horizontal line

Sportstats Capstone Project

# PREPARING PROPOSAL

## CLIENT/DATASET

I selected this dataset from SportsStats, a sports analysis firm collaborating with local news and elite personal trainers. The goal is to extract valuable insights and determine which physical characteristics enhance the chances of winning a medal. By analyzing the data, we can uncover patterns and trends related to specific groups, events, countries, and more. These insights can be used to develop news stories or discover important health insights

## DATA IMPORT

I imported the data into a pandas DataFrame in a Jupyter Notebook. The data consisted of two csv files - athlete\_events.csv and noc\_regions.csv. I noticed that there were null values in each column. To ensure data integrity, I decided to remove these null values from these columns.

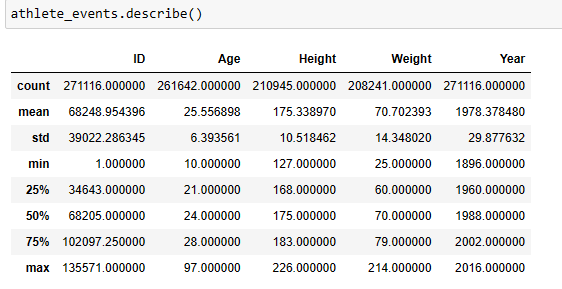
However, I chose not to remove the null values in the "Medals" column as they were necessary for determining whether an athlete had won a medal or not. Similarly, I retained the null values in the "Region" column of the NOC table, as they were required for further analysis.

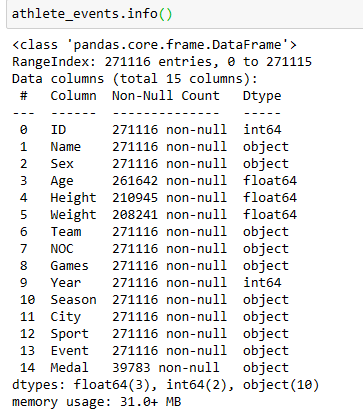
The null values in the "Notes" column were also preserved, as they contained additional information that was deemed important.

In summary, I removed the null values from the "Age", "Height", and "Weight" columns, while retaining the null values in the "Medals" and "Notes" columns, as well as the "Region" column in the NOC table.

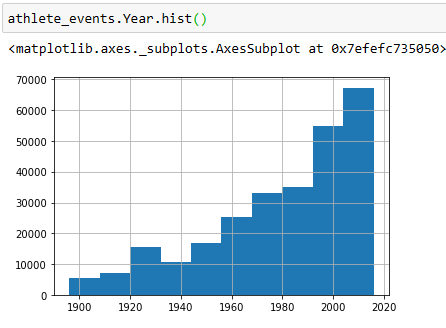
## INITIAL DATA EXPLORATION

Basic Information of the dataset:

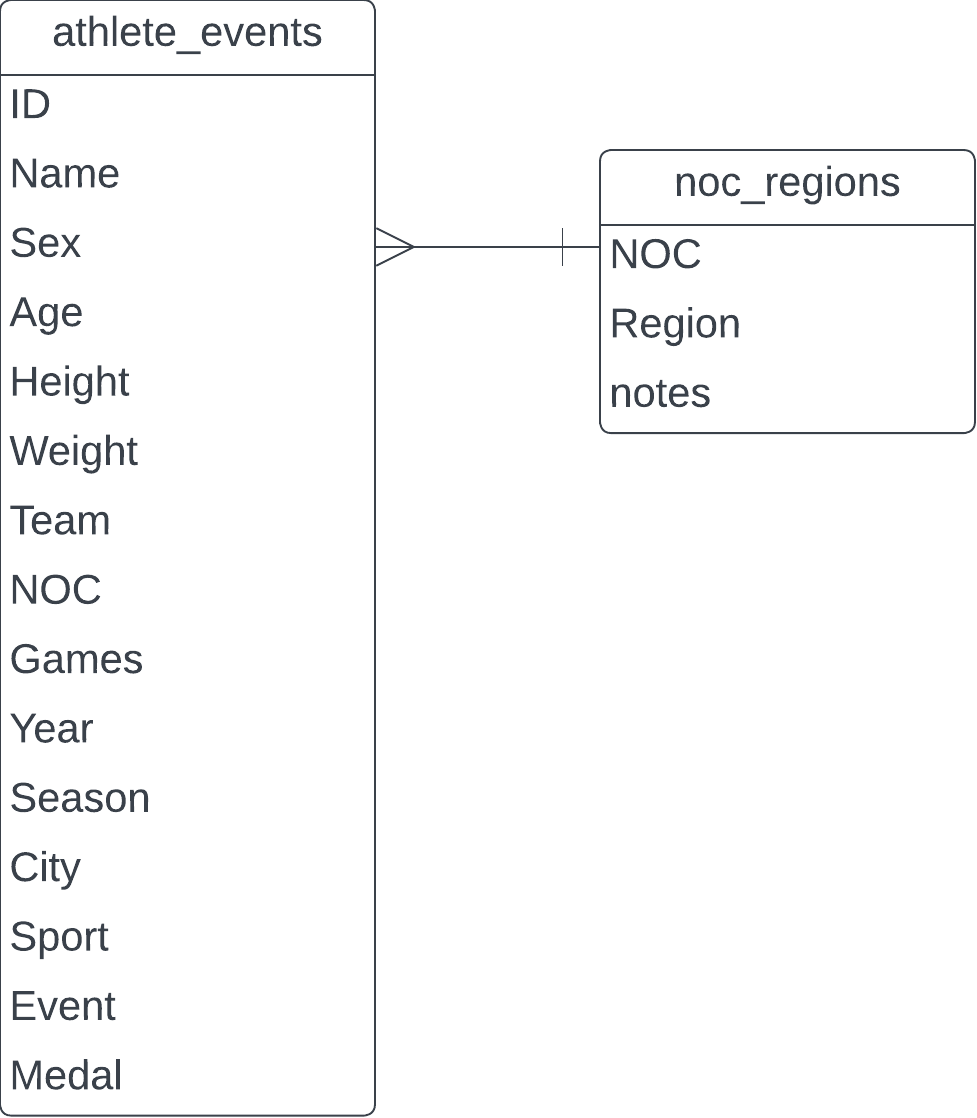




A Histogram of the years of the dataset:



Proposed ERD:



# DEVELOP PROPOSAL

## DESCRIPTION

To determine the physical body characteristics required for winning a medal, an analysis of various factors should be conducted. These factors may include but are not limited to height, weight, muscle mass, body fat percentage, endurance, strength, agility, and flexibility. By studying successful medalists in different sports, patterns and correlations between these characteristics and medal-winning performances can be identified. Additionally, considering the specific requirements and demands of each sport or event would be crucial in determining the necessary

## QUESTIONS

To find the answers to the following questions in the data, I would like to know:

1. The average height and weight characteristics for medal winners in each game, which will help me discover the ideal body type for a certain game.
2. The average age of medal winners in each game, which will help me discover the ideal age to win a game.
3. Which country has won the most medals in each game.

## HYPOTHESIS

1. The BMI values of the medal-winning athletes correspond to the ideal values on average.
2. Developed countries have achieved more medals due to their advanced sports infrastructure

## APPROACH

To find the answer, I will apply SQL queries to the given dataset. By using the Where and GROUP BY clauses, I can narrow down the results and analyze the data.