## *ETL PROJECT REPORT*

## COUNTRY GDP VERSUS COUNTRY OLYMPIC MEDALS

## Team Members: Stephanie Salvona and Samantha van Wyngaarden

## Introduction

## 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 Olympic 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 first international Olympic games were held in 1896 with 15 countries and GDP records from the chosen database start from 1960.

**E**xtract

Two .csv sources were extracted from <https://data.world/> and [https://www.kaggle.com](https://www.kaggle.coms)

1. <https://data.world/nilanshuramteke/gdp-by-country/workspace/file?filename=GDP+By+Country.csv>
2. <https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results>

and cleaned using Jupyter Notebook and Pandas

**T**ransform

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

For example:

* DataFrames were created with reduced number of columns
* 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
* Finally a lambda and merge to extract medal colour from single column and create a new table

**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 first so the Foreign key would be present for the Olympics Data
* After making tables in pgAdmin the data was imported into SQL and tables checked by running SELECT \* and creating views

## Summary

A single table is presented in SQL specifically for 2016 and the gold, silver and bronze medal count. Given more time we would further investigate no medals and low GDP and compare both.