ETL Project

By Matthieu Dubois and Serge Guilao

**Extraction**

We used 2 datasets from the public platform Kaggle and Data World. All of our data was based on 1998 FIFA world cup. The sources for our dataset are as follows

* FIFA\_Ranking from Kaggle.
* FIFA\_Countries\_Audience from Data World.

**Transformation**

Our first steps in cleaning up the datasets involved figuring out which variables were not relevant. For FIFA Countries Audience (Figure1) , we dropped the columns population\_share and tv\_audience\_share and renamed it FIFA\_countries\_df(Figure2).

Figure1

| **country** | **confederation** | **population\_share** | **tv\_audience\_share** | **gdp\_weighted\_share** |
| --- | --- | --- | --- | --- |
| **0** | Afghanistan | AFC | 0.4 | 0.3 | 0.0 |
| **1** | Albania | UEFA | 0.0 | 0.1 | 0.1 |
| **2** | Algeria | CAF | 0.5 | 0.4 | 0.3 |
| **3** | Andorra | AD | 0.0 | 0.0 | 0.0 |
| **4** | Angola | CAF | 0.3 | 0.1 | 0.1 |

Figure2

| **country** | **gdp\_weighted\_share** |
| --- | --- |
| **0** | Afghanistan | 0.0 |
| **1** | Albania | 0.1 |
| **2** | Algeria | 0.3 |
| **3** | Andorra | 0.0 |
| **4** | Angola | 0.1 |

For the FIFA\_Ranking dataset, we clean it up and rename it to FIFA\_Ranking\_df. The two CSV files were combined into one universal table called merged( Figure3).

Fig3

| **country** | **gdp\_weighted\_share** | **rank** |
| --- | --- | --- |
| **0** | Afghanistan | 0.0 | 193 |
| **1** | Albania | 0.1 | 120 |
| **2** | Algeria | 0.3 | 60 |
| **3** | Andorra | 0.0 | 188 |
| **4** | Angola | 0.1 |  |

**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 PostgresSQL 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.