Tural Novruzov 28632

Project Step 1

Introduction

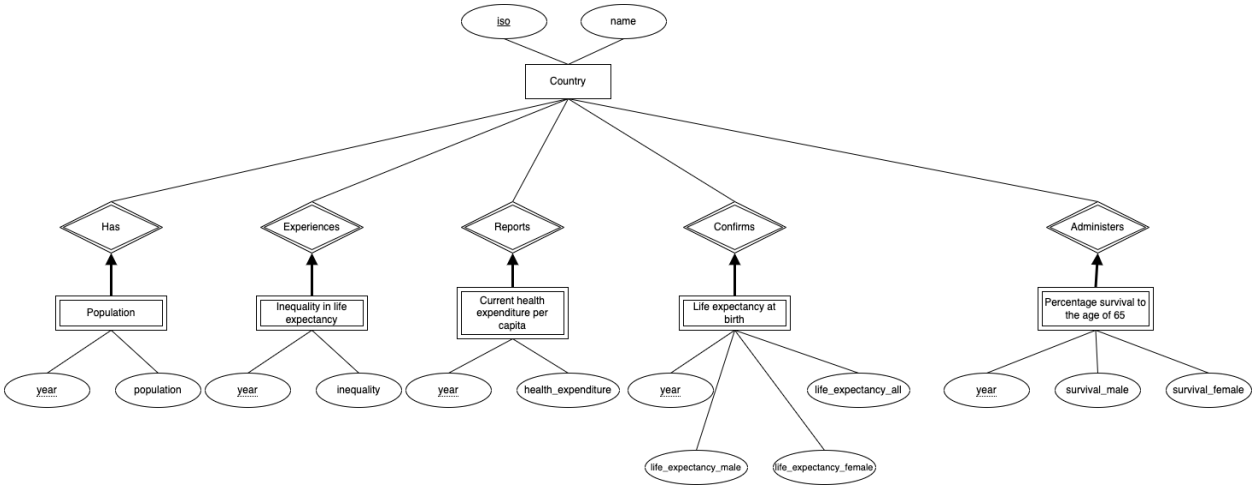
My project is titled "Life Expectancy and Inequality" where I will develop a database application to analyze inequality in life expectancy between countries and genders. A secondary objective of the application is to explore the relationship between life expectancy and population in different countries. The github repository for the project: https://github.com/turalnovruzov/cs306-project. **I will have every step in different branches for easier browsing and grading experience. Please refer to the branch "step1" to see my project's state when I finished Step 1.**

Data

In this project, I am using 6 CSV files. I have ensured that there are no duplicates in these files. To clean the data, I removed all rows that had empty values in the columns used across all files. Additionally, I deleted rows that did not contain a country code value, mainly representing continents.

- The "countries.csv" file stores the country codes and names of countries. I created this sheet by copying the country code and name columns from "population.csv" and then removed duplicates to have unique country entries.
- The "inequality-in-life-expectancy-vs-health-expenditure-per-capita.csv" file contains information on inequality in life expectancy percentage and health expenditure per capita of countries, primarily for the years 2010-2019. I eliminated rows without values for these columns.
- The "life-expectancy-of-women-vs-life-expectancy-of-women.csv" file provides data on the life expectancy of women and men in countries from 1950 to 2021.
- The "life-expectancy.csv" file contains information on the life expectancy of people in countries from 1950 to 2021.
- The "survival-to-age-65-of-cohort.csv" file includes data on the percentage of males and females who survived until the age of 65 in countries from 1960 to 2020.

ER Diagram



The ER diagram for the database application comprises a total of 6 entities and 5 relationship sets. Out of the six entities, five are weak entities. I have enforced a participation constraint on the weak entity side for each relationship set, and all relationships are considered weak entity relationships. The diagram includes 5 key constraints. All relationships are one-to-many relationships, with the "many" side being the Country entity. The weak entities have a weak primary key named "year".

I merged the "life-expectancy-of-women-vs-life-expectancy-of-women.csv" and "life-expectancy.csv" files into the "Life expectancy at birth" entity. This entity now has three attributes representing life expectancy for all people, males, and females. Conversely, I separated the data from the "inequality-in-life-expectancy-vs-health-expenditure-per-capita.csv" file into two distinct entities: "Inequality in life expectancy" and "Current health expenditure per capita". This decision was made based on the distinct information presented in the two columns of the CSV file.