Engeto Datová akademie SQL projekt.

[**CVIČENÍ PRO DATOVOU AKADEMII**](https://engeto.com/cs/kurz/cviceni-pro-datovou-akademii)​/[**PROJEKTY**](https://engeto.com/cs/kurz/cviceni-pro-datovou-akademii/studium/ki_EaCh4TgqvSpEFYMIUPA/projekty)​/[**SQL**](https://engeto.com/cs/kurz/cviceni-pro-datovou-akademii/studium/DbO2pMl8SIykg9ucGXCxdA/projekty/sql)

**Zadání:** Od Vašeho kolegy statistika jste obdrželi následující email:

Dobrý den,

snažím se určit faktory, které ovlivňují rychlost šíření koronaviru na úrovni jednotlivých států. Chtěl bych Vás, coby datového analytika, požádat o pomoc s přípravou dat, která potom budu statisticky zpracovávat. Prosím Vás o dodání dat podle požadavků sepsaných níže.

Výsledná data budou panelová, klíče budou stát (country) a den (date). Budu vyhodnocovat model, který bude vysvětlovat denní nárůsty nakažených v jednotlivých zemích. Samotné počty nakažených mi nicméně nejsou nic platné - je potřeba vzít v úvahu také počty provedených testů a počet obyvatel daného státu. Z těchto tří proměnných je potom možné vytvořit vhodnou vysvětlovanou proměnnou. Denní počty nakažených chci vysvětlovat pomocí proměnných několika typů. Každý sloupec v tabulce bude představovat jednu proměnnou. Chceme získat následující sloupce:

1. Časové proměnné
   * binární proměnná pro víkend / pracovní den
   * roční období daného dne (zakódujte prosím jako 0 až 3)
2. Proměnné specifické pro daný stát
   * hustota zalidnění - ve státech s vyšší hustotou zalidnění se nákaza může šířit rychleji
   * HDP na obyvatele - použijeme jako indikátor ekonomické vyspělosti státu
   * [GINI koeficient](https://en.wikipedia.org/wiki/Gini_coefficient) - má majetková nerovnost vliv na šíření koronaviru?
   * dětská úmrtnost - použijeme jako indikátor kvality zdravotnictví
   * medián věku obyvatel v roce 2018 - státy se starším obyvatelstvem mohou být postiženy více
   * podíly jednotlivých náboženství - použijeme jako proxy proměnnou pro kulturní specifika. Pro každé náboženství v daném státě bych chtěl procentní podíl jeho příslušníků na celkovém obyvatelstvu
   * rozdíl mezi očekávanou dobou dožití v roce 1965 a v roce 2015 - státy, ve kterých proběhl rychlý rozvoj mohou reagovat jinak než země, které jsou vyspělé už delší dobu
3. Počasí (ovlivňuje chování lidí a také schopnost šíření viru)
   * průměrná denní (nikoli noční!) teplota
   * počet hodin v daném dni, kdy byly srážky nenulové
   * maximální síla větru v nárazech během dne

Napadají Vás ještě nějaké další proměnné, které bychom mohli použít? Pokud vím, měl(a) byste si vystačit s daty z následujících tabulek: countries, economies, life\_expectancy, religions, covid19\_basic\_differences, covid19\_testing, weather, lookup\_table.

V případě nejasností se mě určitě zeptejte.

S pozdravem, Student (a.k.a. William Gosset)

**Výstup:** Pomozte Vašemu kolegovi s daným úkolem. Výstupem by měla být tabulka na databázi, ze které se požadovaná data dají získat jedním selectem. Tabulku pojmenujte t\_{jméno}\_{příjmení}\_projekt\_SQL\_final. Na svém GitHub účtu vytvořte repozitář (může být soukromý), kam uložíte všechny informace k projektu - hlavně SQL skript generující výslednou tabulku, popis mezivýsledků, informace o výstupních datech (například kde chybí hodnoty apod.). Případné pomocné tabulky neukládejte na DB jako view! Vždy vytvořte novou tabulku (z důvodu anonymity).

## Řešení

## 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.

-- Spojovací tabulka: 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),

population **integer**

)

**as**

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

**case** **when** cov.country = "Taiwan\*" **then** "Taiwan"

**when** cov.country = "Kosovo" **then** "Kosovo"

**else** cou.country **end** **as** "country\_2",

cou.capital\_city, loo.population

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

**left** **join**

(**select** country, iso3, population **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 5,6 sekundy, 190 řádků.

**První řádky pomocné tabulky:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **country\_1** | **iso3** | **country\_2** | **capital\_city** | **population** |
| Afghanistan | AFG | Afghanistan | Kabul | 38928341 |
| Albania | ALB | Albania | Tirana | 2877800 |
| Algeria | DZA | Algeria | Alger | 43851043 |
| Andorra | AND | Andorra | Andorra la Vella | 77265 |

## t\_jiri\_hartmann\_projekt\_SQL\_eco

Pomocná spojovací tabulka **t\_jiri\_hartmann\_projekt\_SQL\_eco** obsahuje sloupec country\_2 = country z tabulky covid19\_basic\_differences.

Jsou v ní uloženy nejčerstvější údaje gini, GDP a mortality\_under5. Informace o tom, z kterého roku údaje jsou je možno vyčíst z přiložené tabulky Chybějící data.

-- Spojovací tabulka: t\_jiri\_hartmann\_projekt\_SQL\_eco

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

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

gini **float**(5,2),

GDP\_mil **integer**,

mortaliy\_under5 **float**(5,2)

)

**as**

(**select** lin.country\_2 **as** "country", gin.gini, **round**(gdp.GDP/1000000) **as** "GDP\_mil", mor.mortaliy\_under5 **from** t\_jiri\_hartmann\_projekt\_SQL\_link lin

**left** **join**

(**select** country, first\_value(gini) **over** (**partition** **by** country **order** **by** `year` ) **as** "gini"

**from** economies

**where** gini **is** **not** **NULL**

**group** **by** country ) **as** gin

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

**left** **JOIN**

(**SELECT** country, first\_value(GDP) **over** (**partition** **by** country **order** **by** `year` ) **as** "GDP"

**FROM** economies

**where** GDP **is** **not** **NULL**

**group** **by** country ) **as** gdp

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

**left** **join**

(**SELECT** country, first\_value(mortaliy\_under5) **over** (**partition** **by** country **order** **by** `year` ) **as** "mortaliy\_under5", `year`

**FROM** economies

**where** mortaliy\_under5 **is** **not** **NULL**

**group** **by** country ) **as** mor

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

**where** lin.country\_2 **is** **not** **NULL**

**order** **by** lin.country\_2

);

## 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.

-- Náboženství: 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\_1** | **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).

-- Počasí: 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 |

# Populace

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. Protože je to jediný údaj z tabulky lookup\_table, který ještě potřebuji, zkopíroval jsem sloupec population do pomocné tabulky t\_jiri\_hartmann\_projekt\_SQL\_link:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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 mezi tabulkami

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

není obsažen stát Congo (Kinshasa) iso3= COD, Kosovo iso3=XKS, Sri Lanka iso3=LKA, Taiwan iso3 = TWN,

některé státy nemají vyplněné údaje.

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

Tabulka obsahuje duplicity nad sloupci date a ISO. Výsledný soubor má být na tyto sloupce unikátní. To je zajištěno primárním klíčem přes oba tyto sloupce. Počty testů jsem v duplicitních řádcích sečetl.

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

Stát Bahamas má jako jméno vyplněno Bahamas, The a nedohledá se. To stejné platí i pro Brunei Darussalam, Cabo Verde, Fiji. Některé státy nemají vyplněné údaje, některé státy zcela chybí. Vhodná je korekce dat.

**life\_expectancy** – **iso3** (žádný klíč, obsahuje country2)

Chybí údaje pro Kosovo a lodě.

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

Chybí údaje pro Kosovo a lodě a země začínajíci na Saint.

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

# Výsledná tabulka

-- Výsledná tabulka: t\_jiri\_hartmann\_projekt\_SQL\_final

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

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

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

confirmed **integer**,

tests\_performed **integer**,

is\_weekend **integer**,

season **integer**,

surface\_area **integer**,

median\_age\_2018 **float**(5,2),

GDP\_mil **integer**,

gini **float**(5,2),

mortaliy\_under5 **float**(5,2),

life\_expectancy\_difference\_2015\_1965 **float**(5,2),

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),

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

rainy\_hours **integer**,

max\_gust **integer**,

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

)

**as**

(**select** cov.`date`, cov.country, cov.confirmed, tes.tests\_performed,

**case** **when** **WEEKDAY**(cov.`date`) **in** (5,6) **then** 1 **else** 0 **end** **as** is\_weekend,

**case** **when** **month** (cov.`date`) **in** (1,2,3) **then** 0

**when** **month** (cov.`date`) **in** (4,5,6) **then** 1

**when** **month** (cov.`date`) **in** (7,8,9) **then** 2

**else** 3 **end** **as** season,

cou.surface\_area, cou.median\_age\_2018,

eco.GDP\_mil, eco.gini, eco.mortaliy\_under5,

**round**(lif.life\_expectancy\_difference\_2015\_1965,2) **as** "life\_expectancy\_difference\_2015\_1965",

rel.buddhism, rel.christianity, rel.folk\_religions, rel.hinduism, rel.islam, rel.judaism, rel.other\_religions, rel.unaffiliated\_religions,

wea.avg\_daily\_temp, wea.rainy\_hours, wea.max\_gust

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

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

**on** cov.country = lin.country\_1

**left** **join**

(**select** iso3, surface\_area, median\_age\_2018 **from** countries) **as** cou

**on** lin.iso3 = cou.iso3

**left** **join**

(**select** country, GDP\_mil, gini, mortaliy\_under5 **from** t\_jiri\_hartmann\_projekt\_SQL\_eco ) **as** eco

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

**left** **join**

(**select** le2015.iso3, le2015.life\_expectancy-le1965.life\_expectancy **as** "life\_expectancy\_difference\_2015\_1965" **from**

(**select** iso3, life\_expectancy **from** life\_expectancy **where** `year` = 2015) **as** le2015

**left** **join**

(**select** iso3, life\_expectancy **from** life\_expectancy **where** `year` = 1965) **as** le1965

**on** le2015.iso3 = le1965.iso3) **as** lif

**on** lin.iso3 = lif.iso3

**left** **join** t\_jiri\_hartmann\_projekt\_SQL\_religion rel

**on** cov.country = rel.country\_1

**left** **join** t\_jiri\_hartmann\_projekt\_SQL\_weather wea

**on** cov.`date` = wea.`date` **and** cov.country = wea.country\_1

**left** **join**

(**select** `date`, ISO, tests\_performed **from** covid19\_tests **group** **by** `date`, ISO) **as** tes

**on** lin.iso3 = tes.ISO **and** cov.`date` = tes.`date`

)

;

Doba běhu 28 sekund, 55.670 řádků.

### Další možné ukazatele:

avg\_height – country – možný vliv výšky

continent – country – možný vliv odlišné rasy

elevation – country – možný vliv nadmořské výšky