Průvodní listina k SQL projektu

Autor: Veronika Skálová

Datová akademie 13.9.2023

Dostupné online prostřednictvím GitHubu

Úvod

V této průvodní zprávě jsou představeny výsledky analýzy ekonomických dat týkajících se vývoje cen potravin a mezd v průběhu času z datasetu "Engeto_2023_09_13". Cílem tohoto projektu, realizovaného prostřednictvím jazyka SQL, je zodpovědět na výzkumné otázky a získat tak hlubší porozumění vývoje české ekonomiky v průběhu několika let.

1. Představení použitých tabulek

Pro zodpovězení výzkumných otázek byly využity následující pohledy (views), které kombinují data z tabulek obsažených v původním datasetu:

1. **t_veronika_skalova_project_SQL_primary_final.sql** (primary_final_v4):
   Toto view se skládá z následujících částí:
   - czechia_payroll_avg_by_yr_v2 (cpaby): Tento CTE shromažďuje a filtruje údaje o mzdách z tabulky czechia_payrol a vypočítává průměrnou mzdu pro každé odvětví a rok. Byly vyřazeny řádky, kde je hodnota (value) null, a value_type_code se nerovná 316 (údaje pro průměrný počet zaměstnanců). Následně byly použity tyto sloupce:
     - *payroll_year:* Rok, za který se vykazují zprůměrované mzdní údaje.
     - *industry_code:* Kód představující průmysl nebo sektor.
     - *average_salary:* Průměrná mzda v daném odvětví a roce.
     - *currency:* Kód měny, ve které je mzda vykazována.
   - czechia_payroll_unit (cpu): Tato tabulka je připojena k cpaby pomocí sloupce měny k získání názvu měny.
   - czechia_payroll_industry_branch (cpib): Tato tabulka je připojena k cpaby pomocí sloupce industry_code k získání názvu odvětví.
   - czechia_price_avg_by_yr (cpaby2):
     - *price_year:* Rok, za který se vykazují údaje o průměrných cenách potravin. Je spojen s cpaby pomocí sloupce payroll_year.
     - *category_code:* Kód představující kategorii potravin.
     - *average_value:* Průměrná hodnota/cena kategorie potravin.
   - czechia_price_category (cpr): Tato tabulka je připojena k cpaby2 pomocí category_code k získání názvu a jednotky kategorie potravin.

Tabulky czechia_price a czechia_payroll původně obsahovaly data v odlišném formátu, která musela být sjednocena prostřednictvím funkce YEAR.

2. **t_veronika_skalova_project_SQL_secondary_final.sql** (secondary_final_v2):
   Toto view se skládá z následujících tabulek, propojených JOIN na základě země:
   - countries (c): Tabulka poskytuje informace o zemích, včetně názvu země.
     - *country:* Název země.
   - economies (e):
     - *country:* Název země.
     - *YEAR:* Rok, za který jsou vykazovány příslušné ekonomické údaje.
     - *GDP:* Hrubý domácí produkt (HDP) za daný rok a zemi.
     - *population:* Počet obyvatel země za daný rok.
     - *gini:* GINI index, měřítko příjmové nerovnosti pro daný rok a zemi (tento sloupec nakonec nebyl potřebný pro analýzu).

2. Výzkumné otázky a odpovědi

Zadání projektu obsahuje tyto výzkumné otázky. Postup řešení je vysvětlen prostřednictvím odpovědí a zároveň je podrobněji ilustrován ve souborech nahraných do repozitáře, které obsahují poznámky ke kódu.

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

Odpověď:

Analýza mzdových údajů ukazuje, že mzdy ve všech odvětvích v průběhu let trvale rostly, s výjimkou občasných poklesů napříč odvětvími, které však nikdy netrvaly déle než rok.

Meziroční výsledky jsou ilustrovány v CTE "salary_changes_q1", ve kterém jsou obsaženy vypočítané sloupce "salary_change_percentage" a "salary_change_status". K získání meziročních výsledků byla použita funkce "LAG", která umožňuje získat hodnotu "average_salary" z předchozího řádku, seřazenou podle "payroll_year". Dále byla použita klauzule "PARTITION BY", která zajišťuje, že funkce "LAG" počítá pouze s předchozím rokem v rámci stejného odvětví.

payroll_year	industry_code	industry_name	average_salary	previous_yr_avg_salary	salary_change_percentage	salary_change_status
2,006	A	Zemědělství, lesnictví, rybářství	14,619.38	[NULL]	[NULL]	Data not available
2,007	A	Zemědělství, lesnictví, rybářství	15,974.38	14,619.38	9.27	Increase
2,008	A	Zemědělství, lesnictví, rybářství	17,528.13	15,974.38	9.73	Increase
2,009	A	Zemědělství, lesnictví, rybářství	17,418.75	17,528.13	-0.62	Decrease
2,010	A	Zemědělství, lesnictví, rybářství	18,233.88	17,418.75	4.68	Increase
2,011	A	Zemědělství, lesnictví, rybářství	18,829.88	18,233.88	3.27	Increase
2,012	A	Zemědělství, lesnictví, rybářství	19,683.38	18,829.88	4.53	Increase

Řešení dále obsahuje dodatečné CTE, které ukazuje průměrný nárůst mezd za dostupné časové údaje napříč odvětvími. Toto je zaznamenáno v sloupci "average_salary_change_percentage", který ukazuje, že nejpomaleji v průměru rostly mzdy v oboru Peněžnictví a pojišťovnictví.

industry_code	industry_name	average_salary_change_percentage
K	Peněžnictví a pojišťovnictví	2.59
S	Ostatní činnosti	3.05
B	Těžba a dobývání	3.31
N	Administrativní a podpůrné činnosti	3.41

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

Odpověď:

Analýza ukázala, kolik je možné si koupit mléka a chleba za první a poslední srovnatelné období v dostupných datech cen a mezd. Výsledky jsou uvedeny v CTE "bread_milk_avg_salary", jak je ukázáno na následujícím screenshotu.

Sloupec "payroll_year" obsahuje první a poslední srovnatelné období (2006 a 2018), které je založeno na datech dostupných z tabulek "czechia_price" a "czechia_payroll".

Sloupec "average_price" ukazuje průměrnou cenu mléka a chleba za daný rok.

Aby bylo dosaženo požadovaného zobrazení, byla do CTE přidána subquery "payroll_years", která vypočítává pomocí funkcí "MIN" a "MAX" první a poslední srovnatelný rok. Hlavní dotaz spojuje CTE "bread_milk_avg_salary" a subquery "payroll_years" pomocí "CROSS JOIN" (ON 1=1), což umožňuje kombinovat data každého roku s minimálním a maximálním počtem let nalezených v původním dotazu. Zobrazení by nebylo možné bez klauzule "WHERE", která filtruje první a poslední srovnatelný rok.

payroll_year	average_salary	currency	food_type	average_price	currency	food_purchasing_power	measure_unit
2,006	20,753.79	Kč	Chléb konzumní kmínový	16.12	Kč	1,287.46	kg
2,018	32,535.86	Kč	Chléb konzumní kmínový	24.24	Kč	1,342.24	kg
2,006	20,753.79	Kč	Mléko polotučné pasterované	14.44	Kč	1,437.24	l
2,018	32,535.86	Kč	Mléko polotučné pasterované	19.82	Kč	1,641.57	l

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

Odpověď:

Nejpomaleji zdražovala kategorie "Rajská jablka červená kulatá." Pro tuto odpověď byla použita funkce "LAG" pro výpočet procentuálního meziročního nárůstu cen potravin.

Abychom mohli identifikovat potravinu, která zdražuje nejpomaleji, byl použit vzorec, který počítá procentuální nárůst průměrných cen z předchozího roku do aktuálního roku a zároveň zjišťuje minimální nárůst v rámci každého typu potravin:

MIN(ROUND((average_value - previous_year_value) / previous_year_value * 100, 2))

Výsledek bylo ještě potřeba seřadit podle nejnižšího výsledku s použitím funkce "LIMIT 1," aby byl vybrán pouze jeden řádek s nejnižším procentuálním nárůstem.

food_code	food_type	lowest_increase
117,101	Rajská jablka červená kulatá	-30.28

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

Odpověď:

Analýza neprokázala existenci roku, ve kterém byl meziroční nárůst cen potravin výrazně vyšší než růst mezd (větší než 10 %). Výsledky byly získány porovnáním průměrných ročních mezd a cen potravin ("yearly_averages").

Nejprve byla vytvořena subquery "percentual_changes," kde byly přidány kalkulované sloupce, které prostřednictvím vzorce počítají procentuální změnu průměrných mezd a cen potravin oproti minulému roku s použitím funkce "LAG."

payroll_year	avg_salary	salary_change_percent	price_year	avg_food_price	food_change_percent
2,006	20,753.79	[NULL]	2,006	45.52	[NULL]
2,007	22,172.75	6.84	2,007	48.59	6.74
2,008	23,918.28	7.87	2,008	51.6	6.19
2,009	24,674	3.16	2,009	48.29	-6.41
2,010	25,156.19	1.95	2,010	49.23	1.95
2,011	25,735.92	2.3	2,011	50.88	3.35
2,012	26,516.09	3.03	2,012	54.3	6.72

Dále byl použit "LEFT JOIN" pro spojení "yearly_averages" (Y) za současný rok a "yearly_averages" (P) za předchozí rok (Y.payroll_year = P.payroll_year + 1).

Finální SELECT filtruje výsledky tak, aby se zobrazily pouze řádky, kde je změna cen potravin větší než 10 % ve srovnání se změnou mezd z jednoho roku na druhý. Žádný takový rok nebyl v datasetu nalezen.

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?

Odpověď:

Zdá se, že tomu tak je – HDP má vliv na vývoj mezd a cen.

V poslední otázce byla nejprve připojena k hlavní tabulce s daty o mzdách a cenách potravin tabulka "secondary," která obsahuje ekonomické údaje filtrované pro Českou republiku. Následně byly přidány tři vypočítané sloupce s použitím funkce "LAG" pro výpočet meziročních procentuálních změn pro HDP, mzdy a ceny potravin.

Při prvním pohledu na data je patrné, že pokud HDP vzroste např. nad 5 %, má to tendenci ovlivnit růst mezd a cen potravin ve stejném nebo následujícím roce, neboť ty se zpravidla také zvyšují.

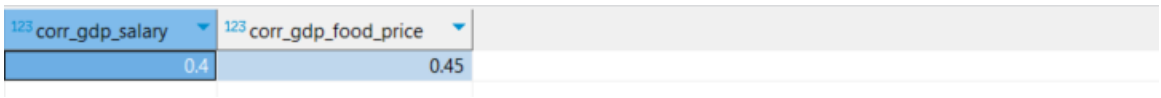
country	economy_year	GDP_mil_dollars	GDP_change_percent	salary_change_percent	food_price_change_percent
Czech Republic	2,006	197,470.14	[NULL]	[NULL]	[NULL]
Czech Republic	2,007	208,469.9	5.57	6.84	6.74
Czech Republic	2,008	214,070.26	2.69	7.87	6.19
Czech Republic	2,009	204,100.3	-4.66	3.16	-6.41
Czech Republic	2,010	209,069.94	2.43	1.95	1.95
Czech Republic	2,011	212,750.32	1.76	2.3	3.35
Czech Republic	2,012	211,080.22	-0.79	3.03	6.72
Czech Republic	2,013	210,983.33	-0.05	-1.56	5.1
Czech Republic	2,014	215,755.99	2.26	2.56	0.74
Czech Republic	2,015	227,381.75	5.39	2.51	-0.54
Czech Republic	2,016	233,151.07	2.54	3.65	-1.21
Czech Republic	2,017	245,202	5.17	6.28	9.63
Czech Republic	2,018	253,045.17	3.2	7.62	2.16

Bohužel, funkce "CORR" pro výpočet korelačního koeficientu, která by to mohla přímo potvrdit, v DBeaveru nefunguje (zdá se, že není standardní SQL funkcí).

Jako alternativní řešení byla vytvořena další CTE s názvem "gdp_correlation," která obsahuje vzorec pro Pearsonův korelační koeficient (měřící lineární vztah mezi dvěma proměnnými). Pro výpočet korelačního koeficientu byl použit vzorec, který zahrnuje "STDDEV_POP" funkci, jež vrací směrodatnou odchylku. Tím lze nahradit funkci "CORR."

```
(SUM((GDP_change_percent - avg_gdp_change) * (salary_change_percent - avg_salary_change))
    / (n * STDDEV_POP(GDP_change_percent) * STDDEV_POP(salary_change_percent)))
```

Výsledný korelační koeficient se pohybuje v rozmezí od -1 do 1, kde kladné hodnoty naznačují pozitivní provázanost mezi změnami HDP, mezd a cenami potravin. V našem případě výsledky činily 0.4 a 0.45, což svědčí o provázanosti HDP s cenami a mzdami.



corr_gdp_salary	corr_gdp_food_price
0.4	0.45

3. Závěr

Tento projekt prostřednictvím analýzy SQL dat poskytl odpovědi na stanovené výzkumné otázky, a tím přinesl hlubší pochopení vývoje české ekonomiky v průběhu několika let. Doufám, že se podařilo dostatečně popsat průběh postupu při řešení prostřednictvím SQL dotazů.