

explore

July 9, 2023

Analyse exploratoire

1 Intro

```
[1]: # Import des bibliothèques
import os
os.chdir("..")
import pandas as pd
import numpy as np
import re
from querympics.utils.helper import parse_name, give_short_name

# Import des données
from querympics.data.cleaning import df, athlet, games, region, result
```

2 Athlètes

2.0.1 Athlètes les plus médaillés tous jeux confondus

```
[2]: # Les 20 athlètes les plus médaillés
n = 20
(athlet
 .join(result.groupby("athlete").sport.first())
 .join(result.groupby("athlete").medal.count())
 .rename(columns={"medal": "n_medals"})
 .sort_values(by="n_medals", ascending=False))[:n]
```

```
[2]:
```

	first_name	last_name	gender	birth_year	latteest_noc	\
id						
207089493835	Michael	PHELPS	M	1985	USA	
027503151640	Larysa	DIRIY	F	1935	URS	
779052604582	Nikolay	ANDRIANOV	M	1953	URS	
027629514478	Borys	SHAKHLIN	M	1932	URS	
473622409020	Ole	BJRNDALEN	M	1974	NOR	
206027787982	Edoardo	MANGIAROTTI	M	1919	ITA	
354705440133	Gustaf	CARLBERG	M	1880	SWE	
470290582797	Paavo	NURMI	M	1898	FIN	

429998356546	Sawao	KATO	M	1947	JPN
692484174312	Birgit	SCHMIDT	F	1962	GER
632866972353	Dara	MINAS	F	1967	USA
123462359385	Jennifer	CUMPELIK	F	1973	USA
151171364282	Takashi	ONO	M	1931	JPN
189908206539	Aleksey	NEMOV	M	1976	RUS
747407348886	Natalie	HALL	F	1983	USA
024676404810	Ryan	LOCHE	M	1984	USA
858118795892	Carl	OSBURN	M	1885	USA
650415523039	Allyson	FELIX	F	1986	USA
897714493094	Mark	SPITZ	M	1950	USA
317143000542	Vra	ODLOILOV	F	1942	TCH

	sport	n_medals
id		
207089493835	Swimming	28
027503151640	Gymnastics	18
779052604582	Gymnastics	15
027629514478	Gymnastics	13
473622409020	Biathlon	13
206027787982	Fencing	13
354705440133	Shooting	13
470290582797	Athletics	12
429998356546	Gymnastics	12
692484174312	Canoeing	12
632866972353	Swimming	12
123462359385	Swimming	12
151171364282	Gymnastics	12
189908206539	Gymnastics	12
747407348886	Swimming	12
024676404810	Swimming	12
858118795892	Shooting	11
650415523039	Athletics	11
897714493094	Swimming	11
317143000542	Gymnastics	11

2.1 Analyse sur la correction des âges (Tokyo 2020)

```
[3]: df[df["ShortName"] == "Leon MARCHAND"]
```

```
[3]:
```

	Name	Sex	Age	Team	NOC	Games	Year	Season	\
index									
S230677	MARCHAND	Leon	M	19	France	FRA	2020	Summer	2020 Summer
S230678	MARCHAND	Leon	M	19	France	FRA	2020	Summer	2020 Summer
S230679	MARCHAND	Leon	M	19	France	FRA	2020	Summer	2020 Summer
S230680	MARCHAND	Leon	M	19	France	FRA	2020	Summer	2020 Summer

	City	Sport	Event	Medal	FirstName	\
index						
S230677	Tokyo	Swimming	Men's 200m Butterfly	NaN	Leon	
S230678	Tokyo	Swimming	Men's 200m Individual Medley	NaN	Leon	
S230679	Tokyo	Swimming	Men's 400m Individual Medley	NaN	Leon	
S230680	Tokyo	Swimming	Men's 4 x 100m Medley Relay Team	NaN	Leon	

	LastName	ShortName	BirthYear	ApproxBirthYear	AthleteID	\
index						
S230677	MARCHAND	Leon MARCHAND	2002	2002	233890689901	
S230678	MARCHAND	Leon MARCHAND	2002	2002	233890689901	
S230679	MARCHAND	Leon MARCHAND	2002	2002	233890689901	
S230680	MARCHAND	Leon MARCHAND	2002	2002	233890689901	

	GamesID
index	
S230677	S2020
S230678	S2020
S230679	S2020
S230680	S2020

```
[9]: athlet.count()
```

```
[9]: first_name      147426
last_name          147426
gender              147426
birth_year          141092
latteest_noc         147426
dtype: int64
```

```
[4]: athlet.count()
```

```
[4]: first_name      146452
last_name          146452
gender              146452
birth_year          140118
latteest_noc         146452
dtype: int64
```

```
[5]: # Différence en nombre d'athlètes uniques
147426 - 146452
```

```
[5]: 974
```

```
[6]: # Différence en nombre d'athlètes uniques
# qui ont une année de naissance renseignée
141092 - 140118
```

[6] : 974