Projekt z SQL

Na analytickém oddělení nezávislé společnosti, která se zabývá životní úrovní občanů, jsme se dohodli, že se pokusím 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 odpovědi na jednotlivé otázky tiskovému oddělení. Toto oddělení bude výsledky prezentovat na následující konferenci zaměřené na tuto oblast.

Potřebují k tomu připravit robustní datové sady, konkrétně dvě tabulky, ve kterých bude možné vidět porovnání dostupnosti potravin na základě příjmů za určité časové období.

Jako dodatečný materiál i tabulku s HDP, GINI koeficientem a populací dalších evropských států ve stejném období jako primární přehled pro ČR.

Datové sady, ze kterých jsem mohl čerpat pro vytvoření primární a sekundární tabulky:

- **czechia_payroll** – Informace o mzdách v různých odvětvích za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR. Datová sada má sloupce id, value, value_type_code, unit_code, calculation_code, industry_branch_code, payroll_year a payroll_quarter.

- **czechia_payroll_calculation** - Číselník kalkulací v tabulce czechia_payroll. Číselník má sloupce name a code, který nabývá hodnot 100 nebo 200.

- **czechia_payroll_industry_branch** - Číselník odvětví v tabulce czechia_payroll. Číselník má sloupce name a code, který nabývá hodnot A až S.

- **czechia_payroll_unit** - Číselník jednotek hodnot v tabulce czechia_payroll. Číselník má sloupce name a code, který nabývá hodnot 200 nebo 80 403.

- **czechia_payroll_value_type** - Číselník typů hodnot v tabulce czechia_payroll. Číselník má sloupce name a code, který nabývá hodnot 316 nebo 5 989.

- **czechia_price** - Informace o cenách vybraných potravin za několikaleté období. Datová sada pochází z Portálu otevřených dat ČR. Datová sada má sloupce id, value, category_code, date_from, date_to a region_code.

- **czechia_price_category** - Číselník kategorií potravin v tabulce czechia_price. Číselník má sloupce name, price_value, price_unit a code, který nabývá 27 různých možností.
- **czechia_region** - Číselník krajů České republiky dle normy CZ-NUTS 2.

- **czechia_district** - Číselník okresů České republiky dle normy LAU.

- **countries** – Všemožné informace o zemích světa, například hlavní město, kontinent, měnu, národní jídlo, průměrnou výšku populace, telefonní předvolbu nebo náboženství.

- **economies** - Další informace o zemích světa v jednotlivých letech, například HDP, GINI, daňová zátěž nebo počet obyvatel.

Výzkumné otázky od kolegů zní:

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?

Vytvoření primární tabulky - t_Lukas_Vesely_project_SQL_primary_final

Pro vytvoření primární tabulky, kde měla být všechna data o mzdách a cenách za Českou republiku sjednocených na totožné porovnatelné období, jsem se nejdříve jako první krok podíval na všechny datové sady, abych věděl, v čem budu pracovat a jaké číselníky v hlavních tabulkách czechia_payroll a czechia_price mám.

První dotazem jsem se podíval na číselník czechia_payroll_unit, kde jsem chtěl vidět řádky, kde byl kód 80403, který měl být dle číselníku spojený s jednotkou Kč. A v hodnotách jsem dostával čísla jako 154, 152, 140, 58, 50, což jsem vyhodnotil, že moc nesedí, aby byla měsíční mzda. A naopak, když jsem dal druhou hodnotu v číselníku, která měla být spojena s tisíci lidmi zaměstnanými v daném odvětví, do podmínky (tedy 200), dostával jsem hodnoty 8793, 11521, 13430 a vyšší. To opět nesedělo na tisíce obyvatel zaměstnaných v jednotlivých oborech, proto jsem

usoudil, že tyto dva kódy jsou jen navzájem otočené a v dalších krocích jsem to bral v potaz.

Dále jsem věděl, že musím pro srovnatelné období ve výsledné tabulce dostat rok i z tabulky czechia_price. Nad tímto krokem jsem v tabulce czechia_payroll přemýšlet nemusel, jelikož tam byl přesně daný. V tabulce czechia_price však byly dva datumy date_from a date_to, které vyjadřovaly, kdy se začalo měřit a kdy měření skončilo. Pro jeden řádek bylo první datum například 2017-11-13 01:00:00.000 +0100 a druhé 2017-11-19 01:00:00.000 +0100. Pomocí funkce date_part jsem si vytáhl jednotlivé roky ze sloupce date_from.

Pro jistotu jsem se ještě podíval na to, zda nejsou nějaké rozdíly ve výstupech z date_from a date_to (v letech). To jsem si vytáhl právě roky z obou sloupců a přidal dynamický sloupec, který by mi načetl NULL hodnoty, pokud by tyto dva roky spolu neseděly. Tím, že funkce COUNT potom nepočítá NULL hodnoty, stačilo si udělat tento dotaz a ten by měl být stejný s počtem, na který si kliknu ještě v dotazu předtím (bez COUNT) dole u tabulky “Calculate total row count” a to se stalo, byl stejný.

Nakonec jsem vytvořil výslednou tabulku t_Lukas_Vesely_project_SQL_primary_final pomocí několika LEFT JOINů s všemi daty o mzdách a cenách potravin s tím, že výsledná moje tabulka obsahovala 11 sloupců a přes 15 milionů řádků. To z toho důvodu, že při spojování jednotlivých tabulek jsem dostal id jiná pro mzdy, ale pro potraviny jsem dostal vždy několik stejných id, s čímž jsem musel potom na další kroky s tabulkou myslet, jelikož 15 milionů měření neproběhlo. Určitě tabulka šla udělat lépe a bez méně řádků, protože 15 milionů řádků je opravdu hodně, ale takhle mi to přišlo taktéž v pohodě.

Celkově jsem dával i podmínky, aby kód regionu v sobě neměl NULL hodnoty, pak to byly hodnoty a odvětví (industry_branch_code).

Vytvoření sekundární tabulky - t_Lukas_Vesely_project_SQL_secondary_final

Pro vytvoření sekundární tabulky, kde měla být data o dalších evropských státech s HDP, GINI koeficientem a populací, jsem věděl, že budu především spojovat tabulky countries a economies. Vytáhl jsem si tedy z obou tabulek to nejdůležitější pomocí WITH a JOINu – country, year, gdp, gini, population, continent. Na následné vytvoření celkové tabulky jsem už jen přidal podmínku, aby kontinent byl Evropa a data o evropských státech byla na světě. Výsledná tabulka obsahovala 6 sloupců a 548 řádků.

Celkově jsem dával podmínku, aby HDP nemělo v sobě NULL hodnoty.

Otázka 1 (Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?)

Pro první otázku jsem používal primární tabulku t_Lukas_Vesely_project_SQL_primary_final. Tím, že jsem měl pracovat pouze s daty o mzdách, vytáhl jsem si pouze potřebná data – id pro mzdy, rok, mzdu, kód odvětví a název odvětví. Pro jistotu jsem si dal id pro mzdu do funkce DISTINCT, aby se mi nestalo, že některá měření se mi v tabulce objeví vícekrát. Na toto jsem si vytvořil pohled salary_industry.

Dále jsem potřeboval průměrnou mzdu v jednotlivých odvětvích a v jednotlivých letech. Proto jsem použil do SELECTu funkci avg, která mi dá aritmetický průměr mezd. Pokud se v SELECTu objeví něco jiného než pouze agregační funkce, musí se v dotazu použít GROUP BY – tady jsem musel do GROUP BY dát tři sloupce, které se mi objevily v SELECTu a to year_salary, industry_branch_code_salary a industry_name_salary. Aby se mi sloupec avg(value_salary) nepojmenoval čistě avg, dal jsem mu konkrétní název average_salary.

Rád pracuji s Excelem, proto jsem si data vyexportoval do Excelu takhle sloučených do jednotlivých let a odvětví. Pomocí flagu (dalšího sloupce) jsem si jen rychle rozřadil, ve kterých letech mzdy v jednotlivých odvětvích rostou a kdy klesají (zároveň jsem si vymazal přechody mezi odvětvími - určitě by to šlo složitějším KDYŽ, ale tím, že jsme neměli stovky řádků, rozhodl jsem se to udělat ručně). Použil jsem funkci COUNT pro spočítání, kolik mám celkově “Roste” a kolikrát se mi ve flagu objevilo “Klesá”. Vyšlo mi 205krát “Roste” a 23krát “Klesá”.

Zkusil jsem se podívat na to, kde a jak moc se mi v jednotlivých objeví “Klesá” a udělal si jen pomocnou tabulku. Hodnotil jsem počet i jak moc dramatický pohyb byl. Většina “Klesá” byla v roce 2012, to mohl být finální dopad finanční krize v letech 2008 až 2012 na českou ekonomikou.

Celkově mi vyšlo, že několikrát mzda klesla v odvětví “Těžba a dobývání”. Jinak hodnoty, kdy třeba dvakrát lehce mzdy klesly jsem nepovažoval za dostatečné pro uznání, že mzdy v tomto odvětví klesají.

Otázka 2 (Kolik je možné si koupit litrů mléka a kilogramů chleba za první a poslední srovnatelné období v dostupných datech cen a mezd?)

Pro druhou otázku jsem používal primární tabulku t_Lukas_Vesely_project_SQL_primary_final. Pokud jsem dal funkci DISTINCT pro vytáhnutí jedinečných id (měření) pro mzdy, bylo jasné, že pro potraviny to bude důležitější z důvodu opakování řádků v primární tabulce (viz. Vytvoření primární tabulky). Tím, že jsem měl pracovat pouze s daty o potravinách, vytáhl jsem si opět jen data, co potřebuji - id pro potraviny, rok (sloupec se jmenuje year_salary, ale tento rok je v primární tabulce spojovníkem jak pro mzdy, tak pro potraviny), cenu,

kód kategorie, název kategorie, kód regionu a region. Na toto jsem si vytvořil pohled, který jsem pojmenoval food.

Abych zjistil první a poslední srovnatelné období, pro která budu výslednou odpověď na otázku dělat, musel jsem nejdříve zjistit, jaký rok mám pro chleba (kód 111301) nejnižší a nejvyšší a pro mléko (kód 114201) nejnižší a nejvyšší. To jsem docílil tak, že jsem si jen z celkového pohledu food dal podmínku nejdřív jednoho kódu kategorie a funkce min a max, které hledaly nejnižší a nejvyšší hodnotu ve sloupci rok. Oba výsledky s min vyšly pro rok 2006 a oba výsledky s max vyšly pro rok 2018. Zjištění nejnižší hodnoty a nejvyšší hodnoty mě čekaly i pro mzdy, jelikož, kdyby ve funkci min vyšel například rok 2007, nemůžu dále mít spojovník rok 2006 jako první srovnatelné období. To se však nestalo a nejnižší hodnota byl rok 2006 a nejvyšší hodnota byl rok 2018. Věděl jsem tedy, že mé první srovnatelné období bude rok 2006 a poslední srovnatelné období bude rok 2018.

Vypočítal jsem si průměrnou cenu chleba a mléka za roky 2006 a 2018 a měl jsem 4 hodnoty, které potom použiju pro výsledný výpočet. To stejné jsem si udělal pro mzdy - vypočítal průměrnou mzdu ve všech odvětvích pro roky 2006 a 2018.

Věděl jsem, že pro vypočítání, kolik kilogramů chleba a litrů mléka si mohu koupit budu potřebovat mít průměrné hodnoty v řádku, abych je potom mohl mezi sebou vydělit a dostat celkový výsledek. Proto jsem použil WITH pro nadefinování 5 sloupců (as6 - průměrná mzda v letech 2006 a 2018, b6 - průměrná cena chleba v roce 2006, b18 - průměrná cena chleba v roce 2018, m6 - průměrná cena mléka v roce 2006 a m18 - průměrná cena mléka v roce 2018).

Takhle jsem dostal vše potřebné v řádcích a k dotazu jsem ještě přidal 4 dynamické sloupce a upravil je na to, že pokud b6 průměrná cena je nenulová, udělej “průměrná mzda děleno průměrná cena” a takhle stejně pro b18, m6 a m18.

Potom mi došlo, že abych tam neměl NULL hodnoty, tak mohu upravit dotaz, aby tam vždy bylo IN(2006, 2018), a tak jsem dotazy ještě takhle upravil.

Celkově mi vyšlo, že v roce 2006 jste si mohli průměrně koupit asi 1287 kilogramů chleba nebo 1437 litrů mléka a v roce 2018 asi 1342 kilogramů chleba a 1642 litrů mléka.

Jsou to samozřejmě výsledky s nebo. Odpověď na otázku by šlo třeba pochopit i tak, že kolik je možné si koupit kilogramů chleba a litrů mléka tak, aby byly obě hodnoty stejné nebo třeba by to šlo pochopit i tak, že chceme, abychom za obě kategorie zaplatili v nějakém poměru.

A také by se mohla odpověď udělat konkrétněji pro všechna odvětví zvlášť pro rok 2006 a 2018.

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

Pro třetí otázku jsem používal primární tabulku t_Lukas_Vesely_project_SQL_primary_final. Pracoval jsem s již vytvořeným pohledem food, který vycházel z primární tabulky. Malá poznámka ještě, že celkové nárůsty jsem dělal v jednotkách stejných, jako jsou v jednotlivých měřeních, že jsem neupravoval např. aby pivo a mléko obě kategorie byly v litrech, ale pivo jsem nechal v půlitrových lahvích.

Přemýšlel jsem nad tím tak, že budu potřebovat pro nějaké procentuální rozdíly mít všechna data v řádcích, abych potom vypočítal v dynamickém sloupci procentuální rozdíly.

Proto jsem si dal nejdříve do SELECTů rok, kategorii potravin, název kategorie a průměrnou cenu potravin (avg(price_food)). Rok, kategorii potravin a název kategorie jsem opět dával do GROUP BY a vždy jsem si dal SELECT pro daný rok – 2006 až 2018.

Ze všech SELECTů, které jsem dal následně do WITH, jsem si vytvořil nový pohled food_average_prices.

Teď jsem měl všechna data potřebná v řádcích a mohl jsem s tím dál pracovat. Potřeboval jsem udělat procentuální rozdíly mezi jednotlivými lety, proto jsem si udělal 12 dynamických, kde jsem měl vždy rozdíl “roku + 1” / “roku”. Kdybych měl jen tento formát úpravy, dostával bych čísla jako 0,123; 0,111 a podobně. Proto jsem ještě přidal krát 100, abych dostal procenta. Takhle jsem měl však celkovou hodnotu včetně základu původního roku. Proto jsem odečetl 100 a dostal už jen rozdíl mezi jednotlivými lety v procentech. Pro toto jsem vytvořil pohled food_differences.

Dále jsem si řekl, že potřebuji dostat z těchto všech rozdílů dostat jedno - průměr. To jsem udělal pomocí:

```
SELECT
category_code_food,
name_food,
(difference_2007_and_2006 + difference_2008_and_2007 + difference_2009_and_2008 +
difference_2010_and_2009 +
difference_2011_and_2010 + difference_2012_and_2011 + difference_2013_and_2012 +
difference_2014_and_2013 +
difference_2015_and_2014 + difference_2016_and_2015 + difference_2017_and_2016 +
difference_2018_and_2017) / 12 AS average_difference
FROM food_differences;
```

Dále jsem si ve sloupečku našel nejnižší hodnotu, abych zjistil, kde se nachází potravina, která zdražuje meziročně nejpomaleji. To jsem udělal pomocí funkce min, do které jsem dal opět sloupec average_difference. Vyšlo mi jedno číslo, to jsem dal do podmínky s WHERE a chtěl jsem, aby mi dotaz ukázal už jen pouze název kategorie.

Celkově vyšlo, že kategorie, u níž je nejnižší percentuální meziroční nárůst je Cukr krystalový s tím, že meziročně podle dat jeho cena dokonce lehce klesá.

Určitě by se otázka mohla dělat i konkrétněji například, v některých regionech něco zdražilo výrazně víc než v jiném, tedy dalo by se to udělat zvlášť pro jednotlivé regiony.

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

Pro čtvrtou otázku jsem používal primární tabulku t_Lukas_Vesely_project_SQL_primary_final. Nejdříve jsem si vytvořil dotaz, který by mi ukázal, průměrný rozdíl v jednotlivých letech z pohledu food_differences. Data jsem měl ideálně připravená i pro to, kdyby se mělo zjistit, jaké konkrétní potraviny měli víc jak 10% meziroční nárůst oproti mzdám nebo mzdám v konkrétních odvětvích.

To pro potraviny bylo třeba i pro mzdy. Vytvořil jsem si stejnou tabulku jako pro potraviny (pohled food_average_prices), tedy pro všechna odvětví průměry mezd pro jednotlivé roky. Dále pomocí opět 12 dynamických sloupců rozdíly “roku + 1” / “roku” vynásobené 100 a od toho odečteno 100. Toto opět dal do pohledu a to salary_differences.

Pak jsem si udělal průměrné rozdíly v jednotlivých odvětvích, ale to jsem nakonec nijak nepoužil.

Vytáhl jsem si průměrné rozdíly pro všechny potraviny (to, co jsem udělal v prvním dotazu pro tuto otázku), průměrné rozdíly pro všechna odvětví. Teď jsem dostal dvakrát 2 řádky po 12 sloupcích a řekl jsem si, že abych dále nějak porovnával hodnoty mezi mzdami a potravinami, musím dostat sloupce do řádků a naopak. Uvědomil jsem si, že toto mohu udělat v Excelu. Tudíž jsem si exportoval oba SELECTy pro rozdíly všech potravin a mezd do CSV a otevřel v Excelu. Řádky jsem si přehodil do sloupců pomocí vložení “Transponovat”. Udělal jsem si Flag – difference, kde jsem odečítal vždy hodnotu salary od food, abych měl odpověď na otázku, zda existuje rok, kdy byl meziroční nárůst cen potravin výrazně vyšší než růst mezd (větší než 10 %). Potom se dalo jen třeba podívat se na data a říct, že nikde není, ale aby to bylo za mě kompletní, tak jsem se vrátil do DBeaveru a vytvořil novou tabulku a materializovaný pohled a pomocí CASE jsem si dal, **jestli existuje rok, kdy byl tento nárůst větší než 10 % a celkově vyšlo, že žádný takový rok neexistuje.**

Fakt, že by ceny potravin vzrostly o 10 % více než mzdy je už tak celkem odvážný, avšak tím, že se v datech mezd nepočítalo s inflací, přeci jen by to v nějakém roce

mohlo nastat. Největší rozdíl mezi potravinami a mzdami byl v roce 2013, konkrétně se jednalo asi o 6,8 %.

Otázka 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 pátou otázku jsem používal sekundární tabulku t_Lukas_Vesely_project_SQL_secondary_final. Jestli má HDP vliv na změny ve mzdách a cenách potravin jsem si ze sekundární tabulky vytáhl data, kde country = ‘Czech Republic’, abych dostal pouze informace o České republice. Nejdříve jsem se pustil do toho, abych vliv zjistil vůči mzdám. Proto jsem si vytáhl z již vytvořeného pohledu salary_differences odvětví a jednotlivé rozdíly mezi lety. Data jsem měl ideálně připravená i pro to, kdyby se mělo zjistit, vůči kterým jednotlivým odvětvím má například větší vliv a na která odvětví růst HDP vliv nemá.

Přemýšlel jsem nad tím tak, že opět potřebuji dostat data do řádků, abych s nimi mohl počítat dál a opět udělat procentuální rozdíl mezi dvěma roky. Dal jsem si tedy do SELECTu například toto a poté jen měnil roky (tím, že společná období dat mezd a potravin byly od roku 2006 do roku 2018, tyto dotazy jsem dělal taktéž od roku 2006 do roku 2018):

```
SELECT
"year",
gdp AS gdp_2006
FROM t_lukas_vesely_project_sql_secondary_final
WHERE country = 'Czech Republic'
AND "year" = 2006;
```

Potom pomocí WITH a LEFT JOINů jsem si všechny informace spojil (a vytvořil pohled gdp_years) a dále pomocí 12 dynamických sloupců vypočítal jednotlivé rozdíly mezi dvěma sousedními roky a toto celé uložil do pohledu gdp_differences.

Teď už to stačilo porovnat jen se mzdami. Tím, že jsem měl data pro jednotlivá odvětví, udělal jsem sumu “roku + 1” / “roku” a jako v otázkách předtím krát 100 mínus 100. Vycházel jsem z pohledu, kde mám uložené rozdíly v jednotlivých letech – slaray_differences.

To stejné jsem udělal pro potraviny. Tedy jsem vycházel z pohledu, kde mám uložené rozdíly v jednotlivých letech – food_differences. Tím, že jsem měl data pro jednotlivé potraviny, udělal jsem sumu “roku + 1” / “roku” a jako předtím krát 100 mínus 100.

Všechny 3 výsledky jsem si exportoval do Excelu, abych si je opět přesunul ze sloupců do řádků pomocí vložení jinak – Transponovat. Jednu tabulku jsem si

vytvořil pro mzdy a HDP v jednotlivých rozdílech let a druhou pro potraviny a HDP v jednotlivých rozdílech let.

Když jsem se podíval na růsty a poklesy v HDP, stanovil jsem si, že pojem "výraznější růst" pro mě bude růst o více jak 4,5 %. Takový růst byl zaznamenán ve třech letech, a to v roce 2007, v roce 2015 a v roce 2017. Zároveň aby se jednalo projev na cenách potravin a mzdách, řekl jsem si, že to pro mě bude hranice opět 4,5 % v tom samém roce nebo v roce následujícím.

Člověk tím, že měl jen 12 rozdílů v 13 letech, potom celkovou odpověď mohl pouze vyčíst z porovnání, že se podívá na tyto tři roky a jestli je procentuální rozdíl taktéž větší než 4,5 v to samém roce nebo v tom následujícím.

Chtěl jsem však mít vše v DBeaveru kompletní, proto jsem chtěl sestavit dotaz, který by i pro větší počet dat fungoval bez toho, aby bylo potřeba to analyzovat manuálně. Vytvořil jsem si dvě tabulky, jednu pro mzdy a HDP, druhou pro ceny potravin a HDP.

Tím, že jsem neměl ještě nikde data (ty jsem chtěl až potom vkládat z Excelu), vytvořil jsem ji jen na základě nuly. Pak jsem tam přidal vždy dané sloupce – u mezd to byly sloupce years, salary, gdp a gdp_plus_one_year, u potravin years, food, gdp a gdp_plus_one_year s tím, že sloupce salary a food obsahovaly právě průměrné procentuální rozdíly mezi jednotlivými lety z pohledů salary_differences a food_differences (konkrétně ze sum všeho v "roce + 1" / "roce").

Do tabulky jsem teď musel vložit data z Excelu. V Excelu jsem tedy začal pracovat v dalším sloupci, kde jsem pomocí spojování textů dohromady a odkazování na čísla a texty v buňkách vytvořil 12 dotazů (vytvořil jsem jen jeden a ostatní se potom podle toho zkopírovaly podle buněk) pro mzdy a 12 dotazů pro ceny potravin. Tyto dotazy jsem použil klausuli INSERT INTO a název tabulky (pro mzdy salary_gdp, pro potraviny food_gdp) a data byla opět v DBeaveru.

Ještě nutno říct, že pro sloupec gdp_plus_one_year bylo nutné mít první řádek bez dat (NULL), abych vlastně porovnával HDP z jednoho roku a mzdy a potraviny v roce následujícím, aby se mi všechna tato data o HDP posunula o jeden rok dolů.

Teď už jen nějak rozřadit, kdy tedy měl růst HDP vliv na mzdy a ceny potravin. To jsem opět využil dvou dynamických sloupců s tím, že první ukazoval, zda HDP růst měl vliv na mzdy a potom i ceny potravin v tom samém roce, druhý dynamický sloupec to dělal pro rok následující.

Vytvořil jsem takový dotaz pro mzdy a pro potraviny byl téměř identický:

```
SELECT
*,
CASE
	WHEN gdp > 4.5 AND salary > 4.5 THEN 1
	ELSE 0
END AS rise_same_year,
CASE
	WHEN gdp_plus_one_year > 4.5 AND salary > 4.5 THEN 1
	ELSE 0
END AS rise_plus_one_year
FROM salary_gdp;
```

Potom podle 0 a 1 by se dalo i ve větším množství dat říct, v jakých letech vliv HDP měl a kdy ne.

Z tabulky pro mzdy a HDP nakonec vyšlo, že pokud porovnáme mzdy a HDP, vliv HDP z roku 2007 na mzdy byl v tom samém roce (2007) i v roce následujícím (2008) a vliv HDP z roku 2017 na mzdy taktéž v tom samém roce (2017) i v následujícím (2018). Z tabulky pro ceny potravin a HDP vyšlo, že pokud porovnáme potraviny a HDP, vliv HDP z roku 2007 na ceny potravin byl v tom samém roce (2007) i v tom následujícím (2008) a vliv HDP z roku 2017 byl pouze v tom samém roce. Ovšem rok 2015 z tohoto přehledu absolutně vypadl, tedy by se dalo říct, že v těchto 24 rozdílech mělo HDP dopad na ceny a potraviny ve 4 z 6 případů.

V celém projektu šlo určitě několik věcí udělat jednodušeji (například porovnávání rozdílů nebo vytváření tabulky pro porovnání mezd a HDP a cen potravin a HDP), ale věřím, že i tyto cesty, kterými jsem se vydal jsou správné.

Lukáš Veselý