

**Olympic Games History**

**ETL Project**

**Cynthia juarez**

**jorge cavazos**

**lynn leifker**

**kellye rennell**

Data Visualization – July 2019 Cohort | October 17, 2019

# **Extract**

Our team utilized the following resources for data extraction:

1. Kaggle - 120 Years of Olympic history: athletes and results

* [www.kaggle.com](http://www.kaggle.com)
* Bio data on athletes and medal results from Athens 1896 to Rio 2016

1. Organization for Economic Co-operation and Development by Angus Maddison (The World Economy- A Millennial Perspective)

* <http://theunbrokenwindow.com/Development/MADDISON%20The%20World%20Economy--A%20Millennial.pdf>

Data sets were provided in csv and pdf formats

|  |  |
| --- | --- |
| PDF | CSV |
| * Africa GDP * Africa population * Asia GDP * Asia population * Caribbean GDP & population | * Athlete events * Country GDP * Country Population * NOC Regions |
| * Europe GDP * Europe population * Latin America GDP * Latin America population * USSR GDP & population |  |

# **Transform**

All of the datasets were cleaned up and imported into Python using Pandas.

Athlete\_events csv contained 15 total columns of data and the National Olympic Committee regions csv contained 3 columns. These two csv’s were joined to create the Olympic data cleaned csv which contains the columns, grouped by: Sex, Team(country), Year, Medal, and ID, needed for this project.

* all nan values were dropped within the medals column which eliminated all of the rows that did not receive medals, leaving all of the athletes that received gold, silver, bronze medals left for accurate analysis.
* Further data cleaning measures were used by dropping all of the non-country team names, like New York Athletic Club, and changing the country names that had a value associated with that country name so as to not cause any skewed results. (example: Austria -1)

Country\_GDP.csv and Country\_population.csv data sets were cleaned up by first converting the pdf’s into csv files. Both csv data sets were imported into Python using Pandas and grouped by country, historical year, and GDP.

* NaN values were left in these data sets as they serve as placeholders in a few of the columns
* Columns were renamed and sorted by Country.
* Group

Results from data transformation

* Clean csv’s for both GDP and Population data
* GDP\_and\_Population – Jupyter Notebook csv cleaning process
* Dataset Transformation – Jupyter Notebook for athlete csv cleaning process
* Olympic Dataset Transformation – additional clean up on teams

# Load

For our last step in the ETL process, we created a database using PostGres / PgAdmin

Data frames were created from the csv files, so those dataframes were loaded into the database by using Pandas to\_sql method.

# Challenges

* Merging all of the datasets from our resources via country name
* Some of the countries were missing values for one or more years from data set
* Some of the Teams are not countries, so grouping by teams created a challenge
* Total number of team names is more than the total number of countries in the country GDP dataset

Total number of teams : **150**

Number of total teams represented: **136**

* Population number and GDP

**Examples for reference**

1. Country name changes such as Bohemia -> Czechoslovakia which split to Czech Republic and Slovakia.
2. If data analysis were to take place, the decision on how to align the teams with GDP and population would have to be considered.
3. Total number of team names is more than the total number of countries in the country GDP dataset :

Total number of teams : **498**

Number of total teams represented: **176**