Luka Žeravica

Croatian Local Administration ReadMe

Project Objective

The objective is to present a descriptive statistical analysis of Croatian local administrative units, i.e., counties and municipalities. The analysis includes the population (migration) and economy of each county. From a technical standpoint, one of the project objectives was to create a county database and a semantic model in Power BI that contains multiple data sources (in this case MS Excel and PostgreSQL database) and stores it in one database. This project was made from scratch.

Methods Used

Data Import

Connection to Database

Web Scraping

Database Design

Descriptive Statistics

Data Visualization

Technologies

PostgreSQL

Python

Power BI

Source of Datasets

All of the datasets (csv tables) were taken from Croatian Statistics Bureau website and Wikipedia for web scraping.

Croatian Statistics Bureau:

https://podaci.dzs.hr/en/statistics/population/ (Visited: 12.12.2024.)

Wikipedia:

Member States of European Union

https://en.wikipedia.org/wiki/Member_state_of_the_European_Union

Counties of Croatia

https://en.wikipedia.org/wiki/Counties_of_Croatia

Code Description

Python

Datasets (MS Excel and CSV files) df17 and df21 were imported in Python. Connection to PostgreSQL was established and SQL queries for table creation were executed. The tables were filled with the imported data (df17 and df21). Web scraping was used to scrap European Union Countries and counties' area tables from Wikipedia. The table was transformed into a dataframe and a new table (eu_countries and county_area) in PostgreSQL was made and filled with values. Columns were renamed and unnecessary ones were removed. Likewise, the county_area table was scraped from Wikipedia and was created in PostgreSQL. In the end, the connection to PostgreSQL was closed.

PostgreSQL

After both tables (administrationdata 2017 and 2021) were imported, the names of the columns were changed to be distinguished from each other. These columns are: county_municipality_or_city_mayor, deputy_mayor, second_deputy_mayor, minority_mayor, and president_of_representitative_body. Each column has a year abbreviation at the end of its name (e.g., deputy_mayor21). Furthermore, county names and administrative unit types were translated into English. Tables counties, administrative_units, and administrative_units_incumbents were created out of administrationdata17/21 tables. Id primary keys were created in the counties and administrative_units tables. Each county received the id value according to its number (e.g., City of Zagreb is the 21st county, thus, the id for this county is 21).

The administrative_unit_incumbents table consists of persons_name, admin_unit_id, function_id, and year. Column persons_name contains data for each incumbent, admin_unit_id is the id of the administrative unit (county, city, and municipality), and function_id is the id of the incumbents' function. Incumbent function id is mayor (id=1), deputy mayor (id=2), second deputy mayor (id=3), minority deputy mayor (id=4), and president of representitative body (id=5). Person_id column was added as well as foreign keys (admin_unit_id). Duplicates were deleted. The people table was made out of the administrative_unit_incumbents table and the person_id column was transferred as a primary key (of the people table). A unique index was made on persons_name in the people table so duplicates would be avoided.

Regarding data manipulation language, count as aggregate function was used for each administrative unit (city and municipality) and incumbent function (mayor, deputy mayor, second deputy mayor, minority mayor, and president of representative body). Also, there was a query about reelected and newly elected incumbents.

Power BI

The tables used have been imported from two sources: PostgreSQL and MS Excel. Regarding the first, Power BI was connected with the database and the next tables are imported: counties, county_area, and administrative_units. Table county_area was web scrapped (via Python) from the internet. It was split into 2 columns: county_area_km2 and county_area_sq_mi. Values were replaced; commas (,) and brackets were removed so column data type could be changed to decimal numbers, such as the unit of distance names (but were split to the name of the column). Counties and administrative_units tables have been transformed from CSV to data frame and then, after connection with PostgreSQL, were transferred into the PostgreSQL Croatian Local Administration database. Excel tables are: county_population_count, county_population_table, county_brutto_salaries, county_netto_salaries, county_no_employment, county_registred_unemployment and county_gdp. Some of the tables were combined in Power Query; county_salary_table (county brutto and netto tables) and county_employment_table (county total number of employment and registered unemployment tables). The date table was made by DAX.

Data types have been changed mostly for columns that are normally numerical or decimal. Also, data categories were updated (e.g., county has been marked as County data category). In the administrative_unit table (fax column) there were some empty rows that were replaced by the "N/A" value.

There are a few calculated columns: net_migration_rate (number of immigrants subtracted by the number of emigrants) and rate_of_natural_increase (number of live births subtracted by death) from county_population_table.

There are 4 pages: County Summary, Population, Top 5 Migration, County Salaries, County Economy. Each page contains a slicer for counties. The first one presents summary data for counties, i.e., number of administrative units and, county area (km2), 5 cards containing the number of administrative units (counties and municipality). The second page is about population and has 2 line charts of natural increase and net migration rate. The third page contains 2 clustered bar charts (top 5 county emigration and immigration), a table with conditional formatting, and a line chart of the net migration rate. The fourth page is about county salaries and contains a line chart with shade area so the difference between salaries can be easily detected. Also, there is a line and stacked column chart that presents the difference between brutto and netto salary among the top 5 migration counties. Finally, there is the fifth page that contains two

line charts (county GDP and total number of employed) and again a line and stacked column chart.

It may seem that some visualizations present incorrect data. This goes especially for the county salaries and economy pages. For instance, on the County Salary page, one may notice that the brutto netto difference amounts to 200 % which seems like a wrong datum. The reason for that is the sum of all values. In order to perceive the right data, a slicer needs to be used, i.e., a county needs to be chosen.

Based on the data we have, it can be concluded that Croatia is a very centralized state and there are few trends. The biggest migration flow occurs in the capital city and its vicinity and Split as being the second biggest city. The trends can be seen around 2008 and 2011 due to Great Recession and 2013 when Croatia became a member of European Union, i.e., when labor market became more accessible.