DATA SCIENCE CAPSTONE PROJECT

SPORTS STATS

INSIGHTS INTO THE OLYMPICS

Author - Bhavesh Khamesra

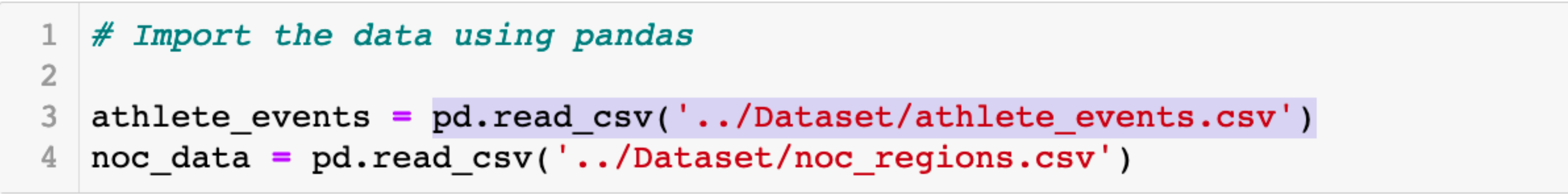
CONTENT

- Preparation -
  - Dataset Description
  - Data Cleaning
  - ERD Diagram
- Development
  - Questions
  - Initial Analysis
  - Hypothesis
  - Approach

DATASET DESCRIPTION

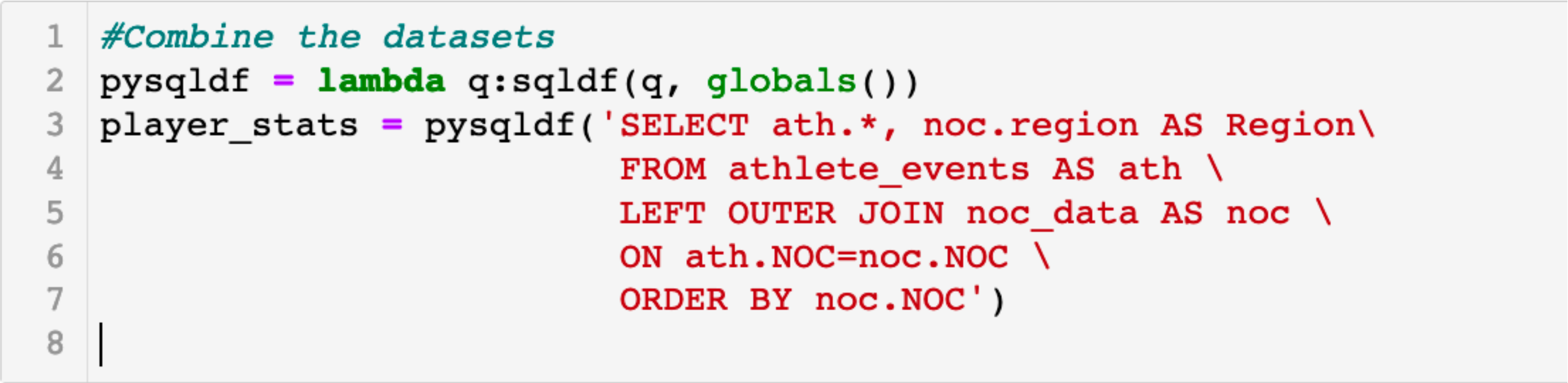
- **Dataset** - SportStats - a sport analysis firm partnering with local news and elite personal trainers to provide “interesting” insights to help their partners.
- **Goal** - Analyse the data for interesting patterns/trends which could be useful to partners in areas of health/journalism etc.
- **Dataset Details - Two separate datasets provided.**
- **Dataset 1 contains the following fields**
  - **Player’s Personal Details - ID, Name, Age, Sex, Height, Width**
  - **Player’s affiliation - Team, NOC**
  - **Player’s Sports Records - Games (participated), Sport, Event, Medal**
  - **Games Details - Year, Season, City**
- **Dataset 2 contains list of NOC and corresponding Region.**

DATA CLEANING



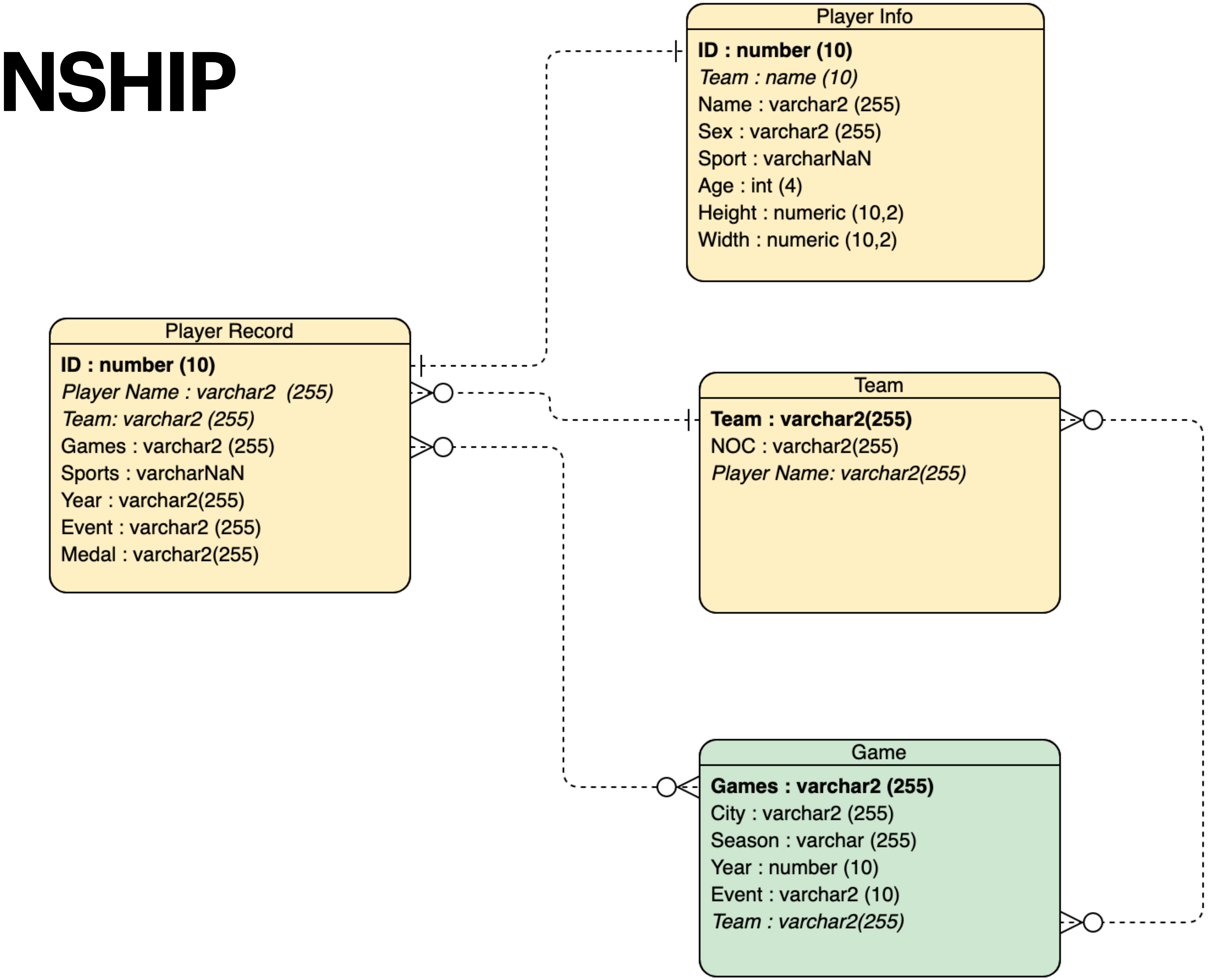
```
# Import the data using pandas

athlete_events = pd.read_csv('../Dataset/athlete_events.csv')
noc_data = pd.read_csv('../Dataset/noc_regions.csv')
```

- Importing Dataset using pandas
- Combine two datasets
- Check for null values - Remove null values from the dataset, if needed.
- Data segregation - Isolate relevant data columns to analyze specific questions
- Create metric functions - Like number of male and female players in same team, number of medals awarded to some players or some regions.



```
#Combine the datasets
pysqldf = lambda q:sqldf(q, globals())
player_stats = pysqldf('SELECT ath.*, noc.region AS Region\
                        FROM athlete_events AS ath \
                        LEFT OUTER JOIN noc_data AS noc \
                        ON ath.NOC=noc.NOC \
                        ORDER BY noc.NOC')
```

ENTITY RELATIONSHIP DIAGRAM



Player Info

ID : number (10)
Team : name (10)
Name : varchar2 (255)
Sex : varchar2 (255)
Sport : varcharNaN
Age : int (4)
Height : numeric (10,2)
Width : numeric (10,2)

Player Record

ID : number (10)
Player Name : varchar2 (255)
Team: varchar2 (255)
Games : varchar2 (255)
Sports : varcharNaN
Year : varchar2(255)
Event : varchar2 (255)
Medal : varchar2(255)

Team

Team : varchar2(255)
NOC : varchar2(255)
Player Name: varchar2(255)

Game

Games : varchar2 (255)
City : varchar2 (255)
Season : varchar (255)
Year : number (10)
Event : varchar2 (10)
Team : varchar2(255)

DEVELOPMENT - Questions & Assumptions

Questions -

1. Are female athletes equally represented as males in different regions?
2. Does there exist a bias between player’s age and the sports?
3. Are all sports dominated by a few countries or the distribution varies depending on the sports?

Assumptions -

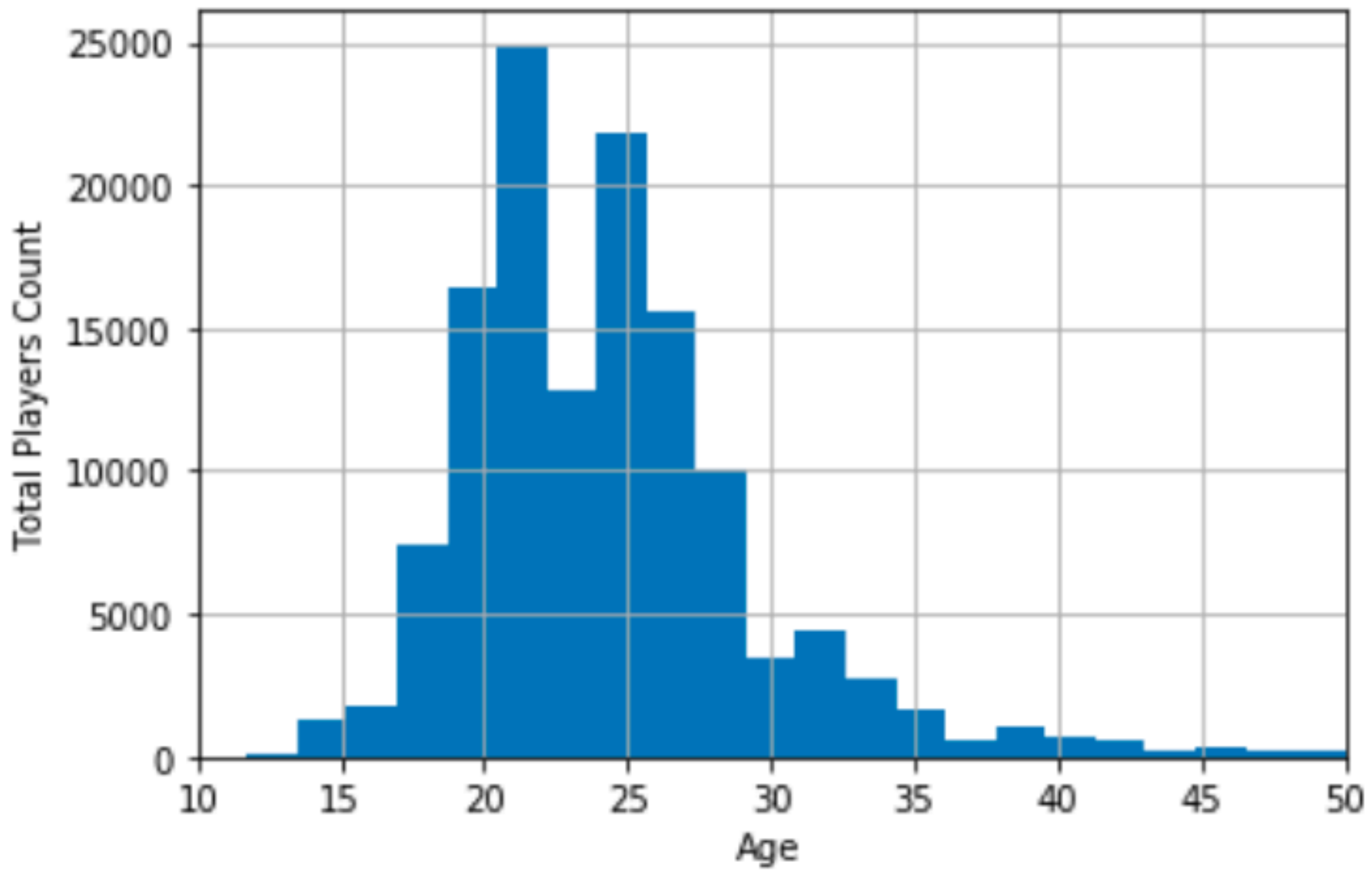
- The provided dataset holds no bias against players of any countries, sex, age, sports or any other catagories. The data is not misrepresentated or false.
- The sports itself provides equal opportunities to players of all age, sex and nationality.

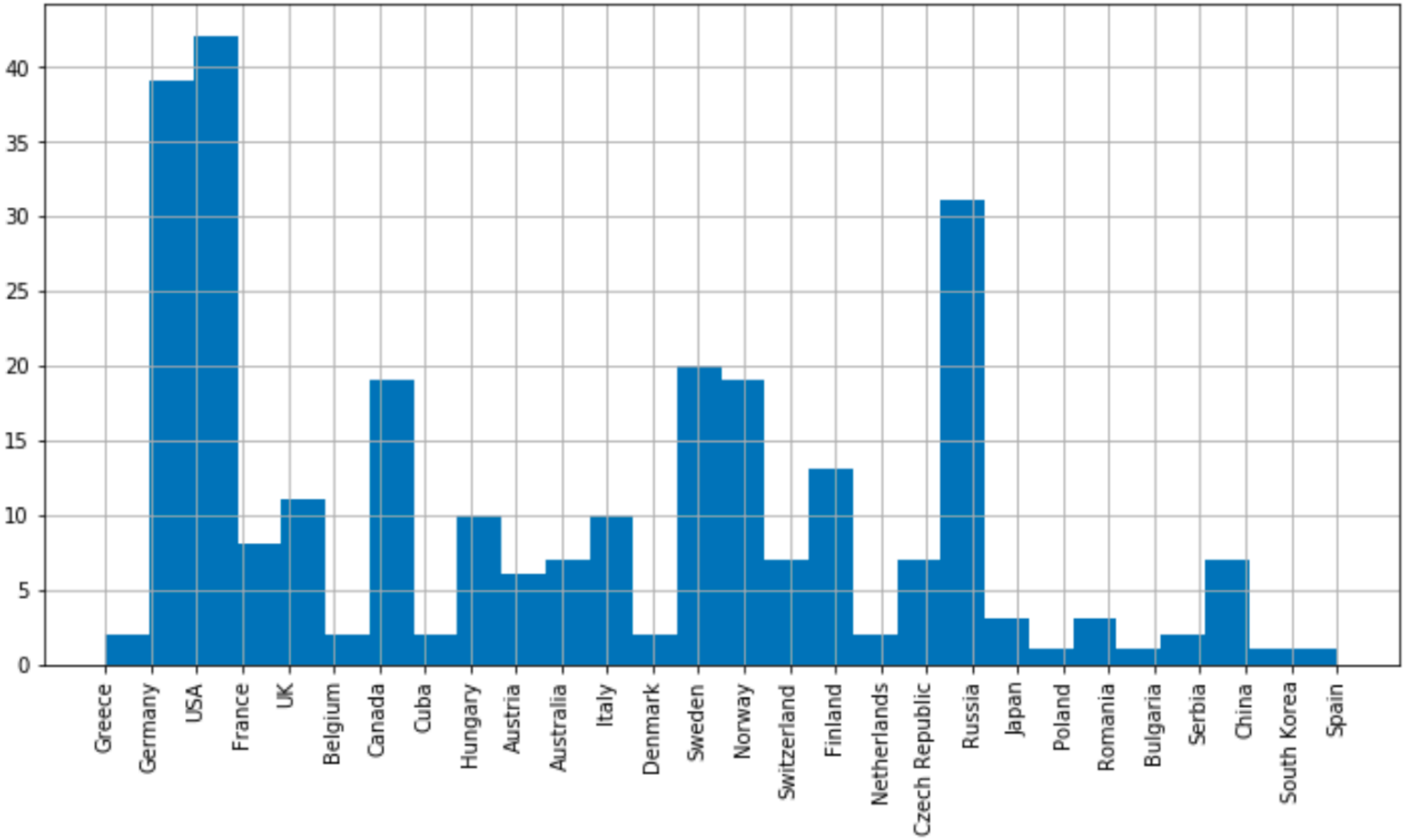
INITIAL EXPLORATION

- Male vs Female Participation - Total male players = 101590, Total Female players - 33981

- Player’s age - Consider age of all players over the years which have participated in these games. We have a double peaked gaussian which suggests that the optimal age of players varies across the sports.

- Dominating Nations - Consider the top five award winning regions in every games. Few nations have dominated these games over the years in most of the sport events.





Hypothesis

1. Men vs Women - The representation of males and females may be more fair in developed nations compared to developing nations, where opportunities are not equally available to females.

2. Agility vs Experience - Performance of player depends on both agility and experience. Hence, for every sport, there would be an optimal age at which player's peak performance can be expected.

3. Factor of Height - Certain sports such as basketball, relays etc might prove more advantageous for taller people.

4. Sport domination - More developed nations would have higher dominance in majority of sports due to better infrastructure and opportunities provided to their players.

ANALYSIS METHOD

1. Create separate tables based on ERD.
2. Metrics -
    1. Men vs Women - Find number of males and females for each team for each games they participated. This will also provide insight into how this ratio has changed over the years.
    2. Agility vs Experience - Look for all events under same sports category and compare the age of top players across the sports.
    3. Sport Domination - Compare number of total medals/awards won by each nation/teams across different sports for each game.

THE
END