**Dokumentace SQL projekt Datová Akademie Engeto Hana Voznicová**

**Cílem projektu bylo vytvořit podklady pro určení faktorů, které ovlivňují rychlost šíření koronaviru na úrovni jednotlivých států.**

**Pro zpracování jsem využila tabulky:**

1. covid19\_basic\_differences

2. covid19\_testing

3. countries

4. economies

5. religions

6. lookup\_table

7. weather

8. life\_expectancy

**Požadovanou tabulku jsem vytvořila těmito kroky:**

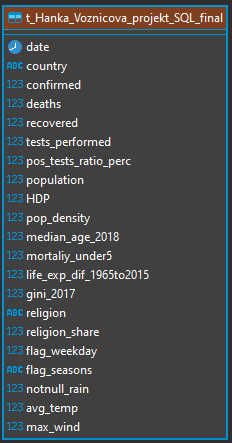
1. vytvořila jsem si dílčí selekty v rámci jednotlivých tabulek,

2.využila jsem klausule WITH pro vytvoření dočasných tabulek, které tvořily dílčí selekty,

3.dočasné tabulky jsem spojila,

4. opravila jsem data ve zdrojových tabulkách z důvodů nekonzistentnosti, více informací níže,

5. z výsledného scriptu jsem vytvořila tabulku **t\_Hanka\_Voznicova\_projekt\_SQL\_final**.



**Informace k tabulce t\_Hanka\_Voznicova\_projekt\_SQL\_final**

Za hlavní zdrojovou tabulku jsem považovala tabulku Covid19\_basic\_difference. Bylo to z důvodu, že za nejdůležitější data považuji údaje o **potvrzených případech, úmrtích a uzdravených**. V této tabulce jsou data do 3.6.2021.

Další použitou tabulkou byla tabulka covid19\_testing, ze které jsem využila **počet provedených testů**. V této tabulce jsou data do 21.11.2020.

Následně jsem data z těchto dvou tabulek využila pro výpočet „ **pos\_tests\_ratio\_perc**“, což je denní poměr pozitivní testů. .

Ve výsledné tabulce pak následují sloupečky s údaji:

* **celkové populaci**,
* **HDP**-zde jsou nejnovější data z roku 2019,
* **hustotě zalidnění**,
* **mediánu věku**(data z roku 2018),
* **děstké úmrtnosti do 5 let**-zde jsou také data z roku 2019,
* **rozdíl mezi očekávanou dobou dožití v roce 1965 a v roce 2015**,
* **GINI koeficint** - data pro GINI koeficient jsem použila z roku 2017, protože toto je rok, kdy je záznam alespoň pro 68 zemí. V pozdějších letech jich je o dost méně.

Dalšími jsou údaje **o poměrech náboženství** v jednotlivých zemích. Pro výpočet poměru jsem nevyužila údaj o populaci z tabulky countries, ale součet počtu jednotlivých nábožeství v dané zemi. Např. u Německa je součet různých nábožeství v tabulce religions 80 540 000, oproti údaji z tabulky countries, kde je populace v Německu 82 905 782. Což je rozdíl cca 2 500 000 obyvatel. Tzn. pokud bych použila údaj o populaci z tabulky countries, poměry nábožeství by byly zkreslené.

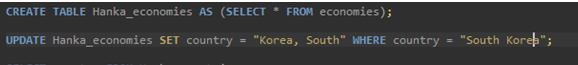
Následuje sloupce **flag\_weekday**, který nám říká zda daný den je dnem v pracovním týdnu (pondělí až pátek), označeno 1, nebo víkendovým dnem, který je označeno 0.

**Flag\_seasons** udává o jaké roční období se podle data jedná. Při určení jsem nezohlednila zeměpisnou délku a šířku. Jaro je označeno 0, léto 1, podzim 2 a zima je 3.

Tabulku ukončují sloupce s daty z tabulky weather. Jedná se o:

1. **počet hodin v daném dni, kdy byly srážky nenulové**,
2. **průměrnou denní teplotu** – pro výpočet jsem použila hodnoty z časů 6, 9, 12, 15, 18, 21hod. a vypočítala jejich průměr,
3. **maximální nárazy větru**.

Po spojení dočasných tabulek, jsem při kontrole zjistila problém, že ne všechny údaje z dílčích selektů se spojily a odhalila jsem, že problém je ve spojování přes sloupec country. Některé země měly názvy rozdílné v různých tabulkách. Proto jsem si pro každou tabulku vytvořila seznam zemí a pomocí excelu a funkce Vlookup jsem hledala shodu v názvu, resp. kde je název jiný. Zemí bylo více, ale za nejdůležitější jsem pro opravu vybrala Českou republiku, Rusko, USA a Severní Koreu. Jelikož jsem v tu chvíli už měla joiny hotové, úpravu dat jsem provedla ne v hlavní tabulce covid19\_basic\_difference, ale v tabulkách countries, economies, religions,life\_expectancy, covid\_tests. Následně jsem si vytvořila nové tabulky s označením Hanka\_, které jsou kopie původních a v nich nahradila názvy tímto způsobem.



**Země, které jsem měnila:**

Czech Republic Czechia

Russia Federation Russia

South Korea Korea, South

United States US

Stejná situace s rozdílnými názvy byla i u tabulky weather, kterou jsem přes sloupec city chtěla spojit s tabulkou countries a sloupcem capital\_city. Pro tento případ jsem použila stejný postup jako výše u zemí a nahradila jsem v tabulce countries názvy, tak aby se shodovaly s tabulkou weather. Tabulka weather obsahuje 35 měst, přičemž je jen 34 z nich hlavními městy. Toto je příčinou toho, že u většiny zemí jsou hodnoty ve sloupcích notnull\_rain, avg\_temp, max\_wind rovny NULL.

Cílem bylo dodat tabulku, kde můžu požadovaná data získat jedním selektem, což bylo splněno.