Průvodní listina | SQL projekt

Michal Janečka (ENGETO – *Datová akademie*, 2024)

Níže uvedená průvodní listina se věnuje mzdám a cenám potravin v rámci České republiky, a to na základě celkově pěti stanovených výzkumných otázek.

Úvodním krokem projektu je seznámení se s dostupnými daty, tedy již předpřipravenými primárními a dodatečnými tabulkami a číselníky pro celkový přehled a orientaci. Díky znalosti obsahu jednotlivých tabulek bude následně možné vybrat pouze taková data, která budou sloužit k odpovědím na výzkumné otázky a zároveň bude dosaženo minimální časové i prostorové složitosti jednotlivých scriptů.

S ohledem na výše uvedené jsou tedy návazně na úvodní krok vytvořeny dvě nové, samostatné tabulky. První z nich (*primary*) slouží pro data mezd a cen potravin za Českou republiku sjednocených na totožné porovnatelné období, druhá (*secondary*) pro dodatečná data o dalších evropských státech.

Konkrétně se jedná o následující tabulky:

- t_michal_janecka_project_SQL_primary_final
- t_michal_janecka_project_SQL_secondary_final

S ohledem na potřebu dat týkající se mezd a potravin, primární datový zdroj tabulky *czechia_price*, *czechia_price_category, czechia_payroll* a *czechia_payroll_industry_branch*. Pro účely návazné práce, nově vytvořená primární tabulka rovněž vylučuje irelevantní hodnoty v atributu *value_type_code* a to díky specifikaci požadavku na průměrnou hrubou mzdu (kód '5958').

Sekundární datový zdroj žádoucí pro poslední (5.) úkol spojuje tabulku *economies*, poskytující data o evropských státech včetně České republiky, s předešle vytvořenými views, které již obsahují potřebná data o průměrné výši mzdy a průměrných cenách potravin v požadovaných letech.

Výzkumné otázky

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

Výzkumná otázka stanovuje konkrétní požadavek na získání informací o celkově třech ukazatelích – mzda, odvětví a časové období.

Mzda

Primárním zdrojem hodnot je v tomto případě atribut *average_wage*, původně vycházející z tabulky *czechia_payroll* (*value)*.

Odvětví

Jednotlivá odvětví pod konkrétními názvy nese sloupec *industry*, původně vycházející z tabulky *czechia_payroll_industry_branch* (*name*).

Časové období

Období v letech stanovuje sloupec *payroll_year,* který po spojení výše uvedených tabulek stanovuje pomocí mezních hodnot časové rozpětí mezi lety 2006 a 2018. Toto období zároveň bude sloužit jako základ pro zbylé úkoly.

Postup

1) První z uvedených variant pracuje s průměrem hrubé mzdy vždy v konkrétním roce (2006 až 2018), a to vždy na základě průměru všech průměrných hrubých mezd za všechna odvětví dohromady. Výstup požadavku ukazuje, že průměrná hrubá mzda v meziročním srovnání vždy rostla, a to napříč všemi sledovanými odvětvími.

Odpověď na výzkumnou otázku: **Průměrná hrubá mzda *roste* v průběhu let ve všech odvětvích.**

2) Druhá z uvedených variant předkládá rozpad průměrné hrubé mzdy v daném roce (2006 až 2018) a to vždy dle konkrétního odvětví. Při pohledu na výstup se ukazuje, že v průběhu některých let průměrná mzda v meziročním srovnání v rámci daného odvětví neroste. Rozdíl ve výsledku mezi prvním a druhým postupem je způsoben zkreslením dat souhrnným průměrem za všechna odvětví v prvním případě, zatímco druhá varianta bere v potaz průměrné hrubé mzdy za daný rok vždy pouze pro konkrétní odvětví.

Odpověď na výzkumnou otázku: **Průměrná hrubá mzda *neroste* v průběhu let ve všech odvětvích.**

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

Výzkumná otázka stanovuje potřebu získat údaje o celkově třech parametrech – časové období, průměrná mzda a cena položek.

Cena

Pro potřebu výpočtu množství (l/kg) jednotlivých položek za určité období je nejprve nutné získat kódy, pod kterými jsou data k dispozici. Zdrojem je v tomto případě atribut *food_category*, respektive *food_code*, který stanovuje kód 111,301 pro 'Chléb konzumní kmínový' a kód 114,201 pro 'Mléko polotučné pasterované'. Cenu daných položek uvádí atribut *price.*

Časové období

Totožně jako v prvním úkole, v obou případech (chléb i mléko) je první srovnatelné období rok 2006, poslední pak rok 2018.

Mzda

Hodnotu průměrných mezd stanovuje atribut *average_wage*.

Postup

Na základě výše stanovených parametrů a zdrojů dat je první krokem postupu vypočet průměrné hrubé mzdy za všechna odvětví pro roky 2006 a 2018.

Jakmile jsou výsledné hodnoty k dispozici, lze přistoupit k výpočtu průměrné ceny chleba a mléka v konkrétních letech, přičemž výstupem jsou celkově čtyři hodnoty.

Závěrečným krokem je výpočet výsledného množství obou položek, tedy vydělení průměrné měsíční mzdy v daném roce cenou zboží. Pro tento krok je aplikováno dělení beze zbytku s cílem získat přesný, respektive dostupný počet kg/l daného zboží za průměrnou měsíční mzdu.

Odpověď na výzkumnou otázku: **V roce 2006 je možné si za průměrnou měsíční mzdu koupit 1287 kilogramů chleba či 1437 litrů mléka. V roce 2018 je možné si za průměrnou měsíční mzdu koupit 1342 kilogramů chleba či 1641 litrů mléka.**

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

Výzkumná otázka stanovuje potřebu získat údaje o celkově třech parametrech – časové období, průměrná cena, a informace o položkách.

Položky

Zdrojem dat je v tomto případě atribut *food_code*, který bude sloužit pro definici jednotlivých položek.

Časové období

Dostupnost dat stanovuje rozpětí mezi lety 2006 až 2018.

Cena

Hodnotu průměrných cen stanovuje atribut *price.*

Postup

Úvodním krokem je výpočet průměrných cen položek v dostupných letech. Cílem této operace je dostat data v podobě, která umožní jejich porovnání. Výstupem je průměrná cena položky ve vzestupném pořadí dle kategorie potraviny a jednotlivých let (2006 až 2018).

Návazně je pro přehlednost a potřebu navazujících operací vytvořen view, který slouží k výpočtu meziroční změny.

Finálním krokem je na základě dostupných dat výpočet meziročního rozdílu v cenách, a to jak v peněžním, tak v percentuálním vyjádření.

Data na výstupu zobrazují, že *Rajská jablka červená kulatá* (kód 117,101) představují kategorii, která zdražovala nejpomaleji. Jinými slovy, je u ní zaznamenán nejnižší percentuální meziroční nárůst, v tomto případě dokonce pokles ceny, a to konkrétně v roce 2007 oproti roku 2006. Celkový meziroční pokles dosáhnul hodnoty 30,28 %.

Odpověď na výzkumnou otázku: **Nejpomaleji rostoucí kategorií potravin je kategorie Rajská jablka červená kulatá (117,101).**

Poznámka: Rok 2006 je v rámci dat irelevantní, jelikož ve výsledném scriptu vypočítává meziroční pokles/nárůst oproti hodnotě za předchozí kategorii za rok 2018 a zároveň je prvním porovnatelným rokem.

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

Výzkumná otázka stanovuje potřebu získat údaje o celkově třech parametrech – průměrná mzda, průměrná cena, a časové období.

Časové období

Dostupnost dat stanovuje rozpětí mezi lety 2006 až 2018.

Mzda

Hodnotu průměrných mezd stanovuje atribut *average_wage*.

Cena

Hodnotu průměrných cen stanovuje atribut *price.*

Postup

Úvodním krokem je stanovení průměrné hodnoty pro jednotlivé roky 2006 až 2018, a to jak v případě cen potravin, tak pro výši mzdy.

Návazně na tento postup je možné přistoupit k meziročnímu porovnání jednotlivých hodnot a získat tak výstup v percentuálním vyjádření.

Výsledné operace pro oba případy ukazují, že žádná z hodnot meziročního srovnání nedosáhla úrovně přesahující 10 %.

Odpověď na výzkumnou otázku: **Neexistuje 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?

Výzkumná otázka stanovuje konkrétní požadavek na získání informací o celkově třech ukazatelích – HDP, mzda a cena položek.

HDP

Přehled o výši HDP v absolutním vyjádření u jednotlivých zemí nabízí tabulka *economies*. Pro účely výzkumné otázky slouží pro vyjádření meziroční změny tohoto ukazatele v případě České republiky a stanoví, ve kterém roce došlo k nejvyššímu nárůstu pro následné porovnání s vývojem cen potravin a mezd ve stejném a následujícím roce.

Mzda

Hodnotu průměrných mezd stanovuje atribut *average_wage*.

Cena

Hodnotu průměrných cen stanovuje atribut *price.*

Postup

Úvodním krokem je stanovení vhodného časového období, které v případě tabulky *economies* bude respektovat ty roky, které jsou k dispozici v rámci mezd a cen potravin. Výstupem již předcházejících výzkumných otázek je ohraničení tohoto časového rámce mezi roky 2006 a 2018.

Návazně jsou pro relevanci porovnávaných dat vyselektovány v rámci výše uvedené tabulky pouze údaje pro Českou republiku, a to za definované časové rozpětí. Poslední operací je stanovení percentuální meziroční změny pro HDP, mzdy, a ceny potravin a jejich seřazení dle nejvyšší úrovně růstu HDP v sestupném pořadí.

Nejvyšší hodnotou výsledné operace je rok 2007, ve kterém došlo k nejvýraznějšímu růstu HDP (5,57 %) z nabízených let. V souladu s výzkumnou otázkou je tedy pozornost v rámci meziročního růstu cen potravin a mezd následně věnována rokům 2007 a 2008. V případě růstu cen potravin došlo v roce 2007 k navýšení o 6,5 % a v roce 2008 o 6,32 %. Průměrná výše mzdy zaznamenala v roce 2007 nárůst o 6,84 % a rok 2008 růst ve výši 7,87 %.

Zatímco data pro roky 2007/2008 potvrzují tvrzení o růstu cen a průměrné hrubé mzdy v daném a následujícím roce v případě vysokého růstu HDP (2007), další dostupná data tuto hypotézu vyvracejí.

Příkladem může být rok 2015 odpovídající druhé nejvyšší hodnotě růstu HDP (5,39 %), kdy sice došlo k růstu mezd v roce 2015 i 2016, nicméně v případě cen potravin došlo jak v roce 2015, tak v roce 2016 k poklesu. Stejně tak mzdy například rostly i v letech (2009 a 2010), kdy HDP zaznamenalo nejvyšší meziroční pokles (2009).

Odpověď na výzkumnou otázku: **Výrazný růst HDP v daném roce nemá přímý vliv na růst mezd a cen potravin v totožném a následujícím roce.**