SQL Projekt

1) Zadanie projekt

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:

czechia_payroll – Informace o mzdách v různých odvětvích za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.

czechia_payroll_calculation – Číselník kalkulací v tabulce mezd.

czechia_payroll_industry_branch – Číselník odvětví v tabulce mezd.

czechia_payroll_unit – Číselník jednotek hodnot v tabulce mezd.

czechia_payroll_value_type – Číselník typů hodnot v tabulce mezd.

czechia_price – Informace o cenách vybraných potravin za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.

czechia_price_category – Číselník kategorií potravin, které se vyskytují v našem přehledu.

Číselníky sdílených informací o ČR:

czechia_region – Číselník krajů České republiky dle normy CZ-NUTS 2.

czechia_district – Číselník okresů České republiky dle normy LAU.

Dodatečné tabulky:

countries - Všemožné informace o zemích na světě, například hlavní město, měna, národní jídlo nebo průměrná výška populace.

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)?
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 následujícím 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.).

2) Výstup

Prvým krokom k vytvoreniu projektu bolo zoznámenie sa s dátami pre vytvorenie primárnej a sekundárnej tabuľky (*t_Martin_Kmet_project_SQL_primary_final* a *t_Martin_Kmet_project_SQL_secondary_final*). Po vyhodnotení, ktoré informácie sú dôležité som sa tieto tabuľky vytvoril. Navyše som si vyhľadal niekoľko informácií, ktoré by mi mohli pomôcť. Potom som pokračoval vypracovaním výskumných otázok.

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

Pomocou príkazu som vyfiltroval a označil počet odvetví, ktoré rastú, klesajú alebo sú stabilné.

Niekoľko odvetví vykazuje rast (19) ale nie sú to všetky.

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

Pre odpoveď na otázku sme museli zistiť dodatočnú informáciu a to, aký „food_code“ má chlieb a mlieko. Po zistení tejto informácie som si vytvoril príkaz na vyfiltrovanie podstatných dát.

Prvé zrovnateľné obdobie je rok 2006, v ktorom bolo možné si zakúpiť 1.192, kg chleba a 1.331, l mlieka. V poslednom období (rok 2018) sme si mohli kúpiť o niečo viac a teda 1.300, kg chleba a 1.590, l mlieka.

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

Vypracoval som príkaz, kde sa bude odzrkadlovať priemerný medziročný nárast všetkých potravín a následne ich porovnal.

Kryštálový cukor nie len vykazoval najnižší rast ale dokonca sa jednalo o pokles ceny a konkrétne priemerne o 1,92%.

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

Porovnával som rast potravín s rastom miezd a ich vzájomnom vzťahu.

Takýto nárast naozaj nenastal. Ceny potravín rástli viac ako ceny miezd len tri krát ale nikto to nebolo viac než o 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?

Filtroval som si medziročné zmeny HDP, miezd a cien potravín a následne porovnával percentuálne zmeny.

Tieto údaje sú síce blízke ale zasahuje viacero premenných. Na základe našich dát som vplyv zmien HDP na mzdy a ceny potravín nenašiel. Všetky jednotky sa menili bez viditeľného vzťahu.