Finální SQL Projekt pro Engeto – Albert Kereškéni

Tento dokument slouží jako průvodní listina k závěrečnému projektu pro *Kurz Datové Analýzy* od společnosti *Engeto*. Cílem tohoto projektu je odpovědět na zadané výzkumné otázky a prokázat tím znalost jazyka SQL. Otázky znějí:

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?

Tvorba první primární tabulky

Abych mohl na výzkumné otázky odpovědět, potřeboval jsem primární tabulku, ze které bych čerpal data. Podíval jsem se tedy na dostupné tabulky a vybral z nich sloupce, které se mi hodily.

Díval jsem se hlavně na tabulku *czechia_payroll* s informacemi o mzdách a *czechia_price* s informacemi o cenách potravin. Bohužel se mi nepodařilo obě tabulky spojit hned do jedné, jelikož jsem měl potíže s různými formáty času. Proto jsem si vytvořil dvě pomocné tabulky.

Tabulka *t_help_a* obsahuje přehledné informace o mzdách. Z tabulky *czechia_payroll* jsem vybral sloupec „payroll_year“ a „value“. Jelikož tabulka obsahuje i sloupec s kvartály, sloupec „value“ jsem zprůměroval a tím zjistil průměrnou celoroční mzdu. Poté jsem připojil sloupec „name“ z tabulky *czechia_payroll_industry_branch*, abych měl v tabulce celé názvy jednotlivých odvětví. Nakonec jsem ještě přidal filtry, aby mi sloupec „value“ ukazoval jen mzdy, aby se mi nezobrazovaly mzdy bez odvětví a aby se mi zobrazoval *payroll_year* jen od roku 2006 do roku 2018.

Tabulka *t_help_b* obsahuje přehledné informace o cenách potravin. Toto byl trochu oříšek, jelikož tabulka *czechia_price*, ze které jsem čerpal, ukazuje data vždy po týdnech. To bylo pro mé účely příliš podrobné. Vzal jsem tedy sloupce „value“ a ze sloupce „date_from“ extrahoval. Poté jsem připojil tabulku *czechia_price_category* a vybral z ní sloupce „name“, „price_value“ a „price_unit“, abych měl v tabulce celé názvy

potravin a údaje o počítaném množství. Ještě jsem dodal filtr, abych měl celorepublikové údaje.

Nakonec jsem obě tabulky spojil do jedné a tím vytvořil první primární tabulku. Zprůměroval jsem sloupec „value“, abych zjistil průměrné ceny potravin za celý rok, a získal všechny potřebné sloupce. Nyní jsem se mohl odebrat na první cvičení

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

Z dat vidíme, že **pokles cen pocítilo** patnáct odvětví z devatenácti. V devíti odvětvích mzdy meziročně klesly jedenkrát, v pěti odvětvích dvakrát a v odvětví *Těžba a dobývání* mzdy meziročně klesly dokonce čtyřikrát. Nejčastějším rokem poklesu byl rok 2013.

Jediná odvětví, kterých se **pokles mezd nedotknul**, byla *Doprava a skladování*, *Ostatní činnosti*, *Zdravotní a sociální péče* a *Zpracovatelský průmysl*.

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

V této otázce jsem vzal všechna průmyslová odvětví a porovnal, kolik si za svou průměrnou mzdu mohli koupit kilogramů chleba a litrů mléka v letech 2006 a 2018. Z dat jsem zjistil, že u chlebu si pět odvětví mohlo koupit více chleba v roce 2006 a 14 odvětví v roce 2018. Co se týče mléka, tak existuje jen jedno odvětví, které si mohlo v roce 2006 koupit více litrů mléka než v roce 2018, a tím je *Peněžnictví a pojišťovnictví*.

Nejvíce chleba si mohli koupit lidé v odvětví *Peněžnictví a pojišťovnictví* v roce 2006, konkrétně 2462 kilogramů. **Nejméně chleba** si zase mohli dovolit lidé v *Ubytování, stravování a pohostinství* v roce 2006, a to jen 707 kg.

Nejvíce mléka si mohli koupit lidé v odvětví *Informační a komunikační činnosti* v roce 2018, a to 2831 litrů**. Nejméně mléka** si pro změnu mohli dovolit lidé z *Ubytování, stravování a pohostinství* v roce 2006, konkrétně pouhých 789 litrů.

Pro zjištění těchto údajů jsem upravil kód tak, že jsem do ORDER BY přidal na první místo sloupec „quantity“. Následně pak filtroval vzestupně, nebo sestupně.

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

Z dat lze vidět, že **nejnižší percentuálně meziroční nárůst měly banány žluté**, a to průměrně 0,81 %. Potraviny „cukr krystalový“ a „Rajská jablka červená kulatá“ dokonce zlevnily. Cukr měl průměrný meziroční pokles 1,92 % a Rajská jablka 0,74 %.

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

Data nám ukazují, že za žádný rok **nebyl výrazně vyšší narůst cen potravin než růst mezd**. Největší rozdíl v nárůstu byl v roce 2013, kdy průměrný vzrůst ceny potravin byl o 6,78 % větší než vzrůst cen. V roce 2009 se málem stal opak, kdy rozdíl mezi nárustem mezd a cen byl 9,55 %. Nikdy ale nebyl větší než 10 %.

Tvorba sekundární tabulky

K vykonání pátého cvičení byla za potřebí sekundární tabulka, která by obsahovala údaje o HDP. Sekundární tabulku jsem tedy vytvořil spojením primární tabulky, kde jsou všechna zbylá potřebná data, a tabulky *economies*, kde jsou právě údaje o DPH jednotlivých zemích za určité roky.

Stačilo jen z *economies* vyfiltrovat údaje HDP pro Českou republiku za roky 2006–2018 (jelikož to je naše srovnatelné období), spojit je s primární tabulkou a tím vytvořit onu sekundární tabulku. Také jsem si sloupec „gdp“ předělal na **numeric**, abych mohl používat funkce **AVG** a **ROUND** v dalším dotazu.

Cvičení 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?

Z dat lze vyčíst, že **změna HDP nemá vliv na změnu cen potravin, ani na změnu mezd**. Ve výsledku vidíme, o kolik procent se oproti předchozímu roku zvýšilo HDP, ceny potravin a mzdy. Vzhledem k rozdílnosti hodnot jednotlivých sloupců je patrné, že mezi růstem/poklesem HDP, cen a mezd není žádná korelace.

HDP například kleslo v letech 2009, 2012 a 2013, zatímco ceny potravin klesly v letech 2009, 2014, 2015 a 2016 a mzdy klesly jen v roce 2013.