Project SQL – Veronika Kvasničková

Průvodní listina

1. 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.

Výzkumné otázky

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 percentuální 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?

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.

Na svém GitHub účtu vytvořte repozitář (může být soukromý), kam uložíte všechny informace k projektu – hlavně SQL skript generující výslednou tabulku, popis mezivýsledků (průvodní listinu) a informace o výstupních datech (například kde chybí hodnoty apod.).

Neupravujte data v primárních tabulkách! Pokud bude potřeba transformovat hodnoty, dělejte tak až v tabulkách nebo pohledech, které si nově vytváříte.

Datové sady

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

Primární tabulky:

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

Číselníky sdílených informací o ČR:

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

Dodatečné tabulky:

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

2. VÝSTUP

Porozumění dat

Prvním krokem bylo porozumění dat. Prohlédla jsem podrobně všechny relevantní tabulky pomocí SELECT a SELECT DISTINCT. (Byly rozdílné hodnoty mezi lokální databází a online databází; to nakonec ale bylo opraveno). Využila jsem též: https://data.gov.cz/datov%C3%A1-sada?iri=https%3A%2F%2Fdata.gov.cz%2Fzdroj%2Fdatov%C3%A9-sady%2F00025593%2F7ddcb833bddeeb84db39004d7e276b87.

Dále bylo třeba zhodnotit, zda se budou používat údaje pro přepočtený nebo pro fyzický počet zaměstnanců (z tab. czechia_payroll) a průměrných mezd. Rozhodla jsem se pro přepočtený, protože to podle mého názoru přesněji odpovídá průměrné mzdě. Z textu na data.gov.cz plyne, že tabulka nezahrnuje mj. dohody o pracovní činnosti a dohody o provedení práce. Což též podporuje rozhodnutí vyjít z přepočtených mezd, protože dohody mimo pracovní poměr většinou bývají lépe ohodnoceny než standardní pracovní úvazek (při přepočtu na plný úvazek), a tím by docházelo ke zkreslení přepočtených údajů. Nicméně toto není tedy případ tabulky.

Na druhou stranu i volba nepřepočteného průměru by mohla být v některých případech vhodná – zejména pro zjištění, kolik si v průměru člověk může za svou průměrnou mzdu pořídit zboží. Nicméně pro zjednodušení bylo využito pouze první varianty.

Z hlediska region_code byly pro primary table vybrány údaje s nulovou hodnotou, protože to znamenalo, že se jedná o průměr za celou ČR, což je pro nás podstatné.

Z hlediska industry_branch_code byly vybrány nenulové hodnoty, a tedy hodnoty,u kterých odvětví nebylo stanoveno, nebylo bráno v potaz. (Mj. proto, že máme dle výzkumných otázek sledovat odvětví).

Primary table

I. Prvním krokem bylo vytvoření dvou výběrů ze zdrojových tabulek pro lepší kontrolu a porozumění dat. Tyto výběry jsou pouze návodné pro vytvoření primary table. Není s nimi tedy jinak dále pracováno, je možné je tedy smazat. Ponechala jsem je pouze pro ukázku mého postupu. První výběr zobrazuje průměrnou cenu každé potraviny pro každý rok. Druhá výběr zobrazuje průměrný plat v každém odvětví pro každý rok.

II. Primární tabulka s názvem t_veronika_kvasnickova_project_SQL_primary_final byla vytvořena seskupením dat (JOIN) z tabulek a "czechia_payroll" a "czechia_price", a to pomocí společných roků (fce YEAR) (jak bylo v zadání stanoveno).

Dále byly připojeny další relevantní tabulky: czechia_payroll_industry_branch a czechia_price_category.

Zároveň byly vybrány pouze relevantní sloupce, s tím, že byla zprůměrována mzda za každý rok (průměr za čtyři čtvrtletí) a zprůměrována cena potravin v rámci jednoho roku.

Údaje byly seskupeny s srovnány.

Secondary table

I. Nejprve bylo zjištěno, jaký název se používá pro Českou republiku v tabulce economies (výsledek: "Czech Republic"), a to pomocí SELECT DISTINCT byly zjištěny názvy zemí/oblastí a vyhledána ČR. Dále bylo zjištěno náhledem do tabulky, jaké časové rozmezí tabulka zahrnuje (výsledek: delší, než potřebujeme).

II. Byla vytvořena pomocná tabulka t_veronika_kvasnickova_project_SQL_secondary1: z tabulky economies s tím, byly vybrány pouze evropské země (využití vnořeného SELECT) a pouze určité roky (které jsou relevantní pro další srovnávání) a vybrány relevantní sloupce.

III. Tato pomocná tabulka byla spojena sama se sebou a zároveň byl přidán další sloupec, který vypočítal procentuální rozdíl v HDP (který bude využit pro odpověď na otázku č. 5).

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

Postup:

A) Nejprve je vytvořen VIEW pouze s relevantními sloupci za odvětví: v_veronika_kvasnickova_industry.

B) Poté je vytvořen VIEW: v_veronika_kvasnickova_payroll_dif, který vznikne spojením jedné tabulky s tím, že provazba je v roce navýšeném o 1. Propojení dále je též stanoveno na úrovni odvětví.

Dalším krokem je vytvoření sloupce, který vypočítá rozdíl mezi průměrnými mzdami během dvou let, čím se zjistí, zda v některých odvětví mzda rostla/klesala.

Použitím WHERE je zamezeno zobrazení roku 2021 (protože rok 2022 již nenásleduje, a tudíž by nebylo možné vypočítat rozdíl mezi 2021 a 2022): tímto je pouze tabulka upravena. A též jsou vybrána pouze negativní čísla = oblasti, ve kterých mzda klesá.

Výstupem je zobrazení všech odvětví a příslušného roku, kde mzda klesala.

C) Ve finálním SELECT jsou pak vidět všechna odvětví, ve kterých mzda z roku na rok klesla, a počet, kolikrát se tak stalo. Tabulka je seřazena podle počtu poklesů mzdy během sledovaného období.

Výsledek:

V některých letech mzda klesala. Téměř ve všech odvětvích v průběhu let došlo alespoň jednou k poklesu mzdy; v některých odvětvích se tak stalo i vícekrát: nejčastěji pak v těžbě.

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

Postup:

A) Prvním krokem bylo zjištění prvního (MIN=2006) a posledního (MAX=2018) srovnatelného období.

B) Druhým bylo zjištění kódu pro mléko a chléb pomocí SELECT DISTINCT.

C) Též byly ověřeny jednotky v tabulce czechia_price_category – výzkumná otázka vs. tabulka. Závěr: jednotky jsou totožné.

D) Následně byl z primární tabulky vytvořen VIEW s průměrnou cenou pro mléko a chléb pouze pro vybrané dva roky

E) Dále byl z primární tabulky vytvořen druhý VIEW s průměrnou mzdou pouze za dané dva roky (vč. nutné úpravy za všechna odvětví pro každý daný rok).

F) Posledním krokem bylo spojení dvou předchozích VIEW (D+E) a vytvoření nového sloupce, ve kterém bylo spočítáno podílem počet litrů mléka a kg chleba.

Výsledek:

V roce 2006 bylo možné koupit 1313 kg chleba a 1466 l mléka.

V roce 2018 bylo možné koupit 1365 kg chleba a 1670 l mléka.

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

Postup:

A) Vytvoření VIEW v_veronika_kvasnickova_groceries s údaji pouze o potravinách za každý rok, a to pouze za roky, ve kterých údaje k potravinám jsou, tj. bez nulových hodnot (WHERE). Využito SELECT DISTINCT, protože jinak by údajů bylo příliš mnoho (ke každému odvětví, což pro tuto výzkumnou otázku není relevantní).

B) Spojení předchozího VIEW navzájem, a to na základě roku, který je posun o +1 a kódu potravin. Zároveň přidán nový sloupec s vypočteným průměrem rozdílů cen v potravinách. Vybrány pouze relevantní sloupce.

Pomocí WHERE vybrány pouze ty řádky, které jsou nenulové (aby se nezobrazil rok 2019, u kterého nejsou ceny uvedeny).

GROUP BY využito pro seskupení údajů na jednotlivé potraviny.

Zároveň stanoven LIMIT pro 5 nejnižších hodnot.

Výsledek:

U některých potravin dochází dokonce ke snižování ceny (cukr krystalový, rajčata). U dalších potravin dochází v průměru ke zdražení: nejméně zdražují banány, poté vepřová pečeně, minerální vody.

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

Postup:

Mzdy:

A) Vytvoření VIEW v_veronika_kvasnickova_pay_year pouze z průměrnými mzdami pro každý rok. Celková průměrná mzda za rok byla spočítána jako průměr mezd za jednotlivá odvětví. Při výpočtu nebylo tedy rozlišováno, kolik osob v příslušném odvětví pracuje (tato informace není známa). (Tudíž se nejedná o přesný výpočet průměrné mzdy. Pro správný výpočet by se mělo vyjít oficiálních údajů za celou ČR.)

B) Z v_veronika_kvasnickova_pay_year byl vytvořen další VIEW, a to spojením tabulky samé se sebou. Byl vypočítán procentuální rozdíl mezi dvěma po sobě jdoucími roky. Tedy byl zjištěn růst (popř. pokles) zprůměrovaných mezd mezi dvěma roky.

Pozn. též by bylo možné počítat nárůst/pokles mezd v každém odvětví, a tyto údaje pak pro každý rok zprůměrovat.

Potraviny:

C) Byl vytvořen VIEW v_veronika_kvasnickova_groceries_t4 pouze pro potraviny. Pro každý rok a každou kategorii potravin byla vypočítána průměrná cena.

D) Byl vytvořen VIEW v_veronika_kvasnickova_groceries_dift4, a to spojením předchozí tabulky samé se sebou. Navíc byl vytvořen nový sloupec s výpočtem procentuální změny pro každý rok a každou kategorii potravin.

E) V následujícím kroce byl vypočítán průměr procentuální změny za celý rok za všechny kategorie v daném roce. Jedná se o prostý aritmetický průměr. (K dispozici nebyl spotřební koš apod. Správně by mělo být využito oficiálních údajů o inflaci.).

Srovnání obou tabulek:

F) Bylo vytvořeno další VIEW: spojeny dvě tabulky – jedna průměrné navýšení mezd a druhá průměrné navýšení cen potravin. Byl doplněn další sloupec, a to 10% navýšení průměrné mzdy (který bude nutný pro další bod).

G) V posledním kroku byl vytvořen finální VIEW, který předchozí VIEW upravil o nové dynamické sloupce s cílem zjistit, jestli v některém roce ceny potravin výrazněji rostly, než rostly mzdy. Výrazněji znamená o více jak deset procent. A též byl doplněn kontrolní dynamický sloupec (viz dále ***)

Výsledek:

V některých letech (čtyřech) došlo k meziročnímu poklesu cen potravin (záporné hodnoty změn průměrných cen potravin), v mezidobí 2008-2009 k výraznějšímu poklesu. V ostatních období ceny potravin rostly.

Ve všech letech, kromě roku 2012-2013 došlo k meziročnímu nárůstu mezd.

Ve většině případů tak rostly jak mzdy, tak ceny.

Zhruba v polovině případů docházelo k tomu, že nárůst cen potravin byl vyšší než nárůst mezd (sloupec "what_increases_more"). Zároveň ve všech těchto případech se jednalo o více jak 10% nárůst (ve smyslu, že nárůst cen potravin byly o více jak 10 % větší než nárůst mezd). (Např. kdyby mzdy vzrostly o 5 % a potraviny o 5,2 %, pak by tedy docházelo k nárůstu cen potravin, avšak ne o více jak 10 % z mezd. A naopak, kdyby mzdy vzrostly o 5 % a potraviny o 5,6 %, pak dochází k nárůstu cen, a to výraznému, o více jak 10 % k cenám mezd.

Pokud však byla otázkou nárůstu mezd větší jak 10 % myšlena skutečnost, že by v některém roce mzdy vzrostly o více jak 10 %, pak k takové situaci nedošlo (a dynamické sloupce ani by nebylo třeba vytvářet).

*** Pro úplnost doplňuji, že dynamický sloupec nedává smysl pro změnu mezd v rámci let 2012-2013, protože se jedná o zápornou hodnotu. Při připočtení 10 % se tato záporná hodnota navýší, což v tomto případě nedává smysl. Toto jsem již dále neřešila. Pouze jsem vložila nový "kontrolní dynamický sloupec.

5. otázka: 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?

Postup:

A) Vycházím z již dříve vytvořených tabulek a pohledů. Dříve vytvořené pohledy jsou dostačující, protože bereme v potaz pouze roky, ve kterých jsou údaje za mzdy a zároveň i za potraviny, aby byla stejná datová základna z hlediska počtu roků.

B) Nejprve je ze secondary table vytvořena tabulka t_veronika_kvasnickova_gdpcz pouze pro výběr České republiky. Poté je provazba s touto tabulkou a dalšími tabulkami na základě roku: přičemž i) rok není posunut, protože úkolem je zjistit vztah ve stejném roce a dále je ii) rok (u potravin a mezd) posunut, aby byl srovnán i rok následující.

C) Nejvhodnějším způsobem by bylo zjistit korelaci mezi změnou v HDP a změnami v mzdách a potravinách ve stejném a dále i v následujícím roce.

   Jako jednodušší variantu zkouším výpočet prostého poměru mezi HDP a potravinami/mzdou ve stejném/následujícím roce. Pokud podíl napříč celým sloupcem dat bude obdobný, pak se jedná o souvislost mezi HDP a potravinami/mzdami. Pokud odlišný, nejedná se o souvislost. Potíže nastanou při přepočtu se zápornými čísly. Proto je nakonec zvolen další způsob. (V SQL toto ponecháno zakomentované).

   Další způsob je vizuální: vytvoření nových sloupců pro HDP, potraviny a mzdy a odlišení, jak moc dochází k nárůstu (viz SQL kód). Čím vyšší hodnoty, tím více "x". Hodnoty pro stanovení "x" jsou uvedeny fixně. Pokud je mezi hodnocenými údaji (např. GDP vs. potraviny) počet "x" stejný či o jedno "x" menší/větší, pak je zde souvislost. Pokud odlišný, souvislost není. Rozložení 'x' bylo stanoveno úsudkem. Tento způsob je dostačující pro naše potřeby, protože roků k sledování souvislosti je jen 11 a je možné toto tedy vizuálně zhodnotit.

   Třetí varianta je zobrazena ve sloupci "groceries_2" (a pouze pro tento sloupec). Jedná se o nový sloupec, který je odvozen od změn GDP a změn cen potravin v roce následujícím. Pokud je zde souvislost, je uvedena "1". Pokud souvislost není, je uvedena "0". Možných let je 11. Zjištění počtu "1" je možné pomocí GROUP BY a SUM. Výsledkem je "5". To znamená, že v pěti letech byla souvislost a ve zbývajících šesti (tj. 11-5), souvislost nebyla prokázána. Tento způsob zjištění je složitější, nicméně vhodný pro větší objem dat. Každopádně výpočtu korelace se nevyrovná.

Výsledek:

Vizuální srovnání:

Srovnání HDP a cen potravin ve stejném roce: není patrný vliv. V některých letech je HDP vyšší a také rostou ceny (např. 2006-2007). V některých je to ale opačně (např. 2011-2012).

Srovnání HDP a cen potravin v následujícím roce: není patrný vliv. V některých letech je HDP vyšší a také rostou ceny (např. 2006-2007). V některých je to ale opačně (např. 2014-2015).

Srovnání HDP a mezd ve stejném roce: je patrný vliv.

Srovnání HDP a mezd v následujícím roce: je patrný vliv.