In [ ]:

```
"""
Date: 10th December, 2020

The project is about hypothesis "The winners of the best running races in the over the wo
rld has been won by African athletes".

It's going to show with two datasets: first one is about the winners in 120 years of Olym
pic Games in the Sport's History, and the second one is about the six best world marathon
majors.
"""
```

In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns #visualisation
import matplotlib.pyplot as plt #visualisation
```

In [2]:

```
import os.path
print(os.path)
#/Users/ariadnapuigventos/Documents/CURSOS/BRIDGE/DS_Ejercicios_Python/BootCamp_TheBridge
/Proyecto_Navidad_Ariadna/src/utils/folders_tb.py
```

```
<module 'posixpath' from '/Library/Frameworks/Python.framework/Versions/3.9/lib/python3.9
/posixpath.py'>
```

Explain the code organization of this file:

It's going to tell about one of two datasets and show the collecting data to understand our hypothesis. Below all these lines, it will show the second datasets with the best insights of Olympic Games Athletes. Finally, It's going to create a new dataframe to try show some similarities to confirm or not hypothesis.

In [3]:

```
from utils.folders_tb import readcsv
#This is one of two dataframes about Best Marathon Majors in all Sport History.
readcsv()
```

```
     year           winner  gender        country      time marathon
0    2014   Dennis Kimetto    Male          Kenya  02:02:57   Berlin
1    2011   Geoffrey Mutai    Male          Kenya  02:03:02   Boston
2    2016  Kenenisa Bekele    Male       Ethiopia  02:03:03   Berlin
3    2016   Eliud Kipchoge    Male          Kenya  02:03:05   London
4    2013   Wilson Kipsang    Male          Kenya  02:03:23   Berlin
..    ...              ...     ...            ...       ...      ...
531  1966       Bobbi Gibb  Female  United States  03:21:40   Boston
532  1974  Jutta von Haase  Female        Germany  03:22:01   Berlin
533  1969  Sara Mae Berman  Female  United States  03:22:46   Boston
534  1967       Bobbi Gibb  Female  United States  03:27:17   Boston
535  1968       Bobbi Gibb  Female  United States  03:30:00   Boston

[536 rows x 6 columns]
```

In [4]:

```
from utils.mining_data_tb import topandtail, dimention

topandtail()
```

```
   year              winner gender   country      time marathon
0  2014      Dennis Kimetto   Male     Kenya  02:02:57   Berlin
1  2011      Geoffrey Mutai   Male     Kenya  02:03:02   Boston
2  2016     Kenenisa Bekele   Male  Ethiopia  02:03:03   Berlin
3  2016      Eliud Kipchoge   Male     Kenya  02:03:05   London
4  2013      Wilson Kipsang   Male     Kenya  02:03:23   Berlin
```

```
4  2015      Wilson Kipsang    Male      Kenya  02:05:25   Berlin
5  2017       Eliud Kipchoge   Male      Kenya  02:03:32   Berlin
6  2011     Patrick Musyoki    Male      Kenya  02:03:38   Berlin
7  2013       Dennis Kimetto   Male      Kenya  02:03:45  Chicago
8  2017       Wilson Kipsang   Male      Kenya  02:03:58    Tokyo
9  2008  Haile Gebrselassie    Male   Ethiopia  02:03:59   Berlin
......
     year           winner  gender        country      time marathon
531  1966       Bobbi Gibb  Female  United States  03:21:40   Boston
532  1974  Jutta von Haase  Female        Germany  03:22:01   Berlin
533  1969  Sara Mae Berman  Female  United States  03:22:46   Boston
534  1967       Bobbi Gibb  Female  United States  03:27:17   Boston
535  1968       Bobbi Gibb  Female  United States  03:30:00   Boston
```

In [5]:

```
dimention()
```

```
(536, 6)
number of duplicate rows:  Empty DataFrame
Columns: [year, winner, gender, country, time, marathon]
Index: []
     year           winner  gender        country      time marathon
0    2014   Dennis Kimetto    Male          Kenya  02:02:57   Berlin
1    2011   Geoffrey Mutai    Male          Kenya  02:03:02   Boston
2    2016  Kenenisa Bekele    Male       Ethiopia  02:03:03   Berlin
3    2016   Eliud Kipchoge    Male          Kenya  02:03:05   London
4    2013   Wilson Kipsang    Male          Kenya  02:03:23   Berlin
..    ...              ...     ...            ...       ...      ...
531  1966       Bobbi Gibb  Female  United States  03:21:40   Boston
532  1974  Jutta von Haase  Female        Germany  03:22:01   Berlin
533  1969  Sara Mae Berman  Female  United States  03:22:46   Boston
534  1967       Bobbi Gibb  Female  United States  03:27:17   Boston
535  1968       Bobbi Gibb  Female  United States  03:30:00   Boston

[536 rows x 6 columns]
(536, 6)
```

ALERT INFO (STEPS)

The Dataframe has not any duplicates but there are some values equality. It needs to check what it means because it's possible some majors who has already won more than one marathons, that's why it's going to show using the method values_counts by country and winner.

In [6]:

```
from utils.mining_data_tb import repite_pais, repetidores

# Effectively, the most country that's repeat is Kenya on the top, below United States an
d Ethiopia.
# It's considering that Kenya started to compete in 1960 and United States since 1896.

repite_pais()
```

```
Kenya             136
United States     104
Ethiopia           51
Germany            36
United Kingdom     35
Japan              22
Norway             20
Canada             17
Portugal           11
Mexico             10
Finland            10
Poland              8
Russia              8
Brazil              7
Italy               6
Name: country, dtype: int64
```

In [7]:

```
repetidores()
```

```
Grete Waitz            11
Bill Rodgers            8
Ingrid Kristiansen      8
Uta Pippig              7
Clarence DeMar          7
Paula Radcliffe         7
Mary Keitany            6
Rosa Mota               6
Eliud Kipchoge          6
Catherine Ndereba       6
Wilson Kipsang          5
Khalid Khannouchi       5
Joyce Chepchumba        5
Martin Lel              5
Steve Jones             4
Name: winner, dtype: int64
```

In [8]:

```
from utils.mining_data_tb import checkingdata
#It wants to know how are the values because it has seen that there is one about time.
checkingdata()
```

```
year          int64
winner       object
gender       object
country      object
time         object
marathon     object
dtype: object
```

It needs to change some data rows after to see time column in dataframe is an object dtype and it's necessary to be a timedelta to do a sum() in a future:

In [9]:

```
from utils.mining_data_tb import changetype
#With this fuction it changed from object time column with seconds to use it in boxplot f
or detecting outliers.
changetype()
```

```
0      0 days 02:02:57
1      0 days 02:03:02
2      0 days 02:03:03
3      0 days 02:03:05
4      0 days 02:03:23
              ...
531    0 days 03:21:40
532    0 days 03:22:01
533    0 days 03:22:46
534    0 days 03:27:17
535    0 days 03:30:00
Name: time, Length: 536, dtype: timedelta64[ns]
0       7377.0
1       7382.0
2       7383.0
3       7385.0
4       7403.0
         ...
531    12100.0
532    12121.0
533    12166.0
534    12437.0
535    12600.0
Name: time, Length: 536, dtype: float64
```

In [4]:

```
from utils.visualization_tb import detect_outliers

#2 extrems: the first time was 2:02:57 by Kenian Athlete in Berlin Marathon in 2014; and the last time was 3:30:00 by United States Athlete in Boston Marathon in 1968. Althought, 25% Marathon majors got a median around 2 hours and 16 minuts and the most majors with 75 % got 2 hours and 46 minuts.

detect_outliers()
```

```
0      0 days 02:02:57
1      0 days 02:03:02
2      0 days 02:03:03
3      0 days 02:03:05
4      0 days 02:03:23
              ...
531    0 days 03:21:40
532    0 days 03:22:01
533    0 days 03:22:46
534    0 days 03:27:17
535    0 days 03:30:00
Name: time, Length: 536, dtype: timedelta64[ns]
AxesSubplot(0.125,0.125;0.775x0.755)
7783.0
8856.25
1073.25
```

In [ ]:

```
```