PROJEKT

# Studium tabulek

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

|  |  |  |
| --- | --- | --- |
| **Sloupec** | **Zdrojová tabulka** | **Poznámka** |
| datum | covid19\_basic\_differences  (covid19\_tests) | přímo v tabulce |
| země | 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 – 21:00 (tj. údaje z časů 9:00, 12:00, 15:00, 18:00), celkem tedy 12h.
* **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, má se počítat do celkového počtu hodin ten čas jako 1h, nebo se má počítat celý 3h interval? Odpověď od kolegy: celý interval

# Řešení prvotních „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 přes INNER JOIN a porovnáním údajů k population jsem zjistila, že údaje v countries jsou z roku 2018, v economies z 2019 a v lookup\_table pravděpodobně z roku 2020 (jsou větší než údaje z economies z roku 2019). Pro rámcovou kontrolu jsem si ještě doplnila sloupce s meziročním růstem populace, ve kterých vidím, že růst 2019/2018 a 2020/2019 je v náhodně vybraných zemích velmi podobný, takže údaje z lookup\_table jsou s velkou pravděpodobností 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 a covid19\_tests přes country a tabulky countries, economies a religions přes country a pak tyto dvě nové tabulky přes iso3.

# První spojení tabulek

* 1. **Spojení economies s countries a religions – VIEW *v\_joined\_eco\_co\_rel***

Přes LEFT JOIN jsem z tabulek economies a countries vytvořila „stavovou“ (pro každou zemi jeden řádek, tj. bez data) tabulku joined\_economies\_countries se sloupci: country, GDP, gini, mortality\_under5, iso3, population\_density, median\_age\_2018.

*Připojení tabulky religions a výpočet*

Abych mohla stanovit podíly příslušníků jednotlivých náboženství v zemi na celkovém obyvatelstvu, musím napřed spojit tabulku religion s tabulkou obsahující informaci o celkové populaci země (nejlépe lookup\_table, kterou jsem si na začátku určila jako výchozí pro hodnoty počtu obyvatel států). V tabulce religions ale není iso, takže ji budu muset ke druhé tabulce připojit přes název země. Země v tabulce religions mají jiné názvy než v tabulce covid19\_basic\_differences, která je mou výchozí tabulkou (a se kterou mám spojenou lookup\_table), ale stejné názvy jako v tabulkách economies a countries.

Takže tabulku religions nejprve připojím pomocí LEFT JOIN přes sloupec country ke spojeným tabulkám economies a countries a uložím do VIEW ***v\_joined\_eco\_co\_rel.***

* 1. **Spojení covid19\_basic\_diff s lookup a covid19\_tests – VIEW *v\_joined\_cov\_lt\_tests\_eco\_co\_rel***

CHYBĚJÍCÍ DATA: Zjistila jsem, že tabulka covid19\_basic\_differences neobsahuje data pro 3 „významné“ země: Austrálii, Kanadu a Čínu. Data pro tyto 3 země jsem tedy vybrala z tabulky covid19\_detail\_global\_differences (musela jsem použít GROUP BY, abych získala součet všech nakažených v daném dni za všechny provincie) a pomocí UNION je připojila k tabulce covid19\_basic\_differences.

Přes LEFT JOIN jsem k této doplněné tabulce covid19\_basic\_differences připojila z tabulky lookup\_table sloupec s údajem o celkové populaci a iso3. K této tabulce jsem pomocí LEFT JOIN přes datum a ISO připojila tabulku covid19\_tests, abych získala sloupec s denními hodnotami testování.

PROBLÉM: Při kontrole, jestli spojení přes LEFT JOIN proběhlo v pořádku (výchozí tabulka a nová tabulka musí mít stejný počet řádků), jsem zjistila, že nová tabulka (po připojení covid19\_tests) 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 pomocí LEFT JOIN přes sloupec iso3 (ISO) připojila tabulku uloženou ve VIEW v\_joined\_eco\_co\_rel. Dostala jsem tak velkou „vývojovou“ (pro každou zemi řádky s jednotlivými daty) tabulku se sloupci: date (všech 468 dnů), country (všech 192 zemí), iso3, confirmed, tests (hodnoty jen za 110 zemí), population, GDP, gini, mortality\_under5, population\_density, median\_age\_2018, religion, population. Tuto tabulku jsem si uložila do VIEW: ***v\_joined\_cov\_lt\_tests\_eco\_co\_rel.***

CHYBĚJÍCÍ DATA: Jelikož tabulka covid19\_tests obsahuje data pouze pro 110 zemí, zatím co výchozí tabulka covid19\_basic\_differences 192, u 82 zemí ve výsledné tabulce nebudou data k testování.

# Přidání dalších sloupců na základě výpočtů

Na základě hodnot ve sloupci date v tabulce v\_joined\_cov\_lt\_tests\_eco\_rel do tabulky rovnou přidám sloupec s informací o víkendu a ročním období:

* 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. |

# Výpočty v tabulce weather a její připojení na výslednou tabulku

CHYBĚJÍCÍ DATA: Tabulka weather obsahuje data pouze pro 36 měst. 2 z těchto měst nejsou hlavní města žádného státu, takže ve výsledné tabulce jsou údaje o počasí k dispozici pouze pro 34 zemí.

* 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).

Použila jsem:

ROUND(SUM(CASE WHEN `time` IN ('09:00', '12:00', '15:00', '18:00') THEN 1 ELSE 0 END) \* temp) /4

Funkce k údajům z daného intervalu dosadila 1 a k údajům mimo tento interval 0. Průměr pak spočítala jako součet (SUM) všech součinů času a příslušné teploty (1 nebo 0 \* temp) podělený 4 (zajímají mě jen 4 časy).

Ve vzorci jsem pro zajištění správnosti výpočtů u hodnot ze sloupce temp odstranila „ °c“ a do výsledku ho zase přidala zpět.

* 1. **počet hodin v daném dni, kdy byly srážky nenulové**

Použila jsem:

SUM(CASE WHEN rain = '0.0 mm' THEN 0 ELSE 1 END) \* 3

Funkce do sloupce rain dosadila 0 místo 0.0 mm, resp. 1 místo ostatních hodnot. Počet hodin s nenulovými srážkami je pak součet těchto jedniček vynásobený 3 (každý řádek představuje 3h interval)

* 1. **maximální síla větru v nárazech během dne**

Jednoduchá funkce MAX(gust) při použití GROUP BY capital\_city, date.

Opět bylo potřeba pro zajištění správnosti výpočtu ve vzorci u hodnot ze sloupce gust odstranit „km/h“ a do výsledku ho zase přidat zpět (jinak, když byly hodnoty ve sloupci jako string, brala funkce MAX např. hodnotu 8 km/h větší než 42 km/h apod.).

**Připojení sloupců z tabulky weather k výsledné tabulce – VIEW *v\_joined\_cov\_lt\_tests\_eco\_co\_rel\_w***

Tabulka weather neobsahuje názvy zemí ani ISO kódy, jen názvy hlavních měst. Pro připojení k mé výsledné tabulce musím nejprve tabulku weather spojit pomocí INNER JOIN 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 (resp. POZVYHLEDST). 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 tedy novou tabulku weather\_new obsahující všechny úpravy a výpočtu, tu jsem přes INNER JOIN spojila s tabulkou countries (uložila jsem si ji do v\_weather\_new) a následně jsem tuto novou tabulku připojila přes LEFT JOIN k velké výsledné tabulce (uložila jsem si ji do ***v\_joined\_cov\_lt\_tests\_eco\_co\_rel\_w***).

# Pivotování a výpočty v tabulce life\_expectancy

Pomocí MAX(CASE WHEN year = 1965*, resp. 2015*, THEN life\_expectancy END) si přesunu hodnoty dožití z let 1965 a 2015 z řádků do sloupců, abych je od sebe mohla odečíst. Získám tak tabulku se sloupcem ISO, life\_expectancy\_1965 a life\_expectancy\_2015, kterou přes LEFT JOIN připojím k výsledné tabulce, ve které udělám výpočet.

Výslednou tabulku uložím do VIEW ***v\_Petra\_Rohlickova\_projekt\_SQL\_final***.

# Tvorba finální tabulky

CREATE TABLE t\_ Petra\_Rohlickova\_projekt\_SQL\_final AS

SELECT \*

FROM v\_Petra\_Rohlickova\_projekt\_SQL\_final

Vypsat výslednou tabulku je možné prostým:

SELECT \*

FROM t\_ Petra\_Rohlickova\_projekt\_SQL\_final

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