Průvodní listina projektu pro procvičení SQL pro projekt Engeto

Zadání projektu

…

Spojení tabulek

Abych si práci pro sebe udělala přehlednější, mám plán vytvořit nejdřív 2 TABULKY – jednu pro ceny, druhou pro platy, a ty následně spojit do výsledné tabulky

TABULKA pro platy: využiji tyhle sloupce z czechia_payrol:

- Value
- Value_type_code – použiji pouze hodnoty 5958 = Průměrná hrubá mzda na zaměstnance. Druhá hodnota – 316 = průměrný počet zaměstnaných osob – by pro naše zkoumání neměla být k užitku. V nové tabulce tato hodnota není potřeba, použiju ji pouze pro filtrování vstupních dat podmíknou WHERE.
- Unit_code = vždy by měl být 200 = hodnota Kč, můžu použít pro kontrolu, zda mám všechna data správná. Ale tento sloupec by mohl být z nové tabulky vynechán. V nové tabulce tato hodnota není potřeba, použiju ji pouze pro filtrování vstupních dat podmíknou WHERE.
- Calculation_code = budu pracovat s hodnotami 200 = přepočtená hrubá mzda. Považuji to za více vypovídající údaj. V nové tabulce tato hodnota není potřeba, použiju ji pouze pro filtrování vstupních dat podmíknou WHERE.
- Industry_branch_code
- Payroll_year
- Payroll_quarter – tento údaj ve sloučené tabulce vynechám a budu zkoumat pouze celé roky

czechia_payroll_calculation: dává nám informaci o tom, jaké hodnoty pro novou tabulku využít, ale do nové tabulky ji zahrnovat nebudu

czechia_payroll_industry_branch: bude připojena k nové tabulce, přes sloupec code ji připojím k tabulce czechia_payroll

czechia_payroll_unit: dává nám informaci o tom, jaké hodnoty pro novou tabulku využít, ale do nové tabulky ji zahrnovat nebudu

czechia_payroll_value_type: dává nám informaci o tom, které hodnoty pro novou tabulku využít, ale do nové tabulky ji zahrnovat nebudu

TABULKA pro ceny: využiji tyhle sloupce z czechia_price:

- Id – je potřeba? ??
- Value
- Category code
- Date_from – vytvořím si nový sloupec, očištěný o zbytek data, abych dostala formát YYYY-MM-DD
- Region_code – vynechám z nové tabulky, v otázkách nikde není dotaz na dělení podle okresů/krajů

czechia_price_category: tabulka nutná pro další výzkum - přes category code ji připojím k tabulce czech_price, abych měla vedle sebe hodnoty pro ceny potravin včetně slovního popisku

Do jedné tabulky je třeba přidat údaje o HDP v České republice v daném roce, přidala jsem do tabulky pro ceny potravin.

Následně jsem spojila obě tabulky dohromady a vznikla tabulka, ze které věřím, že bude možné informace získat.

TABULKA pro ceny + informace o HDP: t_iva_dvorakova_price_table

TABULKA pro platy: t_iva_dvorakova_payroll_table

Výsledná tabulka: t_iva_dvorakova_project_SQL_primary_final

Původně jsem, inspirována při konzultaci o projektu, chtěla vytvořit 2 VIEW a ty spojit do tabulky.

Napoprvé se mi to podařilo, ale zjistila jsem, že nikde nemám údaje o HDP.

Připojila jsem tento údaj to VIEW o cenách a následně stejně spojila, ale tento příkaz mi ani po 10 minutách žádanou tabulku nevygeneroval. Už jsem nepřišla na nic, co by šlo osekat, agregovat, nějak zmenšit. Takže jsem namísto VIEW vytvořila 2 tabulky a spojila do výsledné jedné tabulky. V řádu minut už tato tabulka šla vytvořit.

Z toho důvodu mám 3 tabulky, 2 pomocné, a jednu výslednou.

Výzkumná otázka č. 1, soubor sql_project_question_1

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

Pro zodpovězení této otázky potřebuji znát průměrnou mzdu v daném odvětví pro jednotlivé roky, které zkoumáme. Pro tento výpočet jsem si vytvořila pomocnou tabulku t_avg_payroll, kde mám vypočítané hodnoty průměrných mezd v jednotlivých obdobích v jednotlivých letech pro všechna odvětví.

Pro výpočet meziročních změn jsem připravila JOIN dvou stejných těchto tabulek a s posunem o 1 rok jsem vypočítala rozdíly mezi jednotlivými lety – vždy dva po sobě následující roky.

Výzkumná otázka č. 2, soubor sql_project_question_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 kroku jsem si určila funkcí min a max první a poslední rok, který sledujeme.

Dále jsem si přes SELECT DISTINCT určila přesné názvy pro mléko a chléb.

Připravila jsem dvě varianty SQL:

- První, kde počítám průměrnou mzdu v odvětví v daném roce (v letech 2006 + 2008), přidávám údaj o průměrné ceně hledaných dvou komodit a následně prostým dělením docházím k výsledku, kolik celých jednotek dané komodity si za průměrnou mzdu v daném odvětví ve zkoumaném roce koupíme
- Druhá varianta vynechává rozdělení na odvětví a dává nám jen odpověď na to, kolik bochníků chleba a litrů mléka si pořídíme v daných letech za průměrnou mzdu bez rozdělení na odvětví.

Výzkumná otázka č. 3, soubor sql_project_question_3

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