ZADÁNÍ ZÁVĚREČNÉHO SQL PROJEKTU – DATOVÁ AKADEMIE

10.12.2024 – 11.3.2025

Projekt : Projekt z SQL

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 požít pro získání vhodnéhp 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)?

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?

Výstupy z 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.

VÝSTUP Z PROJEKTU

Jako pracovník na analytickém oddělení nezávislé společnosti, která se zabývá životní úrovní občanů, jsem dostala za úkol pokusit se odpovědět na pár definovaných výzkumných otázek ohledně **dostupnosti základních potravin široké veřejnosti.**

1) Tvorba primární tabulky

Prvním krokem, který jsem udělala, byla tvorba projektové tabulky pod názvem **t_alena_steinerova_project_SQL_primary_final**, kde jsem čerpala z dalších tabulek a číselníků, které jsem měla k dispozici. Do této tabulky jsem zapracovala údaje o průběhu a vývoji cen a mezd za určité časové období. V této tabulce můžete najít sloupce rok_ceny, id_kategorie, nazev_kategorie, prumerna_cena, merna_jednotka, rok_mzdy, nazev_odvetvi a prumerna_mzda.

Pro tuto tabulku jsem použila následující dostupné primární tabulky a číselníky:

- 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.

Svou projektovou tabulku jsem tvořila pomocí dvou CTE:

- V prvním CTE (ceny_cte) jsem zpracovala data ohledně průměrných cen a let a dalších potřebných sloupců z primární tabulky czechia_price. Zároveň jsem připojila přes funkci JOIN čícelník czechia_price_category, abych získala ID a název kategorie.
- V druhém CTE (mzdy_cte) jsem zpracovala data pro mzdy rozdělené podle let a dalších potřebných sloupců z tabulky czechia_payroll. Zde jsem opět připojila přes JOIN číselník czechia_payroll_industry_branch, pro získání názvu odvětví. Zároveň jsem si přes WHERE vyfiltrovala následující údaje:
  - value_typ_code na 5958 (z číselníku czechia_payroll_value_type, kdy označení 5958 značí Průměrnou hrubou mzdu),
  - unit_code na 200 (z číselníku czechia_payroll_unit , kdy označení 200 by mělo značit Kč – dle našeho SQL lektora),
  - calculation_code na 100 (z číselníku czechia_payroll_calculation, kdy označení 100 značí fyzický.

Tato dvě CTE jsem v záveřečném SELECT spojila přes LEFT JOIN díky společným rokům. Díky tomu mi vyšla celková tabulka jako podklad pro zodpovězení výzkumných otázek.

2) **Odpovědi na výzkumné otázky**

Výzkumná otázka č. 1: Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?

Odpověď: Mzdy v průběhu let 2007 až 2018 v některých odvětvích rostou, v některých klesají. Klesající mzdy byly zjištěny u následujících odvětví a let:

2009 Těžba a dobývání
2009 Ubytování, stravování a pohostinství
2009 Velkoobchod a maloobchod; opravy a údržba motorových vozidel
2009 Zemědělství, lesnictví, rybářství
2010 Veřejná správa a obrana; povinné sociální zabezpečení
2010 Profesní, vědecké a technické činnosti
2010 Vzdělávání
2011 Veřejná správa a obrana; povinné sociální zabezpečení
2011 Kulturní, zábavní a rekreační činnosti
2011 Ubytování, stravování a pohostinství
2013 Výroba a rozvod elektřiny, plynu, tepla a klimatiz. vzduchu
2013 Činnosti v oblasti nemovitostí
2013 Peněžnictví a pojišťovnictví
2013 Profesní, vědecké a technické činnosti
2013 Těžba a dobývání
2013 Zásobování vodou; činnosti související s odpady a sanacemi
2013 Informační a komunikační činnosti
2013 Kulturní, zábavní a rekreační činnosti
2013 Stavebnictví
2013 Velkoobchod a maloobchod; opravy a údržba motorových vozidel
2014 Těžba a dobývání
2015 Výroba a rozvod elektřiny, plynu, tepla a klimatiz. vzduchu
2016 Těžba a dobývání

Postup pro zajištění odpovědi:
Abych mohla zodpovědět na tuto otázku, bylo potřeba čerpat z tabulky t_alena_steinerova_project_SQL_primary_final. Použila jsem CTE rust_mezd_cte – zde jsem si přes SELECT zadala sloupce rok_mzdy, nazev_odvetvi a prumerna_mzda. Dále jsem díky funkci LAG a dalších výpočtů zjistila mzdy za předchozí rok, rozdíl mezd a nárůst procenta mezd (tedy o kolik procent se mzdy meziročně zvyšovaly).

Dále díky funkci CASE EXPRESSION jsem mohla určit, zda mzdy v průběhu let rostly, stagnovaly nebo klesaly.

V závěrečném SELECTU jsem si pak nechala zobrazit potřebné sloupce, zaokrouhlila potřebné hodnoty a zároveň si přes funkci WHERE nechala smazat hodnoty, kde mzda za předchozí rok byla NULL (to bylo u všech odvětví za rok 2006, protože rok 2005 již znám nebyl).

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

Odpověď:

Za první srovnatelné období v roce 2006 je možné si zakoupit 1408.96 litrů mléka a 1261.65 kg chleba.

Za poslední srovnatelné období 2018 je možné si zakoupit 1613.73 litrů mléka a 1319.40 kg chleba.

Postup pro zajištění odpovědi:
Abych mohla zodpovědět na tuto otázku, bylo potřeba čerpat z tabulky t_alena_steinerova_project_SQL_primary_final. Použila jsem následující CTE:

- prvni_obdobi_cte- zde jsem si do SELECT kromě rok_ceny, nazev_kategorie a merna_jednotka dala dále zaokrouhlednou a zprůměrovanou průměrnou cenu a průměrnou mzdu. Dále jsem pro zjištění odpovědi, kolik je možné koupit si potraviny za mzdu udělala výpočet: Průměrná mzda děleno průměrná cena. Dále jsem si klauzulí WHERE vyfiltrovala z názvu kategorie 'Chléb konzumní kmínový', 'Mléko polotučné pasterované'. Aby se mi zobrazilo první srovnatelné období, musela jsem zadat funkci LIMIT pro dva záznamy.
- poslední_obdobi_cte- zde jsem udělala to samé, jako v první_obdobi_cte, jen s rozdílem u klauzule ORDER BY, kde jsem dala ‚descending', aby se mi zobrazilo poslední srovnatelné období.
- Tato dvě CTE jsem nakonec spojila přes UNION ALL.

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

Odpověď: Nejpomaleji (nejnižší percentuální meziroční nárůst) má kategorie potravin 117,101, což jsou Rajská jablka červená kulatá. Meziroční nárůst má -30,28 % v roce 2007.

Postup pro zajištění odpovědi:
Abych mohla zodpovědět na tuto otázku, bylo potřeba čerpat z tabulky t_alena_steinerova_project_SQL_primary_final. Použila jsem opět CTE, tentokrát ceny_potravin_cte, kde jsem si do SELECT zadala potřebné sloupce, a přes funkci LAG spočítala ceny za předchozí rok, dále rozdíl cen a meziroční nárůst cen v procentech. V závěrečném SELECT jsem pak potřebné hodnoty zaokrouhlila na dvě desetinná místa. Aby se mi správně zobrazovaly hodnoty, tak jsem v klauzuli WHERE odstranila ve sloupci cena_predchozi_rok NULL hodnoty. Aby se mi zobrazilo nejpomalejší zdražování, tak jsem v ORDER BY klauzuli seřadila hodnoty podle sloupce narust_procenta_cen a podle klauzule LIMIT dala 1.

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

Ano, existují roky, ve kterých byl meziroční nárůst cen potravin výrazně vyšší než růst mezd (větší než 10 %). Konkrétně v letech 2007, 2008, 2011 a 2012.

Postup pro zajištění odpovědi:
Abych mohla zodpovědět na tuto otázku, bylo potřeba čerpat z tabulky t_alena_steinerova_project_SQL_primary_final. Použila jsem následující CTE:

- mezirocni_narust_cen_cte – zde jsem si v SELECT zadala potřebné sloupce pro ceny, a zároveň pomocí funkce LAG spočítala ceny za předchozí roky, rozdíl cen a meziroční nárůst cen v procentech.
- mezirocni_narust_mezd_cte- zde jsem si v SELECT zadala potřebné sloupce pro mzdy, a zároveň pomocí funkce LAG spočítala mzdy za předchozí roky, rozdíl mezd a meziroční nárůst mezd v procentech.

V závěrečném SELECT jsem si zobrazila potřebné sloupce a zprůměrovala a zaokrouhlila potřebná data. Spojila jsem obě CTE přes LEFT JOIN přes sloupce rok_ceny a rok_mzdy. Zároveň jsem si přes funci CASE EXPRESSION nechala určit, o kolik procent byl meziroční nárůst cen potravin vyšší než růst mezd. Ošetřila jsem si vše přes WHERE pro NULL hodnoty u cena_predchozi_rok a mzda_predchozi_rok (NULL hodnoty byly v roce 2006, protože rok 2005 již znám nebyl).

3) Tvorba Sekundární tabulky – pro odpověď na výzkumnou otázku č. 5

Prvním krokem, který jsem udělala, byla tvorba projektové tabulky pod názvem **t_alena_steinerova_project_SQL_secondary_final**, kde jsem čerpala z dalších tabulek a číselníků, které jsem měla k dispozici. Do této tabulky jsem zapracovala údaje o průběhu a vývoji cen a mezd a hdp za určité časové období. V této tabulce můžete najít sloupce rok_ceny, nazev_kategorie, id_kategorie, prumerna_cena, merna_jednotka, rok_mzdy, nazev_odvetvi, prumerna_mzda, rok_hdp, zeme, hdp, gini a populace.

Pro tuto tabulku jsem použila následující dostupné primární tabulky a číselníky:

- 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.

- 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.

Svou projektovou tabulku jsem tvořila pomocí tří CTE:

- V prvním CTE (ceny_cte) jsem zpracovala data ohledně průměrných cen a let a dalších potřebných sloupců z primární tabulky czechia_price. Zároveň jsem připojila přes funkci JOIN číselník czechia_price_category, abych získala ID a název kategorie.
- V druhém CTE (mzdy_cte) jsem zpracovala data pro mzdy rozdělené podle let a dalších potřebných sloupců z tabulky czechia_payroll. Zde jsem opět připojila přes JOIN číselník czechia_payroll_industry_branch, pro získání názvu odvětví. Zároveň jsem si přes WHERE vyfiltrovala následující údaje:
  - value_typ_code na 5958 (z číselníku czechia_payroll_value_type, kdy označení 5958 značí Průměrnou hrubou mzdu),
  - unit_code na 200 (z číselníku czechia_payroll_unit , kdy označení 200 by mělo značit Kč – dle našeho SQL lektora),
  - calculation_code na 100 (z číselníku czechia_payroll_calculation, kdy označení 100 značí fyzický.
- Ve třetím CTE (hdp_cte) jsem zpracovala data pro evropské státy (klauzule WHERE + continent + Europe), kde sleduji zemi, rok, hdp, gini a populaci z tabulky economies. K tomu jsem přes LEFT JOIN přidala tabulku countries se společnými sloupci country a country.

Tato tři CTE jsem v záveřečném SELECT spojila přes LEFT JOIN díky společným rokům. Díky tomu mi vyšla celková tabulka jako podklad pro zodpovězení poslední výzkumné otázky.

Výzkumná otázka č. 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?

U všech sledovaných let nelze jednoznačně stanovit, zda růst HDP má výrazný vliv na změny ve mzdách a cenách potravin. V některých letech rostly HDP, ceny i mzdy, v některých letech některé ukazatele rostly či naopak klesaly.

Například v roce 2007 je nárůst procenta HDP o 5,57 %, nárůst procenta cen o 9,26 % a nárůst procenta mezd o 6,29 %.

Dále je zde rok kdy sledované ukazatelé poklesly, tedy rok 2009, kdy HDP pokleslo o 4,66 %, mzdy o 6,59 % a mzdy jen mírně rostly o 3,11 %.

Dále naopak v roce 2013 pokleslo HDP o -0,05 %, ceny vzrostly o 6,01 % a naopak mzdy poklesly o 0,78 %.

Postup pro zajištění odpovědi:
Abych mohla zodpovědět na tuto otázku, bylo potřeba čerpat z tabulky t_alena_steinerova_project_SQL_secondary_final. Použila jsem následující CTE:

- ceny_cte – zde jsem si v SELECT zadala potřebné sloupce pro ceny, a zároveň pomocí funkce LAG spočítala ceny za předchozí roky, rozdíl cen a meziroční nárůst cen v procentech.
- mzdy_cte- zde jsem si v SELECT zadala potřebné sloupce pro mzdy, a zároveň pomocí funkce LAG spočítala mzdy za předchozí roky, rozdíl mezd a meziroční nárůst mezd v procentech.
- hdp_cte- zde jsem si v SELECT zadala potřebné sloupce pro hdp, a zároveň pomocí funkce LAG spočítala hdp za předchozí roky, rozdíl hdp a meziroční nárůst hdp v procentech.

V závěrečném SELECT jsem si zobrazila potřebné sloupce a zprůměrovala a zaokrouhlila potřebná data. Spojila jsem všechny CTE přes LEFT JOIN přes společné sloupce let. Přes klazuli CASE EXPRESSION jsem si vytvořila další soupec rust_hdp_v_case, abych viděla, zda HDP rostlo nebo klesalo oproti předchozímu roku. Dále jsem si ošetřila přes WHERE hodnoty NULL u narust_procenta_cen, narust_procenta_mezd, narust_procenta_hdp (NULL hodnoty byly v roce 2006, protože rok 2005 již znám nebyl). Zároveň jsem přes WHERE vyfiltrovala zemi jen Českou republiku.