VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

COVID-19

UKLADANIE A PRÍPRAVA DÁT (UPA)

Bc. Matej Otčenáš (xotcen01), Bc. Ján Jakub Kubík (xkubik32), Bc. Ján Kačur (xkacur04)

16. decembra 2021

Obsah

1 Zvolené dotazy

- **Dotazy zo skupiny A**
  - **Dotaz 1:** *Vytvořte čárový (spojnicový) graf zobrazující vývoj covidové situace po měsících pomocí následujících hodnot: počet nově nakažených za měsíc, počet nově vyléčených za měsíc, počet nově hospitalizovaných osob za měsíc, počet provedených testů za měsíc. Pokud nebude výsledný graf dobře čitelný, zvažte logaritmické měřítko, nebo rozdělte hodnoty do více grafů.*
  - **Dotaz 2:** *Vytvořte krabicové grafy zobrazující rozložení věku nakažených osob v jednotlivých krajích.*
- **Dotaz zo skupiny B:** *Sestavte 4 žebříčky krajů "best in covid"za poslední 4 čtvrtletí (1 čtvrtletí = 1 žebříček). Jako kritérium volte počet nově nakažených přepočtený na jednoho obyvatele kraje. Pro jedno čtvrtletí zobrazte výsledky také graficky. Graf bude pro každý kraj zobrazovat celkový počet nově nakažených, celkový počet obyvatel a počet nakažených na jednoho obyvatele. Graf můžete zhotovit kombinací dvou grafů do jednoho (jeden sloupcový graf zobrazí první dvě hodnoty a druhý, čárový graf, hodnotu třetí).*
- **Dotaz zo skupiny C:** *Hledání skupin podobných měst z hlediska vývoje covidu a věkového složení obyvatel. Atributy: počet nakažených za poslední 4 čtvrtletí, počet očkovaných za poslední 4 čtvrtletí, počet obyvatel ve věkové skupině 0..14 let, počet obyvatel ve věkové skupině 15 - 59, počet obyvatel nad 59 let. Pro potřeby projektu vyberte libovolně 50 měst, pro které najdete potřebné hodnoty (můžete např. využít nějaký žebříček 50 nejlidnatějších měst v ČR).*
- **Vlasté dotazy**
  - **Vlastný dotaz 1:** Na základe vybraného datasetu o počte hospitalizovaných s ohľadom na vykázané očkovania budú zostrojené spojnicové a stĺpcové grafy s dennou incidenciou, ktoré budú zobrazovať závislosti medzi počtami hospitalizovaných bez očkovania, s očkovaním po prvej a druhej dávke. Podobne budú vytvorené grafy pre počty úmrtí pre ľudí, ktorí neboli očkovaní alebo boli očkovaní jednou či dvomi dávkami. Následne bude prevedená analýza dennej incidencie počtu nakazených, ktorí neboli očkovaní, alebo boli očkovaní jednou či dvomi dávkami. Z týchto štatistík bude vytvorený záver, za akých okolností je väčšia šanca hospitalizácie či úmrtia na ochorenie COVID-19.
  - **Vlastný dotaz 2:** Budem skúmať vývin pozitivity testov a jej závislosť na počte testov na obyvateľa. Analýza bude obsahovať grafy s vývinom pozitivity v čase, ako aj hľadanie korelácie medzi počtom testov a pozitivitou. Tieto parametre

budú skúmané ako celoštátne, tak aj na úrovni okresov. Z okresov potom vyberiem zopár takých, v ktorých sa pozitivita vymyká normálu, budem tam rovnako skúmať, či je korelácia podobná, ako v prípade celého štátu. Potom na základe toho vyvodím záver, ktoré okresy dostatočne, resp. nedostatočne testujú, ako aj napr. efektivitu trasovania kontaktov, atď.

2 Dátové sady a ich uloženie do DB

2.1 Spôsob uloženia dátových sád do databáze

Spôsob uloženia dát je realizovaný pomocou InfluxDB. InfluxDB je open-source databáza na ukladanie údajov v časových radoch. Hlavným dôvodom je, že väčšina dostupných dátových sád je do veľkej miery závislá na čase.

Všetky dátové sady sú uložené v databáze **covid-19.** Každá dátová sada používa vlastný *measurement*. Measurement by sa dalo prirovnať k tabuľke v SQL databázach.

2.2 Stručná charakteristika zvolených dátových sád

Zvolené dátové sady boli tematicky rozdelené podľa dotazov do 5 celkov.

- **1. celok - dotaz 1 zo skupiny A**
  - COVID-19: *Přehled hospitalizací*
  - COVID-19: *Celkový (kumulativní) počet osob s prokázanou nákazou dle krajských hygienických stanic včetně laboratoří, počet vyléčených, počet úmrtí a provedených testů (v2)*
- **2. celok - dotaz zo skupiny B**
  - *Obyvatelstvo podle pětiletých věkových skupin a pohlaví v krajích a okresech*
  - *COVID-19: Přehled epidemiologické situace dle hlášení krajských hygienických stanic podle okresu*
- **3. celok - vlastný dotaz 1**
  - COVID-19: *Přehled úmrtí s ohledem na vykázaná očkování*
  - COVID-19: *Přehled hospitalizací s ohledem na vykázaná očkování*
- **4. celok - vlastný dotaz 2**
  - COVID-19: *Přehled epidemiologické situace dle hlášení krajských hygienických stanic podle okresu*
  - COVID-19: *Celkový (kumulativní) počet provedených testů podle krajů a okresů ČR*

  - *Obyvatelstvo podle pětiletých věkových skupin a pohlaví v krajích a okresech*

- **5. celok - dotaz 2 zo skupiny A dotaz zo skupiny C**
  - *COVID-19: Přehled vykázaných očkování podle krajů ČR*
  - *COVID-19: Přehled osob s prokázanou nákazou dle hlášení krajských hygienických stanic*

3 Implementácia

Projekt je realizovaný pomocou nástroja *Docker*[1], kde sú vytvorené tzv. kontajnery a ich štruktúra je zadefinovaná pomocou súboru *docker-compose.yaml*.

InfluxDB

InfluxDB slúži na vytvorenie a beh inštancie nerelačnej databázy. Do tejto databázy sa ukladajú všetky dáta cez tzv. servisu *scraper*. Nad dátami sa robia vizualizácie znázorňujúce jednotlivé dotazy zo zadania, a to pomocou služby *visualizer*.

Scraper

Scraper je systém, ktorý má na starosti stiahnutie, predspracovanie a uloženie dát do NoSQL databázy InfluxDB.

Visualizer

Služba *visualizer* vytvára konkrétne vizualizácie nad uloženými dátami v InfluxDB.

Dotazy a vizualizácie nad dátami

Všetky úlohy sú vypracované pomocou programovacieho jazyka Python, pomocou ktorého sú všetky dáta extrahované z NoSQL databázy. Základom pre extrakciu dát je formulácie dotazov pre InfluxDB sprostredkovaná pomocou knižnice v Pythone, z ktorých sú výsledné dáta *parsované* už pomocou klasických dátových štruktúr charakteristických pre tento jazyk. Vizualizácie dát sú realizované pomocou knižníc *Matplotlib*, *Plotly* a *Seaborn*.

3.0.1 Dotazy pre skupinu A

Na vypracovanie dotazov zo skupiny A boli použité nástroje zo sekcie 3. Spôsob vypracovania tejto úlohy sa dá sekvenčne rozdeliť do štyroch krokov, ktorých výsledkom sú výstupné CSV súbory a grafy, vizuálne popisujúce riešenie konkrétnej úlohy.

[1] https://www.docker.com

Vypracovanie prvej úlohy pre skupinu A

Úlohou je vytvoriť spojnicové grafy, ktoré zobrazujú vývoj covidovej situácie po jednotlivých mesiacoch. V grafoch majú byť počty novo nakazených za mesiac, novo vyliečených za mesiac, novo hospitalizovaných za mesiac a počty prevedených testov za mesiac. Celkovo boli vytvorené 3 grafy. Pre grafy boli použité datasety:

- COVID-19: Celkový (kumulativní) počet osob s prokázanou nákazou dle krajských hygienických stanic včetně laboratoří, počet vyléčených, počet úmrtí a provedených testů (v2)
- COVID-19: Přehled hospitalizací

Extrakcia dát z NoSQL

Extrakcia dát prebieha pomocou dotazov so syntaxou InfluxDB, ktorá je sprostredkovaná pomocou knižnice v Pythone. Použité dotazy:

- `select sum("prirustkovy_pocet_nakazenych"), sum("prirustkovy_pocet_vylecenych"), sum("prirustkovy_pocet_provedenych_testu") from "group_1-nakazeni-vyleceni-umrti-testy"where time >='2020-03-01T00:00:00Z' and time <= '2021-11-30T00:00:00Z' group by time(30d).`
- `select sum("pocet_hosp") from "group_1-hospitalizace" where time >='2020-03-01T00:00:00Z' and time <='2021-11-30T00:00:00Z' group by time(30d)).`

Spracovanie dát

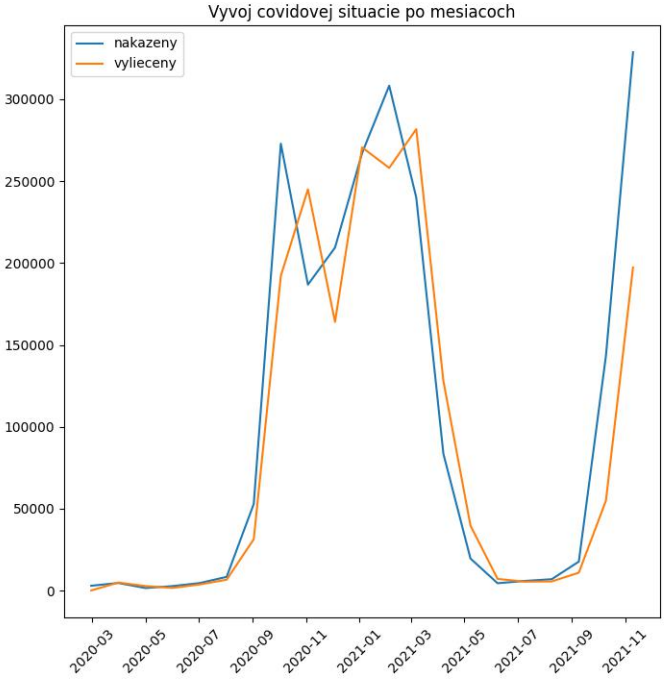
Dáta z dotazov sú spojene do jedného slovníka s klúčom mesiac.

Uloženie dát do formátu CSV

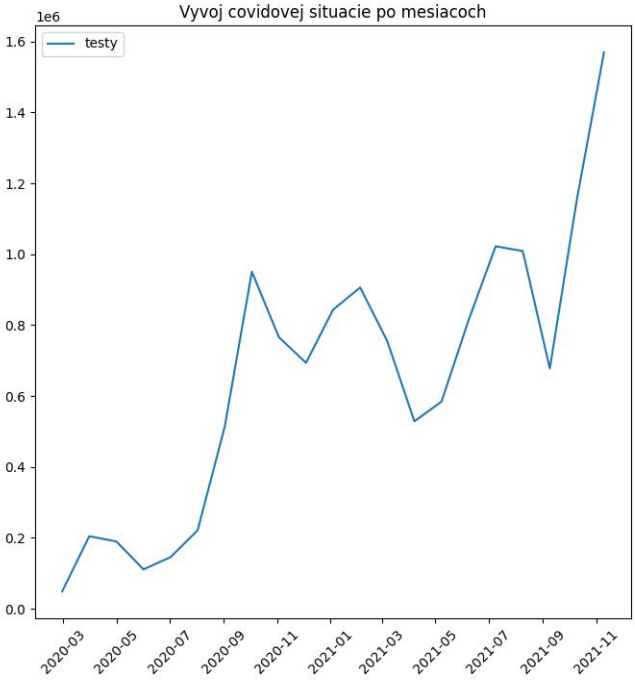
Dáta zo slovníku sú uložené do podaresára `csvs/xkubik32` ako súbor s názvom `task_a_1.csv`.

Vizualizácia dát z CSV súborov

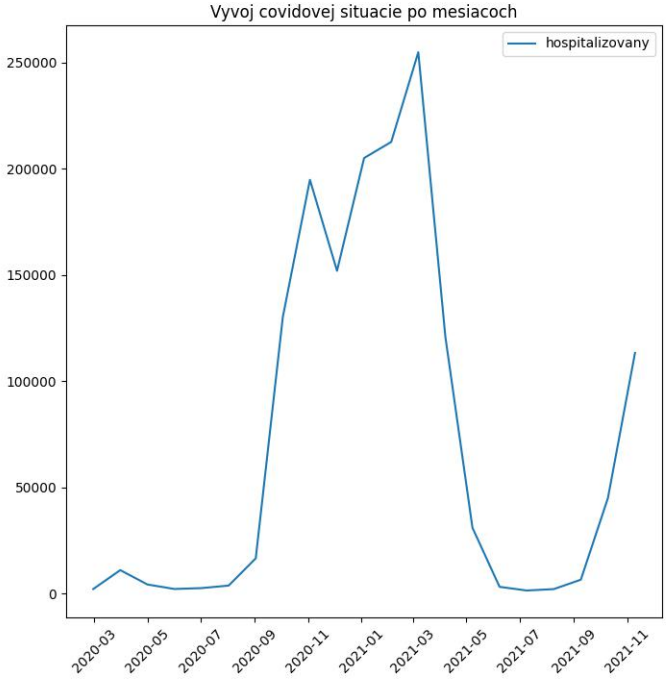
Zo súboru `task_a_1.csv` sú vytvorené 3 grafy:



Obr. 3.1: Počty novo nakazených a vyliečených pre jednotlivé mesiace.



Obr. 3.2: Počty testov pre jednotlivé mesiace.



Obr. 3.3: Počty hospitalizovaných pre jednotlivé mesiace.

Vypracovanie druhej úlohy pre skupinu A

Úlohou je vytvoriť krabicový graf zobrazujúci rozdelenie veku nakazených osôb v jednotlivych krajoch. Celkovo bol vytvorený 1 graf. Pre graf bol použitý dataset:

- COVID-19: Přehled osob s prokázanou nákazou dle hlášení krajských hygienických stanic

Extrakcia dát z NoSQL

Extrakcia dát prebieha pomocou dotazov so syntaxou InfluxDB, ktorá je sprostredkovaná pomocou knižnice v Pythone. Použité dotazy:

- `select count(id) from "group_5-osoby"where "vek"<= 15 group by "kraj_nuts_kod"`
- `select count(id) from "group_5-osoby"where "vek"> 15 and "vek"<= 30 group by "kraj_nuts_kod"`
- `select count(id) from "group_5-osoby"where "vek"> 30 and "vek"<= 45 group by "kraj_nuts_kod"`
- `select count(id) from "group_5-osoby"where "vek"> 45 and "vek"<= 60 group by "kraj_nuts_kod"`
- `select count(id) from "group_5-osoby"where "vek"> 60 group by "kraj_nuts_kod"`

Spracovanie dát

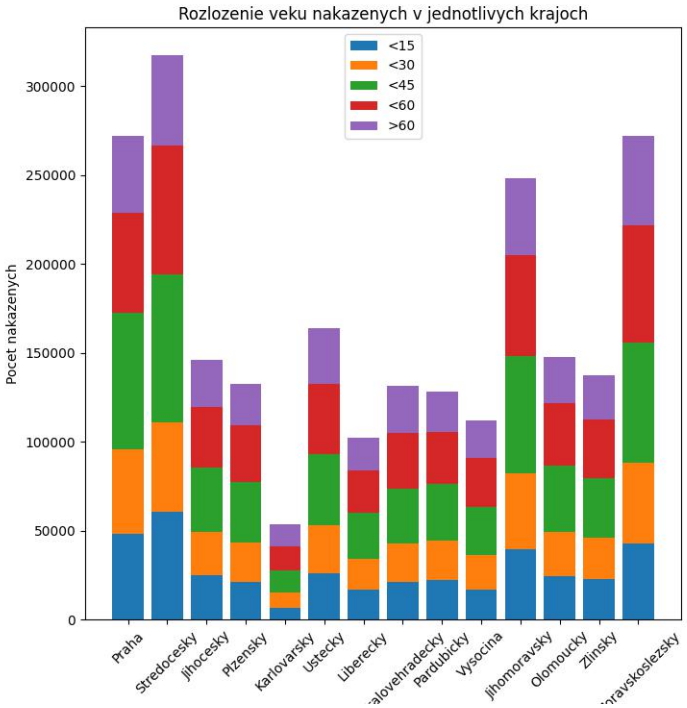
Dáta z dotazov sú spojene do jedného slovníka s klúčom názov kraja.

Uloženie dát do formátu CSV

Dáta zo slovníku sú uložené do podadresára `csvs/xkubik32` ako súbor s názvom `task_a_2.csv`.

Vizualizácia dát z CSV súboru

Zo súboru `task_a_2.csv` je vytvorený 1 graf:



Obr. 3.4: Rozloženie veku nakazených osôb v jednotlivých krajoch.

3.0.2 Dotaz pre skupinu B

Na vypracovanie dotazov zo skupiny B boli použité nástroje zo sekcie 3. Spôsob vypracovania tejto úlohy sa dá sekvenčne rozdeliť do štyroch krokov, ktorých výsledkom sú výstupné CSV súbory a grafy, vizuálne popisujúce riešenie konkrétnej úlohy.

Vypracovanie prvej úlohy pre skupinu B

Úlohou je vytvoriť tzv. rebríček krajov v ČR za posledný rok, symetricky rozdelený na štyri kvartály (časové rozmedzie 3 mesiacov) skúmajúci priebeh nákazy ľudí ochorením COVID-

19. Tento rebríček bude graficky vyhodnocovať, ktoré kraje mali v prepočte na obyvateľa najhorší alebo najlepší priebeh. Rovnako bude v grafoch možné skúmať priebeh nákazy s dennou incidenciou, a teda každodenné prírastky počtu ochorení či kumulatívne počty prípadov v časovom vývoji.

Extrakcia dát z NoSQL

Extrakcia dát prebieha pomocou dotazov so syntaxou InfluxDB, ktorá je sprostredkovaná pomocou knižnice v Pythone. Dotazy bolo potrebné formulovať s prihliadaním na časové rozmedzia. Taktiež bolo potrebné preskúmať bližšiu charakteristiku uložených dát, čo viedlo k použitiu klauzule `GROUP BY` a `SUM` aby bolo možné získať súčty denných počtov nakazených pre daný kraj, keďže dáta boli diverzifikované ešte aj podľa jednotlivých okresov v kraji.

Príkladom takéhoto dotazu je napríklad: `query('SELECT SUM("kumulativni_pocet_nakazenych") FROM "group_2-kraj-okres-nakazeni-vyleceni-umrti"WHERE time > now() - 182d AND time <= now() - 91d GROUP BY time(1d), "kraj_nuts_kod‘‘)`.

Spracovanie dát

Spracovanie extrahovaných dát prebieha následne pomocou klasickej práce s dátovými štruktúrami jazyka Python, akými sú napríklad *list*, *dictionary* alebo *tuple*, ktoré sú výstupom realizovaného dotazu. Príkladom takéhoto spracovania môže byť napríklad kód 3.1.

Listing 3.1: Spracovanie dát.

```
for result in result_query.get_points():
    record["datum"] = result['time']
    record["umrti_celkom"] = result['zemreli_celkem']
```

V tejto časti je taktiež potrebné dáta upraviť tak, aby bolo možné získať denné prírastky medzi jednotlivými dňami.

Uloženie dát do formátu CSV

Spracované dáta sú následne jednoduchým spôsobom ukladané do jednotlivých CSV súborov podľa potreby danej úlohy a požadovaných výsledkov na výsledné grafy. Záhlavie je volené kontextovo podobné ako pri dátach uložených v databáze, kde stĺpce sú atribúty danej problematiky a riadky sú konkrétne hodnoty. V prípade tejto podúlohy sú výsledné CSV súbory ukladané do osobitnej zložky `csvs/xotcen01` s názvami *first_quartal.csv*, *second_quartal.csv*, *third_quartal.csv* a *fourth_quartal.csv*, kde každý z týchto súborov obsahuje dáta požadované pre jedno štvrťročie z posledného roku. Súbory sú následné použité pre vizualizáciu dát.

Vizualizácia dát z CSV súborov

Vizualizácia dát prebieha prostredníctvom knižnice Plotly, ktorá je vhodná na komplexnejšie vykreslenie grafov. Je potrebné dáta načítať z vytvorených CSV súborov. Tento krok

je realizovaný pomocou knižnice *Pandas*, ktorá načíta dáta zo súborov, nad ktorými je potom možné pracovať a zobraziť požadované výsledky. Načítanie dát prebieha následným spôsobom: `data = pd.read_csv(name)`.

Kvôli objemnosti výsledných grafov, ktoré popisujú situáciu v jednotlivých krajoch z hľadiska dennej incidencie v prepočte na jedného obyvateľa v kraji, kumulatívneho nárastu v porovnaní s počtom obyvateľov v kraji a iným sú výsledné grafy priložené v prílohe A.

Vyhodnotenie

Výsledné grafy zobrazujú obdobie štvrtého kvartálu, a teda obdobie posledných troch mesiacov od súčasnosti. Na základe veľkého množstva údajov bolo neprehľadné zobraziť väčšie množstvo údajov do jedného grafu, a preto bolo vyhotovených viacero náhľadov na konkrétnu problematiku. Ako kritérium na vyhodnotenie rebríčku bol volený počet nakazených v prepočte na jedného obyvateľa, z čoho vyplynulo, že v tomto období je v súčasnosti Královohradecký, Pardubický a Zlínský kraj s najhorším vývojom nákazy, kde v prepočte hodnoty dosahujú nákazu približne 0.22 nakazeného obyvateľa v kraji. Naopak najlepšie na tom je kraj Karlovarský s hodnotou 0.18. Grafické výsledky pre zvyšné kvartály sú uložene v zložke `plots/xotcen01`.

3.0.3 Dotazy pre skupinu C

Vypracovanie prvej úlohy pre skupinu C

V tejto úlohe som sa zameral na zhlukovanie okresov do skupín na základe istých atribútov. Tieto atribúty sú:

- Podiel nakazenej populácie za posledný rok
- Počet dávok vakcín na obyvateľa za posledný rok
- Podiel obyvateľov od 0 do 14 rokov k celkovej populácii
- Podiel obyvateľov od 15 do 59 rokov k celkovej populácii

Kategóriu od 60 rokov vyššie som sa rozhodol vynechať z dôvodu, že je závislá na ostatných 2 vekových kategóriách (teda je nimi jednodznačne určená). Pre každý okres sa vypočítajú tieto 4 atribúty. Tieto atribúty slúžia ako súradnice v 4D priestore pre zhlukovací algoritmus. Potom sa vypočítajú zhluky okresov pomocou metódy **K-Means**. Pre túto metódu bola využitá knižnica **scikit-learn**. Rozhodol som sa pre rozdelenie okresov do 4 zhlukov. Pre nedostupnosť dát o očkovaní v rámci jednotlivých okresov som sa rozhodol spriemerovať počty podaných dávok v okresoch na základe počtu obyvateľov v rámci daného kraja.

Extrakcia dát z NoSQL

Bolo použitých celkovo 6 dotazov, resp. skupín dotazov. Tieto dotazy/skupiny boli nasledovné:

- Počet obyvateľov v okresoch a krajoch
- Počet obyvateľov v okresoch od 0 po 14 rokov
- Počet obyvateľov v okresoch od 15 do 59 rokov
- Kumulatívny počet nakazených v okresoch pred 1 rokom
- Kumulatívny počet nakazených v okresoch k dňu vypracovania
- Počet podaných vakcín v jednotlivých krajoch

Počty obyvateľov v rámci vekových skupín boli rozložené na viac dotazov, konkrétne po 5-ročných intervaloch, keďže tieto boli v databázi uložené.

Spracovanie dát

Prvá úprava dát spočívala v započítaní kraja Praha ako okresu, keďže v databáze bol uložený len ako kraj. Ďalej boli dáta o očkovaniach upravené do jednoduchšej štruktúry pomocou operácií so slovníkmi. Podobne boli vyfiltrované a upravené aj dáta o počte obyvateľov v krajoch. Ďalšia časť pozostávala zo prepočtu podielu podaných dávok k počtu obyvateľov v krajoch. Tento podiel bol priradený každému okresu z daného kraja. Následne sa pre každý okres vypočítali podiely počtu obyvateľov v jednotlivých vekových skupinách, ako aj podiel nakazených za dané obdobie. Počet nakazených bol vypočítaný ako rozdiel čísel z dotazov 4 a 5, ktorý sa následne vydelil celkovým počtom obyvateľov. Takto boli získané všetky 4 atribúty.

Zhlukovanie

Ako už bolo spomínané, na zhlukovanie bola použitá metóda **K-Means**. Táto rozdelila okresy do predom nastaveného počtu 4 zhlukov. Výsledok bol potom vypísaný ako slovník do textového súboru.

Vyhodnotenie

Jednou zo zaujímavostí pri výsledných zhlukoch bolo, že okres Praha bol v samostatnom zhluku. Dá sa predpokladať, že je to spôsobené ako vekovým zložením mesta, tak aj nadpriemerným počtom podaných dávok vakcín. Rozdelenie okresov do skupín je na obrázku 3.5.

```
{'0': ['České Budějovice', 'Český Krumlov', 'Jindřichův Hradec', 'Písek', 'Prachatice', 'Strakonice', 'Tábor', 'Domažlice',
'Klatovy', 'Plzeň-město', 'Plzeň-jih', 'Plzeň-sever', 'Rokycany', 'Tachov', 'Hradec Králové', 'Jičín', 'Náchod',
'Rychnov nad Kněžnou', 'Trutnov', 'Blansko', 'Brno-město', 'Brno-venkov', 'Břeclav', 'Hodonín', 'Vyškov', 'Znojmo'],
'1': ['Cheb', 'Karlovy Vary', 'Sokolov', 'Děčín', 'Chomutov', 'Litoměřice', 'Louny', 'Most', 'Teplice', 'Ústí nad Labem',
'Česká Lípa', 'Jablonec nad Nisou', 'Liberec', 'Semily', 'Chrudim', 'Pardubice', 'Svitavy', 'Ústí nad Orlicí', 'Havlíčkův Brod'
'Jihlava', 'Pelhřimov', 'Třebíč', 'Žďár nad Sázavou', 'Jeseník', 'Olomouc', 'Prostějov', 'Přerov', 'Šumperk', 'Kroměříž',
'Uherské Hradiště', 'Vsetín', 'Zlín', 'Bruntál', 'Frýdek-Místek', 'Karviná', 'Nový Jičín', 'Opava', 'Ostrava-město'],
'2': ['Praha'],
'3': ['Benešov', 'Beroun', 'Kladno', 'Kolín', 'Kutná Hora', 'Mělník', 'Mladá Boleslav', 'Nymburk', 'Praha-východ',
'Praha-západ', 'Příbram', 'Rakovník']}
```

Obr. 3.5: Rozdelenie okresov do skupín.

3.0.4 Vlastné dotazy

Vypracovanie vlastného dotazu 1

V tejto úlohe je podstatou preskúmať a vizualizovať aký vplyv má vakcína v prípade hospitalizácie alebo prípadného úmrtia. Aj keď nie je možné túto otázku objektívne pomocou takéhoto prístupu plne vyhodnotiť, úlohou je čo možno najpresnejšie priblížiť efekt očkovania pri spomínaných prípadoch.

Extrakcia dát z NoSQL

Extrakcia dát prebieha veľmi obdobne ako v sekcií 3.0.2. Na vyhodnotenie dát o hospitalizáciach a úmrtiach s preukázaným očkovaním boli formulované následné dotazy, ktoré zozbierajú dáta za posledný rok:

```
query('SELECT"hospitalizovani_celkem",
"hospitalizovani_bez_ockovani",
"hospitalizovani_nedokoncene_ockovani",
"hospitalizovani_dokoncene_ockovani"
FROM "group_3-ockovani-hospitalizace"
WHERE time > now() - 365d')

query('SELECT "zemreli_celkem",
"zemreli_bez_ockovani",
"zemreli_nedokoncene_ockovani",
"zemreli_dokoncene_ockovani"
FROM "group_3-ockovani-umrti"
WHERE time > now() - 365d')
```

Spracovanie dát

Následne sú výsledky dotazov spracované pomocou jazyka Python obdobne ako v sekcií 3.0.2 pre spracovanie dát.

Uloženie dát do formátu CSV

Podobne ako v sekcií 3.0.2 pre uloženie dát do formátu CSV sú výsledné semi-štruktúrovné hodnoty v jazyku Python uložené pre potreby úlohy do požadovaných CSV súborov, z ktorých sú potom vytvorené výsledné grafy.

Vizualizácia dát z CSV súborov

Vizualizácia dát je prevedená prvotným načítaním CSV súborov *deaths.csv* a *hospitalizations.csv* pomocou knižnice Pandas a následnou aplikáciou knižnice Plotly. Dáta zobrazujú pomocou stĺpcových a čiarových grafov dennú incidenciu počtu hospitalizovaných a zomrelých na základe požadovaného faktoru vakcinácie. Incidencia je sledovaná v období

posledného roku po súčasnosť no taktiež sú vyobrazené dáta od začiatku Novembra po súčasnosť, ktoré poskytujú bližší pohľad na jednotlivé pomery počas tretej vlny pandémie v ČR, ktorá práve v tomto období najviac zasiahla Českú republiku. Výsledné grafy sú pre svoju obsiahlosť uložené v prílohách B.

Vyhodnotenie

Na základe výsledných grafov a s prihliadnutím, že v ČR sa v aktuálnom období nachádza vyše 60% očkovaných osôb je možné vidieť, že väčší podiel úmrtí a hospitalizácií majú ľudia, ktorí očkovaní neboli. Taktiež by bolo možné z výsledných grafov usúdiť, že ľudia iba s jednou dávkou sú vo výsledku na tom najlepšie, ale je potrebné zobrať do úvahy fakt, že títo ľudia sú po krátkej dobe plne očkovaní, a teda majú len krátky časový interval, počas ktorého by sa mohli nakaziť[2].

Vypracovanie vlastného dotazu 2

Pre túto úlohu som sa rozhodol skúmať pozitivitu testov v jednotlivých okresoch a jej závislosť od počtu testov na obyvateľa. Pre každý okres bol vypočítaný pomer súčtu pozitivít testov pre každý deň k súčtu počtov testov na obyvateľa pre každý deň. Tieto boli normalizované ako 7-dňový kĺzavý priemer. Z týchto pomerov sa vybrali 3 okresy s najnižími a najvyššími pomermi. Pre týchto 6 okresov boli potom zhotovené grafy priebehu pozitivity testov a počtu testov na obyvateľa v čase.

Extrakcia dát z NoSQL

Boli použité celkovo 3 dotazy:

- Kumulatívny počet nakazených pre jednotlivé dni pre jednotlivé okresy
- Počet testov pre jednotlivé dni pre jednotlivé okresy
- Počet obyvateľov v jednotlivých okresoch

Spracovanie dát

Prvá úprava dát spočívala v započítaní kraja Praha ako okresu, keďže v databáze bol uložený len ako kraj. Následne boli vypočítané prírastkové počty nakazených v okresoch podľa kumulatívnych počtov. Potom boli z prírastkových počtov nakazených a testov spravené 7-dňové kĺzavé priemery, pre vyhladenie rozdiel prírastkov cez víkendy a pracovný týždeň. Z týchto počtov sa potom vypočítala pozitivita, ako aj testy na obyvateľa. V ďalšom cykle boli vypočítané pomery pozitivity ku počtu testov na obyvateľa. Potom sa vybrali 3 najlepšie (najmenší pomer) a 3 najhoršie (najväčší pomer) okresy. Pre tieto boli potom zostrojené grafy.

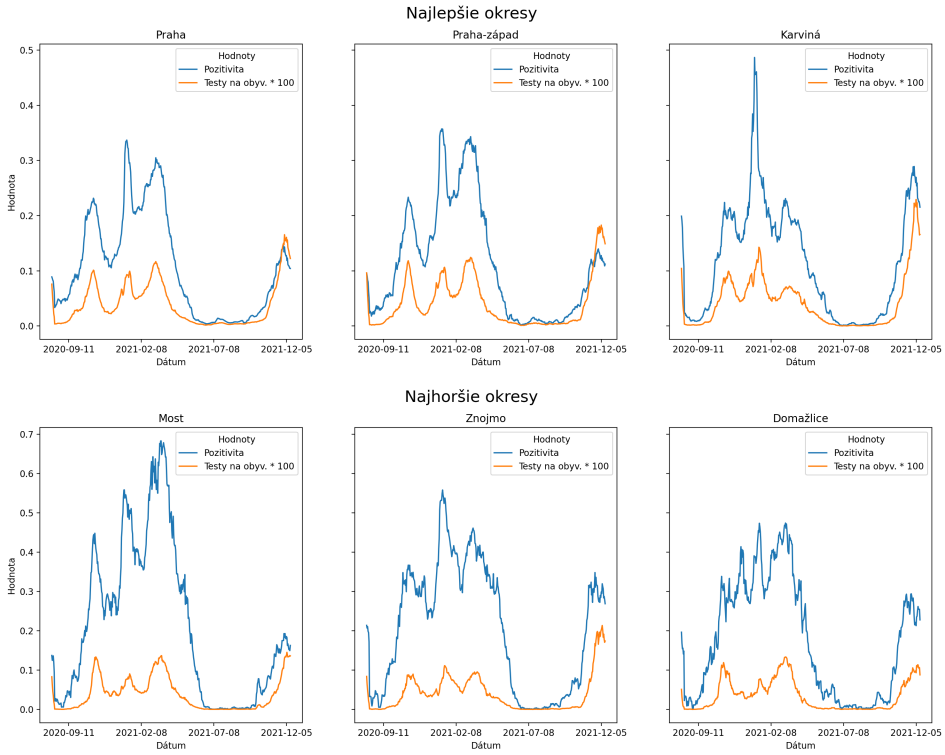
[2]Informácie boli čerpané z oficiálnej stránky https://onemocneni-aktualne.mzcr.cz/covid-19.

Vizualizácia dát

V tejto časti bol vizualizovaný priebeh pozitivity testov a počtu testov na obyvateľa vo vybraných 6 okresoch. Počet testov na obyvateľa bol vynásobený 100, aby bol na grafe viditeľný v pomere k pozitivite.

Vyhodnotenie

Z grafov je možné vidieť, že intenzita testovania má súvis s pozitivitou testov. Vysoká pozitivita totiž značí ako aj mohutnosť infekcie, tak aj nedostatočné testovanie. Zároveň je vidno, že tam, kde sa testovalo málo, bola pozitivita testov veľmi vysoká. Tento graf je možné vidieť na obrázku 3.6.



Obr. 3.6: Porovnanie najlepších a najhorších okresov z hľadiska pomeru pozitivity a počtu testov na obyvateľa.

4 Lokálne spustenie projektu

Požiadavky na software

- Docker engine 20.10 a vyššie
- Docker compose 1.29 a vyššie

Vytvorenie všetkých služieb

- **docker-compose build**

Spustenie jednotlivých služieb

- Najskôr je potrebné spustiť službu pre vytvorenie databázy, používateľa a jej spustenie:

  docker-compose up influxdb

- Následne je potrebné spustiť službu pre naplnenie databáze:

  docker-compose up scraper

- Po dobehnutí servisy scraper je potrebné spustiť službu *visualizer*, ktorý má na starosti vytvorenie jednotlivých grafov:

- Pred vytváraním grafov je potrebné exportovať premennú pre lokálne umiestnenie súboru:

  LOCAL_PROJECT_DIR=PATH kde **PATH** je lokálna cesta k repozitáru.

  docker-compose up visualizer

Zastavenie všetkých bežiacich služieb

- **docker-compose down**

Oddelený mód

Všetky služby je možné spustiť na pozadí v takzvanom oddelenom móde. Pomocou **-d** prepínaču pre príkaze **docker-compose**.

Priamy prístup do databázy

Do influx databáze bežiacej v službe sa dá pristupovať priamo a to pomocou príkazov:

- **docker exec -it influxdb bash**
- **influx**

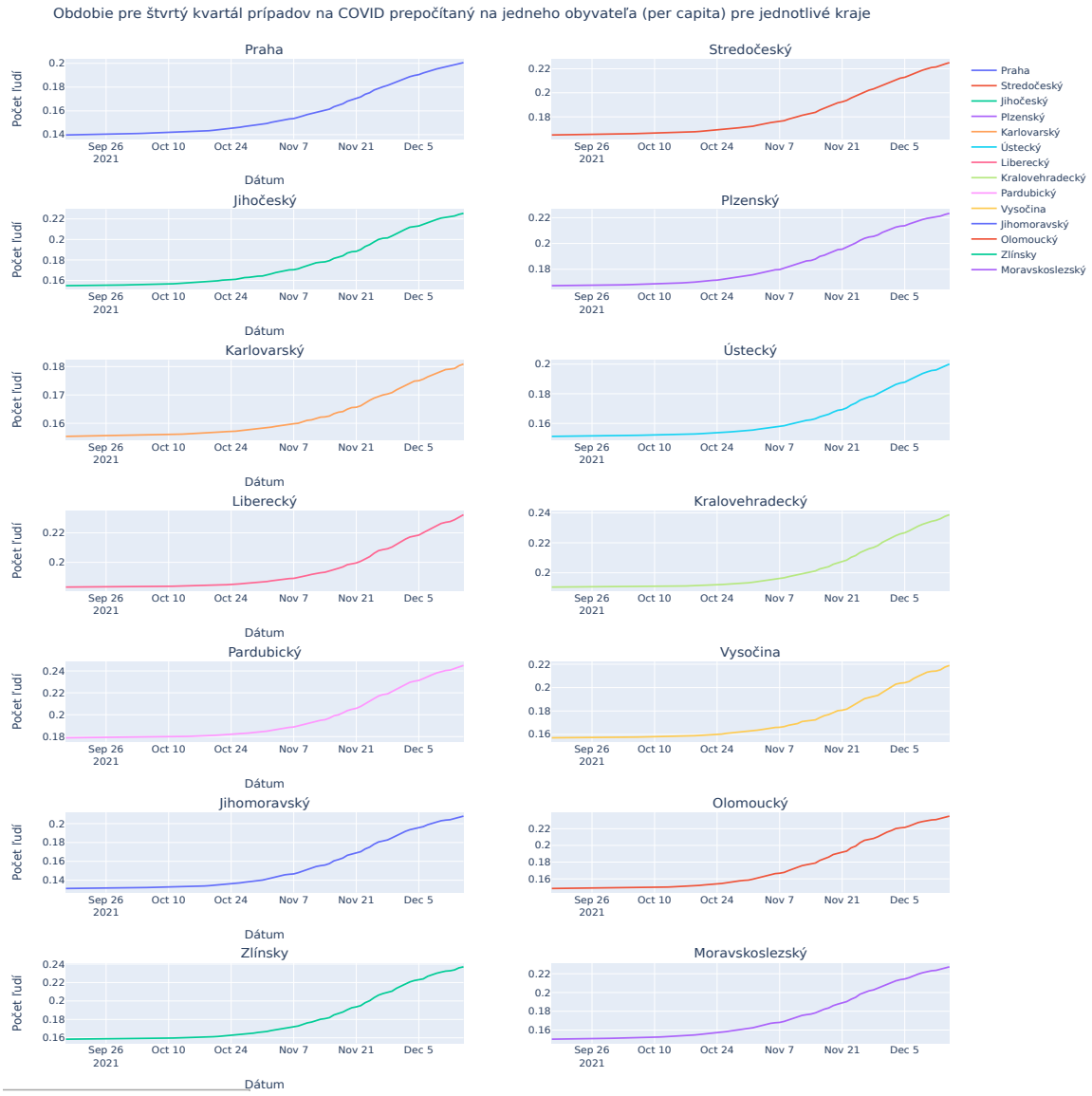
Následne môže robiť používateľ rôzne dotazy nad uloženými dátami.

Literatúra

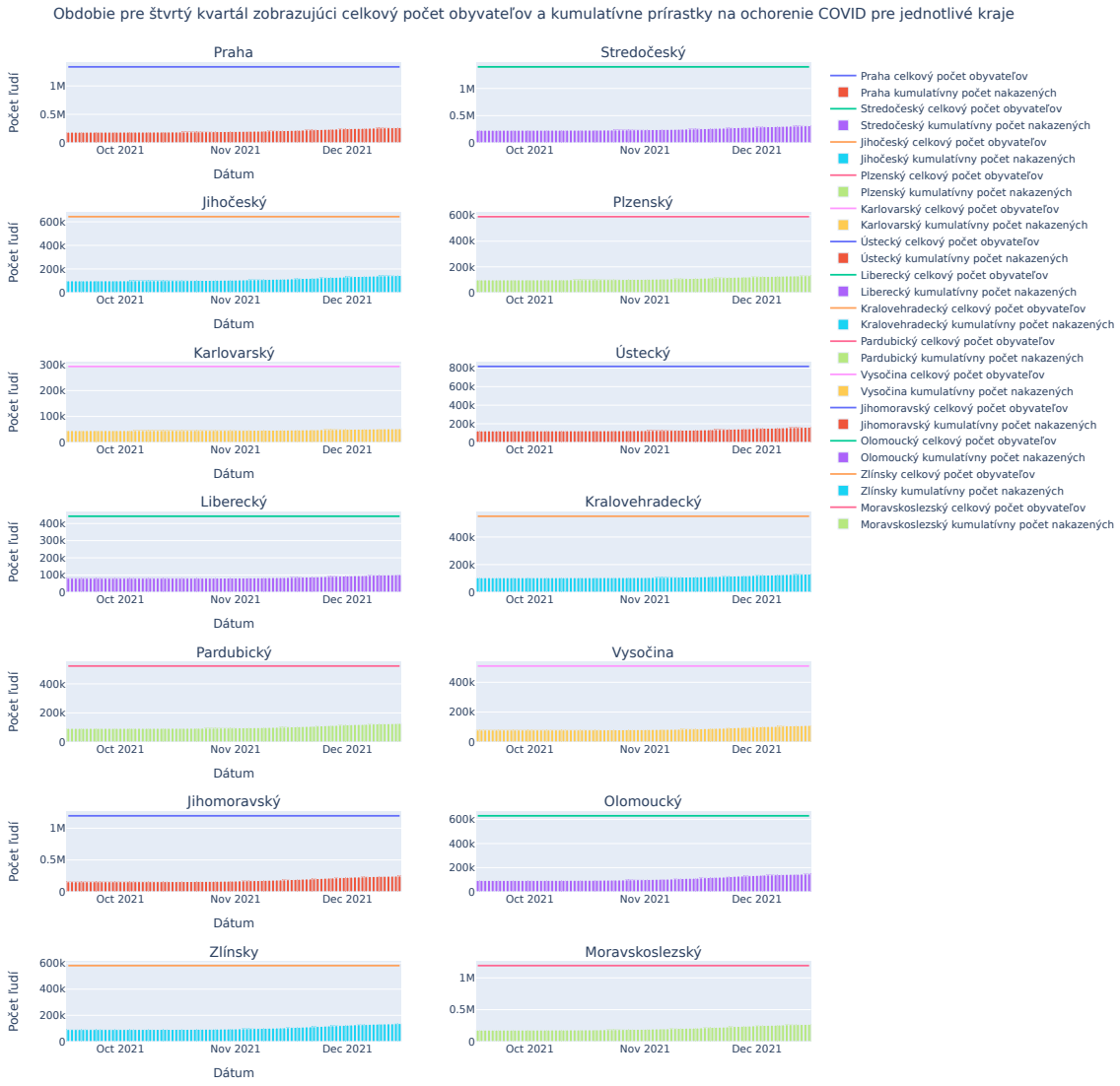
[1] Zadanie projektu [online]. Dostupné z https://wis.fit.vutbr.cz/FIT/st/cwk.php.cs?title=Projects-topics&csid=768274&id=14826

[2] InfluxDB oficiálná dokumentácia [online]. Dostupné z https://docs.influxdata.com/influxdb/v2.0/

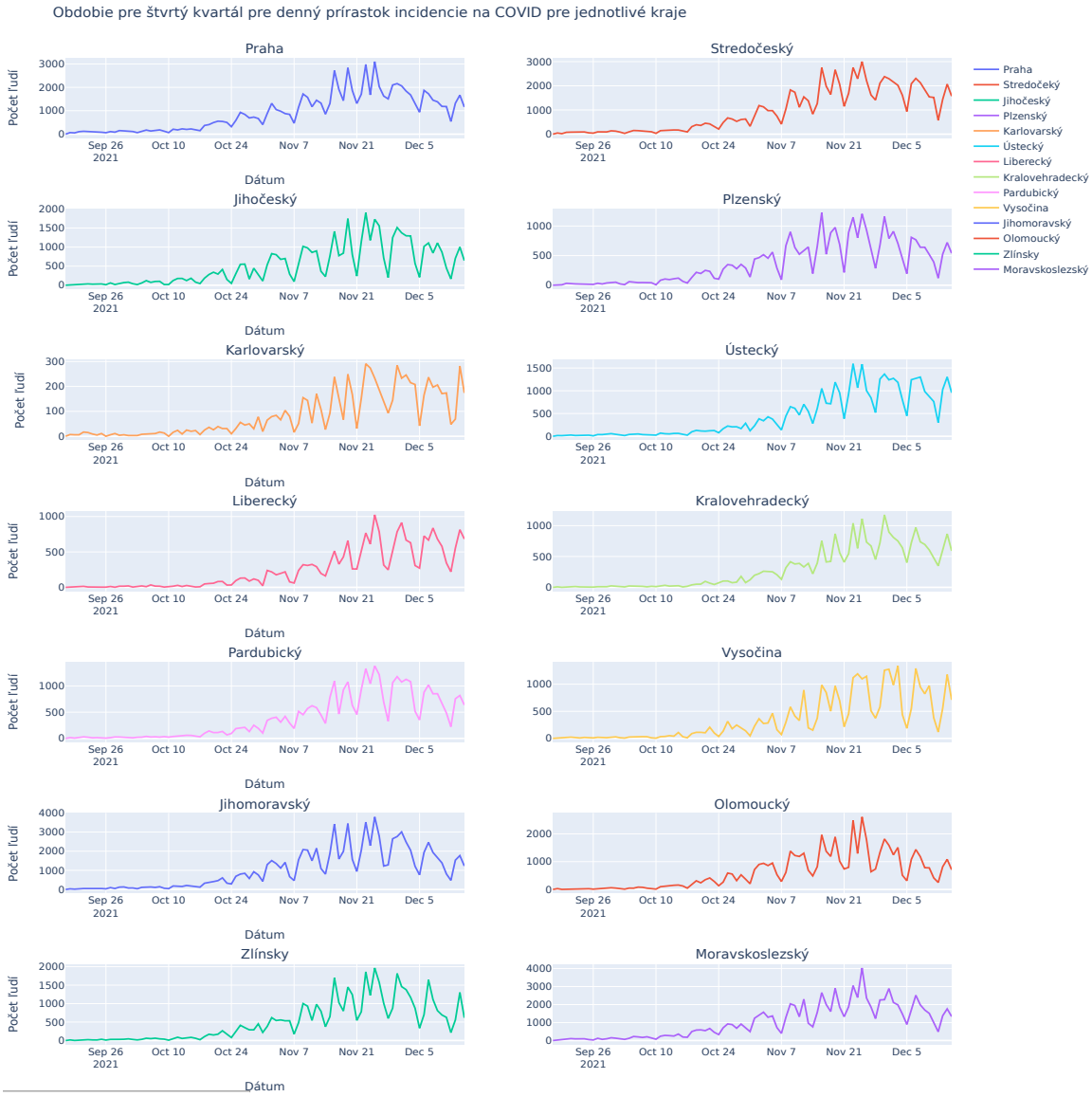
A Prílohy k úlohe B zo zadania



Obr. A.1: Obdobie štvrtého kvartálu zobrazujúce kumulatívny počet prípadov prepočítaný na jedného obyvateľa v kraji.



Obr. A.2: Obdobie štvrtého kvartálu zobrazujúce pomer medzi počtom obyvateľov v kraji a kumulatívnym prírastkom počtu prípadov.



Obr. A.3: Obdobie štvrtého kvartálu zobrazujúce dennú incidenciu prírastkov v jednotlivých krajoch.

B Prílohy k úlohe B pre vytvorené zadanie

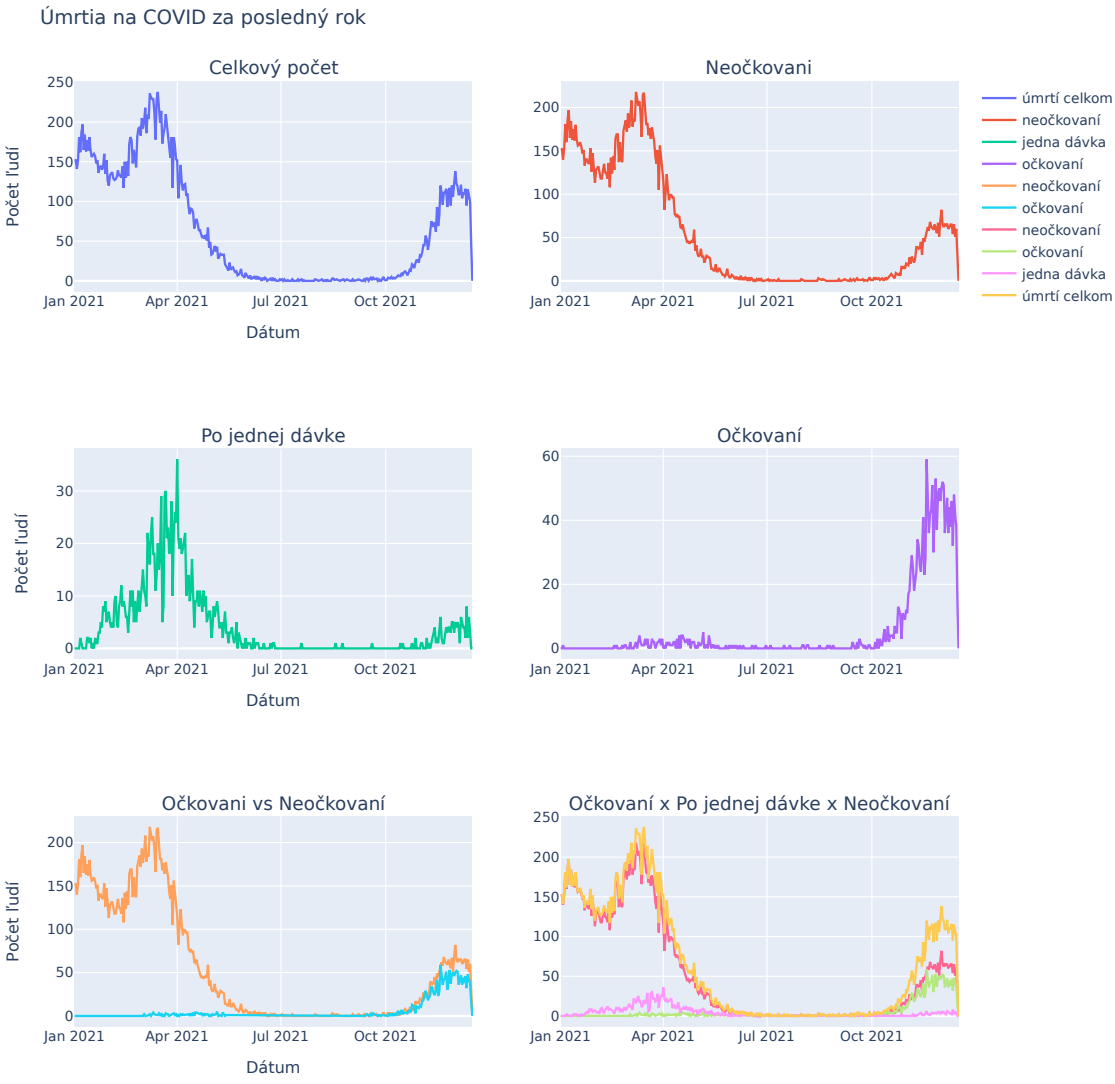
Obr. B.1: StĺpcovÉ grafy zobrazujúce pomer hospitalizovaných ľudí, ktorí neboli očkovaní, boli očkovaní čiastočne alebo boli plne očkovaní.

Obr. B.2: Spojnicové grafy zobrazujúce pomer hospitalizovaných ľudí, ktorí neboli očkovaní, boli očkovaní čiastočne alebo boli plne očkovaní.

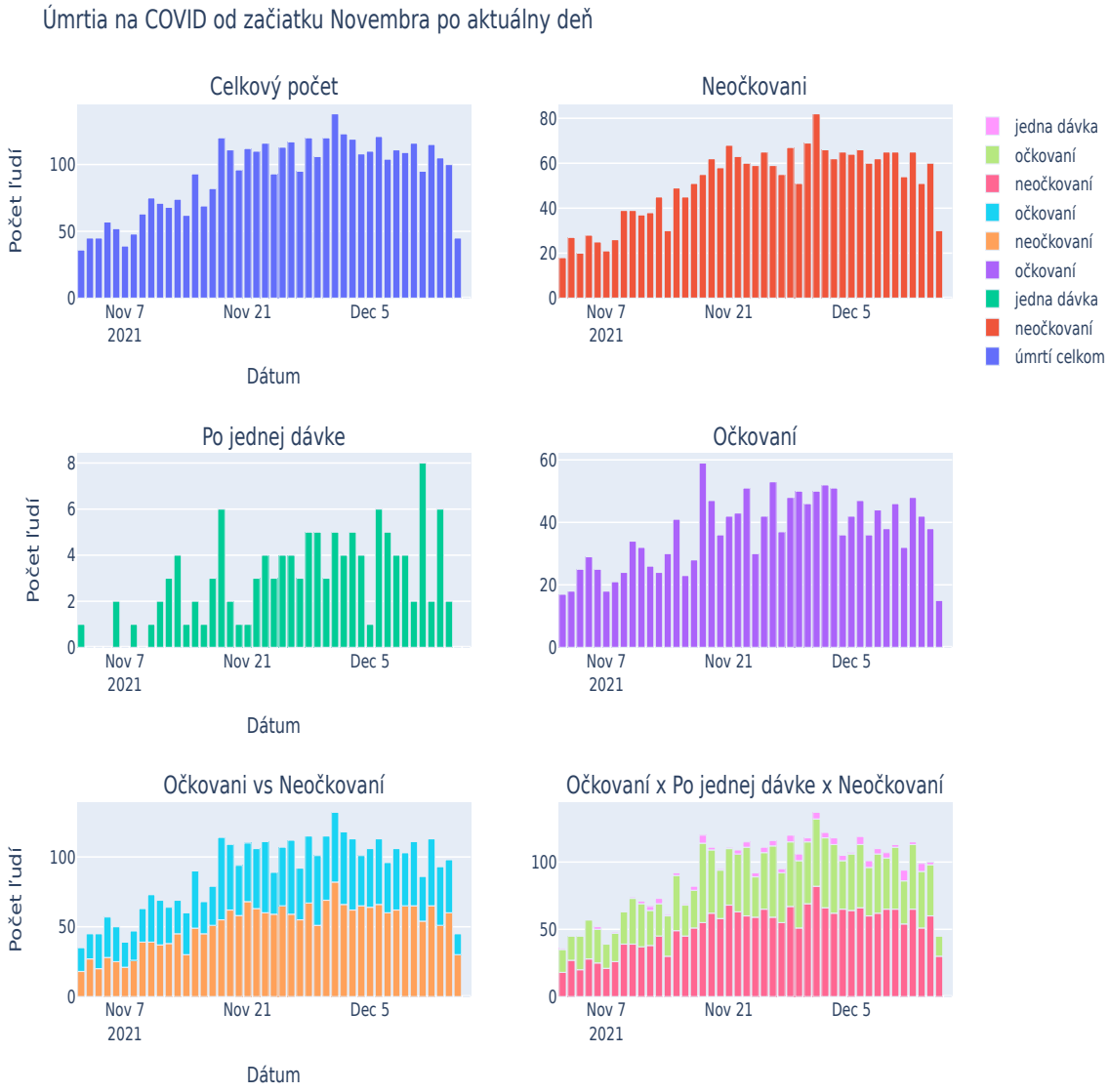
Obr. B.3: Stĺpcové grafy zobrazujúce pomer hospitalizovaných ľudí, ktorí neboli očkovaní, boli očkovaní čiastočne alebo boli plne očkovaní za obdobie od Novembra 2021 po súčasnosť.

Obr. B.4: Spojnicové grafy zobrazujúce pomer hospitalizovaných ľudí, ktorí neboli očkovaní, boli očkovaní čiastočne alebo boli plne očkovaní za obdobie od Novembra 2021 po súčasnosť.

Obr. B.5: Stĺpcové grafy zobrazujúce pomer úmrtí ľudí, ktorí neboli očkovaní, boli očkovaní čiastočne alebo boli plne očkovaní.



Obr. B.6: Spojnicové grafy zobrazujúce pomer úmrtí ľudí, ktorí neboli očkovaní, boli očkovaní čiastočne alebo boli plne očkovaní.



Obr. B.7: Stĺpcové grafy zobrazujúce pomer úmrtí ľudí, ktorí neboli očkovaní, boli očkovaní čiastočne alebo boli plne očkovaní za obdobie od Novembra 2021 po súčasnosť.

Obr. B.8: Spojnicové grafy zobrazujúce pomer úmrtí ľudí, ktorí neboli očkovaní, boli očkovaní čiastočne alebo boli plne očkovaní za obdobie od Novembra 2021 po súčasnosť.