# Příprava

1. Nejprve jsem si prohlédla všechny tabulky, abych zjistila, které z požadovaných proměnných (sloupců nové tabulky) se vyskytují ve kterých tabulkách.

|  |  |  |
| --- | --- | --- |
| **Sloupec** | **Zdrojová tabulka** | **Poznámka** |
| country | covid19\_basic\_differences  (covid19\_tests) | přímo v tabulce |
| date | covid19\_basic\_differences  (covid19\_tests) | přímo v tabulce |
| denní nárůst nakažených | covid19\_basic\_differences | přímo v tabulce |
| počty provedených testů | covid19\_tests | přímo v tabulce |
| počet obyvatel státu | countries  economies  lookup\_table | přímo v tabulce, ale musím vybrat, kterou tabulku použiju |
| binární proměnná pro víkend / pracovní den | covid19\_basic\_differences  (covid19\_tests) | zjistím na základě data |
| roční období daného dne[[1]](#footnote-1) | covid19\_basic\_differences  (covid19\_tests) | zjistím na základě data |
| hustota zalidnění státu | countries | přímo v tabulce |
| HDP na obyvatele státu | economies | dopočítám ze sloupců přímo v tabulce |
| GINI koeficient státu | economies | přímo v tabulce |
| dětská úmrtnost státu | economies | přímo v tabulce |
| medián věku obyvatel v roce 2018 | countries | přímo v tabulce |
| podíly jednotlivých náboženství (pro každé náboženství v daném státě procentní podíl jeho příslušníků na celkovém obyvatelstvu) | religion | dopočítám na základě sloupce population a celkové populace z tabulky countries, economies nebo lookup\_table |
| rozdíl mezi očekávanou dobou dožití v roce 1965 v roce 2015 | life\_expectancy | přímo v tabulce, ale musím transponovat |
| průměrná denní (nikoli noční!) teplota | weather | dopočítám na základě sloupců time a temp |
| počet hodin v daném dni, kdy byly srážky nenulové | weather | dopočítám na základě sloupců time a rain |
| maximální síla větru v nárazech během dne | weather | dopočítám na základě sloupce gust |

Zjistila jsem tyto „issues“ v datech:

* **GDP, gini koeficient** **a mortality under 5** jsou v různých zemích vyplněny různě (např. gini koeficient má v některých zemích nejaktuálnější hodnotu z roku 2018, v jiných z 90. let), bude tedy asi potřeba pro každou zemi vzít hodnotu z jiného roku.
* **population** je v tabulce countries a lookup\_table statická (nevím, ze kterého roku), v tabulce economies dynamická (vyplněná do roku 2019, v roce 2020 NULL). Musím se rozhodnout, ze které tabulky údaj k population vezmu.
* **Názvy některých států** (ČR, USA, Russia) jsou v různých tabulkách různé, resp. v tabulce covid19\_basic\_differences jsou stejné jako v tabulce lookup\_table, a v tabulce countries a covid19\_tests jsou stejné jako v economies

Narazila jsem na tyto nejasnosti:

* **průměrná denní (nikoli noční!) teplota** – jaký časový interval je chápán jako „den“ a jaký jako „noc“? Zvolila jsem si den 9:00 – 20:00 (tj. údaje z časů 9:00, 12:00, 15:00, 18:00) a noc 21:00 – 8:00 (tj. údaje z časů 21:00, 00:00, 03:00, 6:00).
* **počet hodin v daném dni, kdy byly srážky nenulové** – čas v tabulce je uvedený ve 3hodinových intervalech (0:00, 3:00, 6:00 atd.). Když jsou pro daný čas srážky nenulové, počítám do celkového počtu hodin ten čas jako 1h, nebo počítám celý 3h interval?

1. Vyřešila jsem prvotní „issues“

* **GDP, gini koeficient, mortality under 5** jsem vzala v každé zemi nejaktuálnější, který je k dispozici, tj. našla jsem poslední rok (MAX(year)), ve kterém byly tyto ukazatele NOT NULL
* **population** – spojením tabulek a porovnáním údajů k population jsem zjistila, že údaje v countries jsou z roku 2018, v economies z 2019 a v lookup\_table z 2020. Vzhledem k tomu, že data o průběhu COVID19 jsou z let 2020-2021, vezmu údaje k population z roku 2020, tj. **z tabulky lookup\_table**
* **Názvy některých států (ČR, USA) jsou v různých tabulkách různé** – vyřeším tím, že přes JOIN spojím tabulku covid19\_basic\_differences s lookup\_table přes country, tabulky countries a economies přes country a pak tyto dvě nové tabulky přes iso3

1. Spojila jsem si tabulky economies, countries, covid19\_basic\_diff, lookup a covid19\_tests:
   1. Přes LEFT JOIN jsem z tabulek economies a countries vytvořila „stavovou“ (pro každou zemi jeden řádek, tj. bez data) tabulku se sloupci: country, GDP, gini, mortality\_under5, iso3, population\_density, median\_age\_2018 a uložila ji do VIEW: ***v\_joined\_economies\_countries***
   2. Přes LEFT JOIN jsem k tabulce covid19\_basic\_differences připojila z tabulky lookup\_table sloupec s údajem o celkové populaci a iso3. K této tabulce jsem chtěla pomocí LEFT JOIN přes datum a ISO připojit tabulku covid19\_tests, abych získala sloupec s denními hodnotami testování. Toto spojení trvalo strašně dlouho, pokud jsem chtěla mít data v tabulce nějak seřazena (buď podle země, nebo podle data)!!!

PROBLÉM: Při kontrole, jestli spojení přes LEFT JOIN proběhlo v pořádku (výchozí tabulka a nová tabulka mají stejný počet řádků), jsem zjistila, že nová tabulka má víc řádků. Myslela jsem, že je problém v tom, že některé země (CZE, USA) mají v tabulce covid19\_tests jiný název než v covid19\_basic\_diff. Při kontrole pro CZE to bylo OK, u USA jsem zjistila, že problém je ve sloupci entity tabulky covid19\_tests, který obsahuje dvě různé kategorie (test performed a units unclear), přičemž v daný den jsou uvedeny hodnoty testování pro každou entitu zvlášť (takže 2 řádky se stejným datem). Takže i ve výsledné tabulce bude v těchto případech jedno datum na dvou řádcích, což ale nechci, takže je potřeba rozhodnout, který vybrat, nebo jestli je sečíst.

Přes SELECT COUNT(DISTINCT(entity)) jsem zjistila, že zemí se 2 entitami je 8 (Francie, Indie, Itálie, Japonsko, Polsko, Singapore, Švédsko a USA). Našla jsem si data ke COVID19 na <https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.xlsx> a namátkově oba dataset pro tyto země porovnala. Na základě tohoto porovnání jsem vybrala tyto entity:

* + 1. Francie – people tested a tests performed (ve druhém zdroji je uvedeno použití people tested, ale čísla odpovídají spíš tests performed--> vybrala jsem **tests performed**
    2. Indie – people tested a samples tested (některé dny jsou hodnoty stejné, ke konci období už uvádí jen samples tested stejně jako v jiném zdroji) --> vybrala jsem **samples tested**
    3. Itálie – people tested a tests performed --> vybrala jsem **tests performed**
    4. Japonsko – people tested a tests performed (ve druhém zdroji jiná čísla, ale velkou část dnů uvedeno jen people tested) --> vybrala jsem **people tested**
    5. Polsko – people tested a samples tested (čísla ve druhém zdroji sice odpovídají people tested, ale jsou za kratší období a v našem zdroji na začátku období uvádí jen samples tested) --> vzala jsem **samples tested**
    6. Singapore – hodnoty tests\_performed jsou NULL, ale i tak jsou zdvojené řádky, takže jsem vzala jen **samples tested**
    7. Švédsko – má sice 2 entity, ale nepoužívali je současně, takže nedochází ke zdvojování řádků
    8. USA – tests performed a units unclear (incl. non-PCR) (čísla ve druhém zdroji odpovídají spíše tests performed) --> vybrala jsem **tests performed**

Připravila jsem si nové tabulky covid19\_tests pro "problematické" země, ve kterých jsem určila jen jednu entitu. Pomocí UNION jsem tyto tabulky spojila spolu navzájem a také s tabulkou covid19\_tests, ze které jsem tyto země vyjmula. Tím jsem získala novou tabulku covid19\_tests\_new (uložila jsem si ji do VIEW *v\_covid19\_tests\_new*), která už u těchto "problematických" zemí nemá zdvojené záznamy pro žádné datum. Ke spojené tabulce joined\_covid\_lookup tedy připojuju tuto novou covid19\_tests\_new.

K této nové tabulce jsem opět přes LEFT JOIN připojila tabulku uloženou ve VIEW v\_joined\_economies\_countries. Dostala jsem tak velkou „vývojovou“ (pro každou zemi řádky s jednotlivými daty) tabulku se sloupci: date (všech 468 dnů), country (všech 189 zemí), iso3, confirmed, tests (hodnoty jen za 110 zemí), population, GDP, gini, mortality\_under5, population\_density, median\_age\_2018. Tuto tabulku jsem si uložila do VIEW: ***v\_joined\_covid\_lookup\_tests\_economies\_countries***

!!! Ve výsledné tabulce jsou hodnoty z v\_joined\_economies\_actual\_countries vypsány u každé země jen pro prvních 23 řádků, pak už jsou NULL. Ale pokud v tabulce použiju nějaké filtry (např. vyberu jen jednu zemi, nebo jen jedno datum), pak mi to normálně vypíše všechny údaje.!!!

1. Připojení tabulky religion a výpočet
2. Do výsledné tabulky jsem na základě výpočtů a podmínek doplnila další sloupce:
   1. binární proměnná pro víkend / pracovní den na základě CASE WHEN WEEKDAY(date) IN (5, 6) THEN 1 ELSE 0 END
   2. roční období daného dne – zjistila jsem si hraniční data astronomických ročních období ve sledovaném období (22. 1. 2020 – 3. 5. 2021) a na základě CASE WHEN vytvořila sloupec s ročními obdobími

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| rok | kdy začíná jaro | kdy začíná léto | kdy začíná podzim | kdy začíná zima |
| 2020 | 20. 3. | 20. 6. | 22. 9. | 21. 12. |
| 2021 | 20. 3. | 21. 6. | 22. 9. | 21. 12. |

1. Výpočty v tabulce weather

Tabulka weather neobsahuje názvy zemí ani ISO kódy, jen názvy hlavních měst. Pro připojení k mé výsledné tabulce musím nejprve tabulku weather spojit přes sloupec city s tabulkou countries (sloupec capital\_city), přičemž ale musím vyřešit problém, že ve weather jsou hlavní města uvedena anglicky a v tabulce countries v národních jazycích. Nevím, jak pomocí SQL zjistit, které názvy jsou shodné v obou tabulkách, takže jsem to udělala v excelu pomocí funkce MATCH. Tím jsem zjistila 10 zemí, které mají v tabulce weather název jiný než v tabulce countries. V excelu jsem si ručně dohledala názvy z tabulky countries a v mariaDB je nahradila pomocí CASE WHEN. Vytvořila jsem si tedy novou tabulku weather\_new, kterou jsem si uložila do VIEW v\_weather\_new

* 1. průměrná denní (nikoli noční!) teplota – počítala jsem za interval 9:00 – 20:00 (tj. údaje z časů 9:00, 12:00, 15:00, 18:00).
  2. počet hodin v daném dni, kdy byly srážky nenulové
  3. maximální síla větru v nárazech během dne

1. Pivotování a výpočty v tabulce life\_expectancy

1. zakódovat jako 0 až 3, přičemž se berou astronomická roční období: <https://vimjakna.cz/dny/kdy-zacina-zima/> [↑](#footnote-ref-1)