1. 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:

1.1. 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.

1.2. 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.

1.3. 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.

2. 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?

2.1. 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 zodpovězení na vytyčené výzkumné otázky. Pozor, otázky/hypotézy mohou vaše výstupy podporovat i vyvracet! Záleží na tom, co říkají data.

Na svém GitHub účtu vytvořte repositář (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.

3. Generování tabulek

3.1. Primární tabulka

Data v této tabulce budeme potřebovat k zodpovězení otázek jedna až čtyři. Abychom tyto otázky dokázali zodpovědět, potřebujeme znát průměrnou cenu jednotlivých kategorií potravin v jednotlivých letech a průměrnou mzdu ve všech odvětvích v daných letech.

Výsledná tabulka spojuje data z původních tabulek *czechia_payroll* a *czechia_price* pomocí funkce **UNION ALL** do jedné, finální tabulky. Ve výsledné tabulce je taktéž potřeba pomocí funkce **JOIN** rozklíčovat přiřazené kódy jednotlivých druhů potravin z původní tabulky *czechia_price_categor*.

Dále se v tabulce dopočítává průměrná mzda a průměrná cena jednotlivých položek v jednotlivých letech v letech 2006 až 2018 (kdy se máme data v obou spojovaných tabulkách).

Výsledná finální tabulka bude obsahovat následující data:

- *data_type* – salary/price
- *year* – rok, ke kterému se daná hodnota vztahuje (v tomto případě se pohybujeme v rozmezí let 2006 až 2018)
- *code* – kódy jednotlivých odvětví, kterých se daný řádek týká
- *avg_value* – průměrná hodnota mzdy/ceny pro daný rok a danou kategorii (počítaná hodnota, vždy v CZK)

3.2. Sekundární tabulka

Tato tabulka bude potřeba k zodpovězení páté otázky. Tato tabulka nám vznikne spojením tabulek *countries* a *economies*, kde nás budou zajímat následující sloupce:

- year (z *economies)*
- country (z *countries)*
- continent (z *countries)*
- GDP (z *economies)*
- population (z *economies)*
- avg_height (z *countries)*
- life_expectancy (z *countries)*

Tabulky si spojíme pomocí funkce **JOIN** na základě sloupce country, který se nám vyskytuje jak v tabulce *countries*, tak v tabulce *economies*. Nakonec si do finální tabulky necháme pomocí funkce **WHERE** vypsat pouze země, které leží v Evropě.

4. Zodpovězení výzkumných otázek

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

Vyjma odvětví s kódy H, N, Q, S (kde mzdy v průběhu sledovaných let pouze stoupají) ve všech dalších odvětvích dochází k výkyvům a v průběhu sledovaných let vždy alespoň jednou dojde k meziročnímu poklesu mzdy.

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

Za první srovnatelné období je možné si koupit cca 1.404 litrů mléka a 1.257 kilogramů chleba. Za poslední srovnatelné období je možné si koupit cca 1.611 litrů mléka a 1.317 kilogramů chleba.

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

Nejpomaleji zdražuje kategorie 116103 (banány žluté), a to tempem 0,81 %. Ještě můžeme říct, že položky 118101 (cukr krystal) a 117101 (rajská jablka červená kulatá) dokonce meziročně zlevňují (mají záporný meziroční nárůst ceny).

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

Ne, takovýto rok neexistuje. Rok, s nejvyšším meziročním nárůstem cen oproti mzdám, byl rok 2008, kdy tento nárůst dělal 8,05 %.

4.5. 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?

Otázkou je, co znamená "výraznější růst". Ale berme to jako v předešlé otázce a výraznějším růstem berme růst nad 10 %. A takovýto případ v našem datasetu nemáme. Navíc v případě změn GDP nemáme data dostupná pro všechny roky.