

Data Analysis for SportsStats

Data Analysis Project Proposal for Olympics Dataset

Noviembre / 2021

Emer Isau Morales Vega

Overview

The data used for this analysis is the "SportsStats (Olympics Dataset - 120 years of data)" data set.

This dataset is made up of two files: `athlet_events.csv` and `noc_regions.csv`.

This data set was chosen because it reveals the Olympic medal award records for different sports categories, the countries that participated, names of the athletes, age, medals, etc.

Likewise, it can be said that the data is useful for news agencies that report on the different feats in the 120 years of the existence of the Olympic games.

Also, it can be of great importance for countries aspiring to improve their performance in subsequent Olympic events.




Import Data

Getting and Cleaning Data


Data analysis is done using python. The necessary resources are uploaded to google colab.

Here are some screenshots of how the data was imported into a Sqlite database



 athlete_events_coursera-task-1.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda [Se guardaron todos los cambios](#)

Comentar Compartir ⚙️ 

+ Código + Texto

✓ RAM Disco Editando

1 s

```
import pandas as pd
url = 'https://raw.githubusercontent.com/Emermv/sql-for-data-science/master/athlete_events.csv'
athlete_events = pd.read_csv(url)
athlete_events.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 15 columns):
#   Column      Non-Null Count  Dtype
---  -
0    ID          271116 non-null  int64
1    Name        271116 non-null  object
2    Sex         271116 non-null  object
3    Age         261642 non-null  float64
4    Height      210945 non-null  float64
5    Weight      208241 non-null  float64
6    Team        271116 non-null  object
7    NOC         271116 non-null  object
8    Games       271116 non-null  object
9    Year        271116 non-null  int64
10   Season      271116 non-null  object
11   City        271116 non-null  object
12   Sport       271116 non-null  object
13   Event       271116 non-null  object
14   Medal       39783 non-null   object
dtypes: float64(3), int64(2), object(10)
memory usage: 31.0+ MB
```

Import Data



athlete_events_coursera-task-1.ipynb ☆



Comentar



Compartir



Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda [Se guardaron todos los cambios](#)



+ Código + Texto



RAM
Disco



Editando



0 s



```
from pathlib import Path  
Path('olympics.db').touch()
```

▼ Creating a sqlite database

sqlite is a lightweight database that can be started as an empty text file. You can create the file with touch my_data.db or with this equivalent Python code:

Create database



0 s



```
link.execute('''create table regions(  
    NOC    varchar(45),  
    regions varchar(150),  
    notes text  
    ) ''')
```

```
<sqlite3.Cursor at 0x7fda48f8af80>
```

▼ Creating sqlite table

Create a database connection and cursor to execute queries.



```
import sqlite3  
conn = sqlite3.connect('olympics.db')  
link = conn.cursor()  
link.execute('''create table athletes(  
    ID int,  
    Name varchar(250),  
    Sex char(1),  
    Age float,  
    Height float,  
    Weight float,  
    Team    varchar(100),  
    NOC     varchar(45),  
    Games   varchar(100),  
    Year    int,  
    Season  varchar(45),  
    City    varchar(100),  
    Sport   varchar(100),  
    Event   varchar(100),  
    Medal   varchar(45)  
    ) ''')
```

Create tables



athlete_events_coursera-task-1.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda [Se guardaron todos los cambios](#)

Comentar



Compartir



+ Código + Texto

RAM
Disco

Editando



athlete_events.head(5)



	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

view data using python



+ Código + Texto

✓ RAM
Disco

Editando



```
[12] # write the data to a sqlite table
athlete_events.to_sql('olympics', conn, if_exists='append', index = False)
```



```
pd.read_sql("select * from olympics limit 10", conn)
```



	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	None
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	None
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	None
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	None
5	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 1,000 metres	None
6	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 500 metres	None

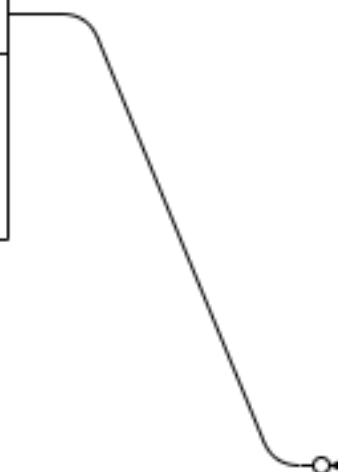
✓ 0 s se ejecutó 13:49



view data using SQL

Regions	
PK	<u>NOC varchar</u>
	Region varchar
	Notes varchar

Athletes	
PK	<u>Id int</u>
	Name varchar
	Sex char
	Age int
	Height float
	Weight float
	Team varchar
	NOC varchar
	Games varchar
	Year int
	Season varchar
	City varchar
	Sport varchar
	Event varchar
	Medal varchar



ERD

Questions to Answer

- **How many Peruvians won a gold medal?**
- **Has Cristiano Ronaldo (CR7) won any medals?**
- **Which sporting event gathered the most athletes?**
- **In which year were the most gold medals awarded?**
- **In the 120 years of the Olympic Games, which country's team won the most medals?**
- **Which sport wins the most gold medals?**





Initial Hypothesis

- **How many Peruvians won a gold medal?**

Of course there are Peruvians who won gold medals

- **Has Cristiano Ronaldo (CR7) won any medals?**

I think that CR7 has won a gold medal

- **Which sporting event gathered the most athletes?**

Without a doubt, football is the sporting event that gathers more athletes

- **In which year were the most gold medals awarded?**

I believe that more gold medals were awarded between the years 2006 to 2010

- **In the 120 years of the Olympic Games, which country's team won the most medals?**

From my perspective, Poland has more medals

- **Which sport wins the most gold medals?**

I believe that the sport that has won the most gold medals in the 120 years of the Olympics is Boxing

Data analysis Approach

The working environment will be in google colab with the python programming language and sqlite will be used to analyze the data.

In the first instance, the files `athlet_events.csv` and `noc_regions.csv` that are related through the NOC column will be loaded. That is, the column exists in both files.

Then it will be necessary to remove or replace the NA values for better analysis.

Finally, to answer the questions posed, statistical inference and graphic visualization will be used to determine if there is a relationship between the columns.

