**PRŮVODNÍ DOKUMENTACE K SQL PROJEKTU**

31. 10. bylo na úvod zahájeno zkoumání dat ze všech pomocných tabulek pomocí funkce **describe**. Primární klíč country se nachází ve všech tabulkách kromě **weather** odkud bude potřeba na základě hlavního města převést údaje o počasí pomocí tabulek **lookup\_table** a **countries**. Druhý primární klíč date se pak nachází pouze v tabulkách **covid19\_basic\_differences** a **weather.** Projekt zahájen tvorbou parciálních selectů ke každému jednotlivému bodu v zadání. Z parciálních selektů budou následně vytvořeny pomocné tabulky, které budu na závěr spojovat jedním skriptem ve výslednou tabulku s panelovými daty.

**Popis parciálních selektů:**

* **Binární proměnná pro víkend / pracovní den:** Pro situace, kdy je datum pracovní den zvoleno číslo 0 a pro víkend číslo 1 přidáním pomocného sloupce a to za pomoci funkce **case**. Vytvořeno na základě dat z tabulky **covid19\_basic\_differences.**
* **Roční období daného dne:** Nastaveno dle dat jarní a podzimní rovnodennosti, resp. dle letního a zimního slunovratu. 0=zima, 1=jaro, 2=léto, 3=podzim, funkce **case** opětovně použita ke tvorbě unikátního sloupečku **weekend**.
* **Hustota zalidnění:** Převzatá z tabulky **countries** a pomocí funkce **round** zaokhrouhlená na dvě desetinná místa.
* **HDP na obyvatele:** Vypočítáno na základě podílu celkového HDP dané země na jejím počtu obyvatel. Data získána z tabulky **economies**, přičemž jsem vycházel z nejaktuálnějších dat, která byla za rok **2020**.
* **Gini koeficient:** Vypočítán jako průměr všech dostupných hodnot giniho koeficientů u daného státu. Data za Giniho koeficient pro jednotlivé státy jsou totiž bohužel nekompletní, resp. v některých letech u některých států naprosto chybí a mnohdy chybí i nejaktuálnější údaje z posledních 10 let. Průměrem ze všech dostupných dat z mého pohledu získáme relativně reprezentativní průřez ke každému státu.
* **Dětská úmrtnost:** Ze stejných důvodů jako Giniho koeficient vypočítána jako průměr všech dostupných hodnot dětské úmrtnosti z tabulky **economies**.
* **Medián věku obyvatel v roce 2018:** Vybrán sloupec přímo z upravené tabulky **covid19\_countries** bez nutnosti dalšího dopočtu.
* **Podíl náboženství:** Nejprve vypočítán podíl populace věřících jednotlivým náboženstvím na celkovém počtu obyvatel v dané zemi to vše z dat tabulky **religions** a za nejaktuálnější dostupný rok 2020. Následně pomocí funkcí **case a sum** převedeny jednotlivá náboženství z řádků do sloupců.
* **Rozdíl mezi očekávanou dobou dožití v roce 2015 a 1965:** Vypočítán z dat tabulky **life\_expectancy** za pomoci tvorby dvou vnořených selectů s dobou dožití za roky 2015 a 1965 a vytvořením sloupce s jejich výsledným rozdílem.
* **Průměrná denní teplota:** Data jsou dostupná v tabulce **weather**, ve sloupci temp, nicméně jsou formátována jako text, takže nejprve bylo nutné si je převést na čísla. Což jde za pomoci kombinace funkcí **cast** a **replace**. Nejprve funkcí **replace** oddělíme °C a následně pomocí funkce **cast** převedeme na číslo – datový typ **decimal**. Tím, že se jedná o průměrnou denní teplotu tak uvažujeme pouze naměřené teploty mezi 09:00 a 18:00, samotné hodnoty jsem pak spočítal pomocí fce AVG a zvolil GROUP BY po datech a zemích.
* **Počet hodin v daném dni, kdy byly srážky nenulové:**  Nejprve si tříhodinové časové intervaly, které jsou součástí tabulky **weather** pomocí funkce **case** rozdělíme za podmínky, když je déšť větší než 0,0 mm napiš 3, když ne napiš 0. Je to z toho důvodu, že údaje o času jsou v tabulce ve tříhodinových intervalech a zároveň toto rozdělení zahrnuje zjednodušující předpoklad, že celé tři hodiny pršelo – granularita dat však více neumožní. Tím si vytvoříme podpůrný sloupec pro kalkulaci, kolik hodin denně pršelo. Následně tyto hodnoty vysčítáme pomocí **SUM** a zvolíme GROUP BY po jednotlivých datech a zemích.
* **Maximální síla větru v nárazech:** Podobně jako u teploty jsem si i u větru za pomocí funkcí **cast** a **substring** nejprve oddělil číslo od textu a následně převedl číslo ve formě textu na číslo – datový typ **decimal.** Následně už stačí pouze zjistit výsledné hodnoty pomocí agregační funkce **MAX** a zvolit **GROUP BY** dle dat a zemí.

**Popis tvorby parciálních tabulek:**

Následně jsem z parciálních selektů vytvořil parciální (pomocné) tabulky, které jsem na závěr spojil v jednu. Všechny tabulky jsem tvořil tak, aby obsahovaly primární klíč **country** a tam kde bylo možno i primární klíč **date**. Některé z pomocných tabulek jsou součástí finálního skriptu, některé jsou pak pomocnými tabulkami pro tyto pomocné tabulky.

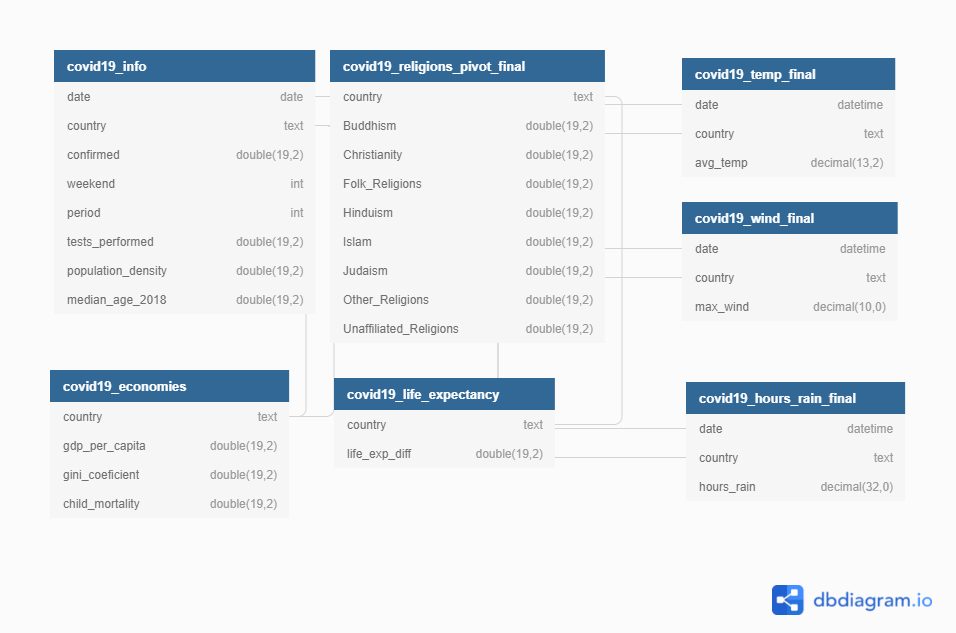
Vzhledem k tomu, že jsem si zvolil covid19\_basic\_differences jako hlavní tabulku, na kterou pomocí left join v rámci finálního skriptu páruji ostatní pomocné tabulky, řešil jsem také, aby finální výstup byl, co nejvíce granulární. Proto jsem si původní tabulky countries a covid19\_tests převedl přes klíče ISO a tabulku lookup\_table.

V tabulce weather, ze které získáváme průměrnou denní teplotu, počet hodin v daném dni se srážkami a maximální sílu větru v nárazech však není dostupný žádný z primárních klíčů, proto musíme za pomocí tabulek countries a lookup\_table, které spojíme na základě primárního klíče iso 3 převést hlavní města z tabulky weather na jednotlivé státy.

**Seznam všech tabulek:**

**covid19\_info, covid19\_economies, covid19\_religion\_rel\_share (pomocná), covid19\_religions\_pivot (pomocná), covid19\_religions\_pivot2 (pomocná), covid19\_religions\_pivot\_final, covid19\_life\_expectancy, covid19\_temp\_wind\_convert (pomocná), covid19\_countries (pomocná), covid19\_tests\_final (pomocná), covid19\_temp\_final, covid19\_wind\_final, covid19\_hours\_rain\_conversion (pomocná), covid19\_hours\_rain\_final**

**Vztahy mezi tabulkami:**



ERD odkaz: <https://dbdiagram.io/d/619a9f0002cf5d186b60eac1>.

**Zhodnocení výsledných dat:**

Ve výsledné tabulce se nachází poměrně vysoké množství NULL hodnot. Ve sloupci **confirmed** je to způsobeno tím, že u některých datumů a zemí z tabulky **covid19\_basic\_differences** nám v databázi chybí údaje o počtu potvrzených případů, podobně je na tom i sloupec **tests\_perfmored,** který čerpá data z tabulky **covid19\_tests**.

Naopak ve sloupcích **weekend** a **period** máme doplněna všechna data, protože tato byla dopočítána na základě datumů.

U geograficko-ekonomického údaje hustota zalidnění **population\_density** pak chybí údaje za následující země **Diamond Princess (loď), Kosovo, Montenegro, MS Zaandam (loď), Namibia, Taiwan\*, Timor-Leste, West Bank and Gaza.** U mediánu věku obyvatel **median\_age\_2018** pak chybí země **Bahamas, Cabo Verde, Diamond Princess (loď), Fiji, Holy See (chrám), Kosovo, Libya, Micronesia, MS Zaandam (loď), Namibia, Taiwan\*, Timor-Leste, West Bank and Gaza, Yemen.**

U HDP na obyvatele **gdp\_per\_capita** chybí údaje u následujících zemí **Bahamas, Brunei, Burma, Congo (Brazzaville), Congo (Kinshasa), Cote d'Ivoire, Cuba, Czechia, Diamond Princess, Djibouti, Eritrea, Eswatini, Holy See, Japan, Korea South, Kuwait, Liechtenstein, Marshall Islands, Micronesia, Monaco, MS Zaandam, Oman, Russia, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, San Marino, Somalia, South Sudan, Syria, Taiwan\*, United Arab Emirates, US, Venezuela, West Bank and Gaza, Yemen**, protože ve zdrojové databázi chybí údaje o jejich celkovém **HDP v roce 2020**.

Podobně i u giniho koeficientu **gini\_coeficient** chybí u některých zemí v databázi hodnoty, z naší výsledné databáze se to konkrétně týká **zemí Afghanistan, Andorra, Antigua and Barbuda, Bahamas, Bahrain, Barbados, Brunei, Burma, Cambodia, Congo (Brazzaville), Congo (Kinshasa), Cote d'Ivoire, Cuba, Czechia, Diamond Princess, Dominica, Equatorial Guinea, Eritrea, Eswatini, Grenada, Holy See, Korea, South, Kuwait, Libya, Liechtenstein, Marshall Islands, Micronesia, Monaco, MS Zaandam, New Zealand, Oman, Qatar, Russia, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, San Marino, Saudi Arabia, Singapore, Taiwan\*, US, West Bank and Gaza**. Je to z toho důvodu, že u těchto zemí v databázi nejsou žádné údaje.

Dětská úmrtnost **child\_mortality** v naší výsledné databázi chybí u zemí **Bahamas, Brunei, Burma, Congo (Brazzaville), Congo (Kinshasa), Cote d'Ivoire, Czechia, Diamond Princess, Eswatini, Holy See, Korea South, Kosovo, Liechtenstein, Micronesia, MS Zaandam, Russia, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Taiwan\*, US, West Bank and Gaza**, protože u těchto zemí rovněž nemáme k dispozici žádné historické údaje.

Podobně jako u dětské úmrtnosti nám i u **očekávané doby dožití mezi lety 2015 a 1965** v databázi chybí údaje za země **Burma, Cabo Verde, Congo (Brazzaville), Congo (Kinshasa), Cote d'Ivoire, Czechia, Diamond Princess, Eswatini, Holy See, Korea South, Kosovo, Micronesia, MS Zaandam, Russia, Taiwan\*, Timor-Leste, US, West Bank and Gaza**.

U **podílů náboženství** nejsou v databázi dostupné údaje za následující země: **Burma, Cabo Verde, Congo (Brazzaville), Congo (Kinshasa), Cote d'Ivoire, Czechia, Diamond Princess, Eswatini, Holy See ,Korea South, Micronesia, MS Zaandam, Russia, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Taiwan\*, US, West Bank and Gaza.** Opět je to z části způsobeno i diskrepancemi v názvech států v tabulkách **covid19\_basic\_differences** a **religions**.

Údaje o **počasí** máme pouze u 23 z celkových 189 zemí, které máme v datasetu. Problémem je opět to, že se v tabulce weather nacházejí údaje jen za 35 světových zemí. Údaje o počasí máme dostupné pouze u následujících zemí: **Albania, Belarus, Bulgaria, Denmark, Estonia, France, Germany, Hungary, Ireland, Latvia, Lithuania, Moldova, Netherlands, North Macedonia, Norway, Russia, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Kingdom**.

Doufám, že Ti to takto k průvodní dokumentaci bude prozatím stačit. Případně jsem k dispozici pro další dotazy.

S pozdravem,

Martin