ENGETO

SQL PROJEKT

Zpracovala: Michaela Kosová

michaela.kosova@gmail.com

Srpen 2023

Obsah

1 Úvod

1.1 Účel dokumentu

V rámci kurzu Datová Akademie, pořádaném společností ENGETO s.r.o., byl v závěru první části kurzu zadán SQL projekt. Tento dokument má za cíl shrnout zadání projektu, popsat postupy při řešení jednotlivých otázek a na tyto otázky odpovědět.

2 Zadání projektu[1]

2.1 Úvod do projektu

Na vašem analytickém oddělení nezávislé společnosti, která se zabývá životní úrovní občanů, jste se dohodli, že se pokusíte odpovědět na pár definovaných výzkumných otázek, které adresují **dostupnost základních potravin široké veřejnosti**. Kolegové již vydefinovali základní otázky, na které se pokusí odpovědět a poskytnout tuto informaci tiskovému oddělení. Toto oddělení bude výsledky prezentovat na následující konferenci zaměřené na tuto oblast.

Potřebují k tomu **od vás připravit robustní datové podklady**, ve kterých bude možné vidět **porovnání dostupnosti potravin na základě průměrných příjmů za určité časové období**.

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

Datové sady, které je možné použít pro získání vhodného datového podkladu

2.2 Primární tabulky

1. czechia_payroll – Informace o mzdách v různých odvětvích za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.
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.

[1]Převzato z: https://learn.engeto.com/cs/kurz/projekt-z-sql/studium/7oWBZzRmTOeC9Y7Vz9gzOQ/zadani-projektu-data-o-mzdach-a-cenach-potravin-a-jejich-zpracovani-pomoci-sql/zadani-projektu?originId=n7w-wcCDQNuaSwesN57eJA&originCourse=Data%20Academy&originLesson=5.3:%20Zad%C3%A1n%C3%AD%20projektu:%20Data%20o%20mzd%C3%A1ch%20a%20cen%C3%A1ch%20potravin%20a%20jejich%20zpracov%C3%A1n%C3%AD%20pomoc%C3%AD%20SQL

6. czechia_price – Informace o cenách vybraných potravin za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.
7. czechia_price_category – Číselník kategorií potravin, které se vyskytují v našem přehledu.

2.3 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.

2.4 Dodatečné tabulky:

1. countries - Všemožné informace o zemích na světě, například hlavní město, měna, národní jídlo nebo průměrná výška populace.
2. economies - HDP, GINI, daňová zátěž, atd. pro daný stát a rok.

2.5 Výzkumné otázky

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?

2.6 Výstup projektu

Pomozte kolegům s daným úkolem. Výstupem by měly být dvě tabulky v databázi, ze kterých se požadovaná data dají získat. Tabulky pojmenujte t_{jmeno}_{prijmeni}_project_SQL_primary_final (pro data mezd a cen potravin za Českou republiku sjednocených na totožné porovnatelné období – společné roky) a t_{jmeno}_{prijmeni}_project_SQL_secondary_final (pro dodatečná data o dalších evropských státech).

Dále připravte sadu SQL, které z vámi připravených tabulek získají datový podklad k odpovězení na vytyčené výzkumné otázky. Pozor, otázky/hypotézy mohou vaše výstupy podporovat i vyvracet! Záleží na tom, co říkají data.

Na svém GitHub účtu vytvořte repozitář (může být soukromý), kam uložíte všechny informace k projektu – hlavně SQL skript generující výslednou tabulku, popis mezivýsledků (průvodní listinu) a informace o výstupních datech (například kde chybí hodnoty apod.).

Neupravujte data v primárních tabulkách! Pokud bude potřeba transformovat hodnoty, dělejte tak až v tabulkách nebo pohledech, které si nově vytváříte.

3 Zdrojová data a software

Projekt je zpracován nad lokálně staženou databází poskytnutou společností ENGETO. Pro přístup k databázi byl použit program DBeaver. Výstupy projektu jsou odevzdány prostřednictvím sdíleného repozitáře webové aplikace GitHub.

4 Postup zpracování

Po prostudování primárních a dodatečných tabulek zmíněných v bodě 2.2, 2.3 a 2.4 jsem se pokusila vytvořit dvě souhrnné tabulky, v mém případě jsem se rozhodla pro pohledy („views“), které by agregovaly relevantní data dle zadání v bodě 2.6:

1. v_michaela_kosova_project_SQL_primary_final – tabulka shrnující průměrné roční hrubé mzdy dle jednotlivých průmyslových odvětví a zároveň průměrné roční ceny potravin dle jednotlivých kategorií potravin. Data jsou sjednocena na totožná srovnatelná období, tj. společné roky 2006-2018.
2. v_michaela_kosova_project_SQL_secondary_final – tabulka zahrnující dodatečná data o evropských státech, zejména HDP pro účely zodpovězení otázky č.5, pro stejné období let 2006-2018.

Nad těmito tabulkami (pohledy) jsem se následně pokusila formulovat SQL dotazy tak, aby jejich výsledek dával relevantní odpovědi na výzkumné otázky formulované v bodě 2.5.

5 Odpovědi na výzkumné otázky

5.1 Vytvoření tabulky „Primary“

Při vytvoření tabulky v_michaela_kosova_project_SQL_primary_final jsem se rozhodla pro použití pohledu („view“), protože pouze potřebuji přehledně uspořádat některá vybraná data z primárních tabulek a dále nad tímto pohledem definovat navazující SQL dotazy.

Do pohledu jsem zahrnula následující sloupce (viz soubor v_michaela_kosova_project_SQL_primary_final.sql):

- Rok / year
- Průmyslové odvětví / industry_branch
- Průměrná hrubá roční mzda na zaměstnance / average_wage – výpočet průměrné hrubé roční mzdy pro dané průmyslové odvětví pomocí funkce „average“
- Rok pro potraviny / year_food
- Kategorie potravin / food_category
- Jednotka potravin / food_unit
- Průměrná roční cena potraviny / average_price – výpočet průměrné roční ceny pro danou kategorii potravin pomocí funkce „average“

5.2 Vytvoření tabulky „Secondary“

Při vytvoření tabulky v_michaela_kosova_project_SQL_secondary_final jsem se opět rozhodla pro použití pohledu („view“), protože pouze potřebuji přehledně uspořádat některá vybraná data z tabulek „countries“ a „economies“ a dále nad tímto pohledem definovat navazující SQL dotazy. Z tohoto pohledu budu čerpat zejména HDP pro Českou republiku v letech 2006-2018 pro odpověď na otázku č.5.

Do pohledu jsem zahrnula následující sloupce (viz soubor v_michaela_kosova_project_SQL_secondary_final.sql):

- Země / country
- Počet obyvatel / population
- Rok / year
- HDP / GDP
- GINI koeficient

5.3 Výzkumná otázka č. 1: Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?

Pro odpověď na tuto otázku jsem formulovala SQL dotaz uložený v souboru michaela_kosova_SQL1.sql.

Cílem bylo zjistit, zda v některých odvětvích a letech došlo i k poklesu průměrné hrubé mzdy mezi lety. K tomuto účelu jsem z pohledu v_michaela_kosova_project_SQL_primary_final nechala pro každé odvětví a rok vypsat meziroční změnu v hrubých mzdách oproti roku přechozímu (wage_growth_percent) a podmínkou WHERE jsem výpis omezila na taková odvětví a roky, ve kterých byla meziroční změna menší než 0, tj. došlo k poklesu průměrných hrubých mezd. Pro výpočet meziroční změny ve mzdách jsem použila vzorec:

(mzda v roce Y+1 – mzda v roce Y) / mzda v roce Y *100, zaokrouhleno na dvě desetinná místa.

Z tohoto dotazu mě vyšla níže zmíněná odvětví a roky, ve kterých došlo k meziročnímu poklesu mezd.

123 year	ABC industry_branch	123 wage_growth_percent
2,013	Administrativní a podpůrné činnosti	-0.36
2,013	Činnosti v oblasti nemovitostí	-1.7
2,013	Informační a komunikační činnosti	-1.01
2,011	Kulturní, zábavní a rekreační činnosti	-0.05
2,013	Kulturní, zábavní a rekreační činnosti	-1.37
2,013	Peněžnictví a pojišťovnictví	-8.91
2,010	Profesní, vědecké a technické činnosti	-0.61
2,013	Profesní, vědecké a technické činnosti	-2.91
2,013	Stavebnictví	-2.13
2,009	Těžba a dobývání	-3.74
2,013	Těžba a dobývání	-2.85
2,014	Těžba a dobývání	-0.79
2,016	Těžba a dobývání	-0.59
2,009	Ubytování, stravování a pohostinství	-1.2
2,011	Ubytování, stravování a pohostinství	-1.11
2,013	Velkoobchod a maloobchod; opravy a údržba motorových vozidel	-0.94
2,010	Veřejná správa a obrana; povinné sociální zabezpečení	-0.33
2,011	Veřejná správa a obrana; povinné sociální zabezpečení	-2.24
2,013	Výroba a rozvod elektřiny, plynu, tepla a klimatiz. vzduchu	-4.37
2,015	Výroba a rozvod elektřiny, plynu, tepla a klimatiz. vzduchu	-1.31
2,010	Vzdělávání	-1.84
2,013	Zásobování vodou; činnosti související s odpady a sanacemi	-0.38
2,009	Zemědělství, lesnictví, rybářství	-0.62

Odpověď na otázku:

V průběhu let mzdy ve výše zmíněných odvětvích a letech klesají, tj. mzdy ve všech odvětvích v průběhu let jen nerostou, ale i klesají.

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

Pro odpověď na tuto otázku jsem formulovala SQL dotazy uložené v souboru michaela_kosova_SQL2.sql.

V tomto bodě pracuji se zjednodušujícím předpokladem, že průměrná hrubá mzda se rovná průměrné čisté mzdě, tj. neberu v úvahu daně, odvody na sociální a zdravotní pojištění a případné další srážky ze mzdy.

Dotazem "SELECT DISTINCT YEAR FROM v_michaela_kosova_project_sql_primary_final;" jsem zjistila, že prvním srovnatelným obdobím je rok 2006 a posledním srovnatelným obdobím je rok 2018.

Následně jsem formulovala dva separátní dotazy, jeden pro rok 2006 a druhý pro rok 2018.

Pro každý rok jsem nechala vypsat průměrnou hrubou mzdu a průměrnou cenu mléka a chleba. Vydělením průměrné mzdy průměrnou cenou mléka (chleba) jsem získala litry mléka (kilogramy chleba) zaokrouhlené na dvě desetinná místa, které je možné za průměrnou mzdu koupit.

Odpověď na otázku:

V roce 2006 bylo možné za průměrnou mzdu koupit 1.437,26 litrů mléka a 1.287,47 kilogramů chleba.

123 average_wage_2006	123 average_price_milk_2006	123 average_price_bread_2006	123 l_milk_per_wage_2006	123 kg_bread_per_wage_2006
20,754	14.44	16.12	1,437.26	1,287.47

V roce 2018 bylo možné za průměrnou mzdu koupit 1.641,57 litrů mléka a 1.342,24 kilogramů chleba.

123 average_wage_2018	123 average_price_milk_2018	123 average_price_bread_2018	123 l_milk_per_wage_2018	123 kg_bread_per_wage_2018
32,536	19.82	24.24	1,641.57	1,342.24

5.5 Výzkumná otázka č. 3: Která kategorie potravin zdražuje nejpomaleji (je u ní nejnižší percentuální meziroční nárůst)?

Pro odpověď na tuto otázku jsem formulovala SQL dotaz uložený v souboru michaela_kosova_SQL3.sql.

Pro každou kategorii potravin jsem vypočítala meziroční změnu ceny, ze které jsem nechala spočítat průměr za celé sledované období let 2006-2018 zaokrouhlený na dvě desetinná místa. Výsledky jsem nechala srovnat od nejmenšího po největší, tj. od nejnižšího průměrného procentuálního meziročního nárůstu po nejvyšší za celé sledované období.

food_category	average_price_growth_2006_2018
Cukr krystalový	-1.92
Rajská jablka červená kulatá	-0.74
Banány žluté	0.81
Vepřová pečeně s kostí	0.99
Přírodní minerální voda uhličitá	1.02
Šunkový salám	1.85
Jablka konzumní	2.01
Pečivo pšeničné bílé	2.2
Hovězí maso zadní bez kosti	2.53
Kapr živý	2.6
Jakostní víno bílé	2.7
Pivo výčepní, světlé, lahvové	2.85
Eidamská cihla	2.92
Mléko polotučné pasterované	2.98
Rostlinný roztíratelný tuk	3.23
Kuřata kuchaná celá	3.38
Pomeranče	3.6
Jogurt bílý netučný	3.95
Chléb konzumní kmínový	3.97
Konzumní brambory	4.18
Rýže loupaná dlouhozrnná	5
Mrkev	5.24
Pšeničná mouka hladká	5.24
Těstoviny vaječné	5.26
Vejce slepičí čerstvá	5.55
Máslo	6.67
Papriky	7.29

Odpověď na otázku:

Nejpomaleji zdražuje Cukr krystalový, jehož průměrná meziroční změna ceny je -1,92 % za celé sledované období let 2006-2018, tj. v průměru ročně zlevnil o 1,92 %.

5.6 Výzkumná otázka č. 4: Existuje rok, ve kterém byl meziroční nárůst cen potravin výrazně vyšší než růst mezd (větší než 10 %)?

Pro odpověď na tuto otázku jsem formulovala SQL dotazy uložené v souboru michaela_kosova_SQL4.sql.

V SQL dotazu jsem spojila dvě tabulky, první tabulka počítá průměrný meziroční nárůst hrubých mezd bez ohledu na odvětví, druhá tabulka počítá průměrný meziroční nárůst cen potravin bez ohledu na kategorii potravin.

Průměrný meziroční nárůst mezd a cen potravin je počítán dle vzorce:

(mzda v roce Y+1 – mzda v roce Y) / mzda v roce Y *100, zaokrouhleno na dvě desetinná místa.
(cena potravin v roce Y+1 – cena potravin v roce Y) / potravin v roce Y *100, zaokrouhleno na dvě desetinná místa.

První SQL dotaz zobrazuje prázdné řádky, protože je zde definována podmínka, že rozdíl mezi průměrným meziročním nárůstem cen potravin a mezd musí být v daném roce větší než 10 %. Žádný takový rok neexistuje.

123 year	123 average_wage_growth_percent	123 year_food	123 average_price_growth_percent

Pro kontrolu jsem definovala druhý SQL dotaz, který zobrazuje roky, ve kterých byl meziroční nárůst cen potravin vyšší než růst mezd, ale rozdíl mezi nimi byl vždy nižší než 10 %.

123 year	123 average_wage_growth_percent	123 year_food	123 average_price_growth_percent
2,007	6.91	2,007	9.26
2,008	7.24	2,008	8.92
2,011	2.24	2,011	4.84
2,012	2.72	2,012	7.47
2,013	-0.78	2,013	6.01
2,017	6.57	2,017	7.06

Odpověď na otázku:

Neexistuje rok, ve kterém byl meziroční nárůst cen potravin výrazně vyšší než růst mezd, tj. rozdíl mezi nimi větší než 10 %. Pouze v letech 2007, 2008, 2011, 2012, 2013 a 2017 byl nárůst cen vyšší než nárůst mezd, avšak vždy menší než 10 %.

5.7 Výzkumná otázka č. 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?

Pro odpověď na tuto otázku jsem formulovala SQL dotaz uložený v souboru michaela_kosova_SQL5.sql.

V SQL dotazu jsem spojila tři tabulky, první tabulka počítá průměrný meziroční nárůst hrubých mezd bez ohledu na odvětví, druhá tabulka počítá průměrný meziroční nárůst cen potravin bez ohledu na kategorii potravin. Obě tyto tabulky vychází z pohledu v_michaela_kosova_project_SQL_primary_final. Třetí připojená tabulka vychází z pohledu v_michaela_kosova_project_SQL_secondary_final a počítá meziroční nárůst HDP v České republice.

Výstupem je následující tabulka porovnávající meziroční nárůst hrubých mezd, cen potravin a HDP v ČR.

year	average_wage_growth_percent	average_price_growth_percent	GDP_growth_percent
2,007	6.91	9.26	5.57
2,008	7.24	8.92	2.69
2,009	2.97	-6.59	-4.66
2,010	2.17	1.52	2.43
2,011	2.24	4.84	1.76
2,012	2.72	7.47	-0.79
2,013	-0.78	6.01	-0.05
2,014	2.52	-0.62	2.26
2,015	2.84	-0.69	5.39
2,016	3.95	-1.4	2.54
2,017	6.57	7.06	5.17
2,018	7.78	2.41	3.2

Odpověď na otázku:

Z dostupných dat nelze jednoznačně prokázat, zda se výraznější nárůst HDP projeví výraznějším růstem mezd a cen potravin. Toto platí např. pro roky 2007 a 2017, kdy byl růst HDP 5,57 %, resp. 5,17 %, nárůst mezd byl 6,91 %, resp. 6,57 % a nárůst cen potravin 9,26 %, resp. 7,06 %. Daný závěr však neplatí pro rok 2015, kdy byl růst HDP 5,39 %, tj. zhruba ve stejné výši jako v letech 2007 a 2017, ale růst mezd byl pouze 2,84 % a u cen došlo dokonce k poklesu o 0,69 %. Ani efekt HDP na mzdy a ceny v následujícím roce nelze z dostupných dat jednoznačně potvrdit.

Jednoznačnou korelaci mezi růstem HDP a růstem mezd a cen potravin není možné konstatovat.

6 Závěr

V tomto dokumentu jsem se pokusila pomocí definice sady SQL dotazů nalézt odpovědi na definované výzkumné otázky v rámci SQL projektu Datové Akademie ENGETO. Odpovědi na některé otázky bylo možné z dostupných dat získat jednoznačně, na jiné otázky jednoznačně odpovědět nebylo možné.