Etl Project

By Nikunj Patel

# Project Summary

The world of sports betting requires companies to have current and historical data to create new markets for customers. The ability to access all this information in one source can be challenging for a data analyst, the aim was to create a single table of information that can used to conduct the following analysis

Can a prediction of a countries medal count be made using the following factos.

1. GDP per capita
2. Size of Olympic team
3. Size of population
4. Home advantage

Top performing countries with the most medals won?

Are certain countries better at certain sports?

The following tables outlines the process that was undertaken during the data sourcing and collection process.

|  |  |  |  |
| --- | --- | --- | --- |
| Steps | Data Needs | Observations | Actions |
|  | Significant amount of historical data on the Olympics. | After searching on multiple web pages was able to find a few csv and excel files that covered large segments of history. | Non |
|  | Clear similar identifier. | Majority of the data sets have differing identifiers as a result main data set had a ID that was used in the final dataset. | Noted down the same columns for each data set. |
|  | Explored the data sets for missing values. | Found that significant data was missing in different parts of each data set. Example age weight height have signiciant missing values. | Made the decision to change the Nana values in the medals column to DNW “did not win’ as this would allow for further analysis if required |
|  | GDP data | Due to the historical nature of this task. A dataset was found for GDP data, but significant amount of information was missing for many countries. | Due to the importance of this data set it was hard not to include it and a note was made of the quality of the data set. |
|  | Population’s data | Like the GDP data set significant missing sections of data were noted down for investigations. | Conclusion was made that the data set was important, and a note was made. |

# Extraction

The following data sets were used form the public platform Kaggle, liste in order.

1. <https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results>
2. <https://www.kaggle.com/resulcaliskan/countries-gdps>
3. <https://www.kaggle.com/centurion1986/countries-population>

Unfortunately, due to the lack of time a decision was made to exclude the recently complete Tokyo Olympic games. Although there was data sets available the actual informations such as individual winners and athletes names wasn’t easly sources able. It was attempted to use web scraping and try finding information but due to the limited time this was abandoned.

# Transformation

To get all the data into one single database the fellowing had to be conducted.

1. The use of Pandas function in Jupiter notebooks to load all CSV files

Athlete\_events.csv

A screenshot of a computer

Description automatically generated with medium confidence

Noc\_regions.csv

Text

Description automatically generated

w\_gdp.xls

Text

Description automatically generated

WorldPopulation.csv

Text

Description automatically generated

1. Each file was converted to a data frame but were loaded in sequences this allowed for a cleaner merge into one table.
2. In the ‘athlete events’ the medal column was adjusted where all the null values were changed to DWN ‘Did Not Win’ illustrated below.

A screenshot of a computer

Description automatically generated with medium confidence

1. Both the World Populations and GDP files had significant formatting issues as a result a a pandas function called “Pd.Melt” was used to swape the years from column to rows.as illustrated in the image. A few columns had to be dropped as they had no significant value.

A picture containing text, road, screenshot

Description automatically generated

Graphical user interface, text

Description automatically generated with medium confidence

1. As a result of merging all the information into one data frame a significantly large table was create. As illustrated below. This table would allow the team of data analysis at the company to answer there question around trends.

Graphical user interface

Description automatically generated

# Load

After merging the data frame into one the data frame was exported both in a csv and a excel file and loaded into a sql data base.

Postgres database

Graphical user interface

Description automatically generated

Text

Description automatically generated