SQL PROJECT

Data academy ENGETO

Vypracoval:

Radek Ivaniškin

A. ZADÁNÍ PROJEKTU

Cílem modelového projektu bylo připravit **robustní datové podklady** pro tiskové oddělení společnosti, která se zabývá životní úrovní občanů. V podkladech musí být vidět **porovnání dostupnosti potravin na základě průměrných příjmů za určité časové období**.

Jako dodatečný materiál tiskové oddělení požaduje také tabulku s HDP, GINI koeficientem a populací **dalších evropských států** ve stejném období, jako primární přehled pro ČR.

Na základě těchto podkladů (tabulek) je nutné odpovědět na soubor základních otázek tiskového oddělení:

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?

B. ANALÝZA DAT

Pro získání vhodného datového podkladu byly k dispozici následující datové sady:

Primární tabulky:

1. **czechia_payroll** – informace o mzdách v odvětvích za několikaleté období
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
7. **czechia_price_category** – číselník kategorií potravin

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** - informace o zemích na světě, např. hlavní město, měna, atd.
2. **economies** - HDP, GINI, daňová zátěž, atd. pro daný stát a rok.

Součástí prvotní analýzy dat bylo zejména:

- zjištění parametru pro spojení datových sad czechia_price a czechia_payroll (year)

- identifikace vybraných hodnot datových sad czechia_price a czechia payroll a jejich propojení skrze vybrané hodnoty na doprovodné tabulky a číselníky

- zjištění průniku časového období (roků) datových sad czechia_payroll a czechia price (období 2006 – 2018, kdy máme data pro mzdy i ceny)

- identifikace parametru pro spojení tabulek countries a economies (country – s omezením na evropské země)

- Link: https://github.com/Radis44/Project1_SQL/blob/main/first_data_analysis.sql

C. TVORBA ZÁKLADNÍCH TABULEK

Primary table: t_radek_ivaniskin_project_SQL_primary_final

- Prvním krokem bylo určení společných dat tabulek czechia price a czechia payroll (year).
- Časové období jsem omezil na roky 2006-2018, tedy období, pro které byla data úplná z obou datových sad.
- Následně jsem propojil obě tabulky s návaznými tabulkami obsahujícími názvy kategorií zboží (pro czechia price) a odvětvími (pro czechia payroll).
- Pro každou tabulku jsem vyfiltroval relevantní data pro dané časové období a pro jednotlivé kategorie zboží (pomocí GROUP funkce) a mezd jsem vypočítal hodnoty průměrných cen zboží a průměrných mezd za jednotlivá odvětví.
- Následně jsem pomocí jednotlivých funkcí zkontroloval správnost vypočtených hodnot pro jednotlivé roky – z ní vyplynulo, že 1 kategorie zboží (Víno jakostní) obsahovala hodnoty až od roku 2015.
- Výsledkem byly dvě nové tabulky s průměrnými cenami a průměrnými mzdami, které jsem následně pomocí funkce JOIN sloučil do finální tabulky (na základě parametru year).
- Link: https://github.com/Radis44/Project1_SQL/blob/main/primary_table_final.sql

Secondary table: **t_radek_ivaniskin_project_SQL_secondary_final**

- Pro vytvoření tabulky s dodatečnými informacemi (GDP, gini, population) jsem využil základní datové sady countries a economies (propojení funkcí JOIN)
- Vzhledem k tomu, že primární tabulka obsahuje data pro období 2006 -2018, zvolil jsem stejné období i pro secondary table (omezení podmínkou WHERE)
- Veškeré požadované informace obsahovala již tabulka economies, jejíž součástí ale nebylo rozdělení kontinentů, proto jsem zvolil spojení s tabulkou countries, kde bylo možné země pomocí podmínky WHERE omezit pouze na evropský kontinent
- Link: https://github.com/Radis44/Project1_SQL/blob/main/secondary_table.sql

D. FINALIZACE ODPOVĚDÍ NA ZÁKLADNÍ OTÁZKY

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

Odpověď:

Ve všech odvětvích průměrné mzdy **ve sledovaném období 2006 – 2018 kontinuálně nerostou**. Pokud porovnáme pouze počáteční (2006) a koncový (2018) rok, tak se mzdy zvýšily ve všech odvětvích, meziročně ale ve většině odvětvích v některých letech klesaly.

Existuje 5 odvětví (C – Zpracovatelský průmysl, H – Doprava a skladování, N – Administrativní a podpůrné činnosti, Q – Zdravotní a sociální péče, S – Ostatní činnosti), kde se mzdy navyšovaly v každém roce sledovaného období.

Link: https://github.com/Radis44/Project1_SQL/blob/main/answer1.sql

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

Odpověď:
Z průměrných cen a mezd za období 2006 – 2018 vyplývá, že:

- **v roce 2006** bylo možné průměrně koupit **1262 kg chleba a 1409 l mléka za průměrnou měsíční mzdu**, tedy **15143 kg chleba a 16905 l mléka** průměrně za celý rok

- **v roce 2018** bylo možné koupit **1319 kg chleba a 1614 l mléka za měsíční průměrnou mzdu**, tedy **15832 kg chleba a 19362 l mléka** průměrně za celý rok

Link: https://github.com/Radis44/Project1_SQL/blob/main/answer2.sql

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

   Odpověď:

   Jde o **kategorii 115 201 – Rostlinný roztíratelný tuk**, která mezi lety 2008 a 2009 zvýšila pouze o 0,01 procenta.

   Link: https://github.com/Radis44/Project1_SQL/blob/main/answer3.sql

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

   Odpověď:

   Z meziročního porovnání průměrných ročních cen potravin (všech kategorií) a průměrných mezd (za všechna odvětví) vyplývá, že v žádném roce ze sledovaného období (2006 – 2018) **nárůst cen potravin nepřekročil růst mez o více než 10 procent**.

   Link: https://github.com/Radis44/Project1_SQL/blob/main/answer4.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?**

   Odpověď:

   Na základě porovnání procentního růstu průměrných cen potravin, mezd a HDP v období 2006 – 2018 v České republice nelze konstatovat, že by výrazný růst HDP měl vliv růst cen a mezd.

   Například v roce 2007 vzrostlo HDP o 5,5 % a ceny i mzdy o 6,7 %. Ale v roce 2015 HDP vzrostlo o 5,4 %, jenže ceny poklesly o 0,5 % a mzdy vzrostly jen o 2,4 %.

   Link: https://github.com/Radis44/Project1_SQL/blob/main/answer5.sql