Engeto s.r.o

Datová akademie 6/2023

Projekt SQL

(průvodní listina)

2023 Bc. Veronika Svitáková

Obsah

[1. Zadání projektu 3](#_Toc143286713)

[2. Výstup 5](#_Toc143286714)

[Výzkumná otázka č. 1: 5](#_Toc143286715)

[Výzkumná otázka č. 2: 5](#_Toc143286716)

[Výzkumná otázka č. 3: 5](#_Toc143286717)

[Výzkumná otázka č. 4 5](#_Toc143286718)

[Výzkumná otázka č. 5 5](#_Toc143286719)

[Zdroje 5](#_Toc143286720)

[Přílohy 6](#_Toc143286721)

# Zadání projektu

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:

czechia\_payroll – Informace o mzdách v různých odvětvích za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.

czechia\_payroll\_calculation – Číselník kalkulací v tabulce mezd.

czechia\_payroll\_industry\_branch – Číselník odvětví v tabulce mezd.

czechia\_payroll\_unit – Číselník jednotek hodnot v tabulce mezd.

czechia\_payroll\_value\_type – Číselník typů hodnot v tabulce mezd.

czechia\_price – Informace o cenách vybraných potravin za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.

czechia\_price\_category – Číselník kategorií potravin, které se vyskytují v našem přehledu.

Číselníky sdílených informací o ČR:

czechia\_region – Číselník krajů České republiky dle normy CZ-NUTS 2.

czechia\_district – Číselník okresů České republiky dle normy LAU.

Dodatečné tabulky:

countries - Všemožné informace o zemích na světě, například hlavní město, měna, národní jídlo nebo průměrná výška populace.

economies - HDP, GINI, daňová zátěž, atd. pro daný stát a rok.

Výzkumné otázky

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

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

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

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

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?

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.

# Výstup

Na základě zadaného projektu byla vytvořena finální tabulka s potřebnými daty. Pro vytvoření této tabulky byly použity dvě dílčí tabulky. První z nich pracuje s daty spojenými s hrubým měsíčním příjmem pracovníků v jednotlivých odvětvích. Druhá je založena na cenách jednotlivých potravin v určitém čase. Následně byly obě dílčí tabulky spojeny pomocí funkce JOIN do jedné finální tabulky.

## Výzkumná otázka č. 1:

Dle získaných dat dochází obecně v průběhu let k růstu mezd. Nicméně se vyskytují v některých odvětvích roky, kde hrubá měsíční mzda meziročně klesla. Takovým případem je například meziroční pokles v roce 2009 a 2018 v odvětví *Zemědělství, lesnictví a rybářství*. Dále došlo k poklesu v odvětví zabývající se těžkou a dobýváním v roce 2009 a 2013. Souhrnně lze říci, že ve všech odvětvích došlo v letech 2006-2018 k nárůstu hrubé měsíční mzdy s několika málo lety, kdy došlo výjimečně k poklesu. Více porovnání viz dodatečně vytvořená tabulka *t\_yearly\_payroll\_increase* (zdroj: SQL script: research\_question\_1)

## Výzkumná otázka č. 2:

Data k zodpovězení další výzkumné otázky ukazují na rozdílné množství mléka a chleba v různých odvětvích. Nejprve bylo zjištěn *category\_code* pro tyto dvě potraviny a následně i určeny klíčové období/roky. Na základě průměrné ceny potravin a hodnoty průměrné hrubé mzdy bylo spočítáno množství (*amount*) jednotlivých potravin v letech 2006 a 2018. Získaná data poukazují na fakt, že odvětví *Informační a komunikační činnosti* disponuje v roce 2018 nejvyšší průměrnou hrubou mzdou, a tudíž je možné si koupit největší množství potravin.

## Výzkumná otázka č. 3:

Třetí výzkumná otázka byla řešena pomocí dotazu, který sloučil dvě totožné tabulky a následně byla vytvořena pomocná tabulka k získání dat k meziročnímu nárůstu cen potravin. Opět byla využita pomocná tabulka k zjednodušení dotazu a rychlejšímu provedení. Z výsledků dotazu je vidět, že v roce 2009 došlo u kategorie *Pečivo bílé pšeničné* k nejmenšímu meziročnímu nárůstu, respektive poklesu ceny (-31 %).

## Výzkumná otázka č. 4:

Výrazně větší nárůst cen na úkor hrubé mzdy v různých odvětvích se vyskytuje v každém sledovaném roce minimálně v jednom odvětví. Dle zadaného dotazu ukazují výsledky na fakt, že průměrná hrubá mzda roste v zásadě pomaleji než průměrné ceny potravin. Díky Case Expression vznikl dodatečný sloupec, který ukazuje rozdíl mezi nárůstem cen a mezd. Pro bližší představu lze zmínit příklad ceny paprik v roce 2007. V tomto roce došlo k enormnímu nárustu ceny potraviny (+ 89 %). Naproti tomu nárůst hrubé mzdy ve většině odvětví se v roce 2007 pohyboval pod 10 %.

## Výzkumná otázka č. 5

Vliv HDP na růst cen potravin a průměrných hrubých mezd lze zkoumat na základě porovnání meziročního růstu HDP a meziročního růstu průměrných hrubých mezd a cen konkrétních potravin. Pro dosažená výsledků byla vytvořena další pomocná tabulka ve spojitosti s HDP (meziroční růst HDP). Následně vznikl dataset porovnávající tyto faktory v jednotlivých letech a v konkrétních odvětvích/kategoriích. Pokud by došlo následujícím dotazem k zprůměrování meziročního nárůstu mezd ze všech odvětví pro dané roky 2007-2018, lze dostat nepříliš validní výsledek. Lepším způsobem by bylo porovnávat první dataset, který bere ohled na jednotlivé kategorie mezd i potravin. Nicméně po srovnání výsledků pro jednotlivé roků lze vypozorovat, že v roce 2009 došlo k poklesu HDP a stejně tak i k meziročnímu poklesu cen potravin. Následně dochází k nárůstu v roce 2010 v rámci HDP i průměrného meziročního růstu cen potravin. Významným milníkem je pak rok 2012, kdy dochází k poklesu HDP, a naopak většímu procentuálnímu meziročnímu růstu cen. Rozdíl mezi těmito hodnotami přesahuje 8 %. Do tří let se naopak situace obrátila a meziroční růst HDP výrazně vzrostl na úkor meziročnímu růstu cen.

Pro názorné srovnání je nutné doporučit vložení dat do grafu.

# Zdroje

Seznam použitých zdrojů:

Zadání projektu, Engeto s.r.o., <https://learn.engeto.com/cs/>, 14.8.2023

Zdroj dat, tabulky z John Hopkins University, dostupné zde: <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data>

# Přílohy

1. SQL script: Project\_part\_1 (finální tabulka)
2. SQL scripts: Research\_question\_1-5