Engeto Datová akademie SQL projekt.

## t\_jiri\_hartmann\_projekt\_SQL\_link

Pomocná spojovací tabulka **t\_jiri\_hartmann\_projekt\_SQL\_link** obsahuje sloupec country\_1 = country z tabulky covid19\_basic\_differences, sloupec iso3 z tabulky lookup\_table, sloupec country\_2 = country z tabulky countries a sloupec capital\_city z tabulky countries .

V tabulce lookup\_table nejsou informace o „státech“ [Diamond Princess](https://en.wikipedia.org/wiki/COVID-19_pandemic_on_Diamond_Princess) a [MS Zaandam](https://en.wikipedia.org/wiki/MS_Zaandam), což jsou výletní lodě.

V tabulce countries nejsou státy „Kosovo“ a „Taiwan\*“. Kosovo je pouze částečně uznaný stát, pro který nejsou k dispozici výsledky v tabulce covid19\_tests (pravděpodobně zahrnut v Srbsku). V případě státu Taiwan\* jde o chybu, kterou vyřeším ve skriptu, který vytvoří spojovací tabulku (Taiwan\* = Taiwan). Sloupec capital\_city obsahuje chyby, které ale nezasahují do států, které obsahuje tabulka weather. Sloupec capital\_city bude sloužit s propojením s tabulkou weather, kde jsou údaje pouze pro 24 států. V případě rozšíření záznamů v tabulce weather bude nutné zrevidovat a doplnit údaje o hlavních městech.

-- t\_jiri\_hartmann\_projekt\_SQL\_link

**create** **or** **replace** **table** **t\_jiri\_hartmann\_projekt\_SQL\_link** (

country\_1 **varchar**(127) **not** **null** **primary** **key**,

iso3 **varchar**(127),

country\_2 **varchar**(127),

capital\_city **varchar**(127)

)

**as**

(**select** **distinct** cov.country **as** "country\_1", loo.iso3,

**if**(cov.country = "Taiwan\*", "Taiwan", cou.country) **as** "country\_2",

cou.capital\_city

**from** covid19\_basic\_differences **as** cov

**left** **join**

(**select** country, iso3 **from** lookup\_table **where** province **is** **null**) **as** loo

**on** cov.country = loo.country

**left** **join**

(**select** country, capital\_city, iso3 **from** countries) **as** cou

**on** loo.iso3 = cou.iso3

**order** **by** cov.country

);

Doba běhu 6,6 sekundy, 190 řádků.

**První řádky pomocné tabulky:**

|  |  |  |  |
| --- | --- | --- | --- |
| **country\_1** | **iso3** | **country\_2** | **capital\_city** |
| Albania | ALB | Albania | Tirana |
| Algeria | DZA | Algeria | Alger |
| Andorra | AND | Andorra | Andorra la Vella |
| Andorra | AND | Andorra | Andorra la Vella |

## t\_jiri\_hartmann\_projekt\_SQL\_religion

Další pomocná tabulka **t\_jiri\_hartmann\_projekt\_SQL\_religion** obsahuje setříděné údaje o náboženstvích dle zadání. V pomocné tabulce je nahrazen název státu - názvem státu ve formátu používaném v tabulce covid19\_basic\_differences (country\_1).

Při kontrole dat jsem nalezl nekonzistenci v názvech zemí:

|  |  |
| --- | --- |
| **country\_2** | **country v tabulce religions** |
| Fiji Islands | Fiji |
| Holy See (Vatican City State) | Vatican City |
| Saint Kitts and Nevis | St. Kitts and Nevis |
| Saint Lucia | St. Lucia |
| Saint Vincent and the Grenadines | St. Vincent and the Grenadines |

Jedná se o malé země, které budou mít zanedbatelný vliv na výsledky. V důležité tabulce s počty testů jsou uvedena data pouze pro Fiji.

-- t\_jiri\_hartmann\_projekt\_SQL\_religion

**create** **or** **replace** **table** **t\_jiri\_hartmann\_projekt\_SQL\_religion** (

country\_1 **varchar**(127) **not** **null** **primary** **key**,

buddhism **float**(5,2),

christianity **float**(5,2),

folk\_religions **float**(5,2),

hinduism **float**(5,2),

islam **float**(5,2),

judaism **float**(5,2),

other\_religions **float**(5,2),

unaffiliated\_religions **float**(5,2)

) **as**

(**select** lin.country\_1, rel.buddhism, rel.christianity, rel.folk\_religions,

rel.hinduism, rel.islam, rel.judaism, rel.other\_religions, rel.unaffiliated\_religions

**from**

(**select** country,

**round**(**sum**(**if**(religion = "Buddhism", population, **Null**))/**sum**(population)\*100, 2) **as** "buddhism",

**round**(**sum**(**if**(religion = "Christianity", population, **Null**))/**sum**(population)\*100, 2) **as** "christianity",

**round**(**sum**(**if**(religion = "Folk Religions", population, **Null**))/**sum**(population)\*100, 2) **as** "folk\_religions",

**round**(**sum**(**if**(religion = "Hinduism", population, **Null**))/**sum**(population)\*100, 2) **as** "hinduism",

**round**(**sum**(**if**(religion = "Islam", population, **Null**))/**sum**(population)\*100, 2) **as** "islam",

**round**(**sum**(**if**(religion = "Judaism", population, **Null**))/**sum**(population)\*100, 2) **as** "judaism",

**round**(**sum**(**if**(religion = "Other Religions", population, **Null**))/**sum**(population)\*100, 2) **as** "other\_religions",

**round**(**sum**(**if**(religion = "Unaffiliated Religions", population, **Null**))/**sum**(population)\*100, 2) **as** "unaffiliated\_religions"

**from** religions **where** `year` = 2020 **and** country **in** (**select** country\_2 **from** t\_jiri\_hartmann\_projekt\_SQL\_link)

**group** **by** country

) rel

**left** **join** t\_jiri\_hartmann\_projekt\_SQL\_link **as** lin

**on** rel.country = lin.country\_2

**order** **by** lin.country\_1 )

;

Doba běhu 100 ms, 182 řádků.

**První řádky pomocné tabulky:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **country** | **buddhism** | **christianity** | **folk\_religions** | **hinduism** | **islam** | **judaism** | **other\_religions** | **unaffiliated\_religions** |
| Afghanistan | 0,00 | 0,10 | 0,00 | 0,02 | 99,80 | 0,00 | 0,07 | 0,00 |
| Albania | 0,00 | 16,82 | 0,00 | 0,00 | 82,24 | 0,00 | 0,00 | 0,93 |
| Algeria | 0,00 | 0,17 | 0,02 | 0,00 | 97,96 | 0,00 | 0,00 | 1,84 |
| Andorra | 0,00 | 100,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |

## t\_jiri\_hartmann\_projekt\_SQL\_weather

Poslední pomocná tabulka **t\_jiri\_hartmann\_projekt\_SQL\_weather** obsahuje vypočítané údaje o počasí v hlavních městech států dle zadání. V pomocné tabulce je nahrazeno hlavní město názvem státu ve formátu používaném v tabulce covid19\_basic\_differences (country\_1).

-- t\_jiri\_hartmann\_projekt\_SQL\_weather

**create** **or** **replace** **table** **t\_jiri\_hartmann\_projekt\_SQL\_weather** (

`date` **date** **not** **null**,

country\_1 **varchar**(127) **not** **null**,

avg\_daily\_temp **float**(5,2),

rainy\_hours **integer**,

max\_gust **integer**,

**primary** **key** (`date`, country\_1)

) **as**

(**select** w.`date`, l.country\_1 , w.avg\_daily\_temp, w.rainy\_hours, w.max\_gust **from**

(**select** `date`, city, **sum**(**if**(`hour` **in** (6, 9, 12, 15), temp, **Null**)) / 4 **as** "avg\_daily\_temp",

**sum**(**if**(rain <> 0, 3, 0)) **as** "rainy\_hours", **max**(gust) **as** "max\_gust"

**from** weather

**where** "2020-01-21" < `date` -- 2020-01-22 is first day in incovid19\_basic\_differences table

**group** **by** `date`, city ) **as** w

**left** **join** t\_jiri\_hartmann\_projekt\_SQL\_link **as** l

**on** w.city = l.capital\_city

**where** l.capital\_city **is** **not** **null**

**order** **by** w.`date`, l.country\_1)

;

Doba běhu 1,3 sekundy, 9.656 řádků.

**První řádky pomocné tabulky:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **`date`** | **country\_1** | **avg\_daily\_temp** | **rainy\_hours** | **max\_gust** |
| 22.01.2020 | Albania | 8,50 | 0 | 8 |
| 22.01.2020 | Austria | 0,50 | 0 | 31 |
| 22.01.2020 | Belarus | 2,75 | 21 | 46 |
| 22.01.2020 | Belgium | 3,75 | 0 | 16 |

Informace o populaci jsou uloženy ve 3 tabulkách lookup\_table, country a economies. Porovnal jsem údaje na vybraném vzorku zemí s údaji na webu worldometers.info. Nejpřesnější údaje jsou v tabulce lookup\_table:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Population | **lookup\_table** | | countries | | economies | | **worldometers.info** |
| Brazil | **212 559 409** | -0,27 | 209 469 333 | -1,72 | 211 049 527 | -0,98 | **213 139 380** |
| Czechia | **10 708 982** | -0,07 | 10 629 928 | -0,81 | 10 669 709 | -0,44 | **10 716 648** |
| Germany | **83 783 945** | -0,12 | 82 905 782 | -1,17 | 83 132 799 | -0,90 | **83 887 186** |
| India | **1 380 004 385** | -0,39 | 1 352 617 328 | -2,36 | 1 366 417 754 | -1,37 | **1 385 343 756** |
| China | **1 404 676 330** | -2,55 | 1 392 730 000 | -3,38 | 1 397 715 000 | -3,04 | **1 441 464 310** |
| US | **329 466 283** | -0,69 | 326 687 501 | -1,53 | 328 239 523 | -1,06 | **331 748 637** |

Dále jsem porovnal vypočtenou hustotu obydlení uloženou v tabulce countries s nově vypočítanou pomocí sloupce population v tabulce lookup\_table a sloupcem surface\_area z tabulky countries. Rozdíly jsou nevýznamné:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **surface\_area** | **population\_density** | **vypočtená population density** |
| Brazil | 8 547 403 | 25 | 25 |
| Czechia | 78 866 | 138 | 136 |
| Germany | 357 022 | 237 | 235 |
| India | 3 287 263 | 455 | 420 |
| China | 9 572 900 | 148 | 147 |
| US | 9 363 520 | 36 | 35 |

vazby na tabulku **t\_jiri\_hartmann\_projekt\_SQL\_covid19**:

**countries** – **iso3** (existuje ve všech řádcích tabulky, žádný klíč, obsahuje country2, capital\_city)

**covid19\_tests** – **ISO** (existuje ve všech řádcích tabulky, žádný klíč, obsahuje country2)

**economies** – **country2** (existuje ve všech řádcích tabulky, klíč index, neobsahuje iso3)

**life\_expectancy** – **iso3** (existuje null pro kontinenty (vyloučit), žádný klíč, obsahuje country2)

**religions** – **country2** (existuje ve všech řádcích tabulky, klíč year, region, country, religion, neobsahuje iso3)

**weather** – **city** (existuje ve všech řádcích tabulky, klíč city, date, hour, neobsahuje iso3 ani country)

Population:

lookup\_table, country, economies

Surface area - population

Další možné ukazatele:

avg\_height – country – možný vliv výšky

continent – country – možný vliv odlišné rasy

elevation – country – možný vliv nadmořské výšky

t\_jiri\_hartmann\_projekt\_SQL\_final