**PRŮVODNÍ DOKUMENTACE K SQL PROJEKTU**

31. 10. bylo na úvod zahájeno zkoumání dat ze všech pomocných tabulek pomocí funkce **describe**. Primární klíč country se nachází ve všech tabulkách kromě **weather** odkud bude potřeba na základě hlavního města převést údaje o počasí pomocí tabulek **lookup\_table** a **countries**. Druhý primární klíč date se pak nachází pouze v tabulkách **covid19\_basic\_differences** a **weather.** Projekt zahájen tvorbou parciálních selectů ke každému jednotlivému bodu v zadání. Z parciálních selektů budou následně vytvořeny pomocné tabulky, které budu na závěr spojovat ve výslednou tabulku.

**Popis parciálních selektů:**

* **Binární proměnná pro víkend / pracovní den:** Pro situace, kdy je datum pracovní den zvoleno číslo 0 a pro víkend číslo 1 a to za pomoci funkce **case**. Vytvořeno na základě tabulky covid19\_basic\_differences.
* **Roční období daného dne:** Nastaveno dle dat jarní a podzimní rovnodennosti, resp. dle letního a zimního slunovratu. 0=zima, 1=jaro, 2=léto, 3=podzim opětovně použita ke tvorbě unikátního sloupečku funkce **case**.
* **Hustota zalidnění:** Převzatá z tabulky countries a zaokhrouhlená na dvě desetinná místa.
* **HDP na obyvatele:** Vypočítáno na základě podílu celkového HDP dané země a jejího počtu obyvatel. Data získána z tabulky economies, přičemž jsem vycházel z nejaktuálnějších dat, která byla za rok 2020.
* **Gini koeficient:** Vypočítán jako průměr všech dostupných hodnot giniho koeficientů u daného státu. Data za Giniho koeficient pro jednotlivé státy jsou totiž bohužel nekompletní, resp. v některých letech u některých států naprosto chybí a průměrem z mého pohledu získáme relativně reprezentativní průřez ke každému státu.
* **Dětská úmrtnost:** Vypočítána jako průměr všech dostupných hodnot z tabulky economies.
* **Medián věku obyvatel v roce 2018:** Vybrán přímo z tabulky countries bez nutnosti dalšího dopočtu.
* **Podíl náboženství:** Nejprve vypočítán podíl populace věřících jednotlivým náboženstvím na celkovém počtu obyvatel v dané zemi to vše z dat tabulky religions a za dostupný rok 2010. Následně pomocí funkcí **case a sum** převedeny jednotlivá náboženství z řádků do sloupců.
* **Rozdíl mezi očekávanou dobou dožití v roce 2015 a 1965:** Vypočítán za pomoci tvorby dvou vnořených selectů s dobou dožití za roky 2015 a 1965 a jejich následným rozdílem.
* **Průměrná denní teplota:** Data jsou v tabulce weather, ve sloupci temp formátována jako text, takže nejprve bylo nutné si je převést na čísla. Což jde za pomoci kombinace funkcí cast a replace. Nejprve funkcí replace oddělíme °C a následně pomocí funkce cast převedeme na číslo. Tím, že se jedná o průměrnou denní teplotu tak uvažujeme pouze naměřené teploty mezi 09:00 a 18:00, samotné hodnoty jsem pak spočítal pomocí fce AVG a zgrupoval po datech a zemích.
* **Počet hodin v daném dni:**  Nejprve si tříhodinové časové intervaly, které jsou součástí tabulky weather pomocí funkce **case** rozdělíme za podmínky, když je déšť větší než 0,0 mm napiš 3, když ne napiš 0. Tím si vytvoříme podpůrný sloupec pro kalkulaci, kolik hodin denně pršelo. Následně tyto hodnoty vysčítáme a zgrupujeme po jednotlivých datech a zemích.
* **Maximální síla větru v nárazech:** Podobně jako u teploty jsem si i u větru za pomocí funkcí cast a substring nejprve oddělil v rámci textu číslo a následně převedl číslo ve formě textu na čísla. Následně už stačí pouze vyMAXovat a zgrupovat dle dat a zemí.

Následně jsem z parciálních selektů vytvořil parciální (pomocné) tabulky, které jsem na závěr spojil v jednu.

Z tabulky countries bude potřeba pro vyřešení projektu získat kromě sloupce country také sloupce **population\_density** a **median\_age\_2018**.

Z tabulky economies jsou pro splnění zadání klíčové kromě primárního klíče (country) potřeba také informace o HDP na osobu, které získáme jako celkové **HDP dané země / počet obyvatel** a dále také **giniho koeficient** a **dětská úmrtnost**.

Pro zisk podílů jednotlivých náboženství je potřeba nejdříve převést údaje z tabulky religions na

Pro výpočet průměrné denní teploty je potřeba propojit tabulku **weather** s tabulkou **countries** na základě hlavního města, tak abychom ve výsledném seznamu dostali seřazené průměrné denní teploty podle našich hlavních klíčů **date** a **country**. Ne všechna hlavní města v tabulce countries jsou dostupná i v tabulce **weather**. V tabulce **countries** však některé země mají jiný název než v tabulce **covid19\_basic\_differences**. Bude potřeba postupovat přes propojení iso3 z lookup\_table.

Tabulka **religions** obsahuje větší množství zemí než je celkový počet zemí na světě, protože jako zemi bere např. i ostrovy patřící některým zemím.

**Každá z propojovaných tabulek má jiný počet zemí.**

Stejně tak jsou zde rozdíly mezi tabulkami Covid 19 basic differences a countries, kdy se v jednotlivých tabulkách některé země jmenují jinak, např. v Covid 19 basic differences se ČR jmenuje Czechia, zatímco v countries je to Czech Republic.

Při napojení tabulek countries a weather jsou viditelné rozdíly, kdy v countries chybí 11 ze 35 hlavních měst uvedených v tabulce weather. Tím pádem data za tyto země budou ve výsledné tabulce chybět. Celkově tady data o počasí po dnech jsou k dispozici jen u 24 zemí z výsledné tabulky.

Země z tabulky countries je nutno převést přes lookup table (primární klíč iso3), která poskytuje vyšší granularitu dat vůči základní tabulce covid19\_basic differences.

Roční období jsem zakódoval jako (0=jaro, 1=léto, 2=podzim,3=zima).

Weather: **Truncated incorrect DOUBLE value: '1 °c'**

Pomocí describe, Engeto pokročilé konstrukty – projít tabulky, se kterými pracuji.

**INFORMACE O VÝSTUPNÍCH DATECH**

Součet počtu věřících z tabulky **religions** dle jednotlivých náboženství neodpovídá celkové velikosti populace jednotlivých států dostupné z tabulek ostatních. Proto ve výsledné tabulce procentuální součet věřících dle náboženství u jednotlivých států nedává 100 %.

Tabulka weather obsahuje údaje pouze za 35 hlavních měst a tabulka countries za 235 (zjištěno pomocí count či select distinct), nebylo tedy možné získat údaje o počasí za všechny státy.