DATOVÁ AKADEMIE – ENGETO

PROJEKT Č. 1

"It is a capital mistake to theorize before one has data."
~ Sherlock Holmes in "A study in Scarlet" by A. C. Doyle

1. Popis projektu

Cílem projektu bylo odpovědět na pět výzkumných otázek, které se zabývají dostupností základních potravin široké veřejnosti. Za tímto účelem bylo potřeba připravit dvě datové tabulky. První (Primary_Final) slouží k porovnání dostupnosti potravin s průměrnými příjmy. Druhá (Secondary_Final) obsahuje základní ekonomické ukazatele pro země světa za stejné období jako tabulka první.

Součásti projektu jsou také scripty PostgreSQL, které slouží k tvorbě výše zmíněných tabulek i k odpovědím na výzkumné otázky.

2. Tvorba tabulek

Tabulky, které jsou jedním z výstupů tohoto projektu, by měly být sestaveny tak, aby pomocí nich bylo možné zodpovědět všechny otázky a výsledný SQL kód vycházel pouze z nich.

PRIMARY FINAL

První tabulka sjednocuje data týkající se mezd a cen potravin v České republice za totožné porovnatelné období. Rozhodla jsem se vytvořit dvě separátní tabulky pro ceny „t_price_light" a mzdy „t_payroll_light", pročistit a sjednotit jejich data, a až poté přistoupit ke spojení těchto dvou tabulek v jednu.

t_payroll_light

Základem pro tabulku „t_payroll_light" byla tabulka „czechia_payroll" v kombinaci s tabulkou „industry_branch_code", které jsem pomocí operátoru LEFT JOIN sjednotila tak, aby se napříště místo kódu ukázalo jméno průmyslového odvětví.

V původní tabulce „czechia_payroll" jsou údaje pro mzdy za každý kvartál sledovaného roku. Takto jemné rozdělení dat je pro naše účely zbytečné, navíc se z něj obtížněji získávají údaje k odpovědím na výzkumné otázky. Místo nich jsem použila průměrné hodnoty za každý rok (sloupec salary), a protože se jedná o relativně vysoké hodnoty v řádech desítek tisíc, zaokrouhlila jsem je na celé počty bez desetinných míst.

Původně jsem zamýšlela ponechat v tabulce nejen údaje o mzdě (value_type_code = 5958), ale také údaje o průměrném počtu zaměstnanců (value_type_code = 316), neboť by nám tato hodnota umožnila vytvářet u výzkumné otázky 2, 3 a 4 vážený průměr mzdy. Po kontrole dat jsem zjistila, že jsou tyto údaje nepoužitelné, protože z celkového počtu 3 440 údajů o počtu zaměstnanců tvoří 3096 hodnotu NULL. Vytvářet s tak nepřesnými daty vážený průměr nelze. Při hodnocení získaných výsledků je však nutné mít na paměti, že už zde vzniklo první výrazné zkreslení.

Z dat jsem dále odstranila všechny řádky, kde nebylo uvedeno průmyslové odvětví (industry_branch_code IS NULL) nebo kde chyběl samotný údaj o mzdě (value IS NULL). Časové rozpětí tabulky (payroll_year) jsem nechala v původním rozsahu s tím, že k jeho omezení dojde až po sjednocení s tabulkou t_price_light.

t_price_light

Obdobně jako u tabulky „t_payroll_light“ jsem postupovala také u tabulky „t_price_light“, jejímž základem byly tabulky „czechia_price“ a „czechi_price_category“, sjednocené pomocí LEFT JOIN tak, aby se místo kódu jednotlivých komodit zobrazovaly přímo jejich názvy.

V původní tabulce „czechia_price“ jsou měření po jednotlivých okresech, tabulka však obsahuje i celorepublikový průměr (region_code IS NULL), a tato hodnota je pro nás zajímavá. Různé komodity přitom mají různý počet celorepublikových hodnot, podle počtu provedených měření. Ty bylo potřeba sjednotit na průměrnou cenu za daný rok a komoditu, kterou jsem zaokrouhlila na jedno desetinné místo (price), tedy tak, jak to bývá běžné u cen v obchodech.

V případě „t_price_light“ bylo také potřeba upravit datum. Jako nejpříhodnější mi přišlo vytáhnout rok (year) z data začátku měření (date_from) .

V tabulce jsem dále kromě sloupce s názvem komodity (name), data (year) a ceny (price) ponechala také sloupce price_unit a price_value, které udávají množství dané komodity, např. 100g jogurtu. Tyto sloupce ve výpočtech nepoužívám, ale k datům pro přesnost a následnou kontrolu patří. Závěrem jsem v rámci čištění dat vyřadila ceny, které neměly přidělenou komoditu (name IS NULL).

t_Miloslava_Erika_Kaderabkova_project_SQL_primary_final

Finální tabulku „t_Miloslava_Erika_Kaderabkova_project_SQL_primary_final“ jsem vytvořila spojením obou předchozích tabulek (t_price_light a t_payroll_light). V textu ji budu nadále nazývat „Primary_Final“.

Tabulky „t_price_light“ a „t_payroll_light“ jsem spojovala pomocí LEFT JOIN. Na tabulku cen jsem nasadila tabulku platů tak, aby se shodovaly v časovém období. Tento moment je důležitý, protože v tabulce došlo v jeho důsledku k duplikaci řady hodnot, což je pak potřeba ošetřit při dalších operacích (většinou pomocí SELECT DISTINCT). Tabulky se spojily na stejné časové období 2006 – 2018, čímž došlo k zásadnímu omezení rozsahu tabulky t_payroll_light.

Ve výsledné tabulce „Primary_Final“ zůstaly řádky odpovídajícím letům (year), průměrným ročním cenám (price), komoditám a jejich jednotkám (comodity, price_value, price_unit), platům (salary) a odvětvím (branch). Tabulku jsem ještě pro pořádek seřadila vzestupně podle let a názvů komodit.

SECONDARY FINAL

Údaje pro „Secondary_Final“ obsahuje beze zbytku tabulka „economies“. Nerozlišuje však země a regiony, takže ve stejném sloupci sdružuje celky jako je např. Česká republika a region východní Asie a Pacifiku. Abych vyřešila tento neduh, spojila jsem ji vnořeným SELECTem s tabulkou „countries“, která obsahuje údaje pouze k státům.

Z tabulky „economies“ jsem pro „Secondary_Final“ vybrala sloupce s názvem země (country), rokem (year), populací (population), HDP (gdp) a Giniho koeficientem (gini). Časové rozmezí jsem omezila na roky 2006 až 2018 včetně, jako je to u Primary_Final. Tabulku jsem seřadila podle let a názvů zemí vzestupně. Za povšimnutí stojí, že u Giniho koeficientu máme velmi malé množství hodnot. Z více než dvou a půl tisíc řádků chybí ve více než jednom a půl tisíci z nich hodnota.

Problém s názvem státu a nutnost předělat tabulku:

Tabulku „economies“ jsem omezila na státy z tabulky „countries“ na základě názvu země, což je problematické. Tabulka „economies“ totiž na rozdíl od tabulky „countries“ neobsahuje dvoupísmenný mezinárodní kód ISO (sloupec abbreviation), který by byl ideálním řešením. Názvy zemí se mohou v obou tabulkách lišit, podle toho, jestli jde o oficiální nebo nejčastěji používaný název. Problematická je ČR i některé velmi významné země např.:

- Česká republika (Czech Republic nebo Czechia),
- Spojené království (United Kingdom, ale také UK),
- Spojené státy (United States, USA, US),
- Čína (China, Peoples Republic of China nebo PRC).

Abych prověřila, jestli nedošlo kvůli rozdílně zapsanému jménu státu ke zkreslení, spočítala jsem počet zemí v tabulce „countries“ a v „Secondary_Final“ pomocí SELECT COUNT (DISTINCT country). Zjistila jsem, že rozdíl činí 46 zemí, které jsem si následně nechala vypsat pomocí EXCEPT. U některých (např. Sultanát Brunej nebo Východní Timor) došlo k vyřazení kvůli odlišnému zapsání názvu, s jinými (např. Antarktida, Západní Sahara) tabulka „economies“ vůbec nepočítá. V případě Západní Sahary už do hry vstupuje také definice státu a otázka mezinárodního uznání. S výjimkou Severní Koreji a Sultanátu Brunej se jedná o velmi malé nebo ekonomicky nevýznamné státy, a proto jsem se rozhodla, že jejich opomenutí nezpůsobí v „Secondary_Final“ potíže. Severní Koreu a Sultanát Brunej jsem se ale rozhodla do tabulky dodatečně zařadit. V zásadě bylo možné tyto státy buďto ručně vložit, nebo „Secondary_Final“ smazat a vytvořit ji nově s upraveným kódem, který bude obě země obsahovat. Zvolila jsem druhou možnost.

3. Odpověď na otázku č. 1

„Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?“

Tabulka Primary_Final pokrývá období 12 let mezi roky 2006 až 2018. To znamená, že nemá cenu srovnávat platy na začátku a konci tohoto období, protože je logické, že za dvanáct let dojde téměř s jistotou k jejich růstu. Rozhodla jsem se proto propočítat meziroční nárůst platů a podívat se na to, ve kterých odvětvích došlo někdy ve sledovaném období k meziročnímu poklesu.

K srovnávání s předchozím obdobím se dá využít operátor LAG, který umožňuje srovnávat hodnoty mezi řádky v rámci jednoho sloupce. Vzhledem k tomu, že je tabulka Primary_Final kvůli sjednocení cen a mezd na jeden řádek složitá, rozhodla jsem se využít jinou možnost, a to spojení tabulky Primary_Final se sebou samou, posunutou datově o jeden rok (Pomocí LEFT JOIN). Vzorcem pro výpočet meziročního růstu jsem si spočítala růst pro každé odvětví a rok, s výjimkou roku 2006, kde chybí hodnoty. Následně jsem si vytvořila pomocný VIEW v_salary_growth, který předchozí SELECT omezuje na rok, odvětví a meziroční růst (growth).

Za pomocí vnořeného SELECT DISTINCT a pomocného view jsem si nechala zobrazit odvětví, v nichž došlo alespoň v jednom roce došlo (growth < 0) nebo nedošlo (growth >0) k poklesu.

V poklesu tak mezi lety 2006 – 2018 došlo alespoň jednou v následující odvětvích.

- Těžba a dobývání
- Profesní, vědecké a technické činnosti
- Zásobování vodou; činnosti související s odpady a sanacemi

- Informační a komunikační činnosti
- Administrativní a podpůrné činnosti
- Stavebnictví
- Zemědělství, lesnictví, rybářství
- Veřejná správa a obrana; povinné sociální zabezpečení
- Vzdělávání
- Peněžnictví a pojišťovnictví
- Velkoobchod a maloobchod; opravy a údržba motorových vozidel
- Činnosti v oblasti nemovitostí
- Ubytování, stravování a pohostinství
- Výroba a rozvod elektřiny, plynu, tepla a klimatiz. vzduchu
- Kulturní, zábavní a rekreační činnosti

Naopak, platy mezi lety 2006 – 2018 nikdy neklesly v:

- Ostatní činnosti
- Doprava a skladování
- Zdravotní a sociální péče
- Zpracovatelský průmysl

Největší pokles zaznamenalo v r. 2013 Peněžnictví a pojišťovnictví (-8,9%). Naopak k největšímu meziročnímu růstu došlo v r. 2008, a to u odvětví Těžba a dobývání a Výroba a rozvod elektřiny, plynu, tepla... (přes 13%) a u Profesních vědeckých a technických činností (téměř 13%). Nejčastěji klesaly meziročně mzdy v oblasti Těžba a dobývání, a to v letech 2009, 2013, 2014 a 2016. Přičemž v roce 2009 šlo s -3,7 % o třetí největší meziroční pokles ze všech odvětví za sledované období.

4. Odpověď na otázku č. 2

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

Pro přesné zodpovězení této otázky by bylo potřeba znát vážený průměr mezd za první (2006) a poslední (2018) sledované období. Bohužel však k tomu nemáme dostatek dat. Můžeme tak vytvořit pouze běžný průměr mezd, který je ale velmi zkreslený, neboť nezohledňuje počty osob, které pracovaly v jednotlivých odvětvích.

Z toho důvodu bych doporučovala vypočítat kolik litrů mléka a kilogramů chleba bylo možné koupit za průměrnou mzdu v r. 2006 a v r. 2018 v jednotlivých odvětvích. Tato tabulka je nejen přesnější, ale nabízí i zajímavý pohled na rozptyl kupní síly mezi různými profesemi. Vyjde nám, že v roce 2006 bylo možné za průměrnou měsíční mzdu koupit něco mezi 791 až 2756 l mléka a něco mezi 707 a 2465 kg chleba v závislosti na odvětví, přičemž nejnižší kupní síla byla u zaměstnanců Ubytování, stravování a pohostinství a administrativních pracovníků. Nejvyšší kupní síla pak u zaměstnanců IT, peněžnictví a pojišťovnictví. V roce 2018 bylo možné si za průměrný měsíční plat koupit něco mezi 948 a 2 833 litry mléka a 776 a 2318 kg chleba, přičemž nejslabší a nejsilnější odvětví vzhledem ke kupní síle zůstala totožná. Zajímavé je, že co se chleba týče, poklesla u administrativních pracovníků kupní síla mezi lety 2006 a 2018 o 15 kg. Obecný trend je ale opačný, tj. kupní síla u většiny odvětví vzrostla.

Kupní síla vypočítala v litrech mléka a kg chleba 2006 a 2018 podle odvětví

	Milk 2006	Bread 2018	Milk 2018	Bread 2018
Ubytování, stravování a pohostinství	791	707	948	776
Administrativní a podpůrné činnosti	967	865	1038	850
Zemědělství, lesnictví, rybářství	1015	908	1277	1045
Ostatní činnosti	1092	976	1135	929
Kulturní, zábavní a rekreační činnosti	1122	1003	1393	1140
Stavebnictví	1232	1102	1414	1157
Velkoobchod a maloobchod...	1241	1110	1485	1215
Zpracovatelský průmysl	1276	1142	1604	1313
Činnosti v oblasti nemovitostí	1278	1143	1393	1140
Zásobování vodou...i	1284	1148	1439	1178
Zdravotní a sociální péče	1285	1150	1654	1353
Vzdělávání	1327	1187	1502	1229
Doprava a skladování	1328	1188	1480	1211
Veřejná správa a obrana...	1602	1433	1819	1488
Profesní, vědecké a technické činn.	1658	1483	1923	1573
Těžba a dobývání	1670	1493	1818	1487
Výroba a rozvod elektřiny, plynu...	2003	1792	2332	1908
Informační a komunikační činnosti	2456	2197	2833	2318
Peněžnictví a pojišťovnictví	2756	2465	2733	2236

Pokud by byla přeci jenom potřeba jedna hodnota, doporučila bych vzít průměrný plat z dat ČSÚ, o němž se lze důvodně domnívat, že se bude jednat o kvalitní údaj. V tom případě by to bylo 19 546 CZK měsíčně za rok 2006 a za rok 2018 32 051 CZK měsíčně za rok 2018.

Údaje z našich nepřesných dat nám vypočítají, že za rok 2006 si bylo možné koupit za „průměrný“ plat 1441 l mléka a 1 289 kg chleba. Za rok 2018 to pak bylo 1643 l mléka a 1344 kg chleba. Kupní síla tedy zcela jasně vzrostla. Pro srovnání, pokud použijeme u platů hodnoty ČSÚ, vyjde nám, že v r. 2006 si bylo možné koupit za průměrný plat 1357 l mléka a 1214 kg chleba. V roce 2018 to pak bylo 1618 l mléka a 1324 kg chleba. Zmíněné hodnoty se tedy od sebe tolik neliší a růstový trend kupní síly dokazují i data z ČSÚ.

Kolik l mléka a kg chleba bylo možné si koupit v r. 2006 a 2018 za průměrný plat ?

	Mléko 2006	Mléko 2018	Chleba 2006	Chleba 2018
Plat (ENGETO)	1441	1543	1289	1344
Plat (ČSÚ)	1357	1618	1214	1324

Je zajímavé se podívat, jestli existují komodity, u kterých kupní síla klesala. Komoditou s největším meziročním růstem cen je máslo (viz další otázka) a u něj došlo skutečně k poklesu kupní síly. V roce 2006 si bylo možné za průměrný plat koupit 198,8 kg másla. V roce 2018 pouze 157 kg másla.

5. Odpověď na otázku č. 3

„Která kategorie potravin zdražuje nejpomaleji (je u ní nejnižší percentuální meziroční nárůst)?“

K získání percentuálního meziročního nárůstu jsem porovnala ceny z let 2006 s cenami z let 2018 pro každou komoditu zvlášť. Z výsledků vychází, že nejpomaleji zdražily banány, a to o 7,3 %. Nicméně, existují dvě komodity, u kterých došlo za pozorované období k poklesu ceny, a to cukr krystal (o 27,2 %) a rajčata (o 2,3 %). Do 20 % zdražila meziročně ještě vepřová pečeně s kostí, minerální voda, pečivo a jablka. Nejvyšší meziroční nárůst zaznamenalo máslo (98,4 %), těstoviny (83,5 %), papriky (71,4 %) a rýže (70%).

Vzhledem k tomu, že nebývá obvyklé, aby komodita za 12 let zlevnila téměř o 1/3, zkontrolovala jsem si vývoj cen cukru za dané období. Cena cukru v r. 2018 byla nejnižší za celé sledované období, což mohlo být způsobeno nenadálou událostí. Nicméně, za poslední čtyři sledované roky se cukr držel pod cenou z roku 2006.

6. Odpověď na otázku č. 4

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

Pro zodpovězení otázky potřebujeme spočítat růst mezd a cen potravin pro každý rok a tyto mezi sebou porovnat. Pro výpočet jsem použila vzorec pro meziroční procentuální změnu:

((Σ(cena)-Σ(lag))/Σ/(lag)) x 100
((Σ(mzda)-Σ(lag))/Σ/(lag)) x 100
přičemž lag = cena/mzda za předchozí rok

Meziroční růst platů/cen a jejich porovnání

rok	salary_growth	price_growth	difference	higher than 10%
2007	6,84	6,77	-0,07	0
2008	7,87	6,21	-1,66	0
2009	3,16	-6,41	-9,57	0
2010	1,95	1,93	-0,02	0
2011	2,3	3,34	1,04	0
2012	3,03	6,74	3,71	0
2013	-1,56	5,1	6,66	0
2014	2,56	0,72	-1,84	0
2015	6,45	3,29	-3,16	0
2016	3,65	-1,18	-4,83	0
2017	6,28	9,62	3,34	0
2018	7,62	2,18	-5,44	0

Hodnoty meziročního růstu cen a mezd jsem spojila do jedné tabulky, vypočítala jejich rozdíl a časové rozmezí omezila na roky 2007 – 2018, protože pro rok 2006 nemáme hodnoty. Ze spojených tabulek jsem vytvořila pomocné view „v_salary_price_growth“, které se hodí i v odpovědi i na otázku č. 5. Pomocí operátoru CASE jsem si nechala vyhledat ty případy, kdy je růst cen o 10 procentních bodů vyšší než růst mezd (sloupec higher than 10%). V tom případě je hodnota sloupce 1, v ostatních případech 0. Z tabulky tedy vyplývá, že situace, kdy by růst cen byl o více než 10 p.b. vyšší než růst mezd nenastala. Naopak, v roce 2009 nastala situace

téměř opačná, kdy došlo k poklesu cen, ale platy nadále mírně rostly, takže rozdíl mezi nimi dosáhl téměř 10 p.b. Nejvíce vzrostly ceny oproti platům v roce 2013, a to o 6,6 p.b.

7. Odpověď na otázku 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 následujícím roce výraznějším růstem?“

Meziroční nárůst cen a mezd již máme k dispozici ve „v_salary_price_growth“, které jsem vytvořila v rámci odpovědi na předchozí otázku. Meziroční růst GDP jsem vypočítala vzorcem pro meziroční růst.

((GDP-lagGDP)/lagGDP)*100
přičemž lagGDP = hodnota GDP z předchozího roku

Všechny hodnoty jsou zaokrouhlené na 2 desetinná místa pro snazší porovnávání. Časový rámec je omezený na období od roku 2007, abychom se vyhnuli NULL hodnotám.

Meziroční růst HDP, mezd a cen v ČR

year	gdp_growth	salary_growth	price_growth
2007	5,57	6,84	6,77
2008	2,69	7,87	6,21
2009	-4,66	3,16	-6,41
2010	2,43	1,95	1,93
2011	1,76	2,3	3,34
2012	-0,79	3,03	6,74
2013	-0,05	-1,56	5,1
2014	2,26	2,56	0,72
2015	5,39	6,45	3,29
2016	2,54	3,65	-1,18
2017	5,17	6,28	9,62
2018	3,2	7,62	2,18

Pouhým pohledem na data v tabulce není patrný žádný trend. Proto jsem se rozhodla, že použiju k odhalení možného trendu korelaci. Pro snaží manipulaci z daty jsem si výše uvedenou tabulku uložila jako dočasnou tabulku t_correlation a použila ji k výpočtu korelačního koeficientu.

Nejvyšší míru korelace s růstem HDP vykazuje růst mezd v následujícím roce. Korelační koeficient (dále R) činí 0,66. V ekonomii se obecně považuje za silnou korelaci absolutní hodnota R nad 0,7. Při hodnotě R 0,66 tedy můžeme mluvit tvrdit, že mezi oběma proměnnými je relativně silná korelace. Mírnou korelaci vykazuje růst HDP také s růstem cen (R=0,59) a růstem mezd (R = 0,56) v tomtéž roce. Ve všech zmíněných případech se jedná o pozitivní korelaci, tzn. roste-li HDP, porostou i mzdy a ceny. Naopak na vliv růstu cen v následujícím roce podle našich dat růst HDP již vliv nemá (R = -0,02).

Závěrem je potřeba upozornit, že korelace sama o sobě neznamená kauzalitu. Bez hlubšího zkoumání a znalostí tedy není možné tvrdit, že růst HDP vede k růstu mezd v následujícím roce. Za růstem HDP i mezd může stát např. jiná proměnná, která oba růsty ovlivňuje.

8. Závěr

Cílem projektu bylo zodpovědět pět ekonomických otázek na základě údajů ze dvou datových tabulek, přičemž jsme zkoumali období mezi lety 2006 a 2018.

Data potvrdila, že i přes postupný nárůst průměrných mezd dochází téměř ve všech odvětvích k občasnému meziročnímu poklesu mezd. Nejnáchylnější na to je odvětví Těžby a dobývání. Data také potvrdila postupný nárůst kupní síly mezi lety 2006 a 2018 u většiny běžných komodit jako je mléko a chleba. U některých komodit, u kterých došlo v daném období k nejdramatičtějšímu nárůstu ceny (např. máslo), však kupní síla poklesla.

Za zkoumané období zdražovaly z vybraných komodit nejpomaleji banány (7,3%), a u cukru krystal a rajčat dokonce došlo ke snížení ceny, přičemž u cukru poměrně dramaticky (o více než 27%). Naopak nejvíce zdražilo máslo (98,4 %), těstoviny (83,5 %), papriky (71,4 %) a rýže (70%).

Meziroční růst cen nebyl nikdy výrazně vyšší než meziroční růst mezd, to jest rozdíl nikdy nepřekročil 10 procentních bodů. Nejvíce se mu přiblížil v r. 2013 (6,7 p.b.).

Růst HDP vykazuje relativně silnou korelaci s růstem mezd v následujícím roce. S růstem cen a mezd v daném roce koreluje pouze mírně, a z růstem cen v následujícím roce vůbec.