Průvodní listina projektu SQL

Zadání projektu

Projekt je zaměřený na zmapování dostupnosti základních potravin široké veřejnosti. K dispozici máme tyto informace:

- Informace o mzdách v různých odvětvích za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR. K této datové sadě máme další pomocné tabulky vysvětlující data v tabulce mezd.
- Informace o cenách vybraných potravin za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.
- Číselníky krajů a okresů České republiky dle norem CZ-NUTS 2 a LAU.
- Informace o dalších zemích světa - populaci, HDP, měně, daňové zátěži, aj.

Jsou definovány následující 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 násdujícím roce výraznějším růstem?

Spojení tabulek

Pro zodpovězení otázek vytvoříme dvě tabulky:

Primární tabulka: t_iva_dvorakova_project_SQL_primary_final

Sekundární tabulka: t_iva_dvorakova_project_SQL_secondary_final

t_iva_dvorakova_project_SQL_primary_final

Nejprve jsem se rozhodla vytvořit dvě tabulky – jednu pro platy, druhou pro ceny. Ty následně spojit do jedné finální tabulky, obsahující údaje zároveň o platech i cenách.

Tabulka pro platy: využiji tyto sloupce z tabulky 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:

- 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.

V obou tabulkách jsem jednotlivé údaje o mzdách a cenách spojila do jedné informace o průměrné ceně/mzdě – v daném roce, odvětví, artiklu spotřebního koše.

Pokud jsem tento údaj do prví tabulky nedopočítala, nebyla jsem schopná řešit následující výzkumné otázky bez dalších pomocných tabulek a výpočtů.

Podle zadání by jednotlivé hodnoty platů a cen neměly být nikde dále potřeba, proto budu už vycházet z průměrných hodnot.

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 do 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.

t_iva_dvorakova_project_SQL_secondary_final

V druhé tabulce jde o informace pro ostatní státy Evropy ve srovnatelném období.

Spojila jsem tabulky countries a economies a zahrnula tyto sloupce:

- Název země „country“
- Populace dané země „population“
- Rok „YEAR“
- HDP „GDP“
- Gini koeficient „gini“

a výběr ohraničila podmínkou WHERE na země v Evropě + data z let 2006 – 2018.

Výzkumná otázka č. 1, soubor sql_project_question_1_2023_10_31.sql

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. V základní tabulce už mám průměrné hodnoty pro platy v daných letech.

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

Pro snazší orientaci jsem přidala sloupec increase_decrease_flag, který nám říká, zda mzdy mezi lety rostly, nebo klesaly.

Výzkumná otázka č. 2, soubor sql_project_question_2_2023_10_31.sql

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_2023_10_31.sql

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

Pro tuto otázku jsem nejprve spojila dvě primární tabulky k sobě, s posunem o 1 rok, a spočítala průměrnou změnu cen mezi danými dvěma lety – údaje pro všechny roky, všechny položky.

V dalším kroku jsem spočítala průměr z těchto ročních průměrů a seřadila výsledky vzestupně.

Z toho mi vychází, že nejnižší meziroční procentuální růst je u položky „Cukr krystalový“, kde se cena dokonce snížila a proto průměrná procentuální roční změna má zápornou hodnotu.

Výzkumná otázka č. 4, soubor sql_project_question_4_2023_10_31.sql

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

Postupně jsem přidávala informace, které potřebuji porovnávat.

Prvním krokem byly průměrná mzda v jednotlivých letech.

Dalším krokem byl JOIN se stejnou tabulkou posunutou o rok, abychom viděli rozdíl mezi lety vyjádřený absolutní hodnotou i procentuální rozdíl.

Pro otázku ale není důležité, o jaké potraviny nebo odvětví průmyslu se jedná, proto třetím krokem bylo srovnání celkové průměrné mzdy za všechna odvětví a jejich meziroční změny. To stejné i pro ceny potravin.

Krok 4 následoval – spojení údajů o změnách mezd + cen + procentuální vyjádření.

V posledním 5.kroku jsem přidala sloupec počítající rozdíl mezi % změnu cen a mezd (diff). Dále také sloupec high_increase_flag, který detekuje rozdíl větší než 10%. Vyšel v jednom roce.

Pro zajímavost jsem přidala SQL, kde výsledkem jsem tyto změny, ale pro jednotlivá odvětví průmyslu. To může vypovídat o tom, v kterém odvětví

Výzkumná otázka č. 5, soubor sql_project_question_5_2023_10_31.sql

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?

V tomto příkladě je potřeba spojit více sloupců a výpočtů dohromady.

První nápad mi nevyšel, zkusila jsem na sebe napojit 3 * primární tabulku, ale asi po 15 minutách výsledek nebyl.

Proto jsem se rozhodla vytvořit 2 pomocné tabulky, které následně spojím do jedné a dají nám všechny informace, které potřebujeme.

V tom případě musí v jednom řádku být všechny tyto údaje (musí mít vypočítané tyto sloupce):

- Roční změna HDP
- Roční změna cen v témže roce
- Roční změna mezd v témže roce
- Roční změna cen v následujícím roce
- Roční změna mezd v následujícím roce

+ dopočítané, kde je změna HDP vůči cenám/platům výrazná

Postupovala jsem s pomocnými tabulkami.

Nejdříve t_GDP_change, která zobrazuje změnu HDP mezi jednotlivými lety.

Druhá t_PP_change_3_years, která zobrazuje změny cenách a mzdách mezi 3 lety – rozdíl mezi lety 1 a 2 a mezi lety 2 a 3. Tu jsem připravila ve dvou krocích, opět mi nešlo napoprvé sloučit dvě tabulky, takže vznikla jedna další pomocná tabulka t_PP_change_2_years.

Spojením t_GDP_change + t_PP_change_3_years mi vzniknul přehled toho, jak se mezi dvěma lety změnilo HDP, jak se ve stejném období změnily ceny a platy a zároveň, jak tomu bylo v roce následujícím.

Pro větší přehlednost porovnám HDP a růsty nadvakrát – nejdříve pro stejné období, potom pro následující období. Přidám tedy sloupce

- overall_situation_2x3 – hodnotí změnu cen a mezd ve vztahu ke změně HDP – ve stejném období, jako se měnilo HDP. Jako podmínku pro „overall growth“ zadám růst všech 3 parametrů o 2 a více %. Pokud by byly všechny hodnoty záporné, označíme řádek „overall decrease“.
- overall_situation_1x3 – hodnotí změnu cen a mezd ve vztahu ke změně HDP – např. změn HDP mezi lety 2006/2007 vůči změně cen a mezd v letech 2007/2008. Podmínka pro vyhodnocení je stejná - pro „overall growth“ zadám růst všech 3 parametrů o 2 % a více. Pokud by byly všechny hodnoty záporné, označíme řádek „overall decrease“.

Nakonec jsem dotaz spojila do jednoho SQL dotazu. Snažila jsem se minimalizovat názvy sloupců, aby data byla přehlednější.

Výsledky všech otázek jsou v přeložených souborech.