

# Průvodní listina projektu SQL

## Zadání projektu

Projekt je zaměřený na zmapování dostupnosti základních potravin široké veřejnosti. K dispozici máme tyto informace:

- Informace o mzdách v různých odvětvích za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR. K této datové sadě máme další pomocné tabulky vysvětlující data v tabulce mezd.
- Informace o cenách vybraných potravin za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.
- Číselníky krajů a okresů České republiky dle norem CZ-NUTS 2 a LAU.
- Informace o dalších zemích světa - populaci, HDP, měně, daňové zátěži, aj.

Jsou definovány následující 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 procentuální 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?

## Spojení tabulek

**Pro zodpovězení otázek vytvoříme dvě tabulky:**

**Primární tabulka:** t\_iva\_dvorakova\_project\_SQL\_primary\_final

**Sekundární tabulka:** t\_iva\_dvorakova\_project\_SQL\_secondary\_final

### t\_iva\_dvorakova\_project\_SQL\_primary\_final

Nejprve jsem se rozhodla vytvořit dvě tabulky – jednu pro platy, druhou pro ceny. Ty následně spojit do jedné finální tabulky, obsahující údaje zároveň o platech i cenách.

**Tabulka pro platy:** využiji tyto sloupce z tabulky czechia\_payrol:

- Value
- Value\_type\_code – použiji pouze hodnoty 5958 = Průměrná hrubá mzda na zaměstnance. Druhá hodnota – 316 = průměrný počet zaměstnaných osob – by pro naše zkoumání neměla být k užitku. V nové tabulce tato hodnota není potřeba, použiju ji pouze pro filtrování vstupních dat podmínkou WHERE.
- Unit\_code = vždy by měl být 200 = hodnota Kč, můžu použít pro kontrolu, zda mám všechna data správná. Ale tento sloupec by mohl být z nové tabulky vynechán. V nové tabulce tato hodnota není potřeba, použiju ji pouze pro filtrování vstupních dat podmínkou WHERE.
- Calculation\_code = budu pracovat s hodnotami 200 = přepočtená hrubá mzda. Považuji to za více vypovídající údaj. V nové tabulce tato hodnota není potřeba, použiju ji pouze pro filtrování vstupních dat podmínkou WHERE.
- Industry\_branch\_code
- Payroll\_year
- Payroll\_quarter – tento údaj ve sloučené tabulce vynechám a budu zkoumat pouze celé roky

czechia\_payroll\_calculation: dává nám informaci o tom, jaké hodnoty pro novou tabulku využít, ale do nové tabulky ji zahrnovat nebudu

czechia\_payroll\_industry\_branch: bude připojena k nové tabulce, přes sloupec code ji připojím k tabulce czechia\_payroll

czechia\_payroll\_unit: dává nám informaci o tom, jaké hodnoty pro novou tabulku využít, ale do nové tabulky ji zahrnovat nebudu

czechia\_payroll\_value\_type: dává nám informaci o tom, které hodnoty pro novou tabulku využít, ale do nové tabulky ji zahrnovat nebudu

**TABULKA pro ceny:** využiji tyhle sloupce z czechia\_price:

- Value
- Category code
- Date\_from – vytvořím si nový sloupec, očištěný o zbytek data, abych dostala formát YYYY-MM-DD
- Region\_code – vynechám z nové tabulky, v otázkách nikde není dotaz na dělení podle okresů/krajů

czechia\_price\_category: tabulka nutná pro další výzkum - přes category code ji připojím k tabulce czech\_price, abych měla vedle sebe hodnoty pro ceny potravin včetně slovního popisku

Do jedné tabulky je třeba přidat údaje o HDP v České republice v daném roce. Přidala jsem do tabulky pro ceny potravin.

V obou tabulkách jsem jednotlivé údaje o mzdách a cenách spojila do jedné informace o průměrné ceně/mzdě – v daném roce, odvětví, artiklu spotřebního koše.

Pokud jsem tento údaj do první tabulky nedopočítala, nebyla jsem schopná řešit následující výzkumné otázky bez dalších pomocných tabulek a výpočtů.

Podle zadání by jednotlivé hodnoty platů a cen neměly být nikde dále potřeba, proto budu už vycházet z průměrných hodnot.

Následně jsem spojila obě tabulky dohromady a vznikla tabulka, ze které věřím, že bude možné informace získat.

TABULKA pro ceny + informace o HDP: `t_iva_dvorakova_price_table`

TABULKA pro platy: `t_iva_dvorakova_payroll_table`

Výsledná tabulka: `t_iva_dvorakova_project_SQL_primary_final`

Původně jsem, inspirována při konzultaci o projektu, chtěla vytvořit 2 VIEW a ty spojit do tabulky.

Napoprvé se mi to podařilo, ale zjistila jsem, že nikde nemám údaje o HDP.

Připojila jsem tento údaj do VIEW o cenách a následně stejně spojila, ale tento příkaz mi ani po 10 minutách žádanou tabulku nevygeneroval. Už jsem nepřišla na nic, co by šlo osekát, agregovat, nějak zmenšit. Takže jsem namísto VIEW vytvořila 2 tabulky a spojila do výsledné jedné tabulky. V řádu minut už tato tabulka šla vytvořit.

Z toho důvodu mám 3 tabulky, 2 pomocné, a jednu výslednou.

## **t\_iva\_dvorakova\_project\_SQL\_secondary\_final**

V druhé tabulce jde o informace pro ostatní státy Evropy ve srovnatelném období.

Spojila jsem tabulky countries a economies a zahrnula tyto sloupce:

- Název země „country“
- Populace dané země „population“
- Rok „YEAR“
- HDP „GDP“
- Gini koeficient „gini“

a výběr ohraničila podmínkou WHERE na země v Evropě + data z let 2006 – 2018.

## **Výzkumná otázka č. 1, soubor sql\_project\_question\_1\_2023\_10\_31.sql**

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

Pro zodpovězení této otázky potřebuji znát průměrnou mzdu v daném odvětví pro jednotlivé roky, které zkoumáme. V základní tabulce už mám průměrné hodnoty pro platy v daných letech.

Pro výpočet meziročních změn jsem připravila JOIN dvou stejných těchto tabulek a posunem o 1 rok jsem vypočítala rozdíly mezi jednotlivými lety – vždy dva po sobě následující roky.

Pro snazší orientaci jsem přidala sloupec `increase_decrease_flag`, který nám říká, zda mzdy mezi lety rostly, nebo klesaly.

**Výsledek: V některých odvětvích mzda meziročně klesla, v následujících letech opět rostla. V žádném odvětví mzda neklesala déle než 2 období po sobě. Poté vždy následoval růst.**

## **Výzkumná otázka č. 2, soubor sql\_project\_question\_2\_2023\_10\_31.sql**

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

V prvním kroku jsem si určila funkcí min a max první a poslední rok, který sledujeme.

Dále jsem si přes SELECT DISTINCT určila přesné názvy pro mléko a chléb.

Připravila jsem dvě varianty SQL:

- První, kde počítám průměrnou mzdu v odvětvích v daném roce (v letech 2006 + 2008), přidávám údaj o průměrné ceně hledaných dvou komodit a následně dělením docházím k výsledku, kolik celých jednotek dané komodity si za průměrnou mzdu v daném odvětví ve zkoumaném roce koupíme
- Druhá varianta vynechává rozdělení na odvětví a dává nám jen odpověď na to, kolik bochníků chleba a litrů mléka si pořídíme v daných letech za průměrnou mzdu bez rozdělení na odvětví.

**Výsledek: Průměrně, napříč odvětvími, byla průměrná mzda v roce 2006 21 165 Kč, za kterou si pracovníci mohli koupit 1313 bochníků chleba (průměrná cena 16,12 Kč), nebo 1466 litrů mléka (průměrná cena 14,44 Kč).**

**V roce 2018 byla průměrná mzda za všechna odvětví 33 091 Kč, pracovníci mohli koupit buď 1 365 bochníků chleba s průměrnou cenou 24,24 Kč, nebo 1 670 litrů mléka s průměrnou cenou 19,82 Kč.**

**Za všechna odvětví společně se koupěschopnost zlepšila.**

### **Výzkumná otázka č. 3, soubor sql\_project\_question\_3\_2023\_10\_31.sql**

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

Pro tuto otázku jsem nejprve spojila dvě primární tabulky k sobě, s posunem o 1 rok, a spočítala průměrnou změnu cen mezi danými dvěma lety – údaje pro všechny roky, všechny položky.

V dalším kroku jsem spočítala průměr z těchto ročních průměrů a seřadila výsledky vzestupně.

Z toho mi vychází, že nejnižší meziroční procentuální růst je u položky „Cukr krystalový“, kde se cena dokonce snížila a proto průměrná procentuální roční změna má zápornou hodnotu.

**Výsledek: Nejpomaleji zdražuje cukr krystalový, jeho cena dokonce klesala. Průměrně ročně o 1,92 %.**

#### **Výzkumná otázka č. 4, soubor sql\_project\_question\_4\_2023\_10\_31.sql**

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

Postupně jsem přidávala informace, které potřebuji porovnávat.

Prvním krokem byly průměrná mzda v jednotlivých letech.

Dalším krokem byl JOIN se stejnou tabulkou posunutou o rok, abychom viděli rozdíl mezi lety vyjádřený absolutní hodnotou i procentuální rozdíl.

Pro otázku ale není důležité, o jaké potraviny nebo odvětví průmyslu se jedná, proto třetím krokem bylo srovnání celkové průměrné mzdy za všechna odvětví a jejich meziroční změny. To stejné i pro ceny potravin.

Krok 4 následoval – spojení údajů o změnách mezd + cen + procentuální vyjádření.

V posledním 5.kroku jsem přidala sloupec počítající rozdíl mezi % změnu cen a mezd (diff). Dále také sloupec high\_increase\_flag, který detekuje rozdíl větší než 10%. Nevyšel nikdy.

Pro zajímavost jsem přidala SQL, kde výsledkem jsou tyto změny, ale pro jednotlivá odvětví průmyslu.

**Výsledek: V žádném roce nebyl růst cen potravin o více než 10% větší než růst mezd. Největší kladný rozdíl (ceny potravin rostou nejvíc oproti mzdám) byl mezi lety 2012 a 2013, kdy ceny rostly o 5,1%, ale mzdy klesly o 1,56 %. Výsledný rozdíl byl tedy 6,65%.**

**Na opačné straně spektra stojí roky 2008 a 2009, kdy meziročně ceny klesly o 6,41 %, ale mzdy vzrostly o 3,07 %. Rozdíl byl tedy téměř 10% - 9,49%, ale záporných. Dá se říci, že v tomto období byl růst mezd výrazně vyšší než růst cen. Ale není to právě ten případ, který hledá naše zadání.**

**Výsledku našeho zadání nebylo dosaženo v žádném roce.**



## Výzkumná otázka č. 5, soubor sql\_project\_question\_5\_2023\_10\_31.sql

**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?**

V tomto příkladě je potřeba spojit více sloupců a výpočtů dohromady.

První nápad mi nevyšel, zkusila jsem na sebe napojit 3 \* primární tabulku, ale asi po 15 minutách výsledek nebyl.

Proto jsem se rozhodla vytvořit 2 pomocné tabulky, které následně spojím do jedné a dají nám všechny informace, které potřebujeme.

V tom případě musí v jednom řádku být všechny tyto údaje (musí mít vypočítané tyto sloupce):

- Roční změna HDP
- Roční změna cen v témže roce
- Roční změna mezd v témže roce
- Roční změna cen v následujícím roce
- Roční změna mezd v následujícím roce

+ dopočítané, kde je změna HDP vůči cenám/platům výrazná

Postupovala jsem s pomocnými tabulkami.

Nejdříve t\_GDP\_change, která zobrazuje změnu HDP mezi jednotlivými lety.

Druhá t\_PP\_change\_3\_years, která zobrazuje změny cenách a mzdách mezi 3 lety – rozdíl mezi lety 1 a 2 a mezi lety 2 a 3. Tu jsem připravila ve dvou krocích, opět mi nešlo napoprvé sloučit dvě tabulky, takže vznikla jedna další pomocná tabulka t\_PP\_change\_2\_years.

Spojením t\_GDP\_change + t\_PP\_change\_3\_years mi vzniknul přehled toho, jak se mezi dvěma lety změnilo HDP, jak se ve stejném období změnily ceny a platy a zároveň, jak tomu bylo v roce následujícím.

Pro větší přehlednost porovnám HDP a růsty nadvakrát – nejdříve pro stejné období, potom pro následující období. Přidám tedy sloupce

- overall\_situation\_2x3 – hodnotí změnu cen a mezd ve vztahu ke změně HDP – ve stejném období, jako se měnilo HDP. Jako podmínku pro „overall growth“ zadám růst všech 3 parametrů o 2 a více %. Pokud by byly všechny hodnoty záporné, označíme řádek „overall decrease“.
- overall\_situation\_1x3 – hodnotí změnu cen a mezd ve vztahu ke změně HDP – např. změn HDP mezi lety 2006/2007 vůči změně cen a mezd v letech 2007/2008. Podmínka pro vyhodnocení je stejná - pro „overall growth“ zadám růst všech 3 parametrů o 2 % a více. Pokud by byly všechny hodnoty záporné, označíme řádek „overall decrease“.

Nakonec jsem dotaz spojila do jednoho SQL dotazu. Snažila jsem se minimalizovat názvy sloupců, aby data byla přehlednější.

**Výsledek: Stanovila jsem si sama, bez větší znalosti problematiky, ukazatel, který vyhodnotí, zda ceny, mzdy a HDP rostou nebo klesají „společně“ a ovlivňují se.**

Z mých výsledků vychází, že v letech 2006/2007 byla souvislost mezi růstem HDP/cen a mezd, kdy všechny ukazatele rostly o více než 2%. V roce 2008 ale došlo k poklesu cen a růstu platů, proto jsem závislost identifikovala jen v letech 2006/2007, kdy vzrostlo HDP, ceny i mzdy zároveň.

V letech 2008/2009 došlo k poklesu na všech frontách – HDP/ceny i mzdy. V roce 2010 ale opět platy rostly, ceny dál klesaly.

Našla jsem tedy souvislost pouze ve dvou případech, v ostatních nikoliv. Z toho vyvozuji, že změny cen a mezd ovlivňuje více faktorů, než pouze změna HDP.