Autor: Martina Palusgová

E-mail: Fajkisovamartina@seznam.cz

Datum vytvoření: 22.01.2025

Název projektu: Projekt SQL Engeto.sql

Účel: SQL projekt vznikl v rámci kurzu Datová Akademie (ENGETO) k vyzkoušení si získaných znalostí a jako nutná součást k získání certifikace.

V rámci projektu jsem v GitHubu vytvořila repozitář Projekt_SQL_ENGETO, který obsahuje tyto soubory:

1. **Primary table.sql**
   SQL soubor s vytvořenou tabulkou *t_martina_palusgova_project_SQL_primary_final* z několika datových sad, která obsahuje základní data pro zodpovězení výzkumných otázek.
2. **Secondary table.sql**
   SQL soubor s vytvořenou tabulkou *t_martina_palusgova_project_SQL_secondary_final* obsahující doplňující data k zodpovězení 5. výzkumné otázky.
3. **Projekt SQL Engeto.sql**
   Základní SQL soubor obsahující datové podklady k odpovězení vytyčených výzkumných otázek. Příkazy v tomto souboru vycházejí z tabulek vytvořených v souborech Primary table.sql a Secondary table.sql.
4. **Zadání projektu.pdf**
   Oficiální zadání projektu od ENGETO.
5. **Průvodní listina.pdf**
   Wordovský soubor popisující vypracovaný projekt (soubory, postup, …).
6. **Projekt SQL úkol 5.xlsx**
   Excelovský soubor k 5. výzkumné otázce, obsahující výslednou tabulku a grafy.
7. **README.md**
   Soubor obsahující základní informace o projektu.

Použité datové sady jsou popsány v souboru Zadání projektu.

Primary table.sql

Dle mého názoru nejsložitější část projektu. Způsobů, jak získat výslednou tabulku je více. Rozhodla jsem se pro způsob, kdy se data ze dvou základních datových sad vypíší pod sebe pomocí operátoru UNION, s minimalizací počtu sloupců. Název každého sloupce byl zvolen tak, aby odpovídal charakteru dat z obou datových sad:

- year – období je rozděleno po letech

- value – ve sloupci value jsou obsaženy ceny potravin, průměrná hrubá mzda a průměrný počet zaměstnaných osob
- area_of_data – obsahuje kódy krajů ČR (nechala jsem jen kódy, jelikož jsem nepotřebovala konkrétní názvy krajů), název Průměrná hrubá mzda na zaměstnance a název Průměrný počet zaměstnaných osob
- name – obsahuje názvy potravin, názvy odvětví (vyskytuje se i NULL, jedná se o souhrnnou hodnotu za všechna odvětví)
- unit_value_or_industry_code – uvádí množství potravin a zkratky odvětví (NULL je při souhrnné hodnotě za všechna odvětví)
- unit_name – obsahuje jednotky hodnot

Využila jsem Common Table Expression, kde jsem každou datovou sadu spojila pomocí LEFT JOIN se souvisejícími číselníky. Abych odstranila nesoulad v časových intervalech, zprůměrovala jsem hodnoty na roční úroveň. Z důvodů odlišnosti období datových sad jsem je sjednotila na porovnatelné období 2006-2018.

Výzkumné otázky a odpovědi

1. **Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?**
   U této otázky mám vypracovány 3 SQL dotazy:
   - Vývoj jednotlivých odvětví v průběhu let.
   - Ve druhém dotazu jsem nechala vypsat ke každému odvětví jeho minimální hodnotu increase. V případě hodnoty menší než 0, nemůžeme říct, že mzdy v průběhu let v daném odvětví rostly.
   - V posledním dotazu jsem pak nechala vypsat odvětví, ve kterých nedošlo k meziročnímu poklesu mezd.

   Z hodnot vyplývá, že ke každoročnímu navýšení platů v letech 2006 až 2018 došlo jen ve 4 odvětvích: Zpracovatelský průmysl, Zdravotní a sociální péče, Ostatní činnosti, Doprava a skladování.
2. **Kolik je možné si koupit litrů mléka a kilogramů chleba za první a poslední srovnatelné období v dostupných datech cen a mezd?**
   Z hodnot vyplývá, že v roce 2018 jsme si mohli z platu koupit větší množství mléka a chleba, oproti roku 2006. Konkrétní množství je uvedeno ve sloupci number_of_commodity.

year	czech_salary	average_value_of_commodity	number_of_commodity	name
2006	19218,9	17,4	**1107**	Chléb konzumní kmínový
2006	19218,9	14,5	**1325**	Mléko polotučné pasterované
2018	31520,5	24,8	**1273**	Chléb konzumní kmínový
2018	31520,5	20,3	**1552**	Mléko polotučné pasterované

3. **Která kategorie potravin zdražuje nejpomaleji (je u ní nejnižší percentuální meziroční nárůst)?**
   K této otázce jsem napsala 2 SQL dotazy:

- V prvním dotazu mám vypsán meziroční vývoj cen jednotlivých potravin.
- Druhý dotaz udává percentuální meziroční změnu cen jednotlivých kategorií potravin souhrnně a to pomocí součtu i průměru.

Meziroční vývoj cen má u jednotlivých kategorií velké výkyvy a tak ze souhrnného pohledu (součet i průměr hodnot) má nejnižší percentuální meziroční nárůst cukr krystal. Druhý v pořadí by byl banán žlutý.

4. **Existuje rok, ve kterém byl meziroční nárůst cen potravin výrazně vyšší než růst mezd (větší než 10 %)?**
   Meziroční nárůst cen potravin vyšší než 10% byl v roce 2007. V tomto roce byl zároveň druhý největší rozdíl mezi nárůstem cen potravin a platů (největší byl v roce 2013).
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?**
   K tomuto úkolu je vypracován excelovský soubor *Projekt SQL úkol 5*, kde je jak tabulka s datovými podklady, tak i grafy znázorňující porovnání ročního vývoje HDP, cen potravin a mezd.
   Výška HDP má vliv na změny ve mzdách a cenách potravin. Dle grafů má větší spojitost HDP se mzdami než HDP s cenami potravin. Ceny potravin mají časté výkyvy vzhledem k HDP a tak na druhou otázku nelze s jistotou odpovědět. Domnívám se, že změny HDP se projevují na mzdách již ve stejném roce.

Závěr

SQL projekt byl sice náročný, ale pro vyzkoušení získaných znalostí velice užitečný. Vyzkoušela jsem si v něm příkazy, které jsem se naučila v kurzu, ale musela jsem také hledat v dokumentaci MariaDB jiné vhodné příkazy. V rámci projektu jsem se také musela naučit základy Gitu. Při vyhotovování projektu jsem měla delší přestávku. Nevýhodou této přestávky bylo, že jsem téměř se vším musela začít znovu. Naopak výhodou bylo, že jsem si mohla příkazy zopakovat, přičemž některé nabyté znalosti do sebe lépe zapadly.