DATOVÁ AKADEMIE BY ENGETO	**Termín kurzu: 26.10.2023 - 25.01.2024**
Project 1	**PROJECT SQL**
Účastník: Lenka Kuligová	**GitHub: LeniK1206**
Datum zpracování: 30.11.2023	

1. POPIS PROJEKTU

Pro přípravu prezentace na následující konferenci vztahující se k životní úrovni občanů je potřeba dodat tiskovému oddělení několik dat o dostupnosti základních potravin široké veřejnosti. Za tímto účelem kolegové z analytického oddělení vydefinovali následujících 5 otázek, na které je potřeba odpovědět.

1. Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?
2. Kolik je možné si koupit litrů mléka a kilogramů chleba za první a poslední srovnatelné období v dostupných datech cen a mezd?
3. Která kategorie potravin zdražuje nejpomaleji (je u ní nejnižší procentuální meziroční nárůst)?
4. Existuje 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?

Pro získání odpovědí na výše uvedené dotazy je potřeba zpracovat 2 tabulky, a to:

1. Primární tabulku obsahující data o mzdách a cenách potravin v České republice za srovnatelné období: *t_lenka_kuligova_project_SQL_primary_final*
2. Sekundární tabulku obsahující HDP, GINI koeficient a populaci dalších evropských států za stejné období jako je zpracováno v primární tabulce: *t_lenka_kuligova_project_SQL_secondary_final*

Vstupní data pro získání datového podkladu jsou k dispozici v tabulkách:

czechia_payroll	czechia_price	economies
czechia_payroll_calculation	czechia_price_category	countries
czechia_payroll_industry_branch		
czechia_payroll_unit	czechia_region	
czechia_payroll_value_type	czechia_district	

Výstupem bude 7 sad SQL, a to 2 sady pro tabulky a 5 sad pro získání datového podkladu k zodpovězení na uvedené dotazy.

2. POPIS PRIMÁRNÍ A SEKUNDÁRNÍ TABULKY

2.1. PRIMARY TABLE

2.1.1. POUŽITÉ ZDROJOVÉ TABULKY

1. **TABLE czechia_payroll cpa**

   Columns použité pro primary table:
   - **value:** jeden z ukazatelů, který má být hodnocen (POZOR: sloupec může nabývat NULL, nutno prověřit pro sledované období – dále viz bod 2.1.2)
   - **value_type_code:** potřebný pro specifikaci, že budu vybírat údaj o průměrných mzdách, tj. `WHERE cpa.value_type_code = 5958`
   - **calculation_code:** pro analýzu bude použita průměrná mzda uvedená pro přepočtený počet osob, protože její výpočet zohledňuje pracovní úvazky `WHERE cpa.calculation_code = 200`
   - **industry_branch_code**: údaj potřebný pro řešení otázky 1
   - **payroll_year:** údaj potřebný pro meziroční srovnání a spojování tabulek

   Columns nepoužité pro primary table:
   - **id:** nepotřebný údaj
   - **unit_code**: protože volím value_type_code, tak automaticky volím unit_code = 200 (tj. Kč)
   - **payroll_quarter**: při řešení otázek bude posuzována průměrná mzda na roční bázi, přičemž pro její získání bude použita agregační funkce avg()

2. **TABLE czechia_payroll_industry_branch cpib**
   - **name:** pomocí left join (přes sloupec code) bude k údajům v table czechia_payroll doplněn i název příslušných odvětví (pro přehlednost a možnost komentování)

3. **TABLE czechia_payroll_unit cpu**
   - **name:** pomocí left join (přes sloupec code) bude k údajům v table czechia_payroll doplněna měrná jednotka pro value, resp. avg_salary)

4. **TABLE czechia_price cpr**

   Columns použité pro primary table:
   - **value:** jeden z ukazatelů, který má být hodnocen (POZOR: sloupec může nabývat NULL, nutno prověřit pro sledované období – dále viz bod 2.1.2)
   - **category_code**: údaj potřebný pro řešení otázky 3 + ceny je vhodné analyzovat na úrovni kategorií potravin, neboť jsou mezi nimi velké cenové rozestupy
   - **date_to:** potřebný pro meziroční srovnání a spojování tabulek

   Columns nepoužité pro primary table:
   - **id:** nepotřebný údaj
   - **date_to**: protože měřené hodnoty jsou v týdenních intervalech a vždy končí kolem 16.12. daného roku, neexistuje kombinace, kdy by date_from = y a date_to = y+1 → stačí tak pracovat pouze s údajem date_from
   - **region_code**: při výpočtu průměrných cen potravin abstrahováno od tohoto údaje, neboť není potřeba k řešení některé z úloh a v tabulce mezd nejsou údaje za jednotlivé regiony, že by bylo pak adekvátní pracovat s regionem na úrovni obou dat (mezd a cen)

5. **TABLE czechia_price_category cpc**
   - **name:** pomocí left join (přes sloupec code) bude k údajům v table czechia_price doplněn i název příslušných potravin (pro přehlednost a možnost komentování)
   - **price_code / price_unit:** doplňující informace o měrných jednotkách jednotlivých potravin (+ použití v řešení úkolu 2)

2.1.2. STRUKTURA PRIMARY TABLE

value_type	year	value	category_code	name	unit	unit_code
200	Jednotlivé roky za srovnatelné období	Prostý průměr za daný rok (viz níže bod Value)	Kód odvětví	Název odvětví	Množství k měrné jednotce	Měrná jednotka
100	Jednotlivé roky za srovnatelné období	Prostý průměr za daný rok (viz níže bod Value)	Kód potravin	Název potravin	Množství k měrné jednotce	Měrná jednotka

Popis sloupců:

- *Value_type:*
  200 – údaje o mzdách v jednotlivých odvětvích (data ve sloupcích naplněna z table 1-3)
  100 – údaje o cenách kategorií potravin (data ve sloupcích naplněna z table 4-5)
- *Year:*

	czechia_payroll	**czechia_price**
Min (year)	2000	2006 (od 02.01.)
Max (year)	2021	2018 (cca do 16.12. každý rok)

  → **srovnatelné období 2006 - 2018**

- *Value:*

	Value_type = 200	Value_type = 100
Úplnost dat	Za sledované období jsou v czechia_payroll uvedeny hodnoty za každý kvartál a každé odvětví.	Za sledované období jsou v czechia_price uvedeny ceny na týdenní bázi, přičemž u některých druhů potravin (např. kapr) probíhá sledování jen v části roku.
Zjednodušení	Uvedeny hodnoty i pro odvětví = NULL → protože není znám původ (zda se nemixují různé další obory v meziročních hodnotách), bylo abstrahováno od této NULL kategorie.	Category_code 212101 má měření až od roku 2015, tj. pouze 4 roky sledovaného období → vyloučení z analýzy.
Výpočet hodnoty	Ve zdrojové tabulce czechia_payroll nejsou naplněny hodnoty pro průměrný počet zaměstnaných osob (value_type_code = 316) za většinu let / kvartály / odvětví, tj. není možné počítat vážený průměr, ale pracovat pouze s prostým.	Prostý průměr za daný rok pro danou kategorii potravin (abstrahováno od regionů).

Počet řádků v primary table:

- v table czechia_payroll je 19 odvětví (vyjma NULL)
- v table czechia_price je 26 potravin (vyjma 212101)

→ **celkový počet záznamů v primary table 585** ((19 odvětví + 26 potravin) * 13 let)

2.2. SECONDARY TABLE

2.2.1. POUŽITÉ ZDROJOVÉ TABULKY

1. TABLE countries c

Columns použité pro secondary table:

- **country:** pro podmínku vytvoření secondary table (v subselectu)
- **continent:** pro zpracování dodatečné sady dat o HDP, GINI a populaci **evropských států**
  ```
  WHERE continent = 'Europe'
  ```

  Pozn: hodnoty v daném sloupci můžou nabývat NULL – po selectu `WHERE continent IS NULL` jsou na výstupu 4 země bez uvedení kontinentu (Guernsey - Isle of Man – Jersey - Timor-Leste) → check, zda tyto země mají data v table economies:
  ```
  SELECT DISTINCT country
  FROM economies e
  WHERE
        country IN (
              SELECT c.country
              FROM countries c
              WHERE c.continent IS NULL)
        AND e.year BETWEEN 2006 AND 2018;
  ```
  → na výstupu 2 země (Isle of Man – Timor-Leste), přičemž na evropský kontinent spadá Isle of Man
  → přidán do podmínky pro vytvoření secondary table

Columns nepoužité pro secondary table:

- **population:** přestože je v této tabulce údaj o population, vztahuje se pouze k jednomu časovému období (rok 2018), ale secondary table bude zpracovávána za roky 2006 – 2018 (viz primary table), tj. pro každý rok je potřeba mít i aktuální hodnotu, která je k dispozici v table economies
- **všechny ostatní**

2. TABLE economies e

Columns použité pro secondary table:

- **country:** pro řešení otázky 5 vytáhnu data o HDP pouze pro „Czech Republic“
  ```
  WHERE country = 'Czech Republic'
  ```
- **year:** potřebný pro spojování s primary table a meziroční srovnání v otázce 5
  Pozn: data jsou zde v rozmezí let min(year) = 1960, max(year) = 2020, nicméně secondary table má být za stejné období jako primary table, tj. při vytváření tabulky bude omezeno:
  ```
  WHERE e.year BETWEEN 2006 AND 2018
  ```
- **GDP:** jeden z ukazatelů, který má být obsažen v secondary table a zároveň bude použit pro řešení otázky 5
  Pozn. jedná se o číslo s desetinnými místy – při zaokrouhlení na celé číslo nedojde k výraznému zkreslení při výpočtu %-ní meziroční změny v otázce 5 → použití funkce round()
- **GINI:** jeden z ukazatelů, který má být obsažen v secondary table (dále se s ním nikde nepracuje)
- **population:** jeden z ukazatelů, který má být obsažen v secondary table (dále se s ním nikde nepracuje

Columns nepoužité pro secondary table:

- **taxes**
- **fertility**
- **mortality_under5**

2.2.2. STRUKTURA SECONDARY TABLE

country	**year**	**GDP**	**gini**	**population**
Všechny evropské státy	Roky za srovnatelné období jako primary table, tj. 2006 - 2018	Hodnota zaokrouhlená na celá čísla	Hodnota převzatá z table economies	Hodnota převzatá z table economies

Počet řádků v secondary table:

- v table countries je 48 zemí s kontinentem Europe a 4 země, které nemají vyplněný kontinent
- v table economies jsou data za sledované období (2006-2018) pro 45 evropských zemí a pro 2 země bez vyplněného kontinentu (Isle of Man (Europe) – Timor-Leste (Asia))

→ **celkový počet záznamů v secondary table 598** (46 evropských zemí * 13 let)

3. OTÁZKY A ODPOVĚDI

Poznámka na úvod: při hledání odpovědí na položené otázky bylo k analýzám mezd a cen přistupováno následovně:

- Analýza mezd: protože průměrná mzda je běžný údaj zveřejňovaný za ekonomiku jako celek, bylo tam, kde odpověď nevyžadovala porovnání mezi jednotlivými odvětvími, od odvětví abstrahováno, tj. použit průměr. Protože ve zdrojových datech nebyly naplněny průměrné počty zaměstnanců v jednotlivých oborech, jedná se o prostý průměr, který může být méně přesný než průměr vážený.
- Analýza cen: vždy jsou zohledňovány jednotlivé kategorie potravin, při jejich abstrahování by průměrná cena za daný rok představovala nekorektní číslo pro interpretaci výsledků.

3.1. Otázka 1

Znění:
Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?

Odpověď:
Ve sledovaném období 2006 – 2018 rostly každý rok mzdy pouze v odvětvích:

- Zdravotní a sociální péče,
- Zpracovatelský průmysl,
- Ostatní činnost.

V ostatních odvětvích byl minimálně jeden rok, kdy došlo k poklesu průměrných mezd.

name	years_of_growth
Zdravotní a sociální péče	12
Zpracovatelský průmysl	12
Ostatní činnosti	12
Stavebnictví	11
Velkoobchod a maloobchod; opravy a údržba motorových vozidel	11
Doprava a skladování	11
Informační a komunikační činnosti	11
Peněžnictví a pojišťovnictví	11
Administrativní a podpůrné činnosti	11
Vzdělávání	11
Zemědělství, lesnictví, rybářství	11
Kulturní, zábavní a rekreační činnosti	11
Zásobování vodou; činnosti související s odpady a sanacemi	11
Ubytování, stravování a pohostinství	10
Činnosti v oblasti nemovitostí	10
Profesní, vědecké a technické činnosti	10
Veřejná správa a obrana; povinné sociální zabezpečení	10
Výroba a rozvod elektřiny, plynu, tepla a klimatiz. vzduchu	9
Těžba a dobývání	8

Years_of_growth: počet kolikrát se ve srovnatelném období 2006-2018 navyšovaly mzdy v daném odvětví, přičemž 12 je max

3.2. Otázka 2

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

Odpověď:
V prvním roce sledovaného období (2006) je možné koupit si 15,755 kg chleba nebo 17,588 l mléka.
V posledním roce sledovaného období (2018) je možné koupit si 16,382 kg chleba nebo 20,035 l mléka.

year	pcs	unit_code	name
2006	15,755	kg	Chléb konzumní kmínový
2018	16,382	kg	Chléb konzumní kmínový
2006	17,588	l	Mléko polotučné pasterované
2018	20,035	l	Mléko polotučné pasterované

3.3. Otázka 3

Znění:
Která kategorie potravin zdražuje nejpomaleji (je u ní nejnižší procentuální meziroční nárůst)?

Odpověď:
Z výstupních dat plyne, že z kategorií potravin, jejichž cena ve srovnávacím období rostla, zdražovaly nejpomaleji banány žluté (7,4%). Mezi kategoriemi potravin jsou však 2 kategorie, a to cukr krystal a rajská jablka červená kulatá, jejichž cena ve srovnávacím období dokonce klesla, tj. potraviny zlevňovaly.

name	price_last_year	price_first_year	price_growth
Cukr krystalový	15,75	21,73	-27,52
Rajská jablka červená kulatá	44,49	57,83	-23,07
Banány žluté	29,32	27,30	7,40
Vepřová pečeně s kostí	116,85	105,18	11,10
Přírodní minerální voda uhličitá	8,64	7,69	12,35
Pečivo pšeničné bílé	43,84	38,60	13,58
Jablka konzumní	36,17	30,71	17,78
Šunkový salám	144,88	116,77	24,07
Konzumní brambory	15,08	12,07	24,94
Eidamská cihla	142,45	110,95	28,39
Hovězí maso zadní bez kosti	223,26	166,34	34,22
Kapr živý	93,46	69,35	34,77
Mléko polotučné pasterované	19,82	14,44	37,26
Pivo výčepní, světlé, lahvové	11,81	8,45	39,76
Rostlinný roztíratelný tuk	99,40	69,45	43,12
Kuřata kuchaná celá	69,31	47,47	46,01
Pomeranče	36,50	24,73	47,59
Chléb konzumní kmínový	24,24	16,12	50,37
Pšeničná mouka hladká	11,44	7,41	54,39
Mrkev	22,45	14,41	55,79
Jogurt bílý netučný	9,17	5,83	57,29
Vejce slepičí čerstvá	38,39	23,49	63,43
Rýže loupaná dlouhozrnná	36,18	21,29	69,94
Papriky	60,47	35,31	71,25
Těstoviny vaječné	47,87	26,10	83,41
Máslo	207,08	104,39	98,37

price_last_year: průměrná cena kategorie potravin v posledním roce srovnávacího období, tj. 2018
price_first_year: průměrná cena kategorie potravin v prvním roce srovnávacího období, tj. 2006
price_growth: procentuální nárůst cen kategorií potravin mezi prvním a posledním rokem srovnávacího období

3.4. Otázka 4

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

Poznámka k odpovědi: Ačkoliv otázka zní obecně co se nárůstu cen potravin, odpověď na ni respektuje jednotlivé kategorie potravin, a to z důvodu, že mezi nimi existuje velký cenový rozptyl a každá kategorie se v oblasti cen chová jinak.

Odpověď: Neexistuje žádný rok, ve kterém by u všech kategorií potravin došlo k meziročnímu nárůstu cen o více než 10%. Nejvíce kategorií potravin zaznamenalo nárůst cen nad 10% v letech 2007 a 2008 (celkem 12 z 26 sledovaných).

year	category_count
2007	12
2008	12
2011	9
2017	8
2013	6
2012	5
2010	4
2014	2
2018	2
2009	1
2015	1
2016	1

category_count: počet kategorií potravin, u kterých byl v daném roce meziroční nárůst cen větší než 10%

3.5. Otázka 5

Znění:
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?

Poznámka k odpovědím: Aby bylo možné porovnat vliv ve stejném nebo následujícím roce, nejsou ve výsledku obsažena data pro rok 2006 a 2018, neboť pro rok 2006 nejsou k dispozici data o mzdách / cenách 2005 (vliv ve stejném roce) a pro rok 2018 nejsou k dispozici data o mzdách / cenách 2019 (vliv v následujícím roce).

A/ vliv HDP na mzdy

Odpověď: Na základě dostupných výsledků nelze za sledované období potvrdit hypotézu, že by změna HDP měla za následek výrazný nárůst průměrných mezd ve stejném nebo následujícím roce. Tento trend se projevil pouze v letech 2007, 2008, 2011, 2014, 2014 a 2017. V letech 2010 a 2015 byl růst HDP doprovázen růstem průměrných mezd, nicméně změna HDP byla vyšší. V letech 2009, 2012 a 2013 došlo k poklesu HDP a v těchto letech průměrné mzdy klesaly také nebo naopak rostly.

year	GDP_growth	salary_growth_ current_year	salary_growth_ next_year	impact_GDP_confirmation
2017	5,17	6,17	7,70	1
2016	2,54	3,64	6,17	1
2014	2,26	2,57	2,60	1
2011	1,76	2,33	2,93	1
2008	2,69	7,69	3,07	1
2007	5,57	6,88	7,69	1
2015	5,39	2,60	3,64	0
2010	2,43	1,91	2,33	0
2013	-0,05	-1,56	2,57	-1
2012	-0,79	2,93	-1,56	-1
2009	-4,66	3,07	1,91	-1

GDP_growth: meziroční změna HDP (v %)
salary_growth_current_year: změna průměrných mzdy v daném roce oproti roku předchozímu (v %)
salary_growth_next_year: změna průměrné mzdy v následujícím roce oproti danému roku (v %)
impact_GDP_confirmation: vyhodnocení, zda má změna HDP vliv na změnu mezd, přičemž:

1…růst HDP v daném roce má vliv na výraznější růst průměrných mezd ve stejném nebo následujícím roce
0…růst HDP v daném roce nemá vliv na výraznější růst průměrných mezd ve stejném nebo následujícím roce
-1…v daném roce došlo k poklesu HDP

B/ vliv HDP na ceny potravin

Odpověď: Na základě dostupných výsledků nelze za sledované období potvrdit hypotézu, že by změna HDP měla za následek výrazný nárůst cen potravin ve stejném nebo následujícím roce. Pouze v roce 2011 se projevil nárůst HDP oproti roku 2010 ve výraznějším nárůstu cen 2011 nebo 2012 u 25 kategorií potravin z celkového počtu 26. V ostatních letech nebyl nárůst cen doprovázen za sledované období nebo rok následující výraznějším nárůstem cen (tj. větším než růst HDP). V letech 2009, 2012 a 2013 došlo k poklesu HDP.

year	category_count
2011	25
2010	21
2007	20
2016	20
2008	18
2017	17
2014	11
2015	3
2012	-26
2009	-26
2013	-26

category_count: počet kategorií potravin, u kterých byla změna cen ve sledovaném roce nebo roce následujícím vyšší než meziroční změna HDP (v %)