Data o mzdách a cenách potravin a jejich zpracování pomocí SQL

Projekt v rámci kurzu „Datová akademie“ Engeto

(Běh kurzu od 26. 9. 2023)

11. listopadu 2023

Taťána Dudáčková

tanuldab@gmail.com

Obsah (struktura projektu)

PŘEDSTAVENÍ PROJEKTU A JEHO CÍL 3

INFORMACE Z TVORBY PROJEKTU + INFORMACE O DATECH 3

PRIMÁRNÍ TABULKA 3

SEKUNDÁRNÍ TABULKA 7

ODPOVĚDI NA VÝZKUMNÉ OTÁZKY 7

VÝZKUMNÁ OTÁZKA Č. 1 7

VÝZKUMNÁ OTÁZKA Č. 2 8

VÝZKUMNÁ OTÁZKA Č. 3 8

VÝZKUMNÁ OTÁZKA Č. 4 9

VÝZKUMNÁ OTÁZKA Č. 5 9

Představení projektu a jeho cíl

Cílem tohoto projektu je připravit datové podklady týkající se dostupnosti základních potravin v České republice. Součástí datových podkladů je analýza údajů o mzdách ve vybraných odvětvích a o cenách potravin. Celkově bude zodpovězeno pět výzkumných otázek, přičemž zjištěné skutečnosti by mohly být později prezentovány na tiskové konferenci.

Podkladem pro analýzu se staly datové sady pocházející z Portálu otevřených dat ČR. Z dostupných datových sad byly vytvořeny dvě tabulky (primární a sekundární tabulka) s pomocí jazyka SQL, databáze Maria DB a programu DBeaver.

Informace z tvorby projektu + informace o datech

Primární tabulka

Jeden z hlavních problémů, se kterým jsem se při zpracovávání setkala, byla značná výpočetní náročnost příkazů při tvorbě primární tabulky. Původně jsem měla v úmyslu vytvořit primární tabulku v rámci jednoho příkazu, bohužel se mi to ale nepodařilo – s příkazem jako takovým můj počítač problém neměl, jakmile jsem ale z již sestaveného příkazu chtěla vytvořit novou tabulku, narazila jsem. Příkaz se prováděl stovky sekund a nakonec jsem jej musela přerušit. Z tohoto důvodu jsem tvorbu primární tabulky rozdělila do několika mezikroků, jak bude ještě popsáno dále. Na základě řady různých pokusů si myslím, že „problémová“ je při spojování tabulka czechia_price vzhledem ke své obsáhlosti.

Primární tabulku jsem vytvořila propojením několika tabulek z dostupných datových sad. Konkrétně šlo o tyto tabulky:

- Czechia_payroll
- Czechia_price
- Economies
- Czechia_payroll_industry_branch
- Category_code
- Czechia_region

Kromě toho jsem ještě připojila z tabulek Czechia_payroll a Czechia_price sloupce s hodnotami za předchozí a následující rok.

Jak jsem již zmínila, primární tabulku jsem tvořila v několika mezikrocích, ve kterých jsem tvořila pomocné tabulky, které jsem postupně slučovala. Rovněž jsem v průběhu tvorby tabulky odstranila několik sloupců, které ve finální tabulce vzhledem k zadání nebudou potřeba. Sloupce jsem také přejmenovala tak, aby bylo na první pohled více zřejmé, jaká data jsou jejich obsahem. Přejmenování navíc bylo nutné i vzhledem k tomu, že se v původních slučovaných tabulkách nacházely stejně pojmenované slupce (např. value).

Nyní popíšu jednotlivé mezikroky.

Mezikrok č. 1a

Propojila jsem tabulku czechia_payroll s totožnou tabulkou tak, abych ve výsledné tabulce získala nový sloupec s hodnotami za předchozí rok (bude nutné pro následné meziroční porovnávání). Použila jsem příkaz LEFT JOIN, přičemž u podmínky ON jsem vybrala co nejvíce společných sloupců, aby se tabulky spojily správně. Posunutí o rok jsem dosáhla pomocí podmínky +1 u sloupců payroll_year.

Vybrané sloupce jsem již v tomto kroku přejmenovala. Rovněž jsem si stanovila příkazem WHERE další podmínky:

- Hodnota 5958 u sloupce value_type_code: z pomocné tabulky czechia_payroll_value_type je možné zjistit, co dvě hodnoty, které mohou být obsahem sloupce, vlastně znamenají. Hodnota 5958 znamená průměrnou hrubou mzdu na zaměstnance.
- Hodnota 200 u sloupce calculation_code: z pomocné tabulky czechia_payroll_calculation zjistíme, že hodnota 200 znamená přepočtený údaj, v tomto případě přepočtenou mzdu na plné úvazky (nebudou započteny částečné úvazky).
- Dále jsem zvolila podmínku, že sloupec industry_branch_code nesmí obsahovat prázdné buňky (podmínka IS NOT NULL).

Mezikrok č. 1b

K tabulce z mezikroku 1a jsem připojila znovu tabulku czechia_payroll, tentokrát ale tak, abych získala v novém sloupci hodnoty za následující rok, opět za účelem možnosti porovnávat. Postup byl analogický jako u mezikroku 1a.

Mezikrok č. 1c

Stejně jako jsem v předchozích dvou mezikrocích propojovala tabulky czechia_payroll samy na sebe, připojovala jsem v tomto mezikroku (zcela odděleně od mezikroků 1a a 1b) tabulku czechia_price samu na sebe. Opět kvůli tomu, abych získala dva nové sloupce s daty posunutými o rok zpět a o rok dopředu.

Rovněž jsem přejmenovala sloupce. Datum jsem převedla na jiný formát, respektive jsem původní datum rozdělila na dílčí údaje: rok, čtvrtletí, měsíc, týden, každý ve zvláštním sloupci. Datový typ je integer. Toto bylo potřeba kvůli následnému propojení s tabulkou mezd. Sloupce měsíc a týden bychom teoreticky k odpovědím na výzkumné otázky nepotřebovali, ale pro jistotu jsem je v tabulce nechala kvůli následnému spojování i tomu, že se celkově může hodit vědět, k jakému týdnu se údaje vztahují.

Mezikrok č. 1d

Analogicky jako v předchozím případě jsem připojila k tabulce z mezikroku 1c znovu tabulku czechia_price, posun byl ale na opačnou stranu než v mezikroku 1c.

Mezikrok č. 2

Propojuji obě pomocné tabulky z mezikroku 1b (czechia_payroll „obohacená“ o sloupce s hodnotami za předchozí a následující rok) a 1d (czechia_price opět se dvěma novými sloupci pro hodnoty za předchozí a následující rok). Přebytečné sloupce odstraním až později. Podmínka u spojení tabulek (join) byla, že se musí rovnat roky a čtvrtletí v obou tabulkách. Použila jsem LEFT JOIN s výchozí tabulkou pro ceny, tj. ve výsledné tabulce zůstanou hodnoty z tabulky pro ceny a připojí se data o mzdách. Ve výsledné tabulce nebudou hodnoty z několika počátečních let z tabulky o mzdách, což je dané tím, že mzdy se začaly měřit dříve (a v tabulce o cenách nemáme hodnoty pro dané roky a čtvrtletí zjištěné).

Mezikrok č. 3

Zbavuji se přebytečných sloupců. Odstranila jsem sloupce pro některé číselníky (value_type_code, unit_code, calculation_code) – jsou již zbytečné, jelikož jsme v prvním mezikroku definovali, které hodnoty z nich potřebujeme. Pro jistotu jsem si ale pomocí SELECT DISTINCT

ověřila, že buňky v daném sloupci, který odstraňuji, obsahují vždy už jen jednu hodnotu (že není ještě nějaká jiná možnost).

Dále jsem odstranila duplicitní sloupce o datumech a sloupce mzdy_id a ceny_id (dále je již nebudu potřebovat).

Mezikrok č. 4

K pomocné tabulce z mezikroku 3 jsem připojila údaje o HDP z tabulky economies, opět pro základní rok a pro rok posunutý o 1, čímž získám údaje o HDP za předchozí rok.

Musela jsem si stanovit podmínku rovnosti na základě společného roku, podmínkou je rovněž název země „Czech Republic“. Další sloupce z tabulky economies pro účely projektu nepotřebuji.

Tvorba finální verze tabulky:

K pomocné tabulce z mezikroku č. 4 jsem ještě připojila tabulky czechia_payroll_industry_branch, category_code a czechia_region = pro větší přehlednost jsem chtěla, aby se ve výsledné tabulce nezobrazovaly kódy zboží, kódy odvětví a kódy krajů, ale jejich názvy. Propojení jsem provedla na základě podmínky rovnosti kódů ve spojovaných tabulkách.

Výsledkem je finální primární tabulka s názvem t_tatana_dudackova_project_sql_primary_final, která obsahuje tyto sloupce:

- Nazev_odvetvi
- Rok
- Ctvrtleti
- Mesic
- Tyden
- Vyse_mezd
- Vyse_mezd_prev_year
- Vyse_mezd_next_year
- Vyse_cen
- Vyse_cen_prev_year
- Vyse_cen_next_year
- Nazev_zbozi
- Kraj
- Gdp
- Gdp_prev_year

Na základě primární tabulky je možné zodpovědět výzkumné otázky.

Sekundární tabulka

Sekundární tabulka obsahuje dodatečná ekonomická data pro jiné evropské státy. Název této tabulky je: t_tatana_dudackova_project_sql_secondary_final.

Tabulka vznikla propojení vybraných sloupců z tabulek countries a economies.

Tabulka obsahuje tyto sloupce:

- Country
- Population
- Year
- Gdp
- Gini
- Taxes

Propojení vzniklo na základě podmínky rovnosti názvů zemí v obou spojovaných tabulkách, dále bylo přidáno omezení, že název kontinentu z tabulky countries musí být „europe“, tedy Evropa (data z jiných kontinentů nejsou pro účely tohoto projektu podstatná).

Bylo také ještě nutné přidat podmínku s časovým rozmezím, aby výsledná tabulka obsahovala údaje za stejné časové období, jako primární tabulka. Toho jsem docílila přidáním podmínky WHERE na konci příkazu. Zde jsem narazila asi na jediný problém při tvorbě této tabulky. Původně jsem měla v úmyslu provést výběr roků prostřednictvím IN, vnořeného ORDER BY a LIMIT, systém ale vrátil hlášku, že tato verze MariaDB zvolenou kombinaci příkazů zatím nepodporuje (z toho soudím, že jsem příkaz měla správně, jen jsem narazila na omezení databáze MariaDB). Přišla jsem ale na alternativní možnost, která fungovala.

Odpovědi na výzkumné otázky

Výzkumná otázka č. 1

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

Bylo zjištěno, že v průběhu let v některých odvětvích a obdobích mzdy klesly. O která odvětví jde, zjistíme ze sloupce „nazev_odvetvi“, pokud sloupec „mezirocni_srovnani“ obsahuje text „mzdy klesly“. Tabulka je seřazena právě podle tohoto sloupce (mezirocni_srovnani), takže na prvních místech se

zobrazí odvětví, kde došlo někdy k poklesu. Dále je tabulka srovnána dle názvů odvětví a časových údajů. Celkově došlo k poklesu 114x (hodnoty s textem „mzdy vzrostly“ začínají na řádku 115). V některých odvětvích došlo k poklesu víckrát.

Výzkumná otázka č. 2

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

Zjistila jsem, že prvním srovnatelným obdobím je první čtvrtletí roku 2006, posledním srovnatelným obdobím je pak poslední čtvrtletí roku 2018.

Zjistila jsem, že za průměrnou mzdu 20014 Kč v 1. čtvrtletí roku 2006 si bylo možné koupit 1357,56 bochníků chleba a 1405,95 litrů mléka. Průměrná cena chleba byla 14,74 Kč, průměrná cena mléka byla 14,24 Kč.

V roce 2018 (4. čtvrtletí) pak byly průměrné ceny 23,86 Kč (chléb) a 19,47 Kč (mléko), průměrná mzda činila 35104 Kč a bylo možné si za ni koupit 1471,32 kg chleba či 1802,67 litrů mléka.

U průměrných mezd jsem nerozlišovala různá odvětví.

Údaje o mzdách jsou zaokrouhlené na celá čísla, data o cenách jsou zaokrouhlená na dvě desetinná místa.

Výzkumná otázka č. 3

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

Vytvořila jsem dvě varianty příkazů, pomocí obou z nich jsem ale došla ke stejnému závěru. Nejnižší percentuální meziroční nárůst nastal v roce 2014 u potraviny s názvem „vepřová pečeně s kostí“ (procentuální nárůst tehdy činil pouhých 0,0089221984 %.

Z druhé varianty dotazu je pak zřejmé, které další potraviny také zdražovaly pomalu, popř. ve kterých letech (na dalších „příčkách nejpomalejšího zdražování“ se ještě 2x umístila vepřová pečeně s kostí (2007, 2009) a poté hovězí maso zadní bez kosti (2010, 2009)).

Nepočítala jsem se zápornými změnami ceny (podmínka, že meziroční změna musí být vyšší než 0) - záporné změny by znamenaly zlevňování a o to v této výzkumné otázce nejde.

Výzkumná otázka č. 4

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

Ne, neexistuje, meziroční nárůst cen potravin nikdy nevyšel vyšší než růst mezd o 10 % a více. Tato skutečnost vyplývá ze sloupce porovnani_rustu_mezd_a_potravin.

Výzkumná otázka č. 5

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?*

Nejprve krátký komentář ke zvoleným hranicím, které říkají, zda HDP, mzdy a ceny rostou výrazněji či nikoli. Určila jsem si tuto hranici u všech tří položek na 5 %. Ze zadání dále nevyplývalo, zda se má významnost růstu mezd/potravin ve stejném a dalším roce posuzovat dohromady nebo zvlášť. Příkaz jsem tedy napsala tak, že podmínka platí, pokud nad stanovenou hranici vzrostla alespoň jedna z obou kategorií (mzdy/ceny). Dále se mohlo stát, že mzdy/ceny vzrostly nad stanovenou hranici 5 %, ale ne v důsledku výrazného růstu HDP – taková varianta nás ale nezajímá, jelikož posuzujeme vliv růstu HDP a ne vliv jiných (nedefinovaných) faktorů. Pokud tedy HDP nerostl tempem 5 % a výše, už nás nezajímá, jak se procentuálně měnily cen a/nebo mzdy (a to i kdyby jejich růst přesáhl určenou hranici). Kromě toho samozřejmě mohl výrazně vzrůst HDP, ale bez toho, že by růst výrazně ovlivnil ceny a mzdy.

Jaké tedy mohly nastat varianty?

- Vlivem výrazného růstu HDP došlo k výraznému růstu mezd a/nebo cen
- Nastal výrazný růst HDP bez dalších vlivů
- Nevýrazný růst HDP, popř. pokles
- (Jiná možnost)

Bylo zjištěno, že:

V letech 2006-2007 HDP vzrostl výrazně a v důsledku toho došlo k výraznému růstu mezd a/nebo cen ve stejném roce i v následujících letech (tj. 2007 i 2008). Stejná situace pak nastala ještě v roce 2017.

V roce 2015 pak došlo k výraznému růstu HDP, ale na mzdách ani cenách se to výrazněji neprojevilo, a to ani v následujícím roce.

V jiných letech nedošlo k růstu HDP nad 5 %.