Zadání projektu

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:

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?

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.).

Vypracování projektu

Vytvoření tabulek:

Pro vypracování tohoto projektu a vyřešení výzkumných otázek jsem pracovala s databází na localhostu. Pouze na něm jsem prováděla veškeré úpravy databází.

Nejprve jsem si připravila dvě tabulky jako datový podklad pro zodpovězení pěti zadaných výzkumných otázek.

1. Tabulka - **t_Marie_Vrbova_project_SQL_primary_final** (pro data mezd a ceny potravin za Českou republiku sjednocených na totožné porovnatelné období – společné roky):
   Vytvořila jsem si dvě dočasné tabulky – jedna dočasná tabulka pojmenovaná **t_Marie_Vrbova_project_SQL_salary** se týká mezd. Vznikla spojením níže vyjmenovaných tabulek přes JOIN:

   - czechia_payroll_calculation,
   - czechia_payroll_industry_branch,
   - czechia_payroll_unit,
   - czechia_payroll_value_type.

   Druhá dočasná tabulka pojmenovaná **t_Marie_Vrbova_project_SQL_price** vznikla spojením následných tabulek opět přes JOIN:

   - czechia_price,
   - czechia_price_category.

   Finální **primary** tabulka vznikla propojením dvou výše zmíněných dočasných tabulek přes INNER JOIN, spojovacím prvkem je rok, proto vymezené období je 2006-2018. Vytvořená tabulka obsahuje tyto sloupečky:
   - payroll_year,
   - salary_average,
   - industry_branch_name,
   - code_industry_branch,
   - product_year,
   - product_name,
   - price_average,
   - price_value,
   - price_unit.

   https://github.com/mvrbova/SQL_Project/blob/main/Project_1_01_Creation_of_primary_table.sql

2. Tabulka t_Marie_Vrbova_project_SQL_secondary_final (pro dodatečná data o dalších evropských státech):

   Finální secondary tabulka vznikla propojením tabulek „countries“ a „economies“ přes INNER JOIN. Spojovacím prvkem je země – country. Vytvořená tabulka obsahuje níže vyjmenované sloupečky:
   - c_country,
   - year,
   - continent,
   - currency_code,
   - GDP,
   - GINI.

   https://github.com/mvrbova/SQL_Project/blob/main/Project_1_02_Creation_of_secondary_table.sql

Vytvoření skriptů pro zodpovězení výzkumných otázek

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

   Pomocí WITH clause jsem vytvořila dočasnou tabulku obsahující pouze data týkající se platů v jednotlivých odvětvích a přes funkci LAG jsem zjistila průměrný nárůst/pokles platů meziročně, včetně procentuálního nárůstu/poklesu.
   - V odvětvích - *Doprava a skladování, Ostatní činnosti, Zdravotní a sociální péče, Zpracovatelský průmysl* – ve sledovaném období 2006-2018 došlo vždy k meziročnímu nárůstu platů.
   - Ve zbývajících odvětvích lze zaznamenat alespoň jeden meziroční pokles platu.

   https://github.com/mvrbova/SQL_Project/blob/main/Project_1_01_1.Question_salary_growth.sql

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

   Pomocí CASE EXPRESSION jsem si stanovila podmínku, aby pro produkty *„Mléko polotučné pasterované“* a *„Chléb konzumní kmínový“* byl průměrný plat pro jednotlivá odvětví vydělen cenou daného produktu. Tak jsem získala množství mléka v litrech a chleba v kilogramech, kolik bylo možné si jich koupit za roční mzdu v letech 2006 a 2018.

   - Mléko – rok 2006: nejméně litrů mléka, tj. 789 l, si mohli koupit zaměstnanci v odvětví *Ubytování, stravování a pohostinství*, a nejvíce, tj. 2749, *v Peněžnictví a pojišťovnictví.*
   - Mléko – rok 2018: nejméně litrů mléka, tj. 947 l, si mohli koupit zaměstnanci v odvětví *Ubytování, stravování a pohostinství*, a nejvíce, tj. 2831, v *Informačních a komunikačních činnostech.*

- Chléb – rok 2006: nejméně kilogramů chleba, tj. 707 kg, si mohli koupit zaměstnanci v odvětví *Ubytování, stravování a pohostinství*, a nejvíce, tj. 2462, *v Peněžnictví a pojišťovnictví.*
- Chléb – rok 2018: nejméně kilogramů chleba, tj. 774 kg, si mohli koupit zaměstnanci v odvětví *Ubytování, stravování a pohostinství*, a nejvíce, tj. 2314, v *Informačních a komunikačních činnostech.*

https://github.com/mvrbova/SQL_Project/blob/main/Project_1_02_2.Question_milch_bread.sql

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

   Vytvořila jsem si pohled pojmenovaný v_price_difference obsahující data týkající se potravin, jejich cen a výpočet meziročního nárůstu cen, včetně procentuálního přes funkci LAG – sloupce: product_year, product_name, price_average, price_average_growth, price_average_perentage_growth. Pomocí funkce AVG ze sloupce price_average_perentage_growth jsem spočítala průměrný nárůst cen pro jednotlivé kategorie ve vymezeném období.
   - Nejpomaleji zdražoval „Cukr krystalový“.

   https://github.com/mvrbova/SQL_Project/blob/main/Project_1_03_3.Question_prices_growth.sql

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

   Vytvořila jsem si dvě tabulky – jednu s daty týkající se cen, druhou s daty týkající se platů s pomocí With clause a tabulky jsem následně propojila přes rok. Pak lze k získaným datům přistupovat dvojím způsobem.
   - Spočítat průměrný nárůst cen všech potravin ročně a ten porovnat s ročním procentuálním nárůstem platů - v tomto případě ve všech letech byl zaznamenám procentuální nárůst minimálně o 10% ve prospěch cen.

   Pozn.: může být problematické, že u některých kategorií jen malý nárůst a u jiných velký nárůst
   - Porovnat meziroční nárůst cen jednotlivých potravin s nárůstem platů – rovněž byl zaznamenán procentuální nárůst minimálně o 10% ve prospěch cen ve všech letech.
   - V obou případech se jednalo o platy v oblasti „Administrativní a podpůrné činnost“. Z toho lze tedy vyvodit, že tam platy v porovnání s cenami potravin rostly nejpomaleji.

   https://github.com/mvrbova/SQL_Project/blob/main/Project_1_04_4.Question_prices_salary_10%25.sql

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?*

   Na základě dat z tabulek **t_Marie_Vrbova_project_SQL_primary_final** a **t_Marie_Vrbova_project_SQL_secondary_final** jsem si spočítala procentuální nárůst HDP, cen a mezd pro Českou republiku v jednotlivých rocích a vytvořila jsem pohled v_gdp_price_salary pro porovnání nárůstu HDP, cen a mezd v České republice v období 2006-2018.

   - Porovnáním procentuálního nárůstu HDP a cen a následně procentuálního nárůstu HDP a mezd v jednotlivých rocích není patrná spojitost mezi nárůstem HDP a nárůstem cen, popřípadě platů v daném nebo následujícím roce.
   - Pro ověření lze porovnat nárůst HDP s pohybem (nárůst či pokles) cen určitého produktu a poté mezd v určitém odvětví. Nicméně v případě cen ani mezd není patrný trvalý růst paralelně s růstem HDP.

   https://github.com/mvrbova/SQL_Project/blob/main/Project_1_05_5.Question_GDP_prices_salary.sql