Project Proposal

UNCOVERING TRENDS IN OLYMPIC HISTORY: AN EXPLORATORY DATA ANALYSIS

SEP 30,2025

RIMSHA IRAM

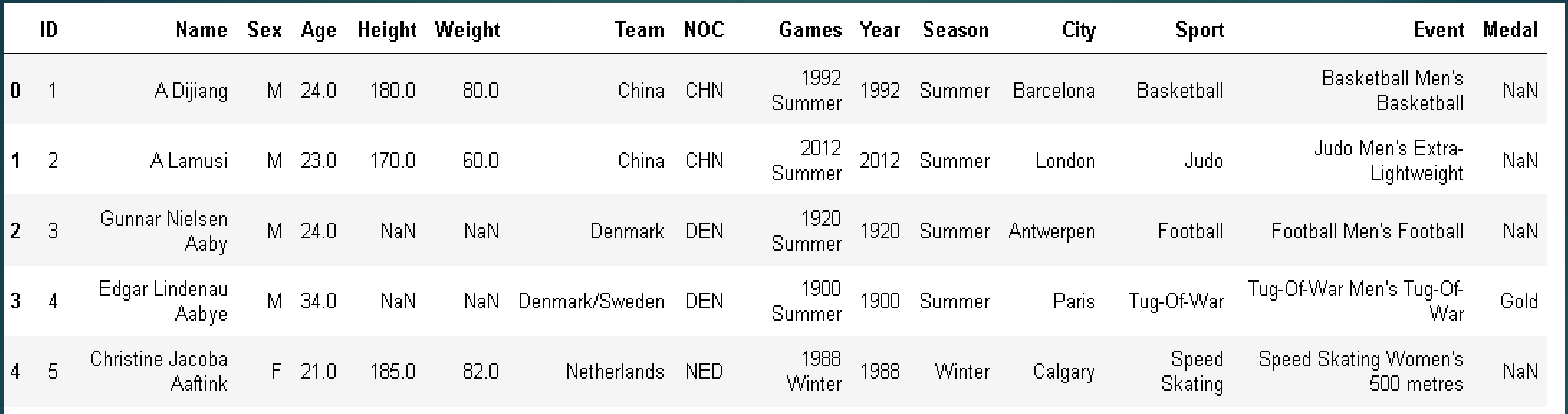
Client/Dataset Selection

- For this project, the client is a **sports analytics consultancy** interested in uncovering trends from the Olympic Games. The dataset chosen is **athlete_events.csv** from Kaggle, which contains detailed information about athletes, events, demographics, and medals across modern Olympic history (1896–2016).

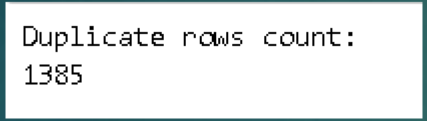
- This dataset was selected because:
- It is well-structured and beginner-friendly.
- It provides demographic, geographic, and performance-related attributes.
- It allows answering questions relevant to **sports federations, national committees, and sponsors** (our "clients").

Import & Cleaning Steps

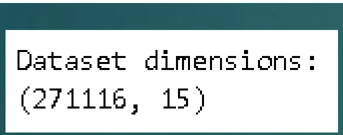
- Imported the dataset into Jupyter Notebook using pandas.
- Previewed the first rows to understand column names and structure.
- Checked dimensions: ~271,000 rows and 15 columns.
- Verified data types and noted categorical vs numerical fields.
- Identified missing values in Age, Height, Weight, and Medal.
- Checked and removed duplicate records based on ID, Year, and Event.
- Converted categorical variables (e.g., Season, Sex) into proper datatypes for analysis.



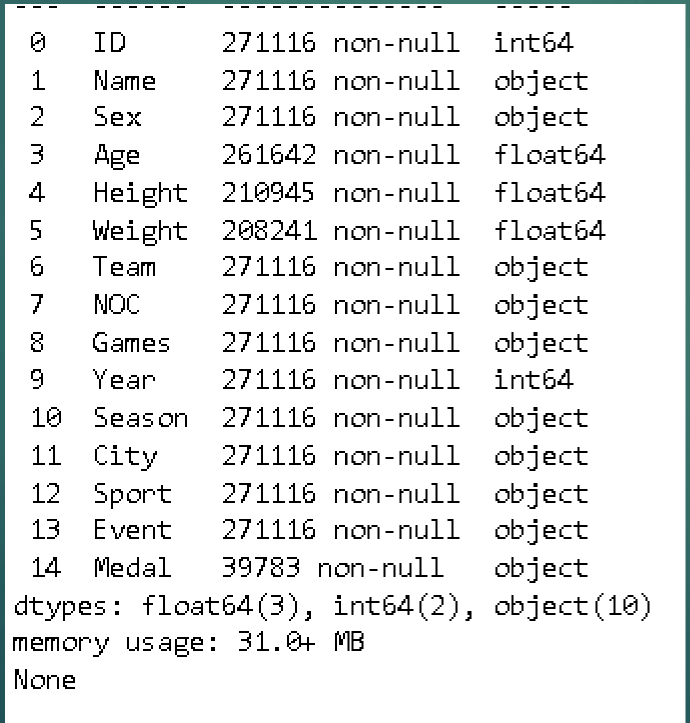
	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN



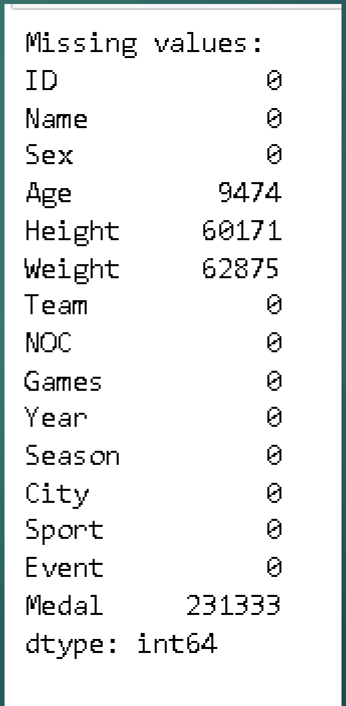
```
Duplicate rows count:
1385
```



```
Dataset dimensions:
(271116, 15)
```



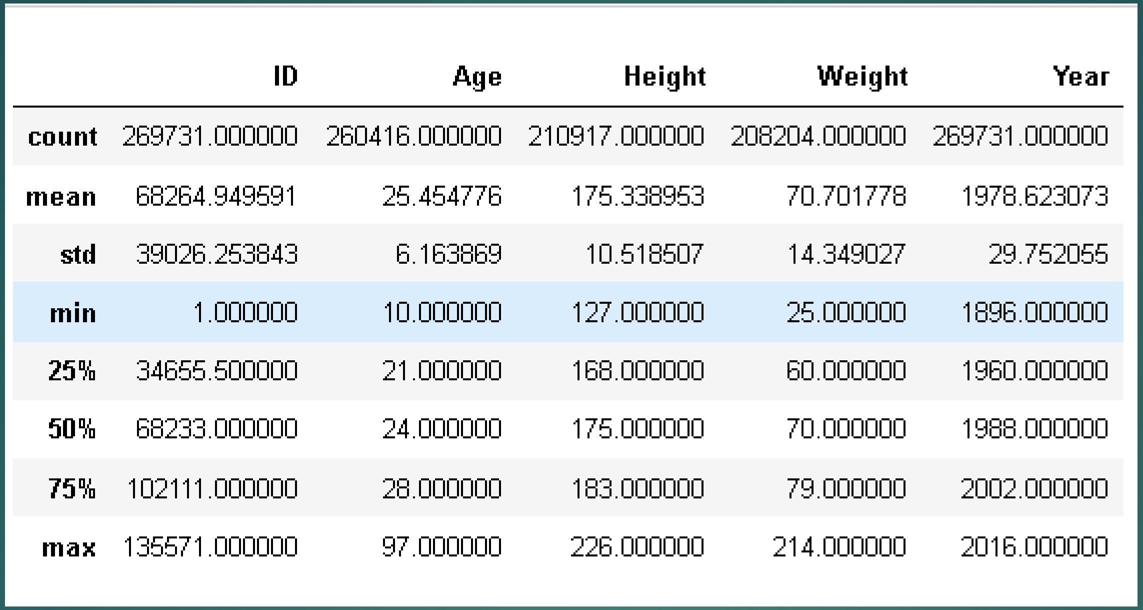
```
---  ------  --------------  -----
 0   ID      271116 non-null  int64
 1   Name    271116 non-null  object
 2   Sex     271116 non-null  object
 3   Age     261642 non-null  float64
 4   Height  210945 non-null  float64
 5   Weight  208241 non-null  float64
 6   Team    271116 non-null  object
 7   NOC     271116 non-null  object
 8   Games   271116 non-null  object
 9   Year    271116 non-null  int64
 10  Season  271116 non-null  object
 11  City    271116 non-null  object
 12  Sport   271116 non-null  object
 13  Event   271116 non-null  object
 14  Medal   39783 non-null   object
dtypes: float64(3), int64(2), object(10)
memory usage: 31.0+ MB
None
```



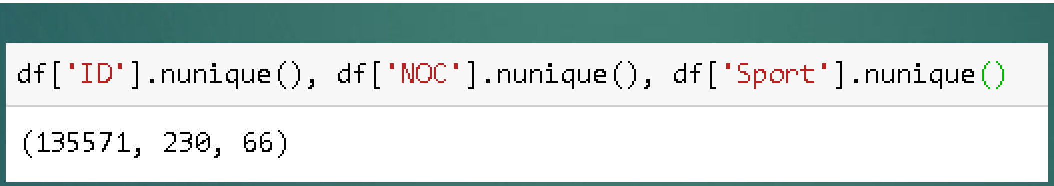
```
Missing values:
ID             0
Name           0
Sex            0
Age         9474
Height     60171
Weight     62875
Team           0
NOC            0
Games          0
Year           0
Season         0
City           0
Sport          0
Event          0
Medal     231333
dtype: int64
```

Initial Exploration

- Dataset spans 1896–2016, covering both Summer & Winter Games.
- Includes ~120,000 unique athletes, 200+ countries, and over 50 sports.
- Gender distribution shows participation of both male and female athletes, with male dominance historically.
- Medal distribution is sparse, with the majority of athletes not winning medals.
- Some athletes appear across multiple Games, making longitudinal analysis possible.

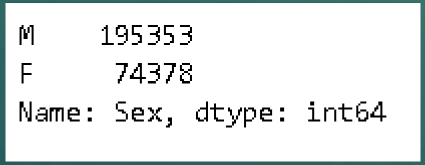


	ID	Age	Height	Weight	Year
count	269731.000000	260416.000000	210917.000000	208204.000000	269731.000000
mean	68264.949591	25.454776	175.338953	70.701778	1978.623073
std	39026.253843	6.163869	10.518507	14.349027	29.752055
min	1.000000	10.000000	127.000000	25.000000	1896.000000
25%	34655.500000	21.000000	168.000000	60.000000	1960.000000
50%	68233.000000	24.000000	175.000000	70.000000	1988.000000
75%	102111.000000	28.000000	183.000000	79.000000	2002.000000
max	135571.000000	97.000000	226.000000	214.000000	2016.000000

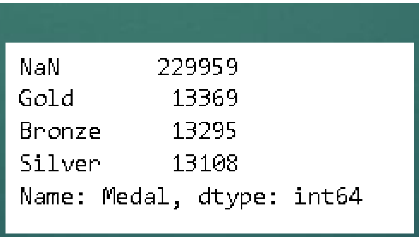


```
df['ID'].nunique(), df['NOC'].nunique(), df['Sport'].nunique()
```

```
(135571, 230, 66)
```



```
M     195353
F      74378
Name: Sex, dtype: int64
```

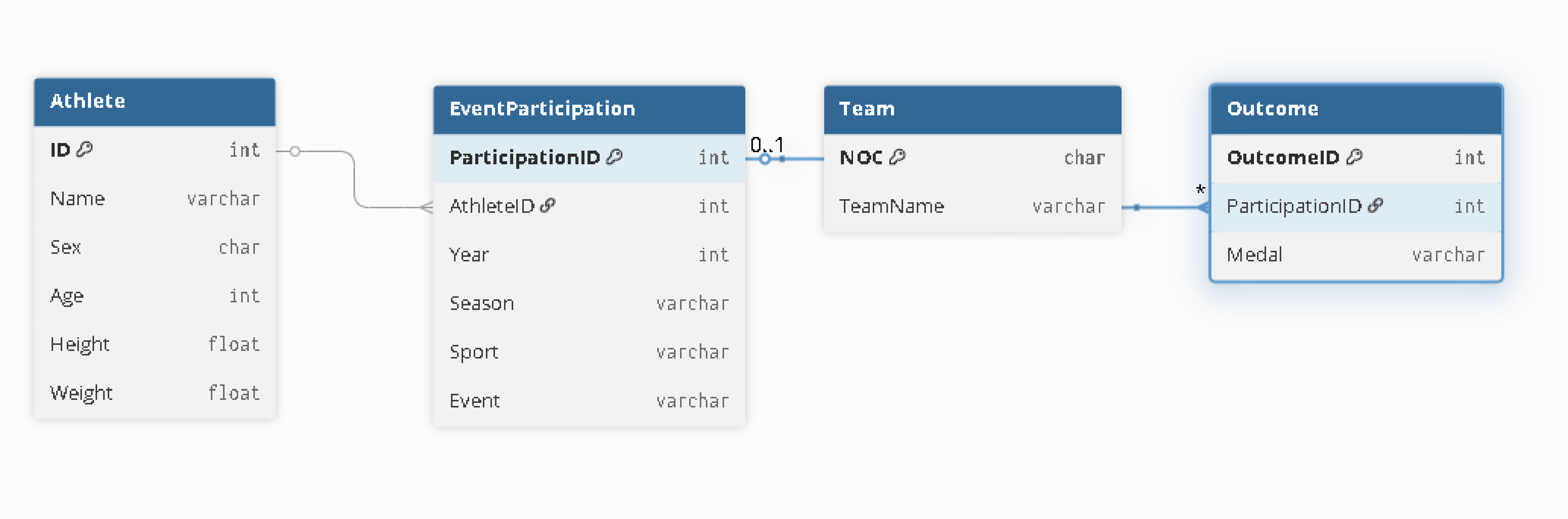


```
NaN       229959
Gold       13369
Bronze     13295
Silver     13108
Name: Medal, dtype: int64
```

Entity Relationship Diagram (ERD)

Although the dataset is a single CSV, the implied relationships are:

- **Athlete** (ID, Name, Sex, Age, Height, Weight)
- **Event Participation** (Year, Season, Sport, Event)
- **Team/Country** (Team, NOC)
- **Outcome** (Medal)



Athlete	
ID	int
Name	varchar
Sex	char
Age	int
Height	float
Weight	float

EventParticipation	
ParticipationID	int
AthleteID	int
Year	int
Season	varchar
Sport	varchar
Event	varchar

Team	
NOC	char
TeamName	varchar

Outcome	
OutcomeID	int
ParticipationID	int
Medal	varchar

Description

- This project explores historical Olympic data to uncover demographic patterns, medal distributions, and participation trends. Stakeholders such as the International Olympic Committee (IOC), national sports committees, coaches, and sponsors may be interested in these findings. The analysis will highlight how countries perform over time, how athlete demographics affect performance, and how participation has evolved.

Questions

- How do athlete demographics (age, gender, height, weight) relate to Olympic success (winning medals)?

- Which countries have historically dominated the Olympics, and how has their performance changed over time?

- What trends can be observed in participation — such as growth of female athletes or the popularity of certain sports?

Hypotheses

- Taller/heavier athletes are more successful in certain sports (e.g., basketball, weightlifting), while lighter athletes excel in others (e.g., gymnastics).

- Developed countries dominate medal counts due to larger investments in training and resources.

- Female athlete participation has steadily increased over the decades, especially after the mid-20th century.

Approach

- Focus initially on demographics (Age, Sex, Height, Weight) and performance (Medal).
- Explore relationships between athletes and countries (Team, NOC) and track over time (Year, Season).
- Use descriptive stats and group-by aggregations (e.g., medals by country, athletes per year).
- Visualize trends with histograms, line plots, and heatmaps.
- Evaluate hypotheses using metrics like:
- Medal counts and proportions (success rates).
- Demographic distributions of medalists vs non-medalists.
- Growth rate of participation (especially by gender).