**ETL Project**

**Team:**

Beth Myers & Kurt Dietrich

**Idea Generation:**

As we are in Women’s History Month (March), we wanted to select use of data of iconic women. We then further narrowed our view to women athletes who have competed in the Olympics – as we are also expecting for the Olympics to take place in Tokyo, Japan later this year.

**Extract / Data Sources:**

We found three separate data sources to extract, which were:

* Athlete\_events.csv from Kaggle.com
* NOC\_regions.csv from Kaggle.com
* Host\_cities from Wikipedia (<https://en.wikipedia.org/wiki/List_of_Olympic_Games_host_cities>)

**Transformation:**

The two csv files were pulled into Pandas, where the NOC\_regions.csv file only required renaming of column titles to align with the design requirements needed for SQL (via PGAdmin). The athlete\_events.csv file required analysis to breakdown data into smaller tables that can be queried in various ways. The athlete\_events.csv was used to create 6 tables / dataframes:

* Teams
* Athlete
* Athlete\_data
* Sport
* Medals
* Games

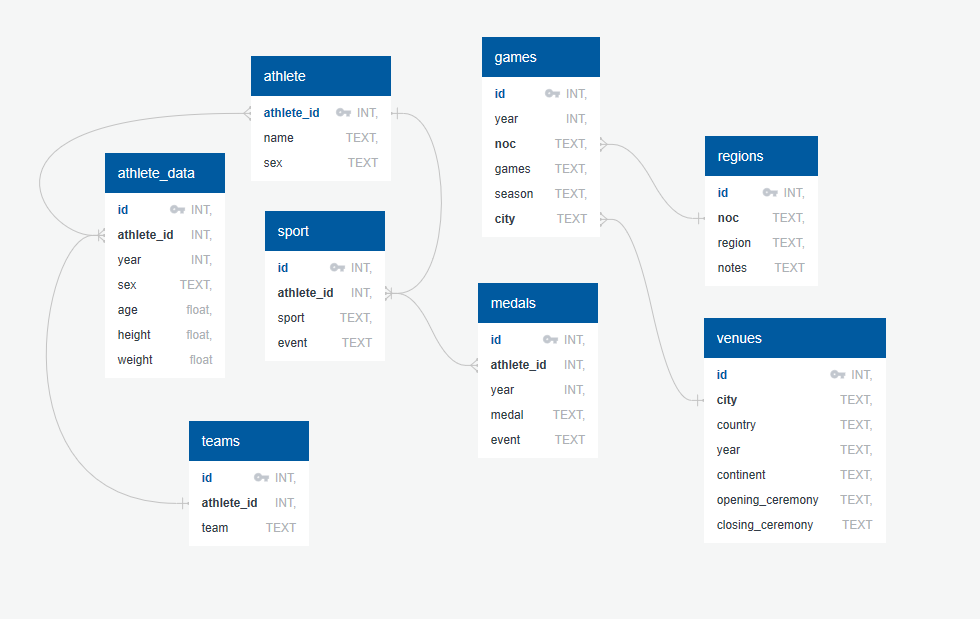
In each of the 6 newly created tables, a primary key was set or added, and all columns required renaming. Additionally, there was the need to remove duplicating information, which was performed through Pandas.

For the table pulled via web-scraping (Host\_cities), there were 3 columns which did not have useful information for database needs and were dropped from use in the tables / dataframes through Pandas. Like the csv files, all column titles required renaming.

**Load:**

Following necessary data clean up, all tables were loaded using Pandas/SQLalchemy.

Below is the table schema for our relational database:



**Conclusion:**

There may be the need to add foreign keys in the future to query and/or data mine information better. We did create a couple queries to test our data, which was verified through using Excel.

Query 1:

Top 12 Women with the most Gold medals



Query 2:

Top Sports with the most Gold medals, in aggregate

