SQL projekt: Covid-19 ve světě

Engeto – Datová Akademie, 18. 6. 2024 – 3. 9. 2024: dobrovolný projekt po skončení kurzu

Dokončeno: 19. 3. 2025

Zpracovala: Adéla Prystaszová

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:

- časové proměnné
    - binární proměnná pro víkend / pracovní den
    - roční období daného dne (zakódujte prosím jako 0 až 3)
- 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 - 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
- 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_tests, weather, lookup_table.

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_{jmeno}_{prijmeni}_projekt_SQL_final. Na svém GitHub účtu vytvořte repozitář, kam uložíte všechny informace k projektu - hlavně SQL skript generující výslednou tabulku, popis mezivýsledků, informace o výstupních datech (například kde chybí hodnoty apod.). Neupravujte data v primárních tabulkách, pokud bude potřeba transformovat hodnoty, dělejte tak až v tabulkách nebo pohledech, které si nově vytváříte.

Data

Vstupní tabulky, jejich obsah relevantní v zadaném úkolu a specifika:

- **covid19_basic_differences**: Obsahuje denní počty nově potvrzených případů covidu 19 v jednotlivých státech v období 22. 1. 2020 až 23. 5. 2021.
- **covid19_tests**: Obsahuje počet vykonaných testů na covid-19 v období od 29. 1. 2020 do 21. 11. 2020. Proměnná entity vyjadřuje, o jaký typ počtu testů se jedná. Může jít o celkový počet vykonaných testů, počet unikátních otestovaných osob nebo počet otestovaných vzorků, zároveň je zohledněno, zda se jedná pouze o PCR testy. Typ může být též nejasný. U převážné většiny států se vyskytují data za pouze jeden typ testů (z největší části jde o celkový počet vykonaných PCR testů), u několika sse ale vyskytují dva typy.
- **lookup_table**: Obsahuje seznam unikátních států a provincií včetně ISO kódů, je možné ji využít jako číselník.
- **countries**: Obsahuje seznam států, jejich hlavních měst a také medián věku v roce 2018.
- **economies**: Obsahuje počet obyvatel jednotlivých států, jejich HDP, giniho koeficient a dětskou úmrtnost v letech 1960–2020, posledním rokem s daty za tyto proměnné je ale většinou rok 2019.
- **life_expectancy**: Obsahuje naději dožití při narození jednotlivých států za roky 1950–2020.
- **religions**: Obsahuje za jednotlivé státy počet osob vyznávajících křesťanství, islám, hinduismus, buddhismus, lidová náboženství, judaismus, ostatní náboženství a osoby bez náboženského vyznání.
- **weather**: Obsahuje za jednotlivá města údaje o počasí měřené denně v 00:00, 03:00, 06:00, 09:00, 12:00, 15:00, 18:00 a 21:00. Tabulka obsahuje za období 1. 5. 2016 až 30. 4. 2021 údaje o teplotě, množství srážek a rychlosti větru v nárazech. Jde nicméně o stringové proměnné obsahující i jednotky.

Postup a mezivýsledky

1) Byla provedena **kontrola potenciálně problematických proměnných**:
    a) **country**: Tabulky je třeba napojovat na číselník lookup_table přes proměnnou country nebo iso/iso3 a bylo tedy ověřováno, jestli se ve všech tabulkách vyskytují státy pod stejnými názvy.
    b) **entity v covid19_tests:** Bylo zjištěno 8 států, k nimž byly v tabulce data za dvě hodnoty entity.
2) **Úprava tabulek**:

a) Z lookup_table byla vytvořena tabulka **lookup_table2**, která je číselníkem pouze za státy, ne za provincie, které se jinde v datech nevyskytují.
b) Z covid19_tests byla vytvořena tabulka **covid19_tests2**, ve které se názvy států shodují s názvy států v tabulce lookup_table a zároveň je v ní ponechán vždy jen jeden typ záznamů podle proměnné entity.
c) Z weather byla vytvořena tabulka **weather2**, která obsahuje potřebné proměnné v numerické, nikoli stringové podobě.

3) **Tvorba výsledné tabulky**:
a) Byl vytvořen view **v1**, v němž byly na lookup_table2 (z ní vybrány proměnné stát, iso3) pomocí left joinu napojeny tabulky countries (hlavni_mesto, hustota_zalidneni, median_veku_2018), economies (populace_2019, HDP_2019, giniho_koeficient_2019, detska_umrtnost_2019), life_expectancy (nadeje_doziti_1965, nadeje_doziti_2015).
b) Byl vytvořen view **v2**, v němž byly na v1 pomocí RIGHT JOINU napojeny tabulky religions (nabozenstvi, nabozenstvi_populace) a covid19_basic_differences (datum, den_v_tydnu, mesic, pocet_novych_pripadu).
c) Byly vytvořeny 3 viewy **pocasi1**, **pocasi2** a **pocasi3**, ve kterých jsou z tabulky weather2 vytvořeny agregované proměnné prumerna_denni_teplota, hodiny_s_destem a max_naraz_vitr.
d) Byl vytvořen view **v3**, v němž byly na v2 pomocí LEFT JOINU napojeny tabulka covid19_tests2 (pocet_testu) a viewy pocasi1 (prumerna_denni_teplota), pocasi2 (pocet_h_s_destem) a pocasi3 (max_naraz_vitr).
e) Byla vytvořena výsledná tabulka **t_adela_prystaszova_projekt_SQL_final**, která čerpala z viewu v3 a v níž jsou všechny potřebné závislé i nezávislé proměnné včetně nově vypočítaných.

Výstup

Výstupem je tabulka **t_adela_prystaszova_projekt_SQL_final** obsahující následující proměnné:

Column Name	#	Data Type	Not Null	Auto Increment	Key	Default
A-Z stat	1	text	[ ]	[ ]		NULL
datum	2	date	[ ]	[ ]		NULL
123 mira_incidence	3	double	[ ]	[ ]		NULL
123 nove_pripady_ku_testum	4	double	[ ]	[ ]		NULL
123 vikend	5	int(1)	[ ]	[ ]		NULL
A-Z rocni_obdobi	6	varchar(1)	[ ]	[ ]		NULL
123 hustota_zalidneni	7	double	[ ]	[ ]		NULL
123 hdp_obyv_2019	8	double	[ ]	[ ]		NULL
123 giniho_koeficient_2019	9	double	[ ]	[ ]		NULL
123 detska_umrtnost_2019	10	double	[ ]	[ ]		NULL
A-Z nabozenstvi	11	text	[ ]	[ ]		NULL
123 podil_nabozenstvi	12	double	[ ]	[ ]		NULL
123 rozdil_e0_1965_2015	13	double	[ ]	[ ]		NULL
123 prumerna_denni_teplota	14	decimal(14,4)	[ ]	[ ]		NULL
123 pocet_h_s_destem	15	bigint(21)	[ ]	[ ]		0
123 max_naraz_vitr	16	int(11)	[ ]	[ ]		NULL

- **stat**: Všechny státy světa pocházející z číselníku lookup_table.
- **datum**: 21. 1. 2020 – 23. 5. 2021
- **závislé proměnné**
    - **mira_incidence**: Míra incidence vypočtená jako podíl denního počtu nových případů covidu-19 a počtu obyvatel. Problematické nicméně je, že počet obyvatel se nevztahuje ke stejnému dni jako počet nových případů, ale k roku 2019.
    - **nove_pripady_ku_testum**: Podíl denního počtu nových případů a počtu provedených testů v daném dni.
- **nezávislé proměnné**
    - **vikend**: Nabývá hodnot 0 (všední den) a 1 (víkend).
    - **rocni_obdobi**: Nabývá hodnot 0 (jaro) až 3 (zima).
    - **hustota_zalidneni**: Hustota zalidnění, zřejmě k roku 2018.
    - **hdp_obyv_2019**: HDP na obyvatele v roce 2019.
    - **giniho_koeficient_2019**: Giniho koeficient vyjadřující míru nerovnosti příjmů, k roku 2019.
    - **detska_umrtnost_2019**: Úmrtnost do 5 let, 2019.
    - **nabozenstvi, podil_nabozenstvi**: 8 náboženských vyznání a podíl jejich vyznavačů na populaci daného státu. Vypočítáno z počtu vyznavačů daného náboženství v roce 2020 a počtu obyvatel daného státu v roce 2019.
    - **rozdil_e0_1965_2015**: Rozdíl naděje dožití při narození mezi roky 2015 a 1965.
    - **prumerna_denni_teplota**: Průměrná teplota naměřená daného dne v hlavním městě daného státu v časech 06:00, 09:00, 12:00, 15:00, 18:00 a 21:00. Jednotkami jsou stupně Celsia.
    - **pocet_h_s_destem**: Počet časů, ve kterých byly srážky vyšší, než 0 mm. Počítáno z 8 časů, v nichž během dne došlo ke měření.
    - **max_naraz_vitr**: Maximální rychlost větru v nárazech během dne. Jednotkami jsou km/h.

Vzhledem k tomu, že tabulka obsahuje data z 8 různorodých výchozích tabulek a jedná se o data o všech státech světa za období delší než rok, není možné, aby byla naprosto kompletní. Téměř u všech proměnných se tedy vyskytují hodnoty NULL, které značí, že hodnota za dané datum a stát chybí.