# Engeto – Datová analytika – SQL projekt

## Obsah

# Zadání

**Úvod do projektu**

Na vašem analytickém oddělení nezávislé společnosti, která se zabývá životní úrovní občanů, jste se dohodli, že se pokusíte odpovědět na pár definovaných výzkumných otázek, které adresují dostupnost základních potravin široké veřejnosti. Kolegové již vydefinovali základní otázky, na které se pokusí odpovědět a poskytnout tuto informaci tiskovému oddělení. Toto oddělení bude výsledky prezentovat na následující konferenci zaměřené na tuto oblast.

Potřebují k tomu od vás připravit robustní datové podklady, ve kterých bude možné vidět porovnání dostupnosti potravin na základě průměrných příjmů za určité časové období.

Jako dodatečný materiál připravte i tabulku s HDP, GINI koeficientem a populací dalších evropských států ve stejném období, jako primární přehled pro ČR.

Datové sady, které je možné použít pro získání vhodného datového podkladu

Primární tabulky:

- czechia_payroll – Informace o mzdách v různých odvětvích za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.
- czechia_payroll_calculation – Číselník kalkulací v tabulce mezd.
- czechia_payroll_industry_branch – Číselník odvětví v tabulce mezd.
- czechia_payroll_unit – Číselník jednotek hodnot v tabulce mezd.
- czechia_payroll_value_type – Číselník typů hodnot v tabulce mezd.
- czechia_price – Informace o cenách vybraných potravin za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.
- czechia_price_category – Číselník kategorií potravin, které se vyskytují v našem přehledu.

Číselníky sdílených informací o ČR:

- czechia_region – Číselník krajů České republiky dle normy CZ-NUTS 2.
- czechia_district – Číselník okresů České republiky dle normy LAU.

Dodatečné tabulky:

- countries - Všemožné informace o zemích na světě, například hlavní město, měna, národní jídlo nebo průměrná výška populace.
- economies - HDP, GINI, daňová zátěž, atd. pro daný stát a rok.

**Výstup projektu**

Pomozte kolegům s daným úkolem. Výstupem by měly být dvě tabulky v databázi, ze kterých se požadovaná data dají získat. Tabulky pojmenujte t_{jmeno}_{prijmeni}_project_SQL_primary_final (pro data mezd a cen potravin za Českou republiku sjednocených na totožné porovnatelné období – společné roky) a t_{jmeno}_{prijmeni}_project_SQL_secondary_final (pro dodatečná data o dalších evropských státech).

Dále připravte sadu SQL, které z vámi připravených tabulek získají datový podklad k odpovězení na vytyčené výzkumné otázky. Pozor, otázky/hypotézy mohou vaše výstupy podporovat i vyvracet! Záleží na tom, co říkají data.

**Výzkumné otázky**

- Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?
- Kolik je možné si koupit litrů mléka a kilogramů chleba za první a poslední srovnatelné období v dostupných datech cen a mezd?
- Která kategorie potravin zdražuje nejpomaleji (je u ní nejnižší percentuální meziroční nárůst)?
- Existuje rok, ve kterém byl meziroční nárůst cen potravin výrazně vyšší než růst mezd (větší než 10 %)?
- 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?

## Zpětná vazba – odevzdání první verze

**Co se mi líbilo:**

Rezervovaná slova v SQL dotazech máš kapitálkami, to zlepšuje přehlednost a čitelnost dotazu.
Finální data (primary i secondary) jsou v požadované struktuře.
SQL jsou dobře formátované.
Doprovdoný Excel je fajn, ale prosím o jeho okomentování, co je kde - docela jsem se v něm ztratil.

**Co by jsi měl/a zlepšit:**

Celkově mi projekt přijde hodně nepřehledný. Kdybych neměl kontext zadání, tak vůbec nepochopím, co to je. Prosím dej tomu nějaký úvod a představ čtenáři, co k tomu přijde prvně o co se jedná.
Absentuje průvodní listina s interpretací odpovědí. Musíš slovně vysvětli tvé závěry. To ti tam úplně chybí. Prosím o tvůj komentář ke každému výsledku u výzkumných otázek a také pokud byly nějaké problémy s daty nebo postupem, tak se k tomu vyjádři. Toto je zásadní věc, která musí v projektu být.
Komentáře uvozené "netuším.." moc nechápu. Stůj si za svým řešením a odůvodni strukturu výsledu jinak než, že netušíš.
https://github.com/milandurek/engeto_datova_akademie_1_projekt-/blob/main/engeto_datova_akademie_1_projekt.sql#L211 pomocné dotazy můžeš z finálního scriptu odstranit.
Jednotlivé dotazy týkající se výzkumných otázek by mohly být v separátních souborch. Myslím, že by to projekt udělalo přehlednějším.

**Závěr:**

Díky za odevzdání první verze. Prosím o zapracování připomínek v komentářích a odevzdání znovu.
Pak se na to kouknu ještě jednou Měj se, Matěj

## Vytvoření tabulek primary a secondary

### Vytvoření primary tabulky (řádky 1-100)

Při tvorbě tabulky jsem vycházel ze zdrojových tabulek:

- czechia_payroll – Informace o mzdách v různých odvětvích za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.
- czechia_payroll_calculation – Číselník kalkulací v tabulce mezd.
- czechia_payroll_industry_branch – Číselník odvětví v tabulce mezd.
- czechia_payroll_unit – Číselník jednotek hodnot v tabulce mezd.
- czechia_payroll_value_type – Číselník typů hodnot v tabulce mezd.
- czechia_price – Informace o cenách vybraných potravin za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR.
- czechia_price_category – Číselník kategorií potravin, které se vyskytují v našem přehledu.

Pro přehlednější práci s těmito tabulkami jsem si vytvořil excelovou tabulku, ve které jsem si napsal specifika jednotlivých tabulek a vzájemné vazby mezi tabulkami. Excelová tabulka je přílohou tohoto projektu (Engeto - finální projekt - popis tabulek).

Tím, že výstupní primarty tabulka v sobě spojuje dvoje rozdílná data (mzdy a ceny potravin). Tak jsem se rozhodl vytvořit nejprve dvě pomocné tabulky, kde následně došlo k jejich spojení do jedné finální tabulky.

První pomocná tabulka se jmenuje pomocná mzdy. Tato tabulka vznikla jako výstup SQL výrazů z řádků 1-27. Při vytváření tabulky bylo nutné zohledit specifika jednotlivých zdrojových tabulek, které jsem vhodně ošetřil v podmínce WHERE. Dále jsem do této tabulky přidal nový sloupec source_table, aby bylo možné jednoduše odlišit o jaká data se jedná (řádky 22-27).

Druhá pomocná tabulka se jmenuje pomocná potraviny. Tato tabulka vznikla jako výstup SQL výrazů z řádků 42-74. Při vytváření tabulky bylo nutné zohledit specifika jednotlivých zdrojových tabulek, kde bylo zejména nutné upravit zdrojovou jednotku g a kg na společnou jednotku. Data v první pomocné tabulce nabízeli maximální možný rozpad do úrovně rok – čtvrtletí. Zatímco data v druhé pomocné tabulce nabízeli maximální možný rozpad od konkténího data platnosti od – do. Data potravin jsem tedy přizpůsobil datům z tabulky mezd na úroveň rok – čtvrtletí. Vycházel jsem z předpodkladu, že mě zajímá datum platnosti od, kde jsem určil do jakého čtvrletí toto datum patří. Dále jsem do této tabulky přidal nový sloupec source_table, aby bylo možné jednoduše odlišit o jaká data se jedná (řádky 68-74).

V dalším kroku došlo k finálním spojení první pomocné tabulky (mzdy) a druhé pomocné tabulky (potraviny) do tabulky primary. Tato tabulka vznikla jako výstup SQL výrazů z řádků 85-97.

Pro moji větší přehlednost jsem se rozhodl ponechat pomocné výrazy, které jsem pro větší přehlednost zakomentoval.

## Vytvoření secondary tabulky (řádky 110-140)

Při tvorbě tabulky jsem vycházel ze zdrojových tabulek:

- countries - Všemožné informace o zemích na světě, například hlavní město, měna, národní jídlo nebo průměrná výška populace.
- economies - HDP, GINI, daňová zátěž, atd. pro daný stát a rok.

Z dostupných dat z tabulek countries a economies jsem vytvořil tabulku výše HDP v jednotlivých regionech světa podle let. Při vytváření tabulky bylo nutné zohledit specifika jednotlivých zdrojových tabulek, které jsem vhodně ošetřil v podmínce WHERE. Tato tabulka vznikla jako výstup SQL výrazů z řádků 110-138.

Pro moji větší přehlednost jsem se rozhodl ponechat pomocné výrazy, které jsem pro větší přehlednost zakomentoval.

## Odpovědi na výzkumné otázky

Pro zodpovězení otázek jsem se rozhodl data z daného SQL výrazu z programu Dbeaver vyexportoval ve formátu csv a data pro větší přehlednost zpracoval v přiloženém excelu (Engeto - SQL projekt - výstup – otázky). Data jsou zde uložena v jednotlivých listech. Kdy list s názvem otázka obsahuje zdrojová data z Dbeaveru a list s názvem odpověď výstupní graf případně tabulku k jednotlivé otázce.

### Otázka 1- Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?

Pro zodpovězení této otázky jsem si připravil SQL výraz na řádcích 152-171, který vychází z tabulky primary. V tomto výrazu jsem vytáhl z tabulky primary pouze údaje s výší mezd, které jsem spojil podle jednotlivých let a oborů. Tedy zprůměroval jsem čtvrtletí v daném roce. Tyto průměry jsem vytáhl jak pro fyzické a tak pro přepočtené mzdy.

Výstupní data jsou v rozmezí let 2000 až 2021 v celkem 19 oborech. Z výstupu je patrné, že průměrné mzdy v ČR neustále rostou. Výjma specifických let 2012 a 2013, kde došlo k mírnému poklesu u oborů s nejvyššími průměrnými příjmy a stagnaci u ostatních oborů. Jednotlivé obory si drží +/- v průběhu let svoji pozici oproti ostatním oborům. Výjma zdravotnictví, které v zažilo výrazný růst v letech 2019 až 2021, které je zřejmě způsobeno odměnami zdravotníkům v rámci epidemie COVID19.

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

Pro zodpovězení této otázky jsem si připravil SQL výrazy na řádcích 174-193, které vychází z tabulky primary. V prvním výrazu jsem vytáhl z tabulky primary pouze údaje s výší mezd, které jsem spojil podle jednotlivých let. Tedy zprůměroval jsem čtvrtletí v daném roce. V druhém výrazu jsem vytáhl průměrnou cenu chleba a mléka podle jednotlivých let. Tedy stejně jako u mezd jsem zprůměroval čtvrtletí v daném roce.

Výstupní data jsou u chleba a mléka v rozmezí let 2006 až 2018. U mezd v rozmezí 2000 až 2021. Tedy průnik dat je v letech 2006 až 2018. Z výstupu je patrné, že bylo možné si v roce 2006 koupit za průměrnou mzdu 1285 kg chleba a 1438 litrů mléka. Oproti tomu v roce 2018 bylo možné si koupit 1341 kg chleba a 1639 litrů mléka.

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

Pro zodpovězení této otázky jsem si připravil SQL výrazy na řádcích 196-216, které vychází z tabulky primary. V prvním výrazu jsem vytáhl z tabulky primary průměrnou cenu potraviny seskupenou podle jednotlivých let. Tedy zprůměroval jsem čtvrtletí v daném roce. V druhém výrazu jsem provedl to stejné s tím, že jsem daný rok odskočil o jeden rok, abych si připravil data pro porovnání mezi sebou. Data jsem dále v excelu pomocí Vlookup funkce spojil do jedné tabulky.

Výstupní data jsem zobrazil v kontingenční tabulce, kde jsem podmíněným formátováním zdůraznil maximální a minimální meziroční nárůst průměrných cen potravin. Z výstupu je patrné, že nejvyššího meziročního nárustu dosáhly papriky v letech 2006-2007 (+ 94%). Nejnižšího meziročního nárůstu (poklesu) dostáhla Rajská jablka červená kulatá v letech 2006-2007 (-30%)

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

Pro zodpovězení této otázky jsem si připravil SQL výrazy na řádcích 218-232, které vychází z tabulky primary. V prvním výrazu jsem vytáhl z tabulky primary průměrnou cenu potravin a mezd seskupenou podle jednotlivých let. Tedy zprůměroval jsem čtvrtletí v daném roce. V druhém výrazu jsem provedl to stejné s tím, že jsem daný rok odskočil o jeden rok, abych si připravil data pro porovnání mezi sebou. Data jsem dále v excelu pomocí Vlookup funkce spojil do jedné tabulky.

Výstupní data jsem zobrazil v tabulce, kde patrné, že v letech 2011-2012 došlo k meziročnímu průměrnému zdražení potravin o 14% a o meziroční průměrnému zvýšení mezd o 3%. Došlo tedy o 11% vyššímu zdražení potravin než růstu mezd.

## Otázka 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?

Pro zodpovězení této otázky jsem si připravil SQL výrazy na řádcích 235-261, které vychází z tabulky primary a secondary. V prvním výrazu jsem výtahl výši HDP podle let a jednotlivých regionů z tabulky secondary. V druhém výrazu jsem provedl to stejné s tím, že jsem daný rok odskočil o jeden rok, abych si připravil data pro porovnání mezi sebou.

V třetím výrazu jsem vytáhl z tabulky primary průměrnou cenu potraviny seskupenou podle jednotlivých let. Tedy zprůměroval jsem čtvrtletí v daném roce. Ve čtvtém výrazu jsem provedl to stejné s tím, že jsem daný rok odskočil o jeden rok, abych si připravil data pro porovnání mezi sebou. Data jsem dále v excelu pomocí Vlookup funkce spojil do jedné tabulky.

Výstupní data výše HDP v jednotlivých regionech jsou v rozmezí 1960 až 2020. Výstupní data pro průměrné ceny potravin v ČR jsou v rozmezí let 2006 až 2018. Tedy průnik dat je v letech 2006 až 2018. Výstupní data jsem zobrazil v kontingenční tabulce, kde jsem podmíněným formátováním zdůraznil maximální a minimální meziroční nárůst průměrných cen potravin oproti růstu HDP v regionech světa. Z výsledků je patrné, že průměrná cena sledovaných potravin vyskočila v letech 2011-2012 a 2016-2017.