ETL report

Extract/Data Utilize

Data sets (csv files) was received from Kaggle

We used three different Olympic games from years 2008, 2012 and 2016

Transformation

All the data was transformed using pandas.

We had to troubleshoot encoding to read the 2008 csv file.

We dropped columns to include only the columns of interest. We also dropped values of NA in the 2012 and 2016 datasets.

We then filtered out the datasets for the year we were interested in

The data was formatted to the data types required for example; we change year type from float to integer to correct for a formatting issue where the year column had a decimal point.

We renamed the columns to be uniform across the datasets

We need to separate out first and last names from the Name column. We used this using a lambda function to split based on delimiters. In one dataset, the names were separated by comma and the other was separated by space. Additionally, the names had to be capitalized and had to be indexed correctly because of inconsistencies between datasets.

For the 2012 and 2016 dataset, we had to filter out winter Olympics to only keep the Summer Olympics data. For this, we had to include an additional column for season, and we filtered for Summer in this column. Once we had the desired data, we dropped the season column.

We had to do a similar correction for another column called event to ensure that the entries matched across the three different datasets.

For the final data frame had columns – First Name, Last Name, Gender, Country, Sport, Event, Medal.

Load

All of the data was loaded into postgres SQL database. The database had three main tables which were olympic\_08, olympic\_12, olympic\_16

Outlier

The 2008 data had 201 events whereas 2012 and 2016 had 302 and 306 respectively.