Průvodní dokumentace k SQL projektu ENGETO

A. PRIMÁRNÍ TABULKA

Tento SQL skript slouží k vytvoření primární tabulky pro analýzu dat spojených s výplatami a cenami zboží v České republice. První část kódu definuje pohled „v_czechia_payroll“, který kombinuje různé informace (a číselníky) o platbách zaměstnanců v České republice, jako jsou hodnoty, jednotky, kalkulace, odvětví průmyslu, rok a čtvrtletí.

Druhá část definuje pohled „v_czechia_price“, který agreguje informace o cenách zboží v České republice. Průměrná hodnota zboží je zaokrouhlena na dvě desetinná místa a spojena s kategorií zboží, názvem, hodnotou ceny a jednotkou ceny, rokem a čtvrtletím.

Poslední část vytváří (CREATE OR REPLACE TABLE) primární tabulku t_radek_komarek_project_SQL_primary_final, která slouží jako primární tabulka pro analýzu dat z obou pohledů. Tato tabulka kombinuje informace z obou pohledů pomocí operace UNION, která spojí výsledky do jedné tabulky. Konkrétně obsahuje hodnoty, kódy, názvy, jednotky, odvětví průmyslu, rok a čtvrtletí z pohledu výplat a hodnoty, kódy, názvy, rok a čtvrtletí z pohledu cen zboží. Bylo nutné také počítat s tím, že pomocí UNION byly spojeny obě tabulky a do stejných sloupců a ve sloupci „industry“ se vytvořily dvě skupiny řádků, tedy „industry“ – průmyslové odvětví, ale také „measurement“ neboli jednotky pro měření jako kilogramy, litry apod. Toto bylo nutné potom zohlednit v následujících úkolech, při nastavení dotazů pro analýzu.

(Pro spojení pohledů jsem zkoušel také spojení pomocí LEFT JOIN, ale bohužel nefungovalo to jako UNION, využil jsem tedy UNION, abych měl správná data pro analýzu dat.)

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

V rámci této výzkumné otázky jsem využil zvolil možnost vypočtu pomocí vloženého selectu, ve kterém jsem využil funkce LAG u sloupce value pro získání hodnoty z předchozího roku řazenou dle price_year a rozdělenou dle industry_name. Pro získání rozdílu mezi hodnotou aktuálního a předchozího roku jsem využil znovu funkce LAG, ale s využitím výrazu „value – LAG(value) OVER…“. Filtrování záznamů jsem provedl pomocí „yearly_diff“, který měl být záporný.

Výstupem tedy je, že v průběhu let mzdy reálně spíše rostou, ale s mírnými výkyvy. U 37 výsledků dotazu je, v průběhu let, zaznamenán také pokles mezd v různých odvětvích.

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

Tento kód provádí výpočet poměru mezi hodnotami zboží a mzdou v jednotlivých odvětvích průmyslu pro dvě kategorie zboží (chléb a mléko) v konkrétních letech 2006 a 2018. Pro každou kategorii zboží jsem použil funkci `ROUND(value / (SELECT value FROM t_radek_komarek_project_sql_primary_final WHERE value_code = ... AND year = ... AND quarter = ...))`, která počítá poměr mezi hodnotou daného zboží a mzdou. Tento poměr odpovídá tomu, kolik jednotek daného zboží (chleba nebo mléka) by bylo možné koupit za jednu jednotku průměrné hrubé mzdy na zaměstnance. Výsledky jsou zaokrouhleny na dvě desetinná místa pomocí funkce `ROUND`. Vytvořil jsem dva dotazy a tyto dotazy posléze spojil pomocí UNION v jednu výslednou tabulku.

V dotazu jsem využil poddotazy (subqueries) pro získání hodnoty zboží pro chléb a mléko v zadaném roce a čtvrtletí. Následně jsou tyto hodnoty využity k výpočtu poměru mezi hodnotou zboží a mzdou.

Ve výsledku vychází, že ve srovnání těchto let u některých průmyslových odvětví rostly mzdy méně než cena komodit (chléb a mléko). Takže zaměstnanec si za průměrnou hrubou mzdu mohl koupit v roce 2018 méně než v roce 2006. Pokud vezmu jako příklad odvětví ‚Doprava a skladování', tak byl rozdíl u chleba takový, že si zaměstnanci mohli dovolit méně chleba v roce 2018 než v roce 2006, zatímco u mléka je rozdíl opačný, tedy, že zaměstnanec si mohl koupit mléka více v roce 2018.

1,202	1,244	17,719	5,958	Průměrná hrubá mzda na zaměstnance	Kč	Doprava a skladování	2,006
1,218	1,261	17,960	5,958	Průměrná hrubá mzda na zaměstnance	Kč	Doprava a skladování	2,006
1,189	1,471	28,791	5,958	Průměrná hrubá mzda na zaměstnance	Kč	Doprava a skladování	2,018
1,200	1,485	29,069	5,958	Průměrná hrubá mzda na zaměstnance	Kč	Doprava a skladování	2,018

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

V tomto SQL dotazu se zaměřuji na vypočet průměrné (a percentuální) míry růstu, minimální míry meziročního růstu cen potravin. Dataset obsahuje informace o cenách nebo hodnotách v průběhu let, s cílem analyzovat míry růstu těchto hodnot v čase.

Vnitřní dotaz vypočítává průměrnou hodnotu pro každý rok spolu s hodnotou předchozího roku a vypočítává procentní růst mezi vybranými roky. Vnější dotaz vypočítává průměrnou míru růstu napříč všemi lety (2007 – 2021), minimální míru růstu (min_year_growth) a seskupuje výsledky podle názvu hodnoty (value_name). Výsledky jsou seřazeny podle průměrné míry růstu (avg_growth).

Filtr a spojení dat: Data jsou filtrována tak, aby byla vyřazena některá nežádoucí data (podle hodnoty value_code a industry), a jsou spojena na základě názvu hodnoty (value_name) a roku (year), tak aby se získaly hodnoty pro aktuální a předchozí rok v rámci vybraných let.

Výsledkem je tedy seskupená sada dat obsahující průměrnou hodnotu růstu ceny (avg_growht), minimální hodnotu růstu z předchozího roku (min_year_growht), ale také název komodity, jednotka míry). Průměrný nárůst se týká většiny komodit, kromě dvou „Cukr krystalový" a „Rajská jablka červená kulatá", která v meziročním srovnání zlevnila a tímto teda zaznamenala nejnižší meziroční nárůst ceny. Do 10% zaznamenaly nárust „Banány žluté" a „víno jakostní bílé". Dalších 23 komodit zaznamenalo více než 10% nárůst. Nejvyšší meziroční nárůst zaznamenaly „Papriky", a to téměř 87%.

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

Tento dotaz slouží k analýze průměrného procentního růstu cen potravin (nebo mezd) v průběhu let. Dotaz začíná vnořeným SELECTem, který vypočítává průměrnou hodnotu (avg_value) určitého ukazatele (např. cena potravin) pro každý rok, a také hodnotu tohoto ukazatele pro předchozí rok (year_value). Poté vypočítává procentuální růst (growth_percent) mezi průměrnými hodnotami těchto ukazatelů mezi aktuálním a předchozím rokem.

Vnořený SELECT se spojuje s druhým SELECTem pomocí LEFT JOIN, pro získání dat z předchozího roku. K odstranění nulových hodnot jsou využity sloupce value_code a industry. K seskupení dat podle roku a názvu ukazatele.

Vnější SELECT pak používá výsledky vnitřního vnořeného SELECTu k výpočtu průměrného procentního růstu cen potravin (price_or_wages_growth_perc) pro každý rok. Tyto výsledky jsou seskupeny podle roku a seřazeny sestupně podle průměrného procentního růstu.

Výsledná tabulka ukazuje, že nárůst cen potravin byl vyšší než růst mezd v případě 8 řádků ve sloupci price_or_wages_growth_perc. Konkrétně nejvyšší nárůst cen potravin byl v roce 2007, a to 9,27 %, následovaný rokem 2008 s nárůstem 8,48 %. Zatímco největší pokles cen potravin byl zaznamenán v roce 2009, konkrétně 6,23 %.

B. SEKUNDÁRNÍ TABULKA

Tento SQL skript vytváří sekundární tabulku s názvem t_radek_komarek_project_sql_secondary_final. Tato tabulka slouží k další analýze dat, která jsou odvozena z primární tabulky (t_radek_komarek_project_sql_primary_) a další externí tabulky s ekonomickými údaji (economies).

Výběr sloupců do sekundární tabulky zahrnuje všechny sloupce z primární tabulky t_radek_komarek_project_sql_primary_final a přidává další informace jako země, hrubý domácí produkt (GDP) a označení typu dat (food_or_wage), které určuje, zda se jedná o údaje o mzda nebo o údaje o potravinách.

Informace o zemi a HDP jsou přidány pomocí LEFT JOIN s externí tabulkou economies, která obsahuje ekonomické údaje o různých zemích. Při spojení jsou použity pouze údaje týkající se České republiky (filtr country LIKE '%czech%').

Navíc jsou použity následující podmínky:

- Řádky s hodnotou value_code rovnou 316 jsou vyloučeny.
- Řádky, kde není uvedeno odvětví průmyslu (industry), jsou také vynechány.

Celkově lze tedy sekundární tabulku chápat jako rozšíření primární tabulky o další ekonomické informace pro Českou republiku a jako filtraci dat, aby byly zahrnuty pouze relevantní záznamy pro další analýzu (v případě tohoto projektu pro úkol číslo 5).

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?

Tento SQL dotaz slouží k vyhodnocení vlivu hrubého domácího produktu (HDP) na změny ve mzdách a cenách potravin v České republice. Dotaz konkrétně zkoumá, zda výraznější růst HDP v jednom roce má vliv na změny ve mzdách a cenách potravin.

Dotaz vypočítává průměrnou hodnotu (avg_value), konkrétně mezd nebo cen potravin, za určitý rok a v rozmezí let 2007 - 2021. Dále vypočítává roční procentní růst (growth_percent) těchto hodnot oproti předchozímu ročníku. Totéž je provedeno i pro HDP, kde se vypočítává procentní růst HDP (gdp_growth) mezi aktuálním a předchozím rokem.

Kromě toho dotaz zahrnuje několik dalších sloupců:

- value_name, unit a industry_or_measurement identifikují název ukazatele, jednotku a příslušné odvětví průmyslu nebo měření.
- YEAR a prev_year označují aktuální rok a předchozí rok.
- food_or_wage určuje, zda se jedná o údaje o mzdách nebo o cenách potravin.
- country identifikuje zemi, v tomto případě Českou republiku.

Dotaz obsahuje také LEFT JOIN (self join tabulky t_radek_komarek_project_sql_secondary_final) s poddotazem t2, který vypočítává průměrnou hodnotu ukazatele, jednotku, odvětví průmyslu, rok a HDP pro každý rok. To umožňuje porovnání hodnot mezi aktuálním a předchozím rokem.

Celkově dotaz analyzuje, zda významné změny v HDP vedou k podobně významným změnám v mzdách a cenách potravin v České republice.

Průměrná hrubá mzda (wages)

Zde je patrné, že vliv zvýšení HDP se propisuje i do nárůstu průměrných mezd, tedy, že pokud je vyšší nárůst HDP v roce 2007, tak v roce 2008 je nárůst také vyšší, zatímco v roce 2008 byl zhruba tříprocentní pokles HDP, tak v roce 2009 je patrné, že je přibližně 4,5 procentní pokles průměrných hrubých mezd.

	avg_value	year_value	growth_percent	GDP	gdp_growth	value_name	unit	industry_or_measurement	YEAR	prev_year	food_or
1	22,172.75	20,753.78	6.84	59,898,850.69	5.57	Průměrná hrubá mzd	Kč	Zemědělství, lesnictví, rybářství	2,007	2,006	wages
2	23,918.28	22,172.75	7.87	0,259,127.502	2.69	Průměrná hrubá mzd	Kč	Zemědělství, lesnictví, rybářství	2,008	2,007	wages
3	24,674	23,918.28	3.16	0,298,391.036	-4.66	Průměrná hrubá mzd	Kč	Zemědělství, lesnictví, rybářství	2,009	2,008	wages
4	25,156.18	24,674	1.95	0,940,963.177	2.43	Průměrná hrubá mzd	Kč	Zemědělství, lesnictví, rybářství	2,010	2,009	wages
5	25,735.92	25,156.18	2.3	0,323,790.759	1.76	Průměrná hrubá mzd	Kč	Zemědělství, lesnictví, rybářství	2,011	2,010	wages

Ceny potravin (food)

Na příkladu potraviny „chléb konzumní kmínový“ lze vidět, že ceny rostly v letech 2007 a 2008 výrazně 16.83, případně 20.97 procenta, když ve zmiňovaných letech byl procentní nárůst HDP 5.57 v roce 2007 nebo 2.69 v roce 2008. Další rok, tedy 2009 HDP kleslo o 4.66 % a zároveň nastal i pokles cen potravin, konkrétně v tomto případě „chléb…“.

	avg_value	year_value	growth_percent	GDP	gdp_growth	value_name	unit	industry_or_measurement	YEAR	prev_year
1	18.88	16.16	16.83	208,469,898,850.69	5.57	Chléb konzumní kmí	1	kg	2,007	2,006
2	22.84	18.88	20.97	214,070,259,127.502	2.69	Chléb konzumní kmí	1	kg	2,008	2,007
3	19.58	22.84	-14.27	204,100,298,391.036	-4.66	Chléb konzumní kmí	1	kg	2,009	2,008
4	18.38	19.58	-6.13	209,069,940,963.177	2.43	Chléb konzumní kmí	1	kg	2,010	2,009
5	21.63	18.38	17.68	212,750,323,790.759	1.76	Chléb konzumní kmí	1	kg	2,011	2,010
6	22.92	21.63	5.96	211,080,224,602.858	-0.79	Chléb konzumní kmí	1	kg	2,012	2,011
7	23.12	22.92	0.87	210,983,331,025.624	-0.05	Chléb konzumní kmí	1	kg	2,013	2,012

Z výše zmíněného lze usuzovat, že nárůst v jednom roce HDP ovlivní nárůst průměrné hrubé mzdy, ale také může vést k nárůstu cen potravin, což může ovlivnit i rok následující po roce měření. V některých případech poměrně výrazně, jak je patrné u příkladu potraviny „chléb…“.