – identifikátor kraje podle klasifikace NUTS 3, ve kterém byla pozitivní nákaza hlášena krajskou hygienickou stanicí,
- `okres_lau_kod` – identifikátor okresu podle klasifikace LAU 1,
- `nakaza_v_zahranici` – příznak, zda došlo k nákaze mimo ČR,
- `nakaza_zeme_csu_kod` – identifikátor státu v zahraničí, kde došlo k nákaze (dvoumístný kód z číselníku zemí CZEM).

Další dvě použité datové sady jsou `COVID-19: Přehled vyléčených dle hlášení krajských hygienických stanic` a `COVID-19: Přehled úmrtí dle hlášení krajských hygienických stanic`, které mají shodnou strukturu:

- `datum` – datum vyléčení nebo úmrtí osoby,
- `vek`,

[1] `https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19`

- `pohlavi`,
- `kraj_nuts_kod`,
- `okres_lau_kod`.

Jelikož tyto datové sady používají identifikátory *NUTS 3*[2] pro kraje a *LAU 1*[3] pro okresy, využíváme dále číselníky od Českého statistického úřadu, které obsahují mapování těchto identifikátorů na odpovídající kraje a okresy a informace o nich.

Číselník okresů[4] a **číselník krajů**[5] jsou oba ve formátu CSV a obsahují:

- `KODJAZ` – kód jazykové verze textů,
- `AKRCIS` – akronym číselníku/klasifikace,
- `KODCIS` – kód číselníku/klasifikace,
- `CHODNOTA` – kód položky,
- `ZKRTEXT` – zkrácený název položky,
- `TEXT` – plný název položky,
- `ADMPLOD` – počátek administrativní platnosti,
- `ADMNEPO` – konec administrativní platnosti.

Zvoleným programovacím jazykem je `Python`[5]. K získání dat použijeme standardní moduly jazyka `Python`[5] `json`[6] a `csv`[7]. K nahrání dat použijeme modul `elasticsearch`[8].

5 Zvolený způsob uložení uložení surových dat

Pro uložení dat jsme zvolili NoSQL řešení `Elasticsearch`[3][9]. `Elasticsearch` je dokumentově orientovaný vyhledávací engine naprogramovaný v jazyce Java s použitím `Lucene`[1], díky kterému je velmi efektivní při komplexním vyhledávání na velkých objemech dat. Schéma struktur kolekcí odpovídá struktuře uložených dat ve vstupních datových sadách. Index je typu klíč-hodnota a pro každý vstupní soubor bude existovat právě jeden index. Klíč nabývá hodnoty `CHODNOTA` u číselníku okresů a krajů a ve zbývajících indexech se generuje automaticky, protože data neobsahují žádnou unikátní hodnotu.

[2] Nomenklatura územních statistických jednotek; úroveň 3 odpovídá krajům

[3] Local Administrative Units; úroveň 1 odpovídá okresům

[4] Číselník okresů: `http://apl.czso.cz/iSMS/cisexp.jsp?kodcis=109&typdat=0&cisvaz=80007_210&datpohl=20.10.2020&cisjaz=203&format=2&separator=,`

[5] Číselník krajů: `http://apl.czso.cz/iSMS/cisexp.jsp?kodcis=100&typdat=0&cisvaz=80007_885&datpohl=20.10.2020&cisjaz=203&format=2&separator=,`

[6] Dokumentace modulu `json` jazyka `Python`[5]: `https://docs.python.org/3/library/json.html`

[7] Dokumentace modulu `csv` jazyka `Python`[5]: `https://docs.python.org/3/library/csv.html`

[8] Modul `elasticsearch` pro `Python`: `https://pypi.org/project/elasticsearch/`

[9] Dokumentace `Elasticsearch`[3]: `https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html`

6 Implementace

Výsledné řešení obsahuje skripty pro:

- nahrání dat do `NoSQL` databáze,
- doplnění, spojení a transformace dat z `NoSQL` do `SQL`,
- prezentaci výsledku ve webovém rozhraní.

Dále obsahuje `Docker`[2] kontejner, ve kterém jsou nainstalované aplikace databází.

Výsledné řešení neobsahuje inkrementální aktualizaci dat, protože mechanismus pro tuto aktualizaci by jednak byl pomalejší než přeplnění a jednak to není ani možné. Pomalejší by byl z toho důvodu, že datová sada ve formátu `CSV` není vytvořená jako inkrementální, ale data jsou vložena náhodně. Pro inkrementální doplnění dat do `NoSQL` databáze (například po měsících) by bylo potřeba načíst celý soubor, seřadit jej a následně plnit rozdílné měsíce/dny (přičemž plnění do `Elasticsearch` je velmi rychlé).

Inkrementální transformace těchto dat do `SQL` není následně možná z důvodu aplikace heuristiky, která mapuje jednotlivé případy vyléčení a úmrtí k záznamu nákazy. Tuto inkrementální aktualizace znemožňuje fakt, že data jsou do datových souborů přidávána i zpětně, tedy když je osoba označena za vyléčenou a za pár dní přibude podobný případ vyléčení, není jisté, s jakou osobou se má spojit.

6.1 Předání vstupních dat

Pro nahrání dat určených ke zpracování a vyhodnocení aplikací slouží webová stránka, která akceptuje pouze oficiální datové formáty (`json`, `CSV`). Vstupní soubor lze také specifikovat pomocí `URL`. Tento mechanismus zjednodušuje práci se skripty a umožňuje spustit celou aktualizaci z UI.

Mechanismus dále validuje vstupní data a upozorňuje na chyby (například neznámý formát) a zobrazuje stav plnění v konzoli.

Díky tomu, že je implementace plnění dat poměrně obecná, bylo by možné nahrát data například z jiného státu a zobrazit si pro něj výsledky.

6.2 Rozšíření datové sady

Pro druhý dotaz byla potřeba informace o sousednosti okresů. Pro tento účel vznikl další soubor ve formátu `CSV`. Obsahuje dva názvu okresů, které spolu sousedí:

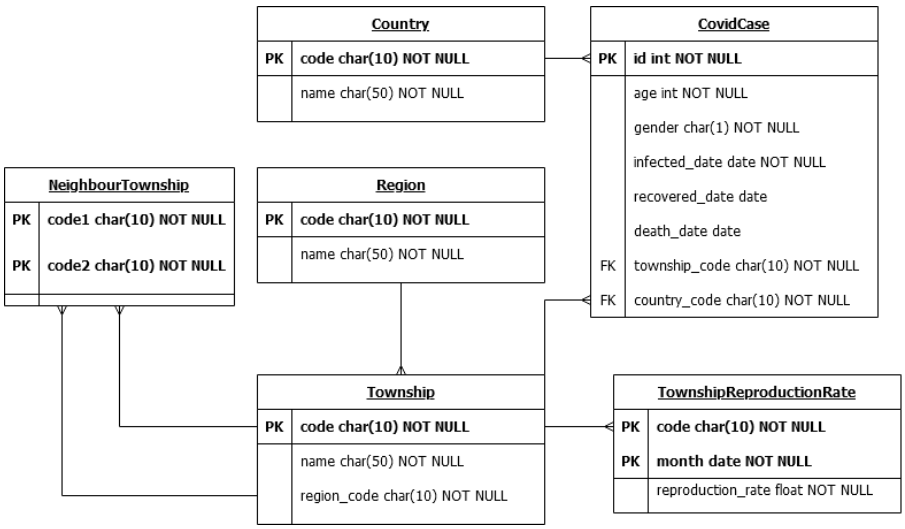
- `HODNOTA1` – název okresu,
- `HODNOTA2` – název okresu.

6.3 NoSQL databáze

Do `NoSQL` databáze `Elasticsearch` jsou nahrány všechny záznamy ze vstupních souborů za použití funkcí knihovny `Pandas` (v případě `CSV` souborů), nebo načtení a vložení vstupního souboru `JSON`.

6.4 SQL databáze

Vytvořená `SQL` databáze má následující schéma:



Postup plnění:

1. vytvoření `SQL` databáze podle definovaného modelu,
2. vyčítání dat z `Elasticsearch`,
3. aplikování heuristiky pro propojení případů vyléčení/úmrtí se záznamem onemocnění,
4. uložení dat do `MySQL`.

Heuristika pro spojení záznamů:

1. Iteruj přes všechny okresy.

2. Vezmi všechny záznamy nakažených pro aktuální okres a seřaď je vzestupně podle data nakažení.

3. Vezmi všechny záznamy uzdravených pro aktuální okres z `NoSQL` databáze a seřaď je vzestupně podle data uzdravení.

4. Vezmi všechny záznamy zemřelých pro aktuální okres z `NoSQL` databáze a seřaď je podle data úmrtí.

5. Iteruj přes všechny získané záznamy nakažených.

6. Pro aktuální záznam nakaženého najdi první odpovídající záznamy (tj. záznamy se stejným věkem a pohlavím) v záznamech uzdravených a zemřelých, pro které platí, že jejich datum uzdravení/úmrtí je mladší, než datum nakažení aktuálního záznamu.

7. Pokud jsou nalezeny záznamy pro obě skupiny, pak:

8. Aktuálnímu záznamu nakaženého přiřaď to datum, které je z nalezených záznamů uzdravení/úmrtí starší, použitý záznam odstraň a pokračuj.

9. Pokud je nalezen pouze jeden záznam, pak:

10. Přiřaď jeho hodnotu data aktuálnímu záznamu nakaženého, odstraň ho a pokračuj.
11. Jinak pokračuj dalším záznamem nakaženého.

7 Prezentace výsledků

Pro prezentaci výsledků byl použit framework `Flask`[10], který umožňuje zobrazeni vytvořených výsledků jednoduše ve webovém rozhraní. Výsledky za celé období nahraných dat jsou perzistentní, lze je aktualizovat až nahráním nových dat. Webové rozhraní dále umožňuje si zobrazit statistiky za konkrétní období, avšak toto chvíli trvá, protože databáze obsahuje mnoho záznamů.

7.1 Dotaz A

> Vytvořte popisné charakteristiky pro alespoň 4 údaje (např. věk, pohlaví, okres, zdroj nákazy) z datové sady COVID-19: Přehled osob s prokázanou nákazou dle hlášení krajských hygienických stanic (využijte krabicové grafy, histogramy, atd.).

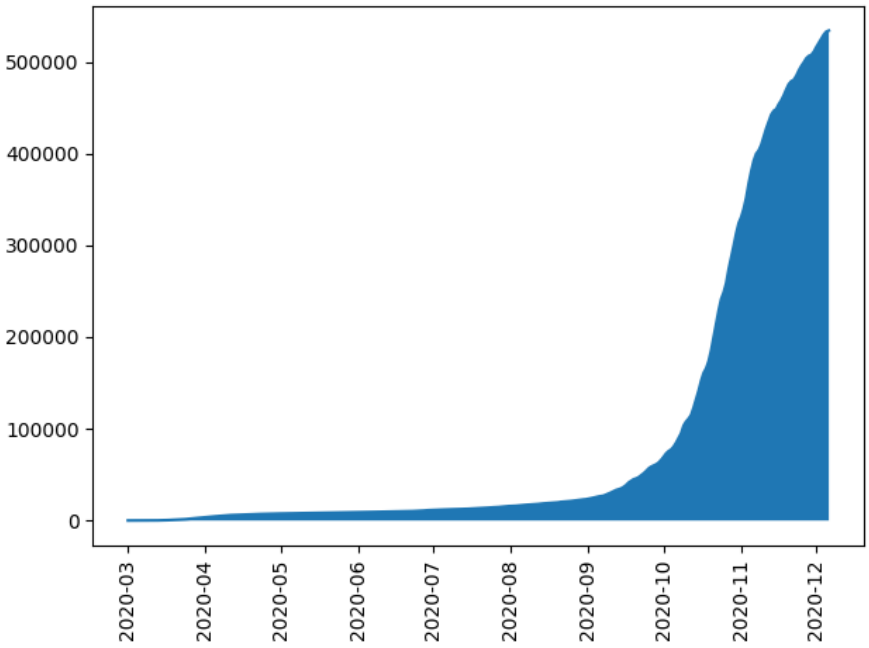
Vytvořili jsme popisné charakteristiky a jejich vývoj v čase pro následující údaje:

- věk,
- pohlaví,
- nákaza,
- úmrtí,
- zdroj nákazy.

[10] Dokumentace modulu `Flask` jazyka `Python`[5]: `https://pypi.org/project/Flask/`

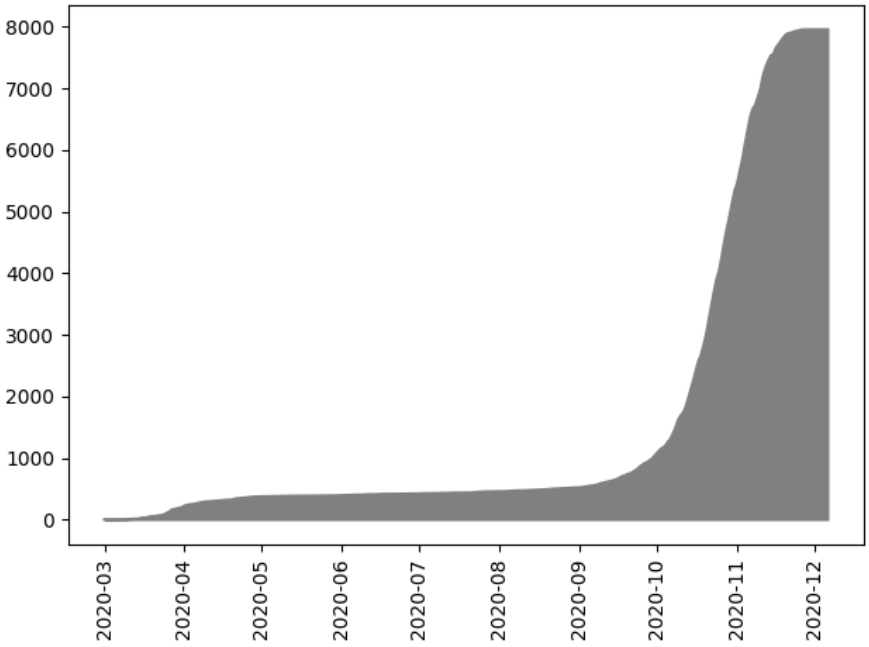
7.1.1 Kumulativní nárůst případů

Jde o vyplněný spojnicový graf, který zobrazuje celkový kumulativní počet případů v čase.



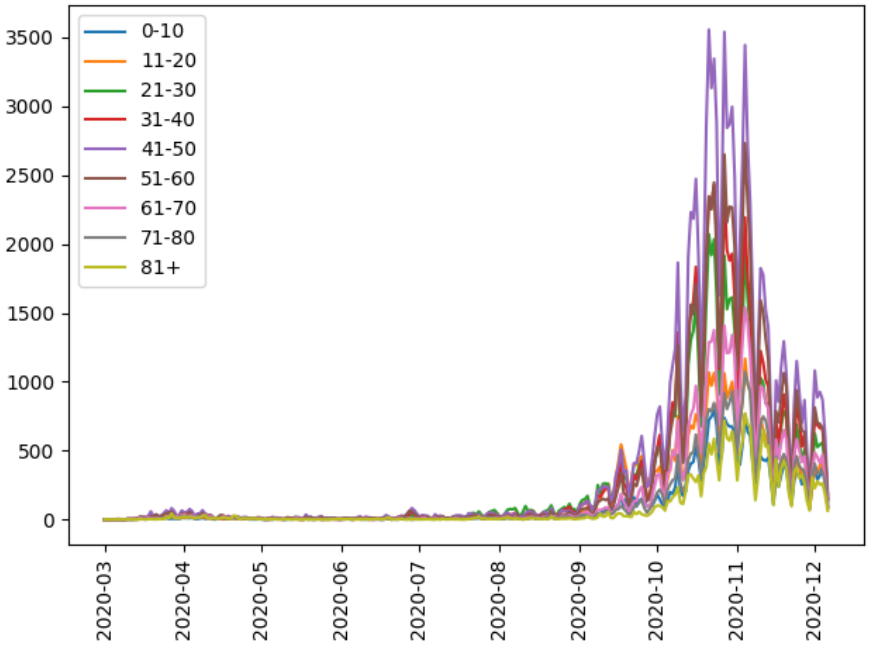
7.1.2 Kumulativní nárůst úmrtí

Jde o vyplněný spojnicový graf, který zobrazuje celkový kumulativní počet případů v čase.



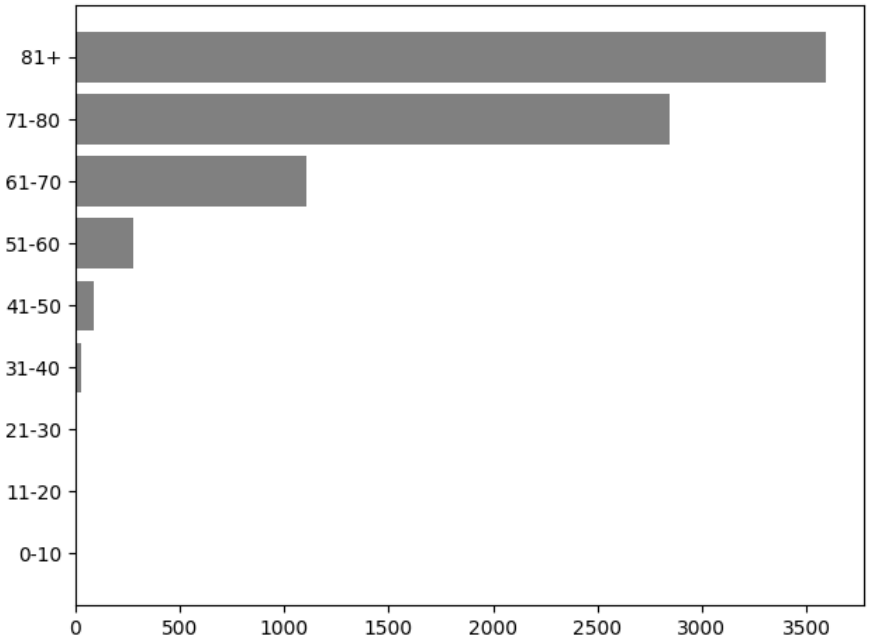
7.1.3 Denní přírůstek podle věkových skupin

Jedná se o spojnicový graf, kde každá spojnice reprezentuje denní přírůstek pro určitou věkovou skupinu. Tento graf je doporučeno zobrazit si na webu pro určité období, protože data jsou čitelnější.



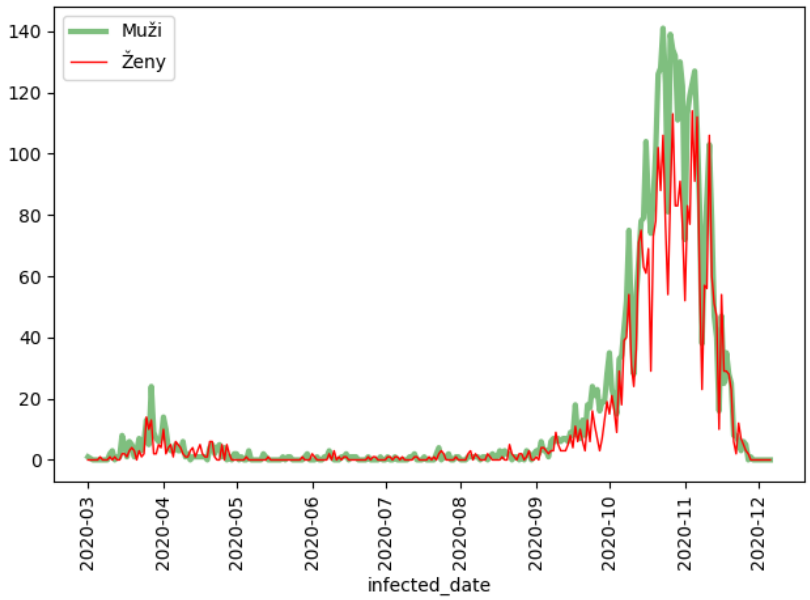
7.1.4 Úmrtnost podle věkových skupin

Jedná se o pruhový graf, který zobrazuje celkový počet obětí, pro danou věkovou skupinu v čase.



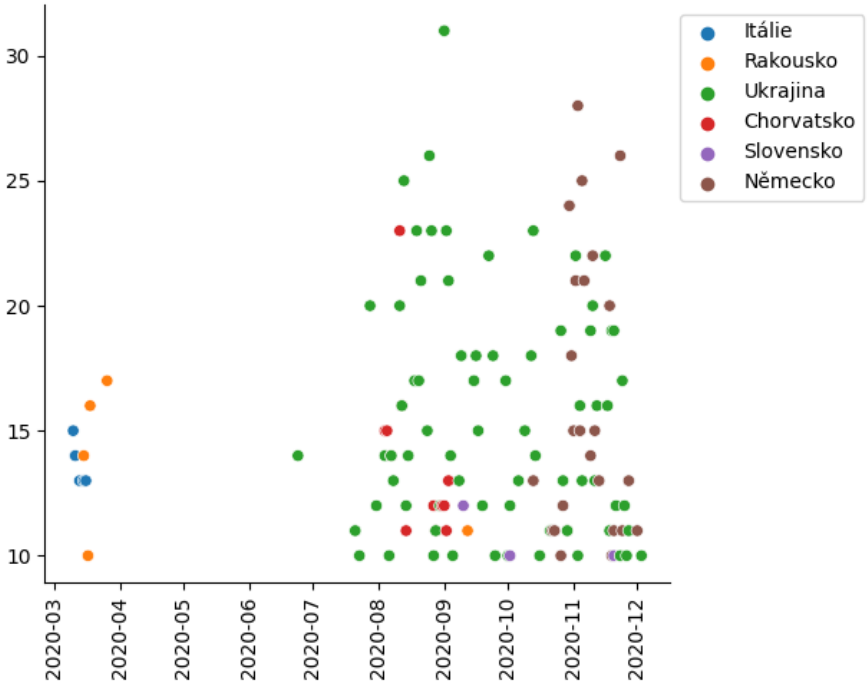
7.1.5 Úmrtnost podle pohlaví

Jde o spojnicový graf, kde každá ze spojnic reprezentuje denní úmrtnost podle pohlaví v čase.



7.1.6 Původ nákazy (pro více než 10 případů)

Jedná se o bodový graf, kde jednotlivé body reprezentují zdroj nákazy českých občanů v zahraničí. Graf pro přehlednost zobrazuje jen státy, kde se v konkrétním dni nakazilo více než 10 osob.



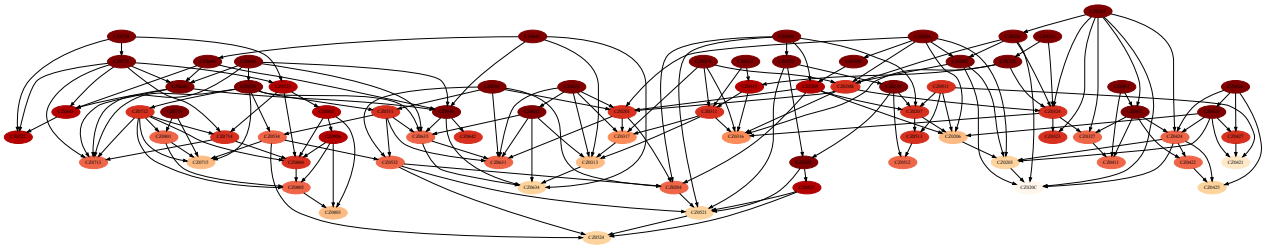
7.2 Dotaz B

> Určete vliv počtu nemocných a jeho změny v čase na sousední okresy (aneb zjistěte jak se šíří nákaza přes hranice okresů).

Dotaz klade důraz na změny v čase, proto zodpovězení dotazu vychází z výpočtu reprodukčního čísla, které vyjadřuje v dotazu právě onu změnu v čase. Výpočet čísla je zjednodušený a jde v podstatě o procentuální změnu přírůstku nakažených.

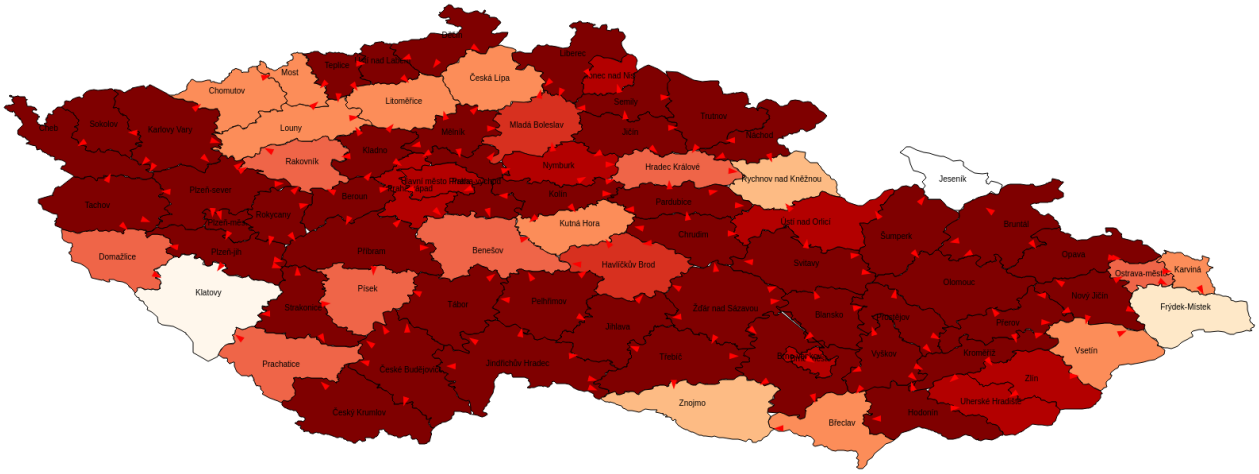
7.2.1 Orientovaný graf

Při znázornění velikosti reprodukčního čísla a vztahu okresů jsme nejdříve použili nástroj graphviz a jeho modul pro Python[5]. Hodnoty použité pro vytvořený orientovaný graf byly použity při tvorbě mapy popsané v 7.2.2. Uzly v grafu znázorňují okresy. Barva uzlů znázorňuje velikost reprodukčního čísla. Hrana vyjadřuje sousednost okresů. Orientace hran grafu je určena porovnáním velikostí reprodukčních čísel dvou okresů. Hrana vychází z uzlu okresu, který má větší hodnotu reprodukčního čísla.



7.2.2 Mapa

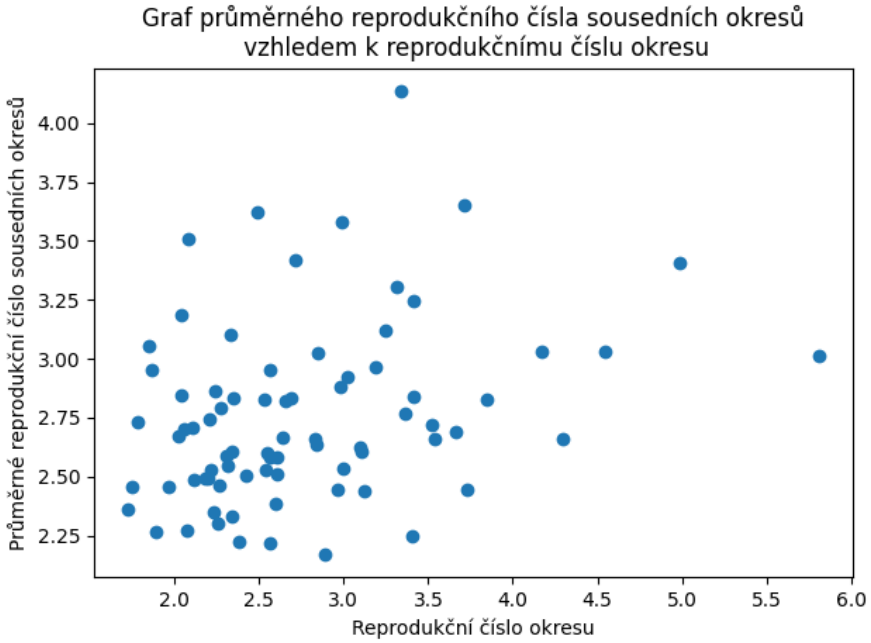
Pro znázornění velikosti reprodukčního čísla a vztahu okresů byla vytvořena mapa, která podobně jako orientovaný graf popsaný v 7.2.1, nastavuje barvu okresu podle reprodukčního čísla. Malé červené trojúhelníky znázorňují šíření nákazy z více nakažených okresů do méně nakažených okresů. Postup nákazy je zobrazen orientací trojúhelníku.



7.2.3 Graf průměrného reprodukčního čísla sousedních okresů vzhledem k reprodukčnímu číslu okresu

Jde o bodový graf, kde pro každý okres je vykreslen jeden bod. Souřadnice `X` vyjadřuje reprodukční číslo okresu. Souřadnice `Y` vyjadřuje průměrné reprodukční číslo okolních okresů.

Z grafu je patrné mírné stoupání hodnot v závislosti na hodnotě osy `X`. Přehledněji vliv sousedních okresů ukazují grafy zobrazené v 7.2.4



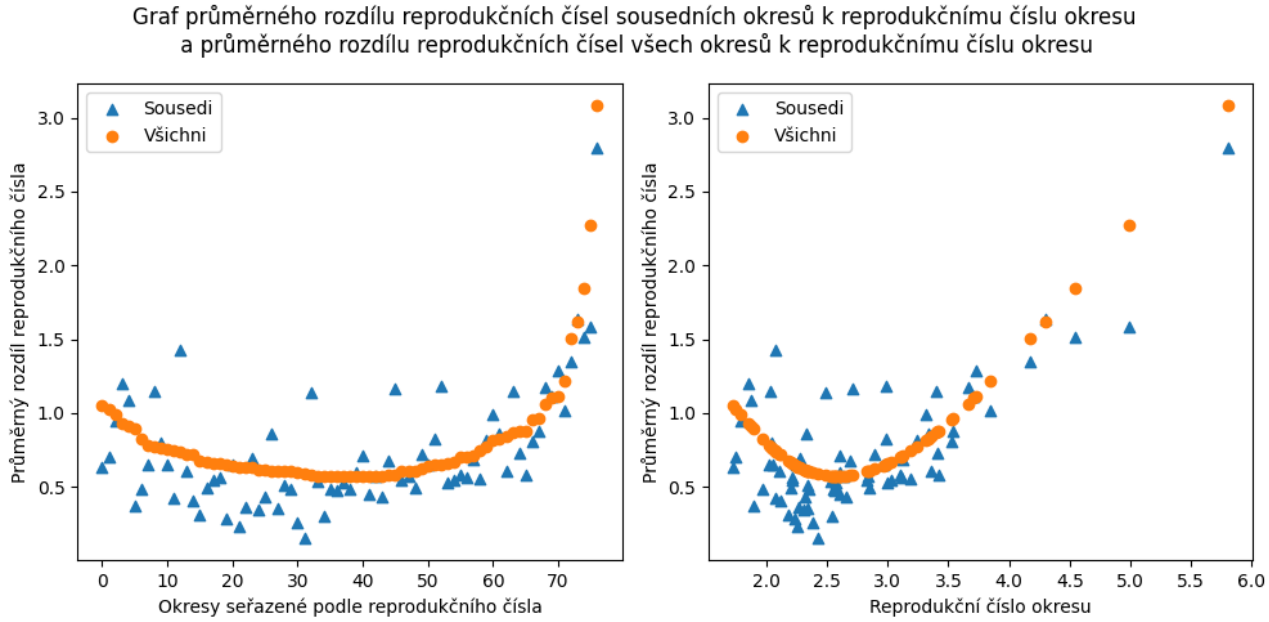
7.2.4 Grafy průměrného rozdílu reprodukčních čísel sousedních okresů k reprodukčnímu číslu okresu a průměrného rozdílu reprodukčních čísel všech okresů k reprodukčnímu číslo okresu

Jde o bodové grafy, kde pro každý okres jsou vykresleny dva body:

- **Oranžové kruhy** - průměrný rozdíl reprodukčních čísel sousedních okresů vzhledem k reprodukčnímu číslu okresu
- **Modré trojúhelníky** - průměrný rozdíl reprodukčních čísel všech okresů vzhledem k reprodukčnímu číslo okresu

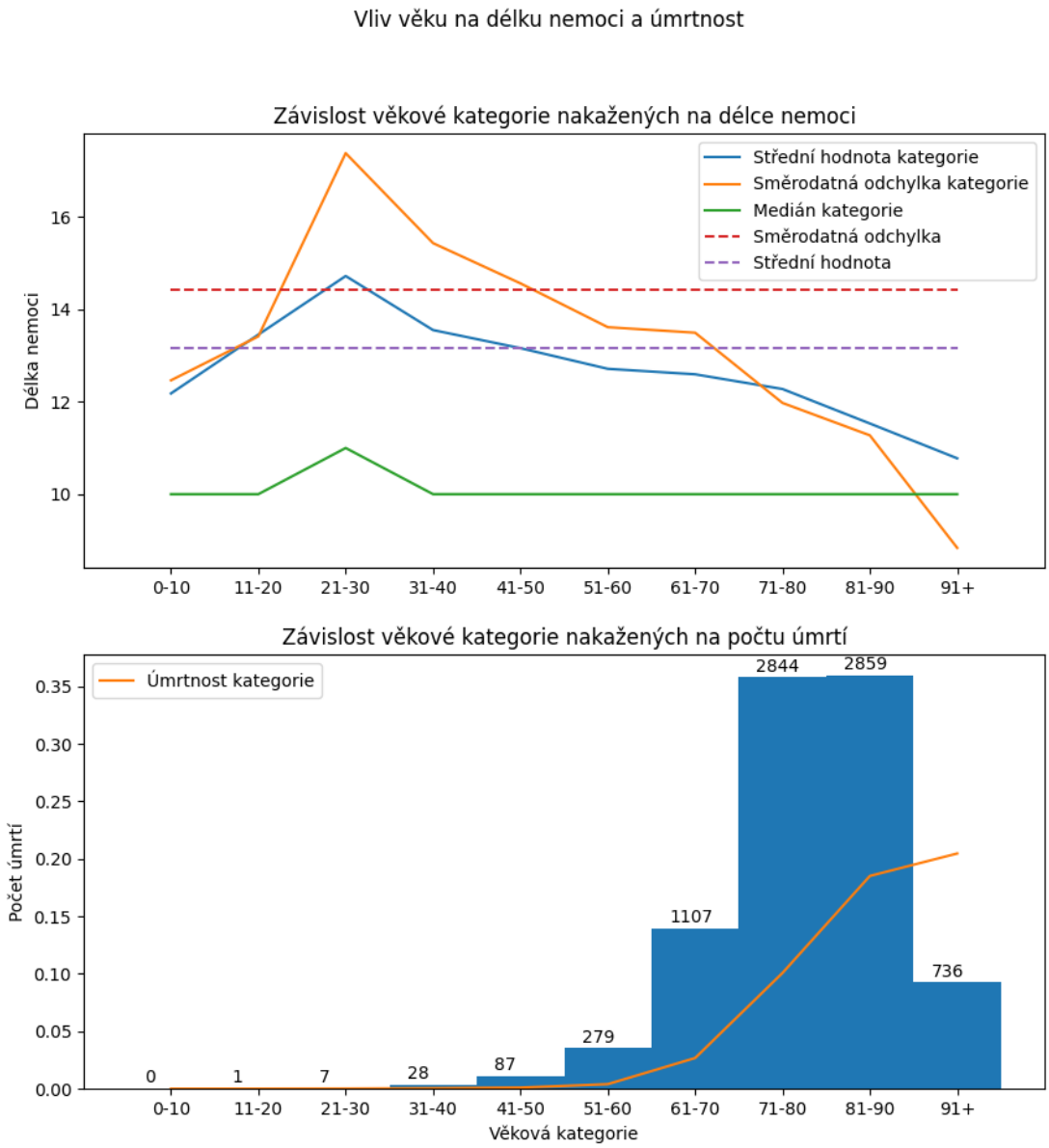
Souřadnice `Y` vyjadřuje průměrný rozdíl reprodukčního čísla vůči reprodukčnímu číslu daného okresu. Tyto dva grafy se liší významem souřadnice `X`. Souřadnice X prvním grafu vyjadřuje pořadové číslo okresu po seřazení okresů dle velikosti reprodukčního čísla a v druhém grafu vyjadřuje reprodukční číslo okresu. Graf s osou `X` znázorňující reprodukční číslo přesněji zobrazuje data, ale protože je okres relativně malé území, tak při menší časové jednotce může při výpočtu vyjít velmi vysoká hodnota (10-12) reprodukčního čísla, při které je naprostá většina hodnot koncentrována na malé ploše, což má za následek nepřehlednost grafu. To je důvod pro vytvoření dvou grafů.

Z grafů je patrné, že rozdíly se sousedními okresy jsou menší než rozdíly se všemi okresy. Takový jev potvrzuje vliv sousedních okresů na nákazu v okrese.



7.3 Vlastní dotaz

Určete vliv věku na délku nemoci a úmrtnost.



8 Návod k použití

Projekt využívá `Docker`[2] a jazyk `Python`[5]. K zajištění závislostí lze spustit:

```
./setup.sh
```

Následující příkaz připraví a spustí docker image:

```
./run.sh -b -esr -sqlr
```

Aplikaci je možné spustit ve dvěma způsoby. Spuštění webové aplikace, kterou lze otevřít v prohlížeči na adrese *localhost*:

```
python run.py --web
```

Spuštění přes příkazovou řádku:

```
python run.py --fill
python run.py --move
python run.py --queries
```

Reference

[1] Apache Software Foundation. *Apache Lucene* [online]. 1999 [cit. 2020-10-20]. Dostupné na: <https://lucene.apache.org/>.

[2] Docker, Inc.. *Docker* [online]. 2013 [cit. 2020-12-7]. Dostupné na: <https://www.docker.com/>.

[3] Elastic NV. *Elasticsearch* [online]. 2010 [cit. 2020-10-19]. Dostupné na: <https://www.elastic.co/>.

[4] Komenda M., Karolyi M., Bulhart V., Žofka J., Brauner T., Hak J., Jarkovský J., Mužík J., Blaha M., Kubát J., Klimeš D., Langhammer P., Daňková Š., Májek O., Bartůňková M., Dušek L.. COVID-19: Přehled aktuální situace v ČR. Onemocnění aktuálně. In [online]. Praha: Ministerstvo zdravotnictví ČR, 2020 [cit. 2020-10-20]. Vývoj: společné pracoviště ÚZIS ČR a IBA LF MU. Dostupné na: <https://onemocneni-aktualne.mzcr.cz/covid-19>. ISSN 2694-9423.

[5] Python Software Foundation. *Python* [online]. 1991 [cit. 2020-10-20]. Dostupné na: <https://www.python.org/>.