Vysoké učení technické v Brně
Fakulta informačních technologií

Ukládání a příprava dat - projekt, část 2
COVID-19

Ondřej Krejčí
Oliver Kuník

16. prosince 2021

1 Úvod

Cílem druhé části projektu je zodpovědět dotazy ke zvolenému tématu a to vytvořením grafů a tabulek, dalším cílem je připravení dat z jednoho dotazu pro dolovací úlohu. Jako téma projektu jsme si zvolili COVID-19 a v první části jsme vytvořili skripty zajišťující stažení dat a jejich uložení do databáze MongoDB. Podrobnější dokumentace k této části projektu, všem vytvořeným kolekcím a zdrojům dat je v souboru `part1/dokumentace.md`. Řešení této části projektu je rozděleno do dvou hlavních částí.

První z nich zajišťuje získání potřebných dat pro řešení úloh z databáze a jejich uložení do souborů ve formátu csv. Tuto část řeší skript `csv_create.py`, který závisí na první části projektu a má tedy podobné požadavky pro spuštění jako řešení první části projektu. Vyžaduje, aby byla spuštěná databáze a v ní dostupná očekávaná data uložená v první části projektu. Dále závisí na několika datových souborech stažených v první části, jedná se o číselníky pro věkové kategorie, kraje atd. Tato data se používají pro získávání identifikátorů potřebných záznamů, získání názvů atd. při dotazech a ukládání dat do souborů. Ze souborů stažených v první části jsou v archivu přiloženy pouze tyto.

Druhá část řešení už pracuje jen se soubory ve formátu csv vytvořenými v předchozí části. Skript `plot_graphs.py` načítá data z csv souborů, případně ještě provede potřebné úpravy a následně vykreslí grafy a uloží je do souboru. Skript `prepare_dm.py` provádí přípravu dat ze vstupního csv souboru pro dolovací úlohu.

2 Načtení dat pro zvolené dotazy

V této části jsou vypsány všechny řešené dotazy. Pro každý z nich je zde popsáno načítání dat potřebných k jejich zodpovězení z databáze a následné uložení těchto dat do souborů ve formátu csv.

2.1 Dotaz A1

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.

Pro vytvoření požadovaného grafu jsou potřebné hodnoty přírůstků nakažených, vyléčených, hospitalizovaných a provedených testů za celou Českou republiku po měsících. Pro účely tohoto dotazu jsme vytvořili přehledovou kolekci `covid_po_dnech_cr`, která obsahuje denní hodnoty přírůstků pro všechny požadované hodnoty.

Jako měsíc, od kterého jsou data načítány, byl zvolen duben 2020, což je první celý měsíc, pro který jsou v databázi data pro všechny potřebné hodnoty. Jako poslední měsíc byl ze stejného důvodu zvolen listopad 2021.

Pro získání požadovaných hodnot jsou sečteny dané přírůstkové hodnoty po jednotlivých měsících (od prvního po poslední den měsíce, včetně) a načtená data jsou uložena do souboru `A1-covid_po_mesicich.csv`.

2.2 Dotaz A2

Vytvořte krabicové grafy zobrazující rozložení věku nakažených osob v jednotlivých krajích.

Pro vytvoření požadovaných krabicových grafů je nutné získat záznamy o případech nákazy jednotlivců s informací o jejich věku a kraji. Data o jednotlivých nakažených jsou dostupná v kolekci `nakazeni_vek_okres_kraj`.

Pro tento dotaz používáme i záznamy o nákaze, které nemají informaci o kraji, navíc ještě odlišujeme nákazy v zahraničí. Z kolekce se načtou všechny záznamy a potřebné hodnoty se uloží do souboru `A2-osoby_nakazeni_kraj.csv`.

2.3 Dotaz B1

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).

Pro účely tohoto dotazu je nutné získat přírůstky nakažených v jednotlivých krajích za celá čtvrtletí. Dále je pro jednotlivé kraje nutné získat jejich celkovou populaci.

Data o přírůstku nakažených je možné získat z kolekce `nakazeni_vyleceni_umrti_testy_kraj`, která mj. obsahuje kumulativní počet nakažených v jednotlivých krajích po dnech. Jako čtvrtletí jsme zvolili poslední celá čtvrtletí, tedy poslední čtvrtletí roku 2020 a tři čtvrtletí roku 2021. Konkrétně se jedná o časová období 1. října až 31. prosince 2020, 1. ledna až 31. března, 1. dubna až 30. června a 1. července až 30. září 2021. Z kolekce se pro všechny kraje načtou hodnoty pro první den každého čtvrtletí a pro první den následujícího čtvrtletí (1. října 2021).

Populaci krajů lze získat z kolekce `obyvatelstvo_kraj`, ze které se pro každý kraj načtou nejnovější hodnoty celkové populace. Údaje o populaci krajů se připojí ke kumulativním hodnotám nakažených pro jednotlivé kraje a jsou uloženy do souboru `B1-nakazeni_kumulativne_kraj.csv`.

2.4 Dotaz C1

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).*

Zadání tohoto dotazu požaduje nalezení dat pro 50 měst, všechny potřebné údaje ale nebyly dostupné, proto jsme pro účely tohoto dotazu nahradili města obcemi s rozšířenou působností (ORP), jak již bylo popsáno v dokumentaci k první části projektu.

Tento dotaz vyžaduje získání přírůstku nakažených a provedených očkování za celá čtvrtletí pro 50 zvolených ORP. Dále je potřeba získat celkovou populaci ORP ve třech daných věkových skupinách. Rozhodli jsme se použít data pro 50 největších ORP (bez Prahy). Pro tento dotaz se používají stejná čtyři čtvrtletí jako u dotazu B1.

Pro získání skupin obyvatelstva byla vytvořena kolekce `obyvatele_orp`, která obsahuje pro každou ORP její populaci rozdělenou do zadaných skupin. Data o počtech nakažených lze získat z kolekce `nakazeni_orp`, která obsahuje přírůstky nakažených na úrovni ORP po jednotlivých

dnech. Data o provedených očkováních jsou dostupná v kolekci `ockovani_orp`, která obsahuje data o počtu očkovaných dávek na úrovni ORP po dnech. Pro účely tohoto dotazu tedy pro hodnotu očkování používáme celkový počet očkovaných dávek (ne celkový počet ukončených očkování).

Získání dat začíná načtením prvních 50 záznamů z kolekce `obyvatele_orp` seřazené podle celkové populace, čímž se získají skupiny obyvatel pro 50 největších ORP. Pro každou ORP se následně pro všechny čtvrtletí provede dotaz do kolekcí s počtem nakažených a počtem dávek očkování, který sečte přírůstky od začátku po konec daného čtvrtletí. Načtené hodnoty jsou uloženy do souboru `C1-orp-ctvrtleti.csv`.

2.5 Vlastní dotaz 1 (dotaz D1)

Vizualizace "nadúmrtí" způsobených covidem za dobu trvání pandemie. Jedná se o spojnicový graf zobrazující podíl úmrtí na covid a celkových úmrtí za celou ČR po týdnech.

Pro tento dotaz se používají údaje o zemřelých v celé ČR z kolekce `umrti_cr`, která byla vytvořena z datové sady ČSÚ *Zemřelí podle týdnů a věkových skupin v České republice*. Data o úmrtích na covid jsou opět získávána z přehledové kolekce `covid_po_dnech_cr`. Bylo zvoleno rozmezí začínající počátkem roku 2020, tedy před vypuknutím pandemie, kdy ještě nebyly zaznamenány žádné úmrtí na covid, a končící týdnem od 6. do 12. září, což je poslední týden, pro který byly do databáze uloženy data o celkových úmrtích.

Načtou se záznamy o úmrtích v celé ČR ve zvoleném rozmezí a následně se z kolekce `covid_po_dnech_cr` načte suma přírůstků úmrtí za daný týden[1]. Výsledné hodnoty pro jednotlivé týdny se uloží do souboru `D1-zemreli_cr.csv`.

2.6 Vlastní dotaz 2 (dotaz D2)

Histogram poměru úmrtí na covid za celou dobu trvání pandemie a počtu obyvatel pro věkové kategorie po deseti letech.

Data o celkových úmrtích se opět získávají z kolekce `obyvatelstvo_kraj`, která obsahuje pro jednotlivé kraje i populaci ve věkových skupinách po 5 letech. Poslední věková kategorie v kolekci je 95 let a více, poslední věková skupina po 10 letech tak bude od 90 let výše. Data o úmrtích jsou dostupná v kolekci `umrti_vek_okres_kraj`, která obsahuje záznamy o jednotlivých úmrtích s informací o věku.

Jako první se načtou počty obyvatel pro desetileté věkové kategorie, které se získají jako suma odpovídajících pětiletých věkových skupin pro všechny kraje. Pro každou věkovou kategorii se následně sečte počet záznamů v kolekci `umrti_vek_okres_kraj`, u kterých hodnota věku spadá do dané kategorie. Data jsou uložena do souboru `D2-zemreli_vekove_kategorie.csv`

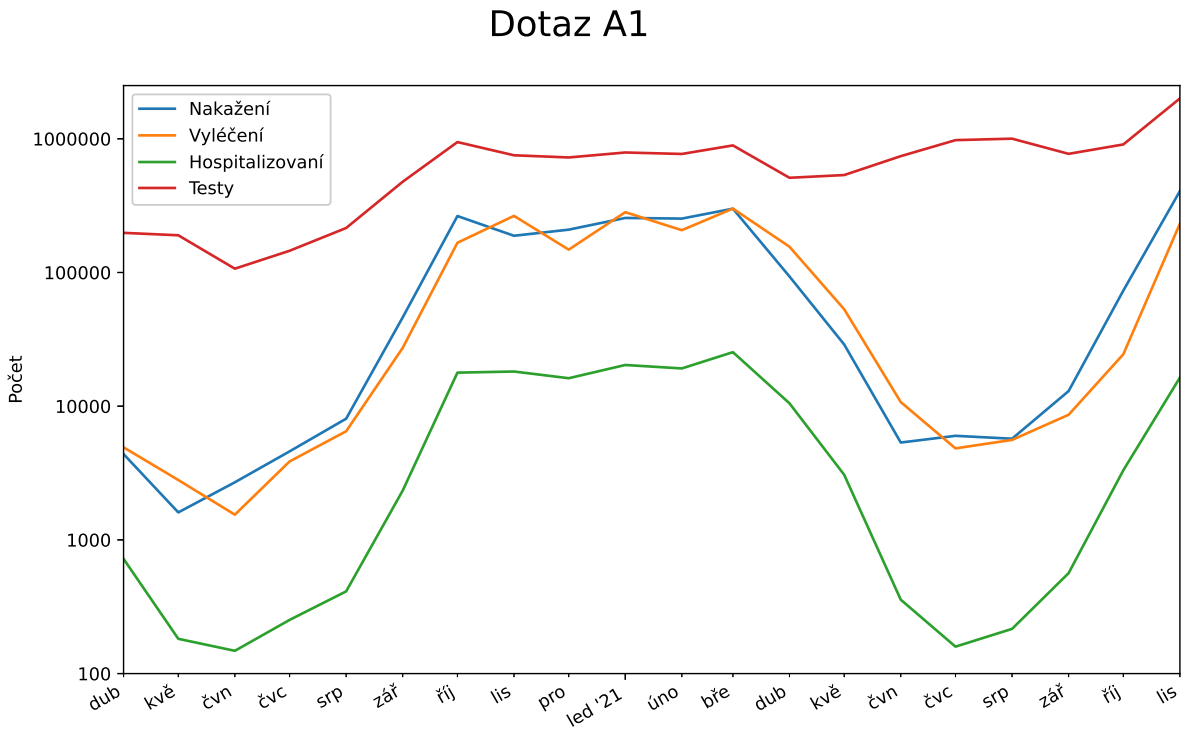
[1] Je přiloženo i řešení pro databázi MongoDB 3.6 a vyšší, které propojení kolekcí a agregaci provede jedním dotazem.

3 Řešení dotazů

Tato sekce popisuje vytváření grafů a tabulek z dat ze vstupních souborů ve formátu csv dle zadání jednotlivých dotazů. Většina souborů již obsahuje potřebná data pro splnění dotazů. Zpracování je proto v této části již minimální. Výjimkou je dotaz B1, kde bylo potřebné přepočítat kumulativní hodnoty na přírůstky.

3.1 Dotaz A1

Pro všechny řádky načtené ze vstupního csv souboru se všechny atributy (nakažení, vyléčení, hospitalizovaní a testy) vykreslí v čárovém grafu 1 v závislosti na čase (po měsících). Pro lepší přehlednost grafu byla pro osu Y zvolena logaritmická stupnice.

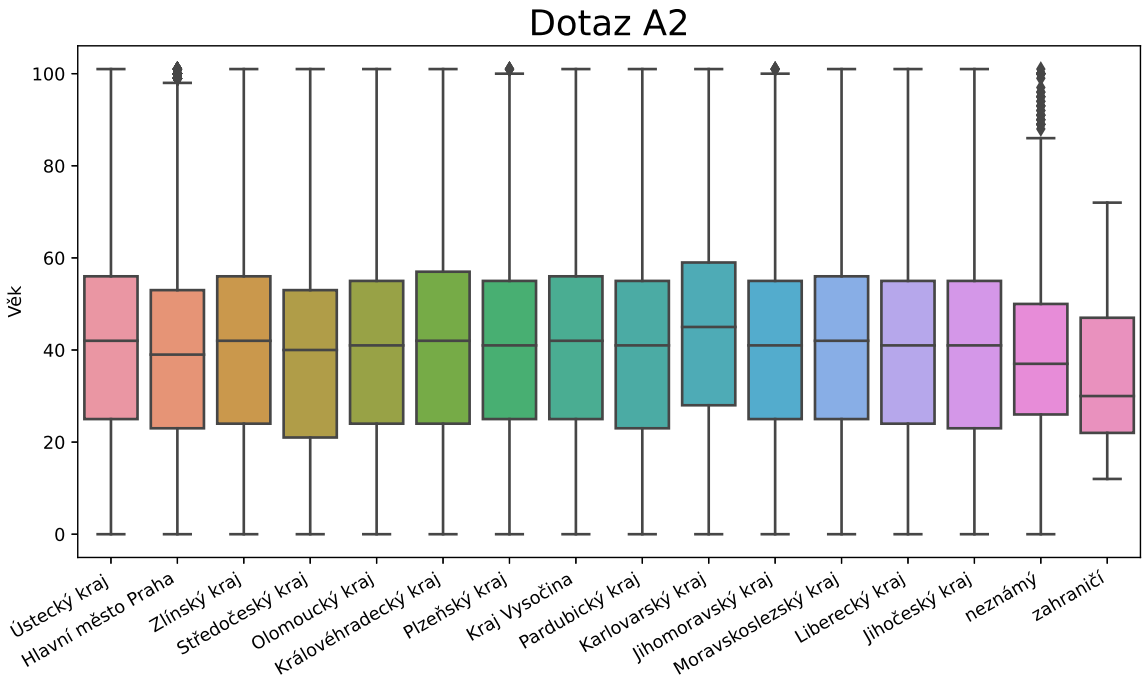


Graf 1: Vývoj covidové situace v ČR po měsících

3.2 Dotaz A2

Ukázalo se, že ve vstupních datech se v atributu věk nachází i několik nesprávných hodnot, jako například věk 119 let. Proto byly odstraněny řádky s odlehlými hodnotami věku pro celý soubor dat (nezávisle na krajích). Jako kritérium pro detekci odlehlých hodnot byla zvolena hodnota 1,5 násobku mezikvartilového rozpětí od prvního (respektive třetího) kvartilu.

Následně byla vytvořena sada grafů 2 složená z krabicových grafů věku nakažených pro každý kraj. Také byly vytvořeny krabicové grafy pro nakažené, u kterých nebyla zjištěna informace o kraji, ze kterých byli dále vyčleněni ti, u kterých došlo k nákaze v zahraničí.



Graf 2: Krabicové grafy zobrazující rozložení věku nakažených osob v jednotlivých krajích (obsahuje také nakažené, u kterých není známý kraj)

3.3 Dotaz B1

Data ve vstupním csv souboru obsahovala kumulativní počet nakažených pro kraje na začátku 5 čtvrtletí. Bylo proto potřeba spočítat přírůstek nakažených pro každé čtvrtletí. Toho bylo docíleno tak, že se pro každé vypočítal rozdíl kumulativního počtu nakažených na jeho začátku a na začátku následujícího čtvrtletí. Získaly se tak hodnoty přírůstku nakažených pro čtyři čtvrtletí.

Kritérium počet nakažených na jednoho obyvatele kraje bylo spočítáno jako podíl počtu nakažených za dané čtvrtletí vůči počtu obyvatel kraje. Pro každé čtvrtletí byly potom kraje seřazené podle tohoto kritéria od „nejlepšího“ po „nejhorší“.

Z těchto hodnot byly následně vytvořeny tabulky. Pro druhé čtvrtletí byla tato data také vykreslena do grafu 3. Počty obyvatel a počty nakažených jsou zde zobrazené jako dva sloupce pro každý kraj. Počet nakažených na jednoho obyvatele kraje je vykreslený jako jako čárový graf s vlastní osou.

	Název kraje	Počet obyvatel	Přírůstek nakažených	Poměr
1	Zlínský kraj	580119	46552	0,08025
2	Královéhradecký kraj	550803	39354	0,07145
3	Kraj Vysočina	508852	35944	0,07064
4	Liberecký kraj	442476	30162	0,06817
5	Pardubický kraj	522856	34673	0,06631
6	Moravskoslezský kraj	1192834	78743	0,06601
7	Olomoucký kraj	630522	40962	0,06497
8	Jihočeský kraj	643551	39002	0,06060
9	Středočeský kraj	1397997	83964	0,06006
10	Plzeňský kraj	591041	34634	0,05860
11	Ústecký kraj	817004	46363	0,05675
12	Jihomoravský kraj	1195327	63054	0,05275
13	Hlavní město Praha	1335084	65445	0,04902
14	Karlovarský kraj	293311	13593	0,04634

Tabulka 1: Tabulka počtu nakažených na jednoho obyvatele od 1. 10. 2020 do 31. 12. 2020

	Název kraje	Počet obyvatel	Přírůstek nakažených	Poměr
1	Královéhradecký kraj	550803	60184	0,10927
2	Karlovarský kraj	293311	28728	0,09794
3	Liberecký kraj	442476	43191	0,09761
4	Pardubický kraj	522856	49494	0,09466
5	Plzeňský kraj	591041	54787	0,09270
6	Středočeský kraj	1397997	120245	0,08601
7	Ústecký kraj	817004	61625	0,07543
8	Jihočeský kraj	643551	46917	0,07290
9	Hlavní město Praha	1335084	88890	0,06658
10	Kraj Vysočina	508852	32826	0,06451
11	Olomoucký kraj	630522	40316	0,06394
12	Moravskoslezský kraj	1192834	74368	0,06235
13	Jihomoravský kraj	1195327	72652	0,06078
14	Zlínský kraj	580119	31930	0,05504

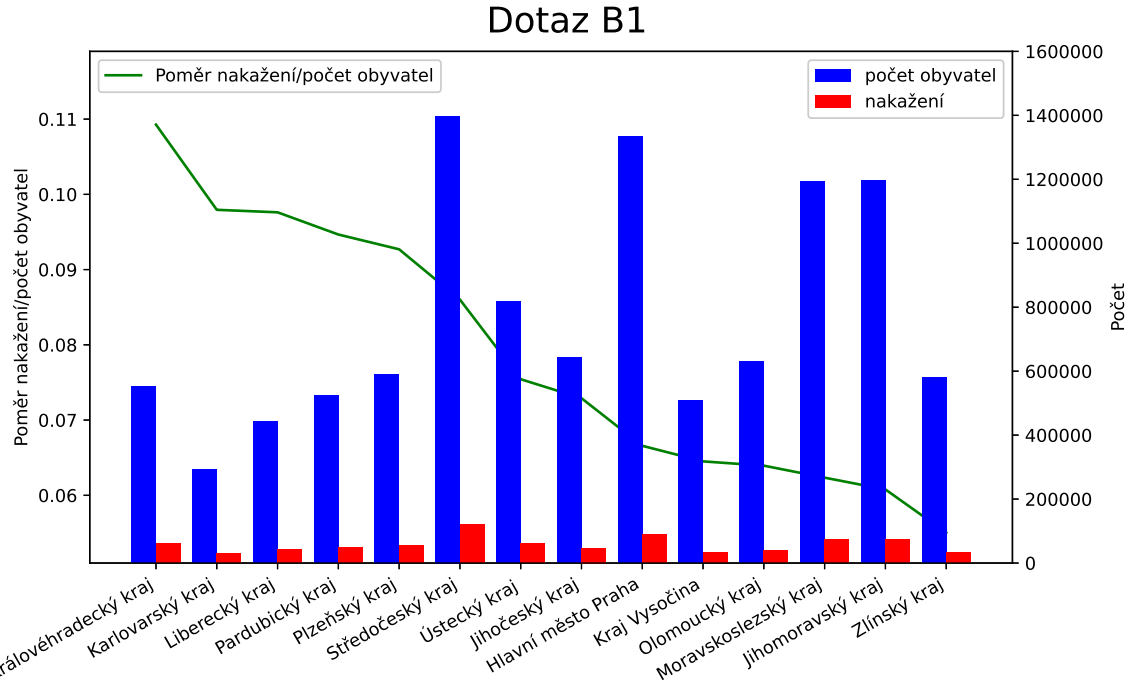
Tabulka 2: Tabulka počtu nakažených na jednoho obyvatele od 1. 1. 2021 do 31. 3. 2021

	Název kraje	Počet obyvatel	Přírůstek nakažených	Poměr
1	Zlínský kraj	580119	9369	0,01615
2	Jihočeský kraj	643551	9871	0,01534
3	Kraj Vysočina	508852	7550	0,01484
4	Ústecký kraj	817004	11420	0,01398
5	Moravskoslezský kraj	1192834	16120	0,01351
6	Olomoucký kraj	630522	8160	0,01294
7	Pardubický kraj	522856	5875	0,01124
8	Jihomoravský kraj	1195327	13153	0,01100
9	Liberecký kraj	442476	4842	0,01094
10	Středočeský kraj	1397997	13586	0,00972
11	Hlavní město Praha	1335084	12350	0,00925
12	Plzeňský kraj	591041	4459	0,00754
13	Královéhradecký kraj	550803	2586	0,00469
14	Karlovarský kraj	293311	1106	0,00377

Tabulka 3: Tabulka počtu nakažených na jednoho obyvatele od 1. 4. 2021 do 30. 6. 2021

	Název kraje	Počet obyvatel	Přírůstek nakažených	Poměr
1	Hlavní město Praha	1335084	6137	0,00460
2	Plzeňský kraj	591041	1656	0,00280
3	Středočeský kraj	1397997	3830	0,00274
4	Jihočeský kraj	643551	1504	0,00234
5	Moravskoslezský kraj	1192834	2700	0,00226
6	Karlovarský kraj	293311	601	0,00205
7	Jihomoravský kraj	1195327	2395	0,00200
8	Pardubický kraj	522856	892	0,00171
9	Kraj Vysočina	508852	828	0,00163
10	Zlínský kraj	580119	906	0,00156
11	Liberecký kraj	442476	674	0,00152
12	Ústecký kraj	817004	1218	0,00149
13	Olomoucký kraj	630522	877	0,00139
14	Královéhradecký kraj	550803	591	0,00107

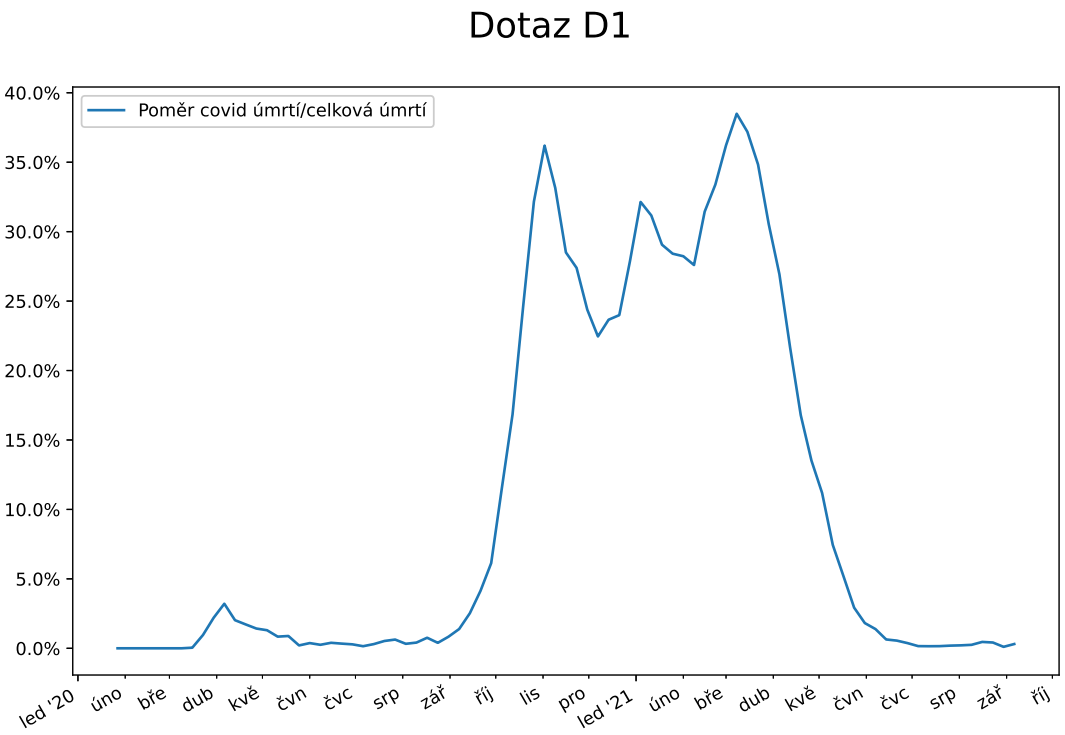
Tabulka 4: Tabulka počtu nakažených na jednoho obyvatele od 1. 7. 2021 do 30. 9. 2021



Graf 3: Graf celkového počtu nově nakažených, celkového počtu obyvatel a počtu nakažených na jednoho obyvatele podle krajů pro čtvrtletí od 1. ledna do 31. března 2021

3.4 Dotaz D1

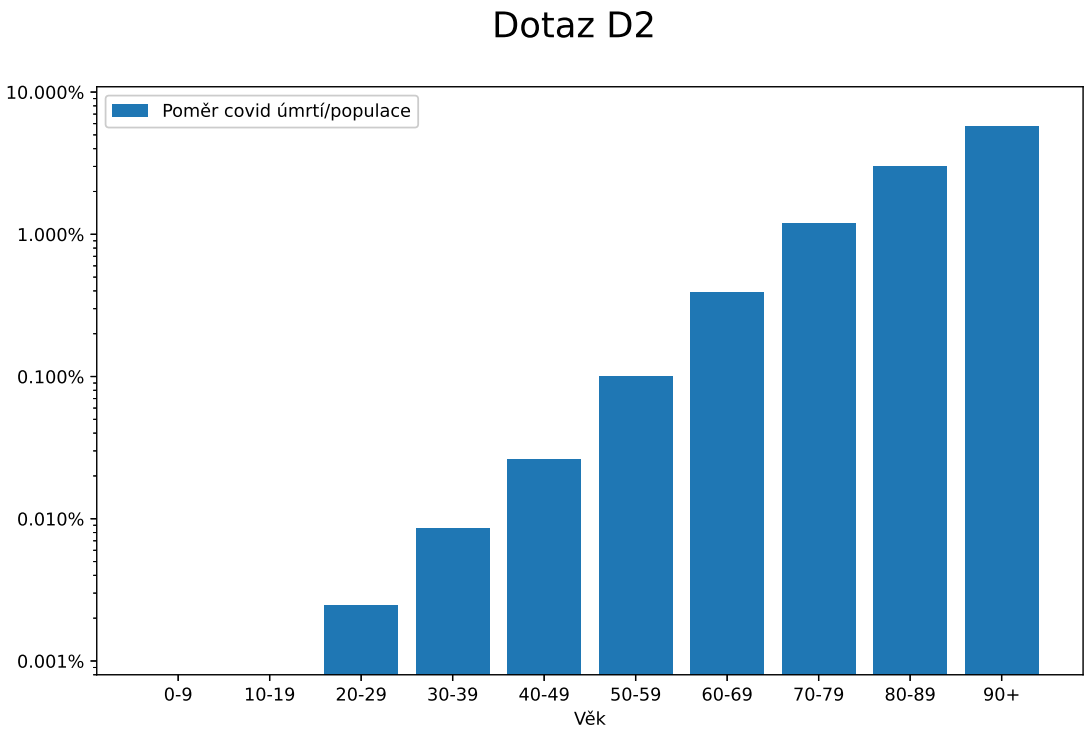
Pro všechny načtené řádky byl spočítán podíl úmrtí na covid vůči celkovým úmrtím za daný týden. Hodnoty byly následně převedeny na procenta a vykresleny ve spojnicovém grafu 4 v závislosti na čase.



Graf 4: Graf podílu úmrtí na covid a celkových úmrtí v celé České republice po týdnech

3.5 Dotaz D2

Vstupní soubor ve formátu csv již obsahuje data o jednotlivých věkových kategoriích. Pro každou z nich bylo procento úmrtí na covid spočítáno jako podíl úmrtí na covid vůči počtu obyvatel v dané věkové kategorii. Data byla následně zobrazena pomocí sloupcového grafu 5. Pro lepší vizualizaci dat byla použita logaritmická stupnice.



Graf 5: Histogram poměru úmrtí na covid za celou dobu trvání pandemie a počtu obyvatel pro věkové kategorie po deseti letech

4 Příprava dat pro dolovací úlohu

Vstupní soubor ve formátu csv již obsahuje potřebná data, tzn. každý řádek obsahuje záznam s potřebnými atributy. Jedním objektem jsou data o nakažených a očkování pro jednu ORP za čtvrtletí, objektů je celkem 200. Na těchto vstupních datech se dále prováděly požadované úpravy.

Načtená data byla zpracovávaná pro každé čtvrtletí zvlášť. V atributech věkových skupin byly všechny odlehlé hodnoty nenacházející se v rozmezí 1,5násobku mezikvartilového rozpětí od prvního nebo třetího kvartilu nahrazeny hraničními hodnotami tohoto rozmezí. Data neobsahovala žádné hodnoty nižší než dané rozpětí, proto zde nedošlo k žádnému nahrazení. Několik hodnot však rozmezí přesahovalo a byly proto nahrazeny maximální hodnotou zvoleného rozmezí. Konkrétně se jednalo o 4 hodnoty atributu věkové kategorie 0–14, 3 hodnoty kategorie 15–59 a 3 hodnoty kategorie 60+.

Hodnoty atributu počet nakažených byly nahrazeny normalizovanými hodnotami v rozmezí 0 až 1. Data byla normalizována pomocí vzorce $(x - x_{min})/(x_{max} - x_{min})$.

Pro hodnoty atributu počet dávek byla provedena diskretizace do tří stejně velkých skupin. Skupina **bad** představuje ORP s malým počtem očkovaných dávek vakcíny, ve skupině **medium** jsou ORP se středním počtem dávek a ve skupině **good** jsou ORP s největším počtem dávek vakcíny.

Vytvořený soubor ve formátu csv `C1-orp_ctvrtleti-upraveno.csv` obsahuje 50 řádků pro každé čtvrtletí a každý řádek odpovídá jedné obci s rozšířenou působností. Celkem tedy stále obsahuje 200 řádků, pouze jsou seřazeny podle čtvrtletí.

V prvním sloupci je datum začátku daného čtvrtletí, ve druhém sloupci je datum jeho konce (poslední den čtvrtletí). Třetí sloupec obsahuje kód obce s rozšířenou působnosti a čtvrtý sloupec její název. Pátý, šestý a sedmý sloupec obsahuje počty obyvatel ORP ve věkových skupinách 0–14, 15–59 a 60+, v těchto sloupcích byly nahrazeny odlehlé hodnoty. V osmém sloupci jsou normalizované hodnoty počtu nakažených. Devátý sloupec obsahuje nezměněné počty očkovaných dávek vakcíny a v desátém sloupci jsou jejich diskretizované hodnoty (tj. skupina, do které byla ORP zařazena dle počtu dávek vakcíny).

5 Přehled obsahu odevzdaného archivu

- `csv_create.py` – skript na vytvoření vstupních csv souborů z dat v databázi
- `plot_graphs.py` – skript na vytvoření grafů a tabulek
- `prepare_dm.py` – skript připravující data pro dolovací úlohu
- `part1_main.py` – main skript pro první část projektu
- `part1\` – řešení první části projektu
  - `data\` – složka pro stažená data z první části projektu, obsahuje pouze číselníky potřebné i pro druhou část projektu
  - `dokumentace.md` – dokumentace první části projektu
- `data_csv\` – složka s vytvořenými soubory ve formátu csv
- `README.md` – readme pro obě části projektu