**ETL Project: Health Spend and Olympic Performance for 10 Countries**

By Brittani Terry, Katia Gonzalez, Aleobe Irotumhe

**Extraction**

We used 2 datasets from the public platform Kaggle and Data World. We wanted to find health spend for a set of countries leading up to the 2021 Olympics, as well as the Olympics performance/medals for the same set of countries. These datasets contained the most information and were also the most recent ones we could find. Data sources are as follows:

* [Tokyo Olympics Dataset](https://data.world/govind6991/1-bi-project/workspace/file?filename=Tokyo+2021+dataset.csv) from Data World.
* [Health Spend Dataset](https://data.oecd.org/healthres/health-spending.htm) from Kaggle.

**Transformation**

To transform, we first analyzed the datasets to conclude which variables were not relevant. For the Tokyo Olympics Dataset we first eliminated columns that were not relevant, then narrowed down the countries to just 10. Finally, we created an additional abbreviation column for the countries, ensuring a similar name format between the two datasets for an easier merge.

Table

Description automatically generated

For the Health Spend Dataset, we filtered the data to only show 2018 data and narrowed down the countries to just 10 to align with the first dataset. We also narrowed down the type of spend to only show the total for each country.

Table

Description automatically generated

**Load**

After all the data was cleaned and sorted, we uploaded the csv files to SQL. In SQL the tables were created so that the data could be uploaded correctly. After both data sets were in SQL, we did an inner join to merge the data sets based on the location while dropping the other unnecessary columns.

Table

Description automatically generated