Engeto – Popis SQL projektu

Študent: Mariana Cagáňová

Postup:

Na základe požiadavkou som si vytriedila, ktoré tabuľky budem potrebovať, ktoré údaje z nich sú pre projekt osožné a na základe čoho budem jednotlivé tabuľky spájať. Všetky tabuľky sú na seba napojené cez atribút Country a niektoré aj Date. Prvý problém nastal pri spájaní. Jednotlivé tabuľky mali rôzne názvy krajín a preto sa dobre nespojili. Aby som predišla tomuto problému porovnávala som názvy krajín v jednotlivých tabuľkách a jednotlivé názvy krajín som upravila. Najskôr som na moju hlavnú tabuľku, ktorou bolo covid19\_basic\_differencies, napojila testy, populáciu a potom pridala požadované stĺpčeky či je víkend alebo deň v týždni (1, 0) a o aké obdobie sa jedná (0, 1, 2, 3).

Ďalej som pripojila tabuľku countries, z ktorej som potrebovala priemerný vek a hustotu populácie. Z tabuľky economies som pripojila vyrátané HDP na osobu, gini koeficient a úmrtnosť detí. Koeficient gini sa v dátach objavoval veľmi sporadicky a to aj v minulých rokoch. Ani úmrtnosť detí nebola uvedené pre rok 2020 ale takmer všetky krajiny mali tento údaj uvedený v roku 2019 preto som sa rozhodla ho použiť.

Ako ďalšie bolo potrebné dodať údaje o percentuálnom podiely náboženstiev v jednotlivých krajinách. Najskôr som zistila aké jednotlivé náboženstvá sú v tabuľke religions uvedené a aby sa mi s nimi lepšie pracovalo vytvorila som si pohľad, kde je vyrátaný podiel jednotlivých náboženstiev na danú krajinu a až tento pohľad som neskôr pridala k hlavnej tabuľke.

Ako posledné bolo potrebné pridať požadované údaje o počasí. Keďže hlavná tabuľka s informáciami o covide obsahuje iba rok 2020 a 2021 aj z počasia som zobrala len tieto hodnoty. Aby bolo možné napárovať hodnoty dátumu musela som zmeniť dátový typ z Datetime na Date a zmeniť teplotu a nárazy vetra na číslo, čomu predchádzalo orezanie textu. Aj pri mestách bolo potrebné porovnať a upraviť názvy miest, pretože nie všetky sa zhodovali. Počet hodín kedy pršalo som zrátala iba hodiny, ktoré sú poskytnuté pretože nešlo určiť či v hodinách, ktoré chýbali pršalo alebo nie. Keďže dážď mal rovnaký formát nebolo potrebného ho meniť na číslo. Pre priemernú teplotu za deň som zvolia čas 6 až 18:00. Z počasia som vytvorila pohľad, ktorý som pridala k výslednej tabuľke.

Popis:

Pre zjednodušenie som si vytvorila 4 pohľady:

* v\_MC\_weather\_year – ktorý obsahuje údaje o počasí v roku 2020 a 2021 a upravuje hodnoty o počasí a nárazoch vetra aby sa s tým dalo pracovať ako s číselnou hodnotou.
* v\_MC\_weather – ktorý obsahuje údaje a medzivýpočty potrebné pre počasie. Počet hodín kedy pršalo je zrátaný počet hodín kedy boli zrážky nenulové keďže inak sa nedalo s presnosťou určiť koľko hodín medzi jednotlivými údajmi pršalo. Aby sa dáta správne napárovali bolo potrebné zmeniť typ Date z Datetime na Date a upraviť názvy krajín a miest z tabuľky countries a weather, ktoré sa používali na spojenie s tabuľkou countries a neskôr CBD.
* v\_MC\_rel\_per\_country – obsahuje podiely náboženstiev jednotlivých krajín za rok 2020
* v\_MC\_life\_expectancy\_diff – obsahuje údaje o rozdieloch medzi life expectancy v rokoch 2015 a 1965

Ako hlavnú tabuľku som použila covid19\_basic\_differencies (ďalej ako CBD), takže všetky ostatné dáta sú napojené na tieto údaje. Keďže názvy krajín medzi jednotlivými tabuľkami sa mierne líšia brala som ako hlavné názvy, názvy krajín z CBD tabuľky a s tými pracovala ďalej. Nie všetky krajiny sa ale nachádzali v potrebných tabuľkách takže údaje k 4-5 záznamom chýbajú. Taktiež tabuľka CBD obsahovala údaje o výletných lodiach, takže tieto záznamy neobsahujú údaje o počasí, krajinách atď.

Údaje v tabuľke CBD nie sú časovo obmedzené keďže obsahujú len údaje z roku 2020 a 2021 a pri rozhodnutí použiť iba rok 2020 je možné tieto údaje neskôr jednoducho vyfiltrovať. Mierne obmedzenie by to mohlo spôsobiť pri údajoch o štátoch alebo náboženstvách.

Ďalej je vo veľa krajinách chýbajúci údaj o GINI indexe. Pre určenie úmrtností detí som použila záznamy z roku 2019, keďže údaje z roku 2020 neboli dostupné.