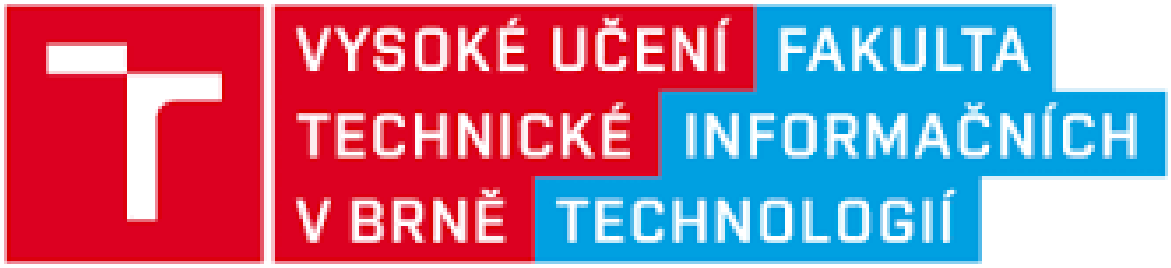
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ



UPA - Ukládání a příprava dat

Dokumentace 2. části

xberdi00

xkocim05

xsetin00

16. prosince 2021 Brno

Obsah

1 Úvod

V první části projektu jsme si zvolili téma **03: COVID-19**. Po zanalyzování dostupných zdrojů pro dané téma jsme upravili datové sady a uložili je do dokumentové databáze MongoDB pro další zpracování ve druhé části projektu. Pro řešení druhé části projektu z UPA jsme si zvolili následující úlohy:

Dotazy ze skupiny A:

- 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 .
- Série sloupcových grafů, které zobrazí:
  1. graf: počty provedených očkování v jednotlivých krajích (celkový počet od začátku očkování).
  2. graf: počty provedených očkování jako v předchozím bodě navíc rozdělené podle pohlaví. Diagram může mít např. dvě části pro jednotlivá pohlaví.
  3. graf: Počty provedených očkování, ještě dále rozdělené dle věkové skupiny. Pro potřeby tohoto diagramu postačí 3 věkové skupiny (0-24 let, 25-59, nad 59).

Dotazy ze skupiny B:

- 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.

Dotazy ze skupiny C:

- Hledání skupin podobných měst z hlediska vývoje covidu a věkového složení obyvatel.
  1. 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.
  2. Pro potřeby projektu vyberte libovolně 50 měst, pro které najdete potřebné hodnoty (můžete např. využít nějaký žebříček 50 nejlidnatějších měst v ČR).

Vlastní dotazy

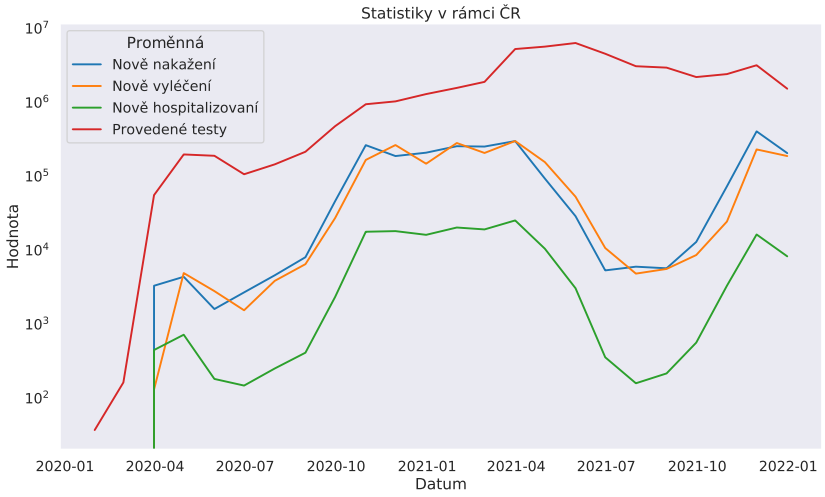
- Statistiky hospitalizovaných v rámci republiky, které zobrazí:
  1. graf: průběh nákazy u hospitalizovaných.
  2. graf: typ hospitalizace.
- Poměr počtu zemřelých na Covid a zemřelých celkově (čtvrtletně a podle věkových skupin).

Před provedením samotných dotazů a úkolů bylo potřeba uložená data v databázi exportovat do `csv` souborů. Export dat je prováděn pomocí skriptu `data_to_csv.py`. Data potřebná k provedení jednotlivých úkolů jsou uvedena v dokumentaci 1. části.

2 Výsledky 2. části

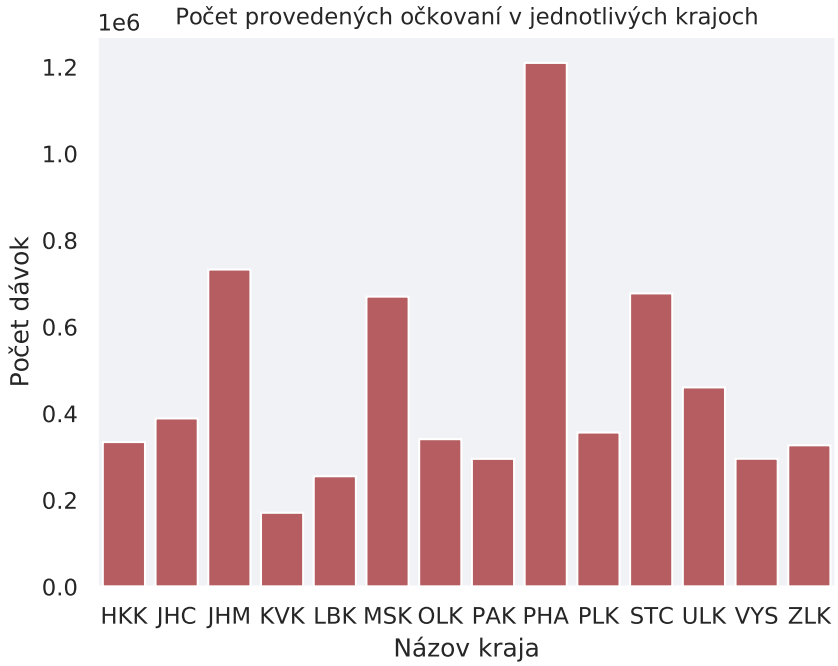
2.1 Dotazy ze skupiny A

První graf na obrázku 1 ukazuje vývoj jednotlivých covidových statistik po měsících v České republice během celé pandemie. Osa x zobrazuje jednotlivé měsíce a pro tyto měsíce jsou vykresleny hodnoty pro celkový počet nově nakažených, celkový počet nově léčených, celkový počet nově hospitalizovaných a celkový počet nově provedených testů za daný měsíc. Tyto hodnoty jsou označeny barvami, které jsou označeny v legendě grafu. Pro přehlednost jsme osu y dali do logaritmického měřítka.



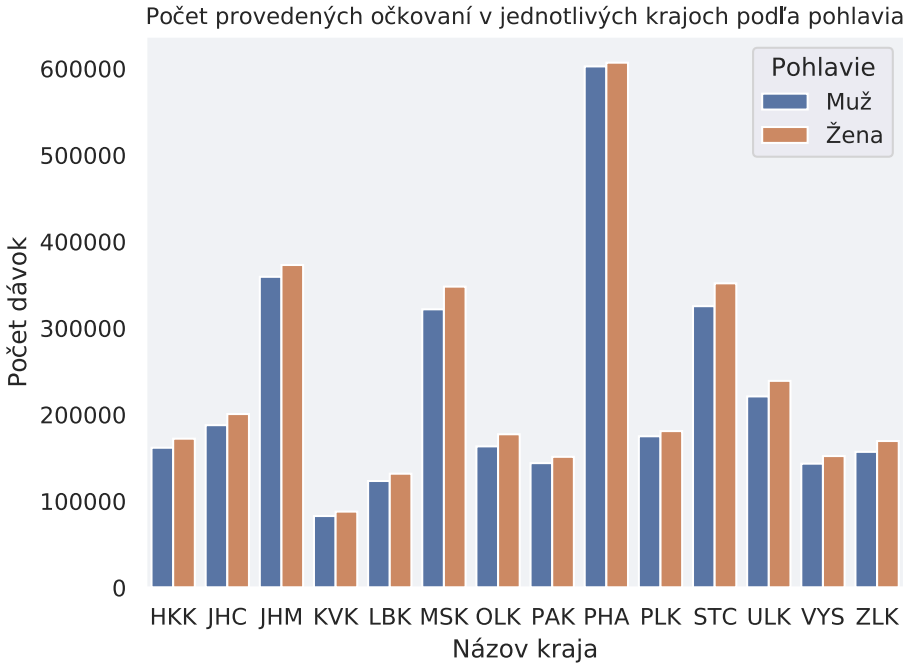
Obrázek 1: Počet nově nakažených, nově vyléčených, nově hospitalizovaných, provedených testů za období pandemie po jednotlivých měsících v logaritmickém měřítku. Použité csv soubory: `A1.csv` (schéma viz 6).

Druhý graf na obrázku 2 ukazuje počet dokončených očkování v jednotlivých krajích České republiky. Počet provedených očkování zahrnuje lidi, kteří byli očkováni jednou dávkou vakcínou od společnosti Johnson & Johnson a dvěma dávkami pro jiné vakcíny. Posilující dávky nebyly započítány. Na ose x vidíte jednotlivé kraje a na ose y počet dokončených očkování.



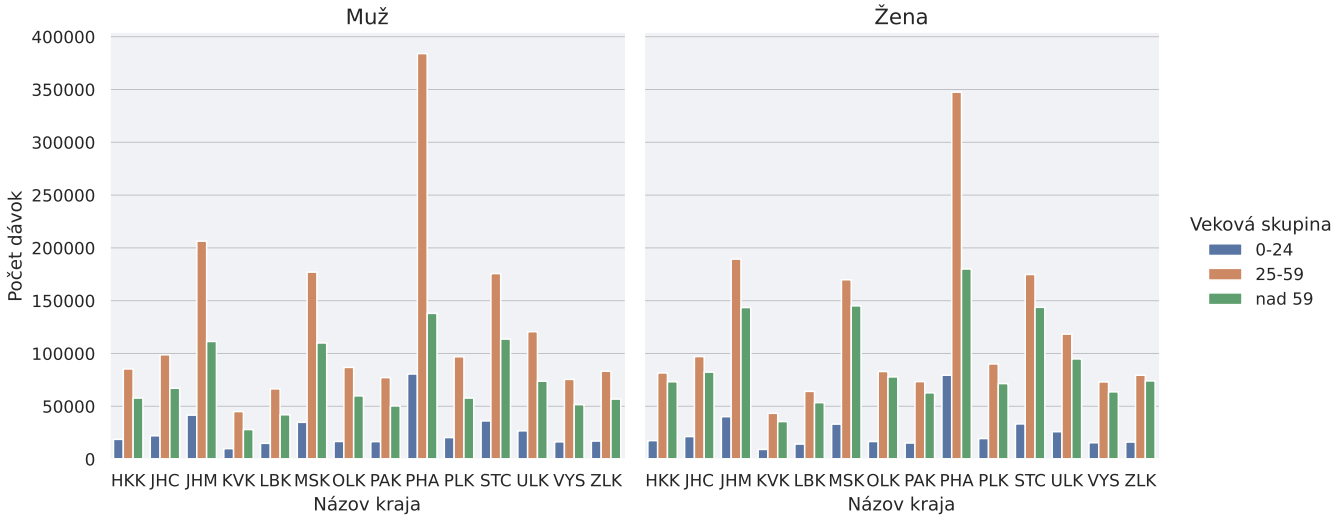
Obrázek 2: Počet provedených očkování v jednotlivých krajích. Použité csv soubory: `A3.csv` (schéma viz 6).

Třetí graf na obrázku 3 je stejný jako graf výše, ale rozdělen na pohlaví.



Obrázek 3: Počet provedených očkování v jednotlivých krajích rozdělených podle pohlaví. Použité csv soubory: `A3.csv` (schéma viz 6).

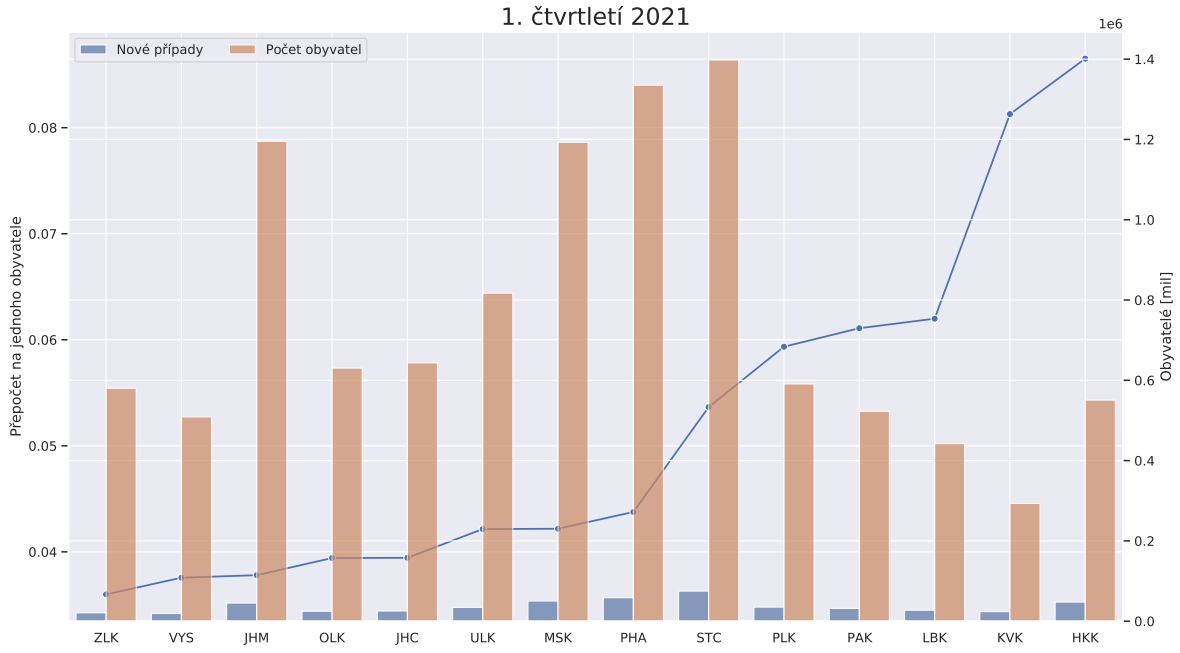
Čtvrtý graf na obrázku 4 je druhý graf rozdělený navíc na věkové skupiny. Tento graf byl již pro přehlednost rozdělen do dvou podgrafů, kde jeden podgraf ukazuje počet očkovaných v každé věkové skupině v daných krajích pro muže a v druhém podgrafu pro ženy.



Obrázek 4: Levý graf: Počet provedených očkování u můžu v jednotlivých krajích rozdělených podle věku. Pravý graf: Počet provedených očkování u žen v jednotlivých krajích rozdělených podle věku. Použité csv soubory: `A3.csv` (schéma viz 6).

2.2 Dotazy ze skupiny B

Úkolem bylo sestavit 4 žebříčky "best in covid"krajů v přepočtu nově nakažených na jednoho obyvatele. Graf prvního kvartálu je zobrazen na obrázku 5. Pro vizualizaci všech informací byl použit sloupcový graf zkombinovaný ze spojnicovým grafem. Zbylé žebříčky krajů budou vyjádřeny pouze pomocí tabulek.



Obrázek 5: Žebříček krajů ukazující počet nově nakažených přepočtený na jednoho obyvatele kraje za jedno čtvrtletí. Použité csv soubory: `B.csv, B_population.csv` (schéma viz 6).

Nyní následuje výpis jednotlivých čtvrtletních žebříčků v tabulkových formátech.

	Kraj	Nové případy	Počet obyvatel	Přepočet na jednoho obyvatele
0	ZLK	20878	580119	0.035989
1	VYS	19114	508852	0.037563
2	JHM	45186	1195327	0.037802
3	OLK	24853	630522	0.039417
4	JHČ	25381	643551	0.039439
5	ULK	34435	817004	0.042148
6	MSK	50312	1192834	0.042179
7	PHA	58424	1335084	0.043761
8	STČ	75006	1397997	0.053652
9	PLK	35076	591041	0.059346
10	PAK	31941	522856	0.061089
11	LBK	27428	442476	0.061988
12	KVK	23841	293311	0.081282
13	HKK	47650	550803	0.086510

Tabulka 1: 1. čtvrtletí 2021

	Kraj	Nové případy	Počet obyvatel	Přepočet na jednoho obyvatele
0	KVK	6772	293311	0.023088
1	HKK	16367	550803	0.029715
2	PHA	46191	1335084	0.034598
3	MSK	41826	1192834	0.035064
4	JHM	42485	1195327	0.035543
5	ZLK	21052	580119	0.036289
6	OLK	24471	630522	0.038811
7	PLK	25452	591041	0.043063
8	VYS	21933	508852	0.043103
9	STČ	61636	1397997	0.044089
10	PAK	24544	522856	0.046942
11	LBK	21474	442476	0.048531
12	ULK	40045	817004	0.049014
13	JHČ	32158	643551	0.049970

Tabulka 2: 2. čtvrtletí 2021

	Kraj	Nové případy	Počet obyvatel	Přepočet na jednoho obyvatele
0	HKK	605	550803	0.001098
1	OLK	891	630522	0.001413
2	ULK	1226	817004	0.001501
3	LBK	685	442476	0.001548
4	ZLK	914	580119	0.001576
5	VYS	853	508852	0.001676
6	PAK	902	522856	0.001725
7	JHM	2417	1195327	0.002022
8	KVK	608	293311	0.002073
9	MSK	2718	1192834	0.002279
10	JHČ	1517	643551	0.002357
11	STČ	3918	1397997	0.002803
12	PLK	1677	591041	0.002837
13	PHA	6388	1335084	0.004785

Tabulka 3: 3. čtvrtletí 2021

	Kraj	Nové případy	Počet obyvatel	Přepočet na jednoho obyvatele
0	KVK	7217	293311	0.024605
1	HKK	25860	550803	0.046950
2	ULK	38659	817004	0.047318
3	LBK	20945	442476	0.047336
4	PLK	32382	591041	0.054788
5	STČ	81825	1397997	0.058530
6	PHA	79246	1335084	0.059357
7	VYS	30667	508852	0.060267
8	PAK	33721	522856	0.064494
9	JHČ	44275	643551	0.068798
10	JHM	89629	1195327	0.074983
11	MSK	89491	1192834	0.075024
12	ZLK	44673	580119	0.077007
13	OLK	53589	630522	0.084991

Tabulka 4: 4. čtvrtletí 2021

2.3 Dotazy ze skupiny C

Jako dolovací úlohu ve skupině C jsme si zvolili hledání skupin podobných měst z hlediska vývoje covidu a věkového složení obyvatel. Na tuto dolovací úlohu jsme potřebovali libovolných 50 měst, pro které bychom uměli najít potřebná data (počet nakažených za poslední 4 čtvrtletí, počet očkovaných za poslední 4 čtvrtletí, počet obyvatel ve věkové skupině 0...14, počet obyvatel ve věkové skupině 15 ...59 a počet obyvatel nad 59 let). Počet očkovaných byl stejně jako v dotazech skupiny A počítán jako počet ukončených očkování, tzn. lidé kteří měli 1 dávku vakcíny od společnosti Johnson & Johnson nebo 2 dávky u ostatních vakcín. Posilovací dávky a započaté se nepočítaly. Formát `csv` souborů `C1-before.csv` a `C1-after.csv` je popsán v tabulce 6. Soubor `C1-before.csv` je původní neupravený csv soubor, který obsahuje hodnoty o počtu nakažených a počtu očkovaných za poslední 4 kvartály a počty pro požadované věkové skupiny pro 50 nejlidnatějších měst v České republice. Soubor `C1-after.csv` je upravený `csv` soubor pro dolovací úlohu. V tomto souboru byly ve všech sloupcích nalezeny odlehlé hodnoty a byly nahrazeny za jinou vhodnou hodnotu tento postup je popsán níže, následně byl vybrán sloupec `pocet_nakažených_Q1`, který byl normalizován a druhý sloupec `0-14` pro diskretizaci do 10 intervalů.

2.3.1 Detekce a náhrada odlehlých hodnot

K detekci odlehlých hodnot v jednotlivých sloupcích byl použit IQR (česky mezikvartilové rozpětí), které se vypočítá jako rozdíl `Q3` a `Q1`.

$$IQR = Q3 - Q1 \tag{1}$$

Následně všechny hodnoty, které nespadají do intervalu $\langle Q1 - 1.5 * IQR; Q3 + 1.5 * IQR \rangle$, jsou brány jako odlehlé hodnoty. Intervaly pro sloupce v souboru `C1-before.csv`:

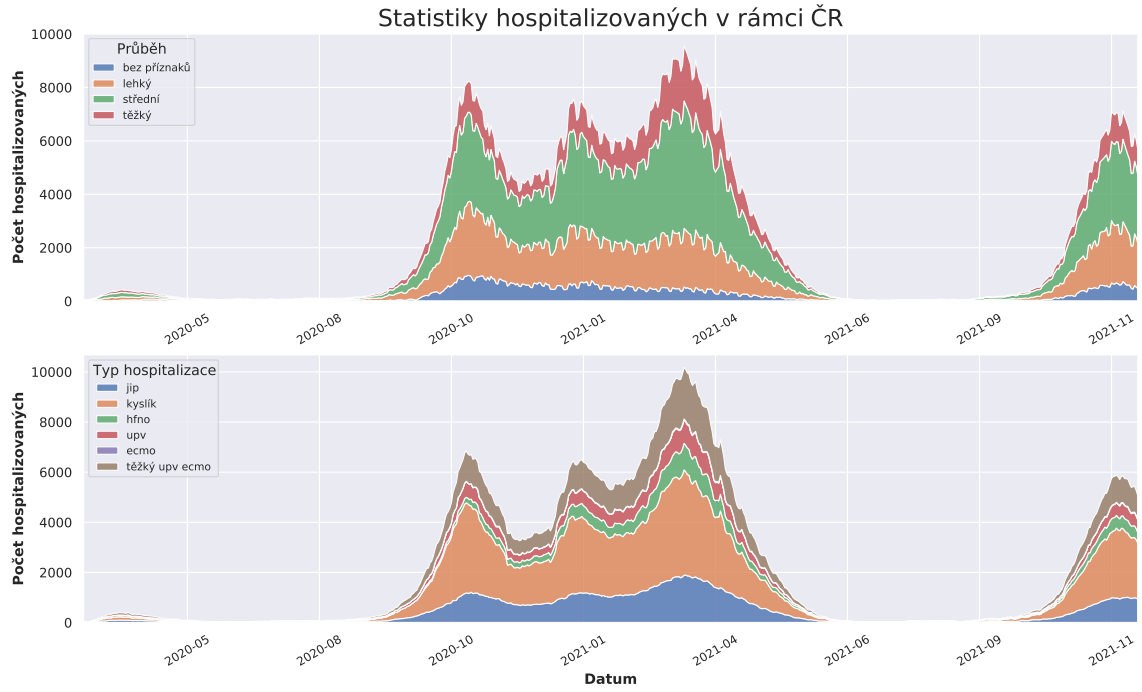
Sloupec	Interval
`pocet_nakazenych_Q1`	$\langle -1258.5; 15161.5 \rangle$
`pocet_nakazenych_Q2`	$\langle -241.25; 2444.75 \rangle$
`pocet_nakazenych_Q3`	$\langle -152.875; 510.125 \rangle$
`pocet_nakazenych_Q4`	$\langle -1364.5; 12497.5 \rangle$
`pocet_ockovanych_Q1`	$\langle -647.25; 9086.75 \rangle$
`pocet_ockovanych_Q2`	$\langle -710.0; 49072.0 \rangle$
`pocet_ockovanych_Q3`	$\langle -2080.875; 44696.125 \rangle$
`pocet_ockovanych_Q4`	$\langle -218.25; 36147.75 \rangle$
`0-14`	$\langle -956.25; 30005.75 \rangle$
`15-59`	$\langle -3947.375; 109445.625 \rangle$
`nad 59`	$\langle 1932.0; 43416.0 \rangle$

Tabulka 5: Tabulka intervalů pro hledání odlehlých hodnot

Na základě těchto intervalů se podařilo detekovat v některých sloupcích odlehlé hodnoty, které jsme následně nahradili pomocí horní a spodní hranice kvantilů. U hodnot, které byly menší než stanovený interval, jsme je nahradili hodnotou 10% kvantilu a pokud hodnota byla větší než stanovený interval nahradili jsme ji hodnotou 90% kvantilu.

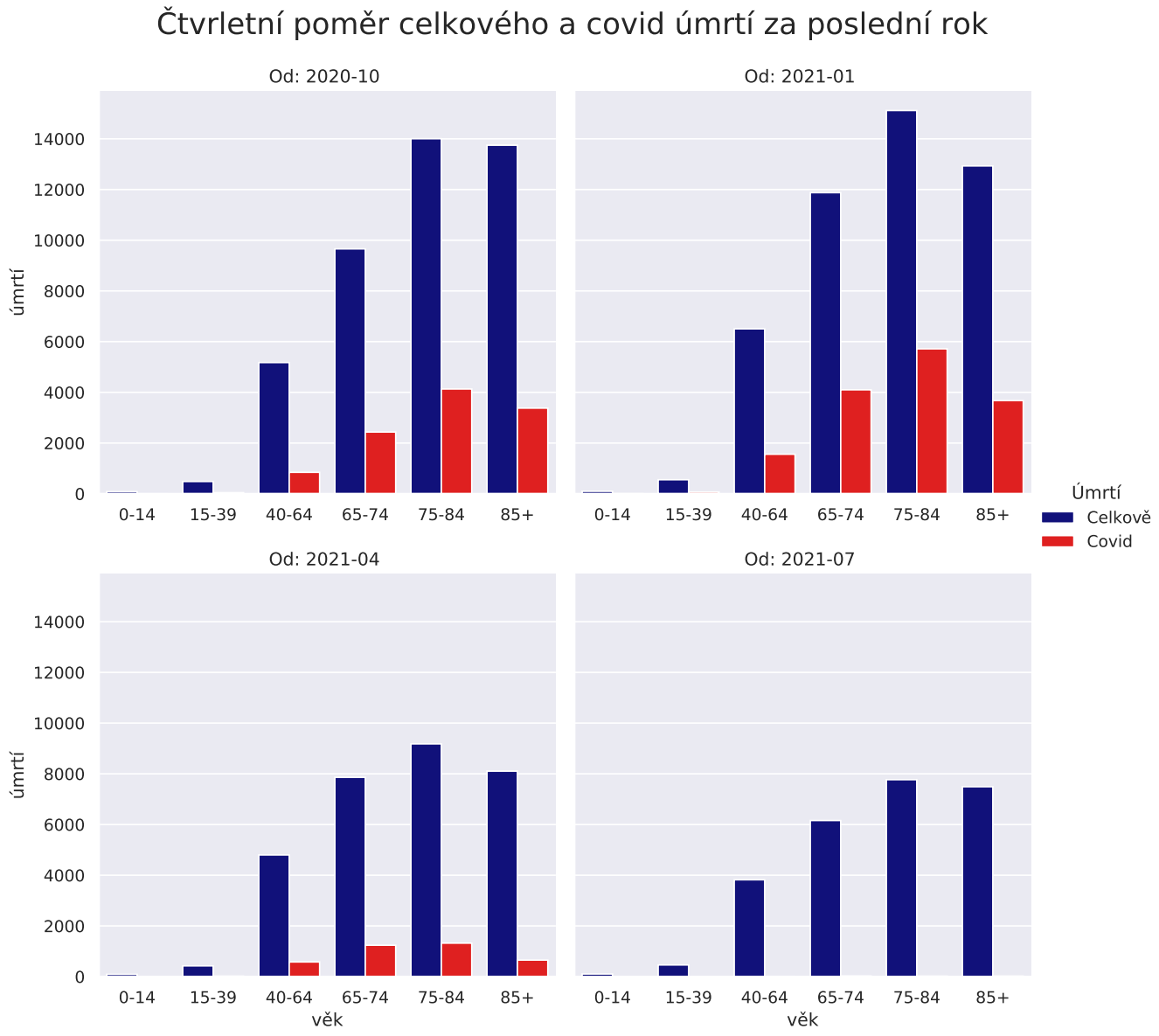
2.4 Vlastní dotazy

První vlastní dotaz bude zaměřený na hospitalizované pacienty během pandemie. Analýza hospitalizovaných bude rozdělena na dvě části (grafy). První část bude porovnávat zastoupení hospitalizovaných podle jejich průběhu nemoci. Máme 4 kategorie (bez příznaků, lehký, střední a těžký průběh). Druhá část na druhou stranu srovná hospitalizované podle způsobu hospitalizace (jip, kyslík, hfno, upv, ecmo, težký upv ecmo). Pro vytvoření výsledných grafů potřebuje pouze přehled hospitalizovaných. Ideálním grafem pro výše uvedené závislosti je kombinovaný spojnicový graf. Grafy jsou vizualizovány na obrázku 6. Osa x bude vždy vyjadřovat časové úseky pandemie a osa y vykazuje počet hospitalizovaných. Kategorické atributy průběhu nemoci nebo typu hospitalizace byly odlišeny pomocí barev, jejichž význam ukazují legendy.



Obrázek 6: Horní graf: průběh nákazy u hospitalizovaných. Dolní graf: typ hospitalizace. Použité csv soubory: `Custom_1_hospitalized.csv` (schéma viz 6).

Druhý vlastní dotaz zkoumá poměr mezi celkovým a covid úmrtím za poslední rok pandemie. Při řešení byly zohledněny i věkové kategorie zemřelých. Pro realizaci dotazu byly potřeba dvě datové sady, jedna obsahující celkové úmrtí a druhá s evidovaným covid úmrtím. Data byla uložena po týdenních intervalech. Kvůli zachování přehlednosti se data zagregovala na čtvrtletní období. Při vizualizaci jsme se zaměřili hlavně na poslední rok pandemie, kde byl poměr úmrtí v největším kontrastu. Množinu grafů můžeme vidět na obrázku 7. Každý graf vyjadřuje 3 měsíce a následně porovnává úmrtí a jednotlivé věkové skupiny. Celkově je zobrazeno období od začátku října 2020 do konce září 2021. Poslední graf je jen do září 2021, z důvodu neaktuálních dat pro celková úmrtí.



Obrázek 7: Poměr počtu zemřelých na Covid a zemřelých celkově (čtvrtletně a podle věkových skupin). Použité csv soubory: `Custom_2_total_deaths.csv, Custom_2_covid_deaths.csv` (schéma viz 6).

2.5 Formát csv souborů

Název souboru	**Schéma souboru**
A1.csv	*datum, nakazenych, vyliecenych, umrti, pcr_testov, ag_testov, hospitalizovany*
A3.csv	*kraj_nazev, vakcina, poradi_davky, vekova_skupina, pohlavi, pocet_davek*
B.csv	*datum, orp_kod, orp_nazev, kraj_nazev, nove_pripady*
B_population.csv	*vuzemi_txt, hodnota*
C1-after.csv	*orp_nazev, pocet_nakazenych_Q1, pocet_nakazenych_Q2, pocet_nakazenych_Q3, pocet_nakazenych_Q4, pocet_ockovanich_Q1, pocet_ockovanich_Q2, pocet_ockovanich_Q3, pocet_ockovanich_Q4, 0-14, 15-59, nad 59*
C1-before.csv	*orp_nazev, pocet_nakazenych_Q1, pocet_nakazenych_Q2, pocet_nakazenych_Q3, pocet_nakazenych_Q4, pocet_ockovanich_Q1, pocet_ockovanich_Q2, pocet_ockovanich_Q3, pocet_ockovanich_Q4, 0-14, 15-59, nad 59*
Custom_1_hospitalized.csv	*id, datum, pocet_hosp, stav_bez_priznaku, stav_lehky, stav_stredni, stav_tezky, jip, kyslik, hfno, upv, ecmo, tezky_upv_ecmo*
Custom_2_total_deaths.csv	*celkove_umrti, datum, vek_txt*
Custom_2_covid_deaths.csv	*datum, vek_txt, covid_umrti*

Tabulka 6: Schéma csv souborů