8/4/2021

Olympics 2021

ETL Project

By Heejun Seo, Nithiya, Litesh Samji

Contents

[Introduction 2](#_Toc78815270)

[Extraction 2](#_Toc78815271)

[Transformation 2](#_Toc78815272)

[Load 2](#_Toc78815273)

[Conclusions and recommendations 2](#_Toc78815274)

[References and appendices 2](#_Toc78815275)

# Introduction

Our project is based on the 2021 Olympics, compared data is list of all athletes participating in the games, their coaches, and the teams participating across the globe.

This contains the details of over 11,000 athletes, with 47 disciplines, along with 743 Teams taking part in the 2021(2020) Tokyo Olympics.

## Extraction

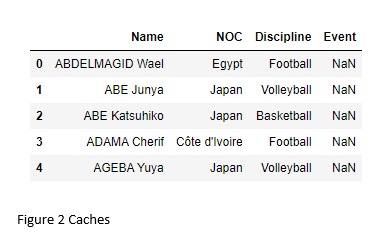
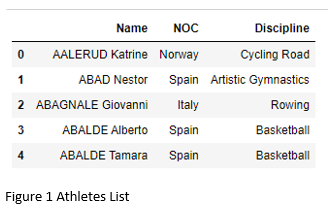
We used 3 datasets from the public Kaggle website.

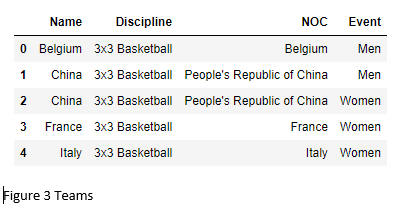
* Countries
* Discipline
* Athletes
* Coaches

Link: <https://www.kaggle.com/arjunprasadsarkhel/2021-olympics-in-tokyo?select=Coaches.xlsx>

In order to transform the public data and use it in our study we performed the following:

Using Jupyter Notebook, we used openpyxl python library to import the Excel files and then load into a Panda’s dataframe





## Transformation

Countries dataframe: Create a primary key country\_id, rename column NOC to country\_name and drop any duplicate values.

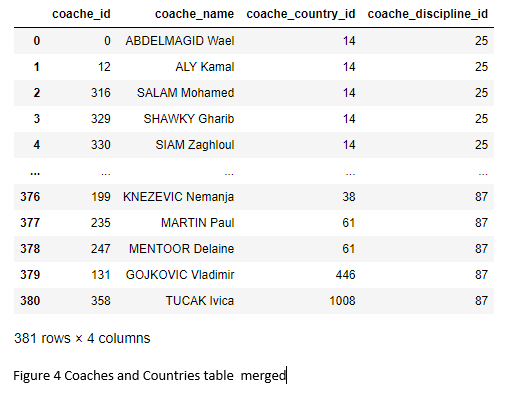
Discipline dataframe: Create a primary key discipline\_id, and drop any duplicate values

Athletes dataframe: Create a primary key athletes\_id, and drop any duplicate vales

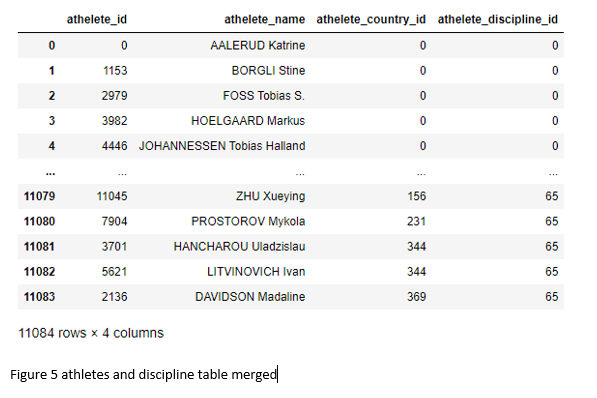
Coaches dataframe: Create a primary key coache\_id, and drop any duplicate values

Merge:

Coaches and country data frame merged:



Athletes and discipline table merged



## Load

We have used a relational database, Postgres SQL to load data. We then establish a connection to the SQL database called “olympics” using create\_engine() in SQLAlchemy were we have created four tables: countries, discipline, athlete, coaches

# Conclusions and recommendations

# References and appendices