**PRŮVODNÍ LISTINA**

## **Data o mzdách a cenách potravin a jejich zpracování pomocí SQL**

## **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.

#### **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?

#### **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.

**Postup zpracování projektu krok za krokem**

**Seznámení se s dostupnými daty**

Pro účely zodpovězení výzkumných otázek jsou klíčové primární tabulky czechia\_payroll, czechia\_price a jejich doplňkové tabulky.

Naopak tabulky czechia\_region a czechia\_district jsem nepoužila, jelikož zodpovídáme na otázky ohledně celé České Republiky. V případě, že bychom se chtěli například podívat na porovnání cen potravin v jednotlivých krajích, by bylo vhodné použít i tabulku czechia\_region.

Dále jsem pak potřebovala data z tabulky economies, která obsahuje data pro zodpovězení poslední výzkumné otázky ve spojení s tabulkou countries.

**Vytvoření primární tabulky t\_denisa\_louzilova\_project\_sql\_primary\_final**

Nejdůležitějším krokem bylo vytvoření hlavní tabulky se správnými daty a hlavně za stejná období.

Tabulka czechia\_payroll obsahuje průměrné mzdy a počty zaměstnanců za každé čtvrtletí sledovaných roků, zatímco tabulka czechia\_price obsahuje ceny potravin podle krajů za týdenní sledované období v průběhu let. Bylo tedy zapotřebí sjednotit data –  
u tabulky czechia\_price jsem použila funkci YEAR pro vybrání pouze roků a u obou tabulek byla třeba funkce AVG, aby mně tyto data zprůměrovala právě pouze za každý rok.

Nejdříve jsem přidala k tabulce czechia\_payroll pomocí LEFT JOIN tabulky czechia\_payroll\_industry\_branch, czechia\_payroll\_unit a czechia\_payroll\_value\_type  
(tabulka czechia\_payroll\_calculation obsahuje způsob získání dat, tudíž nebyla potřeba) a zobrazila jsem pouze potřebné sloupce se zmíněným výpočtem průměru a dále zaokrouhlení pomocí ROUND na 2 desetinná místa. Také jsem musela vyselektovat pouze data s kódem 5958, abych získala pouze data s průměrnými mzdami (kód 316 obsahoval počet zaměstnanců v odvětvích). Po zobrazení dat jsem zjistila, že je zapotřebí odebrat spoustu hodnot, které byly nulové pomocí IS NOT NULL.  
Posledním krokem bylo seskupit data GROUP BY podle odvětví a let.

Dále jsem přidala k tabulce czechia\_price pomocí LEFT JOIN tabulku czechia\_price\_category. Vybrala potřebné sloupce a použila zmíněné funkce YEAR na datum a AVG na ceny potravin, abych získala průměr cen za roky. Tabulku jsem seskupila GROUP BY podle let a druhu potravin a následně je seřadila podle skupin potravin a datumu sestupně DESC.

Tyto dva výběry jsem následně použila jako subselekty ‘a’ a ‘b’ pro finální tabulku, kdy jsem použila JOIN pro spojení těchto dat. Následně jsem pomocí CREATE OR REPLACE TABLE vytvořila tabulku, kterou jsem mohla v případě potřeb i upravit (správně řečeno nahradit = replace).

Z tabulky bylo zřejmé, že budu pracovat s daty z roků 2006-2018, jelikož pro tyto roky mám informace o průměrných mzdách i cenách.

**Vytvoření sekundární tabulky t\_denisa\_louzilova\_project\_sql\_secondary\_final**

Pro vytvoření pomocné tabulky jsem použila JOIN pouze jednou a to pro připojení countries a tabulky economies. Opět jsem vybrala pouze potřebné sloupce - především země, roky, HDP a další. Tabulku countries jsem přidala, jelikož obsahuje data o kontinentu a tím, jsem mohla vybrat do tabulky pouze Evropské země. Další podmínku jsem použila na roky 2006-2018, abychom se dívali na stejné časové období.   
Poté už jsem data pouze seřadila ORDER BY vzestupně podle zemí a let.

**Výzkumné otázky a odpovědi**

**1. Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?**  
  
Pro zodpovězení této otázky bylo především potřeba spočítat meziroční nárůst mezd v procentech. Ten jsem získala spojením primární tabulky s též primární tabulkou, kdy rok = rok + 1 . Tím jsem mohla spočítat meziroční nárůst mezd. A také za použití vzorečku: (p.pay\_value - p2.pay\_value)/p2.pay\_value \* 100

Dalším krokem bylo vzít tato vypočítaná data do subselectu a následně pomocí podmínky WHERE zobrazit pouze ty hodnoty, které jsou menší než nula.

Když jsem seřadila tyto hodnoty od nejmenší (tedy od největší mínusové hodnoty), zjistila jsem, že nejvyšší pokles mzdy byl v odvětví Peněžnictví a pojišťovnictví v roce 2013 - o 8.91%.

Další úpravou subselectu jsem si zobrazila pouze ta odvětví, která v letech 2006-2018 měla alespoň v jednom roce pokles mezd. Zobrazila se mně většina oborů - 15 ze sledovaných 19 oborů.  
Následně jsem si tento výsledek uložila do pohledu v\_pay\_drop, abych s ním mohla ještě pracovat.

Poté jsem si zobrazila odvětví z primární tabulky a odečetla od nich právě vytvořený pohled v\_pay\_drop. Tím jsem získala zbylé 4 obory, u kterých nedošlo ve sledovaných letech k žádnému poklesu mezd: Doprava a skladování, Ostatní činnosti, Zdravotní a sociální péče a Zpracovatelský průmysl.

Závěr: Přestože mzdy meziročně stoupají, v některých letech se s alespoň mírným poklesem mzdy setká většina odvětví.

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

Vzhledem k dostupným datum jsem porovnávala rok 2006 a 2018.

V prvním kroku jsem si našla označení pro mléko a chléb v primární tabulce pro snadnější vyhledávání: chléb code = 111301, mléko code = 114201

Následně jsem si zobrazila všechny potřebné sloupečky z primární tabulky a vydělila průměrnou mzdu cenou sledovaných potravin - round(pay\_value/price, 2) AS food\_amount. Také bylo důležité přidat podmínky pro zobrazení dat pouze pro chleba a mléko v letech 2006 a 2018. Následně jsem hodnoty seřadila tak, aby bylo možné z tabulky přečíst, jaké byly rozdíly v kupní síle pro jednotlivá odvětví.

Následně jsem ještě pomocí subselectu ukázala průměr kupní síly bez ohledu na odvětví.

Závěrem:

Množství cen potravin, které lze nakoupit za srovnatelná období, je závislé na růstu cen potravin a zvyšování mezd. Počítáme-li s průměrnou mzdou, můžeme usoudit, že zaměstnanci většiny (ne všech) sledovaných odvětví si v roce 2018 mohli nakoupit více potravin než v roce 2006.

Vezmeme-li v potaz celou Českou republiku, můžeme říct, že průměrný občan si mohl v roce 2018 (oproti roku 2006) dovolit nakoupit více mléka i chleba. Nicméně cena chleba rostla více než u mléka, tudíž rozdíl v množství nakoupeného mléka v letech je větší.

Pokud počítáme s celorepublikovým průměrem mezd, můžeme říct, že v roce 2006 si zaměstnanci ČR mohli dovolit koupit 1287,5 kg chleba, zatímco v roce 2018 to bylo 1342,2 kg chleba. U mléka byl nárůst ještě vyšší - v roce 2006 bylo možné koupit 1437,2 l mléka a v roce 2018 to bylo již 1641,6 l mléka.

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

Nejdříve jsem spočítala meziroční nárůst cen potravin v procentech. Ten jsem opět získala spojením primární tabulky s též primární tabulkou, kdy rok = rok + 1 . Tím jsem mohla spočítat meziroční nárůst mezd. A také za použití vzorečku: (tp.price - tp2.price ) / tp2.price \* 100. Tento výsledek jsem uložila do pohledu v\_food\_growth. Lze z něj vyčíst meziroční nárůsty cen potravin pro každou sledovanou potravinu za sledované roky.

Dále jsem z takto vytvořeného pohledu spočítala průměr meziročního nárůstu pro každou potravinu a seřadila je od nejmenší.

Závěr:

Mezi lety 2006 a 2018 je nejnižší průměrný percentuální nárůst u potraviny CUKR KRYSTALOVÝ -1.92 %. RAJSKÁ JABLKA ČERVENÁ KULATÁ jsou také v mínusové hodnotě -0.74 %, jedná se tedy spíše o pokles než nárůst.

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

V tomto případě bylo potřeba vzít nárůst cen potravin a mezd a následně tyto nárůsty porovnat.

Pro nárůst cen potravin jsem již měla připravený pohled v\_food\_growth k předešlé otázce. Stačilo tedy spočítat meziroční průměrný nárůst všech potravin s ohledem na roky.

Následně jsem vytvořila pohled v\_pay\_growth a to stejným způsobem jako u předešlého pohledu. Potom jsem spočítala průměr všech mezd na roky bez ohledu na odvětví.

Při spojení obou pohledů pro nárůst cen potravin a mezd jsem meziroční procenta od sebe odečetla a získala jsem tím rozdíl mezi oběma proměnnýma.

Závěr:

V žádném roce nebyl růst cen potravin oproti růstu mezd vyšší než 10% (za sledované období 2006 - 2018). Největší rozdíl byl v roce 2013 - o 6,79 %, kdy došlo k růstu cen potravin a snížení průměrných platů. Naopak v roce 2009 byl rozdíl -9,56 %, kdy došlo k výraznému poklesu cen potravin a mírnému růstu mezd.

1. **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 zjištění růstu HDP jsem ze sekundární tabulky opět vytvořila pohled v\_hdp\_growth, kde jsem opět spojila tabulku se stejnou tabulku, vytvořila year + 1 a použila vzoreček pro výpočet procent růstu.

Následně jsem k pohledu v\_hdp\_growth připojila již dříve používané pohledy v\_food\_growth a v\_pay\_growth. Tyto pohledy jsem opět vložila do subselektů, aby ukazovaly průměrné hodnoty na celou Českou Republiku.

Výsledná tabulka ukazuje meziroční procentuální změny HDP, potravin a mezd pro Českou Republiku v letech 2007 – 2018 (s rokem 2006 nepočítá, protože nemáme údaje pro srovnání s předešlým rokem).

Dodatečně jsem v tabulce použila CASE pro zobrazení společných let růstu HDP a potravin a další sloupec HDP a mzdy.

Závěr:

Po analýze dat lze říct, že při nárůstu HDP stoupají ceny jídla. Pokles HDP může mít i vliv na pokles cen potravin ve stejném roce nebo následujících letech. Mzdy podle dat nemusejí být ovlivněny růstem HDP, jelikož mají spíš rostoucí než klesající trend. U mezd byl klesající trend pouze v roce 2013, což také může souviset se záporným HDP v letech 2012 a 2013.

**Seznam souborů pro projekt:**

t\_denisa\_louzilova\_project\_sql\_primary\_final.csv

t\_denisa\_louzilova\_project\_sql\_secondary\_final.csv

v\_pay\_drop.csv

v\_pay\_growth.csv

v\_food\_growth.csv

v\_hdp\_growth.csv

tabulka\_primary.sql

tabulka\_secondary.sql

question1.sql

question2.sql

question3.sql

question4.sql

question5.sql

Accompanying document.docx