ENGETO

SQL PROJEKT

Zpracovala: Michaela Kosová

[michaela.kosova@gmail.com](mailto:michaela.kosova@gmail.com)

Srpen 2023

Obsah

[1 Úvod 3](#_Toc141738999)

[1.1 Účel dokumentu 3](#_Toc141739000)

[2 Zadání projektu 3](#_Toc141739001)

[2.1 Úvod do projektu 3](#_Toc141739002)

[2.2 Primární tabulky 3](#_Toc141739003)

[2.3 Číselníky sdílených informací o ČR: 4](#_Toc141739004)

[2.4 Dodatečné tabulky: 4](#_Toc141739005)

[2.5 Výzkumné otázky 4](#_Toc141739006)

[2.6 Výstup projektu 4](#_Toc141739007)

[3 Použitý software 5](#_Toc141739008)

[4 Postup zpracování 5](#_Toc141739009)

[5 Odpovědi na výzkumné otázky 5](#_Toc141739010)

[5.1 Vytvoření tabulky „Primary“ 5](#_Toc141739011)

[5.2 Vytvoření tabulky „Secondary“ 6](#_Toc141739012)

[5.3 Výzkumná otázka č. 1: Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají? 6](#_Toc141739013)

[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? 7](#_Toc141739014)

[5.5 Výzkumná otázka č. 3: Která kategorie potravin zdražuje nejpomaleji (je u ní nejnižší percentuální meziroční nárůst)? 8](#_Toc141739015)

[5.6 Výzkumná otázka č. 4: Existuje rok, ve kterém byl meziroční nárůst cen potravin výrazně vyšší než růst mezd (větší než 10 %)? 9](#_Toc141739016)

[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? 10](#_Toc141739017)

[6 Závěr 11](#_Toc141739018)

# Úvod

## Účel dokumentu

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

# Zadání projektu[[1]](#footnote-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**

## 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.
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.

## 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.

## 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.

## 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 násdujícím roce výraznějším růstem?

## 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.**

# Použitý software

Projekt je zpracován nad lokálně staženou databází poskytnutou společností ENGETO. Pro přístup k databázi byl použit program DBeaver.

# 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 jednolivých kategorií potravin. Data jsou sjednocená dle totožných srovnatelných období, společných let 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.

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

# Odpovědi na výzkumné otázky

## Vytvoření tabulky „Primary“

Při vytvoření tabulky v\_michaela\_kosova\_project\_sql\_primary\_final jsem se rozhodla pro použití pohledu („view“), protože pouze potřebuji přehledně uložit pohled na některá vybraná data z primárních tabulek a dále nad 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“

## 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 uložit pohled na některá vybraná data z tabulek „countries“ a „economies“ a dále nad 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

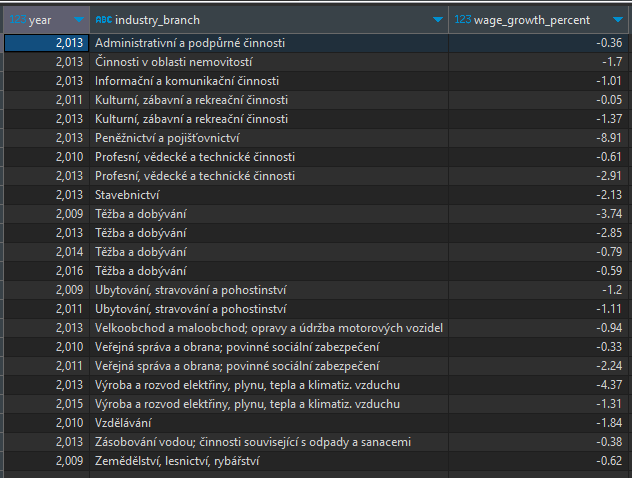
## Výzkumná otázka č. 1: Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?

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. 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.



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.

## 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 dotaz uložený v souboru michaela\_kosova\_SQL2.sql.

V odpovědi na tuto otázku pracuji se zjednodušujícím předpokladem, že průměrná hrubá mzda se rovná průměrné čisté mzdě, tj. neberu v úvahu daně a odvody na sociální a zdravotní pojištění.

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, průměrnou cenu mléka a chleba a 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.



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



## 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 percentuálního meziročního nárůstu po nejvyšší za celé sledované období.



Odpověď na otázku:

Nejpomaleji zdražuje Cukr krystalový, jehož průměrná meziroční změna ceny je -1,92 %, tj. v průměru ročně zlevnil o 1,92 %.

## 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.

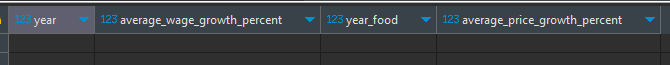
V SQL dotazu jsem spojila dvě tabulky, první tabulka počítá průměrný meziroční nárůst hrubých mezd bez ohledu na odvětví, druhá tabulka počítá průměrný meziroční nárůst cen bez ohledu na kategorii potravin.

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

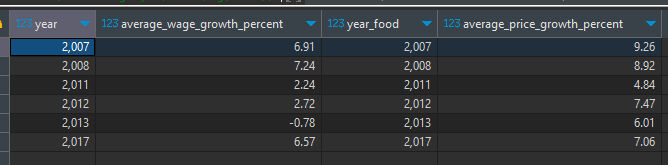
(mzda v roce Y+1 – mzda v roce Y) / mzda v roce Y \*100, zaokrouhleno na dvě desetinná místa.

(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.



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 %.



Odpověď na otázku:

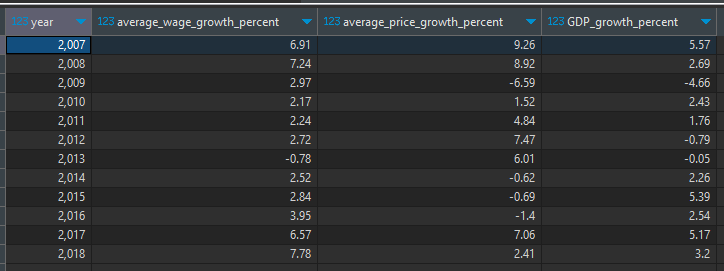
Neexistuje rok, ve kterém byl meziroční nárůst cen potravin výrazně vyšší než růst mezd, tj. rozdíl mezi nimi větší než 10 %.

## 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 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.



Odpověď na otázku:

Z dostupných dat nelze jednoznačně prokázat, zda se výraznější nárůst HDP projeví výraznějším růstem mezd a cen. Toto platí např. pro roky 2007 a 2017, kdy byl růst HDP 5,57 %, resp. 5,17 %, nárůst mez 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.

Korelaci mezi růstem HDP a růstem mezd a cen není možné konstatovat.

# Závěr

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

1. Přezvato 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 [↑](#footnote-ref-1)