
**Projekty SQL**

# Datová Akademie

Běh 16.1.2024 - 2.4.2024


| | |
|---|---|
| **Autor práce:** | Jaroslav Podobský |
| **Rok:** | 2024 |
| **SW1:** | DBeaver 23.3. |
| **SW2:** | MS EXCEL |

# Obsah

**1 Zadání projektu ..... 4**
1.1 Výzkumné otázky ..... 5
1.2 Postup ..... 5
**2 Výstup projektu ..... 6**
2.1 Postup a odpověď na první výzkumnou otázku ..... 7
Postup: ..... 7
Odpověď: ..... 7
2.2 Postup a odpověď na druhou výzkumnou otázku ..... 9
Postup: ..... 9
Odpověď: ..... 9
2.3 Postup a odpověď na třetí výzkumnou otázku ..... 10
Postup: ..... 10
Odpověď: ..... 10
2.4 Postup a odpověď na čtvrtou výzkumnou otázku ..... 12
Postup: ..... 12
Odpověď: ..... 12
2.5 Postup a odpověď na pátou výzkumnou otázku ..... 13
Postup: ..... 13
Odpověď: ..... 13
**Informace o výstupních datech a závěr ..... 15**
**Přílohy ..... 16**

# 1 Zadání 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.

## 1.1 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?

## 1.2 Postup

Nejprve byla vytvořena tabulka t_jaroslav_podobsky_project_SQL_primary_final, z dočasných pohledů -> tabulek (`first_support_document, second_support_document`) tyto tabulky byly následně spojeny pomocí JOIN.

# 2 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.

## 2.1 Postup a odpověď na první výzkumnou otázku

**Postup:**

V souboru 1_otázka.slq jsem vytvořil více dotazů k této otázce:

1) Vytvořena view, z primární tabulky s vybranými sloupci pro následující dotazy.
2) Vytvořen dotaz na tvorbu tabulky pro sledování časových řad. Konkrétně pro sledování růstu všech odvětví v letech 2006 až 2018. Tato tabulka je připravena pro následnou analýzu v Power BI.
3) Výstup tři je určen pro úplnou odpověď na první výzkumnou otázku, kde byl sledován růst během roku 2007, 2018 a poté celkový růst mezi roky 2007 až 2018.

**Odpověď:**

Mezi roky 2007 a 2018 rostly nejvíce průměrné mzdy v odvětví Zdravotní a sociální péče o 67.84 %, Kulturní, zábavní a rekreační činnosti o 64.51 % a Zpracovatelský průmysl o 62%. Přesto, že tato odvětví dosáhla největšího růstu průměrných mezd, nejedná se o odvětví s největší průměrnou mzdou.

V roce 2018 se nejvyšší průměrná mzda dle analyzované databáze nachází v odvětví Informační a komunikační činnosti s hodnotou 58 019 Kč, Peněžnictví a pojišťovnictví 51 769 Kč a Výroba a rozvod elektřiny, plynu, tepla a klimatiz. Vzduchu s 43 429 Kč.

Zajímavostí je, že na druhé příčce nejvyšší průměrné mzdy se v roce 2018 udržuje odvětví Peněžnictví a pojišťovnictví, které mezi roky 2007 a 2018 dosáhlo růstu průměrných mezd pouze o 8,76%.

**Tabulka 1: Odpověď na otázku č.1**

| Název odvětví | mzda_za_rok07 | mzda_za_rok18 | rust_mezd_06_07 | rust_mezd_17_18 | rust_mezd_07_18 |
|---|---|---|---|---|---|
| Zdravotní a sociální péče | 18994 | 31880 | 5.32 | 1.87 | 67.84 |
| Kulturní, zábavní a rekreační činnosti | 16838 | 27701 | 6.91 | 11.01 | 64.51 |
| Zpracovatelský průmysl | 18586 | 30109 | 7.83 | 0.32 | 62.00 |
| Ubytování, stravování a pohostinství | 11640 | 18599 | 5.75 | 5.92 | 59.79 |
| Zemědělství, lesnictví, rybářství | 14221 | 22688 | 10.29 | -15.29 | 59.54 |
| Velkoobchod a maloobchod; opravy a údržba motorových vozidel | 18463 | 28685 | 8.59 | 3.55 | 55.36 |
| Informační a komunikační činnosti | 37536 | 58019 | 6.95 | 12.01 | 54.57 |
| Veřejná správa a obrana; povinné sociální zabezpečení | 23619 | 36374 | 6.89 | 12.31 | 54.00 |
| Profesní, vědecké a technické činnosti | 24666 | 37867 | 10.23 | -3.54 | 53.52 |
| Zásobování vodou; činnosti související s odpady a sanacemi | 18135 | 27724 | 5.32 | -7.02 | 52.88 |
| Administrativní a podpůrné činnosti | 14391 | 21691 | 6.24 | 14.91 | 50.73 |
| Výroba a rozvod elektřiny, plynu, tepla a klimatiz. vzduchu | 28907 | 43429 | 7.81 | -7.26 | 50.24 |
| Stavebnictví | 17459 | 25796 | 11.58 | -1.48 | 47.75 |
| Vzdělávání | 19320 | 28512 | 6.29 | 12.89 | 47.58 |
| Těžba a dobývání | 23225 | 34173 | 8.84 | 13.18 | 47.14 |
| Doprava a skladování | 19339 | 27968 | 7.68 | 3.15 | 44.62 |
| Ostatní činnosti | 16452 | 23462 | 6.04 | 5.30 | 42.61 |
| Činnosti v oblasti nemovitostí | 19909 | 27500 | 8.44 | -0.78 | 38.13 |
| Peněžnictví a pojišťovnictví | 47600 | 51769 | 7.70 | 5.81 | 8.76 |

Zdroj: Databáze společnosti Engeto na základě veřejných dat

## 2.2 Postup a odpověď na druhou výzkumnou otázku

**Postup:**

1) Pomocí *with* jsem vytvořil dočasnou tabulku, kde byla dopočtena průměrná mzda a cena sledovaných komodit (Chléb konzumní kmínový, Mléko polotučné pasterované) To vše čerpané z primární tabulky.
2) Výsledná tabulka zobrazuje roky, název produktu a maximální počty jednotek (kg, l) každé komodity, které lze nakoupit za průměrnou mzdu v těchto letech 2006 a 2018.

**Odpověď:**

Za první sledované období (2006) lze nakoupit Chléb konzumní kmínový v počtu 1 242 ks. Za poslední srovnatelné sledované období (2018) lze nakoupit Chléb konzumní kmínový v počtu 1 311 ks.

Za první sledované období (2006) lze nakoupit Mléko polotučné pasterované v počtu 1 386 L. Za poslední srovnatelné sledované období (2018) lze nakoupit Mléko polotučné pasterované v počtu 1 604 L.

Lze tedy konstatovat, že průměrné ceny sledovaných komodit rostly v průměru pomaleji než průměrná mzda.

**Tabulka 2 – ceny produktů vs průměrná mzda**

| Rok | Název produktu | KG/L |
|---|---|---|
| 2006 | Chléb konzumní kmínový | 1242.0 |
| 2018 | Chléb konzumní kmínový | 1311.0 |
| 2006 | Mléko polotučné pasterované | 1386.0 |
| 2018 | Mléko polotučné pasterované | 1604.0 |

Zdroj: Databáze společnosti Engeto na základě veřejných dat

V roce 2018 lze za průměrnou mzdu zakoupit o 69 ks chleba a 218 litrů mléka více než v roce 2006.

## 2.3 Postup a odpověď na třetí výzkumnou otázku

**Postup:**

1) Nejprve jsem vytvořil *view* third_q, který obsahuje názvy produktů, růst ceny, mezi roky a průměrnou cenu produktu. Tyto hodnoty jsou nutné pro případné zpracování čas.řad v MS Excel či Power BI.
2) Následně jsem pomocí Select vybral název produktu a provedl průměr růstu cen za každý rok se shrnutím dle názvu produktu z view third_q.

**Odpověď:**

Největší meziroční nárůst cen byl zjištěn u produktů rajská jablka červená kulatá o 3,81% , Cukr krystal  o 3,48%. Největší pokles cen byl zaznamenán u těstovin vaječných o -4,86%, másla -4,84% a rýže loupané dlouhozrnné o -3,93 %.

**Tabulka 3 – roční průměrný cenový růst**

| Název produktu | Průměrný růst ceny v % |
|---|---|
| Rajská jablka červená kulatá | 3,81 |
| Cukr krystalový | 3,48 |
| Konzumní brambory | 0,24 |
| Pečivo pšeničné bílé | 0,11 |
| Banány žluté | -0,39 |
| Jablka konzumní | -0,69 |
| Vepřová pečeně s kostí | -0,76 |
| Přírodní minerální voda uhličitá | -0,92 |
| Eidamská cihla | -1,24 |
| Šunkový salám | -1,74 |
| Pšeničná mouka hladká | -2,15 |
| Mléko polotučné pasterované | -2,3 |
| Mrkev | -2,36 |
| Hovězí maso zadní bez kosti | -2,37 |
| Kapr živý | -2,38 |
| Papriky | -2,51 |
| Jakostní víno bílé | -2,63 |
| Pivo výčepní, světlé, lahvové | -2,73 |
| Rostlinný roztíratelný tuk | -2,77 |
| Chléb konzumní kmínový | -2,86 |
| Vejce slepičí čerstvá | -2,89 |
| Pomeranče | -2,92 |
| Kuřata kuchaná celá | -2,94 |
| Jogurt bílý netučný | -3,6 |
| Rýže loupaná dlouhozrnná | -3,93 |
| Máslo | -4,84 |
| Těstoviny vaječné | -4,86 |

Zdroj: Databáze společnosti Engeto na základě veřejných dat

## 2.4 Postup a odpověď na čtvrtou výzkumnou otázku

**Postup:**

1) Z dat primární tabulky jsem vytvořil view fourth_q a vypočítal průměrné mzdy a průměrné ceny, zaokrouhlené na dvě desetinná místa.
2) Dále bylo vytvořeno další view fifth_q, kde byl vypočten růst ceny a růst mezd mezi roky z dat view fourth_q. Tento *„pátý“* pohled byl následně použit pro částečné řešení otázky č.5

**Odpověď:**

V naší tabulce neexistuje v období (2006–2018) nárůst cen o 10% a více než růst mezd ve sledovaném roce. Zajímavostí je, že v letech (2007, 2010, 2014, 2015, 2016, 2018) byl růst mezd vyšší než růst cen. Růst cen vetší, než růst mezd je v tabulce označen dvojitým podtržením.

**Tabulka 4 – růst cen potravin vs růst mezd v letech 2006–2018**

| Rok | Rok+1 | Růst ceny | Růst mezd |
|---|---|---|---|
| 2014 | 2015 | -0,5 | 1,7 |
| 2013 | 2014 | 0,7 | 4,2 |
| 2015 | 2016 | -1,2 | 3,8 |
| 2009 | 2010 | 1,9 | 1,9 |
| 2017 | 2018 | 2,2 | 3,3 |
| 2010 | 2011 | 3,4 | 3,1 |
| 2012 | 2013 | 5,1 | -1,9 |
| 2007 | 2008 | 6,2 | 10,3 |
| 2008 | 2009 | -6,4 | 1,9 |
| 2011 | 2012 | 6,7 | 2,7 |
| 2006 | 2007 | 6,7 | 7,6 |
| 2016 | 2017 | 9,6 | 9,2 |

Zdroj: Databáze společnosti Engeto na základě veřejných dat

## 2.5 Postup a odpověď na pátou výzkumnou otázku

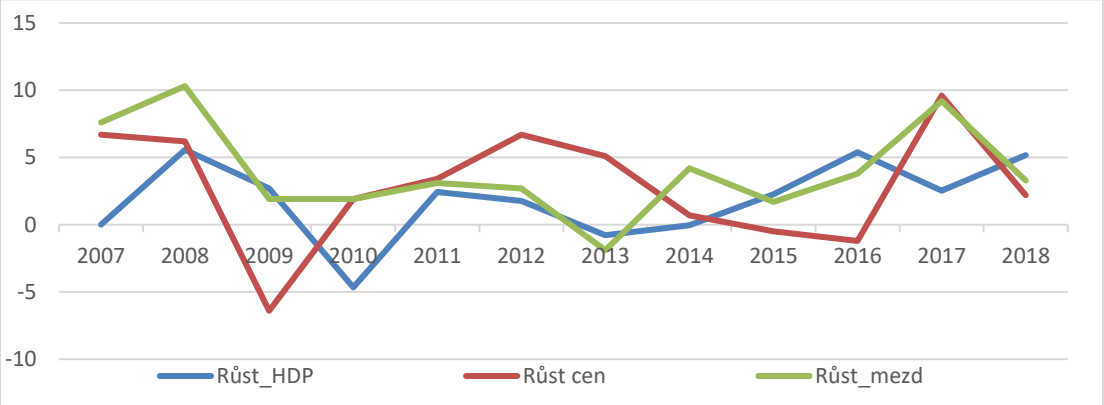
**Postup:**

1) Pomocí formule select jsem vybral nutné položky ze sekundární tabulky a pomocí lag jsem vytvořil sloupec s posunutým HDP vždy o rok. Na základě toho jsem dopočítal růst HDP. Růst cen a růst mezd jsem připojil pomocí JOIN z view fifth_q, který byl primárně použit pro odpověď na otázku č. 4.

**Odpověď:**

Na grafu č.1 je jednoznačné, že mezi růstem mezd a HDP existují nějaké souvislosti. V některých bodech, kde rostlo HDP dochází k růstu mezd. Avšak je nutné dalšího výzkumu.

**Graf 1 – Vzájemný vliv růstu (cen, HDP, mezd)**




Zdroj: vlastní zpracování

Na dynamiku růstu mezd má také vliv nezaměstnanost a restrikce vlády, kdy např. záměrně nedochází k růstu mezd státních zaměstnanců. Tito zaměstnanci jsou dostatečně velkým podílem na trhu práce, který může křivku průměrného růstu mezd ovlivnit.

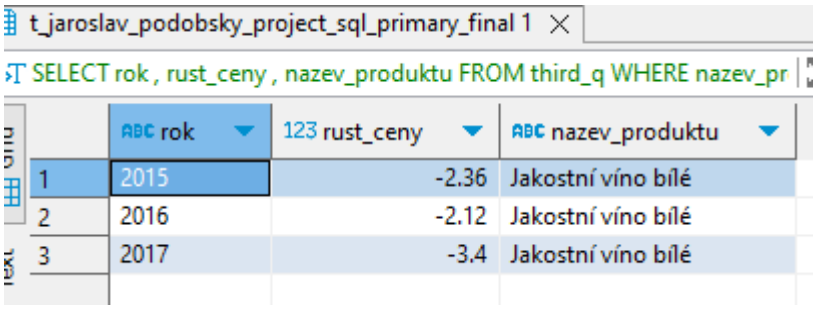
**Tabulka 5 – vliv HDP na cenu potravin a růst mezd**

| Rok | Rok +1 | HDP | HDP_min_rok | růst_HDP | růst_cen | růst_mezd |
|---|---|---|---|---|---|---|
| 2006 | 2007 | 197 470 142 754 | | | 6,7 | 7,6 |
| 2007 | 2008 | 208 469 898 851 | 197 470 142 754 | 5,57 | 6,2 | 10,3 |
| 2008 | 2009 | 214 070 259 128 | 208 469 898 851 | 2,69 | -6,4 | 1,9 |
| 2009 | 2010 | 204 100 298 391 | 214 070 259 128 | -4,66 | 1,9 | 1,9 |
| 2010 | 2011 | 209 069 940 963 | 204 100 298 391 | 2,43 | 3,4 | 3,1 |
| 2011 | 2012 | 212 750 323 791 | 209 069 940 963 | 1,76 | 6,7 | 2,7 |
| 2012 | 2013 | 211 080 224 603 | 212 750 323 791 | -0,79 | 5,1 | -1,9 |
| 2013 | 2014 | 210 983 331 026 | 211 080 224 603 | -0,05 | 0,7 | 4,2 |
| 2014 | 2015 | 215 755 991 069 | 210 983 331 026 | 2,26 | -0,5 | 1,7 |
| 2015 | 2016 | 227 381 745 549 | 215 755 991 069 | 5,39 | -1,2 | 3,8 |
| 2016 | 2017 | 233 151 067 381 | 227 381 745 549 | 2,54 | 9,6 | 9,2 |
| 2017 | 2018 | 245 202 003 266 | 233 151 067 381 | 5,17 | 2,2 | 3,3 |

Zdroj: Databáze společnosti Engeto na základě veřejných dat

# Informace o výstupních datech a závěr

1) Tabulky s cenami produktů zahrnují Jakostní víno, které mohlo být odstraněno, protože máme data omezena.


t_jaroslav_podobsky_project_sql_primary_final 1

SELECT rok , rust_ceny , nazev_produktu FROM third_q WHERE nazev_pr

| | rok | rust_ceny | nazev_produktu |
|---|---|---|---|
| 1 | 2015 | -2.36 | Jakostní víno bílé |
| 2 | 2016 | -2.12 | Jakostní víno bílé |
| 3 | 2017 | -3.4 | Jakostní víno bílé |

2) V primární tabulce byla provedena filtrace dat. Použity pro zpracování byly záznamy, které nemají value_type_code: !=316, ale pouze value_type_code =5958 a unit_code 200. Přepočtené mzdy nám nevadí z toho důvodu, že se jedná o částečné úvazky přepočtené na plné.

```
    ON cp.value_type_code =cpvt.code
WHERE value !=316 AND value IS NOT NULL AND calculation_code = '200'
    AND  unit_code= '200'
GROUP BY cp.payroll_year, cp.industry_branch_code
```

3) Snažil jsem se vyfiltrovat záznamy obsahující hodnotu: NULL

4) Nejvíce úplných dat se nachází v období 2006–2018.

# Přílohy

XLSX soubor vyexportovaných dat, který obsahuje výše uvedené tabulky.