

Primary and Secondary Tables data

1. Primary table creation

We have created a primary table named 't_jelena_puzova_project_SQL_primary_final'. It was compiled by using two different tables: wage data (from 'czechia_payroll') and food price data (from 'czechia_price'). These tables were chosen because they contain all the data needed for analysis. What is more, the rest of the tables ('czechia_payroll_calculation', 'czechia_payroll_industry_branch', 'czechia_price_category', ect.) are linked to them (with foreign keys).

Wage Data Section

- Key columns include: 'id'-useful for identifying individual rows, 'value'- e.g. wage amount, average salary, etc, 'payroll_year' and 'payroll_quarter'- time reference for the wage.
- The columns 'category_name', 'price_value', and 'price_unit' do not exist in wage data but are required to match the structure of the food price data (in the wage data, there is no food category, and wages aren't measured in kilograms or liters).
- 'NULL::type' is used to insert 'NULL' values with the correct data type (e.g. text, numeric).
- Reference tables got joined to get:
 - ✓ the name of the calculation (e.g. "Average gross wage"),
 - ✓ the industry name,
 - ✓ the unit name (e.g. "per employee"),
 - ✓ the value type.
- 'LEFT JOIN' is used to make sure no wage records are lost, even if some codes are missing in the reference tables.

Food Price Section

- Since that table doesn't include a year column directly, we extract the year from the 'date_from' column using 'EXTRACT(YEAR FROM date_from)'. - 'The payroll_quarter' column does not apply to prices, so we set it to 'NULL' (Food prices are not quarterly, so we insert 'NULL').
- Key food price information is included by joining the 'czechia_price_category' table:
 - ✓ 'category_name' (e.g. "Milk"),
 - ✓ 'price_value' (e.g. 1.0),
 - ✓ 'price_unit' (e.g. "l" for liter or "kg" for kilogram).
- Because prices are not related to wage calculations, industries, or units, those columns are set to 'NULL' as well.
- Value type using: 'CONCAT('Food: ', cpc.name) AS value_type_name' to clearly label food price rows (e.g. "Food: Milk").
- Only national average prices are included using the filter 'WHERE region_code IS NULL'.
- We use 'UNION ALL' to combine all rows from both sections (wages + food prices) into one single table.

Abnormalities found:

- Lot's of 'NULLs' (Wage rows: food-related columns = 'NULL', Food price rows: wage-related columns = 'NULL').
- Confusion in time structure (Wages have year + quarter, Food prices have only year no quarter, just 'NULL').
- In the dataset, all records where 'value_type_name' = 'Průměrná hrubá mzda na zaměstnance' represent the average gross monthly wage per employee. However, this is not explicitly stated in the data structure — the values appear without any currency unit. In reality, these values are in Czech koruna (CZK), but the dataset does not indicate this anywhere, which may lead to confusion.

II. II. Secondary table creation

The secondary table 't_jelena_puzova_project_SQL_secondary_final' combines country-level geographic data with economic indicators such as GDP, population, and GDP per person. It is used to explore whether economic growth influences changes in wages and food prices over time. It also helps answer the last question of the project.

Data Section

The table 'countries' is taken as the base since it gives us geographic context like 'region', 'continent', and 'ISO code'. It also ensures that even if a country has no GDP for a certain year, we still keep the structure.

- 'LEFT JOIN' with another table 'economies' is used. This brings in the 'gdp', 'population', and 'year' as it is essential to the research. Joining this way preserves country data and connects it with actual economic performance.
- Selecting 'gdp', 'population', and calculating 'gdp_per_person'. 'Population' is necessary for per-person analysis across the countries. This part of code calculates the average GDP per person by dividing a country's total GDP by its population, converting the result to a numeric type, and rounding it to 2 decimal places. This also allows you to track not just total growth, but how people are affected individually.
- Filtering out NULLs ensures that the table is clean, and only countries and years where we can actually analyze trends are included.

Abnormalities found:

- When joining countries with economies, both tables had a column named 'population'. We used 'population', which comes from 'economies' for the current population for the year.
- Some countries in the 'countries' do not appear in 'economies', or have missing GDP values in certain years (Anguilla, Antarctica, Bahamas, East Timor, French Guiana, ect.).