Online Data Academy
Engeto

SQL Projekt

Lenka Mlýnková

Cíl projektu: Příprava dat k určení faktorů, které mohou ovlivňovat šíření koronaviru na úrovni jednotlivých států.

Použité tabulky:

- *countries,*
- *economies,*
- *life_expectancy,*
- *religions,*
- *covid19_basic_differences,*
- *covid19_tests,*
- *weather*

Postup:

- Jako hlavní jsem si určila tabulku covid19_basic_differences, protože obsahuje hlavní data pro tuto analýzu.
- Vytvořila jsem si selekty pro jednotlivé body (detaily viz níže).
- Bylo nutné opravit zdrojové tabulky, lišily se názvy států.
    - porovnávala jsem pomocí tohoto dotazu:

```
SELECT DISTINCT cbd.country, ct.country
FROM covid19_basic_differences cbd
LEFT JOIN covid19_tests ct
ON cbd.country=ct.country
WHERE ct.country IS NULL;
```

    - zkopírovala jsem si zdrojovou tabulku
    - opravila záznamy v kopii:

```
UPDATE Lenka_covid19_tests
SET country = 'Congo (Kinshasa)'
WHERE country = 'Democratic Republic of Congo';
```

- Jednotlivé selekty jsem spojila do celku pomocí dočasných tabulek (WITH).
- Z dočasných tabulek jsem vytvořila tabulku *t_Lenka_Mlynkova_SQL_projekt.*

Jednotlivé sloupce:

- *Počet provedených testů*
  - Použila jsem hodnotu z tabulky *covid19_tests.* Tato tabulka v porovnání s *covid19_basic_differences* obsahuje podstatně méně zemí, proto některé hodnoty u tests_performed jsou NULL.
- *Časové proměnné*
  - weekend - 0 pro všední den a 1 pro víkend
  - přidala jsem day_name, pro mou kontrolu a určení konkrétního dne
  - season - roční období jsem rozdělila dle meteorologického ročního období viz. tabulka níže (0-jaro, 1 - léto, 2 - podzim, 3-zima)
    - nevzala jsem v potaz rozdíl mezi jižní a severní polokoulí

meteorologické roční období	období	měsíce
jaro	1. března - 31. května	březen, duben, květen
léto	1. června - 31. srpna	červen, červenec, srpen
podzim	1. září - 30. listopadu	září, říjen, listopad
zima	1. prosince - 28. února (v přestupném roce 29. února)	prosinec, leden, únor

- *Proměn*né specifické pro daný stát
  - Hustota zalidnění (population_density) a medián věku obyvatel v roce 2018 (median_age_2018) jsou hodnoty z tabulky *countries.*
  - GDP na obyvatele je hodnota HDP z tabulky *economies* přepočtena na počet obyvatel, za nejaktuálnější rok pro daný stát.
  - GINI koeficient a dětská úmrtnost (mortality_under5) jsou hodnoty z tabulky *economies* za nejaktuálnější rok pro daný stát.
  - Podíl náboženství (Christianity, Islam, Buddhism, Folk_Religions, Hinduism, Judaism, Other_Religions, Unaffiliated_Religions) jsem určila jako hodnotu pro dané náboženství/počet obyvatel * 100.
    - Tabulce *religion* je podle mého názoru, tabulkou s odhady počtů věřících. Jsou tam hodnoty za roky 2010, 2020, 2030 atd. a hodnoty věřících jsou zaokrouhleny na vysoké čísla. Nesedí ani součet všech věřících pro danou zemi, kde převyšuje počet obyvatel až o 2 mil.
    - Proto některé hodnoty budou více jak 100%. Nezměnila jsem nic, svou vypovídající hodnotu tyto údaje mají.
  - Rozdíl mezi očekávanou dobou dožití v roce 1965 a v roce 2015 (LifeExpDif) je rozdíl těchto hodnot v daném roce.
- *Počasí*
  - Průměrnou denní teplotu (averageDayTemp) jsem určila z hodnot v 6 h,9 h,12 h,15 h a v 18 h.
  - Počet hodin s deštěm (RainyHours) jsem určila podle počtu měření s deštěm * 3 (intervaly mezi měřeními).
  - Pro Max hodnotu větru v nárazech (MaxGustWind) jsem použila funkci max ().

- Hodnoty pro počasí jsou z tabulky *weather*, která je jen pro Evropu a pouze do října/listopadu roku 2020
- Tabulku *weather* jsem napojila na *countries*, pomocí hlavních měst a pomocí výše zmíněného dotazu jsem zjistila a opravila názvy měst. Rozdíl byl v jazyce zápisu hlavních měst.

Opravy tabulek:

Toto jsou státy, které jsem opravila z *tabulek countries, economies, life_expectancy, religion.*

	country	country
1	Bahamas	[NULL]
2	Brunei	[NULL]
3	Burma	[NULL]
4	Congo (Brazzaville)	[NULL]
5	Congo (Kinshasa)	[NULL]
6	Cote d'Ivoire	[NULL]
7	Czechia	[NULL]
8	Diamond Princess	[NULL]
9	Eswatini	[NULL]
10	Holy See	[NULL]
11	Korea, South	[NULL]
12	MS Zaandam	[NULL]
13	Micronesia	[NULL]
14	Russia	[NULL]
15	Saint Kitts and Nevis	[NULL]
16	Saint Lucia	[NULL]
17	Saint Vincent and th	[NULL]
18	Taiwan*	[NULL]
19	US	[NULL]
20	West Bank and Gaza	[NULL]

Brunei = Brunei Darussalam
Burma = Myanmar
Congo (Brazzaville) = Konžská republika - francouzská kolonie = Congo
Congo (Kinshasa) = Konžská demokratická republika - belgická kolonie = The Democratic Republic of Congo
Diamond Princess a MS Zaandam = jsou lodě, nepoužila jsem
Eswatini = Swaziland
Holly see = Vatikán
The west bank and gaza - neumím přiřadit
Bahamas = Bahamas, The
Brunei = Brunei Darussalam
Cote d'Ivoire = Ivory Coast
Czechia = Czech Republic
Korea, South = South Korea
Russia = Russian federation
Taiwan - nemám k čemu přiřadit
Timor-Leste = East Timor

Shrnutí:
Bylo by pro mě jednodušší změnit pouze názvy v tabulce *covid19_basic_differences.* To jsem zjistila, když už jsem měla více tabulek změněných. Vidím jako náročné porovnávat hodnoty v tabulkách, když chybí unikátní klíče a jazyk zápisu hodnot se liší. Zaznamenala jsem, že není k dipsozici údaj o počtu nakažených v Číně, jsou pouze data o provedených testech (ve výsledné tabulce nejsou).