## *ETL PROJECT REPORT*

## COUNTRY GDP VERSUS COUNTRY OLYMPIC MEDALS

## Team Members: Stephanie Salvona and Samantha van Wyngaarden

## Summary

## As neither of us particularly like sport we thought the Olympic information would be an interesting topic that would have much data over many years, that we could link up with Gross Domestic Product (GDP) data.

## The final tables or collections that will be used in the production database to investigate if Olymipic Medals, particularly Gold Medals can be correlated to countries with high GDP in contrast to with countries with low GDP would rarely win a medal or have no medals at all. The assumption is medal count is related to the sports budget a country has.

## Some limitations were quickly realised, such as many countries in both datasets are either not counties at all or not recognised countries and the Olympics stared in 1896 with 15 countries and GDP records start from 1960 onwards.

**E**xtract

Two .csv sources were extracted from <https://data.world/>

1. <https://data.world/nilanshuramteke/gdp-by-country/workspace/file?filename=GDP+By+Country.csv>
2. <https://data.world/johayes13/summer-winter-olympic-games>

and cleaned using Jupyter Notebook and Pandas

**T**ransform

Cleaning was required to remove unwanted data, rename columns to recognisable or better understandable headings and to assist with joining for both datasets.

For example:

* A DataFrame was created with 11 columns from Olympics .csv which initially had 33 columns
* Two columns were renamed appropriately
* The code entry for GER (Germany) was replaced with DEU (Deutschland) to match the data for the GDP .csv
* Many codes that had no relevance were not actively sought out, as they would be ‘dropped’ in future joins where the Primary Key and Foreign Keys did not match
* GPD data also had columns removed and renamed, however, required further steps being a much bigger and messier dataset. This included pd.melt and a further clean by dropping all NaNs

**L**oad:

* Using <https://app.quickdatabasediagrams.com/#/d/GPpj1C> for the Entity Relationship Diagram the two cleaned DataFrames were related using the Country\_Code of the GDP data as Primary Key, meaning this table was first in order to be the Foreign key for the Olympics Data
* The GPD data was put first as
* After making tables in pgAdmin the data was imported into SQL and tables checked by running SELECT \*