

**Olympic Games History**

**ETL Project**

**Cynthia juarez**

**jorge cavacos**

**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.

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 need to create a database using PostGres / PgAdmin .

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

\*\*\*\*\*\*Jupyter Notebook with pandas data frames loaded into sql goes here\*\*\*\*\*\*\*\*

# Challenges

* Merging all of the datasets from our resources via country name. Many country names have changed over time, so there is not a one-to-one match between names in datasets.
* Early Olympic teams could represent local organizations rather than countries, so grouping by teams created a data-cleaning challenge. Of 448 teams in the original dataset, only 150 were countries.
* Country names in the GDP and population datasets are modern, whereas Olympic team names are historical. Determining which modern countries to drop from the larger GDP/population datasets would require additional research into historical names.
* Some of the countries were missing GDP or population values for one or more years from data set
* NaN values for GDP and population in some columns would need to be filled if data analysis required mean GDP or population across time or countries. Decisions would need to be made about how to fill them.
* Years did not align between the Olympic dataset and the GDP and population datasets. GDP and population had five values for years (between 1913 and 1998), whereas Olympic dataset had 33 different year values. A decision would need to be made about whether to bin the Olympic year values into five groups or to match year values between datasets in a different way.

**Examples for reference**

1. Country name differences between Olympic and other two datasets, like “Great Britain” vs. “United Kingdom.”
2. Country name overlap like Great Britain, which has competed as a team since 1896 but whose constituent states also competed separately at times. For example, in 1908 Great Britain won 199 medals, England won 11 medals, Ireland won 16 medals, Scotland won 11 medals, and Wales won 11 medals. The dataset does not make it clear if these overlap (meaning they’re already accounted for in the 199 medals) or if they can be combined into 248 medals for Great Britain.
3. Country name changes such as Bohemia, which became part of Czechoslovakia in 1918 and then part of the Czech Republic in 1993. All three country names appear in the Olympic dataset.
4. Country boundary changes such as the USSR, which formed in 1922, expanded in 1945, and then broke up into 15 separate countries from 1991 to 1993. Determining which athletes came from which region during those years would be a challenge requiring additional research. If we decided to keep the USSR as one country, we would need to consolidate GDP and population data from its modern constituent countries into one group for the Soviet period.
5. Occasionally, the same country name refers to different countries. The Macedonia that won medals from 1900 to 1952 is not the same country as the Macedonia that won in the 1992 games and still exists. Decisions would need to be made about which modern country to associate with the historical Macedonia (Serbia or Greece?)