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 from Kaggle were downloaded and saved in our resources folder. These were then pulled into Pandas. The **NOC_regions.csv (National Olympic Committee)** 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. Below is how the data appeared prior to clean-up:

```
csv_file1 = "Resources/athlete_events.csv"
full_athlete = pd.read_csv(csv_file1)
full_athlete.head()
```

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

This athlete_events.csv file 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 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, see below:

```
venues_df = tables[2]
venues_df.head(10)
```

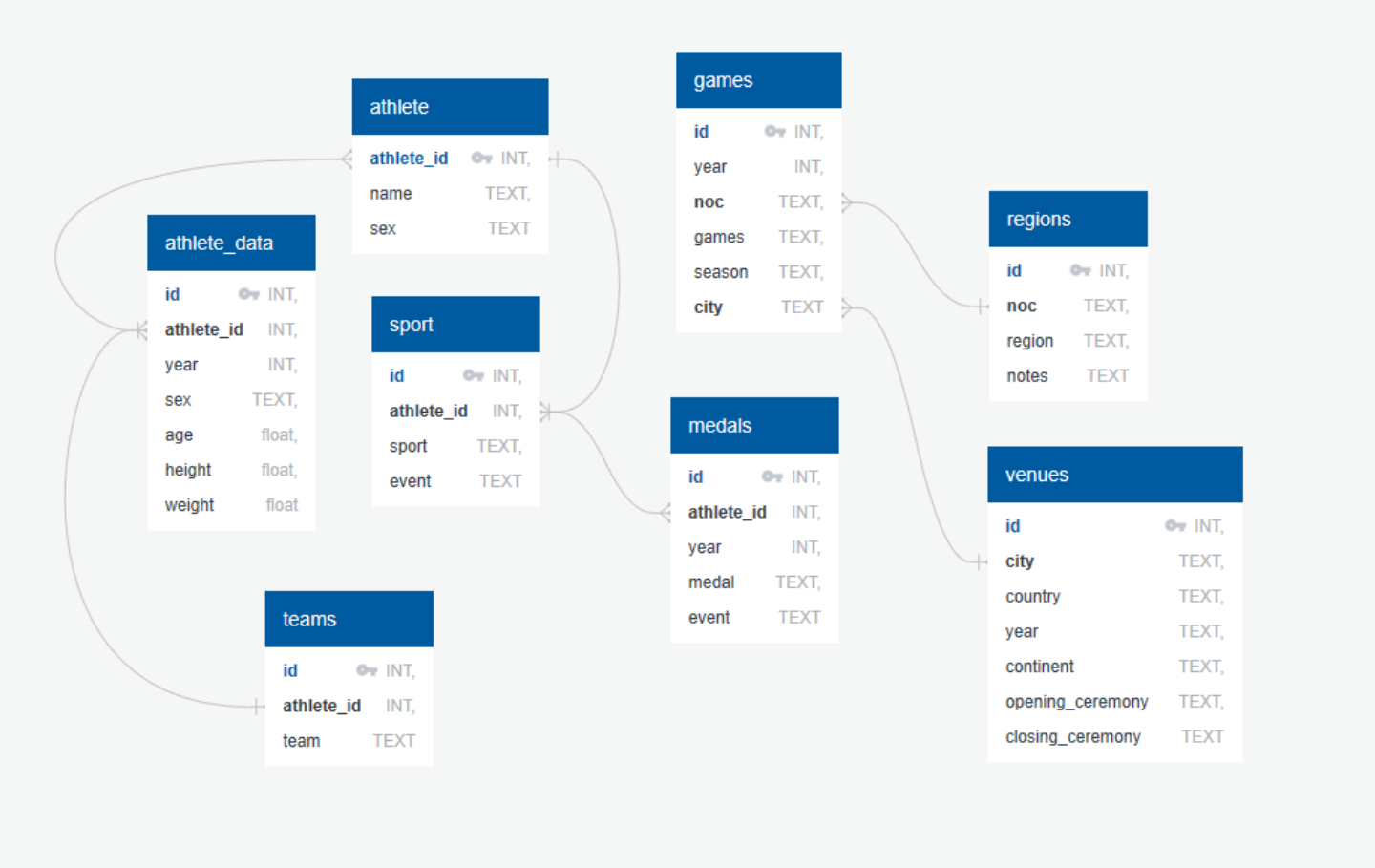
	City	City.1	Country	Year	Continent	Summer	Winter	Opening ceremony	Closing ceremony	Ref
0	NaN	Athens	Greece	1896	Europe	NaN	NaN	6 April 1896	15 April 1896	NaN
1	NaN	Paris	France	1900	Europe	NaN	NaN	14 May 1900	28 October 1900	NaN
2	NaN	St. Louis[a]	United States	1904	North America	NaN	NaN	1 July 1904	23 November 1904	NaN
3	NaN	London[b]	United Kingdom	1908	Europe	NaN	NaN	27 April 1908	31 October 1908	NaN
4	NaN	Stockholm	Sweden	1912	Europe	NaN	NaN	6 July 1912	22 July 1912	NaN
5	NaN	Berlin	Germany	1916	Europe	VI	NaN	Cancelled due to WWI	Cancelled due to WWI	[11]

The Summer, Winter and Ref columns housed unreliable information and were dropped from the dataframe. Like the csv files, all column titles required renaming. No duplicate records existed.

Load:

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

Below is the table schema for our relational database:



Conclusion:

We see that the benefits of this process as a great start in data analytics, where one can run various queries.

Some challenges that arose, where tables and data was created quickly and required re-work to assure queries pulled results as expected, which includes the need to assure duplicates are addressed appropriately.

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

count bigint	name text
9	Larysa Semenivna Latynina (Diriy-)
8	Jennifer Elisabeth "Jenny" Thompson (-Cumpelik)
8	Birgit Fischer-Schmidt
7	Vra slavsk (-Odloilov)
6	Isabelle Regina Werth
6	Marit Bjrgen
6	Maria Valentina Vezzali
6	Lidiya Pavlovna Skoblikova (-Polozkova)
6	Allyson Michelle Felix
6	Amy Deloris Van Dyken (-Rouen)
6	Lyubov Ivanovna Yegorova
6	Kristin Otto

Query 2:
Top Sports with the most Gold medals, in aggregate

count bigint	event text
414	Football Men's Football
407	Ice Hockey Men's Ice Hockey
360	Hockey Men's Hockey
287	Water Polo Men's Water Polo
249	Gymnastics Men's Team All-Around
244	Rowing Men's Coxed Eights
234	Basketball Men's Basketball
194	Handball Men's Handball
166	Volleyball Men's Volleyball
158	Hockey Women's Hockey