AUTOR: ERIKA DANKOVÁ

DÁTUM: 25/06/2023

SQL Projekt - Sprievodná listina

1. ÚVOD

K absolvovaniu 12-týdňového kurzu Dátová akadémia poskytovaným spoločnosťou ENGETO s.r.o. je potrebné vypracovať SQL projekt. Na základe jeho vypracovania dochádza k overeniu znalostí jazyka SQL. Obdržali sme 7 primárnych tabuliek, 2 číselníky zdieľaných informácií a 2 dodatočné tabuľky, ktoré sú nevyhnutné k dokončeniu projektu. Cieľom je zodpovedanie 5 výskumných otázok prostredníctvom vytvorenia dotazov na nami pripravené 2 tabuľky.

2. ZADANIE

Na vašom analytickom oddelení nezávislej spoločnosti, ktorá sa zaoberá životnou úrovňou občanov, ste sa dohodli, že sa pokúsite odpovedať na pár definovaných výskumných otázok, ktoré adresujú dostupnosť základných potravín širokej verejnosti. Kolegovia už definovali základné otázky, na ktoré sa pokúsia odpovedať a poskytnúť túto informáciu tlačovému oddeleniu. Toto oddelenie bude výsledky prezentovať na nasledujúcej konferencii zameranej na túto oblasť.

Potrebujú k tomu od vás pripraviť robustní dátové podklady, v ktorých bude možné vidieť porovnanie dostupnosti potravín na základe priemerných príjmov za určité časové obdobie.

Ako dodatočný materiál pripravte aj tabuľku s HDP, GINI koeficientom a populáciou ďalších európskych štátov v rovnakom období, ako primárny prehľad pre ČR.

A) OBDRŽANÉ DÁTOVÉ SADY

PRIMÁRNE TABUĽKY:

1. **czechia_payroll** – Informácie o mzdách v rôznych odvetviach za niekoľkoročné obdobia. Dátová sada pochádza z Portálu otvorených dát ČR.
2. **czechia_payroll_calculation** – Číselník kalkulácií v tabuľke miezd.
3. **czechia_payroll_industry_branch** – Číselník odvetví v tabuľke miezd.
4. **czechia_payroll_unit** – Číselník jednotiek hodnôt v tabuľke miezd.
5. **czechia_payroll_value_type** – Číselník typov hodnôt v tabuľke miezd.
6. **czechia_price** – Informácie o cenách vybraných potravín za niekoľkoročné obdobie. Dátová sada pochádza z Portálu otvorených dát ČR.
7. **czechia_price_category** – Číselník kategórií potravín, ktoré sa vyskytujú v našom prehľade.

ČÍSELNÍKY ZDIEĽANÝCH INFORMÁCIÍ O ČR:

1. **czechia_region** – Číselník krajov Českej republiky podľa normy CZ-NUTS 2.
2. **czechia_district** – Číselník okresov Českej republiky podľa normy LAU.

DODATOČNÉ TABUĽKY:

1. **countries** - Všemožné informácie o krajinách na svete, napríklad hlavné mesto, mena, národné jedlo alebo priemerná výška populácie.
2. **economies** - HDP, GINI, daňová záťaž, atď. pre daný štát a rok.

B) VÝSKUMNÉ OTÁZKY

1. Rastú v priebehu rokov mzdy vo všetkých odvetviach, alebo v niektorých klesajú?
2. Koľko je možné si kúpiť litrov mlieka a kilogramov chleba za prvé a posledné zrovnateľné obdobie v dostupných dátach cien a miezd?
3. Ktorá kategória potravín zdražuje najpomalšie (je u nej najnižší percentuálny medziročný nárast)?
4. Existuje rok, v ktorom bol medziročný nárast cien potravín výrazne vyšší než rast miezd (väčší než 10 %)?
5. Má výška HDP vplyv na zmeny v mzdách a cenách potravín? Alebo, ak HDP vzrastie výraznejšie v jednom roku, prejaví sa to na cenách potravín či mzdách v rovnakom alebo nasledujúcom roku výraznejším rastom?

C) VÝSTUP PROJEKTU

Pomôžte kolegom s danou úlohou. Výstupom by mali byť dve tabuľky v databáze, z ktorých sa požadované dáta dajú získať. Tabuľky pomenujte t_{jmeno}_{prijmeni}_project_SQL_primary_final (pre dáta miezd a cien potravín za Českú republiku zjednotených na totožné porovnateľné obdobie – spoločné roky) a t_{jmeno}_{prijmeni}_project_SQL_secondary_final (pre dodatočné dáta o ďalších európskych štátoch).

Ďalej pripravte sadu sql, ktoré z vami pripravených tabuliek získajú dátový podklad k zodpovedaniu na vytýčené výskumné otázky. Pozor, otázky/hypotézy môžu vaše výstupy podporovať aj vyvracať! záleží na tom, čo hovoria dáta.

Na svojom github účte vytvorte repozitár (môže byť súkromný), kam uložíte všetky informácie k projektu – hlavne sql skript generujúci výslednú tabuľku, popis medzivýsledkov (sprievodnú listinu) a informácie o výstupných dátach (napríklad kde chýbajú hodnoty apod.).

3. RIEŠENIE ZADANIA

A) ZVOLENÝ POSTUP RIEŠENIA

Pred tvorbou tabuliek, je nutné preskúmať obdržané dátové sady. U tých, ktoré pochádzajú z Portálu otvorených dát je priložená dokumentácia, kde je popísaná dátová štruktúra. Následne je potreba prečítať si výskumné otázky, na základe ktorých vyvodíme záver - ktoré dáta budeme potrebovať z obdržaných tabuliek.

Pre overenie správnosti dotazov som používala Excel – porovnala som si výstupy mojich skriptov s riešením prostredníctvom Excelu. Viem, že tento prístup nie je možné aplikovať pri väčšom množstve dát, ale pre začínajúceho dátového analytika je to dobrý nástroj pre pochopenie fungovania jednotlivých dotazov.

B) TVORBA TABULIEK

1. Primárna tabuľka

Primárna tabuľka **t_erika_dankova_project_SQL_primary_final** obsahuje dáta miezd a cien potravín za ČR zjednotených pre spoločné roky – konkrétne 2006 až 2018. Pripojila som aj dáta k zodpovedaniu otázky č.5 – údaj o HDP na obyvateľa a medziročnú zmenu. Navyše, pre jednoduchšie riešenie otázok č. 1 a 3 som doplnila aj stĺpce k medziročnej zmene miezd na úrovni ekonomických činností a k medziročnej zmene cien potravín na úrovni produktu.

Štruktúra primárnej tabuľky

Tabuľka obsahuje 14 nasledovných stĺpcov:

year – rok

category_code – číselné označenie produktu (potraviny)

name – názov produktu (potraviny)

AVG_price – priemerná cena produktu v konkrétnom roku (t) vyjadrená v mene Kč

AVG_price_prev – priemerná cena produktu v predošlom roku (t -1) vyjadrená v mene Kč

YoY_change_price – medziročná zmena ceny produktu vyjadrená v percentách

industry_branch_code – označenie ekonomickej činnosti písmenom

IB_name - názov ekonomickej činnosti

AVG_wage – priemerná hrubá mzda v Kč v konkrétnom roku (t) na úrovni ekonomickej činnosti vypočítaná podľa prepočítaného stavu zamestnancov (nie fyzického stavu)

AVG_wage_prev – priemerná hrubá mzda v Kč v predošlom roku (t-1) na úrovni ekonomickej činnosti vypočítaná podľa prepočítaného stavu zamestnancov (nie fyzického stavu)

YoY_change_wage – medziročná zmena hrubej mzdy na úrovni ekonomickej činnosti vyjadrená v percentách

GDP_per_capita_act – HDP na obyvateľa v ČR pre konkrétny rok (t) v ČR (mena je irelevantná)

GDP_per_capita_prev – HDP na obyvateľa v ČR pre predošlý rok (t-1) v ČR (mena je irelevantná)

YoY_change_GDP – percentuálna medziročná zmena HDP na obyvateľa

Úprava dát:

Pred spojením primárnych (a dodatočných) tabuliek bolo potrebné upraviť dáta filtrovaním:

Tabuľka miezd (**czechia_payroll**):

- v stĺpci **value_type_code** som vyfiltrovala len hodnoty 5958, aby som získala priemernú hrubú mzdu na zamestnanca
- v stĺpci **calculation_code** je vyfiltrovaná hodnota 200, čím som získala len hodnoty miezd podľa prepočítaného stavu zamestnancov
- v stĺpci **industry_branch_code** som vyfiltrovala nenulové hodnoty (NULLové predstavovali priemer za všetky odvetvia)
- filtrovanie stĺpca **payroll_year** – roky 2006 až 2018 (nutné k zjednoteniu s tabuľkou cien, ktorá má dostupné dáta len pre tieto roky)

Tabuľka cien (**czechia_price**):

- v stĺpci **region_code** som vyfiltrovala NULLové hodnoty, ktoré predstavujú priemery za všetky kraje (údaje za jednotlivé kraje nepotrebujeme)
- stĺpec **category_code** – vylúčila som kategóriu 212101 (Jakostní víno bílé) keďže údaje boli dostupné len pre roky 2015-2018

Z tabuľky **economies** som si prepočítala HDP na obyvateľa, keďže HDP je závislé na počte obyvateľov (pozn. tento ukazovateľ má hlavne význam pri porovnávaní HDP viacerých krajín).

2. Sekundárna tabuľka

Sekundárna tabuľka **t_erika_dankova_project_SQL_secondary_final** obsahuje hodnoty HDP, veľkosť populácie, HDP na obyvateľa a giniho koeficient pre európske krajiny za roky 2006 až 2018.

Tabuľka vznikla z pôvodnej tabuľky **economies**, z ktorej som si vyfiltrovala európske krajiny s použitím subselectu s pomocou tabuľky **countries**.

Štruktúra sekundárnej tabuľky

Tabuľka obsahuje 6 stĺpcov:

year – roky 2006 až 2018

country – názov európskych krajín

GDP – výška HDP v konkrétnej krajine a roku

population – počet obyvateľov v konkrétnej krajine a roku

GDP_per_capita – HDP na počet obyvateľov v konkrétnej krajine a roku

gini – giniho koeficient pre konkrétnu krajinu a rok

C) ZODPOVEDANIE VÝSKUMNÝCH OTÁZOK

Táto časť obsahuje zodpovedanie 5 výskumných otázok uvedených v časti **2. ZADANIE. B) VÝSKUMNÉ OTÁZKY**. Každá otázka je zodpovedaná písomne a takisto je priložená tabuľka, ktorá predstavuje výstup z jednotlivých dotazov, ktoré sú v súbore **queries.sql.**

1. Rastú v priebehu rokov mzdy vo všetkých odvetviach, alebo v niektorých klesajú?

Mzdy v priebehu rokov nerastú vo všetkých odvetviach.

K najväčšiemu medziročnému poklesu došlo v roku 2013 v odvetví Peňažníctvo a poisťovníctvo o 8,83%.

year	industry_branch_code	IB_name	AVG_wage	AVG_wage_prev	YoY_change_wage
2013	K	Peněžnictví a pojišťovnictví	46,316.50	50,800.50	-8.83
2013	D	Výroba a rozvod elektřiny, plynu, tepla a klimatiz. vzduchu	40,761.75	42,657.25	-4.44
2013	B	Těžba a dobývání	31,486.50	32,540.25	-3.24
2009	B	Těžba a dobývání	28,360.50	29,272.50	-3.12
2013	M	Profesní, vědecké a technické činnosti	31,824.75	32,816.75	-3.02

Tab. 1 Top 5 záznamov – poklesov v konkrétnom odvetví a roku

K najvýraznejšiemu poklesu došlo v roku 2013 – pokles v tomto roku zaznamenalo 11 odvetví (pozn. celkový počet odvetví je 19). V tomto roku došlo aj k celkovému poklesu priemernej mzdy o 7 968,5 Kč. V ostatných obdobiach došlo k celkovému nárastu priemernej mzdy. Počet záznamov – medziročných poklesov v súčte za obdobie 2018 až 2018 v konkrétnom odvetví je 25.

year	count_decrease	count_increase	difference
2007	0	19	27,658.50
2008	0	19	33,069.25
2009	4	15	14,220.50
2010	3	16	9,126.50
2011	4	15	11,348.50
2012	0	19	14,584.25
2013	11	8	-7,968.50
2014	1	18	12,977.25
2015	1	18	13,429.25
2016	1	18	19,302.00
2017	0	19	33,914.50
2018	0	19	44,937.00

Tab. 2 Počet odvetví – pokles vs. Nárast

2. Koľko je možné si kúpiť litrov mlieka a kilogramov chleba za prvé a posledné zrovnateľné obdobie v dostupných dátach cien a miezd?

Za prvé zrovnateľné obdobie je možné kúpiť 1 312 ks chleba a 1 465 l mlieka.

Za posledné zrovnateľné obdobie je možné kúpiť 1 365 ks chleba a 1 669 l mlieka.

year	category_code	name	AVG_price	AVG_wage	quantity_available
2006	111301	Chléb konzumní kmínový	16.12	21,165.18	1,312
2018	111301	Chléb konzumní kmínový	24.24	33,091.45	1,365
2006	114201	Mléko polotučné pasterované	14.44	21,165.18	1,465
2018	114201	Mléko polotučné pasterované	19.82	33,091.45	1,669

Tab. 3 – Porovnanie množstva chleba a mlieka (2006 vs. 2018)

Poznámka: hodnoty sú zaokrúhlené nadol.

3. Ktorá kategória potravín zdražuje najpomalšie (je u nej najnižší percentuálny medziročný nárast)?

Toto zadanie mi nebolo úplne jasné, môžeme sa na to pozrieť zo 4 hľadísk:

A) Najnižšia percentuálna medziročná zmena ceny potraviny – porovnanie rokov t vs. t+1:

V tomto prípade došlo k najnižšiemu medziročnému nárastu (resp. poklesu) ceny v roku 2007 u potraviny Rajské jablká červené guľaté – o -30,28%.

year	category_code	name	AVG_price	AVG_price_prev	YoY_change_price
2007	117101	Rajská jablka červená kulatá	40.32	57.83	-30.28
2009	111303	Pečivo pšeničné bílé	38.83	54.31	-28.5
2011	117101	Rajská jablka červená kulatá	30.31	42.21	-28.19
2008	117401	Konzumní brambory	10.78	14.1	-23.55
2009	111201	Pšeničná mouka hladká	9.97	12.98	-23.19

Tab. 4 TOP 5 medziročných poklesov cien potravín

B) Najnižšia percentuálna medziročná **KLADNÁ** zmena ceny potraviny – porovnanie rokov t vs. t+1:

K najnižšiemu medziročnému nárastu (kladná zmena) ceny došlo v roku 2009 u potraviny Rastlinný roztierateľný tuk – o 0,01%.

year	category_code	name	AVG_price	AVG_price_prev	YoY_change_price
2009	115201	Rostlinný roztíratelný tuk	84.41	84.4	0.01
2009	112201	Vepřová pečeně s kostí	106.48	106.46	0.02
2007	115201	Rostlinný roztíratelný tuk	69.5	69.45	0.07
2018	111201	Pšenična mouka hladká	11.44	11.43	0.09
2015	122102	Přírodní minerální voda uhličitá	8.7	8.69	0.12

Tab. 5 TOP 5 medziročných nárastov cien potravín

C) Najnižšia percentuálna zmena ceny potraviny – porovnanie rokov 2006 vs. 2018:

K najnižšej percentuálnej zmeny ceny potraviny (porovnanie 1. a posledného obdobia) došlo u potraviny Cukor kryštálový – o -27,52%.

first_year	last_year	category_code	name	AVG_price_2006	AVG_price_2018	price_change
2006	2018	118101	Cukr krystalový	21.73	15.75	-27.52
2006	2018	117101	Rajská jablka červená kulatá	57.83	44.49	-23.07
2006	2018	116103	Banány žluté	27.31	29.32	7.36
2006	2018	112201	Vepřová pečeně s kostí	105.18	116.85	11.1
2006	2018	122102	Přírodní minerální voda uhličitá	7.69	8.65	12.48

Tab. 6 TOP 5 najnižších % zmien cien potravín 2006 vs. 2018

D) Najnižšia percentuálna **KLADNÁ** zmena ceny potraviny – porovnanie rokov 2006 vs. 2018:

K najnižšej percentuálnej nárastu ceny potraviny (porovnanie 1. a posledného obdobia) došlo u potraviny Banány žlté – o 7,36%.

first_year	last_year	category_code	name	AVG_price_2006	AVG_price_2018	price_change
2006	2018	116103	Banány žluté	27.31	29.32	7.36
2006	2018	112201	Vepřová pečeně s kostí	105.18	116.85	11.1
2006	2018	122102	Přírodní minerální voda uhličitá	7.69	8.65	12.48
2006	2018	111303	Pečivo pšeničné bílé	38.6	43.84	13.58
2006	2018	116104	Jablka konzumní	30.71	36.18	17.81

Tab. 7 TOP 5 najnižších % nárastov cien potravín 2006 vs. 2018

Podľa môjho názoru je najpravdepodobnejšia požadovaná odpoveď na otázku č. 3 riešenie **B)**, keďže v otázke je uvedené „zdražuje“ čo značí, že ide o kladnú zmenu a v prípade, ak zadávateľ chcel porovnanie 1. a posledného obdobia tak to rovno uviedol v otázke (viď. Otázka č.2).

4. Existuje rok, v ktorom bol medziročný nárast cien potravín výrazne vyšší než rast miezd (väčší než 10 %)?

Neexistuje rok, v ktorom bol medziročný nárast cien vyšší než rast miezd o viac ako 10%. Maximum bolo dosiahnuté v roku 2013, kedy bol medziročný pokles miezd o 1,56% a medziročný nárast potravín 6,66%.

year	AVG_wage_prev	AVG_wage	YoY_change_wage	AVG_price_prev	AVG_price	YoY_change_price	diff_YoY_change
2013	26,955.05	26,535.66	-1.56	54.3	57.07	5.1	6.66
2017	28,941.37	30,726.34	6.17	55.03	60.6	10.12	3.95
2012	26,187.46	26,955.05	2.93	50.88	54.3	6.72	3.79
2011	25,590.17	26,187.46	2.33	49.23	50.88	3.35	1.02
2010	25,109.83	25,590.17	1.91	48.29	49.23	1.95	0.04
2007	21,165.18	22,620.89	6.88	45.52	48.59	6.74	-0.14
2008	22,620.89	24,361.38	7.69	48.59	51.6	6.19	-1.5
2014	26,535.66	27,218.67	2.57	57.07	57.49	0.74	-1.83
2016	27,925.47	28,941.37	3.64	55.82	55.03	-1.42	-5.06
2015	27,218.67	27,925.47	2.6	57.49	55.82	-2.9	-5.5
2018	30,726.34	33,091.45	7.7	60.6	61.86	2.08	-5.62
2009	24,361.38	25,109.83	3.07	51.6	48.29	-6.41	-9.48

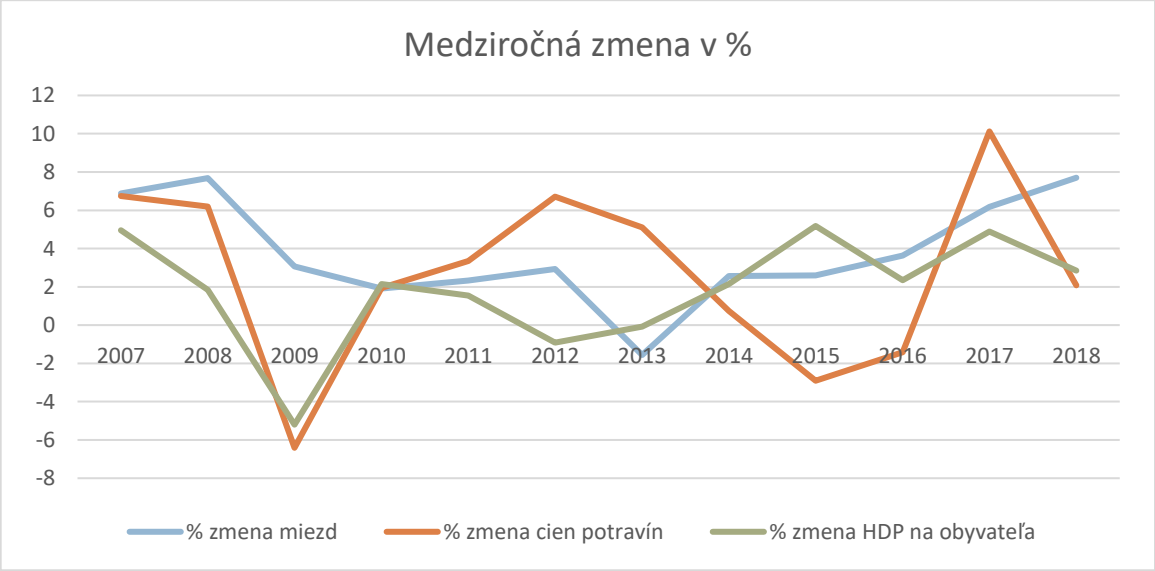
Tab. 8 Porovnanie medziročného rastu miezd a cien potravín

5. Má výška HDP vplyv na zmeny v mzdách a cenách potravín? Alebo, ak HDP vzrastie výraznejšie v jednom roku, prejaví sa to na cenách potravín či mzdách v rovnakom alebo nasledujúcom roku výraznejším rastom?

Na grafe nižšie môžeme vidieť porovnanie medziročnej zmeny HDP na obyvateľa, cien potravín a miezd. Môžeme si všimnúť , že mzdy majú miernejšie výkyvy ako HDP na obyvateľa a ceny potravín.

Počas ekonomickej krízy došlo k vyvrcholeniu poklesu HDP v roku 2009, rovnako došlo aj k poklesu cien potravín, avšak na mzdy to nemalo taký výrazný dopad.

Ďalší pokles HDP bol zaznamenaný v roku 2012, v tomto prípade už došlo k adekvátnemu poklesu miezd, avšak s oneskorením 1 roka. Čo sa týka zmeny cien potravín od roku 2010 sa oneskorila reakcia na zmenu HDP o približne 2 roky.



Graf 1 Medziročná % zmena miezd, cien potravín a HDP na obyvateľa

year	AVG_wage_prev	AVG_wage	YoY_change_wage	AVG_price_prev	AVG_price	YoY_change_price	GDP_per_capita_prev	GDP_per_capita_act	YoY_change_GDP
2007	21,165.18	22,620.89	6.88	45.52	48.59	6.74	19,286.26	20,242.10	4.96
2008	22,620.89	24,361.38	7.69	48.59	51.6	6.19	20,242.10	20,614.20	1.84
2009	24,361.38	25,109.83	3.07	51.6	48.29	-6.41	20,614.20	19,542.47	-5.2
2010	25,109.83	25,590.17	1.91	48.29	49.23	1.95	19,542.47	19,960.07	2.14
2011	25,590.17	26,187.46	2.33	49.23	50.88	3.35	19,960.07	20,269.49	1.55
2012	26,187.46	26,955.05	2.93	50.88	54.3	6.72	20,269.49	20,082.25	-0.92
2013	26,955.05	26,535.66	-1.56	54.3	57.07	5.1	20,082.25	20,066.38	-0.08
2014	26,535.66	27,218.67	2.57	57.07	57.49	0.74	20,066.38	20,498.71	2.15
2015	27,218.67	27,925.47	2.6	57.49	55.82	-2.9	20,498.71	21,560.83	5.18
2016	27,925.47	28,941.37	3.64	55.82	55.03	-1.42	21,560.83	22,065.47	2.34
2017	28,941.37	30,726.34	6.17	55.03	60.6	10.12	22,065.47	23,144.41	4.89
2018	30,726.34	33,091.45	7.7	60.6	61.86	2.08	23,144.41	23,804.98	2.85

Tab. 9 Porovnanie medziročného rastu miezd, cien potravín a HDP na obyvateľa

4. PRÍLOHY

t_erika_dankova_project_SQL_primary_final.sql – vytvorená primárna tabuľka s dátami cien a miezd ČR

t_erika_dankova_project_SQL_secondary_final.sql – vytvorená sekundárna tabuľka s dodatočnými dátami o európskych štátoch

queries.sql – dotazy „SELECT-y“ k zodpovedaniu výskumných otázok