Project Proposal for Olympics data Analysis

Prepared by Amar

July 23, 2023

Data selection and preparation

Selected Dataset

- SportStats Olympic Dataset (120 years of data)

Reason for selecting

- Eager to discover the insights and patterns arising from sports events spanning various countries.

Importing the data

- Imported the csv files (athletes_events, noc_regions) to databricks environment
- Uploaded the csv files to databricks file management system(dbfs)

Create New Table

Data source ❓

Upload File   S3   Other Data Sources

DBFS Target Directory ❓

/FileStore/tables/ (optional)   Select

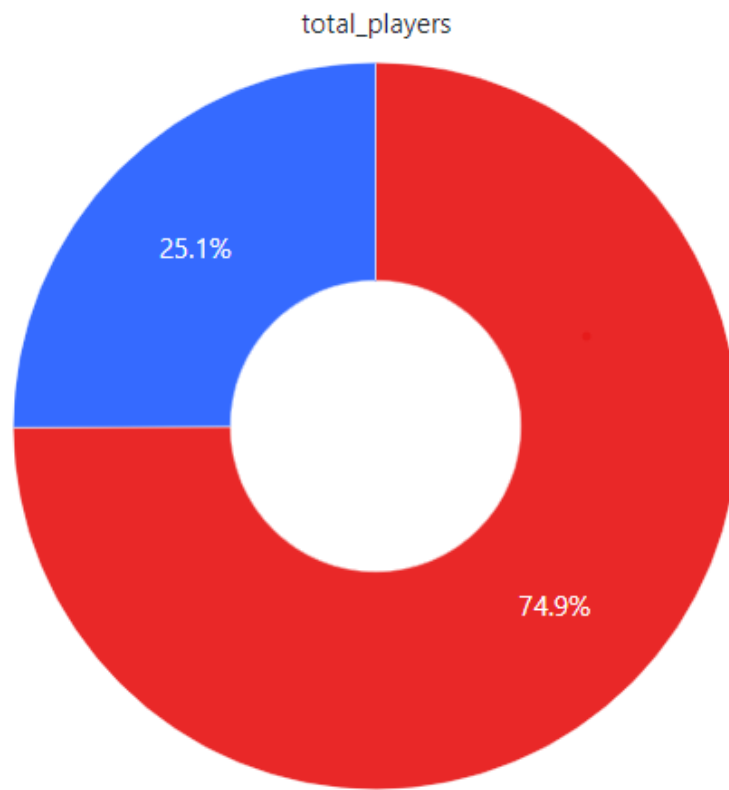Files uploaded to DBFS are accessible by everyone who has access to this workspace. Learn more
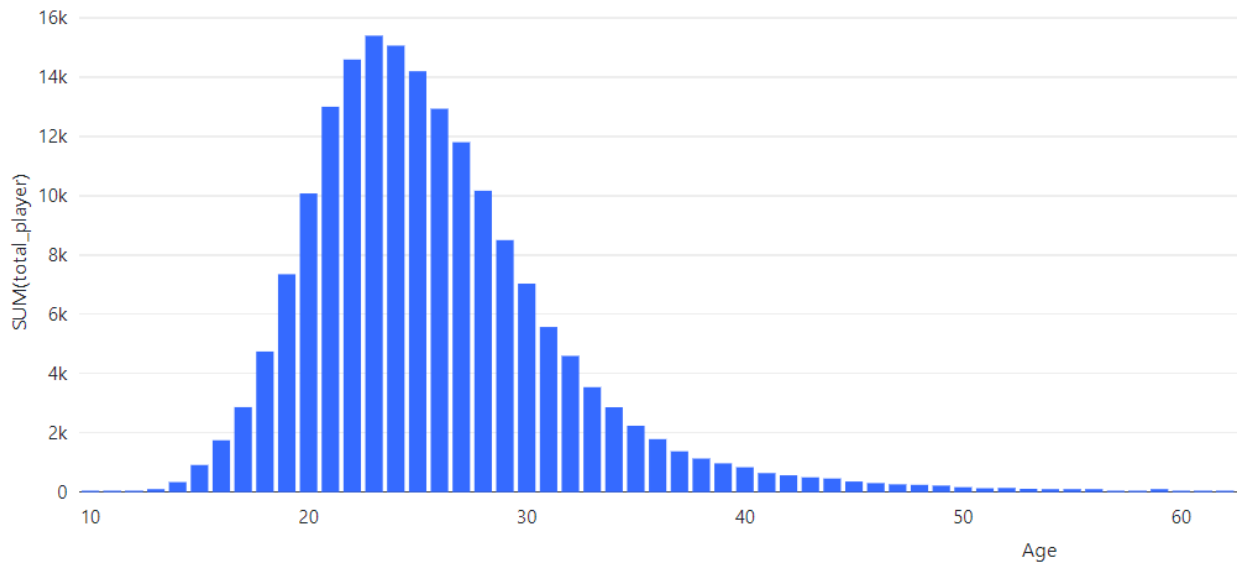
Files ❓

athlete_events.c

41.5 MB
Cancel upload
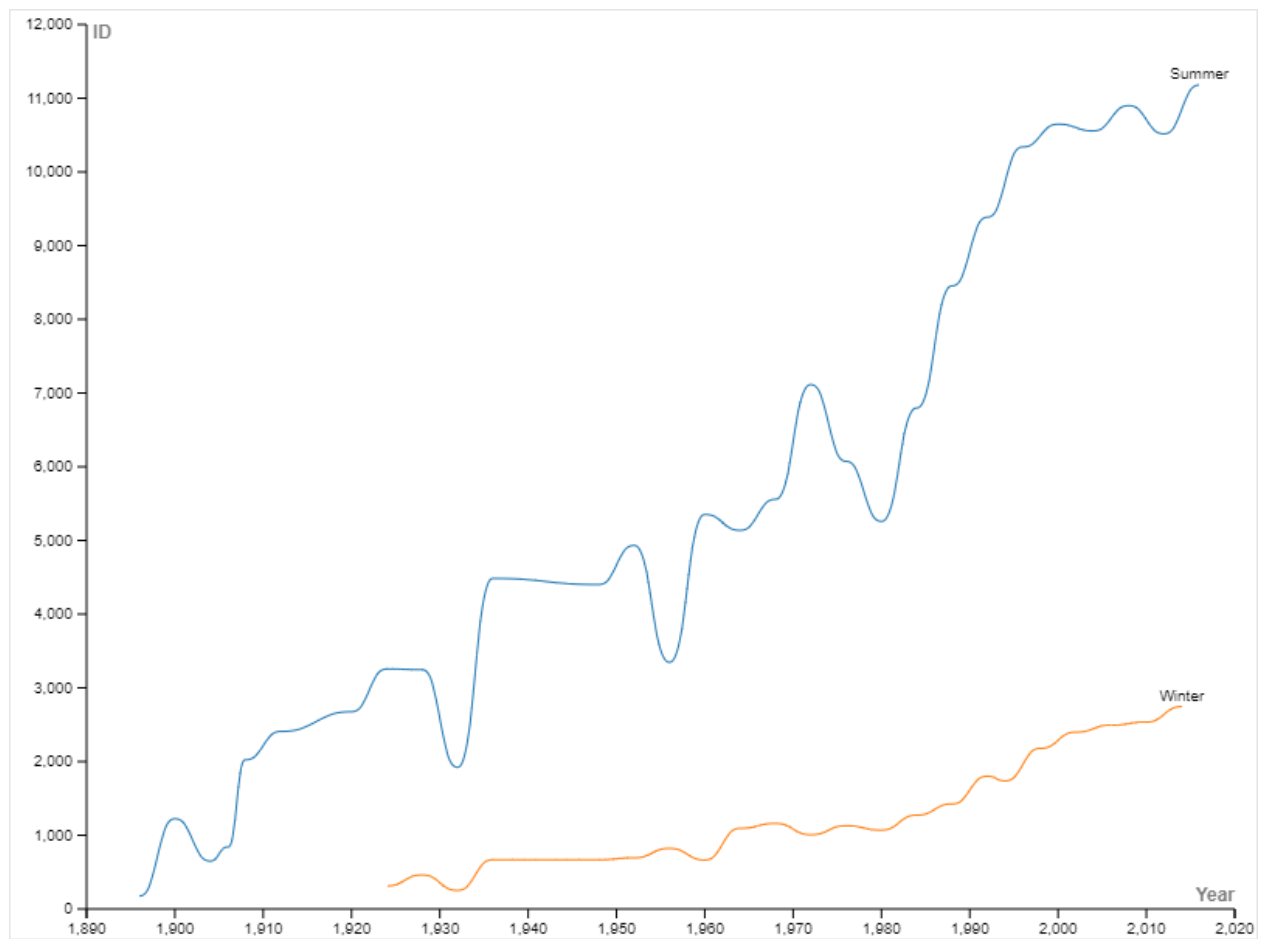
Initial Feature Exploration

- During the initial feature exploration, it was observed that the proportion of male athletes participating is greater than that of female athletes.

total_players

25.1%

74.9%

- There is compelling evidence suggesting that participants aged between 20 and 30 dominate the events.
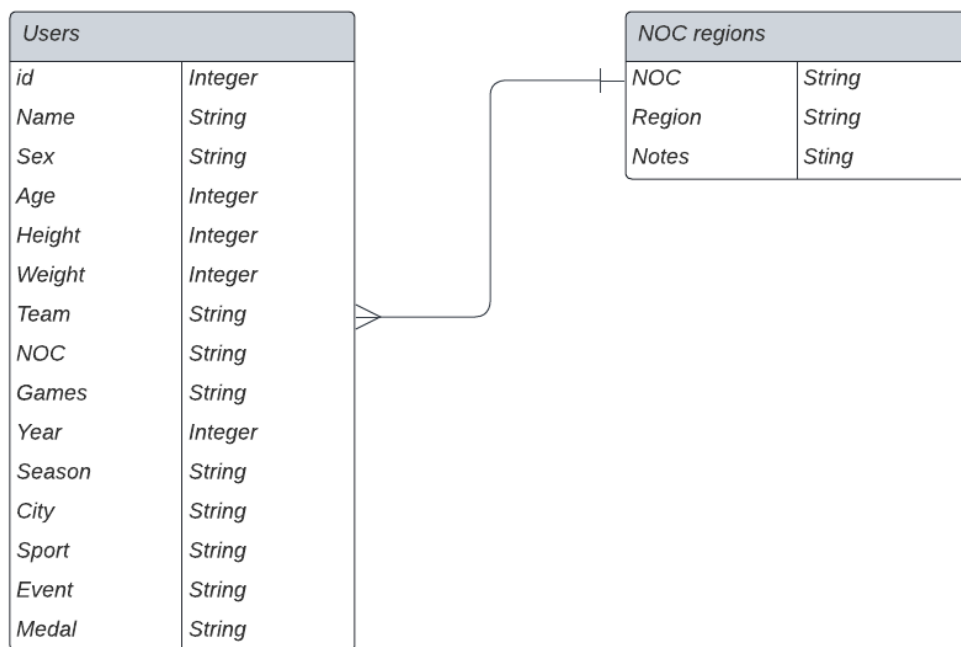
- Based on the observations, it is evident that the number of participations has consistently increased over successive years.
- In the summer season, there is a greater number of participants compared to the winter season..

Proposed E-R Diagram

ER Diagram
Sports Events

Users	
id	Integer
Name	String
Sex	String
Age	Integer
Height	Integer
Weight	Integer
Team	String
NOC	String
Games	String
Year	Integer
Season	String
City	String
Sport	String
Event	String
Medal	String

NOC regions	
NOC	String
Region	String
Notes	Sting

Objective

- The objective of this data analysis is to explore and analyze the Olympics dataset to uncover significant trends, patterns, and insights related to athlete performances, medal distributions, and the overall success of different countries in various sports. By conducting this analysis, we aim to provide valuable information that can be utilized for enhancing athletes' training strategies, understanding the factors influencing medal-winning performances, and making data-driven decisions to improve future Olympic events.

Questions we are going to answer.

- Which country has achieved the highest number of medals?
- Which country has achieved the highest number of gold medals?.
- The top ten athletes who have earned gold, silver, and bronze medals..

Hypothesis

- In the realm of Olympic judo sports, Japan stands as the country with the most gold medals won.
- In the Olympic Games' ice hockey sports, Britain stands as the top gold medal-winning country.

Data Analysis Approach

1. Initial feature explorations

   To understand the distibutin of age and proportions like sex ratio etc.

2. Handling missing values and outliers

Handling missing values and out;iers for higher accuray analysis.

3. Plotting visualizations

Turning data into graphs and charts to understand the data for better decision making.

Tools and environments

Data bricks, RawGraphs.io