# Faktory, které ovlivňují rychlost šíření koronaviru na úrovni jednotlivých států

**Projekt SQL**

#### Kurz: Datová analýza

**Engeto Lenka Baklíková**

## Obsah

Obsah

[Faktory, které ovlivňují rychlost šíření koronaviru na úrovni jednotlivých států 1](#_Toc89702287)

[Obsah 2](#_Toc89702288)

[1 Zadání projektu 89](#_Toc89702289)

[2 Cíl 90](#_Toc89702290)

[3 První část – porozumění datům 90](#_Toc89702291)

[3.1 Definice úkolu, výstupu 90](#_Toc89702292)

[3.2 Zdrojové tabulky 90](#_Toc89702293)

[3.3 Postup 91](#_Toc89702294)

[4 Druhá část – zpracování dat 91](#_Toc89702295)

[4.1 Vysvětlovaná proměnná – Daily\_confirmed\_cases 91](#_Toc89702296)

[4.2 Vysvětlující proměnné 92](#_Toc89702297)

[1.1.1 Časové proměnné 92](#_Toc89702298)

[1.1.2 Proměnné specifické pro daný stát 92](#_Toc89702299)

[1.1.3 Počasí (ovlivňuje chování lidí a také schopnost šíření viru) 93](#_Toc89702300)

[5 Třetí část – kombinace výsledných dat 95](#_Toc89702301)

[6 Čtvrtá část - shrnutí 96](#_Toc89702302)

[1.1 Problémy a možné omezení datové sady 96](#_Toc89702303)

[1.2 Legenda obsažených dat 96](#_Toc89702304)

[1.3 Doporučení pro práci s datovou sadou 97](#_Toc89702305)

## Zadání projektu

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 -** [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 vice
* **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\_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.).

## Cíl

Cílem je porozumět zadání, určit si postup, připravit data, která jsou vhodná pro statistické zpracování a připravit průvodní informacemi ohledně možných datových omezeních.

## První část – porozumění datům

### Definice úkolu, výstupu

Úkol: připravit datovou sadu pro statistické oddělení

Výstup: hodnoty pro zadané vysvětlující a vysvětlované proměnné dle země a data

### Zdrojové tabulky

*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.*

Data: pro zpracování využijeme tabulky s následujícími daty.

Červeně vyznačené jsou sloupce, které se opakují ve vice než jedné zdrojové tabulce**.**

|  |  |  |  |
| --- | --- | --- | --- |
| **countries** | **Economies** | **weather** | **covid19\_tests** |
| country | country | index | country |
| abbreviation | year | weather | entity |
| avg\_height | GDP | time | date |
| calling\_code | population | temp | ISO |
| capital\_city | gini | feels | cumulative |
| continent | taxes | wind | tests\_performed |
| currency\_name | fertility | gust | **covid19\_basic\_differences** |
| religion | mortaliy\_under5 | rain | date |
| currency\_code | **life\_expectancy** | humidity | country |
| domain\_tld | index | cloud | confirmed |
| elevation | country | pressure | deaths |
| north | iso3 | vis | recovered |
| south | year | date |  |
| west | life\_expectancy | city |  |
| east | **religions** | **lookup\_table**. |  |
| government\_type | year | country |  |
| independence\_date | country | province |  |
| iso\_numeric | region | UID |  |
| landlocked | religion | iso2 |  |
| life\_expectancy | population | lookup\_table |  |
| national\_symbol |  | iso3 |  |
| national\_dish |  | code3 |  |
| **population\_density** |  | FIPS |  |
| countries |  | admin2 |  |
| population |  | lat |  |
| region\_in\_world |  | long |  |
| surface\_area |  | combine\_key |  |
| yearly\_average\_temperature |  | population |  |
| median\_age\_2018 |  |  |  |
| iso2 |  |  |  |
| iso3 |  |  |  |

### Postup

Postup zpracování projektu

* Definice proměnných
  + Vyjádření vysvětlované proměnnépo částech a spojení dat
  + Vyjádření vysvětlujících proměnných po částech a spojení dat
  + Kombinace výsledných dat
  + Informace o výstupních datech

## Druhá část – zpracování dat

### Vysvětlovaná proměnná – Daily\_confirmed\_cases

*“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. „*

Vysvětlovaná proměnná: Variable\_Cases\_tests\_population

Denní nárust případů covidu dle země a data.

Vysvětlovaná proměnná je složená ze tří částí

* Počty nakažených (denní nárusty nakažených) - Cases\_per\_day
* Počty provedených testů - Tests\_per\_day
* Počet obyvatel v daných státech - population\_by\_country\_2020

Postup: vybrat dílčí proměnné (view) a kombinace dat do výsledné tabulky.



SQL script je zmíněny v přiloženém excel souboru záložka Vysvětlovaná proměnná

* Zahrnuté data za měsíce 2.2020-6.2020 na základě porovnání zdrojových dat.
* Vybrané pouze státy, kde máme informace o potvrzených případech a testování
* Ve vybraných zemích (tabulka covid19\_basic\_differences) se vyskytl název země, který obsahuje speciální znak \*. - Taiwan\* - hodnota byla nahrazená
* Entity - Info o typu testování- Na základě obsaženého popisu budeme pokračovat s následujícím výběrem.
  + Zahrnuté Tested: people tested, people tested (incl. non-PCR), samples tested, tests performed, tests performed (incl. non-PCR)
  + Nezahrnuté: units unclear, units unclear (incl. non-PCR)
* Data ohledně populace států jsou zahrnuty za rok 2020

Kalkulace vysvětlované proměnné

1. Ratio\_confirmed\_tested - (t.tests\_performed /c.confirmed) poměr mezi provedenými testy a potvrzenými případy
2. Ratio\_Confirmed\_tested\_per\_population\_100000 - poměr mezi provedenými testy a potvrzenými případy na 100 000 obyvatel (t.tests\_performed /c.confirmed)/p.population\_per\_100000

### Vysvětlující proměnné

Vysvětlující proměnné

* Časové proměnné
* Proměnné specifické pro daný stát
* Počasí

### Časové proměnné

Weekend indicator - binární proměnná pro víkend / pracovní den

Season - roční období daného dne (zakódujte prosím jako 0 až 3)



Data ohledně SQL script jsou v záložce excel souboru „Vysvětlující proměnné\_čas“

* Zahrnuté data za měsíce 2.2020-6.2020 na základě porovnání zdrojových dat.

### Proměnné specifické pro daný stát

Proměnné specifické pro stát:

* *population\_density* - hustota zalidnění - ve státech s vyšší hustotou zalidnění se nákaza může šířit rychleji
* *GDP* – HDP na obyvatele-použijeme jako indikátor ekonomické vyspělosti státu
* *GINI* - GINI koeficient - má majetková nerovnost vliv na šíření koronaviru?
* mortaliy\_under5 - dětská úmrtnost - použijeme jako indikátor kvality zdravotnictví
* median\_age\_2018 - věku obyvatel v roce 2018 - státy se starším obyvatelstvem mohou být postiženy vice
* Religion\_ratio – 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
* life\_expectancy\_1965-2015- 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

|  |  |  |
| --- | --- | --- |
| **Popis** | **View** | **Zdrojová tabulka** |
| **Proměnné pro daný stát** | **T\_Lenka\_countries\_all\_variables** | **countries life\_expectancy economies religions** |
| population\_density, median\_age\_2018 | V\_Lenka\_Countries\_variable | countries |
| rozdíl mezi očekávanou dobou dožití v roce 1965 a v roce 2015 | V\_Lenka\_life\_expectancy\_variable | life\_expectancy |
| HDP na obyvatele GINI koeficient  dětská úmrtnost (mortality\_under5) | V\_Lenka\_economies\_variables | economies |
| podíly jednotlivých náboženství | v\_Lenka\_religion\_variable | religions |

SQL script je obsažený v přiloženém excelu sheet „vysvětlující proměnné\_stát“

* Některé státy zahrnuté v tabulce countries variable jsou značené jiným názvem – přejmenováno („Czech Republic“ to „Czechia, Micronesia“, „Federated States of“ to „Micronesia“)

### 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



SQL script je obsažený v přiloženém excelu sheet „vysvětlující proměnné\_čas“

**Přiřazení denní a noční teploty**

Denní čas – 6h-18h (6h, 9h,12h,15)

Noční čas – 18h-6h (18h,21h,0h,3h)

* Zahrnuté data za měsíce 2.2020-6.2020 na základě porovnání zdrojových dat.

Problémy:

1. formát dat sloupečků

-Temp – nahrazeny hodnoty °c za nic

- Rain - Nahrazeno mm za nic

2. Tabulka weather neobsahuje informace o státě (country) ale pouze hlavní město.

- Proto jsme přiřadili k tabulce názvy hlavních měst (tak, aby korespondovali s tabulkou countries) pro možnost vyjádřené státu

Obsah obrázku stůl

Popis byl vytvořen automaticky

Obsah obrázku stůl

Popis byl vytvořen automaticky

|  |  |  |
| --- | --- | --- |
| **Country** | **City (COUNTRIES)** | **City (Weather)** |
| **Austria** | Wien | Vienna |
| **Belgium** | Bruxelles [Brussel] | Brussels |
| **Finland** | Helsinki [Helsingfors] | Helsinki |
| **Greece** | Athenai | Athens |
| **Italy** | Roma | Rome |
| **Luxembourg** | Luxembourg [Luxemburg/L | Luxembourg |
| **Poland** | Warszawa | Warsaw |
| **Portugal** | Lisboa | Lisbon |
| **Romania** | Bucuresti | Bucharest |
| **Ukraine** | Kyiv | Kiev |
| **Czech republic** | Praha | Prague |

## Třetí část – kombinace výsledných dat

V třetí části zkombinujeme data do výsledné tabulky.

*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.).*

Výsledná data jsou kombinace následujících proměnných.

SQL script je obsažený v přiloženém excelu sheet „t\_Lenka\_Baklikova\_projekt\_SQL“



Přehled všech zahrnutých hodnot:



## Čtvrtá část - shrnutí

### Problémy a možné omezení datové sady

1. **Problém s formátem dat** - při vyjádření proměnné bylo potřeba změnit formát dat temp °c - odstranili jsme v datech znak °c, rain odstranili jsme mm
2. **Problém – velké množství dat – čas** – zpracování a spojování tabulek zabírá spoustu času, proto jsem vytvořila několik tabulek dle charakteru zvlášť a následně jsou spojené do jedné finální tabulky.
3. **Problém – odlišně značená klíčová data (státy, města) -** pro přesnou interpretaci a spojení dat bylo potřeba porovnat klíčová data a upravit tabulku tak, aby byly ve stejném formátu a bylo možné přiřadit správné hodnoty ke správnému státu.
4. **Problém pořadí sloučení tabulek -** U tabulky, která se týká počasí jsem měla problém s tím,

### Legenda obsažených dat

Zahrnuté proměnné ve výsledných datech

### Doporučení pro práci s datovou sadou

* **Mínusové hodnoty** (otázka je zda se jedná o validní data dle dní – k tomuto bychom potřebovali vice informací o datech – přidat ke komentáři).
* **Ne všechny proměnné jsou dostupné pro zahrnuté státy** - dle potřeby možné zohlednit zúžení výběru
* **Negativné hodnoty přírustku denních nakazených** tabulka covid19\_basic\_differences a Covid19\_tests obsahují negativní hodnoty. Může se jednat o korekce.
* **Typ provedených testů** V tabulce Tests\_performed je zahrnuté detailnější rozdělení testů ve sloupci entity. Na základě obsaženého popisu budeme pokračovat s následujícím výběrem.Zahrnuté date Tested**:** people tested, people tested (incl. non-PCR), samples tested, tests performed, tests performed (incl. non-PCR)Nezahrnuté data: units unclear, units unclear (incl. non-PCR)
* **Země** Požadovaná data nemáme k dispozici pro všechny státy. Některé ze zemí nejsou porovnatelné. V rámci práce byly některé státy přejmenovány. Pokud bychom ale chtěli přesnější data je možné projít více do detailů.
* **Hodnoty, za rok 2020** Kde bylo možné zahrnuli jsme data z roku 2020 – například population
* **Datum – Při** porovnání dat, které jsou uvedené ve zdrojových tabulkách jsem vypsala pouze data, která jsou v rozmezí 2.2020–6.2020. Zahrnutí dalších hodnot by nemělo vypovídací hodnotu.
* **GINI koeficient – Máme** k dispozici málo dat, zvážila bych vybrání jiné proměnné, která by dokázala vysvětlit majetkovou nerovnost.