Průvodní listina

1. Úvod do projektu a popsání zadání

Tento projekt se zaměřuje na analýzu ekonomických dat za účelem zodpovězení otázek, které se týkaji vývoje cen a mezd v České republice mezi lety 2006-2018.

Zadání bylo vytvořit z dostupných dat tabulku, ze které bude možné odpovědět na pět otázek, které popíši ve třetím odstavci. Druhým úkolem bylo vytvořit doplňkovou tabulku s informacemi o ostatních evropských státech ve stejném období.

2. Popis použitých datasetů

Pro tvorbu finální tabulky jsem použil tabulku **czechia_payroll**, která je napojena na doplňkové tabulky jako jsou **czechia_payroll_industry_branch**, kde jsou napsány kódy a názvy jednotlivých odvětví, které přiléhají k daným kódům. Dále **czechia_payroll_calculation, czechia_payroll_unit, czechia_payroll_value_type**, které obsahují vysvětlení kódů z tabulky **czechia_payroll**. Tyto tabulky slouží k datům o platech v ČR.

Pro data o cenách potravin v ČR jsem použil tabulky **czechia_price** a **czechia_price_category.** Jako poslední jsem použil tabulku **economies** pro zisk hodnoty HDP v různých letech.

K vytvoření sekundární tabulky jsem použil datasety **countries** a **economies**.

3. Popis metod př tvorbě tabulek

K tvorbě primární tabulky jsem došel následovně. Nejdříve jsem si vytvořil view, které obsahuje průměrný plat v určitém roce a určitém odvětví.
Přemýšlel jsem, jestli chci zobrazit jen určitý kvartál v roce nebo udělat průměr ze všech čtyř kvartálů pro každý rok. Jelikož mi druhá možnost přišla více obecná a méně zkreslující, jelikož zohledňovala platové výkyvy mezi sezónami,
tak jsem se rozhodl použít tuto druhou možnost. Pro další view o cenách potravin během let jsem udělal průměr ceny ze všech krajů pro každou potravinu.
V posledním kroce jsem spojil obě view dohromady a ještě jsem k nim přidal informaci o HDP Česka během let. Spojil jsem je na základě let pomocí INNER JOINu, takže jsem získal jen záznamy, kde mají všechny tabulky společný rok.

Rozhodoval jsem se jak implementovat PRIMÁRNÍ KLÍČ a jelikož jsem věděl, že tabulku neplánuji škálovat, přidávat do ní data nebo ji jakkoliv upravovat, tak mi přišlo zbytečné plýtvání místem, abych vytvářel sloupec **“id”**, a proto jsem se rozhodl primární klíč vytvořit na kombinaci třech sloupců,

což mi zajišťovalo jednoznačnost a unikánost každého řádku. Vím, že přístup s vytvořením identifikátoru by byl výkonnější, ale vzhledem k mojí situaci mi lepší řešení přišlo použít kompoziční klíč, protože tabulka je stejnak relativně malá a výkonnostní rozdíl bude opravdu zanedbatelný.

K tvorbě sekundární tabulky jsem použil propojení tabulek economies a countries na základě názvu státu opět pomocí INNER JOINu a v omezení při spojování jsem ještě specifikoval, aby stát byl na kontinentu Evropa.
Tím jsem dostal jen státy, které leží geograficky v Evropě. Poté jsem už jen zahrnul sloupce, které byly v zadání.
Při tvorbě primárního klíče na této tabulce jsem použil stejnou filozofii jen jsem musel ještě změnit datový typ sloupce **“country”**, protože mi to vyhazovalo chybu při přidávání klíče.

4. Odpovědi na otázky

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

Pro zjištění, jestli mzdy rostly nebo klesaly jsem se rozhodl použít geometrický průměr. Rozhodl jsem se hlavně kvůli tomu, že je to poměrně složitější postup, než použití vzorce pro složenou roční míru růstu, kde by mi stačila pouze první a poslední hodnota a počet období. Chtěl jsem však výzvu a věděl jsem, že složitější metoda mě o SQL více naučí.

Abych mohl použít vzorec pro gemoetrický průměr potřeboval jsem zjistit procentuální změnu oproti minulému roku. To jsem vyřešil pomocí funkce LAG(), která mi umožnila provádět aritmetické operace mezi řádky. Dále jsem tyto hodnoty potřeboval převést na koeficient růstu označovaný v tabulce jako growth_factor. Poté už jen stačilo dosadit do vzorce pro geometrický průměr a vypočítat tempo růstu. Vše jsem uložil do temporary table.

$$G(x_1, x_2, \ldots, x_n) = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}.$$

Narazil jsem však na problém, protože SQL neumí vzít hodnoty ze sloupce a mezi sebou je vynásobit. To jsem vyřešil využitím matematických pravidel o logaritmech a převedl jsem hodnoty koeficientu růstu na logaritmy a podle pravidla o logaritmech jsem je mohl nyní sečíst funkcí SUM(). To jsem vydělil počtem období a nakonec jsem výsledek převedl zpět aplikací inverzní funkce, což u logaritmu je exponenciální funkce. Výpočet v SQL vypadá takto:

```
ROUND((EXP(SUM(LOG(growth_factor)) / COUNT(*)) - 1) * 100, 2) AS avg_growth
```

Tímto výpočtem jsem získal přehled o průměrném ročním růstu pro různá odvětví.

ODPOVĚĎ:

Jelikož všechna čísla o průměrném ročním růstu byla u všech odvětví kladná, tak to znamená, že i když mezi lety byly občas poklesy mezd, tak celkově mzdy ve všech odvětví rostly.

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

Tuto otázku jsem zpracoval tak, že jsem vytvořil dvě CTE, které ukazují průměrné mzdy a ceny mléka a chleba za roky 2006 a 2018, protože to jsou první a poslední dostupné roky, které v datasetu jsou. Tyto dvě CTE jsem spojil pomocí JOIN dohromady a vytvořil jsem sloupec purchasing_power, kde je výpočet toho kolik si mohl člověk koupit kilogramů chleba a litrů mléka za průměrnou mzdu v daném roce.

ODPOVĚĎ:

Z tabulky je možné vypozorovat, že v roce 2018 si mohl člověk koupit více litrů mléka kilogramů chleba za průměrnou mzdu než člověk v roce 2006.

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

Zde jsem použil stejný přístup jako v první otázce a to takový, že jsem nejdříve vytvořil temp table a z toho jsem vytvořil dotaz, který vypočítal (úplně stejným způsobem jako v první otázce) geometrický průměr. Výslednou tabulku jsem seřadil vzestupně podle průměrného ročního růstu ceny, abych viděl, která potravina zdražovala nejméně.

ODPOVĚĎ:

Z výsledné tabulky dotazu jsem vypozoroval, že nejméně zdražoval cukr krystalový, který nejenže nejméně zdražoval, ale dokoncé průměrně každý rok zlevňoval.

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

V tomto úkolu jsem pouze spojil pomocí CTE již mnou vytvořené dočasné tabulky, které jsem jen spojil dohromady a vyfiltroval jsem výslednou tabulku tak, aby se mi ukázaly jen řádky, kde růst cen potravin byl větší než růst mezd a kontroloval jsem, jestli je někde záznam, ve kterém rostly ceny o 10 a více procent než mzdy.

ODPOVĚĎ:

Zjistil jsem, že neexistuje záznam mezi lety 2006 a 2018 o tom, že by ceny potravin rostly o 10 a více procent než mzdy.

Otázka č. 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 násdujícím roce výraznějším růstem?

Zde jsem nejdříve, stejně jako v první odpovědi, vytvořil dočasnou tabulku, kde jsem vypočítal procentuální změnu v českém HDP.
Následně jsem pomocí CTE vypočítal průměrný růst skrz všechna odvětví za dané roky a to samé jsem udělal i v cenách potravin. Vše jsem spojil dohromady pomocí JOIN s tabulkou pro růst HDP.

ODPOVĚĎ:

Když jsem se podíval na roky 2008 a 2009, kde byl výrazný pokles HDP, tak jsem zjistil, že klesaly i ceny a mzdy až do roku 2015 po celou dobu toho, co se HDP drželo v klesajícím trendu.

Od roku 2015 až do roku 2018 HDP naopak poměrně výrazně rostlo a lze pozorovat v těchto letech nárůst cen a mezd.

Z toho jsem si tedy vyvodil, že jistá korelace mezi HDP a mzdami a cenami potravin určitě je. Je vidět, že mezi HDP a růstem mezd je korelace vyšší než korelace mezi HDP a růstem cen potravin.

5. Závěr

Z analýzy jsem tedy zjistil a popsal odpovědi na otázky, které byly v zadání.
Vysvětlil jsem postup, kterým jsem došel k tvorbě finálních tabulek.
Naučil jsem se u toho velmi nových věcí a vyzkoušel jsem si své dosavadní schopnosti v SQL.