**Olympic Games: Extract Transform Load**

**Final Report**

**Team members:**

Ariel Dunn

Irina Fedorova

Portia Anyamba

**Motivation**

It is the season for Olympic Games. Different athletes from different countries compete for the event's prestigious medals. Besides the hype about the event, host cities also use the games to showcase their food and culture to attract tourists during the games and beyond! Our team was curious to discover what data is available on the subject besides athletes, countries, cost, sports and medals, and how many times each country hosted Olympics.

**Project description/outline**

This project will consider datasets on where the Olympic Games held, athletes, participating countries, costs and medals won.

The project’s outline is as follows:

1. **Finding Data**

This is a pre-ETL stage that involves data research and screening, as well as developing a project proposal. Our three data files were obtained from the following two sources of data:

The athlete\_events.csv from[*https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results*](https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results)

The Cost\_per\_ event\_and\_cost\_ per\_athlete\_in\_the\_Olympics.csv <https://data.world/sports/olympics/workspace/file?filename=+Cost+per+event+and+cost+per+athlete+in+the+Olympics+.csv>

Costs\_ of\_ the\_ Olympic\_ Games.csv <https://data.world/sports/olympics/workspace/file?filename=outturn+sports-related+costs+of+the+Olympic+Games+.csv>

Based on the evaluation of selected files, our **Hypothesis** is as follows: Olympic Games host countries, particularly the USA, have won more medals than when they compete abroad.

2. **Extracting Files into DataFrames**

\* Extracted CSV files were inspected to identify possible flaws including: different column titles, labelling format, and dates.

\* Files were saved under the "Resources" folder.

\* On Jupyter, dependencies ( pandas, numpy and sqlalchemy )were imported dependencies.

\* A pathway for the csv files was created and the csv files were read onto DataFrames.

3. **Transforming DataFrames**

* Athletes DataFrame

\* Created a Medals DataFrame and split DataFrame to three using groupby function by medal categories: Gold, Silver, and Bronze.

\* The three medal DataFrames were concatenated into a single DataFrame.

\* Additional columns for medals count were inserted. To convert data to integers NaN values were replaced by zero(0) and columns were then changed to (int) data type.

\* Medals and athletes tables were merged. Columns were filtered, renamed and formatted. Also columns with empty rows and duplicate insertions were dropped.

\* The language for several cities was corrected to their English versions.

* **Costs DataFrames**

\*The two cost DataFrames’ columns were filtered and renamed. Duplicates were dropped.

\* Standardized NaN values

\* Merged the two cost DataFrames based on year and city.

4. **Data Loading**

\* First we used an ERD tool to depict relationships between Tables.

\* Initialized engine to connect to postgress (pgAdmin) and checked for tables using engine.table\_names().

Through pgAdmin, we created tables, loaded data and joined them to get the relevant columns to our question. This database was chosen because the Olympics is a very structured event, and we believed that a relational database would be the best way to load the data. The data was structured in a way, that there were many common columns that we could use to merge the data together.

**Query Results**

Results shows that the USA hosted Olympic Games five times since 1960 as shown below:

