Zvolené téma: 4 - COVID-19
Řešitelé: Kateřina Fořtová (xforto00), Jakub Kolb (xkolbj00), Ondřej Pavela (xpavel34)

Zvolené dotazy a formulace vlastního dotazu:

- **Dotaz skupiny 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 skupiny B** - určete vliv epidemie COVID-19 na počet zemřelých v porovnání dle počtu nemocných, počtu hlášených úmrtí na nemoc COVID-19 a v porovnání s minulými lety
- **Vlastní dotaz** - zobrazte vývoj počtu testů a počtu vyléčených, dále zobrazte počty mrtvých dle krajů

Stručná charakteristika zvolené datové sady:

(Zde konkrétně popište jaké soubory budou představovat zdroj dat pro zvolené úlohy. Dále popište, jakým způsobem budou tato data získána a stručně charakterizujte strukturu souborů vybraných pro řešení projektu. Zaměřte se na části souborů, které jsou důležité pro zodpovězení zvolených dotazů.)

Dotaz skupiny A:

Pro dotaz skupiny A bude využito datové sady, kterou poskytuje Ministerstvo zdravotnictví ČR (https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/osoby.json). Tato datová sada je poskytována ve formátech *JSON* a *CSV*, přičemž v našem případě jsme zvolili formát *JSON*. Datový soubor se poté skládá z hlavičky a ze samotného seznamu osob s prokázanou nákazou. Hlavička obsahuje datum a čas poslední modifikace, které využijeme pro odlišení nových záznamů při aktualizacích naší NoSQL databáze.
Formát pro jednu konkrétní osobu s potvrzenou nákazou je následující:

```
{
	"datum": "2020-10-07",
	"vek": 55,
	"pohlavi": "Z",
	"kraj_nuts_kod": "CZ072",
	"okres_lau_kod": "CZ0721",
	"nakaza_v_zahranici": false,
	"nakaza_zeme_csu_kod": ""
}
```

Položka `"datum"` udává, kdy byla nákaza prokázána krajskou hygienickou stanicí.
Položky **`'kraj_nuts_kod'`** a **`'okres_lau_kod'`** udávají okres a jeho nadřízený kraj, v němž byla nákaza prokázána. Položka udávající kraj je zde redundantní v případě, že máme k dispozici informaci o územním dělení České republiky. Poslední dvě položky

`"nakaza_v_zahranici"` a `"nakaza_zeme_csu_kod"`, indikují, zda došlo k nákaze v zahraničí a pokud ano, tak v jaké zemi, dle kódu z číselníku ČSÚ.

Dotaz skupiny B:

Pro dotaz skupiny B je nutno využít více datových sad. První z nich je opět sada poskytovaná MZČR (https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/umrti.json), která je dostupná ve formátech *JSON* a *CSV*. Opět jsme zvolili formát *JSON*, který má stejnou strukturu jako sada z dotazu A.

Formát jedné datové položky je následující:

```
{
        "datum": "2020-03-24",
        "vek": 44,
        "pohlavi": "M",
        "kraj_nuts_kod": "CZ080",
        "okres_lau_kod": "CZ0802"
}
```

Položka `"datum"` udává, kdy bylo úmrtí zaznamenáno. Význam zbylých položek je obdobný jako v předchozím případě.

Dále bude nutné využít datovou sadu z jiného zdroje, která nám bude podávat informace o vývoji mortality v předešlých letech v ČR. Pro tuto skutečnost jsme využili datové sady poskytované Českým statistickým úřadem, která udává týdenní počty zemřelých pro každý týden v roce již od roku 2011 (https://www.czso.cz/documents/62353418/138258837/130185-20data092920.csv). Tyto data jsou zapisována do datové sady s určitým zpožděním (např. v říjnu 2020 máme poslední data dostupná počátkem srpna 2020). Vzhledem ke skutečnosti, že tato datová sada je dostupná pouze ve formátu *CSV*, bylo potřeba vynaložit více úsilí při předzpracování, abychom získali jednotlivé datové objekty ve slovníkovém formátu. Formát datové sady je následující:

```
idhod,hodnota,stapro_kod,vek_cis,vek_kod,vuzemi_cis,vuzemi_kod,rok,tyden,roktyden,casref_od,casref_do,vek_txt
832459719, 9, 5393, 7700, 400000610015000, 97, 19, 2011, 01, 2011-W01, 2011-01-03, 2011-01-09, 0-14
832460200, 55, 5393, 7700, 410015610040000, 97, 19, 2011, 01, 2011-W01, 2011-01-03, 2011-01-09, 15-39
832460678, 457, 5393, 7700, 410040610065000, 97, 19, 2011, 01, 2011-W01, 2011-01-03, 2011-01-09, 40-64
832461162, 1151, 5393, 7700, 410065610085000, 97, 19, 2011, 01, 2011-W01, 2011-01-03, 2011-01-09, 65-84
832461643, 577, 5393, 7700, 410085799999000, 97, 19, 2011, 01, 2011-W01, 2011-01-03, 2011-01-09, 85 a více
832462584, 2249, 5393, , , 9, 19, 2011, 01, 2011-W01, 2011-01-03, 2011-01-09, celkem
```

Stručný popis významu jednotlivých sloupců:

- **idhod:** ČSÚ identifikátor záznamu
- **hodnota:** počet zemřelých
- **stapro_kod:** kód statistické proměnné, zde pouze 5393
- **vek_cis:** číselník pro věkovou skupinu, zde pouze 7700
- **vek_kod:** kód položky číselníku pro věkovou skupinu
- **vuzemi_cis:** kód číselníku pro referenční území, pouze 93 (stát)
- **vuzemi_kod:** kód položky číselníku pro referenční území, pouze 19 (ČR)
- **rok:** rok referenčního období ve formátu RRRR
- **tyden:** pořadové číslo referenčního týdne (dle normy ISO)

- **roktyden:** referenční rok a týden ve formátu RRRR-Wxx (dle normy ISO)
- **casref_od:** datum pondělí referenčního týdne ve formátu RRRR-MM-DD
- **casref_do:** datum neděle referenčního týdne ve formátu RRRR-MM-DD
- **vek_txt:** text udávající věkovou skupinu, 0-14, 15-39, 40-64, 65-84, 85+

Hodnota v posledním řádku daného referenčního týdne vždy udává celkový úhrn úmrtí za všechny věkové skupiny.

Vlastní dotaz

Jako vlastní dotazy jsme si vybrali zobrazení vývoje testování a vyléčených případů na nemoc COVID-19. Dále jsme si vybrali zobrazení počtu mrtvých podle jednotlivých krajů. Pro zjištění počtu testů v jednotlivých dnech jsme využili datovou sadu opět od MZČR (https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/testy.json). Jedná se znovu o data v preferovaném formátu *JSON*, které pro náš projekt byly nejsnazší na zpracování. Níže je ukázka jednoho datového záznamu:


```
{
            "datum": "2020-01-29",
            "prirustkovy_pocet_testu": 5,
            "kumulativni_pocet_testu": 33
}
```


Pro náš graf jsme využili údaj přírustkového počtu testů za dané datum.
Pro informace o vyléčených jsme využili kumulativní statistiky opět dostupné ze stejného zdroje:
(https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/nakazeni-vyleceni-umrti-testy.json). Tato datová sada je opět ve formátu *JSON*. Následuje ukázka jednoho záznamu z této datové sady:


```
{
            "datum": "2020-07-08",
            "kumulativni_pocet_nakazenych": 12829,
            "kumulativni_pocet_vylecenych": 9771,
            "kumulativni_pocet_umrti": 353,
            "kumulativni_pocet_testu": 584946
 }
```


Položky `"datum"` a `"kumulativni_pocet_vylecenych"` budou sloužit jako údaje na základě kterých zobrazíme finální graf.

Pro poslední vlastní dotaz zobrazující počet úmrtí dle jednotlivých krajů jsme využili stejně jako u dotazu B datovou sadu poskytovanou MZČR (https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/umrti.json). Opět jsme využili informací o územním dělení ČR a na základě daného kraje máme tedy k dispozici všechny potřebné informace pro vykreslení grafu.

Získání a zpracování dat

Pro získání a zpracování datových sad jsme zvolili skriptovací jazyk *Python*, jelikož umožňuje jednoduchou práci se soubory formátu *JSON,* které se zde přímo mapují na vestavěné slovníky. Kromě toho existuje pro tento jazyk mnoho vyspělých knihoven pro jednoduché stahování souborů z internetu. Konkrétně jsme zvolili knihovnu *requests* (https://requests.readthedocs.io/en/master/), která podporuje vše, co jsme pro naše účely potřebovali.
Pro datové sady poskytované od MZČR zahrnovalo následné zpracování pouze vyřazení vybraných datových položek (viz dále) a vyfiltrování záznamů na základě data v případě aktualizace již existující databáze.
Při zpracování datové sady od ČSÚ ve formátu *CSV* bylo potřeba data výrazněji transformovat, aby následně odpovídala navrženému schématu naší NoSQL a relační databáze. Datový soubor je proto zpracovávan po jednotlivých týdnech jako celcích. Nejprve jsou vyřazeny nepotřebné sloupce: **idhod, stapro_kod, vek_cis, vek_kod, vuzemi_cis, vuzemi_kod, roktyden, vek_txt**. Následně je všech 6 záznamů za jeden konkrétní týden sloučeno do jednoho datového slovníku, který poté obsahuje počty úmrtí všech věkových skupin (celkový počet je možné dopočítat a není proto zahrnut).

Při práci s daty jsme upřednostňovali data ve formátu JSON díky snadnému stahování a získávání potřebných údajů pomocí skriptu napsaném v jazyce Python.

Zvolený způsob uložení uložení surových dat:

(Zde stručně charakterizujte NoSQL databázi, která bude využita pro uložení zvolených zdrojových dat.)

Pru uložení nestrukturovaných data jsme si zvolili dokumentovou NoSQL databázi MongoDB. Pro naše potřeby jsme poté vytvořili tři datové kolekce pro uložení datových sad od MZČR:

- **`rho_confirmed_cases`:** zde jsou uloženy veškeré osoby s potvrzenou nákazou. Nejprve byly vyřazeny nepotřebné datové položky (v našem případě pouze kód kraje) z jednotlivých záznamů osob reprezentovaných slovníky a výsledný seznam slovníků byl poté uložen do kolekce, což bylo velmi jednoduché, díky přímému mapování Python slovníku na MongoDB záznam.
- **`covid_deaths`:** kolekce uchovávající záznamy o úmrtích na COVID-19. Zpracování je obdobné jako v předchozím případě (opět byl vyřazen pouze kód kraje z jednotlivých záznamů)
- **`daily_statistics`:** kolekce sdružující více MZČR datových sad s charakterem časové řady. Konkrétně jsou zde uloženy záznamy z následujících zdrojů:
    - https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/testy.json
    - https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/nakaza.json
    - https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/nakazeni-vyleceni-umrti-testy.json

Záznamy bylo možné propojit na základě sdílené datové položky **'datum'**, která je v tomto případě jednoznačným identifikátorem. Pro každé datum byl vytvořen jeden záznam, který poté sdružoval vybrané datové položky ze všech sad, konkrétně:

- **'prirustkovy_pocet_testu'**
- **'prirustkovy_pocet_nakazenych'**
- **'kumulativni_pocet_vylecenych'**
- **'kumulativni_pocet_umrti'**

Je zřejmé, že tyto datové sady mají charakter časové řady, nicméně se nám nechtělo kombinovat více NoSQL databází pro různé části dat. Navíc je datová sada rozšiřována s frekvencí jednoho záznamu na jeden den, což výrazně limituje její velikost a využití specializované NoSQL databáze typu InfluxDB je proto dle nás v takovém případě prozatím neopodstatněné.

- **weekly_deaths:** zde je uložena datová sada od ČSÚ, jejíž způsob předzpracování byl popsán v předchozí sekci. Výsledné datové slovníky, jsou zde pouze uloženy do kolekce.

Vzhledem k tomu, že veškerá data o osobách jsou anonymní a neobsahují žádné unikátní identifikátory, nebylo příliš mnoho možností, na základě čeho jednotlivé datové sady propojit. Stále je ovšem možné jednoznačně propojit záznamy na základě časového rozmezí při tvorbě anonymních statistik, což je v našem případě dostačující.
Absence unikátních identifikátorů nicméně představuje problém při pozdější aktualizaci některých kolekcí. Konkrétně datové sady poskytované MZČR jsou průběžně aktualizovány v průběhu dne, což by samo o sobě ještě nebylo problémem, jelikož jsme při importu schopni ignorovat záznamy z aktuálního dne. Bohužel jsou v mnoha případech zpětně vkládány starší záznamy i s několikadenním zpožděním. To má za následek to, že nejsme při aktualizaci vlastní databáze schopni rozpoznat nově vložené záznamy s datem starším než je datum poslední aktualizace. Tím je do určité míry porušena konzistence a jediná možnost, jak tento problém řešit je opětovný import celé datové sady.

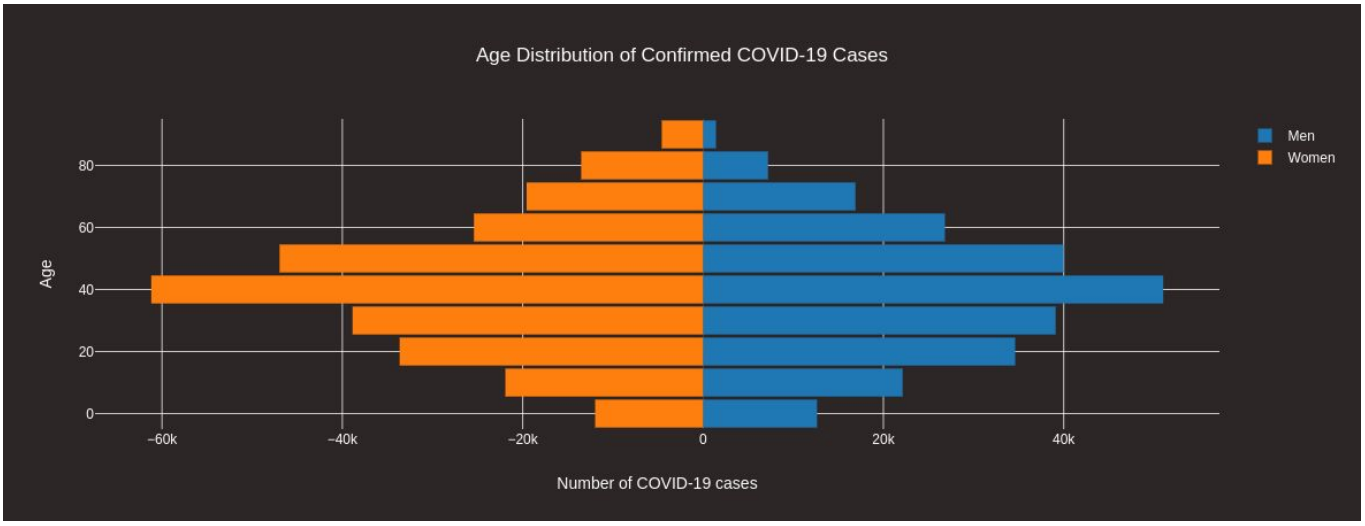
Druhá část projektu - Výsledky řešení dotazů

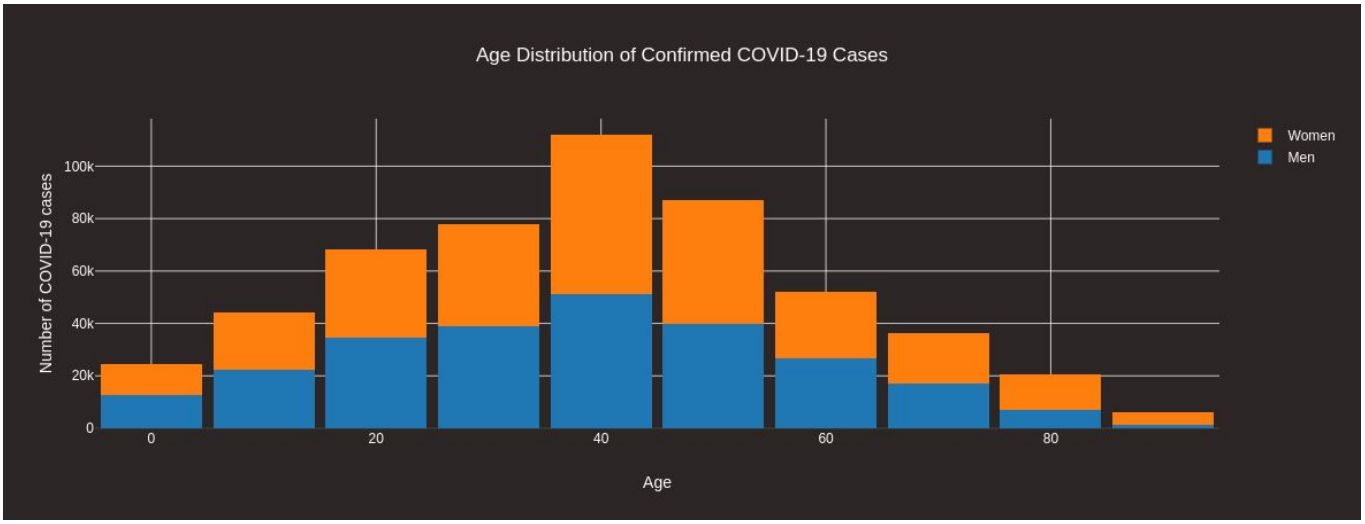
Následující část dokumentace pojednává o výsledcích zvolených dotazů. Pro vizualizaci všech dotazů jsme využili knihovnu Plotly. Ve svých výsledcích jsme se snažili využít více typů grafů. Níže jsou výsledky dotazů zobrazeny s popisem informací, které můžeme z daných grafů zjistit o různých aspektech problematiky pandemie COVID-19.

Dotaz skupiny A

Rozdělení nakažených na základě věku a pohlaví
Pro vizualizaci členění nakažených dle prvních dvou charakteristik - věku a pohlaví - jsme zvolili skládaný pruhový a sloupcový graf. Povedlo se nám tedy do jednoho grafu

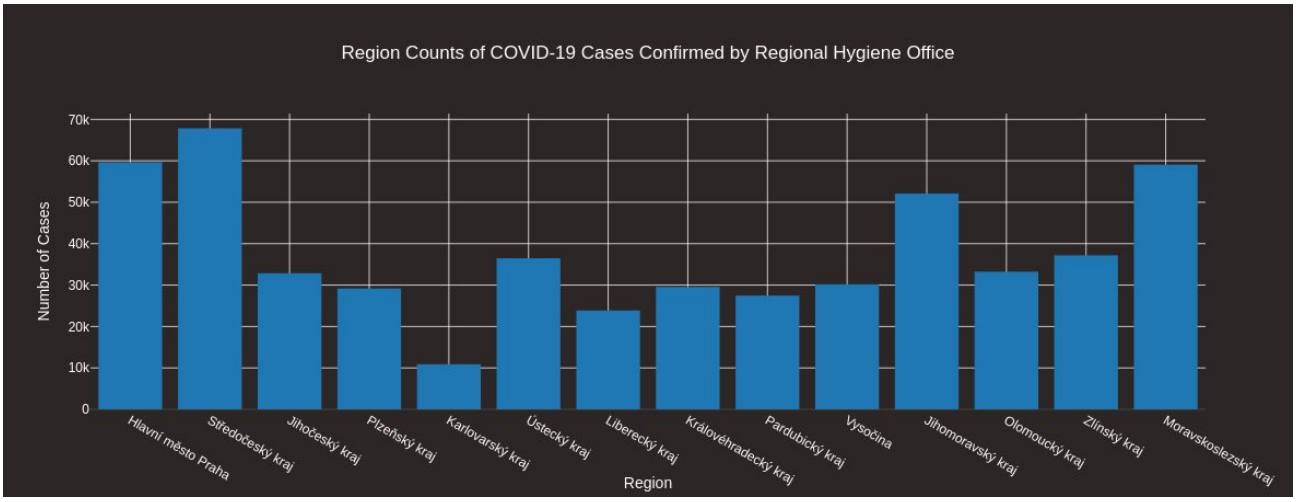
začlenit rozdělení dle obou vlastností. Z grafů můžeme tedy vidět, že nejvíce nakažených mužů i žen je ve věkovém rozmezí 40 - 50 let. Co se týče důchodců od 80 let tak převládá vyšší počet nakažených žen.

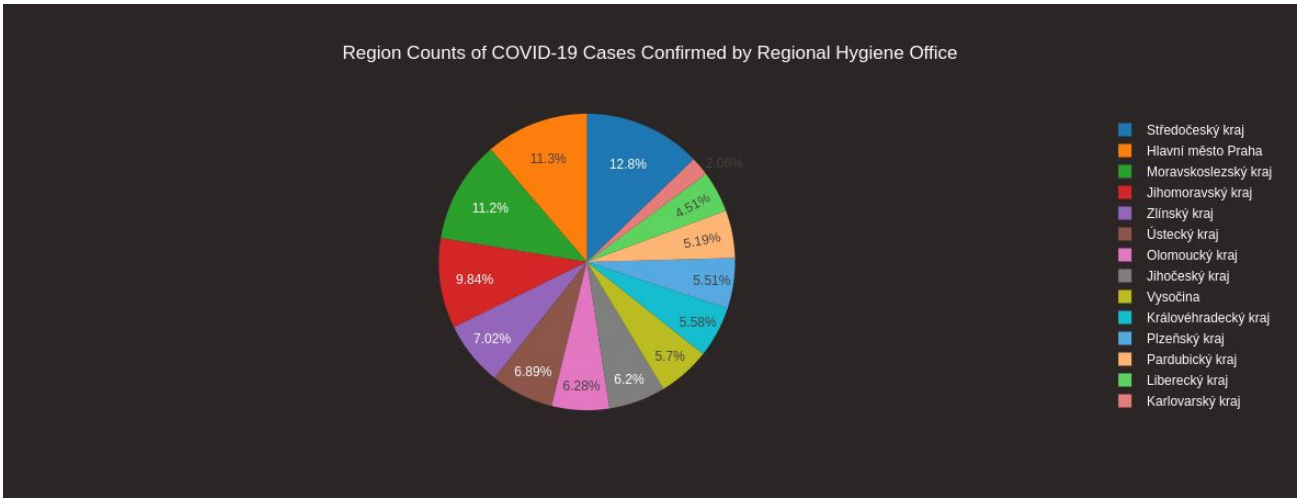




Rozdělení nakažených na základě kraje, kde byla nákaza evidována

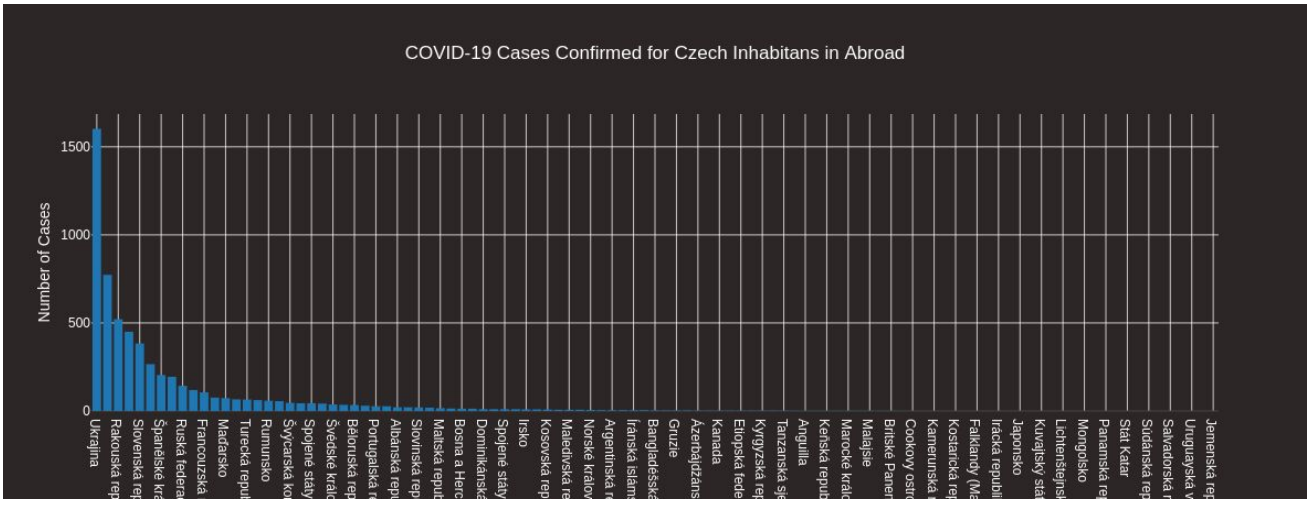
Pro vizualizaci byl vytvořen sloupcový a koláčový graf. Koláčový graf nám zde lépe reprezentuje procentuální zastoupení nakažených pro každý kraj. Vidíme, že nejvíce nakažených má Středočeský kraj, Hlavní město Praha a Moravskoslezský kraj, nejméně nakažených pak Karlovarský kraj. Pro oba grafy bylo potřeba vyfiltrovat data, která mají kódy neznámých regionů a do vizualizace jsme tedy zahrnuli pouze případy, které mají známý region nákazy.



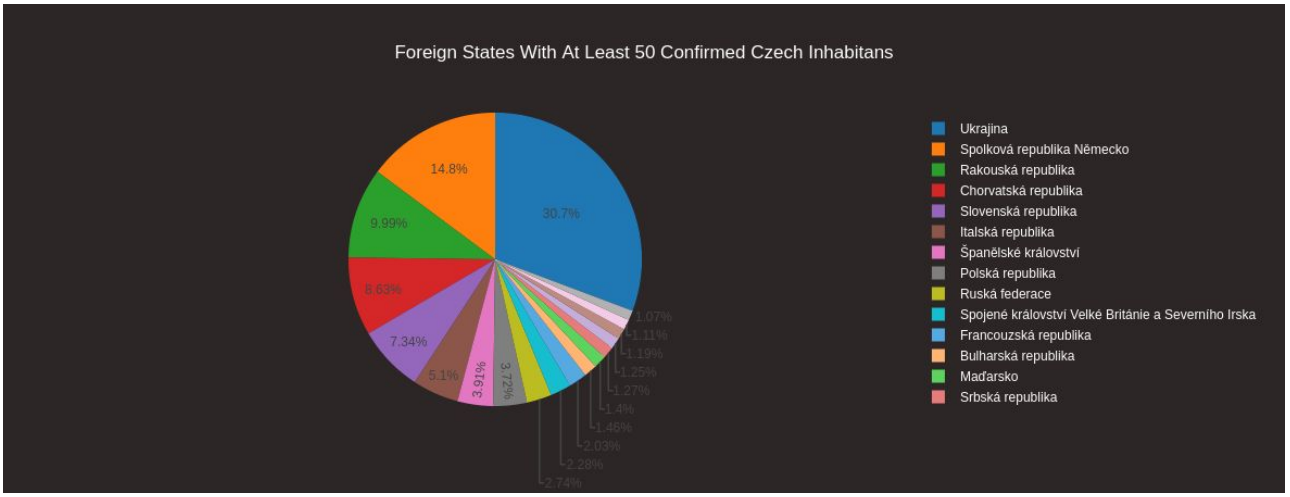


Rozdělení nakažených na základě nákaz v zahraničí a výpis cizích zemí, kde se nakazilo minimálně 50 Čechů

Pro první sloupcový graf bylo potřeba vyfiltrovat množství zahraničních zemí, ve které se žádný z Čechů nenakazil, což napomáhá i samotné přehlednosti grafu. Vidíme však i tak, že mnoho zemí má pouze malé množství evidovaných nakažených Čechů, často se jedná pouze o jednotky případů. Na druhou stranu velmi velký podíl nakažených má Ukrajina ve srovnání s ostatními zeměmi.



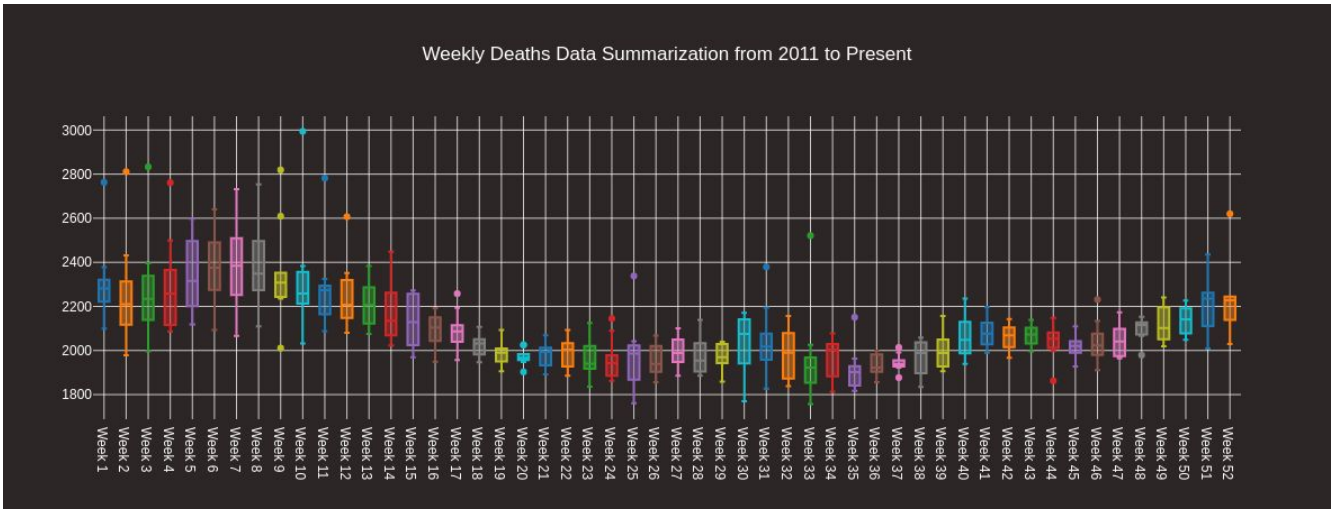
Pro stále vysoké množství dat vyskytující se v prvním grafu i po vyfiltrování zemí s žádným nakaženým Čechem jsme vytvořili koláčový graf, který ukazuje země, kde se nakazilo aspoň 50 Čechů. A vidíme poměrově tedy tyto zahraniční země s největším počtem nakažených obyvatel ČR. Největší procentuální zastoupení zde tvoří již zmíněná Ukrajina a dále Německo a Rakousko.



Dotaz skupiny B

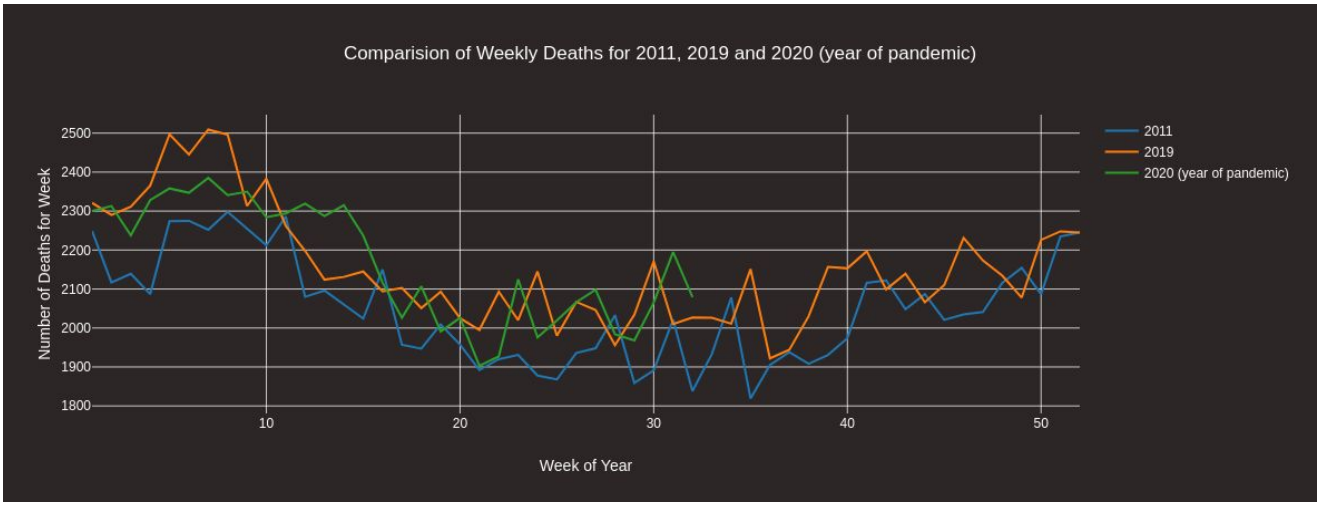
Krabicový graf ukazující týdenní mortalitu od roku 2011 do 2020

Data pro týdenní mortalitu jsou zaznamenávána ve větších časových rozestupech. Dnes (30.11.2020) máme poslední data pro 32. druhý týden tohoto roku (tedy z tohoto srpna). Avšak minimálně nějaký překryv dat máme. Nejvyšší rozpětí mortality od roku 2011 do 2020 je k začátku roku (zejména 4., 5. a 7 týden). Z údajů v tomto grafu a v grafu srovnání let 2011, 2019 a 2020 však vidíme, že minimálně do srpna letošního roku se rok 2020 příliš neodlišoval od jiných, co se týče například velmi vysokého přírůstku mortality vzhledem k pandemii. Bude však určitě zajímavější sledovat vývoj mortality v druhé části letošního roku, až budeme mít dostupné dané informace v datové sadě.



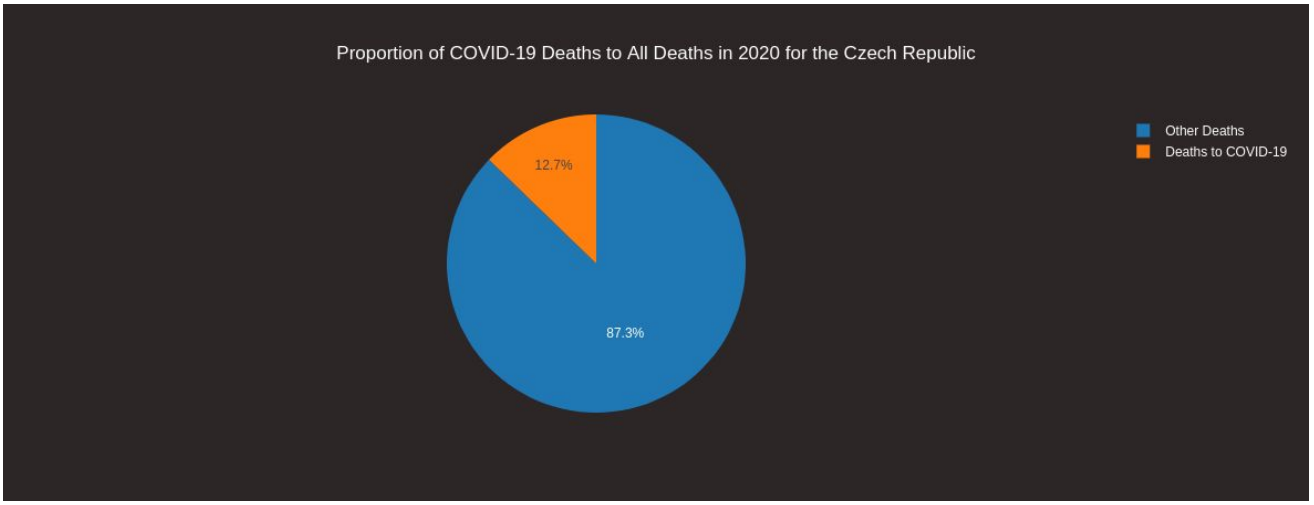
Graf zobrazující porovnání týdenní mortality pro rok 2011, 2019 a 2020 (rok pandemie)

V následujícím grafu byly porovnány data týdenní mortality pro námi zvolené roky. Prvním z nich byl rok 2011, kdy pro tento rok máme díky datové sadě dostupná nejstarší data z dané datové sady. Dalším rokem je 2019, tedy minulý rok, kdy Česko nebylo ještě postihnuto koronavirovou epidemií. A posledním rokem je 2020, tedy rok pandemie. Ačkoliv nemáme ještě známé všechny data pro týdenní mortalitu tohoto roku, můžeme vidět, že minimálně v první části tohoto roku byla týdenní mortalita srovnatelná minimálně s minulým rokem 2019. I když po 10. týdnu tohoto roku (začátek března) na několik týdnů vidíme vyšší přírůstek mortality než v minulém roce. Bylo by určitě zajímavější sledovat nárůst mrtvých v období tohoto října, listopadu, v době druhé vlny pandemie, avšak taková data nám zatím daná datová sada neposkytuje a údaje dostáváme se zpožděním.

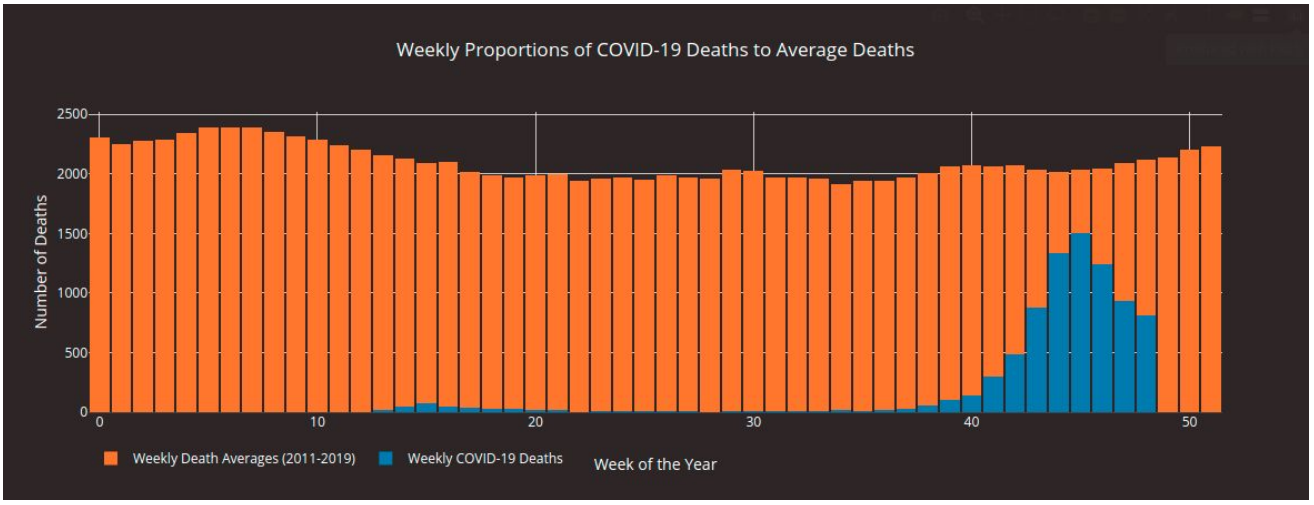


Poměr mrtvých na COVID-19 vůči celkové mortalitě v tomto roce

Koláčový graf ukazuje, jak velké procento zastupuje úmrtí na COVID-19 ve srovnání s všemi zaznamenanými úmrtími (na všechny příčiny) v datové sadě v tomto roce. Jedná se však o průměrné odhadované procento, protože týdenní data o úmrtích v ČR máme k dispozici zatím pouze do srpna tohoto roku. Až bude datová sada za tento rok kompletnější, tak budeme mít k dispozici přesnější procento. Vidíme tedy, že úmrtí na COVID-19 tvořily zatím zhruba 12,7 procent ze všech letošních zaznamenaných úmrtí v ČR, které se v datové sadě nacházejí.

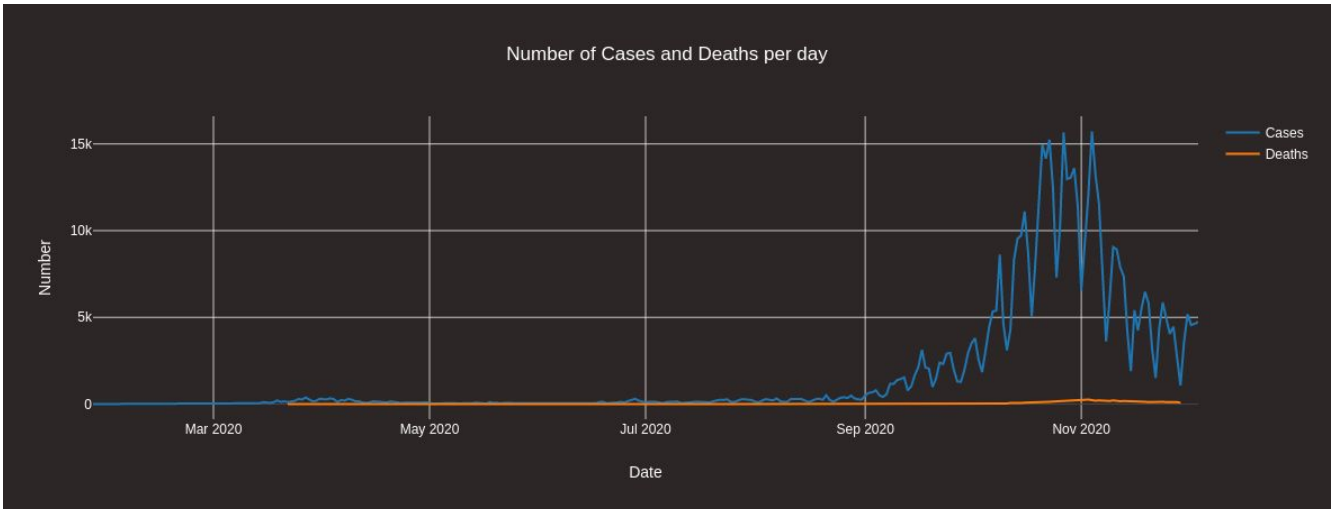


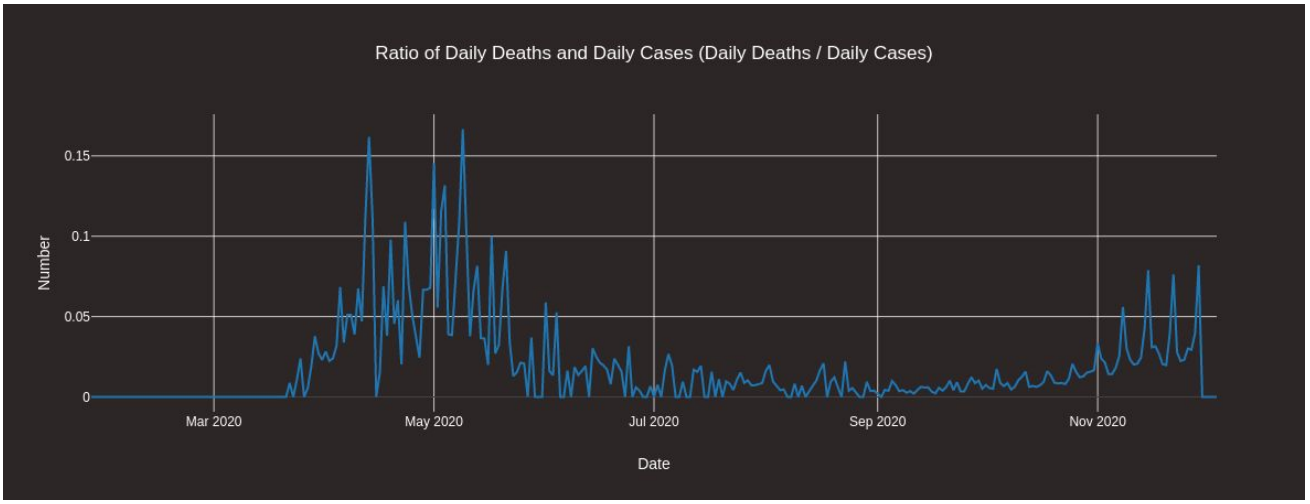
Právě z důvodu nekompletní datové sady jsme se rozhodli vytvořit druhý pohled na data, který pracuje s aktuálně dostupnými informacemi. Následující graf proto vizualizuje poměr mezi průměrným počtem zemřelých za jednotlivé týdny v letech 2011 až 2019 (roky nezatížené úmrtími na COVID-19) a počty zemřelých na COVID-19 v roce 2020 v odpovídajících týdnech. Zejména týdny v závěru roku ukazují znepokojující vývoj, kdy se v některých případech pochybujeme okolo 50% průměrného počtu úmrtí.



Denní vliv COVID-19

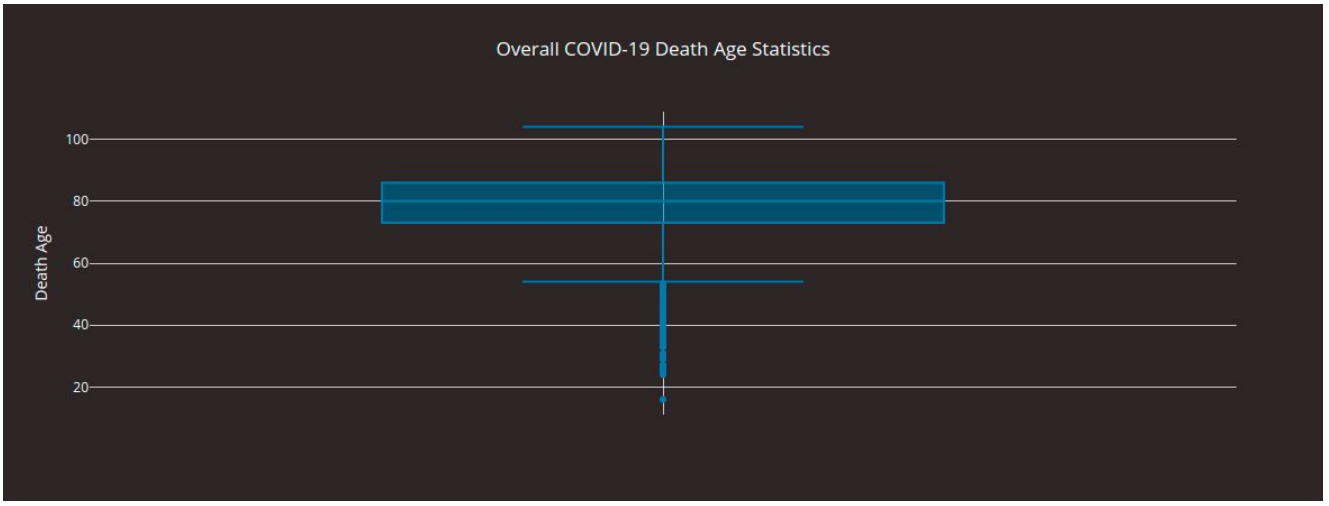
První graf po najetí na požadované datum zobrazuje denní počet nakažených, mrtvých a druhý graf poté zobrazí poté poměr mrtvých za daný den ku nakaženým za daný den. V těchto dnech je podíl mrtvých ku nakaženým docela nízký, což je pravděpodobně dáno vysokým počtem nakažených a poměr se tedy zákonitě snižuje. Nejvyšší hodnota daného poměru byla na začátku pandemie, kdy bylo zatím málo nakažených případů za den a podíl tedy dosahoval vyššího čísla. Například 13. dubna 2020 jsme měli 68 nakažených a 11 mrtvých (tedy poměr 11 / 68 = 0,162). Na druhou stranu v dnešních dnech máme již vysoká čísla nakažených a poměr je nižší (například 29. listopad 2020 - 1074 nakažených a 88 mrtvých - poměr 88 / 1074 = 0,082).



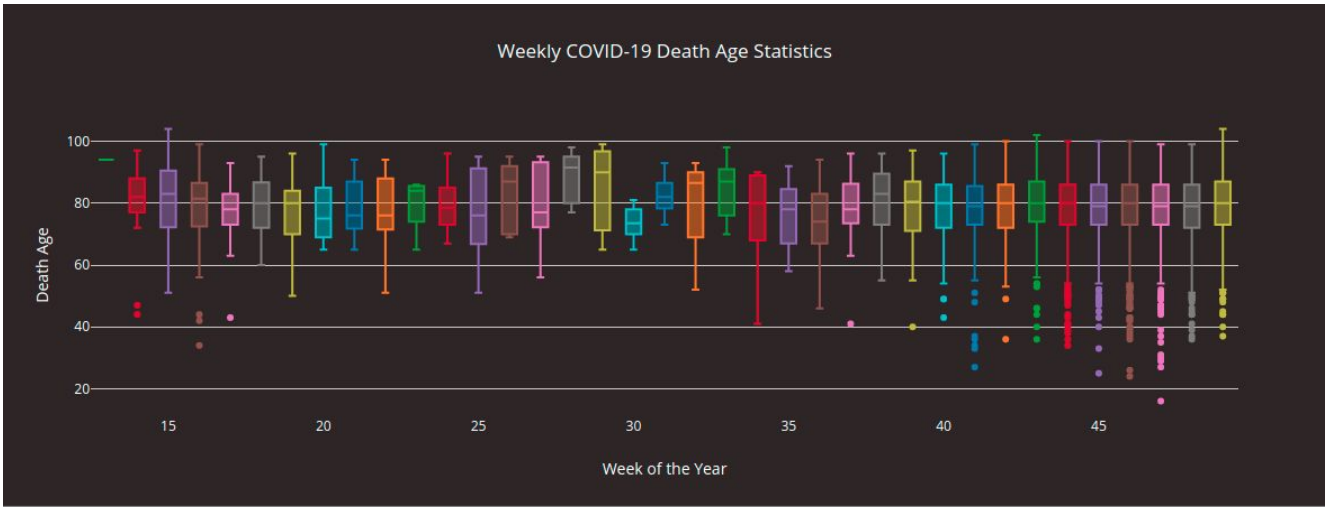


Věk úmrtí na COVID-19

Následující krabicový graf zobrazuje zajímavé informace týkající se věku mrtvých pacientů. Po najetí na graf se nám zobrazí charakteristiky, které nám právě krabicový graf dokáže poskytnout. Minimální věk mrtvého byl 16 a maximální pak 104. Avšak průměrně se věk umírajících pohybuje mezi 73 - 86 lety. Medián poté činí 80 let.



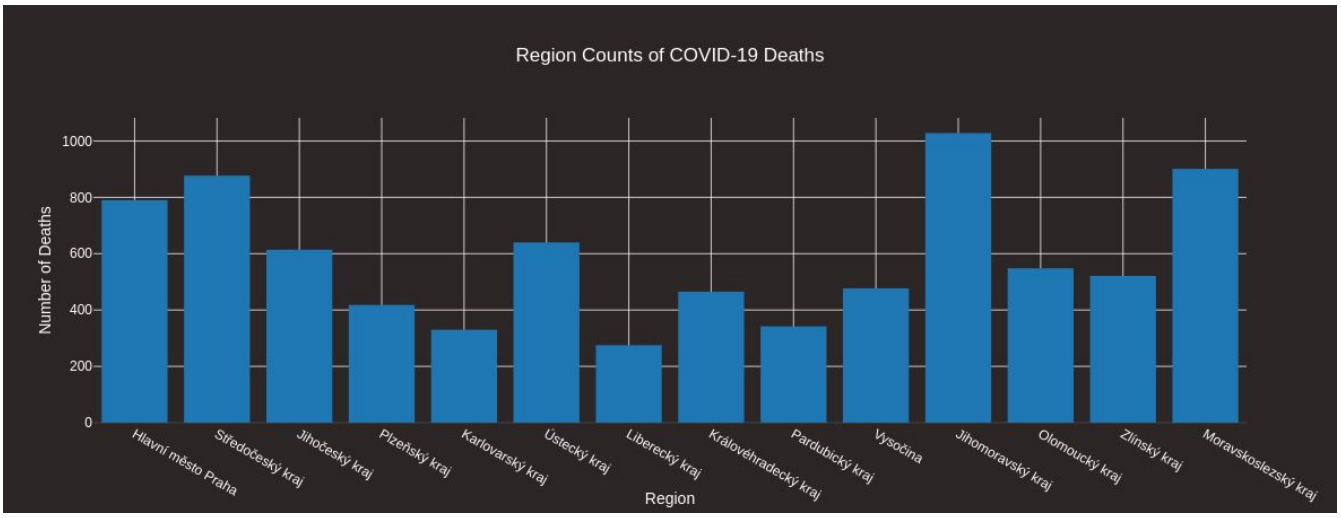
Druhý pohled na stejná data nám poté ukazuje vývoj této statistiky na základě týdnů.

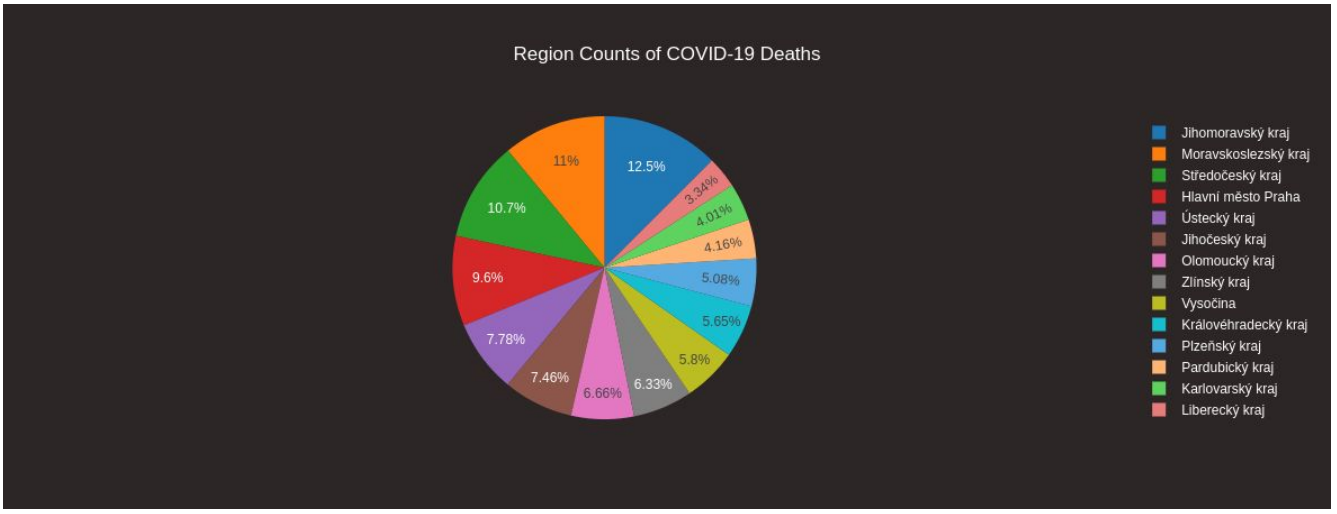


Vlastní dotaz

Počet mrtvých na COVID-19 dle regionu

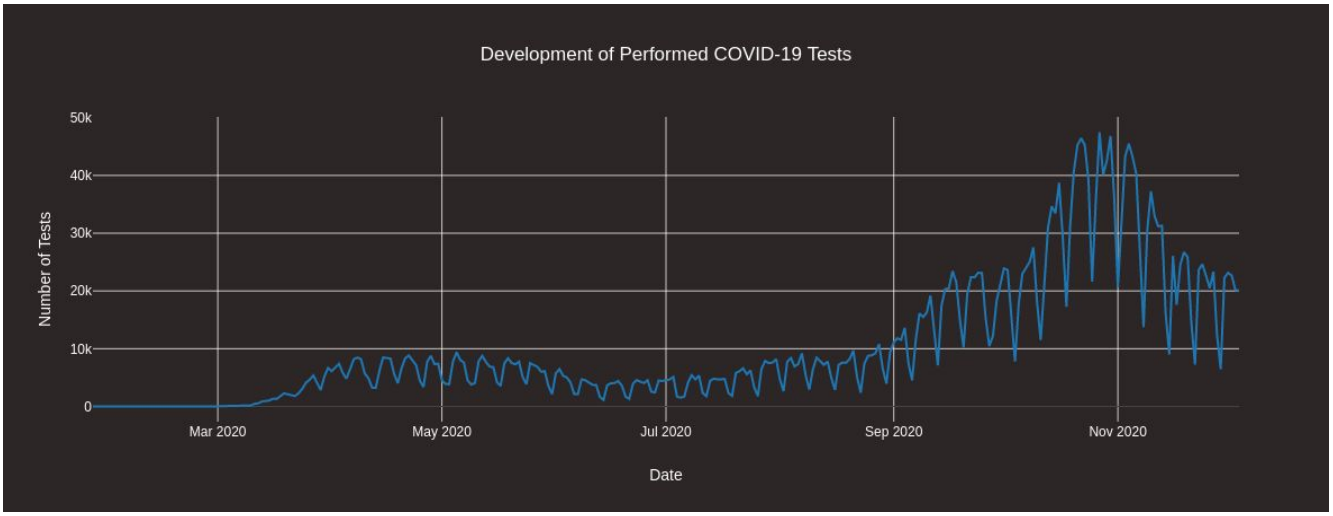
Stejně jako u nakažených dle regionu byl vytvořen sloupcový a koláčový graf. Opět bylo potřeba také vyfiltrovat kód neznámého regionu. Můžeme porovnat, zda kraje, co mají největší počet nakažených mají také nejvyšší počet mrtvých. Vidíme, že Moravskoslezský kraj má více úmrtí jak Středočeský. Avšak nejvyšší počet mrtvých má aktuálně Jihomoravský kraj, který se v počtu nakažených umístil až na 4. místě.





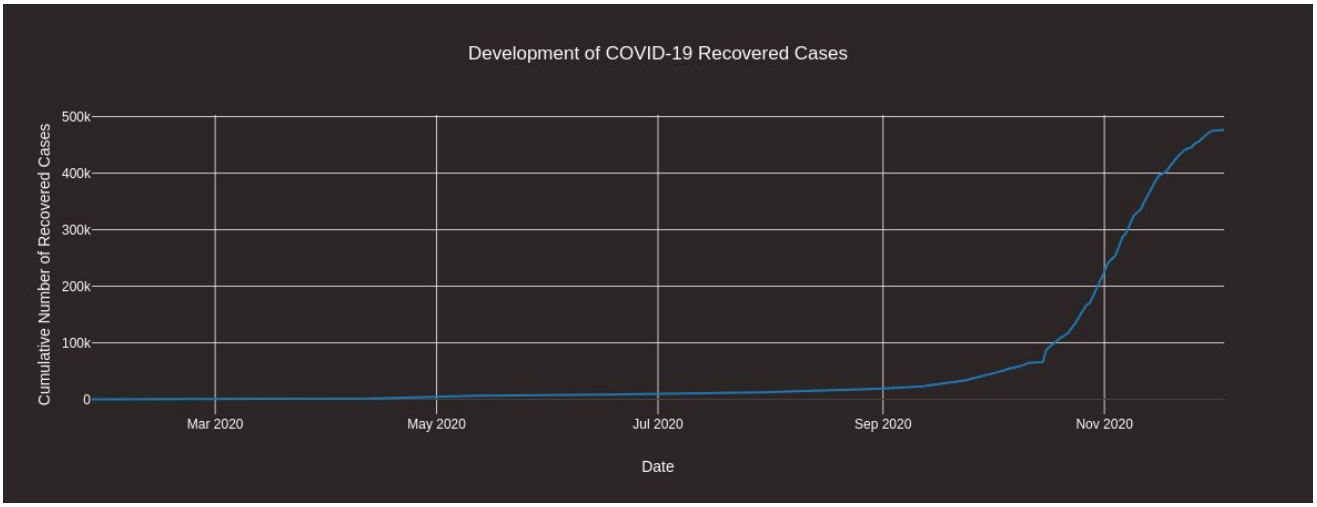
Vývoj provedených testů na COVID-19

Graf zobrazuje vývoj testování na COVID-19 v ČR. Využíváme údaje přírustkového počtu testů. Vidíme, že po zmírnění testování o letních prázdninách opět od září strmě testování narůstalo. Pravidelné propady jsou příčinou víkendů a svátků. Nyní testujeme opět méně než například na konci října, kdy se zatím v ČR testovalo úplně nejvíce.



Vývoj uzdravených na COVID-19

Pro ukázání vývoje vyléčených pacientů je využit údaj kumulativního počtu vyléčených. Vidíme, že tento údaj se stále zvyšuje s (jak víme) vyšším počtem případů nákazy COVID-19.


Development of COVID-19 Recovered Cases
Cumulative Number of Recovered Cases
500k
400k
300k
200k
100k
0
Mar 2020
May 2020
Jul 2020
Sep 2020
Nov 2020
Date