ETL Project

By Sherin Thomas, Therese Reed and Anjali Khare.

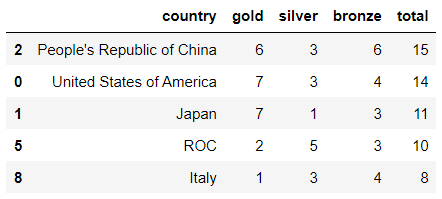
**Extraction**

We used datasets from the public platform Kaggle and following csv files. All of our data is based on Olympics 2016 and 2020. These were the most recent ones we could find. The sources for our dataset are as follows

* 2016 athlete’s data
* 2020 athlete’s data
* 2016 country data
* 2020 medals data

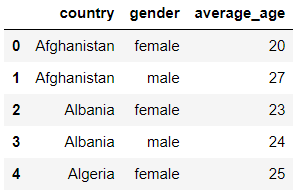
**Transformation**

* Our first step in cleaning up the datasets involved figuring out which variables were not relevant. We Extracted medals data of 2020, dropped unwanted columns and renamed required columns.



**Figure 1:** Dropped unnecessary columns

* We renamed country code column in countries data frame to merge the table with athlete’s data of 2016 in order to get country of each athlete.
* We extracted country, gender and dob column from 2016 data and used it to find average age of men and women for each country.
* We used the dob column to find the age with pandas datetime object. There was a limitation with determined index function when the dob is less than 1969. We created a user defined function to adjust the year when the dob is less than 1969.
* We used group by function with country and gender and calculated the mean to find the average age of players for each country per gender.

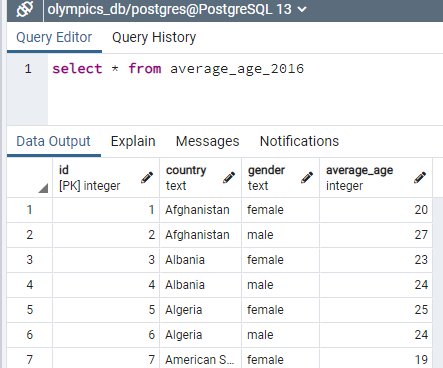


**Figure 2:** Group by function with country and gender

* We used group by country and calculated the sum of gold, silver and bronze medals to find the medals for each country. We used these columns to find the total medal for each country.

**Load**

The last step was to transfer our final output into a Database. Using the final Panda’s Data Frame and SQLAlchemy, we created a PostgreSQL database engine and uploaded data into tables.



**Figure 6: Sample Query**

**Summary**