# **Project 2 ETL - (Extract Transfrom & Load)**

Sarah Casauria [sarah.casauria@hotmail.com](mailto:sarah.casauria@hotmail.com)   
Dhiren Patel [dhiren.p@outlook.com](mailto:dhiren.p@outlook.com)   
Rosalina Indriani [osha\_bu@yahoo.com.au](mailto:osha_bu@yahoo.com.au)   
Michael Dunne [michaeldunneMD17@gmail.com](mailto:michaeldunneMD17@gmail.com)   
   
**Project Proposal**

For this project, our group decided to use two global datasets to utilise our ETL skills. We have used 2 datasets from the public platform Kaggle. The sources for our dataset are as follows:

1. UNICEF’s Out of School Rates Global Data, which aggregated net attendance rate of primary, lower secondary, and upper secondary school students globally based on household survey data. This dataset can be found at: <https://www.kaggle.com/datasets/komalkhetlani/out-of-school-rates-global-data>
2. The Cato Institute’s Human Freedom Index, which measures the overall global freedom by country. This dataset can be found at: <https://www.kaggle.com/datasets/gsutters/the-human-freedom-index>

For this project, our group decided to employ a relational database hosted with PostgreSQL due to its open source, no cost nature.  
  
**Project Report  
  
Github**  
Attached below is the link to our github created to exhibit our findings.   
<https://github.com/sarahcasauria/Project_2_ETL>   
  
Files uploaded include:   
- ETL notebook.ipynb (jupyter)   
- README.md   
- joins.sql   
- schema.sql   
  
Along with the two resource files (csv) we analysed for this project:   
- out\_of\_school\_primary.csv   
- hfi\_cc\_2021.csv  **Extract**

Data was extracted from two .csv files which were downloaded from the data sources cited above. Each .csv file was read into a Pandas dataframe with ISO-8859-1 encoding. The two dataframes were labelled as follows:

* Out of School Rates Global Dataset = `out\_of\_school`
* Human Freedom Index Dataset = `freedom\_index`

### **Transform**

* Out of School dataset
  + The dataframe `out\_of\_school` was reduced to store three columns that we deemed useful for our analysis, which included:
    - “ISO3”
    - “Countries and areas”
    - “Total” (i.e. total percentage of primary school-age students who are out of school)
  + Columns were then renamed to `country\_code`, `country\_name`, and `out\_of\_school` respectively.
* Freedom Index Dataset
* The dataframe `freedom\_index` was reduced to store three columns that we deemed useful for our analysis, which included:
  + `ISO`
  + `countries`
  + `hf\_score` (Human Freedom score)
* Columns were then renamed to `country\_code`, `country\_name` and `hf\_score` respectively.

All null data was then dropped from each dataframe, reducing the `out\_of\_school` dataset down to 112 rows, and the `freedom\_index` to 1867 rows.

Since unique countries were required to load into the relational database, any duplicate data was dropped from both dataframes based on the `country\_code` field. This transformation did not affect the `out\_of\_school` dataframe but reduced the `freedom\_index` dataframe to 165 rows of data.

For each dataframe, the ‘country\_code’ column was then set as the index.

### **Load**

Using PostgreSQL as the host, a database titled `project\_2\_etl` was created with the intention of storing our transformed data. Two tables were created, `out\_of\_school` and `hf\_score` to store the data for the two dataframes. Please refer to the schema(link) provided.

Once the tables were created in Postgres, SQLAlchemy was employed to make a connection to the database from our Jupyter notebook. Data from the `out\_of\_school` dataframe was then loaded into the `out\_of\_school` table and data from the `freedom\_index` dataframe was loaded into the `hf\_index` table.  
  
The reasoning for this topic was to further investigate the relationship between the human freedom index (HFI) & primary education (aged 6-12 levels). The Cato Institute define this as ‘the state of human freedom in the world based on broad measures encompassing personal, civil, and economic freedoms’. This index is divided into two sub-groups:  
1) Economic freedom, 2) Personal freedom.   
It is measured on a scale of 1 to 10 with the average score across the 162 juridictions (countries) analysed being 7.12 / 10.   
In the most index released (2019), Switzerland recorded the highest cumulative index scores  
(Personal: 9.56 , Economic: 8.48), with Venezuela (Personal 4.9, Economic: 2.83), recording the lowest.   
  
More information relating to this index can be found on the link below:   
<https://www.cato.org/human-freedom-index/2021>

## **Summary of findings**

* This method of ETL is most efficient when multiple datasets share similar information. Our datasets share both country name and country ISO3 standard codes. Therefore, we were able to join the data from both tables based on their country codes.
* Since we removed any null data from each dataframe before loading into our PostgreSQL database, only 100 countries were represented when we joined the data, as each country had to have data for both the HFI score and the percentage of students out of school.