# Project Proposal

## Sport Statistics – Olympic Data

September 2020

Antonio Torralba

# Index

- Data Selection
- Questions and Approach
- EDR-Diagram
- Data Preparation
- Data Analysis
- Discovery
- Recommendations

# Data Selection

- **Sports Stats** (Olympics – 120 years of data)

- **Objectives**
  - Identify relations between nacionality, sex and sport.
  - Calculate the weights of each attribute involve in getting a medal.
  - Identify most relevant attributes for each sport.
  - Differences between medalists and non-medalists.

- **Potencial Audience**
  - Couches, Betting companies, General Managers...

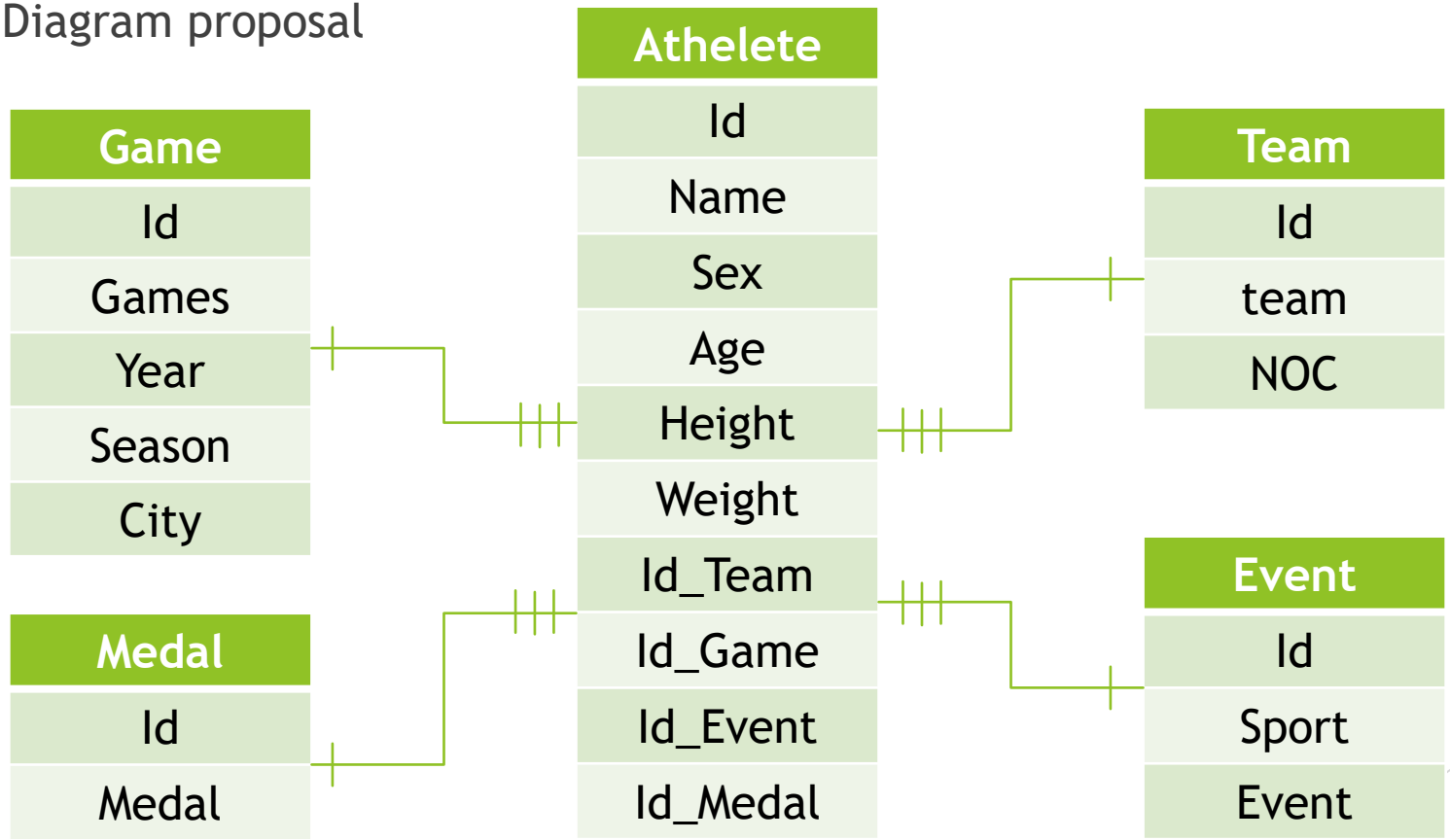
# Questions and Approach

- **Questions**
  - Which are the most popular sports?
  - Does some contries are inherently better at some sports?
  - Are height and weight equally important on each sport?
  - Are medalists and non-medalists actually so different?

- **Approach**
  - Calculate the number of participants in each sport (Every team counts as one)
  - Group medalists by nationality.
  - Calculate height and weight for each sport.

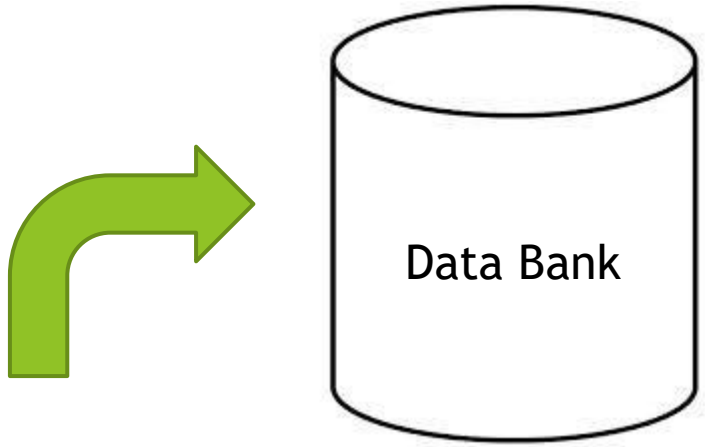
# ERD-Diagram

- ERD-Diagram proposal



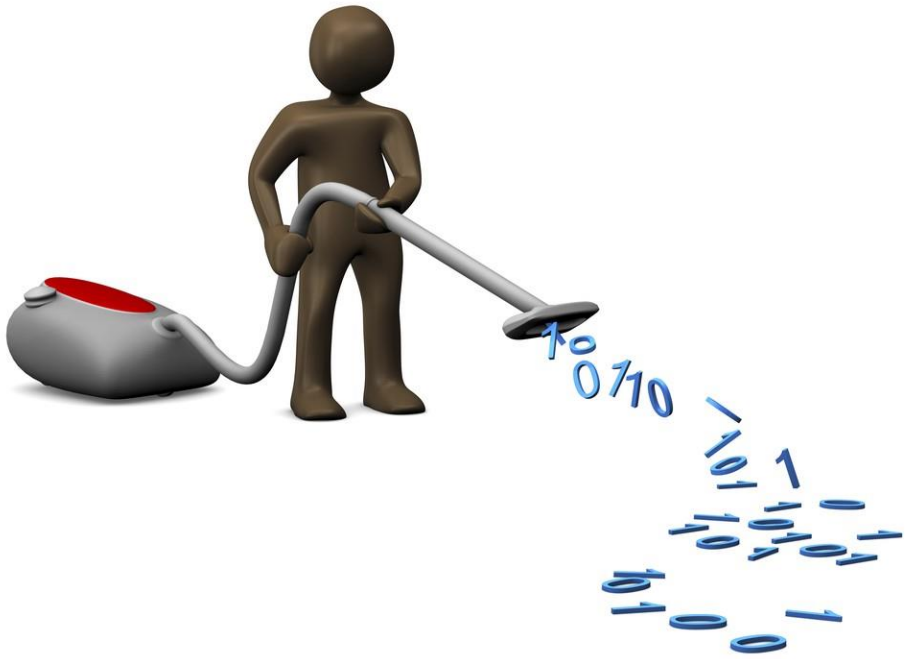
# Data Preparation

- **Import the data**
  - Select the data
  - Upload Files
  - Attach to a cluster table
  - Create the ERD tables




# Data Preparation

- **Clean the data**
  - Check for inconsistencies
  - Eliminate duplicate data
  - Delete incomplete data
  - Drop non correlated columns



# Data Preparation

- **Initial Exploration**
  - Size of the data
  - Number of sports
  - Type of the data
  - Basic statistics

| | region | notes |
|---|---|---|
| **count** | 227 | 21 |
| **unique** | 206 | 21 |
| **top** | Germany | Australasia |
| **freq** | 4 | 1 |

| | Age | Height | Weight | Year |
|---|---|---|---|---|
| **count** | 261642.000000 | 210945.000000 | 208241.000000 | 271116.000000 |
| **mean** | 25.556898 | 175.338970 | 70.702393 | 1978.378480 |
| **std** | 6.393561 | 10.518462 | 14.348020 | 29.877632 |
| **min** | 10.000000 | 127.000000 | 25.000000 | 1896.000000 |
| **25%** | 21.000000 | 168.000000 | 60.000000 | 1960.000000 |
| **50%** | 24.000000 | 175.000000 | 70.000000 | 1988.000000 |
| **75%** | 28.000000 | 183.000000 | 79.000000 | 2002.000000 |
| **max** | 97.000000 | 226.000000 | 214.000000 | 2016.000000 |

# Data Analysis

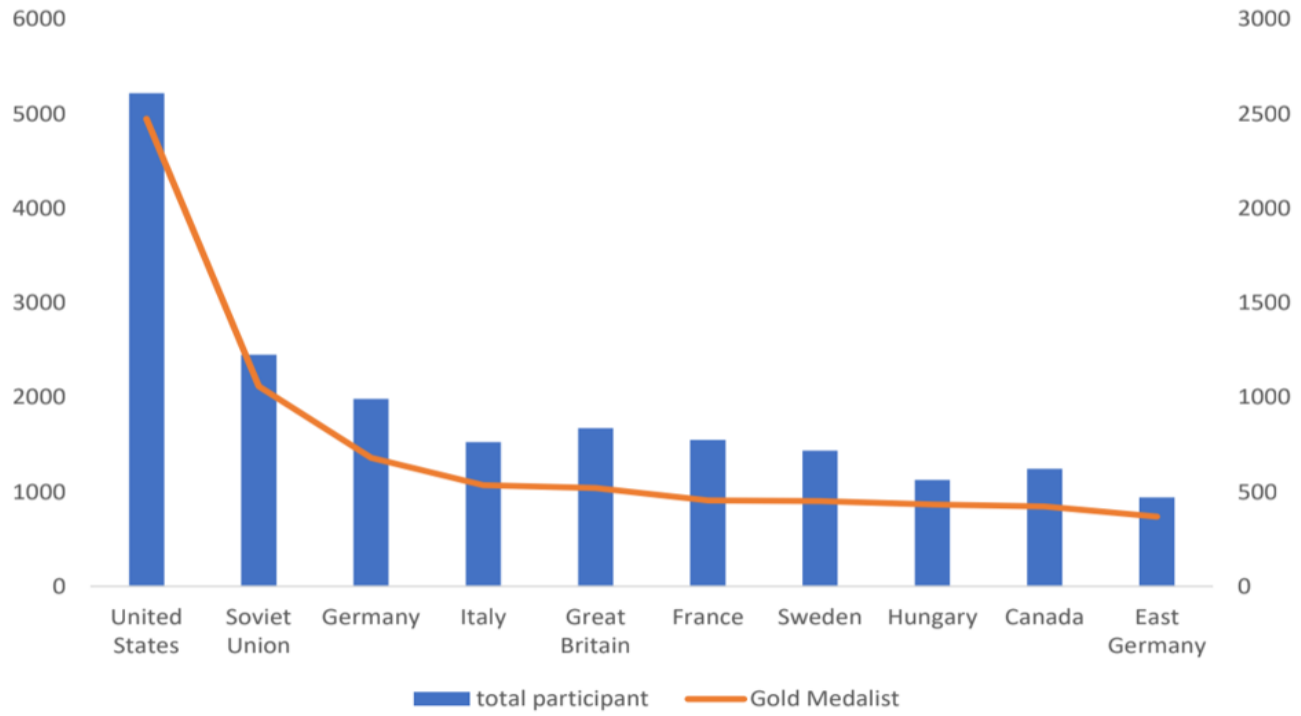
Average Age for medal winning:

$\sim 25\ years$

| | Medal | avg_age |
|---|---|---|
| 0 | None | 25.492289 |
| 1 | Bronze | 25.879210 |
| 2 | Gold | 25.901013 |
| 3 | Silver | 25.996724 |

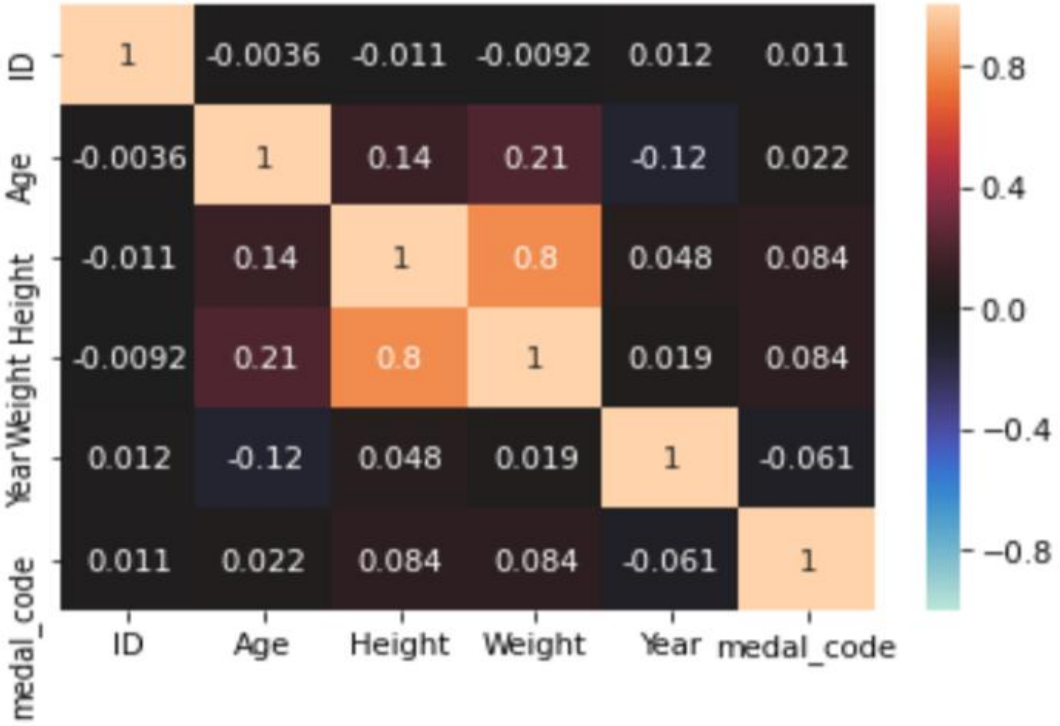
# Data Analysis

Linear correlation between the number of participants per country and the number of medals won by those countries.



# Data Analysis

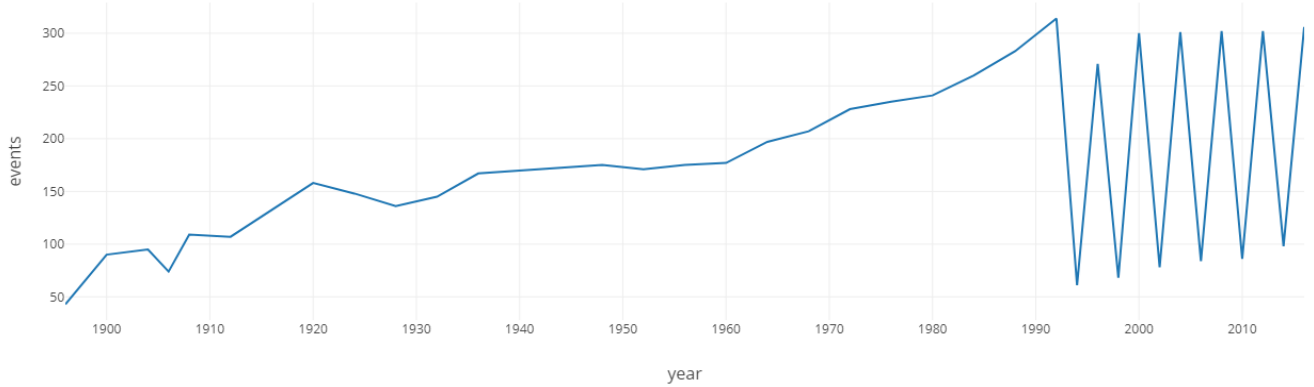
There is a weak correlation between the Weight/Height ratio with the medals:

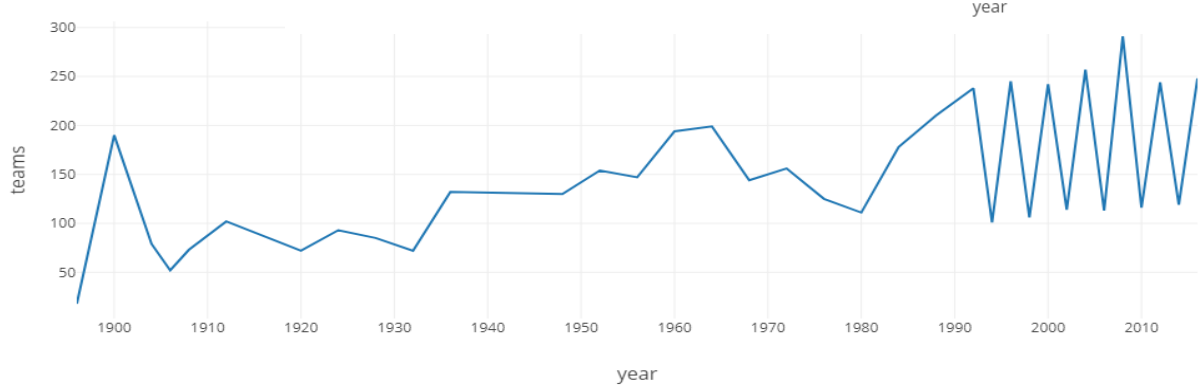



Correlation value of 0.082

# Discovery

Over this period of time, the number of teams and athletes grew with a strong impact in the year of 1924. It could explain why the number of events increase in 1928.



# Recommendations

- **Recommendations**
  - Based on the analysis or insights we discovered we are able to support the countries in order to help them grow in the sports arena
  - We can encourage people into contributing to participate in sports to represent their countries growing their economy
  - We recommend that the productivity and rates obtained in this work could be at people´s disposal