Projektová dokumentace

Implementace systému pro získání, zpracování a ukládání dat

COVID-19

Autoři:
Tomáš Beránek (xberan46)
Daniel Kamenický (xkamen21)
Richard Klem (xklemr00)

Datum odevzdání: **16. prosince 2021**

Obsah

1 Řešení 2. části projektu

Cílem 2. části projektu je:

- extrakce uložených dat z MongoDB do CSV souborů,
- úprava dat pro tvorbu popisných charakteristik a úloh pro dolování z dat.

2 Volba dotazů ze zadání

Pro vhodné předzpracování dat je dobré znát způsob, jakým se budou data využívat. Na základě této informace je pak možné například – identifikovat nepotřebná data, která lze v předzpracování odstranit, zvolit vhodný databázový systém, upravit formát atributů v předzpracování atp. Cílem 2. části projektu je realizace dvou dotazů typu A, jednoho dotazu typu B, dvou vlastních dotazů a příprava dat pro jednu z dolovacích úloh typu C.

Vybrané dotazy typu A:

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.

   Potřebná data:

   – **počet nově nakažených** (minimální granularita – měsíc)

   – **počet nově vyléčených** (minimální granularita – měsíc)

   – **počet nově hospitalizovaných** (minimální granularita – měsíc)

   – **počet nově provedených testů** (minimální granularita – měsíc)

3) Vytvořte sérii sloupcových grafů, které zobrazí:

   – počty provedených očkování v jednotlivých krajích (celkový počet od začátku očkování),

   – počty provedených očkování jako v předchozím bodě navíc rozdělené podle pohlaví. Diagram může mít např. dvě části pro jednotlivá pohlaví,

– počty provedených očkování, ještě dále rozdělené dle věkové skupiny. Pro potřeby tohoto diagramu postačí 3 věkové skupiny (0-24 let, 25-59 let, nad 59).

Potřebná data:

– **počet provedených očkovaní u můžu/žen v daném kraji** (minimální granularita – celkem)

– **počet provedených očkovaní podle věku v daném kraji** (minimální granularita počtu očkování – celkem, minimální granularita věku – 0-24 let, 25-59 let a nad 59 let)

Vybraný dotaz typu B:

2) Vytvořte sérii sloupcových grafů (alespoň 3), které porovnají vývoj různých covidových ukazatelů vámi zvoleného kraje se zbytkem republiky. Jako covidové ukazatele můžete použít: počet nakažených osob, počet hospitalizovaných osob, počet zemřelých, počet očkovaných. Všechny hodnoty uvažujte přepočtené na jednoho obyvatele kraje/republiky. Zobrazte alespoň 12 po sobě jdoucích hodnot (např. hodnoty za poslední rok po měsících).

Potřebná data:

- **celkový počet nakažených podle kraje** (minimální granularita – měsíc)
- **celkový počet zemřelých podle kraje** (minimální granularita – měsíc)
- **celkový počet očkovaných podle kraje** (minimální granularita – měsíc)

Vlastní dotazy zahrnující data alespoň ze dvou zdrojů:

1) Vytvořte sloupcový graf rozdělený podle věku (po 5letých intervalech) obyvatelstva ve vybraném okresu. Každý sloupec vyjadřuje kolik procent obyvatelstva z dané věkové skupiny dostalo COVID-19 za poslední rok (neuvažujte duplicity).

Potřebná data:

– **počet nakažených podle věku a okresu** (minimální granularita věku – 5leté intervaly)

– **počet obyvatel podle věku a okresu** (minimální granularita věku – 5leté intervaly)

2) Vytvořte graf, který ukazuje vývoj počtu nakažených žen a mužů začátku pandemie COVID-19 po měsících. Hodnoty uvádějte v procentech (např. 100% u mužů znamená, že jsou nakaženi všichni muži).

   Potřebná data:

   – **počet nakažených podle pohlaví** (minimální granularita – měsíc)

Vybraná dolovací úloha typu C:

1) Hledání skupin podobných okresů 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 okresů, pro které najdete potřebné hodnoty (můžete např. využít nějaký žebříček 50 nejlidnatějších okresů v ČR).

   Potřebná data:

   – **počet nakažených podle věku a okresů** (minimální granularita věku – 0-24 let, 25-59 let a nad 59 let, minimální granularita datumu – 3 měsíce)

   – **počet očkovaných podle věku a okresů** (minimální granularita věku – 0-24 let, 25-59 let a nad 59 let, minimální granularita datumu – 3 měsíce)

   – **počty obyvatel v jednotlivých okresech podle věku** (minimální granularita věku – 0-24 let, 25-59 let a nad 59 let)

3 Extrahování dat z MongoDB do CSV souborů

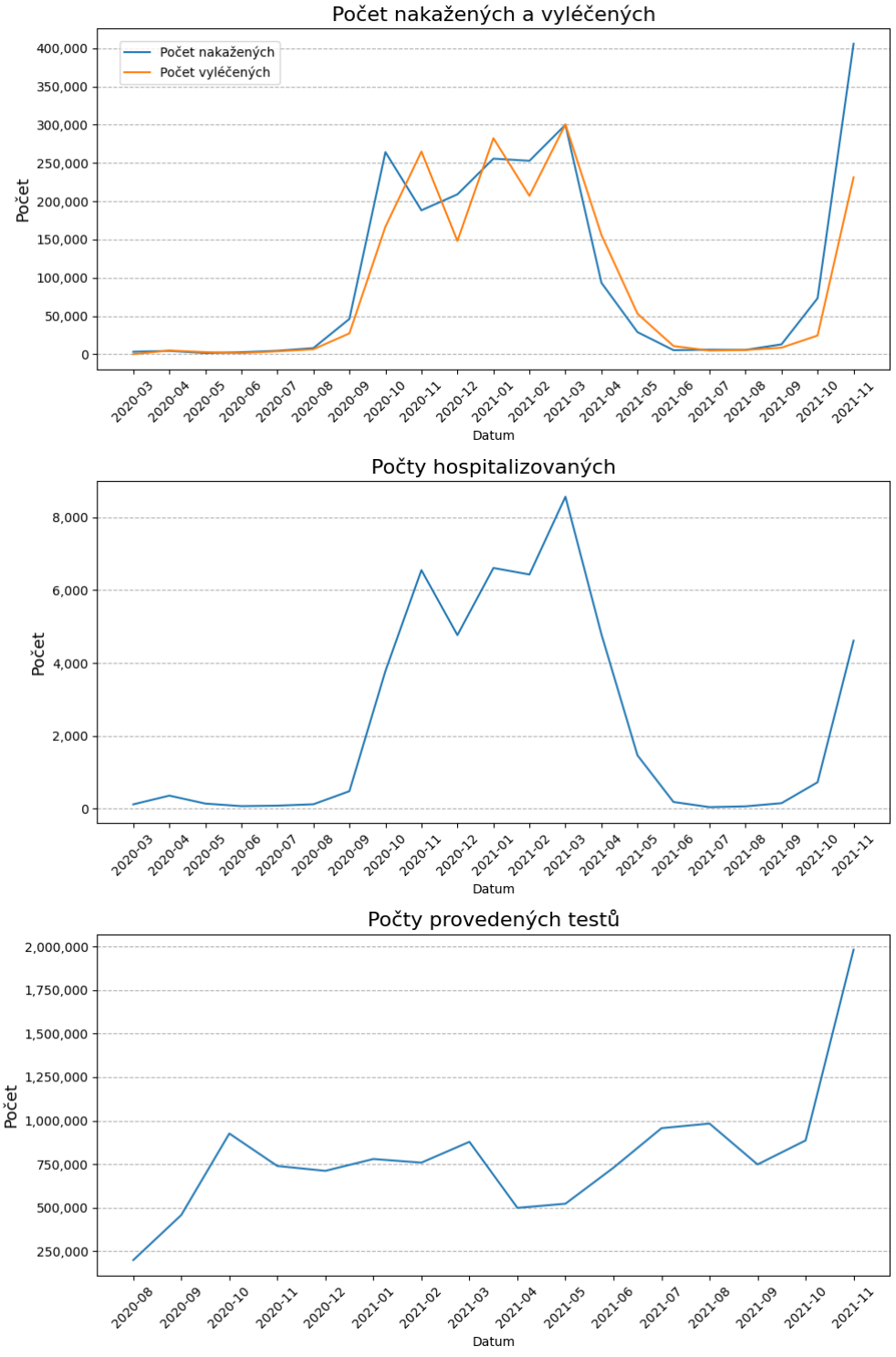
Pro spuštění tohoto kroku je nutné mít vytvořenou a naplněnou databázi z 1. části projektu. Podrobný popis vkládání dat do databáze je uveden v dokumentaci 1. části projektu, která je umístěna v kořenové složce adresáře. Extrahování dat zajišťuje Python3 skript `src/mongo_to_dataframe.py`. Obdobně jako v 1. části je zde využito knihovny `pymongo`. Skript generuje CSV soubory do složky `data-part2/`. Každý soubor má prefix `A1`, `A3`, `B2`, `V1`, `V2` nebo `C1`, který udává pro jaké dotazy bude použit.

4 Dotaz A1

Pro tento dotaz byly z databáze extrahovány čtyři soubory:

- `A1-hospitalizace.csv` s atributy:
  - `datum`
  - `pocet_hosp`
- `A1-kraj-okres-testy.csv` s atributy:
  - `datum`
  - `prirustkovy_pocet_testu_okres`
- `A1-osoby.csv` s atributy:
  - `datum`
- `A1-vyleceni.csv` s atributy:
  - `datum`

U tohoto dotazu (ani u ostatních) není detailně popsána implementace. Je však možno nahlédnout do komentovaných zdrojových kódů v repozitáři. Pro zodpovězení dotazu A1 byl vytvořen obrázek se třemi grafy, viz. obrázek 1. První graf ukazuje vývoj počtu nakažených a vyléčených lidí. Jelikož jsou tyto ukazatele velmi podobné (mají podobný průběh), tak byly vloženy do jednoho obrázku. Druhý graf ukazuje vývoj počtu hospitalizovaných. Jelikož jsme neměli k dispozici přírustkový počet hospitalizovaných (pouze aktuální počet), tak měsiční hodnota je vypočítána jako průměr všech dní v daném měsíci. Tento graf byl oddělen, protože počty hospitalizovaných jsou razantně menší a křivka v grafu by se jevila téměř jako přímka u hodnoty 0. Třetí graf ukazuje počet provedených testů. Tento graf byl také oddělen, jelikož počty provedených testů jsou razantně větší než počty v prvním/druhém grafu a docházelo by tak ke změnšení detailu u předchozích ukazatelů.



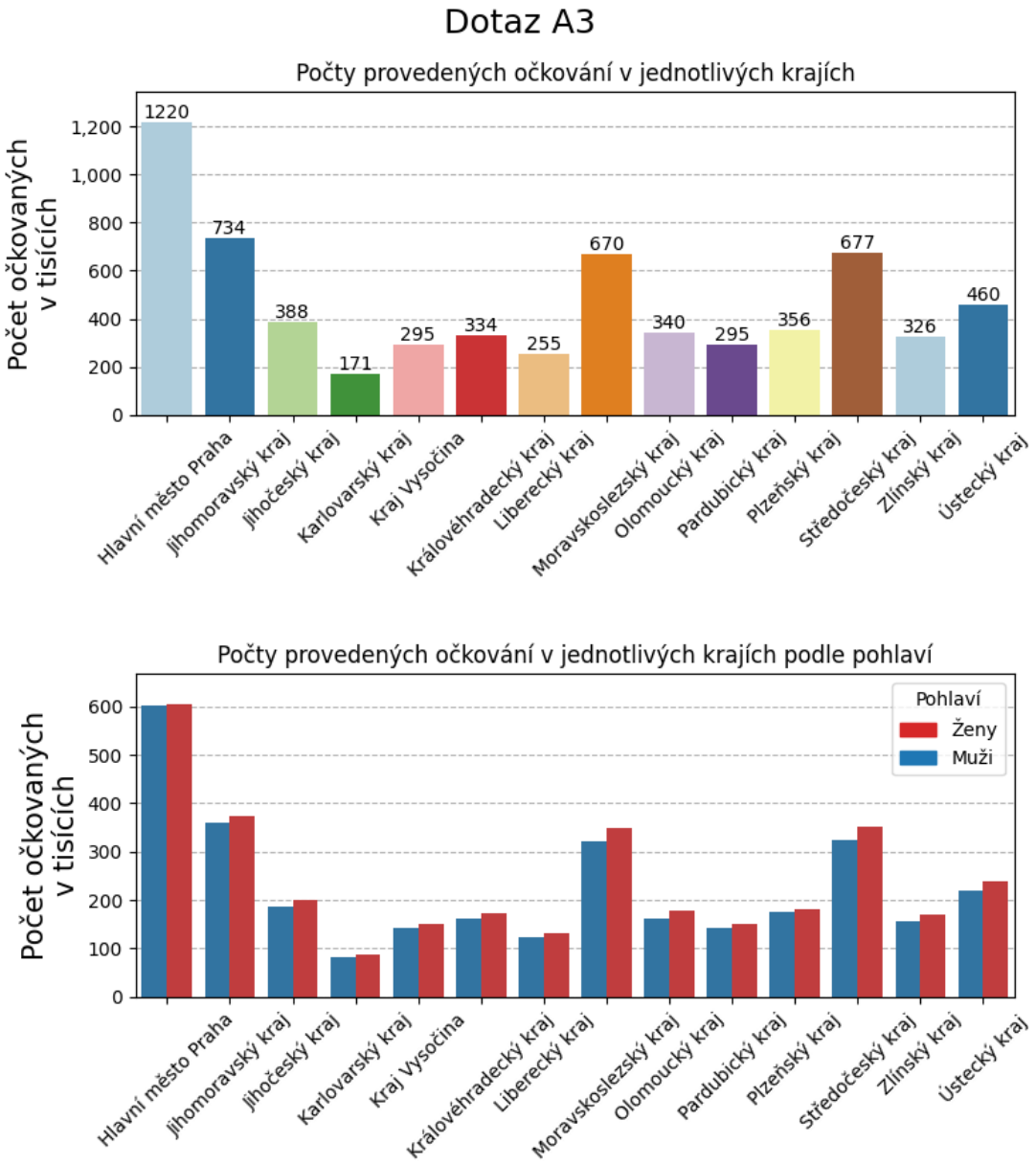
Obrázek 1: Vizualizace dotazu A1. Graf zobrazující vývoj covidové situace.

5 Dotaz A3

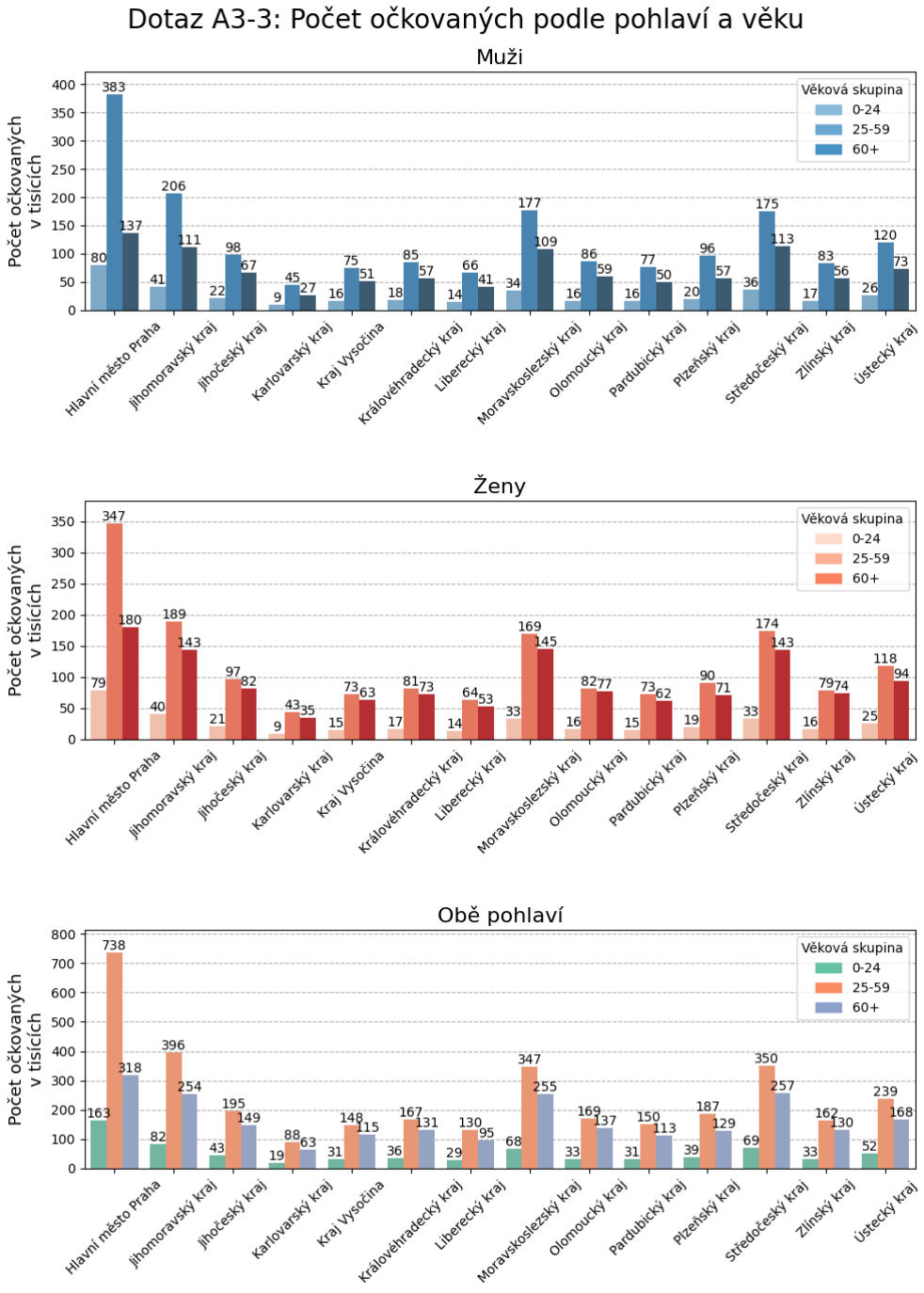
Pro zodpovězení dotazu A3 byl z databáze extrahován jeden soubor:

- `A3--profese.csv` s atributy:
  - `datum`
  - `vakcina`
  - `kraj_nazev`
  - `poradi_davky`
  - `vekova_skupina`
  - `pohlavi`

Pro zodpovězení dotazu A3 byly vytvořeny dva obrázky. Obrázek 2 obsahuje grafy, které popisují počty provedených očkování v jednotlivých krajích navíc rozdělené podle pohlaví. Obrázek 3 obsahuje tři grafy, které tyto statistiky rozdělují ještě podle věku.



Obrázek 2: Vizualizace dotazu A3. Počty provedených očkování v jednotlivých krajích navíc rozdělené podle pohlaví.



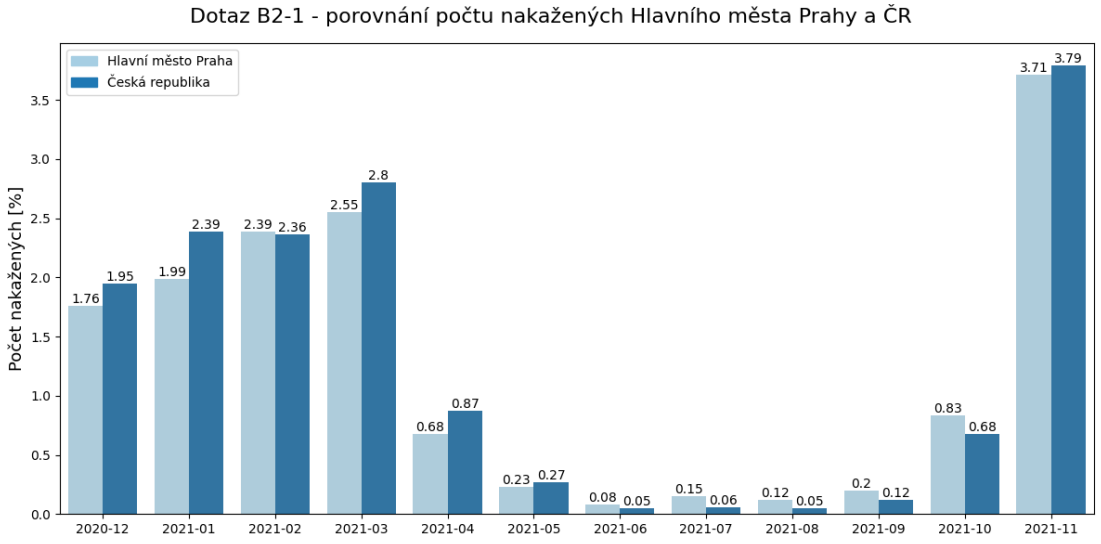
Obrázek 3: Vizualizace dotazu A3. Počty provedených očkování v jednotlivých krajích rozdělené podle pohlaví a podle věku.

6 Dotaz B2

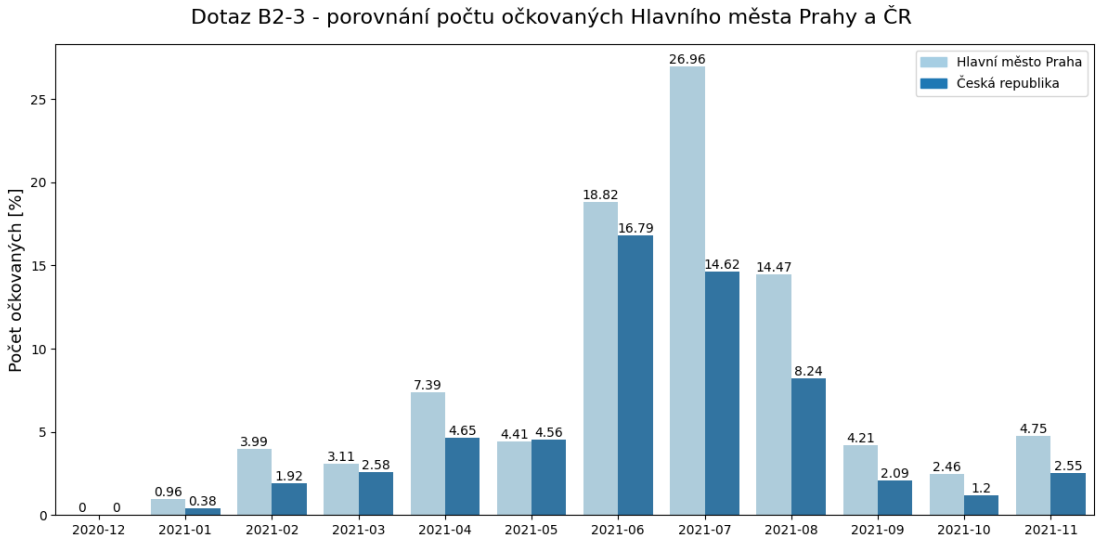
Pro tento dotaz byly z databáze extrahovány tři soubory:

- `B2-ockovani-profese.csv` s atributy:
  - `datum`
  - `vakcina`
  - `kraj_nuts_kod`
  - `poradi_davky`
- `B2-osoby.csv` s atributy:
  - `datum`
  - `kraj_nuts_kod`
- `B2-umrti.csv` s atributy:
  - `datum`
  - `kraj_nuts_kod`

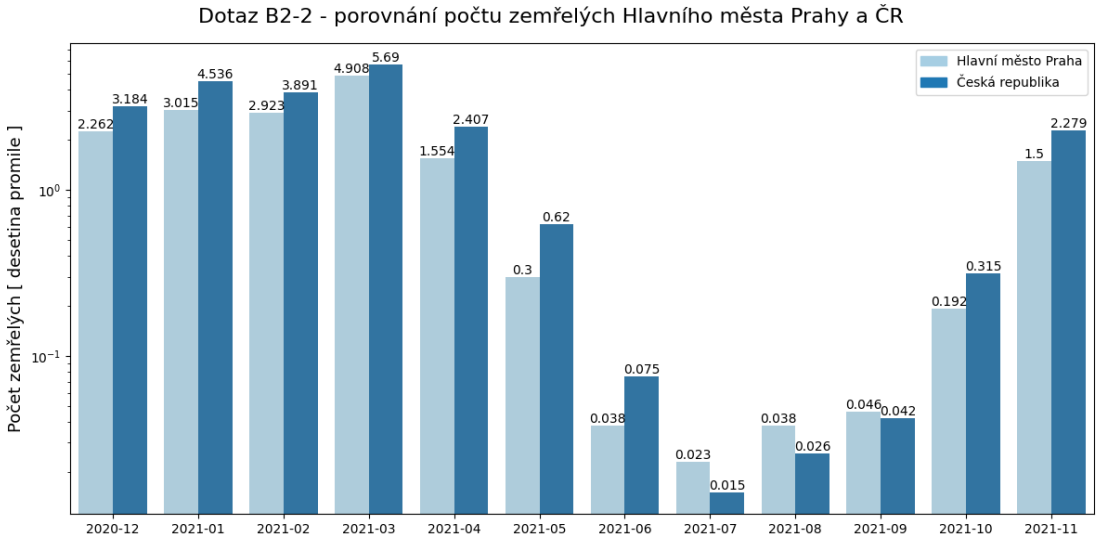
Pro zodpovězení dotazu B2 byly vytvořeny tři grafy, které ukazují porovnání Hlavního Města Prahy se zbytkem republiky. První graf porovnává počty nakažených, viz. obrázek 4. Druhý graf porovnává počty **plně** očkovaných, viz. obrázek 5. A třetí graf porovnává počty úmrtí, viz. obrázek 6.



Obrázek 4: Vizualizace dotazu B2. Porovnání počtu nakažených.



Obrázek 5: Vizualizace dotazu B2. Porovnání počtu očkovaných.



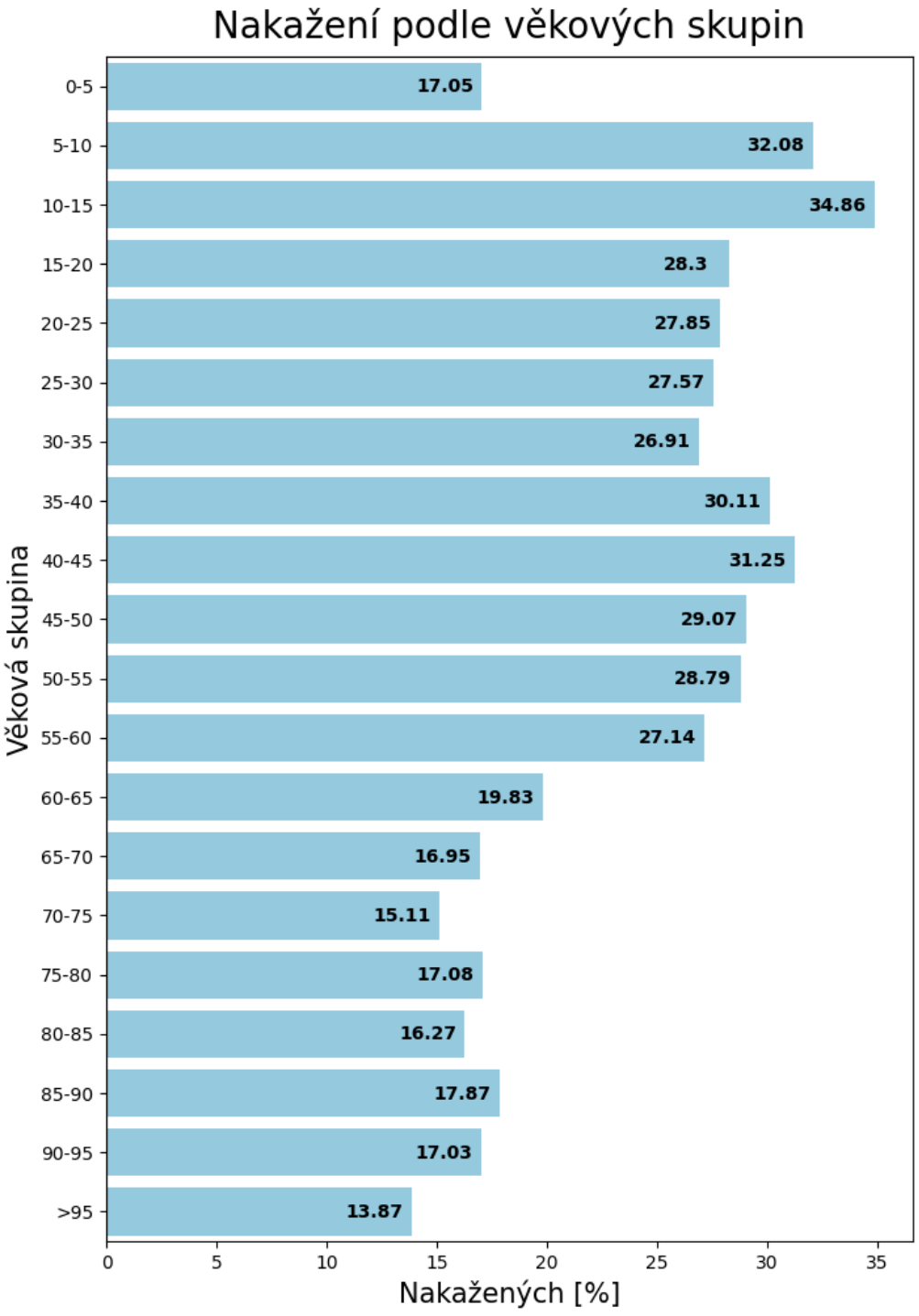
Obrázek 6: Vizualizace dotazu B2. Porovnání počtu úmrtí.

7 Dotaz V1

Pro tento dotaz byly z databáze extrahovány tři soubory:

- `V1-130142-21data043021.csv` s atributy:
  - `hodnota`
  - `vek_txt`
  - `vuzemi_txt`
  - `casref_do`
  - `pohlavi_kod`
- `V1-obce.csv` s atributy:
  - `okres_lau_kod`
  - `okres_nazev`
- `V1-osoby.csv` s atributy:
  - `datum`
  - `vek`
  - `okres_lau_kod`

Pro zodpovězení dotazu V1 byl vytvořen graf, který ukazuje zastoupení nakažených ve věkových skupinách v okresu Opava za poslední rok, viz. obrázek 7.



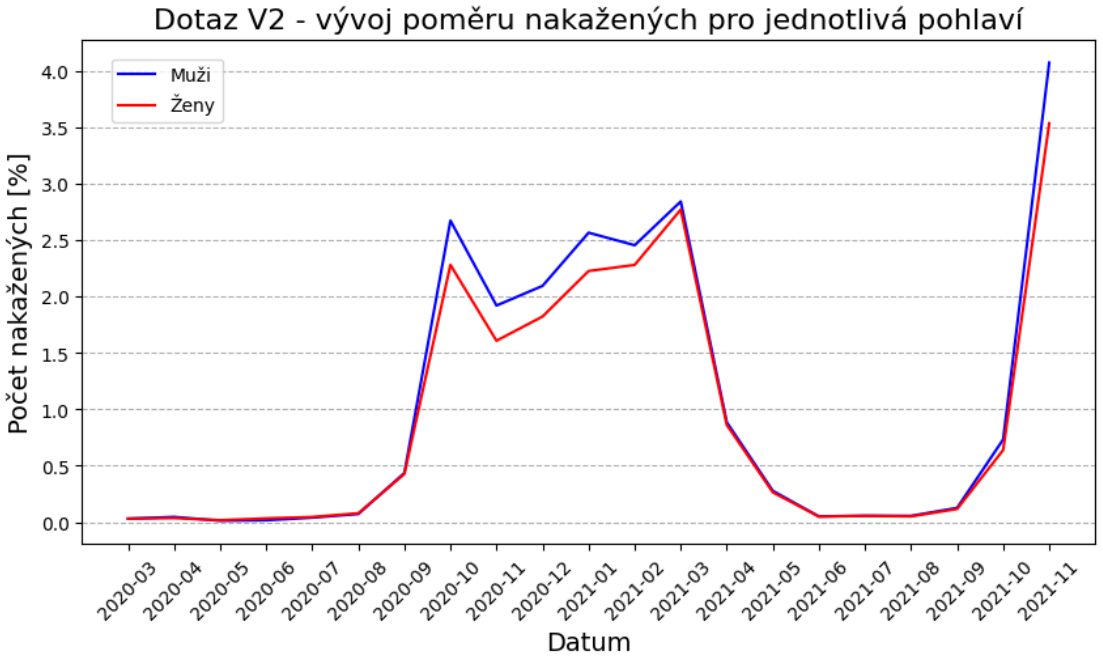
Obrázek 7: Vizualizace dotazu V1. Zastoupení nakažených ve věkových skupinách v okresu Opava za poslední rok.

8 Dotaz V2

Pro tento dotaz byl z databáze extrahován jeden soubory:

- `V2-osoby.csv` s atributy:
  - `datum`
  - `pohlavi`

Pro zodpovězení dotazu V2 byl vytvořen graf, který ukazuje procentuální zastoupení nakažených můžu/žen, viz. obrázek 8. Tento dotaz jsme zvolili, abychom zjistili, zda je některé pohlaví náchylnější na nakažení.



Obrázek 8: Vizualizace dotazu V2. Graf procentuálního zastoupení nakažených mezi muži/ženami.

9 Dotaz C1

Pro tento dotaz byly z databáze extrahovány čtyři soubory:

- `C1-130142-21data043021.csv` s atributy:
  - `hodnota`
  - `pohlavi_kod`

  - `casref_do` – datum, ke kterému řádek platil
  - `vek_txt`
  - `vuzemi_txt` – ČR, kraje a okresy
- `C1-obce.csv` s atributy:
  - `okres_lau_kod`
  - `okres_nazev`
  - `orp_kod`
- `C1-ockovani-profese.csv` s atributy:
  - `datum`
  - `vakcina`
  - `kraj_nuts_kod`
  - `poradi_davky`
  - `vekova_skupina`
  - `orp_bydliste_kod`
- `C1-osoby.csv` s atributy:
  - `datum`
  - `okres_lau_kod`

Výsledkem tohoto dotazu je CSV soubor `data-mining/C1-data.csv` s následujícími atributy:

- `okres_nazev` – název okresu. Bylo vybráno 50 nejlidnatějších okresů v ČR. Do těchto okresů počítáme i Hlavní Město Prahu, ikdyž se technicky vzato nejedná o okres, ale o kraj. Nicméně jsme chtěli vidět statistiky z Prahy.
- `4_nakazeni` – počet nakažených za poslední kvartál. Hodnota v tomto sloupci byla nejdříve transformována na počet nakažených na 1000 obyvatel daného okresu. Poté byly odstraněny odlehlé hodnoty pomocí IQR a nahrazeny průměrnou hodnotou (bez odlehlých hodnot). Nakonec byla tato hodnota normalizována pomocí metody min-max do intervalu 0 až 1.
- `3_nakazeni` – počet nakažených za předposlední kvartál. ...
- `2_nakazeni` – ...

- `1_nakazeni` – ...
- `4_ockovani` – počet očkovaných za poslední kvartál. Hodnota v tomto sloupci byla opět nejdříve přepočtena na 1000 obyvatel daného kraje, opět byly obdobně odstraněny a nahrazeny odlehlé hodnoty. Poté se obdobně provedla normalizace a nakonec diskretizace těchto hodnot do tří kategorií, které udávají jaká je proočkovanost daného okresu vůči ostatním okresům za daný kvartál:
  - `nízká` – hodnoty od 0 (včetně) do 0.3.
  - `střední` – hodnoty od 0.3 (včetně) do 0.7.
  - `vysoká` – hodnoty od 0.7 (včetně) do 1 (včetně).
- `3_ockovani` – počet očkovaných za předposlední kvartál. ...
- `2_ockovani` – ...
- `1_ockovani` – ...
- `0_14_vek` – počet obyvatel okresu ve věku 0 (včetně) až 14 let.
- `15_59_vek` – počet obyvatel okresu ve věku 15 (včetně) až 59 let.
- `60_vek` – počet obyvatel okresu ve věku 60 a více let.

Ve sloupci `3_nakazeni` bylo odstraněno 5 odlehlých hodnot, ve sloupci `1_nakazeni` byly odstraněny 3 hodnoty a ve sloupci `1_ockovani` byla odstraněna 1 hodnota.

CSV soubor `data-mining/C1-data-bez-uprav.csv` obsahuje neupravená (tzn. bez transformace, normalizace, diskretizace a odstranění odlehlých hodnot) data pro data miningovou úlohu.

10 Návod na spuštění

Prerekvizity, specifické vlastnosti chování, potřebná nastavení a návod na spuštění jsou popsány v dokumentaci v souboru `README.md`, který je umístěn v kořenovém adresáři projektu. Celý projekt (1. i 2. část) funguje jako jedna pipeline – stáhne data, transformuje na JSON, vyfiltruje, vloží do Mongo DB, extrahuje do CSV, vygeneruje grafy a vytvoří CSV pro data mining. Jelikož je ve WISu limit na 10MB a naše extrahovaná CSV data mají přibližně 2GB, tak jsou data ke stažení na této adrese. Stažená data vložte i se složkou `data-part2/` do kořenového adresáře repozitáře.

Při vytváření grafů, skript ukládá do složky `dumps/` CSV soubory, které obsahují data, ze kterých jsou grafy vygenerovány.