SportsStats

"**Client 3: SportsStats (Olympics Dataset - 120 years of data)**: SportsStats is a sports analysis firm partnering with local news and elite personal trainers to provide “interesting” insights to help their partners. Insights could be patterns/trends highlighting certain groups/events/countries, etc. for the purpose of developing a news story or discovering key health insights."

I choose this dataset because I want to work with tabular data instead of text mining in the future (maybe it's change, I don't know, but It's the focus right now). I think taht this dataset will bring me more experience about this topic.

This dataset offer 2 csv files:

- athlete_events: This file bring the information about all atheletes that participate of Olympics, with the columns:
  - ID;
  - Name;
  - Sex;
  - Age;
  - Height;
  - Weight;
  - Team;
  - NOC;
  - Games;
  - Year;
  - Season;
  - City;
  - Sport;
  - Event;
  - Medal.
- noc_regions: This file have the data about national Olympics Comitte Regions (learn more: https://en.wikipedia.org/wiki/National_Olympic_Committee (https://en.wikipedia.org/wiki/National_Olympic_Committee)). The columns are:
  - NOC
  - region
  - notes

At first I need to import all python packages and then read the csv files.

In [1]:

```
import pandas as pd
```

In [2]:

```
athdf = pd.read_csv("./athlete_events.csv")
regdf = pd.read_csv("./noc_regions.csv")
```

And now I performed a initial exploration of data. I checked the first line and the types of each column.

In [3]:

```
athdf.head(1)
```

Out[3]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	E

In [4]:

```
athdf.dtypes
```

Out[4]:

```
ID          int64
Name       object
Sex        object
Age       float64
Height    float64
Weight    float64
Team       object
NOC        object
Games      object
Year        int64
Season     object
City       object
Sport      object
Event      object
Medal      object
dtype: object
```

In [5]:

```
regdf.head(1)
```

Out[5]:

	NOC	region	notes
0	AFG	Afghanistan	NaN

In [6]:

```
regdf.dtypes
```

Out[6]:

```
NOC       object
region    object
notes     object
dtype: object
```

Here we can note that two DataFrames have the a commum column: NOC. I will try to merge the DFs using it but first, let's check the number of Null values.

In [7]:

```
athdf.isna().sum()
```

Out[7]:

```
ID             0
Name           0
Sex            0
Age         9474
Height     60171
Weight     62875
Team           0
NOC            0
Games          0
Year           0
Season         0
City           0
Sport          0
Event          0
Medal     231333
dtype: int64
```

Many data for age, height and weight is missing. About the Medal collumn, we need to check if NA means no medal to the athlete.

In [8]:

```
regdf.isna().sum()
```

Out[8]:

```
NOC         0
region      3
notes     209
dtype: int64
```

In both Dataframes, the common collumn (NOC) have no NA values. Here is the ERD:

Description

This project has data about athletes that competed at the Olympics, with physical information about each one, country, the season that they competed, which sport, and how many medals they have.

Maybe the people who could be interested in this dataset are coaches, Federations of each sport, and Companies that want to sponsor some athlete.

Questions

- What is the country that has more medals?
- Which the sex have more amount of medals?
- A higher or a lower BMI interferes in amount of medals?

Hypothesis

- Older athletes have more medals than younger ones;
- Countries with more athletes have more medals.

Approach

For this study, I will use a lot of medal columns. With this, I will use weight and height to find the BMI. Age and sex to answer some questions and country to prove my hypothesis.

My data haven't a direct relation, but using a little of experience with another datas and some readings, I will try to relate them (like using BMI, aggregations and etc).

My metrics are defined inside the questions (50/50, and the best and big results).