Projekt SQL: Data o mzdách a cenách potravin a jejich zpracování pomocí SQL

A. Popis projektu

V tomto projektu se pokusím reagovat na výzkumné otázky, které adresují **dostupnost základních potravin. Abych je mohla zodpovědět, připravím robustní datové podklady**, ve kterých bude možné vidět **porovnání dostupnosti potravin na základě průměrných příjmů za určité časové období**.

Budu vycházet z Datových sad z Portálu otevřených dat ČR.

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.

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, které v projektu budu zodpovídat:

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)?
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

Výstupem jsou **2** tabulky v databázi, ze kterých se požadovaná data dají získat.

Primární tabulka s názvem t_{jitka}_{mikulasova}_project_SQL_primary_final (pro data mezd a cen potravin za Českou republiku sjednocených na totožné porovnatelné období – společné roky) a t_{jitka}_{mikulasova}_project_SQL_secondary_final (pro dodatečná data o dalších evropských státech, jako primární přehled pro ČR).

Dále je připravena sada SQL a několik dodatečných tabulek (u otázek 4 a 5), ze kterých získám datový podklad k odpovědi na **5** výzkumných otázek.

B. Tvorba primární a sekundární tabulky

Pro zodpovězení výzkumných otázek jsem si nejprve musela projít, co se v kterých tabulkách nachází.

Při tvorbě **primární tabulky** jsem k informacím o cenách v tabulce czechia_price na základě funkce na úpravu roku spojila tabulku o mzdách (czechia_payroll), primární tabulka tak není propojena na základě ERD diagramu, ale funkcí na hodnotě YEAR. Na hodnotách mzdy a ceny jsem použila agregační funkci a zprůměrovala je tak.

Dále jsem ještě připojila LEFT JOINEM 3 dodatečné tabulky (czechia_price_category, czechia_payroll_industry_branch, czechia_payroll_value_type), jejíž informace budu v průběhu potřebovat a celou tabulku ukončila klauzulí GROUP BY na všechny sloupce, které nejsou agregované.

Při tvorbě **sekundární tabulky** jsem vycházela z požadavku připravit tabulku s HDP, GINI koeficientem a populací dalších evropských států ve stejném období, jako primární přehled pro ČR. Tabulku countries jsem propojila LEFT JOINEM na základě společného propojení country s tabulkou economies. Tabulka má obsahovat evropské země, proto jsem z tabulky countries vybrala jen Evropu, tím jsem získala data pro EU včetně České republiky. Dále jsem chtěla pracovat s HDP, proto jsem vyfiltrovala země, které HDP uvádí a ještě tabulku seřadila. V sekundární tabulce jsem nepoužila agregační funkce, budu je případně využívat v samotných otázkách.

C. Jak probíhala tvorba 5 výzkumných otázek

- **Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?**

Při tvorbě této otázky a odpovědi na ni jsem si v prvé řadě vyfiltrovala z mé primární tabulky od kterého nejnižšího, po který nejvyšší rok mzdy sledujeme a vytvořila tak poddotaz.

Dále jsem použila agregační funkci na mzdy, kdy je nutné při použití agregační funkce v selectu do klauzule GROUP BY vypsat (shluknout) všechny sloupce, které nejsou agregované. Abych pracovala jen s průměrnou hrubou mzdou na zaměstnance, vybrala jsem v klauzuli WHERE value_type_code 5958, dále tam vypsala všechna odvětví a roky mezi 2006 – 2018 a tohle všechno seřadila. Z těchto informací se dá říct že:

v odvětví *Dopravy, Ostatních činnostech*, *Zdravotní a sociální péče* a *Zpracovatelském průmyslu* v průběhu sledovaných let 2006 až 2018 **lze sledovat růst mezd**.

V odvětví *Činnosti v oblasti nemovitostí*, *Informační a komunikační činnosti*, *Kulturní, zábavní a rekreační činnosti, Peněžnictví a pojišťovnictví, Stavebnictví*, *Velkoobchod a maloobchod; opravy a údržba motorových vozidel*, *Zásobování vodou; činnosti související s odpady a sanacemi a Administrativní a podpůrné činnosti* **mzdy v průběhu sledovaných let klesaly mezi lety 2012 a 2013.**

V odvětví *Těžby a dobývání* zaznamenáváme **pokles** mezd **mezi lety 2008 a 2009** a dále **mezi 2014 a 2016**, v odvětví *Ubytování, stravování a pohostinství* vidíme **pokles** mezi **2008 a 2009** (stejně jako u odvětví *Zemědělství, lesnictví, rybářství*) **a 2010 a 2011**.

Obor Veřejná správa a obrana; povinné sociální zabezpečení zažíval pokles mezd mezi **2010 a 2011**, *Vzdělávání* mezi roky **2009 a 2010** (stejně jako u *Profesní, vědecké a technické činnosti)* a nakonec mzda pro *Výrobu a rozvod elektřiny, plynu, tepla a klimatiz. vzduchu* klesala v období mezi **2012 a 2013 a 2014 a 2015.**

- **Kolik je možné si koupit litrů mléka a kilogramů chleba za první a poslední srovnatelné období v dostupných datech cen a mezd?**

Odpověď na otázku jsem začala skládat výběrem czechia_price_category, kterou mám v mém případě pojmenovanou food_name v klauzuli WHERE. Dále bylo nutné do klauzule WHERE přidat první a poslední sledované období pro mzdy a ceny potravin a po zadání průměrné hrubé mzdy na zaměstnance, kdy jsem vybrala value_type_code 5958 se výběr opět zúžil.

Při výběru sloupečků jsem zprůměrovala hodnoty mzdy a ceny potravin a tyto jsem od sebe ještě v dalším sloupci vydělila a dostala tak cenu za jednotku v prvním a posledním sledovaném období. Z tohoto se dá usoudit, že:

v roce 2006, v prvním srovnatelném období, bylo možné nakoupit 1283 kg chleba a 1432 l mléka v dostupných datech cen a mezd. V roce 2018, posledním srovnatelném období, bylo možné v dostupných datech cen a mezd nakoupit 1340 kg chleba a 1639 l mléka.

- **Která kategorie potravin zdražuje nejpomaleji (je u ní nejnižší percentuální meziroční nárůst)?**

Nejdříve jsem spojila své 2 primární tabulky, aby se mi tak posunula data o 1 rok na operátoru ON a funkci year, dále jsem taky spojila kategorii potravin, abych dostala stejné potraviny. Meziroční nárůst jsem získala po odečtu aktuální a minulé ceny a vydělením ceny minulé, výsledek se vynásobí 100 a získá se procentuální nárůst nebo záporná míra růstu.

K odpovědi na výzkumnou otázku jsem se dostala tak, že vlastně hledáme kategorii, která má největší zápornou míru růstu.

Po seřazení dat mi vyplynulo, že nejnižší percentuální meziroční nárůst neboli nejvyšší zápornou míru růstu zaznamenala za sledované období kategorie Rajská jablka.

- **Existuje rok, ve kterém byl meziroční nárůst cen potravin výrazně vyšší než růst mezd (větší než 10 %)?**

  Odpověď na otázku jsem začala stavět tak, že jsem si spojila na funkci YEAR 2 primární tabulky, dále jsem k tomu spojila na operátoru ON jména potravin a odvětví tak, abychom je mohli stejně porovnávat, dále jsem zprůměrovala hodnoty cen a mezd v minulém a aktuálním roce, vybrala jen kategorii s průměrnou hrubou mzdou, shlukla v klauzuli GROUP BY a seřadila do tabulky **t_question_4**.

  Odtud jsem po vytvoření tabulky **t_question_4a** v dalším kroku mohla získat meziroční procentuální nárůst jak cen tak mezd.

  V posledním kroku jsem z těchto meziročních procentuálních nárůstů odečetla meziroční nárůst cen potravin od meziročního nárůstu cen mezd, dodala podmínku, že musí být větší než 10 a seřadila tento rozdíl od největšího.

  Z tohoto jsem vyvodila odpověď, že největší meziroční nárůst cen potravin v porovnání s meziročním nárůstem mezd lze pozorovat v r. 2006 u paprik.

- **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 následujícím roce výraznějším růstem?**

  Nejprve jsem si vytvořila pomocnou tabulku **t_question_five**, ze které jsem spojením primárních tabulek na sloupečku country a years (o 1 rok se data posunula) získala meziroční procentuální nárůst HDP pro Českou Republiku pro všechny sledované roky (1991 – 2020).

  Tuto tabulku jsem poté použila v klauzuli FROM a spojila ji přes years s tabulkou t_question_4 tak, abych získala společné hodnoty pro sumu HDP, mezd a cen pro ČR ve sledovaném období od r. 2006 do 2017 a získala tak novou tabulku **t_question_five_1**.

  Tuto jsem použila už pro finální výpočet součtů meziročních výpočtů HDP, cen a mezd v jednotlivých letech.

  Z tohoto by se dalo říct, že například v r. 2006, kdy byl meziroční nárůst HDP přes 5%, vzrostly ceny potravin meziročně o vice než 6% a mzdy o vice než 37%. Oproti tomu v r. 2014, kdy HDP v ČR zaznamenal meziročně růst o vice než 5%, ceny potravin meziročně klesaly, zatímco mzdy výrazně rostly.