Milestone 1: Project Proposal and Data Selection/Preparation

Eduardo Félix – 2022/08

Preparing Proposal

1. Which Client/Dataset did you select and why?
    a. I selected the 'Sports Stats' client because when I was a child I had a chance to compete at Military Olympic Game.

2. Describe the steps you took to import the data.
    a. I imported the data on 'Jupyter Notebook' with pandas library - **pd.read_csv ()**.

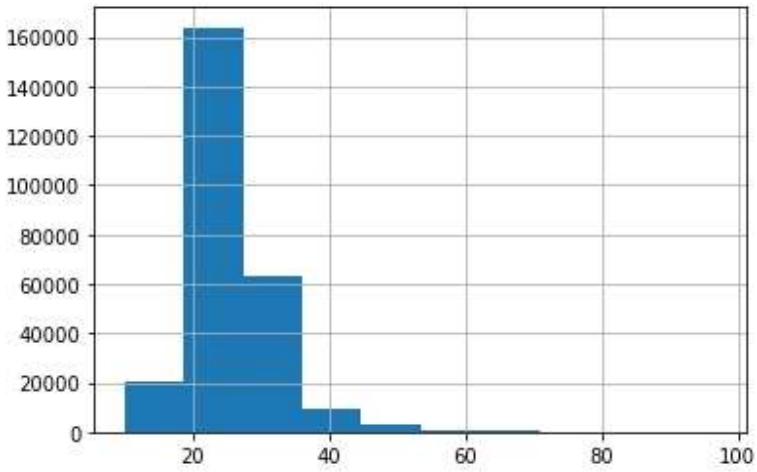
3. Perform Initial Exploration of data and provide some screenshots or display some stats of the data you are looking at.
    a. This image shows the older and younger athlete against Olympic Games between 1896-2016.

```
In [4]: print('Age: ', athleteEvents.Age.max())
        print('Age: ', athleteEvents.Age.min())

        Age:  97.0
        Age:  10.0
```

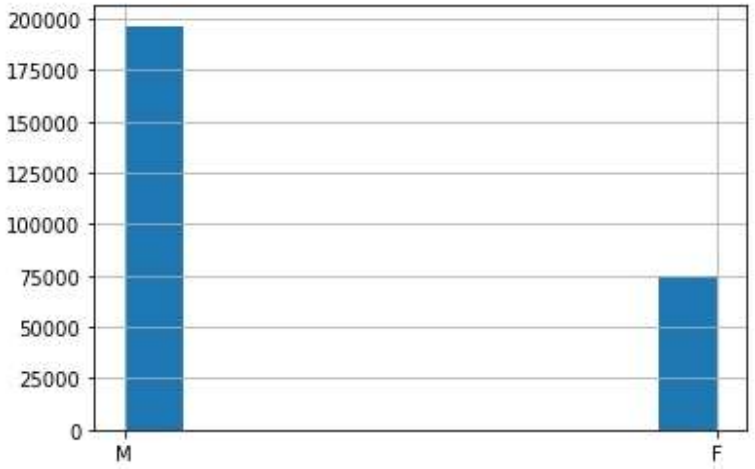
    b. This image shows a graph from the amount of athletes grouped by age.

```
In [7]: athleteEvents.Age.hist()
Out[7]: <AxesSubplot:>
```



c. This image shows a graph from the amount of athletes grouped by sex.

```
In [8]: athleteEvents.Sex.hist()
Out[8]: <AxesSubplot:>
```



d. This image shows the use of **pandassql** library and this module **sqldf**, with this tool we can provide SQL queries against our dataset.

```
In [10]: from pandasql import sqldf
pysqldf = lambda q: sqldf(q, globals())

In [14]: pysqldf('SELECT Count(ID) as "Total of Athletes",Team FROM athleteEvents WHERE Year = "2016" GROUP BY Team ORDER BY
```

Out[14]:

	Total of Athletes	Team
0	1	Equatorial Guinea
1	1	Tuvalu
2	2	Argentina-1
3	2	Argentina-2
4	2	Austria-1
...	...	...
244	504	France
245	510	Australia
246	528	Germany
247	571	Brazil
248	699	United States

249 rows × 2 columns

e. This image shows the amount of athletes from Brazil team at 2016 Olympic Games which happened in Rio de Janeiro Grouped by Sex and Age (It´s only showing the first 10 rows).

```
In [15]: OM athleteEvents WHERE Team = "Brazil" AND Year = 2016 GROUP BY Sex, Age ORDER BY "Total of Athlete" DESC LIMIT 10')
```

Out[15]:

	Total of Athlete	Sex	Age
0	36	M	22.0
1	31	M	26.0
2	29	F	25.0
3	29	M	24.0
4	24	F	29.0
5	24	M	23.0
6	22	M	27.0
7	19	F	27.0
8	19	F	28.0
9	19	M	25.0

4. Create an ERD or proposed ERD to show the relationships of the data you are exploring.

Develop Project Proposal

Description

a. The main project target is to analyze sport pattern in the Olympics. This analysis can be used for anyone who has interest in Olympic Games, who wants to know more about the topic and for athletes. It can also be useful to understand the common characteristics of the athletes and countries that compete in them.

Questions

1. What are the ages of the Olympic athletes?
    a. How is the age distribution against athletes in Olympic Games?

2. What are the gender of the Olympic athletes?
    a. Is the amount of male and female Olympic Athletes similar?

3. What are the country that has more athletes in Rio de Janeiro Olympic Games?

Hypothesis

1. There has been a steady increase of sports included to the Olympic Games.

2. I believe the ages will be mostly between 20-35 years old and the participation will be roughly the same.
    a. Young people are the peak of their physical abilities; therefore, they are the most likely to participate, weather they are male or female

3. I think the country that has more athletes on the Olympics is the country that has the highest number of overall medals.
    b. This is because the more athletes a country sends to the Olympics, the more likely the country supports sports and therefore the training and conditions of their training are better.

Approach

1. I will be looking mainly at the metrics of:
    a. The country Participation and
    b. Retrieving the information of age and gender of the Olympic Athletes.