ENGETO – Datová akademie

Projekt z SQL – průvodní list

Vypracoval: Jiří Broukal

Září - Prosinec 2023

Obsah: Strana

[Výzkumné otázky - 2 -](#_Toc155037077)

[Poznámky k datovým zdrojům - 2 -](#_Toc155037078)

[Odpověď na 1. otázku - 3 -](#_Toc155037079)

[Odpověď na 2. otázku - 3 -](#_Toc155037080)

[Odpověď na 3. otázku - 4 -](#_Toc155037081)

[Odpověď na 4. otázku - 4 -](#_Toc155037082)

[Odpověď na 5. otázku - 5 -](#_Toc155037083)

# 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?

# Poznámky k datovým zdrojům

Data pro zodpovězení otázek byla čerpána z poskytnuté datové sady. V tomto materiálu jsou dále uvedeny popisy, jak s daty bylo zacházeno i komentáře k nim.

Pro zodpovězení otázek 1 – 4 byla byla z datové sady vytvořena tabulka

*t\_jiri\_broukal\_project\_sql\_primary\_final* (dále jen *tjb*),

ze které jsou čerpána data pro odpovědi pomocí SQL scriptů.

V *tjb* jsou data jak o průměrných mzdách, počtu zaměstnanců, tak i o cenách potravin (rozdělena jsou hodnotou sloupce *dataset*. Tato data (pokud ve zdroji existovala), jsou již časově zprůměrována, časovou jednotkou je kalendářní rok.

Do *tjb* jsou data mezd a počtu zaměstnanců přenesena ze zdroje jako „přepočtená“. (Calculation code = 200, lépe odpovídá účelu posuzování.)

Data o počtu zaměstnanců jsou velice neúplná, při vytváření dat pro odpovědi bylo nutno přistoupit k velikým zjednodušením (viz. dále).

Pokud by datová sada byla větší, bylo by lepší neduplikovat v *tjb* některá data v jednotlivých záznamech (názvy odvětví, jméno kategorie potravin) a tím zvětšovat velikost tabulky. Řešením by bylo např. vytvoření a použití další tabulky, která by se ve scriptech připojovala.

Pro zodpovězení otázky 5 byla byla z poskytnuté datové sady vytvořena tabulka

*t\_jiri\_broukal\_project\_sql\_secondary\_final* (dále jen *tjb\_sec*),

ze které jsou čerpána data pro odpověď pomocí SQL scriptu.

V *tjb\_sec* jsou jak potřebná data pro zodpovězení otázky 5 (tedy data týkající se České republiky), tak i data ostatních států Evropy či data o oblastech, kde Evropa (či její část) nějak figuruje.

Druh dat je v *tjb\_sec* rozdělen hodnotou sloupce *dataset* a má následující významy:

* *dataset* = ‘country‘ – hodnoty pro jednotlivé státy v evropě (zdroj dat – tabulka countries)
* *dataset* = ‘country\_eco‘ – hodnoty pro jednotlivé státy v evropě (zdroj dat – tabulka economies)
* *dataset* = ‘area‘ – hodnoty pro různé oblasti (zdroj dat – tabulka economies)

# Odpověď na 1. otázku

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

**Odpověď: V některých odvětvích a letech klesají.**

Odpověď je v datech poskytnutých SQL scriptem *Otazka\_1*.

Odpověď určuje sloupec *grow\_payroll\_perc*, což je procentní poměr růstu průměrné mzdy v odvětví vůči předchozímu roku.

Nejčastěji mzdy napříč sledovanými odvětvími poklesly v roce 2013.

Odpověď by bylo dobré prezentovat grafem (čárovým či sloupcovým) průměrných mezd v jednotlivých letech podle odvětví vytvořeným např. v MS Excelu nebo Power BI.

# Odpověď na 2. otázku

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

**Odpověď:**

**V roce 2006 (první období s daty cen) bylo možné za průměrnou mzdu zakoupit 1353 l mléka nebo 1212 kg chleba.**

**V roce 2018 (poslední období s daty cen) bylo možné za průměrnou mzdu zakoupit 1617 l mléka nebo 1322 kg chleba.**

Odpověď je v datech poskytnutých SQL scriptem *Otazka\_2*.

Při hledání odpovědi bylo nutno z důvodu nedostatku dat o počtech zaměstnanců provést značné zjednodušení. Nebylo možné provést celkový vážený průměr mezd přes odvětví. Jako podklad pro výpočet byla použita průměrná mzda ze záznamů bez uvedeného odvětví, protože v záznamech, které jsou k dispozici a jsou v nich uvedeny počty zaměstnanců, jsou počty zaměstnanců bez uvedeného odvětví nejvyšší (a měly by v případně dostatku dat nejvyšší váhu).

# Odpověď na 3. otázku

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

**Odpověď:**

**Nejpomaleji zdražuje Cukr krystalový. Jeho cena v posuzovaném období (2006-2018) dokonce poklesla (v průměru o 2,3 % za rok).**

Odpověď je v datech poskytnutých SQL scriptem *Otazka\_3*.

Odpověď určuje sloupec *grow\_price\_perc*, což je průměrný roční procentní nárůst ceny dané kategorie potravin (určený z cen za roky 2006 a 2018).

# Odpověď na 4. otázku

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

**Odpověď:**

**Ne, takový rok nebyl.**

**Maximální rozdíl mezi nárůstem cen potravin a mezd byl v roce 2013, kdy činil 5,2%.**

**Naopak v roce 2009 byl nárůst mezd cca o 10% vyšší než nárůst cen (ceny totiž v tomto roce dokonce poklesly).**

Odpověď je v datech poskytnutých SQL scriptem *Otazka\_4*.

Odpověď určuje sloupec *diff\_grow\_price\_grow\_payroll*, což je rozdíl procentního růstu cen a procentního růstu mezd.

Stejně jako při hledání odpovědi na otázku 2 bylo i zde nutno z důvodu nedostatku dat provést značné zjednodušení:

* Nebylo možné provést celkový vážený průměr mezd přes odvětví. Jako podklad pro výpočet byla použita průměrná mzda ze záznamů bez uvedeného odvětví, protože v záznamech, které jsou k dispozici a jsou v nich uvedeny počty zaměstnanců, jsou počty zaměstnanců bez uvedeného odvětví nejvyšší (a měly by v případně dostatku dat nejvyšší váhu).
* Obdobně nebylo možné provést vážený průměr cen v jednotlivých letech. Byl použit normální průměr. Lepší vypovídací schopnost by měl výpočet průměrných cen např. za použití nějaké varianty „spotřebního koše“.

Odpověď by bylo dobré prezentovat grafem (např. sloupcovým) nárůstu mezd, nárůstu cen a rozdílu těchto hodnot v jednotlivých letech např. v MS Excelu nebo Power BI.

# Odpověď na 5. otázku

Otázka 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?

Odpověď je v datech poskytnutých SQL scriptem *Otazka\_5*.

**Odpověď:**

**Určitou drobnou závislost lze vidět mezi růstem HDP a růstem mezd. Mzdy na HDP sice reagují (spíše s ročním zpožděním), ale ne v silné závislosti. Téměř po celé vyhodnocované období však rostou rychleji než HDP.**

**Závislost mezi růstem HDP a růstem cen objevena nebyla.**

Odpověď určují sloupce

* *grow\_GDP\_perc*, což je procentní růst HDP
* *grow\_payroll\_perc*, což je procentní růst mezd
* *lead\_payroll\_perc*, což je procentní růst mezd v následujícím roce
* *grow\_price\_perc*, což je procentní růst cen
* *lead\_price\_perc*, což je procentní růst cen v následujícím roce

Stejně jako při hledání odpovědi na otázku 2 a 4 bylo i zde nutno z důvodu nedostatku dat provést značné zjednodušení:

* Nebylo možné provést celkový vážený průměr mezd přes odvětví. Jako podklad pro výpočet byla použita průměrná mzda ze záznamů bez uvedeného odvětví, protože v záznamech, které jsou k dispozici a jsou v nich uvedeny počty zaměstnanců, jsou počty zaměstnanců bez uvedeného odvětví nejvyšší (a měly by v případně dostatku dat nejvyšší váhu).
* Obdobně nebylo možné provést vážený průměr cen v jednotlivých letech. Byl použit normální průměr. Lepší vypovídací schopnost by měl výpočet průměrných cen např. za použití nějaké varianty „spotřebního koše“.

Velice jednoduchý graf z dat výsledku SQL scriptu:

