1 Cíl projektu

Cílem projektu je na základě poskytnutých dat (několik tabulek dostupných převážně na Portálu ověřených dat ČR) odpovědět na 5 výzkumných otázek:

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 Příprava

Před samotnou tvorbou podkladových tabulek bylo třeba si definovat, jaká data jsou pro zodpovězení otázek potřeba, kde je najít a jak spolu případně data v různých tabulkách souvisí. Analýzou tabulek v data_academy_content a výzkumných otázek vzešlo následující:

- Ve všech tabulkách jsou data členěna alespoň dle jednotlivých let. Na základě let bude probíhat i případné spojování dat
- Pro zodpovězení otázek potřebuji data:
    - Mzdy – rozdělené dle jednotlivých odvětví a let
    - Ceny produktů – opět dle kategorií a let
    - HDP dle jednotlivých let

3 Tvorba primární tabulky

Do primární tabulky t_michal_gabrle_project_SQL_primary_final bylo třeba zahrnout všechna potřebná data viz výše. Jako nejlepší řešení se mi jevilo spojit data z různých tabulek pomocí UNION s tím, že poslední sloupec (*type)* ponese informaci o tom, o jaký typ údaje se na daném řádku jedná. Spojoval jsem data z tabulek czechia_price (*type = price)*, czechia_payroll *(type = payroll)* a economies *(type = GDP_USD)*

Tabulka t_michal_gabrle_project_SQL_primary_final má ve výsledku 5 sloupců, které nesou informace (podle zdrojové tabulky, odkud pocházejí):

Zdroj	**Avg_value**	**Category_code**	**Category_name**	**Year**	**type**
Czechia_payroll	{Průměrná mzda za daný rok}	{Odvětví – kód} Null – republikový průměr	{Odvětví – název}	{Rok}	‚payroll‘
Czechia_price	{Průměrná cena za daný rok}	{Kategorie potravin – kód}	{Kategorie potravin – název}	{Rok}	‚price‘
Economies	{Hodnota daný rok}	‚CZ‘	‚Česká republika‘	{Rok}	‚GDP_USD‘

3.1 Poznámky

Při tvorbě primární tabulky jsem přemýšlel, zda nemám zrovna zahrnout pouze roky, ke kterým jsou dostupné údaje ze všech oblastí (gdp, price, payroll). Nakonec jsem se rozhodl vzít pro každou oblast co nejdelší časové období – zejména s ohledem na výzkumnou otázku č.1, kde se dlouhá časová řada bude hodit. V ostatních výzkumných otázkách, kde naopak potřebuji uvažovat pouze roky, ke kterým vím údaje z více typů dat, si časovou řadu ořízu pomocí join/intersect.

3.1.1 Czechia_payroll

Data v czechia_payroll jsou členěna dle jednotlivých čtvrtetí. Do primární tabulky jsem pro jednotlivá odvětví uvažoval průměr hodnot za daná čtvrtletí.

Data, kde odvětví není vyplněno (dle informací na Portálu ověřených dat ČR se jedná o úhrn za všechna odvětví) jsem neodfiltrovával – třeba se při nějaké výzkumné otázce budou hodit.

Dále jsou v czechia_payroll dle informací na Portálu ověřených dat ČR dostupné 2 typy hodnot (calculation code):

- položka 100 znamená fyzický počet zaměstnanců,
- položka 200 přepočtený počet zaměstnanců na plný úvazek

Rozhodl jsem se vzít pouze calculation code = 200, tedy hodnotu přepočtenou pouze na plné úvazky. Pro představu o platech a další práci s nimi (porovnávání odvětví) mi tato hodnota přijde vhodnější.

Je třeba brát ohled na to, že rok 2021 obsahuje data pouze za Q1 a Q2. Nicméně do tabulky jsem tento rok zprůměroval taktéž. V případné interpretaci dat (pokud by to zásadně ovlivňovalo nějakou z výzkumných otázek, je ale třeba toto zdůraznit).

3.1.2 Czechia_price

Do primární tabulky si beru celorepublikové průměry (region code je null), jelikož v žádné z otázek neřeším ceny v konkrétních krajích.

Data jsou uváděna vždy v časovém rozmezí datum_od a datum_do. Pro sjednocení dat s ostatními tabulkami uvažuji rok z hodnoty datum_od. Toto ošetření by nemělo žádný z údajů ovlivnit, jelikož kontrolou nad zdrojovou tabulkou:

*SELECT * FROM czechia_price WHERE date_part('year', date_from) <> date_part('year', date_to)*

bylo zjištěno, že žádné měření nepřesahovalo rok (nezačalo v jednom roce a neskončilo v dalším)

4 Tvorba sekundární tabulky

Do sekundární tabulky jsem přes join tabulek economies a countries zahrnul údaje:

- Country
- Year
- Gdp
- Gini
- Population

- Continent

Data jsem ořízl pouze do období, ve kterém jsou pro ČR dostupné jak informace v czechia_payroll, tak informace v czechia_price (intersect atributů year)

5 Výzkumné otázky

5.1 Otázka 1

Pro odpověď na otázku, zda v některých oblastech klesají průměrné mzdy jsem si vytáhl sloupce: rok, kód odvětví, název odvětví a průměrná mzda. Dále jsem přidal 2 vypočítané sloupce:

- Last_avg value - pomocí funkce LAG jsem si ke každému roku přidal i hodnotu průměrné mzdy v roce předchozím.
- Trend – v tomto sloupci se zobrazuje informace o tom, zda došlo k poklesu, růstu nebo žádné změně v průměrné mzdě.

Klauzulí where jsem data omezil pouze na type = payroll (údaje o mzdách) a category_code is not null (jelikož mě zajímají údaje po jednotlivých odvětvích, pro přehlednost výsledků jsem úhrny napříč odvětvími odfiltroval).

Pro jednoduché zodpovězení otázky je pak možnost si zobrazit pouze řádky, kde došlo k poklesu (trend = down). Pomocí tohoto selectu lze i velmi jedndouše zodpovědět výzkumnou otázku: Průměrné mzdy ve všech odvětvích nerostou. Lze najít i taková odvětví, kde došlo k meziročnímu poklesu.

Pokud bychom se ale na vývoj mezd podívali s větším odstupem (delším časovým horizontem), situace se změní, což nám zobrazuje poslední select. Ten porovnává vývoj mezd nikoliv meziročně, ale za posledních 5 let. Jelikož na časovém horizontu 5 let se nám nevrátí žádný řádek, kde by trend = down, lze odpověď na otázku 1 trochu upravit:

Došlo k několika případům, kdy mzdy v daném odvětví meziročně poklesly. Nicméně při delším sledovaném období (5 let), lze říci, že mzdy ve všech odvětvích rostou.

5.2 Otázka 2

Zde jsem si potřeboval připravit průměrný plat, název komodity a cenu komodity po jednotlivých letech (řádcích). Využil jsem k tomu možnost joinu na tutéž tabulku – a to na základě roku. Nejprve si beru rok a průměrnou mzdu, joinem si k tomu přidávám do kombinace název a cenu komodity.

Pro každý řádek pak prostým dělením počítám (sloupec amount_buy), kolik které komodity bylo možné v daném roce za průměrný plat nakoupit.

Jelikož nás v otázce zajímá jen první a poslední rok (pro která jsou data), nachystal jsem si nalezení správného roku do klauzule WITH. Nejprve si v *dataset* vyfiltruji všechna data. V CTE části *years* si pomocí UNION spojím první a poslední rok. To pak využívám v hlavním selectu (where year in…).

5.3 Otázka 3

Pro zjištění, která kategorie nejpomaleji zdražuje, jsem si vypočítal průměrný roční rozdíl ceny v procentech. Ve vnořeném selectu jsem si nejprve nalezl kód + název kategorie, rok a průměrnou cenu. K těmto sloupcům jsem přidal 2 vypočítané:

- avg_value_last_year - Pomocí funkce LAG jsem si vypsal průměrnou cenu minulý rok.
- difference_percent – meziroční změna ceny v procentech

Ve vnějším selectu pak už jen z hodnot difference_percent počítám celkový průměr, abych zjistil, která kategorie rostla nejpomaleji. Ve výsledcích s objevují i kategorie se zápornou hodnotou (v průměru zlevnily). Ty jsem pak případně (select od řádku 22) pomocí klauzule HAVING odfiltroval.

5.4 Otázka 4

Pro zodpovězení této otázky jsem si potřeboval dát do souvislostí (po jednotlivých letech) údaj o meziroční změně mezd a meziroční změně cen potravin.

Pomocí CTE jsem spočítal:

- Price_to_year – meziroční procentní změna cen potravin dle jednotlivých kategorií
- Avg_price_year_to_year – průměrná meziroční procentní změna (průměr meziročních změn cen jednotlivých kategorií)
- payroll_year_to_year – meziroční procentní změna mezd dle odvětví. Zde jsem akorát odfiltroval celorepublikové průměry (category code is not null). Varianta by byla uvažovat právě jen republikové průměry, které jsou v primární tabulce též zahrnuté
- avg_payroll_year_to_year – průměrná meziroční změna mezd
- price_payroll_difference – zde už po jednotlivých letech do sloupce price_payroll_difference vracím rozdíl procentních změn průměrných cen a mezd..

Finální select pak už jen nalezne všechny rozdíly, které jsou větší než 10.

5.5 Otázka 5

Podklady pro tuto otázku se ve velké míře podobají podkladům v otázce č.4. Jen je potřeba je rozšířit o podklady s meziroční změnou HDP.

K tabulkám CTE price_year_to_year, avg_price_year_to_year, payroll_year_to_year, avg_payroll_year_to_year jsem přidal avg_gdp_year_to_year, kde ve sloupci year_to_year_percent počítám meziroční změnu HDP.

Vše si pak pomocí joinu těchto tabulek vrátím na jednotlivé řádky po jednotlivých letech. Zde už lze zkoumat jak změna HDP ovlivní změny cen/mezd.