explore

June 24, 2023

Analyse exploratoire

1 Intro

[1]:

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

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

[2]:

```
print("'athlet':")
print(f"{athlet.shape[0]} lignes")
print(f"{athlet.shape[1]} colonnes")
```

```
'athlet':
286237 lignes
19 colonnes
```

[3]:

```
athlet.head(3)
```

[3]:

```
                        Name Sex  Age     Team  NOC        Games  Year  \
index
S000001            A Dijiang   M   24    China  CHN  1992 Summer  1992
S000002             A Lamusi   M   23    China  CHN  2012 Summer  2012
S000003  Gunnar Nielsen Aaby   M   24  Denmark  DEN  1920 Summer  1920

         Season       City      Sport                         Event Medal  \
index
S000001  Summer  Barcelona  Basketball   Basketball Men's Basketball   NaN
S000002  Summer     London        Judo  Judo Men's Extra-Lightweight   NaN
S000003  Summer  Antwerpen    Football       Football Men's Football   NaN
```

```
        FirstName LastName    ShortName  BirthYear  ApproxBirthYear  \
index
S000001   Dijiang        A    Dijiang A       1968             1968
S000002    Lamusi        A     Lamusi A       1989             1988
S000003    Gunnar     AABY  Gunnar AABY       1896             1896

            AthleteID GamesID
index
S000001  995417935682   S1992
S000002  805993218285   S2012
S000003  559659414555   S1920
```

[4]:
```
person
```

[4]:
```
              first_name     last_name gender  birth_year lattest_noc
id
000004934552       Lidia       SORIANO      F        1969         ESP
000007112158     Rosella     CICOGNANI      F        1940         ITA
000008459775  Alessandro        RASINI      M        1919         ITA
000013733620       Peter  SCHLAGENHAUF      M        1961         SUI
000018050719      Tebogo       MOERANE      M        1995         RSA
...                ...         ...  ...       ...       ...
999943434496       Jaime       ROBERTS      F        1990         AUS
999944047748         Ben      SANDFORD      M        1980         NZL
999948773968        Taip      RAMADANI      M        1972         AUS
999958826255    Benjamin       SIRIMOU      M        1969         CMR
999997177985        Jong           UNG      M        1957         PRK

[147426 rows x 5 columns]
```

[5]:
```
games
```

[5]:
```
       year  season                    city
id
S1896  1896  Summer                  Athina
S1900  1900  Summer                   Paris
S1904  1904  Summer               St. Louis
S1906  1906  Summer                  Athina
S1908  1908  Summer                  London
S1912  1912  Summer               Stockholm
S1920  1920  Summer               Antwerpen
S1924  1924  Summer                   Paris
W1924  1924  Winter                Chamonix
S1928  1928  Summer               Amsterdam
W1928  1928  Winter            Sankt Moritz
S1932  1932  Summer             Los Angeles
W1932  1932  Winter             Lake Placid
```

```
S1936  1936  Summer                  Berlin
W1936  1936  Winter  Garmisch-Partenkirchen
S1948  1948  Summer                  London
W1948  1948  Winter            Sankt Moritz
S1952  1952  Summer                Helsinki
W1952  1952  Winter                    Oslo
S1956  1956  Summer               Melbourne
W1956  1956  Winter       Cortina d'Ampezzo
S1960  1960  Summer                    Roma
W1960  1960  Winter            Squaw Valley
S1964  1964  Summer                   Tokyo
W1964  1964  Winter               Innsbruck
S1968  1968  Summer             Mexico City
W1968  1968  Winter                Grenoble
S1972  1972  Summer                  Munich
W1972  1972  Winter                 Sapporo
S1976  1976  Summer                Montreal
W1976  1976  Winter               Innsbruck
S1980  1980  Summer                  Moskva
W1980  1980  Winter             Lake Placid
S1984  1984  Summer             Los Angeles
W1984  1984  Winter                Sarajevo
S1988  1988  Summer                   Seoul
W1988  1988  Winter                 Calgary
S1992  1992  Summer               Barcelona
W1992  1992  Winter             Albertville
W1994  1994  Winter             Lillehammer
S1996  1996  Summer                 Atlanta
W1998  1998  Winter                  Nagano
S2000  2000  Summer                  Sydney
W2002  2002  Winter          Salt Lake City
S2004  2004  Summer                  Athina
W2006  2006  Winter                  Torino
S2008  2008  Summer                 Beijing
W2010  2010  Winter               Vancouver
S2012  2012  Summer                  London
W2014  2014  Winter                   Sochi
S2016  2016  Summer          Rio de Janeiro
S2020  2020  Summer                   Tokyo
```

[6]:

```
result
```

[6]:

```
               games       sport                                     event  \
id
S000001  1992 Summer  Basketball               Basketball Men's Basketball
S000002  2012 Summer        Judo              Judo Men's Extra-Lightweight
S000003  1920 Summer    Football                   Football Men's Football
```

```
S000004  1900 Summer   Tug-Of-War               Tug-Of-War Men's Tug-Of-War
S000005  1932 Summer    Athletics              Athletics Women's 100 metres
…              …          …                                       …
W048560  1976 Winter         Luge                Luge Mixed (Men)'s Doubles
W048561  2014 Winter  Ski Jumping  Ski Jumping Men's Large Hill, Individual
W048562  2014 Winter  Ski Jumping        Ski Jumping Men's Large Hill, Team
W048563  1998 Winter    Bobsleigh                      Bobsleigh Men's Four
W048564  2002 Winter    Bobsleigh                      Bobsleigh Men's Four

              athlete  noc medal
id
S000001  995417935682  CHN   NaN
S000002  805993218285  CHN   NaN
S000003  559659414555  DEN   NaN
S000004  341882753977  DEN  Gold
S000005  027272868027  NED   NaN
…                …  …   …
W048560  382070033368  POL   NaN
W048561  677219328541  POL   NaN
W048562  677219328541  POL   NaN
W048563  579073266868  POL   NaN
W048564  579073266868  POL   NaN

[286237 rows x 6 columns]
```

[8]:
```
athlet.head(3)
```

[8]:
```
                        Name Sex  Age     Team  NOC        Games  Year  \
index
S000001            A Dijiang   M   24    China  CHN  1992 Summer  1992
S000002             A Lamusi   M   23    China  CHN  2012 Summer  2012
S000003  Gunnar Nielsen Aaby   M   24  Denmark  DEN  1920 Summer  1920

         Season       City       Sport                         Event Medal  \
index
S000001  Summer  Barcelona  Basketball   Basketball Men's Basketball   NaN
S000002  Summer     London        Judo  Judo Men's Extra-Lightweight   NaN
S000003  Summer  Antwerpen    Football       Football Men's Football   NaN

        FirstName LastName    ShortName  BirthYear  ApproxBirthYear   \
index
S000001   Dijiang        A    Dijiang A       1968             1968
S000002    Lamusi        A     Lamusi A       1989             1988
S000003    Gunnar     AABY  Gunnar AABY       1896             1896

            AthleteID
index
```

```
S000001  995417935682
S000002  805993218285
S000003  559659414555
```

[67]:

```
person[(person["FirstName"] == "Aaron") & (person["LastName"] == "COOK")]
```

[67]:

```
             FirstName LastName  BirthYear  NOC
AthleteID
909174037505     Aaron     COOK       1991  MDA
```

[43]:

```
athlet[(athlet["Year"] == 1956) & (athlet["Season"] == "Summer")].City.
 ↪value_counts()
```

[43]:

```
City
Melbourne    4829
Stockholm     298
Name: count, dtype: int64
```

[29]:

```
athlet[(athlet["Year"] == 1956) &
       (athlet["Season"] == "Summer") &
       (athlet["City"] == "Stockholm")]
```

[29]:

```
                                               Name Sex  Age           Team  \
index
S045958                           Raimondo D'Inzeo   M   31          Italy
S045942                              Piero D'Inzeo   M   33          Italy
S043256               Eloy Massey Oliveira de Menezes   M   45         Brazil
S180757  Joaquim Miguel Matos Fernandes Duarte Silva   M   32       Portugal
S180758  Joaquim Miguel Matos Fernandes Duarte Silva   M   32       Portugal
…                                               …  ..  …            …
S183992   Patricia Rosemary "Pat" Smythe (-Koechlin)   F   27  Great Britain
S183991   Patricia Rosemary "Pat" Smythe (-Koechlin)   F   27  Great Britain
S213975                                  Hugh Wiley   M   27  United States
S029295                            Adriano Capuzzo   M   28          Italy
S029294                            Adriano Capuzzo   M   28          Italy

         NOC        Games  Year  Season       City         Sport  \
index
S045958  ITA  1956 Summer  1956  Summer  Stockholm  Equestrianism
S045942  ITA  1956 Summer  1956  Summer  Stockholm  Equestrianism
S043256  BRA  1956 Summer  1956  Summer  Stockholm  Equestrianism
S180757  POR  1956 Summer  1956  Summer  Stockholm  Equestrianism
S180758  POR  1956 Summer  1956  Summer  Stockholm  Equestrianism
…        …          …   …     …       …          …
S183992  GBR  1956 Summer  1956  Summer  Stockholm  Equestrianism
S183991  GBR  1956 Summer  1956  Summer  Stockholm  Equestrianism
S213975  USA  1956 Summer  1956  Summer  Stockholm  Equestrianism
```

```
S029295  ITA  1956 Summer  1956  Summer  Stockholm  Equestrianism
S029294  ITA  1956 Summer  1956  Summer  Stockholm  Equestrianism

                                                   Event   Medal FirstName  \
index
S045958                Equestrianism Mixed Jumping, Team  Silver  Raimondo
S045942          Equestrianism Mixed Jumping, Individual  Bronze     Piero
S043256          Equestrianism Mixed Jumping, Individual     NaN      Eloy
S180757  Equestrianism Men's Three-Day Event, Individual     NaN   Joaquim
S180758        Equestrianism Men's Three-Day Event, Team     NaN   Joaquim
…                                                  …     …       …
S183992                Equestrianism Mixed Jumping, Team  Bronze  Patricia
S183991          Equestrianism Mixed Jumping, Individual     NaN  Patricia
S213975                Equestrianism Mixed Jumping, Team     NaN      Hugh
S029295        Equestrianism Men's Three-Day Event, Team     NaN   Adriano
S029294  Equestrianism Men's Three-Day Event, Individual     NaN   Adriano

           LastName          ShortName  BirthYear   YOB     AthleteID
index
S045958     D'INZEO   Raimondo D'INZEO       1925  1924  732169054991
S045942     D'INZEO      Piero D'INZEO       1923  1922  645118408004
S043256  DE MENEZES    Eloy DE MENEZES       1911  1910  762428348578
S180757       SILVA      Joaquim SILVA       1924  1924  969858199022
S180758       SILVA      Joaquim SILVA       1924  1924  969858199022
…             …                …        …   …           …
S183992    KOECHLIN  Patricia KOECHLIN       1929  1928  914348522509
S183991    KOECHLIN  Patricia KOECHLIN       1929  1928  914348522509
S213975       WILEY         Hugh WILEY       1929  1928  667173293696
S029295     CAPUZZO    Adriano CAPUZZO       1928  1928  334486418671
S029294     CAPUZZO    Adriano CAPUZZO       1928  1928  334486418671

[298 rows x 18 columns]
```

[8]:

```
athlet[~athlet.Medal.isnull()].Medal.apply(len).max()
```

[8]:

```
6
```

[5]:

```
athlet.Sport.apply(len).max()
```

[5]:

```
25
```

[4]:

```
person
```

[4]:

```
                  id  first_name     last_name gender  birth_year lattest_noc
0       000004934552       Lidia       SORIANO      F        1969         ESP
6       000007112158     Rosella     CICOGNANI      F        1940         ITA
12      000008459775  Alessandro        RASINI      M        1919         ITA
```

```
13      000013733620       Peter  SCHLAGENHAUF      M        1961         SUI
14      000018050719      Tebogo       MOERANE      M        1995         RSA
…              …         …           …    …         …         …
286225  999943434496       Jaime       ROBERTS      F        1990         AUS
286227  999944047748         Ben      SANDFORD      M        1980         NZL
286230  999948773968        Taip      RAMADANI      M        1972         AUS
286231  999958826255    Benjamin       SIRIMOU      M        1969         CMR
286235  999997177985        Jong           UNG      M        1957         PRK

[147426 rows x 6 columns]
```

2 Athlètes : Nettoyage des données

2.1 Extraire les noms

[4]:
```
person = (pd.DataFrame(athlet
                        .groupby("Name").Event.agg('count')
                        .sort_values(ascending=False)
                        .reset_index()))
person.head()
```

[4]:
```
                       Name  Event
0      Robert Tait McKenzie     58
1  Heikki Ilmari Savolainen     39
2     Joseph "Josy" Stoffel     38
3       Ioannis Theofilakis     36
4               Takashi Ono     33
```

[5]:
```
person[person.Name.str.contains("Phelps")]
```

[5]:
```
                                         Name  Event
15                    Michael Fred Phelps, II     30
4169               Richard Lawson Phelps, Jr.      7
5084                Robert Lawson Phelps, Sr.      6
7151    Jaycie Lynn Phelps (-McClure, -Marus)      6
9352     Monica Kathleen Rutherford (-Phelps)      5
31508                    Mason Elliott Phelps      2
32809                   Harlow Phelps Rothert      2
39453               John Phelps "Jack" George      2
58195                       Brian Eric Phelps      2
72932                  Richard Charles Phelps      1
74357                    Robert Edward Phelps      1
91583                            Peter Phelps      1
138023              Harold Roy "Pete" Phelps      1
```

[6]:

```
person["ShortName"] = person.Name.apply(give_short_name)
person["FirstName"] = person.Name.apply(lambda s: parse_name(s)["first"])
person["LastName"] = person.Name.apply(lambda s: parse_name(s)["last"].upper())
person = person[["Name", "FirstName", "LastName", "ShortName", "Event"]]
person.head()
```

[6]:

```
                       Name FirstName     LastName           ShortName  Event
0      Robert Tait McKenzie    robert     MCKENZIE     Robert MCKENZIE     58
1  Heikki Ilmari Savolainen    heikki   SAVOLAINEN   Heikki SAVOLAINEN     39
2     Joseph "Josy" Stoffel    joseph      STOFFEL      Joseph STOFFEL     38
3       Ioannis Theofilakis   ioannis  THEOFILAKIS Ioannis THEOFILAKIS     36
4               Takashi Ono   takashi          ONO         Takashi ONO     33
```

[7]:

```
short_names = (person
 .groupby("ShortName")
 .sum("Event")
 .sort_values(by="Event", ascending=False))
short_names.to_csv("../../draft/athlete_short_names.csv")
short_names.head(3)
```

[7]:

```
                   Event
ShortName
Robert MCKENZIE       58
Gustaf CARLBERG       49
Heikki SAVOLAINEN     39
```

[8]:

```
person[person["LastName"] == "SCHMIDT"]
```

[8]:

```
                                    Name  FirstName LastName  \
2662                       Iohan Schmidt      iohan  SCHMIDT
7159          Magdalena Schmidt (-Jakob)  magdalena  SCHMIDT
8132        Oscar Daniel Bezerra Schmidt      oscar  SCHMIDT
10367                    Florian Schmidt    florian  SCHMIDT
15015             Ingrid Schmidt (-Naue)     ingrid  SCHMIDT
…                                  …        …      …
133023                   Herbert Schmidt    herbert  SCHMIDT
133885                  Heinrich Schmidt   heinrich  SCHMIDT
134636  Ingrid Ulrike Schmidt (-Koppers)     ingrid  SCHMIDT
136868                    Gustav Schmidt     gustav  SCHMIDT
141995                     Josef Schmidt      josef  SCHMIDT

                ShortName  Event
2662        Iohan SCHMIDT      8
7159    Magdalena SCHMIDT      6
8132        Oscar SCHMIDT      5
10367     Florian SCHMIDT      4
15015      Ingrid SCHMIDT      4
```

```
…                   …    …
133023    Herbert SCHMIDT      1
133885   Heinrich SCHMIDT      1
134636     Ingrid SCHMIDT      1
136868     Gustav SCHMIDT      1
141995      Josef SCHMIDT      1

[89 rows x 5 columns]
```

[9]:

```python
names = (person
            .groupby(["ShortName", "FirstName", "LastName", "Name"])
            .sum("Event").reset_index()
            .sort_values(by="Event", ascending=False))
names.to_csv("../../draft/athlete_names.csv", index=None, chunksize=10000)
names
```

[9]:

```
                  ShortName FirstName     LastName  \
116699      Robert MCKENZIE    robert     MCKENZIE
51535     Heikki SAVOLAINEN    heikki   SAVOLAINEN
69622        Joseph STOFFEL    joseph      STOFFEL
56880   Ioannis THEOFILAKIS   ioannis  THEOFILAKIS
129728          Takashi ONO   takashi          ONO
…                     …       …          …
61070       Janier HERNNDEZ    janier     HERNNDEZ
61069            Janie REED     janie         REED
61067        Janice TROMBLY    janice      TROMBLY
61066       Janice TEIXEIRA    janice     TEIXEIRA
146360       Zzimo CALAZANS     zzimo     CALAZANS

                               Name  Event
116699         Robert Tait McKenzie     58
51535      Heikki Ilmari Savolainen     39
69622         Joseph "Josy" Stoffel     38
56880           Ioannis Theofilakis     36
129728                  Takashi Ono     33
…                             …    …
61070     Janier Concepcin Hernndez      1
61069                    REED Janie      1
61067   Janice Yvonne "Jan" Trombly      1
61066           Janice Gil Teixeira      1
146360         Zzimo Alves Calazans      1

[146361 rows x 5 columns]
```

[10]:

```python
print(146361 - 141513)
```

```
4848
```

[11]:

```
person[person["ShortName"] == "Zsuzsanna JAKABOS"]
```

[11]:

```
                         Name  FirstName LastName          ShortName  Event
1476    Zsuzsanna "Zsu" Jakabos  zsuzsanna  JAKABOS  Zsuzsanna JAKABOS     10
144651        JAKABOS Zsuzsanna  zsuzsanna  JAKABOS  Zsuzsanna JAKABOS      1
```

2.2 Age : passer en entier

[12]:

```
# On s'assure que tous les ages sont entiers
# avant de les convertir en "int"
athlet.Age.apply(lambda x: x.is_integer()).value_counts()
```

[12]:

```
Age
True     276763
False      9474
Name: count, dtype: int64
```

[13]:

```
# On s'assure que les ages non entiers
# ne correspondent qu'aux valeurs vides
athlet["intAge"] = athlet.Age.apply(lambda x: x.is_integer())
athlet.Age.isna().sum() == athlet[athlet["intAge"] == False].shape[0]
```

[13]:

```
True
```

[14]:

```
#assert len(athlete[athlete.Age == 0]) == 0
#athlete["Age"].fillna(0, inplace=True)
#athlete["Age"] = athlete.Age.astype("int64")
#athlete.replace(0, np.nan, inplace=True)
athlet["Age"] = athlet.Age.astype("Int64")
athlet.drop(columns=["intAge"], inplace=True)
athlet.head()
```

[14]:

```
                                     Name Sex  Age            Team  NOC  \
index
S000001                         A Dijiang   M   24           China  CHN
S000002                          A Lamusi   M   23           China  CHN
S000003                Gunnar Nielsen Aaby   M   24         Denmark  DEN
S000004               Edgar Lindenau Aabye   M   34  Denmark/Sweden  DEN
S000005  Cornelia "Cor" Aalten (-Strannood)   F   18     Netherlands  NED

               Games  Year  Season         City       Sport  \
index
S000001  1992 Summer  1992  Summer    Barcelona  Basketball
S000002  2012 Summer  2012  Summer       London        Judo
S000003  1920 Summer  1920  Summer    Antwerpen    Football
S000004  1900 Summer  1900  Summer        Paris  Tug-Of-War
S000005  1932 Summer  1932  Summer  Los Angeles   Athletics
```

```
                                Event Medal
index
S000001   Basketball Men's Basketball   NaN
S000002  Judo Men's Extra-Lightweight   NaN
S000003       Football Men's Football   NaN
S000004   Tug-Of-War Men's Tug-Of-War  Gold
S000005  Athletics Women's 100 metres   NaN
```

[15]:

```
athlet[athlet.Age.isna()]
```

[15]:

```
                          Name Sex   Age          Team  NOC        Games  \
index
S000086  Mohamed Jamshid Abadi   M  <NA>          Iran  IRI  1948 Summer
S000091       Georgi Abadzhiev   M  <NA>      Bulgaria  BUL  1924 Summer
S000092       Georgi Abadzhiev   M  <NA>      Bulgaria  BUL  1924 Summer
S000101        Mohamed Abakkar   M  <NA>         Sudan  SUD  1972 Summer
S000151      Sayed Fahmy Abaza   M  <NA>         Egypt  EGY  1920 Summer
…                        …  ..   …            …  …           …
W047155          Boris Yakimov   M  <NA>  Soviet Union  URS  1956 Winter
W047156          Boris Yakimov   M  <NA>  Soviet Union  URS  1956 Winter
W047713       Luciano Zampatti   M  <NA>         Italy  ITA  1928 Winter
W047757        Ernesto Zardini   M  <NA>         Italy  ITA  1932 Winter
W047758        Ernesto Zardini   M  <NA>         Italy  ITA  1932 Winter

         Year  Season               City            Sport  \
index
S000086  1948  Summer             London           Boxing
S000091  1924  Summer              Paris          Cycling
S000092  1924  Summer              Paris          Cycling
S000101  1972  Summer             Munich           Boxing
S000151  1920  Summer          Antwerpen         Football
…       …     …                …              …
W047155  1956  Winter  Cortina d'Ampezzo    Speed Skating
W047156  1956  Winter  Cortina d'Ampezzo    Speed Skating
W047713  1928  Winter       Sankt Moritz      Ski Jumping
W047757  1932  Winter        Lake Placid      Ski Jumping
W047758  1932  Winter        Lake Placid  Nordic Combined

                                             Event Medal
index
S000086                   Boxing Men's Heavyweight   NaN
S000091        Cycling Men's Road Race, Individual   NaN
S000092              Cycling Men's Road Race, Team   NaN
S000101                     Boxing Men's Flyweight   NaN
S000151                    Football Men's Football   NaN
…                                             …   …
```

```
W047155           Speed Skating Men's 5,000 metres   NaN
W047156          Speed Skating Men's 10,000 metres   NaN
W047713  Ski Jumping Men's Normal Hill, Individual   NaN
W047757  Ski Jumping Men's Normal Hill, Individual   NaN
W047758           Nordic Combined Men's Individual   NaN

[9474 rows x 12 columns]
```

2.3 Age : déterminer l'année de naissance

[ ]:

```python
athlet["YOB"] = athlet["Year"] - athlet["Age"]
```

2.4 Sex

[4]:

```python
print("Répartition des athlètes par genre :")
print(athlet["Sex"].value_counts())
print("en proportion :")
print(athlet["Sex"].value_counts(normalize=True))
```

```
Répartition des athlètes par genre :
Sex
M    204449
F     81788
Name: count, dtype: int64
en proportion :
Sex
M    0.714265
F    0.285735
Name: proportion, dtype: float64
```

[5]:

```python
# On vérifie qu'il n'y a pas de valeurs vides pour le genre
assert athlet[athlet["Sex"].isnull()].shape[0] == 0
```

[6]:

```python
# On vérifie si chaque athlète (regroupé sous le même nom court)
# a un unique genre identifié dans la base
athlet.groupby("ShortName").Sex.nunique().describe()
```

[6]:

```
count    141513.000000
mean          1.000530
std           0.023015
min           1.000000
25%           1.000000
50%           1.000000
75%           1.000000
max           2.000000
Name: Sex, dtype: float64
```

```
[7]: n_gender = athlet.groupby("ShortName").Sex.nunique().reset_index()
     print(n_gender[n_gender["Sex"] == 2].count(), "pers. identifiées sous 2 genres")
     n_gender[n_gender["Sex"] == 2].to_csv("../../draft/two_genders.csv", index=None)
     n_gender[n_gender["Sex"] == 2]
```

```
ShortName    75
Sex          75
dtype: int64 pers. identifiées sous 2 genres
```

```
[7]:             ShortName  Sex
     4486      Alex MORGAN    2
     5118   Alexis THOMPSON    2
     20715        Chan LUI    2
     21672      Chen HONG    2
     21690      Chen JING    2
     …                  …    …
     140705       Zhang LI    2
     140717      Zhang MIN    2
     140719      Zhang NAN    2
     140729     Zhang QING    2
     141020       Zhu TING    2

     [75 rows x 2 columns]
```

N.B : certaines personnes sont peut-êtres identifiées sous un même nom court car elles portent un prénom mixte et un nom courant.

Ex : la joueuse de football américaine Alex Morgan a peut-être un homonyme masculin.

```
[8]: athlet[athlet.ShortName == "Alex MORGAN"]
```

```
[8]:                Name Sex  Age           Team  NOC        Games  Year  Season  \
     index
     S134519  Alex Morgan   M   23        Jamaica  JAM  1996 Summer  1996  Summer
     S231595  MORGAN Alex   F   32  United States  USA  2020 Summer  2020  Summer

                 City      Sport                       Event   Medal    ShortName
     index
     S134519  Atlanta  Athletics  Athletics Men's 800 metres     NaN  Alex MORGAN
     S231595    Tokyo   Football                  Women Team  Bronze  Alex MORGAN
```

Il faudrait un ID par athlète qui nous permetrait de distinguer les 2. Il pourrait être composé comme ça 4 lettres du prénom + 4 lettres du nom + les 3 lettres du NOC + un chiffre détrompeur qui distingue si c'est une femme ou un homme

ainsi on aurait * `ALEXMORGUSA0` pour la joueuse de football américaine * `ALEXMORGJAM1` pour le coureur jamaicain

```
[10]: athlet[athlet.ShortName == "Chan LUI"]
```

```
[10]:                  Name Sex  Age       Team  NOC        Games  Year  Season  \
index
S032033  Chan Fai Lui   M   21  Hong Kong  HKG  1976 Summer  1976  Summer
S032034  Chan Fai Lui   M   21  Hong Kong  HKG  1976 Summer  1976  Summer
S032035  Chan Fai Lui   M   21  Hong Kong  HKG  1976 Summer  1976  Summer
S032083  Chan Tan Lui   F   23  Hong Kong  HKG  1992 Summer  1992  Summer
S032084  Chan Tan Lui   F   23  Hong Kong  HKG  1992 Summer  1992  Summer
S032085  Chan Tan Lui   F   27  Hong Kong  HKG  1996 Summer  1996  Summer
S032086  Chan Tan Lui   F   27  Hong Kong  HKG  1996 Summer  1996  Summer

              City         Sport  \
index
S032033   Montreal       Cycling
S032034   Montreal       Cycling
S032035   Montreal       Cycling
S032083  Barcelona  Table Tennis
S032084  Barcelona  Table Tennis
S032085    Atlanta  Table Tennis
S032086    Atlanta  Table Tennis

                                                Event Medal ShortName
index
S032033           Cycling Men's Road Race, Individual   NaN  Chan LUI
S032034  Cycling Men's 100 kilometres Team Time Trial   NaN  Chan LUI
S032035         Cycling Men's 1,000 metres Time Trial   NaN  Chan LUI
S032083                  Table Tennis Women's Singles   NaN  Chan LUI
S032084                  Table Tennis Women's Doubles   NaN  Chan LUI
S032085                  Table Tennis Women's Singles   NaN  Chan LUI
S032086                  Table Tennis Women's Doubles   NaN  Chan LUI
```

Les noms asiatiques semblent avoir été parsés de manière imparfaite. Pour plusieurs raisons : * Dans plusieurs pays asiatiques (comme en Chine, à Hong-Kong ou en Corée), le nom de famille est placé au début du nom or la fonction `parse_name` pense qu'il s'agit de noms de famille * Dans plusieurs pays asiatiques il y a 3 noms au lieu de 2

Conclusion il faudrait que `parse_name` regarde également la nationalité de l'athlète

Edit : mauvaise idée d'inclure le NOC dans l'id de l'athlète car le nombre d'athlète qui changent de NOC est plutôt courant

2.5 NOC

```
[8]: n_noc = athlet.groupby("AthleteID").NOC.nunique().reset_index()
```

```
[9]: n_noc.shape[0]
```

```
[9]: 147426
```

[10]:

```
print(n_noc.NOC.value_counts())
print("En proportions :")
print(n_noc.NOC.value_counts(normalize=True))
```

```
NOC
1    145756
2      1575
3        90
4         5
Name: count, dtype: int64
En proportions :
NOC
1    0.988672
2    0.010683
3    0.000610
4    0.000034
Name: proportion, dtype: float64
```

[11]:

```
mult_noc = n_noc[n_noc["NOC"] > 1]
mult_noc.sort_values(by="NOC")
```

[11]:

```
                       AthleteID  NOC
65                AARONCOOK11990    2
104578       OKSANARAVILOVA01968    2
104556      OKSANAGRISHCHUK01972    2
104510       OGNJENSOKOLOVI11964    2
103986     NORFALIAVILLEGAS01964    2
…                            …  …
57510            IRINAFURLER01972    4
95789           MICHALIWISKI11970    4
61838          JASNABRAJKOVI01966    4
56374          ILIJALUPULESKU11968    4
86975   MAKHARBEKKHADARTSEV11966    4

[1670 rows x 2 columns]
```

[12]:

```
person[person["AthleteID"] == "AARONCOOK11990"]
```

[12]:

```
          AthleteID FirstName LastName   YOB  NOC
125  AARONCOOK11990     Aaron     COOK  1990  MDA
```

[13]:

```
assert mult_noc.groupby("AthleteID").count().shape[0] == mult_noc.shape[0]


m = mult_noc.merge(person[["AthleteID", "FirstName", "LastName", "YOB", "NOC"]],
                   on="AthleteID", how="left")
m
```

[13]:

```
                    AthleteID  NOC_x FirstName      LastName   YOB NOC_y
0              AARONCOOK11990      2     Aaron          COOK  1990   MDA
1              AASMMMDOV11980      2       Aas        MMMDOV  1980   AZE
2            ABANTRSTENA11964      2      Aban       TRSTENA  1964   MKD
3            ABBASKHAMIS11932      2     Abbas        KHAMIS  1932   EGY
4              ABBASQALI11992      2     Abbas          QALI  1992   KUW
…                       …    …       …            …   …   …
1665       ZORANPRIMORAC11968      2     Zoran      PRIMORAC  1968   CRO
1666       ZORANSOKOLOVI11966      2     Zoran      SOKOLOVI  1966   BIH
1667            ZSOLTBAL11972      2     Zsolt           BAL  1972   HUN
1668     ZULFIYAZABIROVA01974      2   Zulfiya      ZABIROVA  1974   KAZ
1669  ZURABIDATUNASHVILI11990      2    Zurabi  DATUNASHVILI  1990   SRB

[1670 rows x 6 columns]
```

[14]:

```python
athlet[athlet["AthleteID"] == "AARONCOOK11990"]
```

[14]:

```
                      Name Sex  Age           Team  NOC        Games  Year  \
index
S037075  Aaron Arthur Cook   M   25        Moldova  MDA  2016 Summer  2016
S037074  Aaron Arthur Cook   M   17  Great Britain  GBR  2008 Summer  2008

         Season            City      Sport                         Event  \
index
S037075  Summer  Rio de Janeiro  Taekwondo  Taekwondo Men's Welterweight
S037074  Summer         Beijing  Taekwondo  Taekwondo Men's Welterweight

        Medal FirstName LastName   ShortName  BirthYear   YOB       AthleteID
index
S037075   NaN     Aaron     COOK  Aaron COOK       1991  1990  AARONCOOK11990
S037074   NaN     Aaron     COOK  Aaron COOK       1991  1990  AARONCOOK11990
```

[16]:

```python
m.to_csv("../../draft/mult_noc.final.csv", index=False)
```

[17]:

```python
person
```

[17]:

```
                  AthleteID     FirstName  LastName   YOB  NOC
0          AADAMKHAMIS11988         Aadam    KHAMIS  1988  BRN
1        AADJIJATMIKOH01972  Aadjijatmiko         H  1972  INA
3       AADOLFSVANSTRM11886        Aadolf  SVANSTRM  1886  FIN
5          AAFKEHAMENT01970         Aafke    HAMENT  1970  NED
6         AAGEANDERSEN11884          Aage  ANDERSEN  1884  DEN
…                     …           …       …   …  …
286228    ZYGMUNTWEISS11902       Zygmunt     WEISS  1902  POL
286232       ZYTAJARKA01962          Zyta     JARKA  1962  POL
286233         ZZETNCE11980          Zzet       NCE  1980  TUR
286235       ZZETSAFER11990          Zzet     SAFER  1990  TUR
```

```
286236   ZZIMOCALAZANS11932         Zzimo  CALAZANS  1932  BRA

[147426 rows x 5 columns]
```

[17]:

```
athlet[athlet["AthleteID"] == "JOSDE L1"]
```

[17]:

```
                                                      Name Sex   Age  \
index
S004941  Jos Mara lvarez de las Asturias Bohorques y Go…   M    29
S004942  Jos Mara lvarez de las Asturias Bohorques y Go…   M    29
S004943  Jos Mara lvarez de las Asturias Bohorques y Go…   M    33
S004944  Jos Mara lvarez de las Asturias Bohorques y Go…   M    33
S015860                      Jos Reynaldo Bencosme de Leon   M    20
S027705                     Jos de la Cruz Cabrera Gandera   M  <NA>
S027706                     Jos de la Cruz Cabrera Gandera   M  <NA>
S042956                              Jos de la Cruz Guzman   M    29
S042973               Jos Julin Regino de la Fuente Vzquez   M    32
S042974               Jos Julin Regino de la Fuente Vzquez   M    32
S043002                     Jos Kelvin de la Nieve Linares   M    21
S043003                     Jos Kelvin de la Nieve Linares   M    25
S043054                  Jos Guadalupe de la Torre Gonzlez   M    35
S043122                Jos Maria Emirto de Lima y Sintiago   M    42
S043136                    Jos Artur Pratas Rebelo de Lima   M    20
S043137                    Jos Artur Pratas Rebelo de Lima   M    20
S043138                                 Jos Carlos de Lima   M    25
S043139                                 Jos Carlos de Lima   M    25
S057160              Jos Miguel Fernndez de Liencres Flrez   M    21
S057161              Jos Miguel Fernndez de Liencres Flrez   M    21
S063508                          Jos Luis Garca de la Cruz   M  <NA>
S063509                          Jos Luis Garca de la Cruz   M  <NA>
S068250                     Jos Antonio Gonzlez de Linares   M    22
S145336                           Jos Oliveros de la Torre   M    20
S145337                           Jos Oliveros de la Torre   M    24
S145338                           Jos Oliveros de la Torre   M    24
S206218                    Jos Julin Velsquez de la Cuesta   M    22
S206219                    Jos Julin Velsquez de la Cuesta   M    22
S206220                    Jos Julin Velsquez de la Cuesta   M    22
S206221                    Jos Julin Velsquez de la Cuesta   M    26

              Team  NOC        Games  Year  Season         City  \
index
S004941      Spain  ESP  1924 Summer  1924  Summer        Paris
S004942      Spain  ESP  1924 Summer  1924  Summer        Paris
S004943      Spain  ESP  1928 Summer  1928  Summer    Amsterdam
S004944      Spain  ESP  1928 Summer  1928  Summer    Amsterdam
S015860      Italy  ITA  2012 Summer  2012  Summer       London
S027705     Mexico  MEX  1948 Summer  1948  Summer       London
```

```
S027706     Mexico  MEX  1952 Summer  1952  Summer     Helsinki
S042956  Venezuela  VEN  1992 Summer  1992  Summer    Barcelona
S042973    Uruguay  URU  1936 Summer  1936  Summer       Berlin
S042974    Uruguay  URU  1936 Summer  1936  Summer       Berlin
S043002      Spain  ESP  2008 Summer  2008  Summer      Beijing
S043003      Spain  ESP  2012 Summer  2012  Summer       London
S043054     Mexico  MEX  1948 Summer  1948  Summer       London
S043122   Colombia  COL  1932 Summer  1932  Summer  Los Angeles
S043136   Portugal  POR  1928 Summer  1928  Summer    Amsterdam
S043137   Portugal  POR  1928 Summer  1928  Summer    Amsterdam
S043138     Brazil  BRA  1980 Summer  1980  Summer       Moskva
S043139     Brazil  BRA  1980 Summer  1980  Summer       Moskva
S057160      Spain  ESP  1920 Summer  1920  Summer    Antwerpen
S057161    Spain-1  ESP  1920 Summer  1920  Summer    Antwerpen
S063508  Guatemala  GUA  1968 Summer  1968  Summer  Mexico City
S063509  Guatemala  GUA  1968 Summer  1968  Summer  Mexico City
S068250      Spain  ESP  1968 Summer  1968  Summer  Mexico City
S145336     Mexico  MEX  1968 Summer  1968  Summer  Mexico City
S145337     Mexico  MEX  1972 Summer  1972  Summer       Munich
S145338     Mexico  MEX  1972 Summer  1972  Summer       Munich
S206218   Colombia  COL  1992 Summer  1992  Summer    Barcelona
S206219   Colombia  COL  1992 Summer  1992  Summer    Barcelona
S206220   Colombia  COL  1992 Summer  1992  Summer    Barcelona
S206221   Colombia  COL  1996 Summer  1996  Summer      Atlanta

                    Sport                                              Event  \
index
S004941     Equestrianism            Equestrianism Men's Jumping, Individual
S004942     Equestrianism                  Equestrianism Men's Jumping, Team
S004943     Equestrianism            Equestrianism Men's Jumping, Individual
S004944     Equestrianism                  Equestrianism Men's Jumping, Team
S015860         Athletics                 Athletics Men's 400 metres Hurdles
S027705        Basketball                        Basketball Men's Basketball
S027706        Basketball                        Basketball Men's Basketball
S042956            Boxing                          Boxing Men's Welterweight
S042973           Fencing                    Fencing Men's Sabre, Individual
S042974           Fencing                          Fencing Men's Sabre, Team
S043002            Boxing                       Boxing Men's Light-Flyweight
S043003            Boxing                       Boxing Men's Light-Flyweight
S043054          Shooting  Shooting Men's Small-Bore Rifle, Prone, 50 metres
S043122  Art Competitions                       Art Competitions Mixed Music
S043136         Athletics                         Athletics Men's 100 metres
S043137         Athletics                         Athletics Men's 200 metres
S043138           Cycling                Cycling Men's Road Race, Individual
S043139           Cycling           Cycling Men's Team Pursuit, 4,000 metres
S057160            Tennis                               Tennis Men's Singles
S057161            Tennis                               Tennis Men's Doubles
```

```
S063508         Wrestling         Wrestling Men's Featherweight, Greco-Roman
S063509         Wrestling           Wrestling Men's Featherweight, Freestyle
S068250           Cycling       Cycling Men's 100 kilometres Team Time Trial
S145336         Athletics                 Athletics Men's 20 kilometres Walk
S145337         Athletics                 Athletics Men's 20 kilometres Walk
S145338         Athletics                 Athletics Men's 50 kilometres Walk
S206218           Cycling              Cycling Men's 1,000 metres Time Trial
S206219           Cycling           Cycling Men's Team Pursuit, 4,000 metres
S206220           Cycling                          Cycling Men's Points Race
S206221           Cycling           Cycling Men's Team Pursuit, 4,000 metres

        Medal FirstName                               LastName  \
index
S004941   NaN       Jos  DE LAS ASTURIAS BOHORQUES Y GOYENECHE
S004942   NaN       Jos  DE LAS ASTURIAS BOHORQUES Y GOYENECHE
S004943   NaN       Jos  DE LAS ASTURIAS BOHORQUES Y GOYENECHE
S004944  Gold       Jos  DE LAS ASTURIAS BOHORQUES Y GOYENECHE
S015860   NaN       Jos                                DE LEON
S027705   NaN       Jos             DE LA CRUZ CABRERA GANDERA
S027706   NaN       Jos             DE LA CRUZ CABRERA GANDERA
S042956   NaN       Jos                      DE LA CRUZ GUZMAN
S042973   NaN       Jos                    DE LA FUENTE VZQUEZ
S042974   NaN       Jos                    DE LA FUENTE VZQUEZ
S043002   NaN       Jos                    DE LA NIEVE LINARES
S043003   NaN       Jos                    DE LA NIEVE LINARES
S043054   NaN       Jos                    DE LA TORRE GONZLEZ
S043122   NaN       Jos                     DE LIMA Y SINTIAGO
S043136   NaN       Jos                                DE LIMA
S043137   NaN       Jos                                DE LIMA
S043138   NaN       Jos                                DE LIMA
S043139   NaN       Jos                                DE LIMA
S057160   NaN       Jos                      DE LIENCRES FLREZ
S057161   NaN       Jos                      DE LIENCRES FLREZ
S063508   NaN       Jos                             DE LA CRUZ
S063509   NaN       Jos                             DE LA CRUZ
S068250   NaN       Jos                             DE LINARES
S145336   NaN       Jos                            DE LA TORRE
S145337   NaN       Jos                            DE LA TORRE
S145338   NaN       Jos                            DE LA TORRE
S206218   NaN       Jos                           DE LA CUESTA
S206219   NaN       Jos                           DE LA CUESTA
S206220   NaN       Jos                           DE LA CUESTA
S206221   NaN       Jos                           DE LA CUESTA

                                         ShortName AthleteID
index
S004941  Jos DE LAS ASTURIAS BOHORQUES Y GOYENECHE  JOSDE L1
```

```
S004942  Jos DE LAS ASTURIAS BOHORQUES Y GOYENECHE  JOSDE L1
S004943  Jos DE LAS ASTURIAS BOHORQUES Y GOYENECHE  JOSDE L1
S004944  Jos DE LAS ASTURIAS BOHORQUES Y GOYENECHE  JOSDE L1
S015860                               Jos DE LEON  JOSDE L1
S027705             Jos DE LA CRUZ CABRERA GANDERA  JOSDE L1
S027706             Jos DE LA CRUZ CABRERA GANDERA  JOSDE L1
S042956                      Jos DE LA CRUZ GUZMAN  JOSDE L1
S042973                    Jos DE LA FUENTE VZQUEZ  JOSDE L1
S042974                    Jos DE LA FUENTE VZQUEZ  JOSDE L1
S043002                    Jos DE LA NIEVE LINARES  JOSDE L1
S043003                    Jos DE LA NIEVE LINARES  JOSDE L1
S043054                    Jos DE LA TORRE GONZLEZ  JOSDE L1
S043122                     Jos DE LIMA Y SINTIAGO  JOSDE L1
S043136                               Jos DE LIMA  JOSDE L1
S043137                               Jos DE LIMA  JOSDE L1
S043138                               Jos DE LIMA  JOSDE L1
S043139                               Jos DE LIMA  JOSDE L1
S057160                      Jos DE LIENCRES FLREZ  JOSDE L1
S057161                      Jos DE LIENCRES FLREZ  JOSDE L1
S063508                            Jos DE LA CRUZ  JOSDE L1
S063509                            Jos DE LA CRUZ  JOSDE L1
S068250                            Jos DE LINARES  JOSDE L1
S145336                           Jos DE LA TORRE  JOSDE L1
S145337                           Jos DE LA TORRE  JOSDE L1
S145338                           Jos DE LA TORRE  JOSDE L1
S206218                          Jos DE LA CUESTA  JOSDE L1
S206219                          Jos DE LA CUESTA  JOSDE L1
S206220                          Jos DE LA CUESTA  JOSDE L1
S206221                          Jos DE LA CUESTA  JOSDE L1
```

Les premiers caractères du nom ne permettent pas de discriminer les personnes, pareil pour le nom court complet.

solution : se tourner vers le hash d'info complètes tels que 1. le nom 2. l'année de naissance

ex : Aaron Cook a concourru pour la grande bretagne puis l'île de malte puis la Moldavie. Il a été d'abord britannique avant de devenir moldave

[16]:

```python
athlet[athlet["ShortName"] == "Aaron COOK"]
```

[16]:

```
                      Name Sex  Age           Team  NOC        Games  Year  \
index
S037074  Aaron Arthur Cook   M   17  Great Britain  GBR  2008 Summer  2008
S037075  Aaron Arthur Cook   M   25        Moldova  MDA  2016 Summer  2016

         Season            City     Sport                         Event  \
index
S037074  Summer         Beijing  Taekwondo  Taekwondo Men's Welterweight
```

```
S037075  Summer  Rio de Janeiro  Taekwondo  Taekwondo Men's Welterweight

        Medal   ShortName
index
S037074    NaN  Aaron COOK
S037075    NaN  Aaron COOK
```

```
[7]: athlet["lenShortName"] = athlet.ShortName.apply(lambda s: len(s.split(" ")))
     athlet[athlet["lenShortName"] > 2]
```

```
[7]:                                               Name Sex  Age         Team  \
index
S000591                 Abel Carlos da Silva Braga   M   19       Brazil
S000719                        Tarek Abou Al Dahab   M   28      Lebanon
S000720                        Tarek Abou Al Dahab   M   28      Lebanon
S000721                        Tarek Abou Al Dahab   M   28      Lebanon
S000722                        Tarek Abou Al Dahab   M   28      Lebanon
…                                              …  ..  …           …
W045462           Stefanie "Steffi" von Siebenthal   F   24  Switzerland
W045463                   Walter H. von Siebenthal   M   24  Switzerland
W045464                          Hermann von Valta   M   35    Germany-1
W045465                          Hermann von Valta   M   35    Germany-1
W047028  Clifton Hugh Lancelot de Verdon Wrottesley   M   33      Ireland

         NOC        Games  Year  Season                   City         Sport  \
index
S000591  BRA  1972 Summer  1972  Summer                 Munich      Football
S000719  LIB  1968 Summer  1968  Summer            Mexico City       Cycling
S000720  LIB  1968 Summer  1968  Summer            Mexico City       Cycling
S000721  LIB  1968 Summer  1968  Summer            Mexico City       Cycling
S000722  LIB  1968 Summer  1968  Summer            Mexico City       Cycling
…      …          …   …     …                     …           …
W045462  SUI  2002 Winter  2002  Winter         Salt Lake City  Snowboarding
W045463  SUI  1924 Winter  1924  Winter               Chamonix    Ice Hockey
W045464  GER  1936 Winter  1936  Winter  Garmisch-Partenkirchen     Bobsleigh
W045465  GER  1936 Winter  1936  Winter  Garmisch-Partenkirchen     Bobsleigh
W047028  IRL  2002 Winter  2002  Winter         Salt Lake City      Skeleton

                                                   Event Medal  \
index
S000591                          Football Men's Football   NaN
S000719              Cycling Men's Road Race, Individual   NaN
S000720                             Cycling Men's Sprint   NaN
S000721            Cycling Men's 1,000 metres Time Trial   NaN
S000722  Cycling Men's Individual Pursuit, 4,000 metres   NaN
…                                                  …   …
W045462       Snowboarding Women's Parallel Giant Slalom   NaN
```

```
W045463                     Ice Hockey Men's Ice Hockey   NaN
W045464                             Bobsleigh Men's Two   NaN
W045465                            Bobsleigh Men's Four   NaN
W047028                         Skeleton Men's Skeleton   NaN

                            ShortName  lenShortName
index
S000591           Abel DA SILVA BRAGA             4
S000719                Tarek AL DAHAB             3
S000720                Tarek AL DAHAB             3
S000721                Tarek AL DAHAB             3
S000722                Tarek AL DAHAB             3
…                               …           …
W045462       Stefanie VON SIEBENTHAL             3
W045463         Walter VON SIEBENTHAL             3
W045464             Hermann VON VALTA             3
W045465             Hermann VON VALTA             3
W047028  Clifton DE VERDON WROTTESLEY             4

[10345 rows x 14 columns]
```

2.5.1 Prénoms: Correction des caractères manquants

```
[7]: athlet["FirstName"].nunique()
```

```
[7]: 24816
```

```
[8]: athlet["FirstName"].value_counts()[:50]
```

```
[8]: FirstName
     John           2479
     Robert         1890
     Michael        1473
     Peter          1458
     David          1445
     William        1392
     Jos            1348
     Kim            1280
     Thomas         1163
     Paul           1160
     Charles        1083
     James          1082
     Jan             972
     Daniel          953
     Hans            944
     Aleksandr       886
     Juan            879
```

```
Anna            876
Richard         875
Karl            863
Lee             856
George          834
Joseph          823
Mohamed         808
Ivan            754
Martin          744
Jean            725
Vladimir        724
Josef           721
Carlos          711
Sergey          697
Maria           674
Walter          638
Li              623
Albert          604
Andreas         595
Jorge           590
Frank           585
Luis            581
Alfred          572
Carl            568
Christian       564
Andrew          562
Mara            553
Edward          544
Patrick         544
Olga            538
Andrea          528
Christopher     514
Roberto         509
Name: count, dtype: int64
```

Les caractères accentués semblent avoir été effacés dans la colone `Name` Exemple avec le prénom “Jos” qui est surement “José” Regardons le `NOC` des athlètes qui portent ce prénom

[9]:
```
athlet[(athlet["FirstName"] == "Jos")].NOC.value_counts()
```

[9]:
```
NOC
ESP    421
MEX    212
POR    120
BRA    103
CUB     66
ARG     64
```

```
PUR      58
VEN      50
COL      31
GUA      28
URU      20
DOM      18
CHI      18
PER      16
LUX      16
CRC      14
ESA      13
PHI       9
FRA       9
USA       8
BEL       7
NCA       6
ANG       5
BOL       5
PAN       5
PAR       4
ECU       4
HKG       3
HON       3
GEQ       2
COD       2
MOZ       1
MRI       1
ITA       1
TUN       1
AHO       1
BEN       1
ISV       1
NED       1
Name: count, dtype: int64
```

[11]:

```python
athlet[(athlet["FirstName"] == "Jos") & (athlet["NOC"].isin(["BEL", "NED"]))]
```

[11]:

```
                               Name Sex  Age         Team  NOC        Games  \
index
S044576                 Jos Delaval   M   27      Belgium  BEL  1948 Summer
S044577                 Jos Delaval   M   31      Belgium  BEL  1952 Summer
S050173          Jos Nicolas Dumont   M   20      Belgium  BEL  1960 Summer
S050174          Jos Nicolas Dumont   M   24      Belgium  BEL  1964 Summer
S151085  Jos Maria Aloysius Pauwels   M   24      Belgium  BEL  1952 Summer
S203615              Jos Van Baelen   M   33      Belgium  BEL  1960 Summer
S203616              Jos Van Baelen   M   33      Belgium  BEL  1960 Summer
S204997                 Jos van Veen   F   30  Netherlands  NED  2016 Summer
```

```
         Year  Season            City       Sport  \
index
S044576  1948  Summer          London      Hockey
S044577  1952  Summer        Helsinki      Hockey
S050173  1960  Summer            Roma  Water Polo
S050174  1964  Summer           Tokyo  Water Polo
S151085  1952  Summer        Helsinki     Cycling
S203615  1960  Summer            Roma     Fencing
S203616  1960  Summer            Roma     Fencing
S204997  2016  Summer  Rio de Janeiro      Rowing

                                             Event Medal FirstName    LastName  \
index
S044576                        Hockey Men's Hockey   NaN       Jos     DELAVAL
S044577                        Hockey Men's Hockey   NaN       Jos     DELAVAL
S050173                Water Polo Men's Water Polo   NaN       Jos      DUMONT
S050174                Water Polo Men's Water Polo   NaN       Jos      DUMONT
S151085  Cycling Men's Team Pursuit, 4,000 metres   NaN       Jos     PAUWELS
S203615            Fencing Men's Sabre, Individual   NaN       Jos  VAN BAELEN
S203616                  Fencing Men's Sabre, Team   NaN       Jos  VAN BAELEN
S204997                Rowing Women's Coxed Eights   NaN       Jos    VAN VEEN

              ShortName   YOB           AthleteID
index
S044576     Jos DELAVAL  1921     JOSDELAVAL11921
S044577     Jos DELAVAL  1921     JOSDELAVAL11921
S050173      Jos DUMONT  1940      JOSDUMONT11940
S050174      Jos DUMONT  1940      JOSDUMONT11940
S151085     Jos PAUWELS  1928     JOSPAUWELS11928
S203615  Jos VAN BAELEN  1927  JOSVAN BAELEN11927
S203616  Jos VAN BAELEN  1927  JOSVAN BAELEN11927
S204997    Jos VAN VEEN  1986    JOSVAN VEEN01986
```

2.6 Génération d'ID numérique

[11]:

```
import uuid
uuid.uuid4().hex
```

[11]:

```
'0a48cfcdf15f4834b6cccc54b0336a88'
```

[34]:

```
import hashlib
s = "JOSVANVEEN01986"

def genid(person_id):
    id = int(hashlib.sha1(person_id.encode("utf-8")).hexdigest(), 16) % (10**12)
    id = str(id).zfill(12)
    return id

genid(s)
```

```
[34]: '074208717632'
```

```
[35]: len("074208717632")
```

```
[35]: 12
```

```
[36]: person["NumAthleteID"] = person["AthleteID"].apply(genid)
      athlet["NumAthleteID"] = athlet["AthleteID"].apply(genid)
```

```
[37]: person
```

```
[37]:
                  AthleteID     FirstName  LastName   YOB  NumAthleteID
0          AADAMKHAMIS11988         Aadam    KHAMIS  1988  536241055175
1        AADJIJATMIKOH01972  Aadjijatmiko         H  1972  232521346500
3       AADOLFSVANSTRM11886        Aadolf  SVANSTRM  1886  751540828727
5          AAFKEHAMENT01970         Aafke    HAMENT  1970  334896867726
6         AAGEANDERSEN11884          Aage  ANDERSEN  1884  803222010350
…                     …           …       …   …           …
286228    ZYGMUNTWEISS11902       Zygmunt     WEISS  1902  510630496403
286232       ZYTAJARKA01962          Zyta     JARKA  1962  021099040718
286233         ZZETNCE11980          Zzet       NCE  1980  558742942791
286235       ZZETSAFER11990          Zzet     SAFER  1990  084653996587
286236   ZZIMOCALAZANS11932         Zzimo  CALAZANS  1932  000838332997

[147426 rows x 5 columns]
```

```
[38]: athlet[athlet.NumAthleteID == "021099040718"]
```

```
[38]:
               Name Sex  Age    Team  NOC        Games  Year  Season   City  \
index
S088047  Zyta Jarka   F   26  Poland  POL  1988 Summer  1988  Summer  Seoul

          Sport                       Event Medal FirstName LastName  \
index
S088047  Rowing  Rowing Women's Coxed Fours   NaN      Zyta    JARKA

          ShortName  BirthYear   YOB       AthleteID  g  NumAthleteID
index
S088047  Zyta JARKA       1962  1962  ZYTAJARKA01962  0  021099040718
```

```
[39]: athlet.NumAthleteID.nunique()
```

```
[39]: 147426
```

```
[40]: person.NumAthleteID.nunique()
```

```
[40]: 147426
```

```
[41]: person.shape[0]
```

```
[41]: 147426
```

```
[5]: person.FirstName.value_counts()
```

```
[5]: FirstName
     John          1380
     Robert         975
     William        818
     Jos            805
     David          799
                   …
     Jennieffer       1
     Jennet           1
     Jenne            1
     Jenly            1
     Zzimo            1
     Name: count, Length: 24816, dtype: int64
```

```
[6]: person[person.FirstName == "Jos"]
```

```
[6]:
                         AthleteID FirstName              LastName   YOB
     132172          JOSABAITA11903       Jos                ABAITA  1903
     132173          JOSABANDO11946       Jos                ABANDO  1946
     132175        JOSABRANTES11970       Jos              ABRANTES  1970
     132176           JOSABREU11948       Jos                 ABREU  1948
     132183         JOSACCIBAR11945       Jos               ACCIBAR  1945
     …                           …       …                     …     …
     135533        JOSZARIQUEY11949       Jos              ZARIQUEY  1949
     135534           JOSZAYAS11950       Jos                 ZAYAS  1950
     135535         JOSZELEDON11946       Jos               ZELEDON  1946
     135536            JOSZIGA11937       Jos                  ZIGA  1937
     135538  JOSZORRILLA Y SCHNEIDER11913       Jos  ZORRILLA Y SCHNEIDER  1913

     [805 rows x 4 columns]
```

```
[13]: from Levenshtein import distance as lev
```

```
[14]: lev("José", "Jos")
```

```
[14]: 1
```

2.7 Year

[17]:

```
athlet["Year"] = athlet.Year.astype("Int64")
athlet["Year"].sample(n=10)
```

[17]:

```
index
S133381    1972
S187497    2000
S223597    2020
W027020    1968
W010320    1976
S202964    1956
S113226    1980
S096882    1968
S081543    2000
S102519    1996
Name: Year, dtype: Int64
```

[18]:

```
athlet.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 286237 entries, S000001 to W048564
Data columns (total 12 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   Name    286237 non-null  object
 1   Sex     286237 non-null  object
 2   Age     276763 non-null  Int64
 3   Team    286237 non-null  object
 4   NOC     286237 non-null  object
 5   Games   286237 non-null  object
 6   Year    286237 non-null  Int64
 7   Season  286237 non-null  object
 8   City    286237 non-null  object
 9   Sport   286237 non-null  object
 10  Event   286237 non-null  object
 11  Medal   42232 non-null   object
dtypes: Int64(2), object(10)
memory usage: 28.9+ MB
```

2.8 Médailles

[19]:

```
compet = (pd.DataFrame(athlet.groupby(["Games", "Event"]).Medal.count())
          .reset_index())
compet.sort_values(by="Medal", ascending=False)
```

[19]:

```
            Games                                                Event  Medal
6228  2020 Summer                                             Men Team    335
```

```
6381  2020 Summer                                           Women Team    334
354   1908 Summer                   Gymnastics Men's Team All-Around     94
599   1920 Summer                   Gymnastics Men's Team All-Around     77
466   1912 Summer   Gymnastics Men's Team All-Around, Swedish System     74
…          …                                                       …    …
957   1928 Winter                    Speed Skating Men's 10,000 metres      0
1285  1948 Summer     Art Competitions Mixed Painting, Unknown Event      0
1289  1948 Summer   Art Competitions Mixed Sculpturing, Unknown Event      0
6466  2020 Summer                                     Women's Madison      0
1125  1936 Summer   Art Competitions Mixed Sculpturing, Unknown Event      0

[6498 rows x 3 columns]
```

```
[20]: compet.to_excel("../../draft/compet.ods")
```

```
[21]: compet.Medal.describe()
```

```
[21]: count    6498.000000
      mean        6.499231
      std        10.449955
      min         0.000000
      25%         3.000000
      50%         3.000000
      75%         6.000000
      max       335.000000
      Name: Medal, dtype: float64
```

```
[22]: athlet["Medal"].fillna(pd.NA, inplace=True)
```

```
[23]: athlet["Medal"] = athlet["Medal"].fillna("None")
      athlet.Medal.value_counts()
```

```
[23]: Medal
      None      244005
      Gold       14172
      Bronze     14162
      Silver     13898
      Name: count, dtype: int64
```

```
[24]: athlet["Medal"] = athlet.Medal.astype("category")
      athlet.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 286237 entries, S000001 to W048564
Data columns (total 12 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
```

```
 0   Name    286237 non-null  object
 1   Sex     286237 non-null  object
 2   Age     276763 non-null  Int64
 3   Team    286237 non-null  object
 4   NOC     286237 non-null  object
 5   Games   286237 non-null  object
 6   Year    286237 non-null  Int64
 7   Season  286237 non-null  object
 8   City    286237 non-null  object
 9   Sport   286237 non-null  object
 10  Event   286237 non-null  object
 11  Medal   286237 non-null  category
dtypes: Int64(2), category(1), object(9)
memory usage: 27.0+ MB
```

[25]:
```
athlet
```

[25]:
```
                                       Name Sex  Age            Team  NOC  \
index
S000001                           A Dijiang   M   24           China  CHN
S000002                            A Lamusi   M   23           China  CHN
S000003                 Gunnar Nielsen Aaby   M   24         Denmark  DEN
S000004                Edgar Lindenau Aabye   M   34  Denmark/Sweden  DEN
S000005  Cornelia "Cor" Aalten (-Strannood)   F   18     Netherlands  NED
…                                      …  . .  …             …  …
W048560                          Andrzej ya   M   29        Poland-1  POL
W048561                            Piotr ya   M   27          Poland  POL
W048562                            Piotr ya   M   27          Poland  POL
W048563                  Tomasz Ireneusz ya   M   30          Poland  POL
W048564                  Tomasz Ireneusz ya   M   34          Poland  POL

               Games  Year  Season            City        Sport  \
index
S000001  1992 Summer  1992  Summer       Barcelona   Basketball
S000002  2012 Summer  2012  Summer          London         Judo
S000003  1920 Summer  1920  Summer       Antwerpen     Football
S000004  1900 Summer  1900  Summer           Paris   Tug-Of-War
S000005  1932 Summer  1932  Summer     Los Angeles    Athletics
…              …   …     …             …          …
W048560  1976 Winter  1976  Winter       Innsbruck         Luge
W048561  2014 Winter  2014  Winter           Sochi  Ski Jumping
W048562  2014 Winter  2014  Winter           Sochi  Ski Jumping
W048563  1998 Winter  1998  Winter          Nagano    Bobsleigh
W048564  2002 Winter  2002  Winter  Salt Lake City    Bobsleigh

                                            Event Medal
index
```

```
S000001               Basketball Men's Basketball  None
S000002              Judo Men's Extra-Lightweight  None
S000003                   Football Men's Football  None
S000004               Tug-Of-War Men's Tug-Of-War  Gold
S000005              Athletics Women's 100 metres  None
…                                            …   …
W048560                Luge Mixed (Men)'s Doubles  None
W048561  Ski Jumping Men's Large Hill, Individual  None
W048562        Ski Jumping Men's Large Hill, Team  None
W048563                      Bobsleigh Men's Four  None
W048564                      Bobsleigh Men's Four  None

[286237 rows x 12 columns]
```

3 Region

[27]:
```
assert region.shape[0] == region.NOC.nunique()
```

[28]:
```
region
```

[28]:
```
     NOC       region        notes
0    EOR      Refugee          NaN
1    LBN      Lebanon          NaN
2    SGP    Singapore          NaN
3    ROC       Russia          NaN
4    AFG  Afghanistan          NaN
..   …           …          …
229  YEM        Yemen          NaN
230  YMD        Yemen  South Yemen
231  YUG       Serbia   Yugoslavia
232  ZAM       Zambia          NaN
233  ZIM     Zimbabwe          NaN

[234 rows x 3 columns]
```

[29]:
```
region.loc[region["NOC"] == "ROT", "region"] = "Refugee"
region.loc[region["NOC"] == "TUV", "region"] = "Tuvalu"
region.loc[region["NOC"] == "UNK", "region"] = "Unknown"
```

[30]:
```
region.columns = map(str.lower, region.columns)
region
```

[30]:
```
     noc       region        notes
0    EOR      Refugee          NaN
1    LBN      Lebanon          NaN
2    SGP    Singapore          NaN
```

```
3    ROC       Russia          NaN
4    AFG  Afghanistan          NaN
..   …           …          …
229  YEM        Yemen          NaN
230  YMD        Yemen  South Yemen
231  YUG       Serbia   Yugoslavia
232  ZAM       Zambia          NaN
233  ZIM     Zimbabwe          NaN

[234 rows x 3 columns]
```

[31]:

```
region[~region["notes"].isna()]
```

[31]:

```
     noc                      region                        notes
5    AHO                     Curacao         Netherlands Antilles
10   ANT                     Antigua          Antigua and Barbuda
11   ANZ                   Australia                  Australasia
30   BOH              Czech Republic                      Bohemia
55   CRT                      Greece                        Crete
92   HKG                       China                    Hong Kong
97   IOA  Individual Olympic Athletes  Individual Olympic Athletes
103  ISV            Virgin Islands, US               Virgin Islands
147  NBO                    Malaysia                 North Borneo
151  NFL                      Canada                 Newfoundland
172  ROT                     Refugee          Refugee Olympic Team
179  SCG                      Serbia        Serbia and Montenegro
183  SKN                 Saint Kitts     Turks and Caicos Islands
209  TTO                    Trinidad          Trinidad and Tobago
212  TUV                      Tuvalu                       Tuvalu
214  UAR                       Syria         United Arab Republic
217  UNK                     Unknown                      Unknown
227  WIF                    Trinidad       West Indies Federation
228  YAR                       Yemen                  North Yemen
230  YMD                       Yemen                  South Yemen
231  YUG                      Serbia                   Yugoslavia
```

[32]:

```
region[region["region"] == "Tuvalu"]
```

[32]:

```
     noc  region   notes
212  TUV  Tuvalu  Tuvalu
```

[33]:

```
for reg in list(region["region"].unique())[:25]:
    print(reg)
```

```
Refugee
Lebanon
Singapore
```

```
Russia
Afghanistan
Curacao
Albania
Algeria
Andorra
Angola
Antigua
Australia
Argentina
Armenia
Aruba
American Samoa
Austria
Azerbaijan
Bahamas
Bangladesh
Barbados
Burundi
Belgium
Benin
Bermuda
```

4 Bases nettoyées

[11]:

```
print(result.shape)
```

```
(286237, 6)
```