Iveta Kolínská

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

Zadání projektu

Cílem projektu je odpovědět na 5 definovaných výzkumných otázek, které se zaměřují na dostupnost základních potravin široké veřejnosti. Výsledky budou prezentovány tiskovým oddělením na následující konferenci zaměřené na tuto oblast.

K zodpovězení otázek je potřeba připravit robustní datové podklady, ve kterých uvidíme porovnání dostupnosti potravin na základě průměrných příjmů za určité časové období.

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

K dispozici máme tyto tabulky:

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.

Tvorba tabulek

Než jsem začala vytvářet obě tabulky, ze kterých jsem následně čerpala data k zodpovězení otázek, nejprve jsem si důkladně pročetla všechny dostupné datové sady, abych lépe pochopila data v nich obsažená. Zjistila jsem také, že společným časovým obdobím pro všechny primární tabulky jsou roky 2006 – 2018, které pro mě budou zkoumaným časovým obdobím.

První tabulku (t_iveta_kolinska_project_sql_primary_final) jsem se nejprve pokusila vytvořit tak, že jsem na sebe napojila všechny primární tabulky týkající se cen, mezd a regionů v ČR. Výsledná tabulka ale měla dohromady kolem 34 milionů řádků a nebylo téměř možné spustit jakýkoli SQL dotaz. Proto jsem se snažila vymyslet, jak co nejvíce omezit počet řádků. Vyřešila jsem to tak, že jsem zprůměrovala ceny potravin v daném roce pro každý region a sloupec průměrné mzdy / počet zaměstnanců za daný rok. Dále jsem nevybrala všechny sloupce, ale pouze ty, které jsem potřebovala k zodpovězení otázek. A nakonec jsem z tabulky odstranila NULL hodnoty regionů, protože z dat nebylo jasné, ke kterému regionu se vztahují. To samé jsem udělala se sloupci týkajícími se průměrných mezd / počtu zaměstnanců a odvětví, nebyla jsem totiž schopna zjistit význam těchto hodnot pro zodpovězení otázek.

Tvorba druhé tabulky (t_Iveta_Kolinska_project_SQL_secondary_final) byla pro mne mnohem jednodušší. Spojila jsem pouze dvě tabulky (countries a economies), vybrala data za mnou stanovené zkoumané časové období a jako kontinent zvolila Evropu.

Odpovědi na výzkumné otázky

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

Odvětví, ve kterých v průběhu let mzdy klesají, jsou ty, u kterých je growth_index roven 0. Odpověď tedy je, že mzdy vzrostly pouze u těchto odvětví: Doprava a skladování, Ostatní činnosti, Peněžnictví a pojišťovnictví, Zpracovatelský průmysl a Zdravotní a sociální péče.

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

V prvním srovnatelném období, tedy v roce 2006, bylo možné koupit si 1287 kilogramů chleba a 1437 litrů mléka.
V posledním srovnatelném období, v roce 2018, jsme si mohli koupit 1342 kilogramů chleba a 1642 litrů mléka.
Ceny obou položek sice v průběhu let vzrostly, nárůst mezd byl ale rychlejší, tím pádem jsme si v roce 2018 mohli koupit více kilogramů chleba i litrů mléka.

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

Nejpomaleji zdražoval cukr krystalový, meziroční percentuální nárůst byl -1,92%, cena cukru tedy poklesla.

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

Takový rok v rámci zkoumaného období neexistuje.

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 následujícím roce výraznějším růstem?**

Abych mohla odpovědět na tuto otázku, zobrazila jsem si pomocí SQL dotazu do sloupečků vedle sebe procentuální růst mezd, cen potravin a HDP. K lepšímu vyhodnocení výsledků jsem použila následující graf, do kterého jsem přenesla výsledné hodnoty SQL dotazu.

Podle grafu by se dalo říci, že výška HDP má vliv na změny v růstu cen potravin i mezd. Změny ve mzdách se ale projevily s ročním zpožděním a změny v cenách potravin kopírovaly změny HDP s dvouletým zpožděním. Výjimkou byla celosvětová bankovní krize v roce 2009, kdy se naráz citelně zpomalil růst cen potravin, mezd i HDP.