Project 2: ETL Challenge

In this project, we are a team of Australian national policy makers looking into the OECD countries to determine if good Work LIfe Balance can improve a country's overall health and the economy. If yes, we aim to identify a few better performing countries from which Australia can adopt social policies to improve overall health and economy.

Team:

- Miley Kotamee
- Zachary Breen
- Mime Liu
- Ishaan Nigam

Extract: The two data sources used in this project are located in the Resources folder.

1. OECD countries' Better life index data
   https://stats.oecd.org/Index.aspx?DataSetCode=BLI
2. World happiness index data
   https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021

Transform:

- **Data cleaning:**
  - We screened the quality of data in the csv first, we noticed in the OECD csv there were rows of blanks. Thus, we filled in the value of "0" to avoid potential errors in the data manipulation process.
- **Joining:**
  - The OECD csv file lists data of only OECD countries (and three non OECD countries). Whereas the World happiness index data contains up to 148 countries' data. So after reading the csv with Pandas, we dropped the rows of the non-OECD countries in the OECD dataframe. Then, we inner-joined the two dataframes by the countries to make sure the two tables match each other. Eventually we had a combined table for 35 OECD countries with no null/blank values.
- **Summarisation:**
  - Based on our questions leading into the project and our hypothesis, we created an ERD diagram to assist us with mapping out subtables that will help with our data analysis. Please see the last page of this document for

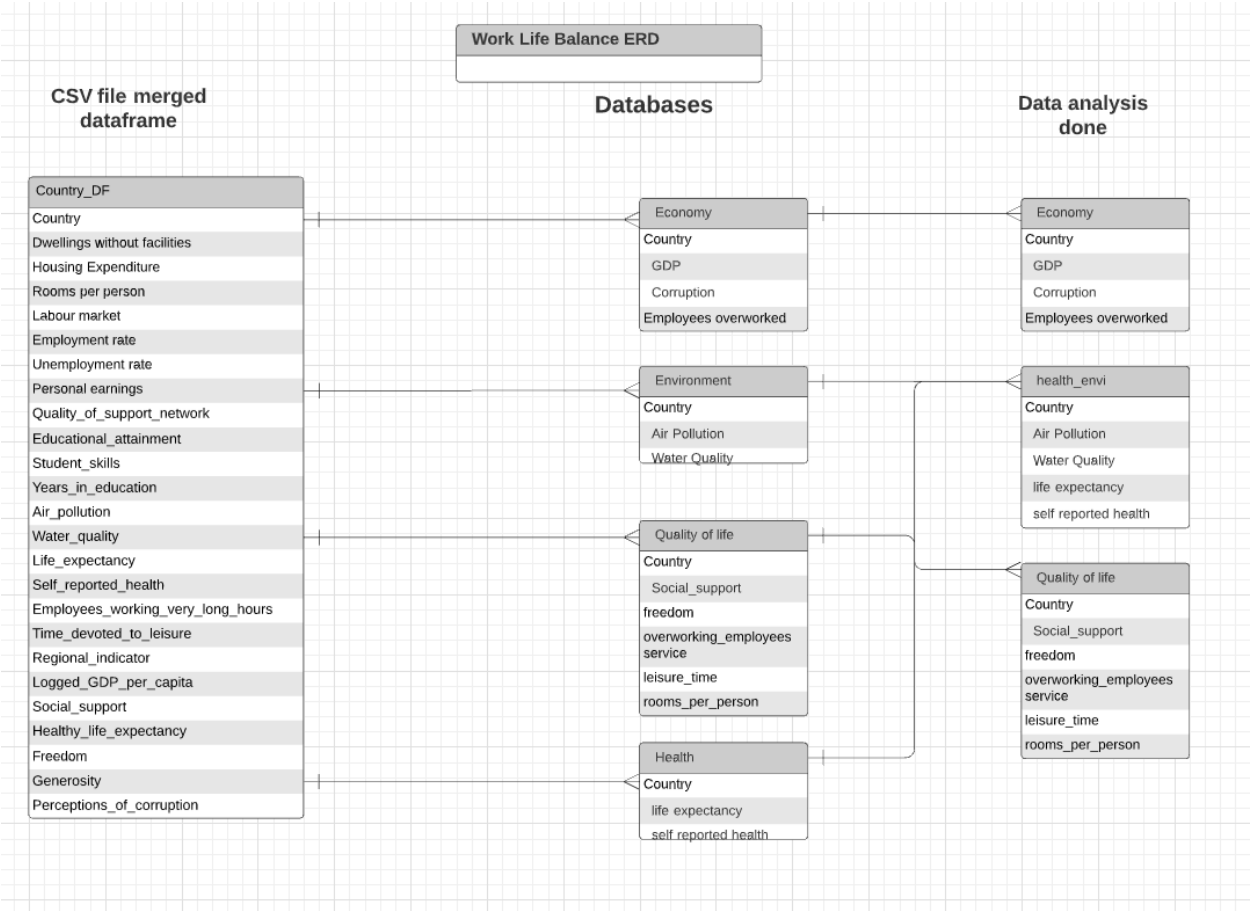
the attached ERD chart.

- **Filtering and selection:**
    - We assigned a sub-topic to each individual team member and completed our analysis. Please refer to our proposal to see our task distributions. For example, Mime's task was to find out correlations between work life balance and sense of happiness and the freedom of making life choices, so she created a table that mainly focuses on the quality of life data, including happiness index, freedom to make life choices,percentage of overworking employees, amount of leisure time and the average rooms per person in that country. Mime did this by selecting the needed columns and creating them into a new dataframe called "qol_df".

Load:

- After creating the data frame that serves our research purpose, we then each created the database named ETLE on PostgreSql, and empty tables to house the data we have transformed. This was achieved using SQLalchemy.
    - For example, Ishaan's focus of research was to find out the correlations between GDP, residents' perception of corruption and their work life balance. So he created a table named economy_df which contains all these three columns of data. He then created an economy table in his PostgreSql. After this, he pushed/loaded his economy_df's data into/with his sql economy table.
- In total, we have created four tables in this ETLE database, namely Employment, Economy, Qol (meaning quality of life) and Health. All these tables were joinable/related by "country".

Analysis: Please locate the Data Analysis document which displays findings.


Work Life Balance ERD
CSV file merged dataframe
Databases
Data analysis done
Country_DF
Country
Dwellings without facilities
Housing Expenditure
Rooms per person
Labour market
Employment rate
Unemployment rate
Personal earnings
Quality_of_support_network
Educational_attainment
Student_skills
Years_in_education
Air_pollution
Water_quality
Life_expectancy
Self_reported_health
Employees_working_very_long_hours
Time_devoted_to_leisure
Regional_indicator
Logged_GDP_per_capita
Social_support
Healthy_life_expectancy
Freedom
Generosity
Perceptions_of_corruption
Economy
Country
GDP
Corruption
Employees overworked
Environment
Country
Air Pollution
Water Quality
Quality of life
Country
Social_support
freedom
overworking_employees service
leisure_time
rooms_per_person
Health
Country
life expectancy
self reported health
Economy
Country
GDP
Corruption
Employees overworked
health_envi
Country
Air Pollution
Water Quality
life expectancy
self reported health
Quality of life
Country
Social_support
freedom
overworking_employees service
leisure_time
rooms_per_person