Název projektu:

Data o mzdách a cenách potravin a jejich zpracování pomocí SQL

Autor:

Pavla Šťastná

Popis:

Projekt se zaměřuje na analýzu ekonomických dat o průměrných mzdách a cenách vybraných potravin. Data pocházejí z veřejně dostupných zdrojů. Výstupem jsou dvě tabulky v databázi:

1. **t_{jmeno}_{prijmeni}_project_SQL_primary_final**
   Obsahuje data o mzdách a cenách potravin za Českou republiku sjednocených na totožné porovnatelné období (společné roky).

2. **t_{jmeno}_{prijmeni}_project_SQL_secondary_final**
   Obsahuje dodatečná data o dalších evropských státech.

Dalším cílem projektu je příprava sady SQL dotazů, které z připravených tabulek získají datový podklad k odpovězení následujících výzkumných otázek:

- Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?
- Kolik je možné si koupit litrů mléka a kilogramů chleba za první a poslední srovnatelné období v dostupných datech cen a mezd?
- Která kategorie potravin zdražuje nejpomaleji (tj. vykazuje nejnižší procentuální meziroční nárůst)?
- Existuje rok, ve kterém byl meziroční nárůst cen potravin výrazně vyšší než růst mezd (větší než 10 %)?
- Má výše HDP vliv na změny ve mzdách a cenách potravin? Tedy, pokud HDP vzroste výrazněji v jednom roce, projeví se to na cenách potravin či mzdách ve stejném nebo následujícím roce výraznějším růstem?

Klíčové funkce projektu:

1. Čištění a příprava dat:

Hlavním zdrojem pro přípravu primární tabulky jsou:

- **CZECHIA PAYROLL** (obsahuje informace o mzdách v různých odvětvích za několikaleté období).
- **CZECHIA PRICE** (obsahuje informace o cenách vybraných potravin za několikaleté období).

Tvorba primární tabulky probíhala ve třech krocích:

1. Seznámení a transformace dat z tabulky Czechia Payroll.
2. Seznámení a transformace dat z tabulky Czechia Price.
3. Propojení těchto dvou tabulek přes atribut rok.

Pro přípravu sekundární tabulky byly hlavním zdrojem tabulky **Economies** a **Countries**. Výstupem je tabulka obsahující informace o HDP, GINI koeficientu a populaci ve všech evropských zemích za roky 1960 až 2020.

Pro obě tabulky byly přes relace napojené číselníky, které jsem propojila a zobrazila pouze názvy kategorií pro větší přehlednost.

2. Problémy při realizaci:

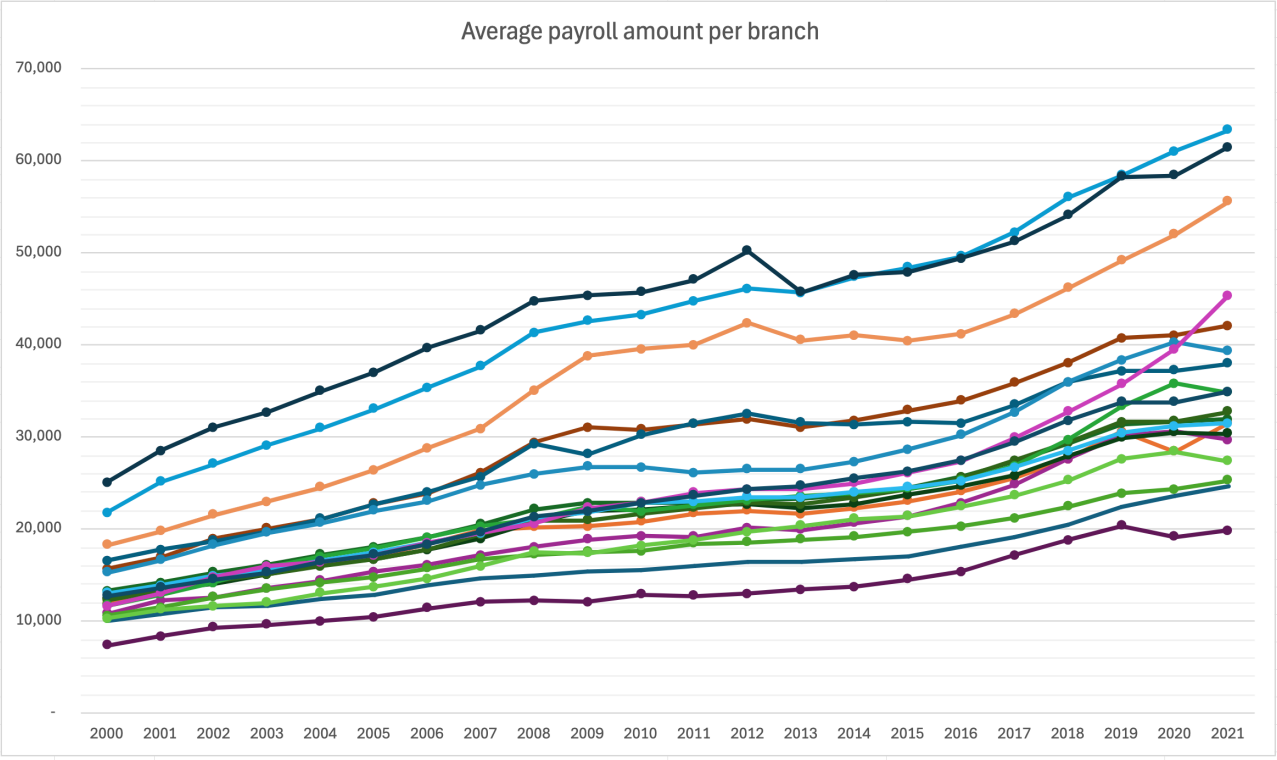
Během tvorby sekundární tabulky jsem omylem opakovaně spustila příkaz `INSERT INTO` a data v tabulce byla duplikována. Abych tomuto problému v budoucnu zabránila, přidala jsem omezení, které zajišťuje unikátnost kombinací hodnot. Při opakovaném spuštění příkazu `INSERT INTO` se nyní zobrazí chyba, pokud se pokusím vložit duplicitní data.

3. Analýza a odpovědi na otázky:

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

  Pro výpočet rozdílu oproti minulému období jsem využila funkci `LAG` spolu s podmínkovým zápisem. Když je hodnota rozdílu menší než 0, řádek je označen jako "descending"; když je hodnota rozdílu větší než 0, je označen jako "ascending".

  Vyhodnocení:
  Po rychlém průzkumu dat jsem identifikovala některé roky s hodnotami označenými jako "descending". Tyto hodnoty naznačují, že v některých letech a odvětvích průměrné mzdy meziročně klesly. Nicméně trend mezi lety 2000 až 2021 ukazuje obecný růst průměrných mezd. Přiložený graf poskytuje přehled, který tento trend vizualizuje.



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

  Při zpracování jsem využila funkci `AVG`, pomocí které bylo možné spočítat průměrnou hodnotu mezd v jednotlivých letech. Dále jsem použila funkce `MAX` a `MIN` v kombinaci s vnořeným dotazem jako filtr pro určení prvního a posledního období.

  Vyhodnocení:

  - **Za první období (rok 2006):** Lze koupit **1 287 kg chleba** a **1 437 l mléka** za průměrnou roční mzdu.
  - **Za poslední období (rok 2018):** Lze koupit **1 342 kg chleba** a **1 642 l mléka** za průměrnou roční mzdu.

- **Která kategorie potravin zdražuje nejpomaleji (s nejnižším procentuálním meziročním nárůstem)?**

  Pro potvrzení/vyvrácení hypotézy jsem vyhodnotila procentuální změny průměrných cen potravin mezi prvním a posledním obdobím (roky 2006 a 2018). Pro výpočet procentní změny jsem využila funkci LAG a dále použila funkce MIN a MAX k určení prvního a posledního období. Z hodnot procentní změny průměrných cen potravin jsem filtrovala pouze kladné hodnoty (indikující růst cen) pomocí subdotazu.

  Vyhodnocení:

  - Nejnižší procentní nárůst mezi sledovanými roky 2006 a 2018 vykazuje kategorie **banány žluté** s **7,4% růstem**.

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

  Vyhodnocení:

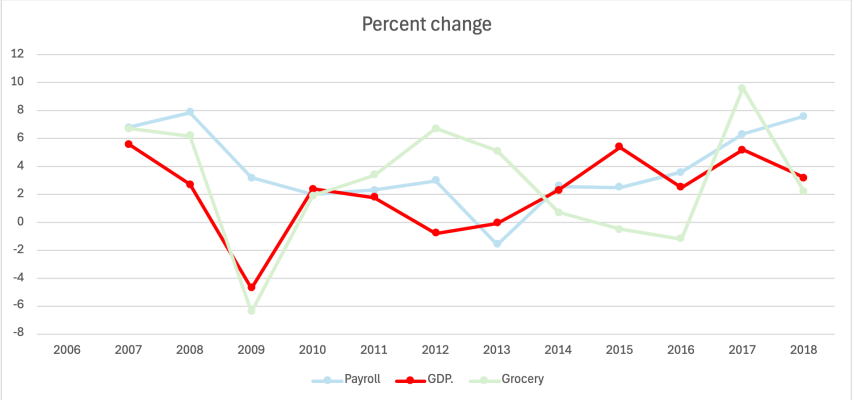
  - Neexistuje rok, ve kterém by byl meziroční nárůst cen potravin vyšší než 10 % ve srovnání s růstem mezd.

- **Má výše 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?**

  Při analýze jsem porovnávala meziroční růsty HDP, průměrných cen potravin a průměrných mezd v jednotlivých letech. Výstupy jsem také zpracovala graficky a vypočítala korelační koeficient.

Vyhodnocení:

- Dle korelačního koeficientu, který se blíží hodnotě **1**, lze konstatovat, že výše HDP má vliv na změny ve mzdách i cenách potravin.



Category	Correlation coefficient
GDP vs Payroll	0.92
GDP vs Grocery price	0.89