ETL Project Proposal

By Nikunj Patel

The aim of this explore, transform and load (ETL) project is to provide a database for data analysts to use at Sportsbet. Allowing the team to explore historical data form past Olympic games, population and GDP data that can be used to analysis trends. The following are example

* Can a predication of a countries medal count be made using the following factors?
  1. GDP per capita
  2. Size of Olympic team
  3. Size of Population
  4. Home Advantage
* Top performing countries with most medals won?
* Are certain countries better at certain sports?

# Data Source:

The data used in this ETL project can be found on Kaggle.com at the following links

* 120 years of Olympic history: athletes and results

Link: <https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results>

* Countries GDPs

Link: <https://www.kaggle.com/resulcaliskan/countries-gdps>

* Countries Population

Link: <https://www.kaggle.com/centurion1986/countries-population>

* 2021 Olympics in Tokyo

Link: <https://www.kaggle.com/arjunprasadsarkhel/2021-olympics-in-tokyo>

# Project Time Breakdown:

Tuesday 31/08/2021

* Create a ETL Project Repository (send link to Daniel)
* Complete Project Proposal upload to GitHub Repository

Start Exploration phase

* Source the data sets from Kaggle.com place in resource folder in project repository.
* Create a notebook file and import all the require csv files
* Start a database schema based on the required data.
* Commit, Push to git hub under new branch ‘first draft’

Thursday 02/08/2021

* Finalize database schema

End of Exploration phase

Start Transformation phase

* Update, remove and combine the required data columns from both Olympic data sets.
* Check the format of the GDP and population data and combine with the required data form Olympic data sets to create one table.
* Any other data clean-up and preparation required.
* Upload changes to GitHub branch ‘draft 2’

End of Transformation Phase

Monday 06/09/2021

Start Load Phase

* Create the needed tables in pgAdmin with ProstgreSQL
* Check on the notebook file if the table has been imported correctly.
* Upload the PostgreSQL queries in pgAdmin to the project folder.
* Push the files to GitHub branch ‘draft 3’

End Load Phase