1) POPIS 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.
Cílem projektu je připravit robustní datové podklady, ve kterých bude možné vidět porovnání dostupnosti potravin na základě průměrných příjmů za určité časové období.
Jako dodatečný materiál je potřeba připravit i tabulku s HDP, GINI koeficientem a populací dalších evropských států ve stejném období, jako primární přehled pro ČR.

Klíčové výzkumné otázky jsou tyto:

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 percentuálně 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?

2) POPIS PŘÍPRAVY PRIMÁRNÍ A SEKUNDÁRNÍ TABULKY

PRIMÁRNÍ TABULKA

- Nejprve jsem si na základě zadání, resp. výzkumných otázek zjistil jaká budu potřebovat data, tabulky, sloupce etc. pro vytvoření tabulky t_jan_blazek_project_sql_primary_final (pro data mezd a cen potravin za Českou republiku sjednocených na totožné porovnatelné období – společné roky)
- Po bližším ohledání jsem zjistil, že klíčovými tabulkami jsou především tabulky:
    - czechia_price
    - czechia_payroll
- A částečně také
    - czechia_price_category
    - czechia_payroll_industry_branch
- Z výzkumných otázek jsem si vypsal klíčová data, která budou třeba k jejich zodpovězení:
    - Roky
    - Mzdy

    - Všechna odvětví
    - Cena za litr mléka
    - Cena za kilogram chleba
    - První a poslední srovnatelné období v datech cen a mezd
    - Kategorie potravin
    - Ceny kategorií potravin
- V průběhu tvorby tabulky jsem se snažil využít především funkcí JOIN, zejména protože v zadání máme získat data za "srovnatelné období".
- Při prvním vytvoření tabulky, jsem si zapomněl vybrat payroll_year z tabulky czechia_payroll.
- Takže jsem smazal prve vytvořenou tabulku pomocí příkazu DROP TABLE
- Zopakoval jsem svůj napsaný sql dotaz na vytvoření tabulky, ale tentokrát s přidaným sloupcem payroll_year
- S takto připravenou tabulkou jsem byl schopen odpovědět na otázky 1 a 2, ale u třetí už se dotaz nerealizoval. Nehlásil sice chybu a běžel "do nekonečně". Tudíž jsem tabulku ještě jednou smazal s využitím příkazu DROP TABLE.
- V následujícím kroku jsem se pokusil při opětovné tvorbě tabulku optimalizovat, i s pomocí tipů od Jana K. (lektor), a to s využitím subqueries pro price a payroll, agregačních funkcí (AVG) pro value sloupce a funkcí GROUP BY. Na doporučení lektora Jana jsem použil LEFT JOIN na jejich spojení. S tím, že jsem začlenil JOINy pro pomocné tabulky, které jsem původně nechal na konci query do dílčích poddotazů. Na doporučení jedné z kolegyň jsem také nazval dílčí poddotazy "a" a "b", protože to působí esteticky hezčeji a čistěji. Po několika pokusech a experimentech se konečně podařilo vytvořit dotaz, který primární tabulku vytvořil během 1-2 vteřin, namísto "nekonečného běžení."

SEKUNDÁRNÍ TABULKA

- t_Jan_Blazek_project_SQL_secondary_final (pro dodatečná data o dalších evropských státech).
- Po bližším ohledání výzkumné otázky č. 5 jsem zjistil, že klíčovými tabulkami jsou především tabulky:
    - economies
    - czechia_price
    - czehia_payroll
- Z výzkumné otázky jsem si vypsal klíčová data, která budou třeba k jejímu zodpovězení:
    - Název země
    - Rok
    - Výška HDP
    - Průměrné mzdy
    - Průměrné ceny potravin
    - Sloupec obsahující kód pro průměrné mzdy
- Původní plán byl vytvořit vše pouze pomocí JOINů spojených prostřednictvím let ze všech tří tabulek a vyfiltrovat si Česko a kód pro průměrné mzdy (5958). Takto zadaný

dotaz ovšem trval věčnost při vytváření. Tudíž mi došlo, že bych měl opakovat podobný postup jako u první tabulky a dotaz optimalizovat pomocí subqueries.
- V následujícím kroku jsem tedy původní dotaz rozdělit do tří subqueries, spojených pomocí LEFT JOIN, přičemž se tabulka vytvořila téměř okamžitě.

3) POPIS POSTUPU A ODPOVĚDI NA VÝZKUMNÉ OTÁZKY

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

POSTUP:

- Nejprve bylo třeba vyfiltrovat z primární tabulky ze sloupce cpr_value_type hodnoty pro průměrnou mzdu (tedy 5958). Proto jsem si vytvořil CTE nazvaný *average_payroll.*
- Z něj jsem pak ve zkoumaném období 2006 až 2018 vyfiltroval nejvyšší a nejnižší mzdu, a vypočítal jejich rozdíl, abychom mohli zhodnotit míru růstu průměrné mzdy v každém odvětví ve zkoumaném období.
- Ve druhém detailnějším kroku jsem si vytvořil SELECT, ve které jsem si vyjel seznam průměrných mezd podle jednotlivých zkoumaných let pro jednotlivá odvětví. A doplnil ho podmínku pomocí funkce CASE s využitím funkce LAG, abych si vytvořil sloupec, který mi naznačí, zdali existovala odvětí, ve kterých došlo oproti předchozímu roku k poklesu.
- Pokusil jsem se vytvořit i podmínku, která by mi přímo spočítala o kolik došlo k poklesu, ale nakonec jsem si dané roky, kde došlo k poklesu našel a spočítal pokles manuálně.

ODPOVĚDI:

- Ve zkoumaném období 2006-2018 sdílí všechna odvětví vzestupnou trajektorii průměrné mzdy.
- Největší růst zaznamenaly tato tři odvětví:
    - Informační a komunikační činnosti - o 20 734 Kč
    - Výroba a rozvod elektřiny, plynu, tepla a klimatiz. vzduchu - o 17 334 Kč
    - Peněžnictví a pojišťovnictví - o 14 421 Kč
- Nejméně rostla naopak tato odvětví:
    - Administrativní a podpůrné činnosti - o 6 631 Kč.
    - Ostatní činnosti - o 6 754 Kč.
    - Ubytování, stravování a pohostinství - o 7380 Kč.
- Celkově platí, že ve všech odvětvích došlo ve zkoumaném období k růstu průměrné mzdy. U části odvětví ale existovaly roky, kdy došlo (většinou k mírnému) poklesu. Rok, ve kterém zaznamenal pokles největší množství odvětví byl rok 2013, ve kterém se propadlo 11 z 19 zkoumaných odvětví.
- Největší propad pak zaznamenala odvětví
    - Peněžnictví a pojišťovnictví - v roce 2013 cca o 4 479 Kč a

- Výroba a rozvod elektřiny, plynu, tepla a klimatiz. vzduchu v roce 2013 - o 1851 Kč
- Těžba a dobývání v roce 2009 - o 1093 Kč

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?

POSTUP:

- Nejprve jsem si pro forma pomocí MIN, MAX zjistil první a poslední rok pro roky u cen a u mezd.
- Pak jsem zaměřil na dvojí krok, a to s využitím CTE (funkce WITH):
  - První CTE nazvanou *average_pay* jsem zaměřil na zjištění průměru z průměrných mezd (ze všech odvětví) v prvním a posledním roce zkoumaného období.
  - Druhé CTE jsem se zaměřil na zjištění průměrné ceny mléka a chleba ve vybraných obdobích, tedy v letech 2006 a 2018.
- Oba CTE jsem pak spojil JOINem a v jeho rámci jsem podělil předem vypočítanou průměrnou mzdu v daném roce, cenou potravin v daném roce, abych zjistil kvantitu mléka a chleba v letech 2006 a 2018.

ODPOVĚDI

- Z výsledných dat pak vyplývá, že v roce 2006 šlo koupit za průměrnou mzdu (z celého roku):
  - 1 287 kg chleba
  - 1 437 l mléka
- V roce 2018 pak šlo koupit za průměrnou (průměrnou) mzdu
  - 1 342 kg chleba
  - 1 642 l mléka

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

POSTUP

- Využil jsem znalostí z lekce 4, kde jsme se naučili na sebe JOINovat tabulku posunutou o jeden rok, abychom zjistili meziroční změnu. Současně jsme využili vzorce pro výpočet procentní změny.

ODPOVĚDI

- Nejnižší růst, resp. dokonce pokles ceny, zaznamenala v roce 2007 - "Rajská jablka červená kulatá - a to o 30 procent.

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

POSTUP

- Vycházel jsem do značné míry z úpravy předchozího dotazu. Jen namísto jedné z hodnot jsem připravil dotaz pro dvě různé hodnoty, a to a) food_growth a b) payrol_growth. Dotaz fungoval, ale nabídl mi velké množství hodnot pro každý rok. Takže jsem musel navrhnout agregaci podle let.
- K agregaci dat jsem využil CTE, ve které jsou použil GROUP BY pro agregaci podle let.

ODPOVĚDI

- Ano, v roce 2008 existoval rozdíl až 34,2 procent v růstu cen potravin a cen mezd.

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?

POSTUP

- Původně jsem si ze sekundární tabulky vybral potřebné sloupce, tedy hlavně HDP, průměrné mzdy, průměrné ceny a rok. S tím, že jsem si vytvořil pomocné sloupce pomocí funkce CASE, ve kterých jsem sledoval pomocí funkce LAG, jestli je hodnota v následujícím roce vyšší než v předchozím. A na základě toho jsem chtěl hrubě odpovědět výzkumnou otázku. Plus jsem výsledek vyčistil o hodnoty „NULL“.
- Dlouho jsem zápasil, jak to udělat přesnější. Při samostudiu mi internet doporučil využít funkci CORR, kterou ale MariaDB nepodporuje. A teprve později mi došlo, že se dá opět využít výpočet pro procentuální růst, kde lze lépe a přesněji sledovat téma korelace.
- Tudíž jsem v druhém kroku rozepsal původní query o další krok, a to vytvoření dalších tří pomocných sloupců, které vypočítávají procentuální růst oproti předchozímu kroku u HDP, průměrné mzdy a cen potravin a doplňují tak sloupce, které definují pomocí "yes"/"no", jestli došlo k růstu či nikoliv.
- Jelikož není zadání uplně přesné, zaměřil jsem pouze na Česko

ODPOVĚDI

- Jednoduchá odpověď je, že neexistuje jasná pozitivní korelace mezi růstem HDP a růstem cen potravin/průměrných mezd. Z dat vyplývá, že je vztah poměrně chaotický a nedá se vyčíst jasný vzor, schéma. V datech pro Česko mezi lety 2007 a 2018 nalezneme všechny možné kombinace. Tedy jinými slovy data nám nabízí situaci, kdy došlo k růstu HDP a došlo k růstu cen i mezd (roky 2007, 2008, 2010, 2011, 2017, 2018). Najdeme ovšem i roky, kdy došlo k růstu HDP a k růstu mezd, zatímco došlo k poklesu cen (2014-2016). V roce 2012 také došlo sice k poklesu HDP, ale růstu mezd i cen (2012). Stejně jako najdeme roky, kdy došlo k poklesu HDP a došlo k poklesu alespoň jedné z kategorií cen/mezd (2009, 2013).
- Nelze tedy vyvodit jasnou korelaci mezi růstem HDP a růstem mezd a cen potravin.