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. **[[dolovaci uloha]]**

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.

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`.

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`.

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`.

[[uprava kumulativnich hodnot]]

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`.

Vlastní dotaz 1 (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`.

Vlastní dotaz 2 (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. 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 (poslední kategorie je 90+), které se získají jako suma přes pětileté věkové skupiny pro všechny kraje. Pro každou věkovou kategorii se následně sečte počet záznamů v kolekci umrti_vek_okre_kraj, u kterých hodnota věku spadá do dané kategorie. Data jsou uložena do souboru `D2-zemreli_vekove_kategorie.csv`

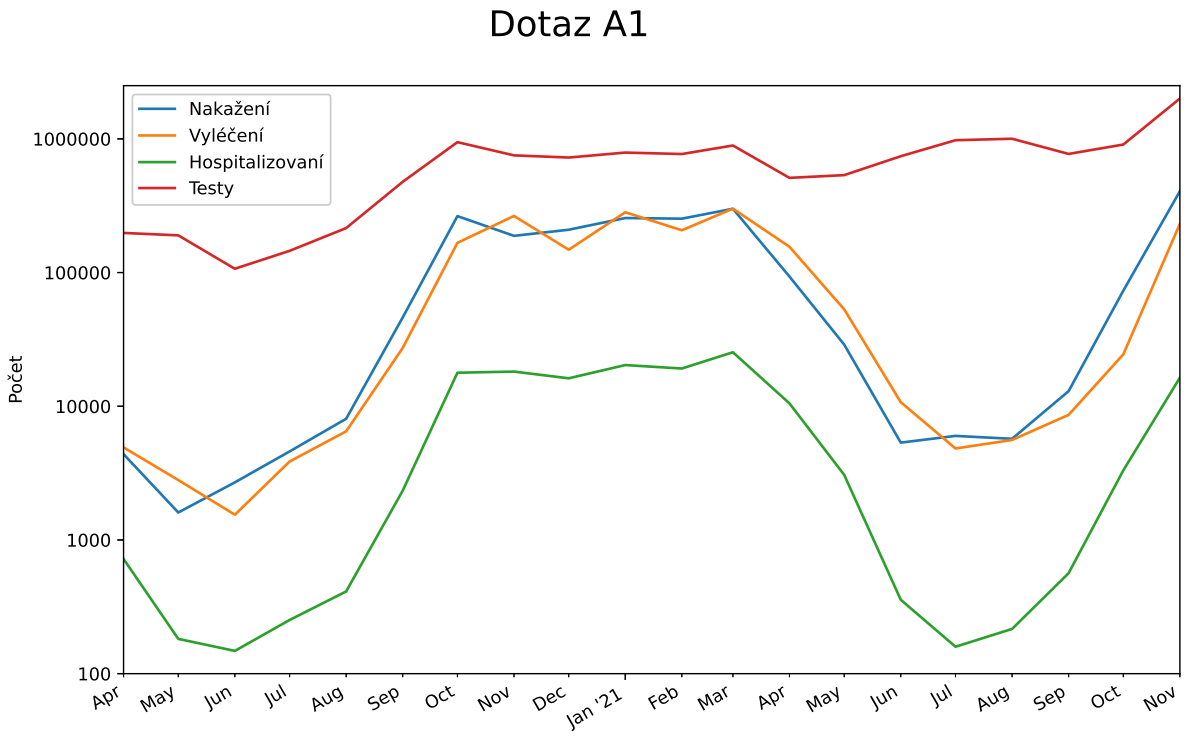
[1] Je přiloženo i řešení pro databázi MongoDB 3.6 a vyšší, které propojení kolekcí a agregaci provede jedním dotazem.

3 Řešení dotazů

[[TODO]]

Dotaz A1

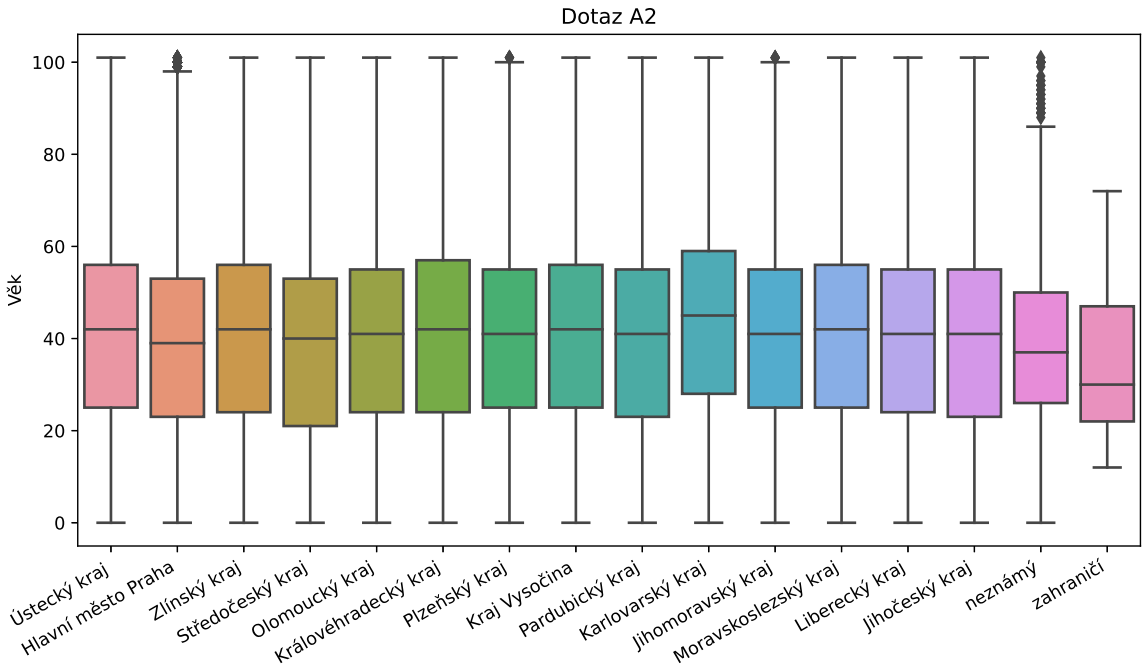
[[A1]]



Graf 1: Vývoj covidové situace po měsících

Dotaz A2

[[A2]]



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

Dotaz B1

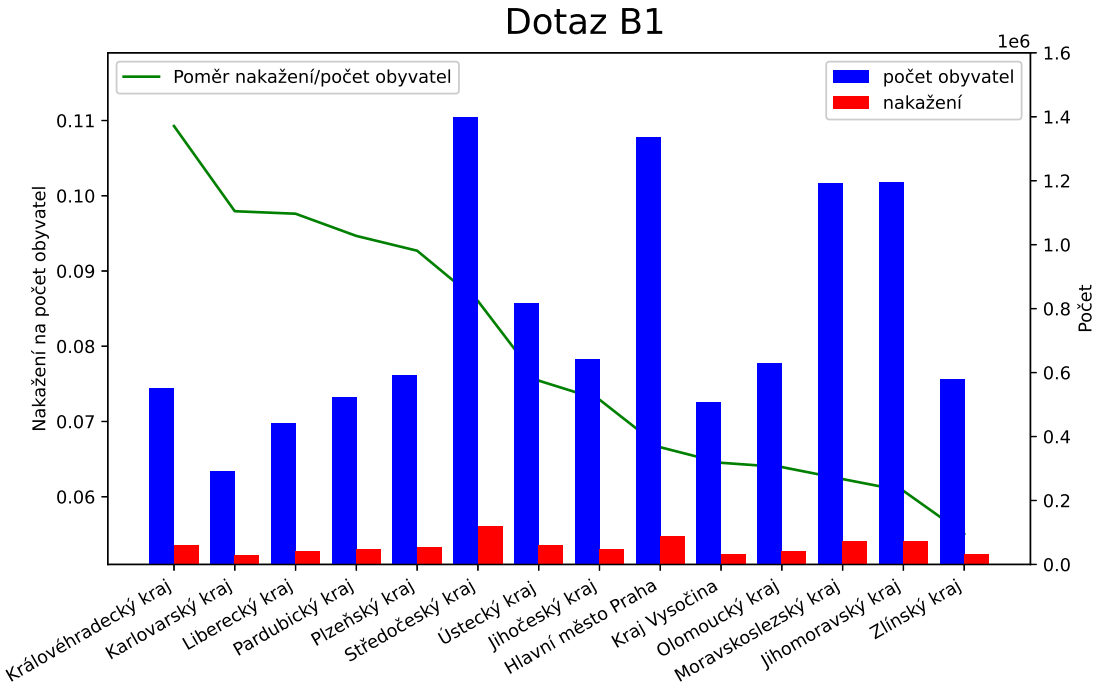
[[B1]]

	kraj_nazev	kraj_populace	nakazeni_prirustek	pomer
1	Zlínský kraj	580119	46552	0.08025
2	Královéhradecký kraj	550803	39355	0.07145
3	Kraj Vysočina	508852	35944	0.07064
4	Liberecký kraj	442476	30161	0.06816
5	Pardubický kraj	522856	34673	0.06631
6	Moravskoslezský kraj	1192834	78742	0.06601
7	Olomoucký kraj	630522	40963	0.06497
8	Jihočeský kraj	643551	38998	0.06060
9	Středočeský kraj	1397997	83961	0.06006
10	Plzeňský kraj	591041	34633	0.05860
11	Ústecký kraj	817004	46363	0.05675
12	Jihomoravský kraj	1195327	63052	0.05275
13	Hlavní město Praha	1335084	65441	0.04902
14	Karlovarský kraj	293311	13593	0.04634

	kraj_nazev	kraj_populace	nakazeni_prirustek	pomer
1	Královéhradecký kraj	550803	60185	0.10927
2	Karlovarský kraj	293311	28729	0.09795
3	Liberecký kraj	442476	43183	0.09759
4	Pardubický kraj	522856	49492	0.09466
5	Plzeňský kraj	591041	54788	0.09270
6	Středočeský kraj	1397997	120241	0.08601
7	Ústecký kraj	817004	61624	0.07543
8	Jihočeský kraj	643551	46914	0.07290
9	Hlavní město Praha	1335084	88894	0.06658
10	Kraj Vysočina	508852	32826	0.06451
11	Olomoucký kraj	630522	40318	0.06394
12	Moravskoslezský kraj	1192834	74367	0.06234
13	Jihomoravský kraj	1195327	72646	0.06078
14	Zlínský kraj	580119	31930	0.05504

	kraj_nazev	kraj_populace	nakazeni_prirustek	pomer
1	Zlínský kraj	580119	9369	0.01615
2	Jihočeský kraj	643551	9870	0.01534
3	Kraj Vysočina	508852	7550	0.01484
4	Ústecký kraj	817004	11418	0.01398
5	Moravskoslezský kraj	1192834	16119	0.01351
6	Olomoucký kraj	630522	8159	0.01294
7	Pardubický kraj	522856	5870	0.01123
8	Jihomoravský kraj	1195327	13153	0.01100
9	Liberecký kraj	442476	4842	0.01094
10	Středočeský kraj	1397997	13587	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

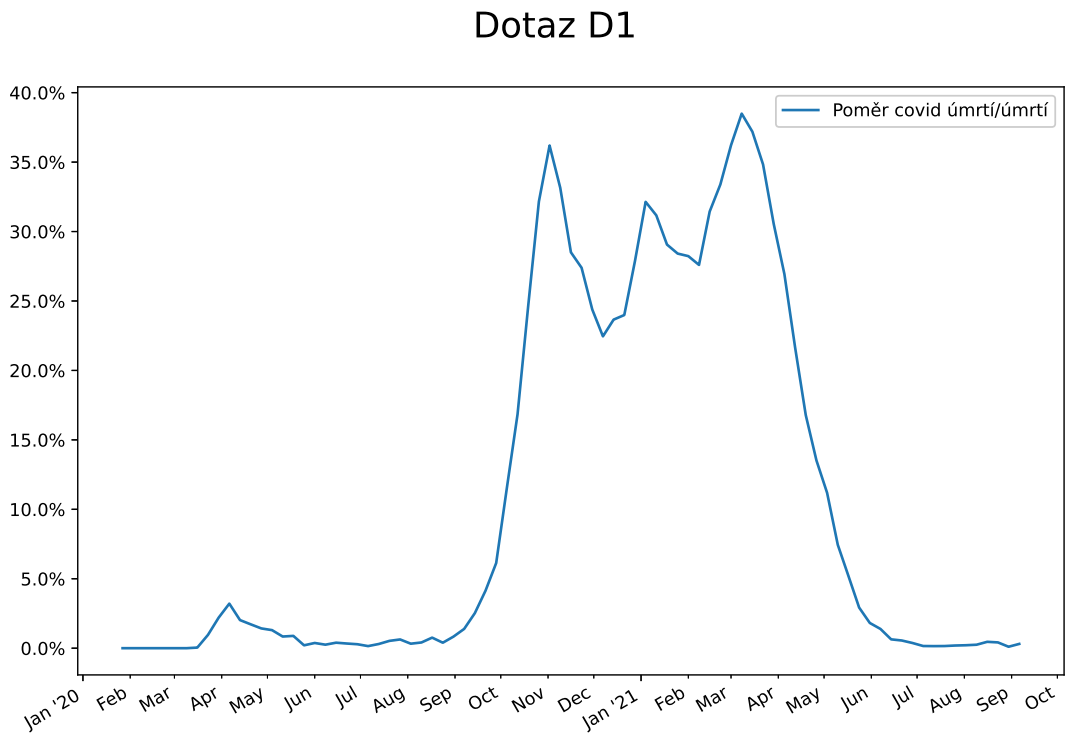
	kraj_nazev	kraj_populace	nakazeni_prirustek	pomer
1	Hlavní město Praha	1335084	6136	0.00460
2	Plzeňský kraj	591041	1655	0.00280
3	Středočeský kraj	1397997	3830	0.00274
4	Jihočeský kraj	643551	1504	0.00234
5	Moravskoslezský kraj	1192834	2700	0.00226
6	Karlovarský kraj	293311	601	0.00205
7	Jihomoravský kraj	1195327	2394	0.00200
8	Pardubický kraj	522856	892	0.00171
9	Kraj Vysočina	508852	827	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	876	0.00139
14	Královéhradecký kraj	550803	591	0.00107



Graf 3: Graf celkového počtu nově nakažených, celkového počtu obyvatel a počtu nakažených na jednoho obyvatele podle krajů pro čtvrtletí **[[obdobi]]**

Dotaz D1

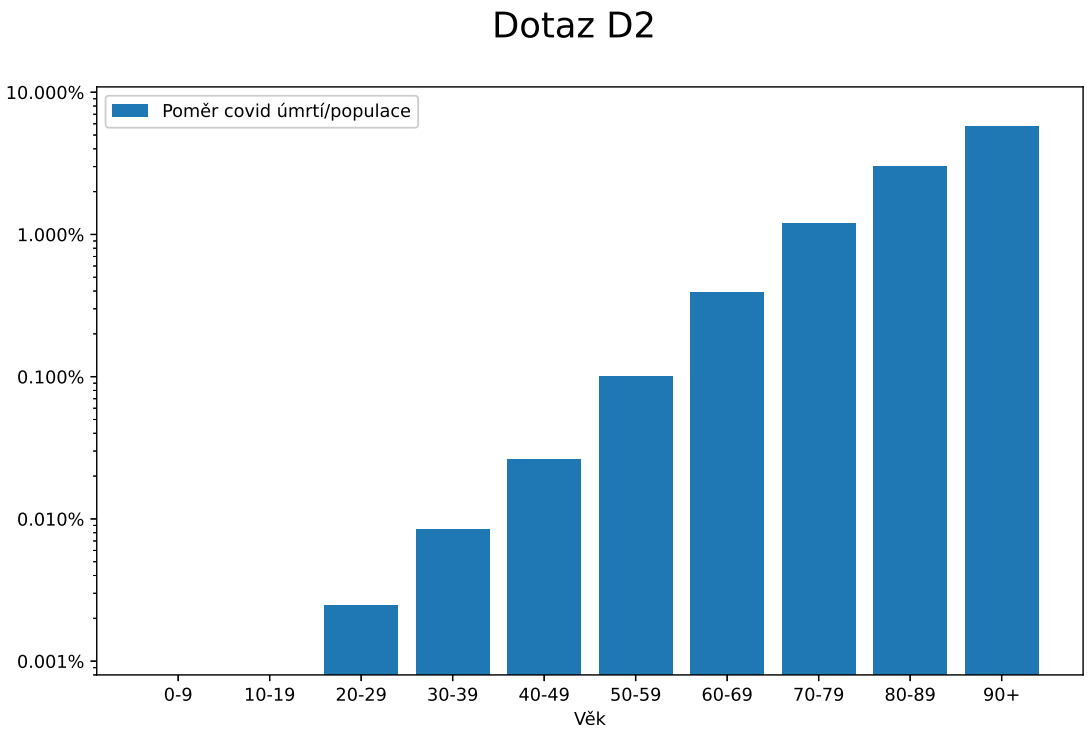
[[D1]]



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

Dotaz D2

[[D2]]



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

4 [[dolovaci uloha]]

5 Přehled obsahu odevzdaného archivu

- `csv_create.py`, `plot_graphs.py`, [[`dolovani`]] – skripty řešící druhou část projektu
- `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