# SQL project

Po přečtení zadání jsem pochopil, že je potřeba vytvořit pomocnou tabulku tří proměnných, ze kterých bude vytvořena velká tabulka, kde se budou nacházet vysvětlované proměnné.

Tři proměnné jsou:

* denní nárůsty nakažených v jednotlivých zemích
* počet provedených testů
* počet obyvatel daného státu

Začal jsem zkoumat jednotlivé tabulky, zjistil jsem, že v zadání mezi tabulkami, které bychom měli používat je napsáno covid19\_testing. Taková tabulka v databázi není, jedná se asi o překlep, byla tím nejspíše myšlena tabulka covid19\_tests.

Rozhodl jsem se, že nejdříve spojím tabulky covid19\_basic\_differences AS cbd a covid19\_tests AS ct. Tyto dvě tabulky mají společné země a datumy. Zjistil jsem, že cbd má:

* 189 zemí (country)
* 488 datumů (date)

ct má:

* 110 zemí (country)
* 329 datumů (date)

Spojovat tyto dvě tabulky budu pomocí funkce LEFT JOIN, přes sloupce `country` a `date`, z tabulky cbd dostanu informace o denních nárůstech a z tabulky ct informace o provedených testech. Výsledná tabulka bude obsahovat mnoho NULL hodnot, z důvodu rozdílného počtu zemí a datumů.

Vytvořil jsem první tabulku, bohužel jsem zjistil, že je to špatně. Obyčejný LEFT JOIN nestačí, bude třeba vymyslet něco komplikovanějšího, protože několik hodnot z tabulky ct není v nové vytvořené tabulce.

Napadá mě vytvoření tabulky, kde budou všechny země a datumy, následně vytřídím státy, které se vyskytují v tabulce 2x akorát pod jinými názvy. Pak začnu na tuhle tabulku připojovat confirmed, tests\_performed a population.

Po zkoušení různých možností, které selhaly mě na padlo, že udělám tabulku LEFT JOIN jako na začátku, k tomu vytvořím tabulku s zeměmi a daty, které jsou obsaženy v tabulce ct a ne v tabulce cbd pomocí EXCEPT. Následně vytvořím sloupce stejné jako má první vytvořená tabulka pomocí LEFT JOIN a z toho pomocí UNION vytvořím novou tabulku, která bude obsahovat všechny země a datumy.

Další den mě napadlo spojit tabulky s využitím CASE pomocí, kterého změním názvy zemí v tabulce ct, které se neshodují s názvy zemí v tabulce cbd. Po konzultaci s Matějem mě to navedlo ještě k vytvoření VIEW, kde jsem CASE uplatnil a doufám, že správně.

Vytvořil jsem pomocné tabulky a VIEW, pomocí který jsem aktualizoval tabulku cbd\_and\_cttpm z důvodu chybějících zemí, které se vyskytovaly v původní tabulce covid19\_tests ale nevyskytovaly se v původní tabulce covid19\_basic\_differences. Vytvořil jsem nejdřív TABLE abych mohl následně udělat EXCEPT (tím zjistit, co chybí v mojí tabulce), vytvořil jsem tabulku s EXCEPT a následně spojil pomocí UNION s tabulkou, kde nám chybělo pár států.

Vše jsem vydropoval a znova založil, protože jsem přehlédl Jiždní Koreu, která se zapisuje dvěma způsoby. Musel jsem upravit původní VIEW, tak jsem jsem vše vydropoval a znova založil. Dále jsem přejmenoval SQL script na pokus omyl (dotazy v tomto scriptu nedávají smysl, je to pouze pro mě, když si ověřuji nějaké info z tabulek).

Vytvořil jsem novou tabulku se sloupcem, který nám říká jestli se jedná o záznam z vikendu nebo z pracovního dne. Použil jsem CASE a premiérově weekday, doufám, že správně.

Vytvořil jsem další novou tabulku, kde jsem přidal pomoci CASE sloupec Rocni\_obdobi.

Vytvořil jsem VIEW, kde jsem upravil stejné státy s rozdílným pojmenováním abych mohl napojit hustotu zalidnění na moji vytvořenou tabulku.

Přidal jsem sloupec do mnou vytvořené tabulky, hustota zalidnění.

Přemýšlím, jak napojit sloupce HDP na obyvatele a GINI koeficient na mnou vytvořenou tabulku. U HDP na obyvatele mě napadá, že budu muset vytvořil pro každý stát 2 ukazatele, protože tabulka obsahuje záznamy s daty, které sahají do roku 2020 a 2021, udělám HDP na obyvatele za rok 2019 a přiřadím k záznamům roku 2020 a HDP na obyvatele za rok 2020 přidělím pro rok 2021.

GINI koeficient je trochu oříšek, protože se počítá kumulativně a nějaký rok se zvedne, někdy klesne … tak nelze použit agregační funkci MAX, nedávalo by to smysl. Další problém je, že chybí hodnoty. U některých sátů není vůbec a někde je například poslední záznam u dané země 4 nebo 5 let zpátky. Nenapadá mě nic lepšího než data ze sloupce gini koeficinet pro každou zemi jednoduše zprůměrovat, abychom měli aspoň nějakou orientační hodnotu a s touto hodnotou dále pokračovat.

Vytvořil jsem VIEW s HDP na obyvatele a následně spojit s mnou vytvořenou tabulkou. Následně jsem to DROPnul, napadl mě elegantnější příkaz bez využití VIEW, výsledek je stejný.

Vytvořil jsem nový sloupec gini koeficient, pomocí vnořeného selectu a LEFT JOINU jsem připojil průmer gini ze všech zaznamů dané zeme k záznamům v mojí tabulce. Nejsem si zde uplně jistý, jestli by se s Gini ukazatelem nemělo pracovat jinak ale vzhledem k chybějícím hodnotám mě nenapadlo jiné řešení.

Připojuji sloupec dětská úmrtnost (mortaliy\_under5), v tomto sloupci za rok 2020 nejsou žádná data, pouze nullové hodnoty. Takže vyberu rok 2019, kde jsou hodnoty skoro téměř všude a zároveň jsou data k dnešnímu datu nejvíce aktuální.

Připojuji sloupec meadian\_age\_2018 k mnou vytvořené tabulce, použil jsem LEFT JOIN a vnořený SELECT abych připojil správná data i na státy, které mají v tabulkách rozdílné označení.

Narazil jsem na úkol, kde musim napojit podíl jednotlivých náboženství na každou zemi. Ze zadání jsem pochopil, že budu muset vytvořit 8 sloupců, kde budou jednotlivé typy náboženství a do každého sloupce přiřadit podíl lidí, kteří dané náboženství vyznávají. Výsledkem tedy bude, že všechny země budou mít uvedený podíl vyznavačů všech náboženství.

Pro každé náboženství jsme si vytvořil pomocnou tabulku, kde je podíl věřících daného náboženství v dané zemi. Následně jsem všech 8 tabulek připojil na mnou vytvořenou tabulku. Respektive jsem vždy vytvořil pomocnou tabulku a hned připojil, takhle jsem to udělal 8x.

Vytvořil jsem sloupec diff\_life\_expectancy, který se mi podařilo napojit na mnou vytvořenou tabulku jedním příkazem. Začínám si uvědomovat, že jsem nejspíš v průběhu projektu jsem vytvořil několik tabulek a meziTabulek zbytečně, hodně věci se dalo zapsat jednodušeji do jednoho obsáhlejšího příkazu.

Napojení sloupce prum\_den\_tep. Bylo potřeba vytahnout informace z tabulky weather, v tabulce nebyly názvy států ale pouze názvy měst, takže jsem musel vytvořit pomocnou tabulku countries\_modified\_2, z tabulky weather a countries, názvy zemí jsem upravil tak aby to následně pasovalo na mou hlavní tabulku. Následně jsem vytvořil tabulku weather\_2, kde bylo potřeba upravit vyjmout z hodnoty znak °c a napojit na countries\_modified\_2. Potom jsem zprůměroval denní teplotu v pomocné tabulce prum\_denni\_tep, kterou jsem v podmínce vymezil pouze na den. Následně jsem napojit na mnou vytvořenou tabulku přes názvy zemí a datumy.

Napojení sloupce počet hodin v daném dni, kdy byly srážky nenulové, jelikož tabulka weather má záznamy o každém dni ve tříhodinových intervalech, tak jsem zvolil postup, že pokud v záznamu byla srážka nenulová, tak to je jakoby 3 hodiny pršelo. Vytvořil jsem sloupec, kde je hodnota 3 když pršelo a hodnota 0 když nepršelo. Následně jsem seSUMoval tento sloupce a pomocí GROUP BY sumy přiřadil ke správným datům a městům. V datech byla chyba, pár měst bylo v původní tabulce se stejným datem 2x. Takže bylo pár dní, kdy pršelo více, jak 24h. Problém jsem vyřešil, že pokud byl počet hodin deště nad 24h za den, tak jsem tuto sumu vydělil 2. Výsledek pak dával smysl číselně i principem. Jelikož byl záznam v původní tabulce 2x. Pak jsem to celé napojil podle data a země na mou původní tabulku.