ENGETO - DATOVÝ ANALYTIK S PYTHONEM – PROJEKT 4 – SQL

POPIS ZADÁNÍ

V rámci zkoumání životní úrovně občanů je úkolem odpovědět na 5 výzkumných otázek z oblasti vývoje mezd a dostupnosti základních potravin široké veřejnosti.

ANALÝZA DAT

Data o cenách vybraných potravin

Při kontrole četnosti dat bylo zjištěno, že dvě položky mají odlišný počet výskytů v porovnání s ostatními položkami:

- Kapr živý – nižší četnost záznamů je způsobená faktem, že se tato položka prodává pouze v předvánočním období. Tento fakt nevede k nekonzistentnosti datového modelu, protože do datového zdroje sql_primary_final je zaznamenávána průměrná cena za položku za daný rok.
- Jakostní víno bílé – nižší četnost záznamů je způsobena tím, že se tato položka začala zaznamenávat až v roce 2015. Pro zajištění konzistentnosti dat v jednotlivých letech a možnost porovnávaní meziročního nárustu cen je nutné mít v každém roce stejné zastoupení potravin. Proto bude tato položka vyloučena z datového zdroje sql_primary_final.
- Ostatní položky – v letech 2006-2010 bylo zaznamenávání údajů ve větší frekvenci, avšak od roku 2011 se četnost ustálila na 1 měření za každý region pro každou potravinu. Tento fakt neovlivňuje konzistenci datového modelu, protože do datového zdroje sql_primary_final je zaznamenávána průměrná cena za položku za daný rok.

Dále je užitečný poznatek, že každé měření pro jednu potravinu v jednom časovém úseku obsahuje 15 záznamů – jeden pro každý ze 14 regionů uvedených v číselníku czechia_region, a jeden pro jejich průměr (záznam s chybějící hodnotou ve sloupci region_code). Jelikož v naší analýze nás nebude zajímat granularita na úrovni regionů, do datového zdroje sql_primary_final využijeme pouze tuto průměrnou hodnotu.

Data o mzdách

Číselník czechia_payroll_unit je chybný, protože průměrná hrubá mzda se vykazuje v jednotkách měny a průměrný počet osob v osobách. Tento číselník z toho důvodu opomeneme a jednotky si domyslíme.

123 count	123 round	A-Z name	A-Z name
3,268	24,647.79	tis. osob (tis. os.)	Průměrná hrubá mzda na zaměstnance
3,268	202.55	Kč	Průměrný počet zaměstnaných osob

Četnost záznamů je pro každé odvětví v každém roce a pro každý typ kalkulace (průměrná hrubá mzda vs. průměrný počet zaměstnanců, fyzický vs. přepočtený) stejná, data tudíž nevykazují žádné nekonzistence. Bylo ověřeno, že záznamy o průměrných mzdách neobsahují nulové hodnoty ve sloupci value.

Data se dělí podle číselníku calculation_code na hodnoty:

- Za zaměstnance ve fyzických osobách
- Za přepočtený počet zaměstnanců = je přepočtem průměrného evidenčního počtu zaměstnanců ve fyzických osobách podle délky jejich pracovních úvazků na zaměstnavatelem stanovenou (plnou) pracovní dobu. Dále v textu bude uvedeno pod zkratkou FTE.

Zadání nespecifikuje, zda je předmětem zájmu statistika průměrných mezd přepočtena na fyzické osoby nebo na FTE. V datovém zdroji sql_primary_final budou zahrnuty obě varianty. Výzkumné otázky budou řešeny pro průměrné mzdy

přepočtené na fyzické osoby, přičemž možnost přepočtu na FTE bude uvedena v jednotlivých scriptech formou poznámky.

Data o HDP, GINI…

Součástí zadání bylo i vytvořit dodatečný materiál obsahující HDP, GINI koeficient a populaci evropských zemí. Tyto data byli zahrnuty v datovém zdroji sql_secondary_final.

DATOVÉ ZDROJE

t_livia_crhova_project_SQL_primary_final

1) Sloupec „year“ – rok
2) Sloupec „branch“ – označení odvětví pro údaje o průměrné mzdě ve struktuře:
    I. Odvětví
    II. Text „fyzický“ nebo „přepočtený“ – podle kalkulace na fyzické osoby nebo FTE

    Příklad „Ostatní činnosti-fyzický“
3) Sloupec „avg_pay“ – hodnota průměrné mzdy v daném roce a v daném odvětví
4) Sloupec „foodstuff“ – název potraviny
5) Sloupec „avg_price“ – hodnota průměrné ceny potraviny v daném roce

Datový zdroj je kombinací průměrných mezd a průměrných cen v letech, kdy jsou dostupná data pro obě skupiny.

t_livia_crhova_project_SQL_secondary_final

1) Sloupec „country“ – krajina
2) Sloupec „year“ – rok
3) Sloupec „gdp“ – hodnota HDP v daném roce
4) Sloupec „gini“ – GINI koeficient v daném roce
5) Sloupec „population“ – velikost populace v daném roce

Datový zdroj je omezen na časové období odpovídající sledovanému období v zdroji sql_primary_final.

VÝZKUMNÉ OTÁZKY

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

V průběhu let se vyskytují v některých odvětvích meziroční poklesy mezd v porovnání s předchozím rokem. Takový pokles byl zaznamenán především v roce 2013, kdy mzdy meziročně klesli v 10 sledovaných odvětvích.

Porovnání průměrných mezd v prvním a posledním sledovaném roce však potvrzuje celkový rostoucí trend ve všech sledovaných odvětvích.

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

Dle analýzy, za průměrnou měsíční hrubou mzdu fyzického zaměstnance (opomeneme-li zdanění a odvody) bylo možné si koupit:

- V prvním srovnatelném období, tedy v roce 2006, buďto 1408 litrů mléka nebo 1261 kg chleba
- V posledním srovnatelném období, tedy v roce 2018, buďto 1613 litrů mléka nebo 1319 kg chleba

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

Porovnáním průměrných cen vybraných potravin v prvním a v posledním sledovaném roce (2006 a 2018) se ukázalo, že nejnižší procentuální meziroční nárůst v sledovaném období zaznamenala kategorie Banány žluté.

Jsou ale dvě kategorie potravin (Cukr krystalový a Rajská jablka červená kulatá), které v sledovaném období zlevnili, a tudíž se nejedná o nárůst.

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

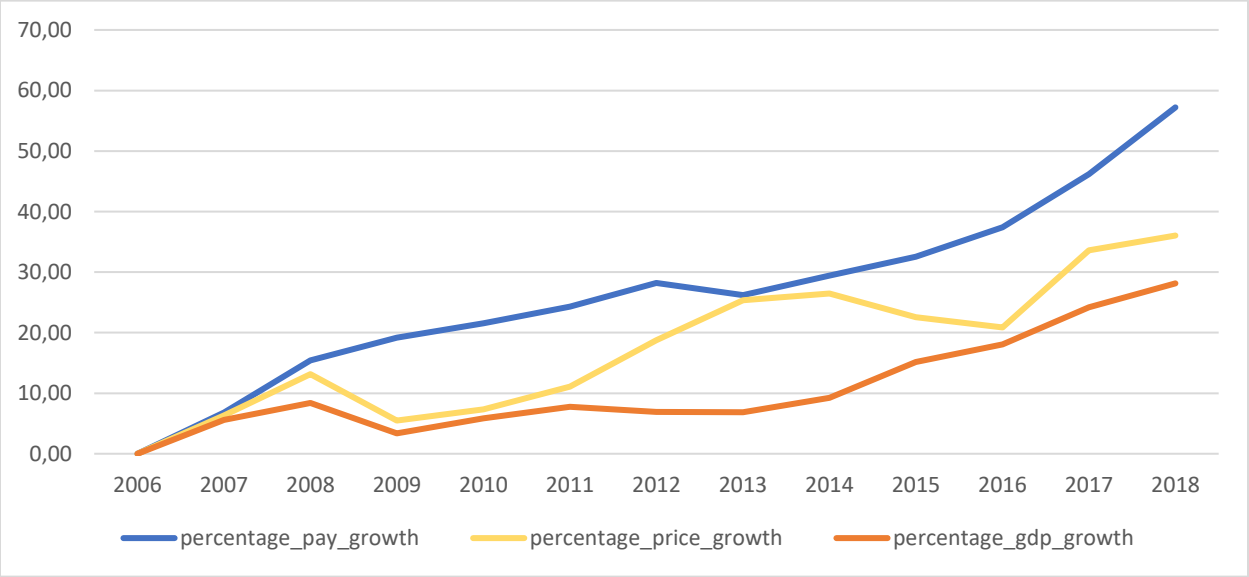
Pro porovnání meziročního nárustu cen s meziročním růstem mezd bylo potřebné zjistit v prvním kroku průměrné ceny a mzdy v jednotlivých letech a jejich meziroční změny.

Tato analýza ukázala, že v žádném ze sledovaných období nenastala situace, kdy by nárůst průměrných cen potravin výrazně překročil (vice než 10 %) růst průměrných hrubých mezd.

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?

Aby bylo možné srovnání tempa nárustu cen, mezd a HDP, byla zvolena indexová metoda. V tomto případě se použil první rok sledovaného období jako základní referenční hodnota, a následně se počítala změna v procentech ve srovnání s tímto základem. To umožňuje porovnávat růst mezi jednotlivými roky na základě konstantní výchozí hodnoty (tedy hodnoty z prvního roku).

Výsledek indexového srovnání tempa růstu průměrných cen, průměrné hrubé mzdy a HDP je zobrazen na grafu:



Tempo růstu mezd dosáhlo vrcholu v roce 2008. Následně bylo toto tempo nízké až do roku 2013, kdy dokonce dosáhlo záporných hodnot. Od roku 2014 však tempo růstu mezd stále rostlo až do konce sledovaného období.

Tempo růstu HDP vykazovalo volatilitu. V období 2007 - 2018 mělo tři vrcholy, kdy meziroční nárůst indexu přesáhl 5 % - v letech 2007, 2015 a 2017.

Vzhledem na tempo růstu mezd je možné odhadnout souvislost, podle které vrchol růstu HDP v roce 2007 mohl mít vliv na vysoké tempo růstu mezd v letech 2007-2008. Dále nízké až záporné tempo růstu HDP v letech 2008-2014 může

souviset s nízkým až záporným tempem růstu mezd v letech 2009-2016. Následně období 2015-2018 vykazuje dva vrcholy tempa růstu HDP vyšší než 5 %, což koreluje s vysokým tempem růstu mezd v letech 2017-2018.

Pokud jde o vývoj cen, tempo jejich růstu v letech 2007-2011 korelovalo s tempem růstu HDP, přičemž křivky meziročních změn cen a HDP vykazovaly podobnou trajektorii. V letech 2012-2016 však vykazovalo tempo růstu cen opačný trend než tempo růstu HDP. V letech 2017-2018 už znovu jejich tempa růstu korelují.

Ve všeobecnosti ale není možné jednoznačně stanovit jasný vztah mezi vlivem vývoje HDP na vývoj cen a mezd, protože takto zjednodušená analýza neposkytuje dostatek informací o směru příčinné souvislosti ani o dalších proměnných vstupujících do tohoto vztahu.