ETL Project

By Anson Ia and Chan Ho Anh

**Extraction**

We used datasets an REST API - <https://restcountries.eu/#api-endpoints-all> and GDP (Gross domestic product) data from Kaggle. All of our data was based on country data received from the API which came in the form of JSON and the GDP information from Kaggle that came in a CSV file, where we looked into the GDP based on each country’s population by country code.

Here’s the list of columns we choose to display:

* Country Name
* Country Code
* Capital
* Region
* Sub Region
* Population
* Language
* Flag (link to image of each country’s flag)
* Native name
* GDP

**Transformation**

Our first steps in cleaning up the datasets involved figuring out which variables were not relevant. There were over 24 columns and we narrowed it down to 7-8 columns (see above) from the REST API. For the GDP data from Kaggle, we used all three columns that were provided. We then set up dataframes for each set of data. We then used the country code to join the two tables.

**Load**

The last step was to transfer our final output into a database. We created a database and respective tables to match the columns from the final Panda’s Data Frame using SQL and then connected to the database using SQLAlchemy and loaded the result. Here we were able to perform multiple queries to suit a desired criterion for GDP, Country, Country code and Population