SQL projekt

Zadání projektu: Data o mzdách a cenách potravin a jejich zpracování pomocí SQL

## Zadání projektu

### Úvod do projektu

Na vašem analytickém oddělení nezávislé společnosti, která se zabývá životní úrovní občanů, jste se dohodli, že se pokusíte odpovědět na pár definovaných výzkumných otázek, které adresují **dostupnost základních potravin široké veřejnosti**. Kolegové již vydefinovali základní otázky, na které se pokusí odpovědět a poskytnout tuto informaci tiskovému oddělení. Toto oddělení bude výsledky prezentovat na následující konferenci zaměřené na tuto oblast.

Potřebují k tomu **od vás připravit robustní datové podklady**, ve kterých bude možné vidět **porovnání dostupnosti potravin na základě průměrných příjmů za určité časové období**.

Jako dodatečný materiál připravte i tabulku s HDP, GINI koeficientem a populací **dalších evropských států** ve stejném období, jako primární přehled pro ČR.

### Datové sady, které je možné použít pro získání vhodného datového podkladu

**Primární tabulky:**

1. czechia\_payroll – Informace o mzdách v různých odvětvích za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.
2. czechia\_payroll\_calculation – Číselník kalkulací v tabulce mezd.
3. czechia\_payroll\_industry\_branch – Číselník odvětví v tabulce mezd.
4. czechia\_payroll\_unit – Číselník jednotek hodnot v tabulce mezd.
5. czechia\_payroll\_value\_type – Číselník typů hodnot v tabulce mezd.
6. czechia\_price – Informace o cenách vybraných potravin za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.
7. czechia\_price\_category – Číselník kategorií potravin, které se vyskytují v našem přehledu.

**Číselníky sdílených informací o ČR:**

1. czechia\_region – Číselník krajů České republiky dle normy CZ-NUTS 2.
2. czechia\_district – Číselník okresů České republiky dle normy LAU.

**Dodatečné tabulky:**

1. countries - Všemožné informace o zemích na světě, například hlavní město, měna, národní jídlo nebo průměrná výška populace.
2. economies - HDP, GINI, daňová zátěž, atd. pro daný stát a rok.

### Výzkumné otázky

1. Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?
2. Kolik je možné si koupit litrů mléka a kilogramů chleba za první a poslední srovnatelné období v dostupných datech cen a mezd?
3. Která kategorie potravin zdražuje nejpomaleji (je u ní nejnižší percentuální meziroční nárůst)? ***!!! pozor na záporná čísla !!!***
4. Existuje rok, ve kterém byl meziroční nárůst cen potravin výrazně vyšší než růst mezd (větší než 10 %)?
5. Má výška HDP vliv na změny ve mzdách a cenách potravin? Neboli, pokud HDP vzroste výrazněji v jednom roce, projeví se to na cenách potravin či mzdách ve stejném nebo násdujícím roce výraznějším růstem?

### Výstup projektu

Pomozte kolegům s daným úkolem. Výstupem by měly být dvě tabulky v databázi, ze kterých se požadovaná data dají získat. Tabulky pojmenujte t\_{jmeno}\_{prijmeni}\_project\_SQL\_primary\_final (pro data mezd a cen potravin za Českou republiku sjednocených na totožné porovnatelné období – společné roky) a t\_{jmeno}\_{prijmeni}\_project\_SQL\_secondary\_final (pro dodatečná data o dalších evropských státech).

Dále připravte sadu SQL, které z vámi připravených tabulek získají datový podklad k odpovězení na vytyčené výzkumné otázky. Pozor, otázky/hypotézy mohou vaše výstupy podporovat i vyvracet! Záleží na tom, co říkají data.

Na svém GitHub účtu vytvořte repozitář (může být soukromý), kam uložíte všechny informace k projektu – hlavně SQL skript generující výslednou tabulku, popis mezivýsledků (průvodní listinu) a informace o výstupních datech (například kde chybí hodnoty apod.).

**Neupravujte data v primárních tabulkách! Pokud bude potřeba transformovat hodnoty, dělejte tak až v tabulkách nebo pohledech, které si nově vytváříte.**

**Otázky:**

1. **Záludnost? vše potřebné najdeme pouze v uvedených tabulkách, nebo je možné, že budeme potřebovat i jiné tabulky? Dá se to zvládnout jenom pomocí toho co jsme se učili?**
2. **Jak máme předat tabulky? Databáze Engeto? Nebo v SQL skriptu bude příkaz na vytvoření tabulek?**
3. Výsledem jsou tedy 2 tabulky, 1 skriptový soubor s dotazy, 1 průvodní list?
4. Pojem meziroční co je tím myšleno, nárůst za rok od měření, nebo rozdíl mezi jednotlivými roky např. 2020 a 2021? Pracujeme s celým rokem?
5. Otázka č. 2 Jak zjistím čistou mzdu? K*olik je možné si koupit litrů mléka a kilogramů chleba za první a poslední srovnatelné období v dostupných datech cen a mezd?*
6. Je možné měření víc let a jejich vzájemné porovnání provést do jedné tabulky? Pokud se mi podaří dat vedle sebe (jako sloupce) hodnoty jednotlivých let, jak je vzájemně porovnám? Můžu data exportovat do excelu a tam porovnat?
7. Nejde mi v engeto databázi smazat tabulka?
8. Aliasy – vnořený selecty, joiny

**Kde jsou vlastně skutečně uložené, např. jako excelové tabulky?**