

Průvodní listina k SQL projektu – Marian Koutný

1. Výpis datových setů potřebných k projektu

Jako první jsem si vypsal všechny potřebné datové sety, které jsou pro projekt relevantní, abych si je postupně všechny prošel, osahal si je a viděl, jak jsou spolu vzájemně propojeny.

2. Vytvoření pomocných tabulek, pomocí kterých se dostaneme k první finální tabulce - t_mk_wage, t_mk_price, t_mk_price_general

Vytvořil jsem si pomocnou tabulku **t_mk_wage**, která obsahuje průměrné roční mzdy v jednotlivých odvětvích od roku 2000 do roku 2021. Využil jsem k tomu tabulku **czechia_payroll** spolu s tabulkou **czechia_payroll_industry_branch**. Zprůměroval jsem mzdy z jednotlivých čtvrtletí skrze **kód 5958** (tedy hodnotou mzdy) a napříč neprázdnými hodnotami **industry_branch_code** (u kterých bych nevěděl, k čemu mzda patří). Výsledek je 418 hodnot. (22 let naskrz 19 odvětvími)

Dále jsem si vytvořil tabulku **t_mk_price**, která obsahuje průměrné ceny jednotlivých potravin v daných letech a krajích. Použity byly sety **czechia_price_category**, **czechia_price** a **czechia_region**. Pro výpočet jsem používal datovou hodnotu **date_from** a následně průměroval týdenní měření na celý rok přes neprázdné hodnoty **regionu**. Výsledek 4788 hodnot. (roky 2006-18 pro 14 krajů a 26 potravin, položka jakostní víno se objevuje pouze v letech 2015-18 -> celkem tedy $14*26*13 + 14*4 = 4788$ hodnot)

Poté jsem si vytvořil poslední tabulku **t_mk_price_general**, kde jsem využil předchozí tabulku a jednotlivé ceny jsem zprůměroval pro všechny kraje. Ve výsledku tedy vznikla tabulka, která obsahuje ceny 26 potravin v letech 2006-18 a jakostního vína v letech 2015-18, celkem tedy 342 hodnot.

3. Vytvoření první finální tabulky t_marian_koutny_project_sql_primary_final:

Nakonec jsem vytvořil konečnou tabulku **t_marian_koutny_project_sql_primary_final**, ke které jsem využil pomocné tabulky **t_mk_wage** a **t_mk_price_general**. Zde jsem obě tabulky propojil na proměnné **roku**. Výsledkem je tabulka s 5 sloupci (**odvětví**, **rok**, **příslušný průměrný plat v daném roce**, **název potraviny** a **její průměrná cena v daném roce** – celkem 6669 hodnot. (26 potravin naskrz 19 odvětvími a 13 lety a jakostní víno naskrz 4 roky a 19 odvětvími a poté roky 2000-05 a 2019-21, tedy 9 let naskrz 19 odvětvími, které mají u potravin hodnotu **NULL**, jelikož cena není k dispozici -> $26*19*13 + 4*19 + 9*19 = 6669$). Důležité pro zachování údajů o platech v letech 2000-05 a 2019-21 bylo použití **LEFT JOIN**. Nakonec proběhlo pár úprav tabulky.

4. Vytvoření pomocné tabulky pro sekundární tabulku projektu - tabulka t_ec

Vytvořil jsem si pomocnou tabulku `t_ec`. Využil jsem k tomu datovou sadu `economies` spolu se sadou `countries`. Ty jsem propojil na základě parametru `country`. Získal jsem tak tabulku s potřebnými údaji o jednotlivých zemích (`název`, `GDP`, `gini`, `population`, `kontinent`).

5. Vytvoření druhé finální tabulky `t_marian_koutny_project_sql_secondary_final`

Nakonec jsem vytvořil finální tabulku `t_marian_koutny_project_sql_secondary_final` za pomoci předchozí tabulky. Propojení bylo provedeno na základě `country` a `roku`. Vzniklá tabulka obsahuje data o 41 evropských zemích z let 2000 až 2021 (jsou-li k dispozici). Byla vyloučena závislá, okupovaná a další území, kde byl např. nedostatek dat. (Vatikán, Svalbard, S. Irsko, Faerské o., Gibraltar, Lichtenštejnsko a Montenegro). Výsledná tabulka obsahuje údaje o `zemi`, `GDP`, `jeho vývoji`, `populaci` a `gini` koeficientu. Tabulka obsahuje 818 řádků.

Jako poslední krok jsem odstranil všechny `pomocné tabulky` a zobrazil `finální tabulky` k projektu.¹

Část druhá – výzkumné otázky pro analytické oddělení

1. ROSTOU V PRŮBĚHU LET MZDY VE VŠECH ODVĚTVÍCH, NEBO V NĚKTERÝCH KLESAJÍ?

1.1 Obecné uvedení do vývoje platů v letech 2000 až 2021 napříč odvětvími.

1.2 Vývoj průměrných platů jako celek ukazuje, že **za celé období 2000-2021 stouply mzdy ve všech odvětvích** (minimálně na 2,25 násobek), nejvýrazněji pak v sektoru zdravotní a sociální péče, kde to byl téměř čtyřnásobek.

1.3 Výpis detailního vývoje mezd v jednotlivých odvětvích po letech.

1.5 Podíváme-li se na detailní data zaměřená na jednoleté změny napříč jednotlivými sektory, zjistíme, že **mzdy v některých sektorech meziročně klesaly**. Takových poklesů bylo celkem 30. Nejvýraznější pokles mezd zaznamenal sektor peněžnictví a pojišťovnictví v roce 2013, a to téměř 9%.

1.6 Výpis let, ve kterých klesaly mzdy, a počet sektorů, kterých se tento pokles týkal. Je patrné, že smutným vítězem je rok 2013, kdy zhoršení platů pocítili zaměstnanci 11 sektorů z celkových 19.

1.7 Záznam ukazuje četnost poklesů mezd v jednotlivých sektorech, nejvíce zasažen byl obor těžba a dobývání. Naopak můžeme vyčíst, že hned 4 odvětví měly za celé období růst mezd každý rok.

¹ Pozn. autora: Je mi jasné, že existují určitě elegantnější řešení než tvorba pomocných tabulek, nicméně tento postup nebyl nikde zakázán, tudíž mi připadá korektní. Na konci postupu jsou výsledkem požadované 2 tabulky a všechny pomocné tabulky jsou odstraněny.

2. KOLIK JE MOŽNÉ SI KOUPIT LITRŮ MLÉKA A KILOGRAMŮ CHLEBA ZA PRVNÍ A POSLEDNÍ SROVNATELNÉ OBDOBÍ V DOSTUPNÝCH DATECH CEN A MEZD?

2.1 Přehled, **kolik kg chleba si můžeme koupit za průměrnou mzdu v letech 2006 a 2018.**

Z výsledku je patrné, že chléb v tomto období reálně „zlevnil“.

2.2 Přehled, **kolik litrů mléka si můžeme koupit za průměrnou mzdu v letech 2006 a 2018.**

Z výsledku je patrné, že mléko v tomto období také reálně „zlevnilo“. V obou případech tedy mezi lety 2006 a 2018 celkové průměrné mzdy stoupaly strmějším tempem než průměrné ceny našich dvou potravin.

2.3 Detailní rozbor kupní síly na jednotlivé obory. Zde můžeme pozorovat, že chléb zlevnil jen pro některá odvětví. Mléko zlevnilo pro všechna odvětví vyjma peněžnictví a pojišťovnictví.

3. KTERÁ KATEGORIE POTRAVIN ZDRAŽUJE NEJPOMALEJI (JE U NÍ NEJNIŽŠÍ PERCENTUÁLNÍ MEZIROČNÍ NÁRŮST)?

3.1 Zde vidíme **jeden nejnižší meziroční růst ceny potravin za celé období** – konkrétně **nejvíce zlevnila rajčata v roce 2007** a to o celých 30%. Pokud z příkazu vyřadíme **LIMIT**, můžeme se podívat na jednotlivé meziroční vývoje cen všech potravin, kde byly ceny k dispozici. (**NOT NULL**)

3.2 Přehled zdražování všech potravin za celkové období (2006-2018). Jak vidno, **nejpomaleji za celé období zdražuje cukr krystal**, jehož cena spadla během 12 let o 27,5%. Na opačném konci žebříčku je máslo. Položku jakostní víno nebereme v potaz, jelikož jeho hodnoty nepokrývají celé časové období a výsledek by neměl relevanci.

4. EXISTUJE ROK, VE KTERÉM BYL MEZIROČNÍ NÁRŮST CEN POTRAVIN VÝRAZNĚ VYŠŠÍ NEŽ RŮST MEZD (VĚTŠÍ NEŽ 10 %)?

4.2 Výsledná tabulka ukazuje několik věcí: nejprve potravinu, její cenu v daném roce a také v roce předchozím, následuje plat v daném roce a v předchozím roce. Tyto údaje slouží pro další sloupcečky v tabulce, kde jsou vidět nárůsty průměrných cen jak potravin v daných letech, tak i průměrných mezd.

Poslední sloupec pak ukazuje rozdíl, zda-li stoupaly rychleji ceny potravin či platy a o kolik. Celkem 36 položek potravin mělo meziroční nárůst o více než 10 procent vyšší než byl růst platu v daném roce. Nicméně se jedná o maximálně 7 položek ročně (rok 2013) z celkových 26 potravin, kde rozdíl překročil oněch 10%. **Tudíž jako závěr můžeme tvrdit, že nepochybně byly roky, kde ceny některých potravin opravdu rostly výrazně rychleji než mzdy, ale byla to vždy menšina z nich. V tomto případě asi bude záležet na úhlu pohledu. Pokud stačí jedna potravina, pak je podmínka splněna, pokud bychom chtěli většinu potravin, tak nikoliv.**

4.3 Detailní přehled jak výrazné zdražování bylo vztaženo na obor, v kterém člověk pracuje.

5. MÁ VÝŠKA HDP VLIV NA ZMĚNY VE MZDÁCH A CENÁCH POTRAVIN? NEBO LI, POKUD HDP VZROSTE VÝRAZNĚJI V JEDNOM ROCE, PROJEVÍ SE TO NA CENÁCH POTRAVIN ČI MZDÁCH VE STEJNÉM NEBO NÁSLEDUJÍCÍM ROCE VÝRAZNĚJŠÍM RŮSTEM?

5.1 Vývoj růstu HDP a průměrné mzdy vztažené na stejný rok. Na první pohled tam lehká závislost nejspíše je, ale pouze z rozdílu hodnot je těžké hodnotit. Důležitý je také aspekt, jak si zvolíme procentuální růst výše HDP, který je pro nás signifikantní. Tabulka s daty je dostupná, nicméně jsem si dovolil přidat i excel, kde je vidět korelační koeficient. Do SQL syntaxe jsem použil HDP růst větší než 5 procent.

5.2 Stejný výstup jako v 5.1, pouze máme souvislost mezi HDP v roce x a mzdy v roce $x+1$. U tohoto příkladu vychází za celé sledované období poměrně slušná korelace 0,65. U let, kdy je HDP růst skutečně vysoký, je korelace minimální a závislost tedy těžko pohledat.

5.3 a 5.4 Zde bude podobná logika jako předchozího výzkumu. Na první pohled moc souvislost nevidím, hledal bych ji asi primárně jinde. Ideálně by bylo každou potravinu rozebrat zvlášť.

Závěr

Na konec bych ještě rád podotknul, že si uvědomuji, že by možná bylo vhodnější během spouštění příkazů používat spíše číselné kódy než slovní spojení, nicméně uživatelsky mi to připadá přívětivější a pokud bych tyto příkazy spouštěl a viděl poprvé, rozhodně by mi to usnadnilo orientaci.