# PROJECT REPORT

# ETL Report - Employment & GDP

We decided to create a useful data pipeline for analyzing international socioeconomic trends. Pulled the standardized data from the OECD, and prepared it for the future analysis

## Extracting Data:

Pulled 2 CSV files from the OECD website:

A. [Employment rate](https://data.oecd.org/emp/employment-rate.htm): Employment rates are defined as a measure of the extent to which available labour resources (people available to work) are being used. They are calculated as the ratio of the employed to the working age population.

B. [Gross domestic product (GDP)](https://data.oecd.org/gdp/gross-domestic-product-gdp.htm): Gross domestic product (GDP) is the standard measure of the value added created through the production of goods and services in a country during a certain period. [measured in US dollars]

## Transforming Data:

#### 1. Loading the csv files to dataframes with pandas

#### 2. Removing the unnecessary data

A) Employment rates:

* Keeping only the total unemployment, removed the metrics segmented by gender
* Isolating the data on single measurement unit (pct of labor force) = removing the employment measured in thousands of persons
* Annualizing all the data - Some of the countries were using monthly or quarterly measurement cycle, some just the yearly one. We needed just the yearly cycle for pairing with the GDP data (yearly only): We annualized the data by calculating the mean yearly value for any country-time period.
* Dropping the redundant columns, keeping only the country, year, percentage of employment.

B) GDP data:

* Removing the redundant columns, keeping only the country, year, GDP (MLN USD).

#### 3. Combining both of the datasets into the single dataframe (pd.merge on country & time)

* Keeping only the years since 2000

## Loading Data:

1. Creating the local db & schema for the tables in PostgreSQL
2. Connecting to the created db using sqlalchemy
3. Loading our cleaned up datatadframes into the prepared db tables via pandas