**Exploratory Data Analysis Report: #Assignment 1**

**Date**- 22/09/2024

### **Group**:

### Lavanya HS (2023UG000157)

### Amrutheshwari V (2023UG000115)

### Tiya Rose Pulikunnel (2023UG000135)

### Priyanka Bhatt (2023UG000127)

### Gajendiran A (2023UG000154)

### Sairaj Bhise (2023UG000117)

**Exploring Determinants of Participation and Outcomes in the Paris 2024 Olympics: A Focus on Gender Disparity and the Performance of Host Nations**

# ***Introduction***

This project aims to investigate the multifaceted influences affecting athlete participation and performance in the Paris 2024 Olympic Games. It emphasizes the exploration of gender disparities and evaluates the historical performance enhancements observed in host nations, aiming to identify systemic advantages or patterns that might benefit the host country. Furthermore, the study will analyze trends in gender representation and representation of marginalized communities. The insights garnered from this project are intended to inform policy decisions, enhance training programs, and provide a more comprehensive understanding of the dynamics at play in one of the world's most prestigious sporting events. Through a combination of basic exploratory data analytics and qualitative assessments, the project seeks to understand the common issues faced during EDA processes while highlighting statistical and inferential information that contribute to academic discourse and promote equitable opportunities across the spectrum of Olympic competitors.

# ***Application of Data Science in Sports***

Data science plays a pivotal role in analyzing player performance and refining game strategies across various sports disciplines. Through the use of sensors, wearables, and video analysis tools, teams can capture data on player movements, biometrics, and tactical decisions during training sessions and live matches. Advanced analytics models can then process this data to identify patterns, trends, and correlations that traditional methods may overlook.

For instance, in basketball, data scientists analyze shooting percentages from different areas of the court to optimize shot selection. In soccer, tracking player positioning and passing accuracy helps coaches devise strategies for maintaining possession and attacking effectively. These insights not only improve individual and team performance but also provide a competitive edge by enabling data-driven decision-making on the field.

Data science revolutionizes sports by transforming decision-making, enhancing player

performance, and engaging fans. Examining its intricate role in reshaping the sports

arena, they uncover key aspects.

o Enhanced Decision-Making

o Performance Optimization

o Injury Prevention and Management

o Game Strategy Evolution

o Fan Engagement and Experience

o Business and Revenue Generation

o Technological Integration

o Global Collaboration and Competition

# ***Challenges in Private Data Acquisition***

* **Outreach and Data Request Challenges**

During the course of our project, a significant and immediate challenge we faced was acquiring private, relevant, and comprehensive data sets. Recognizing the need for detailed and diverse data sources to enhance our analysis, we attempted to procure potential data sets directly from leading multinational corporations like Unilever and Johnson & Johnson.

Our approach involved direct outreach to data scientists and key personnel within these organizations through professional networking platforms such as LinkedIn. An example of this was our communication with Rounak Singh, a data scientist, where we expressed our intent and the academic nature of our project. Despite our proactive efforts and assurance of confidentiality through the willingness to engage in Non-Disclosure Agreements (NDAs), we faced considerable challenges. The data requested often encompassed sensitive information that companies were not prepared to share outside of their internal stakeholders, reflecting industry-wide practices on data confidentiality and security.

* **Alternative Data Sourcing and Its Limitations**

In response to these challenges, we turned to alternative data sources like Kaggle, a platform renowned for its vast repository of datasets across various domains, and the Olympics as a wider scope for our project. While Kaggle provided an avenue to access publicly available data, it introduced its own set of challenges. The datasets available, although extensive, frequently lacked consistency in data features required for our analysis. Moreover, the data often did not align perfectly with the specific metrics or time frames relevant to our project’s focus on the upcoming 2024 Olympics, particularly concerning historical trends and forward-looking predictions.

This inconsistency in data features and the lack of targeted datasets specifically tailored to our project’s goals necessitated additional data processing steps and sometimes led to compromises in the breadth of analysis we could conduct. The variability in data quality and granularity also impacted our ability to draw as robust conclusions as we had initially hoped.

* **Conclusion and Forward Steps**

These experiences highlighted the critical importance of data accessibility and quality in conducting comprehensive research. They also underscored the need for establishing stronger collaborations between academia and industry to facilitate access to high-quality data for research purposes. Moving forward, enhancing partnerships and exploring more open data initiatives could be potential strategies to mitigate such challenges in future research endeavors.

We moved ahead, by using various methods that we discussed further, which aided in producing a dataset that uses various sources including but not limited to Kaggle.

**Dataset Features, Characteristics and Insights**

# ***Datasets***

## ***Dataset***-[athletes.csv](https://drive.google.com/file/d/1nPEexoTjxzDL-Nz07lwbq6b5nXhdo141/view?usp=drive_link)

### **Dimension and Details:**

1. The dataset on Olympic athletes contains **11,110 entries across 35** columns, covering attributes like nationality, height, weight, birth and residence details, and personal factors like hobbies, education, and influences. However, there is significant missing data in many fields: around 2,400 entries lack birth details, over 4,000 miss residence information, and personal details like occupation, nickname, and coach data are missing for thousands of athletes. These missing values likely stem from inconsistencies in data collection, athlete privacy, and historical gender disparities, reflecting broader societal and institutional biases in sports.

### 

### **Numerical and Categorical data:**

1. Quantitative:

→Interval: Age (sometimes, if not considered in a ratio context).

→Ratio: Height, Weight, Age (typically treated as a ratio), Number of Medals (if included).

1. Qualitative:

→Nominal: Gender, Nationality, Sport, Birth Place, Coach Name.

→Ordinal: Medal Type (Gold, Silver, Bronze), Education Level (if included).

### **Data types:**

1. Object (String/Text): Most columns, including categorical or textual data such as:“name”, “name\_short”, “gender”, “country”, “disciplines”, “events”, “birth\_place”, “residence\_place”, “coach”, etc.
2. Integer (int64): Used for numeric data without decimal points, including: “code”, “height”
3. Float (float64): Used for numeric data with decimal points:“weight”

These data types are common for datasets involving a mix of personal details, categorical data, and numeric measurements.

### **Feature details (Knowing our Data) :**

1. Missing Data:

1. **nationality\_full**: 19 missing values

→ The 19 missing nationality\_full values will have minimal impact on gender disparity analysis unless you're comparing it by nationality, where those cases can't be included.

→For host country vs. performance, missing nationalities could exclude a few athletes, slightly affecting performance comparisons.

1. **birth\_place**: 2397 missing values

→Missing values in birth\_place (2397 missing entries) will not directly impact *gender disparity* analysis.

→However, for host country vs. performance, missing birth places could hinder identifying athletes' origins, making it harder to assess how athletes from the host country perform compared to those from other places.

1. **weight**: 18 missing values

→ The 18 missing weight values are unlikely to impact *gender disparity* analysis significantly, unless you're comparing performance by weight categories.

→For host country vs. performance, missing weight data could limit analysis if you're examining the correlation between athlete weight and performance outcomes.

1. **residence\_country**: 2834 missing values

→The 2834 missing residence\_country values may have minimal impact on *gender disparity* analysis.

→However, for host country vs. performance, missing residence data could skew results, as it would make it harder to determine if athletes living in the host country perform better compared to those residing elsewhere.

Many columns, especially personal details such as birth\_place, residence\_place, family, and education, have significant missing values.

This missing data can skew analysis and lead to biased results, particularly when analyzing trends based on demographic factors like nationality, gender, or coaching availability.

### **With Respect to Our Main Goal: Gender Disparity**

1. Gender

→Shows the distribution of the males and females across sports on the basis of country, nationality, and birth\_country. This gives an idea of gender disparity (if any) with a side-by-side comparison with other features in the datasets

2. Country

→Shows the distribution of countries which have sent the number of males, females and others which shows the gender disparity level of the country.

3. Height

→The average height of the males and females may affect the end results which can affect the future inclusion into sports.

4. Disciplines(Sport)

→The ”disciplines” feature will show the gender entries to each sport and show which gender is dominating that sport.

### **With Respect to Our Main Goal: Host Country Performance**

1. Gender

→This feature gives the gender distribution w.r.t. performance and host country. This also gives an idea of how each gender is contributing to the performance in the host country. On the contrary, how gender is distributed in the performance with non-host country.

2. Country

→Country gives an idea about the country from which the athlete is from and if compared with other dataset which gives the data for host country v/s non-host country, we’ll be knowing if the athlete of the host country are performing better than other non-host country athletes.

3. Height

→Height of the athlete may affect performance in sports like running, swimming, badminton, boxing, etc.

4. Disciplines(Sport)

→The “disciplines” consist of the national sports/ the country that has high performance in the sport which can affect the other non-host countries’ performance in that sport.

In the dataset, one of the most obvious relationships between features is likely between height, weight, and the athlete's discipline (or sport). These physical attributes often correlate with the type of sport an athlete competes in. For example:

- Weightlifting: Athletes in strength-based sports like weightlifting tend to have higher body weight relative to their height due to muscle mass.

- Running and Gymnastics: Athletes in endurance sports like running or gymnastics tend to be leaner and shorter to optimize speed, agility, and performance.

- Basketball/Volleyball: Height is a significant advantage in sports like basketball and volleyball, where taller athletes have an edge.This relationship reflects how physical traits influence performance in specific Olympic events. By analyzing height and weight against sports disciplines, patterns emerge that are characteristic of the physical demands of different types of competition.

### 

### Expected and Actual Data Distribution:

In the context of comparing **actual** versus **expected** data for the Olympic athletes dataset, we can think of the distributions based on different features. Here are the relevant distributions you would encounter:

* Gender Distribution:
* Expected Distribution: A binomial distribution might be expected in terms of gender, where the outcomes are either "Male" or "Female" and the proportions should be approximately equal in a gender-balanced dataset. Ideally, we'd expect a gender distribution close to 50% male and 50% female.
* Actual Distribution: The actual data may reflect either a binomial distribution (if male and female athletes are represented equally) or an imbalanced binomial distribution (if there is a significant difference in the representation of genders).
* Height and Weight Distribution:
* Expected Distribution: Physical attributes like height and weight are often modeled using a normal distribution, where most athletes would cluster around the average height and weight for their respective sport.
* Actual Distribution: The actual height and weight distributions could either follow a normal distribution if the athletes' physical attributes are typical for their sports, or could show skewed normal distribution if there is a bias toward certain sports that require specific physical traits (e.g., basketball, weightlifting).
* Nationality Distribution:
* Expected Distribution: Ideally, you might expect a somewhat uniform distribution across countries, assuming equal representation of athletes from different nations.
* Actual Distribution: In reality, the data is likely to follow a Pareto distribution, where a small number of countries (e.g., USA, China, Russia) dominate the athlete representation, while the rest have fewer athletes.
* Sport/Discipline Distribution:
* Expected Distribution: You might expect a uniform distribution if all sports were equally represented in the dataset.
* Actual Distribution: The actual data is likely to follow a right-skewed distribution (or power law distribution), where a few popular sports (like track and field, swimming) have many athletes, while niche sports have fewer.
* Age Distribution:
* Expected Distribution: Age is often modeled using a normal distribution, where most athletes cluster around a typical age range (e.g., 20–30 years for prime performance).
* Actual Distribution: The actual age distribution might still be normal, but could show some skewness depending on the inclusion of younger or older athletes in specific sports.
* Residence Distribution:
* Expected Distribution: If athletes from all over the world are equally distributed across different countries, you might expect a uniform distribution for their residence.
* Actual Distribution: In reality, the distribution could be skewed or multimodal, as certain countries provide better facilities and may attract more athletes.

These distributions help compare how well the actual data reflects expected diversity in gender, physical traits, and nationality among Olympic athletes.

**Statistics and probabilities for the dataset:**

Numerical Statistics:

Height (available for some athletes):

* Mean: 178.97 cm (Excluding the zero/ missing values )

→ The average height alone, without breaking it down by gender, gives a general sense of the population’s overall stature but does not directly indicate gender disparity

* Standard Deviation: 89.51 cm

→The standard deviation for heights (or weights) can help us understand how much individual heights deviate from the average within each gender group. It will be given further in the report while breaking down the performance of athletes based on gender and country.

* Range: 0 cm to 222 cm

Weight (available for some athletes):

* Mean: 77.75 kg (Excluding the zero/ missing values )

→Identifying Physical Differences Between Genders

→Performance Disparities between males and females

→Highlighting Resource and Opportunity Disparities

* Standard Deviation: 13.11 kg

→

* Range: 0 kg to 113 kg

**Probability Distribution:**

* Gender Distribution:

Ideal: Depending on the sport, there might be near equal representation between male and female athletes in most global sporting events, especially in the Olympics. Ideally, we'd expect a gender distribution close to 50% male and 50% female.

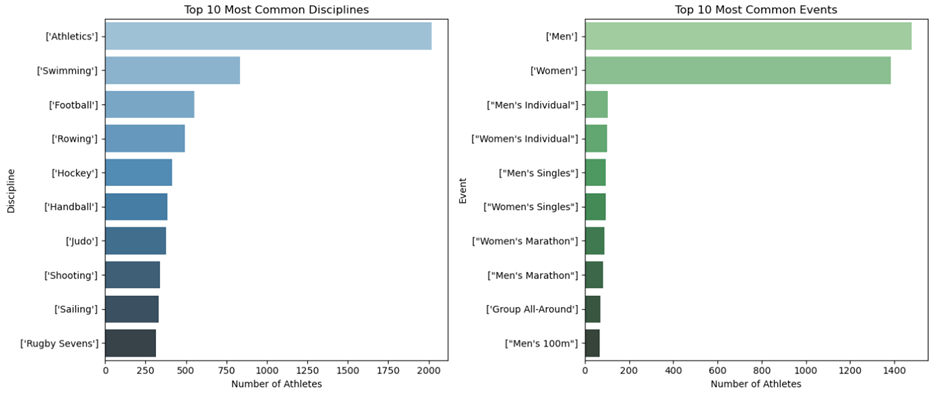
Dataset Compliance: The dataset shows 50.91% male and 49.09% female, which is quite balanced and adheres closely to the ideal distribution.

* Male: 50.91%
* Female: 49.09%

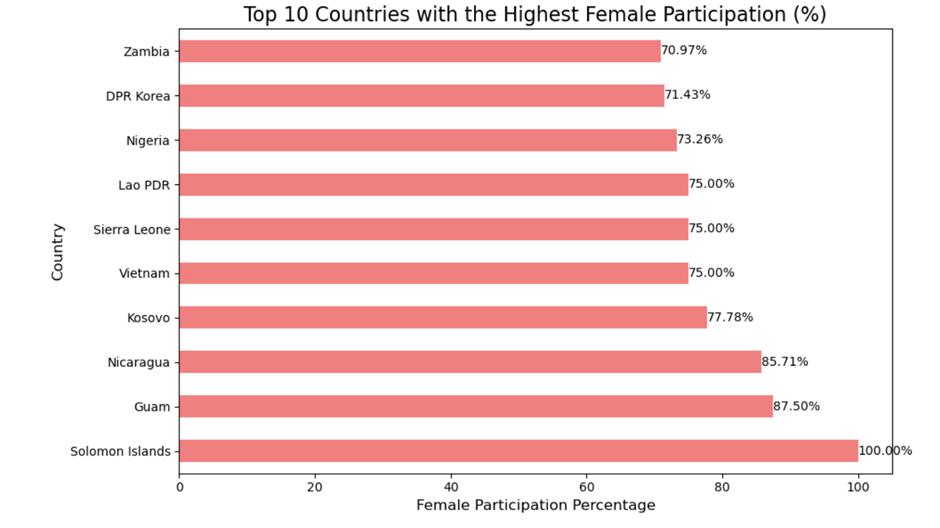
Overall, the dataset largely follows ideal distributions for gender, but it suffers from incomplete attributes like height and weight, which prevents it from fully adhering to ideal distributions.

**Visualization & Insights**

1. **Top 10 Most Common Disciplines vs Sports with the Most Entries:**



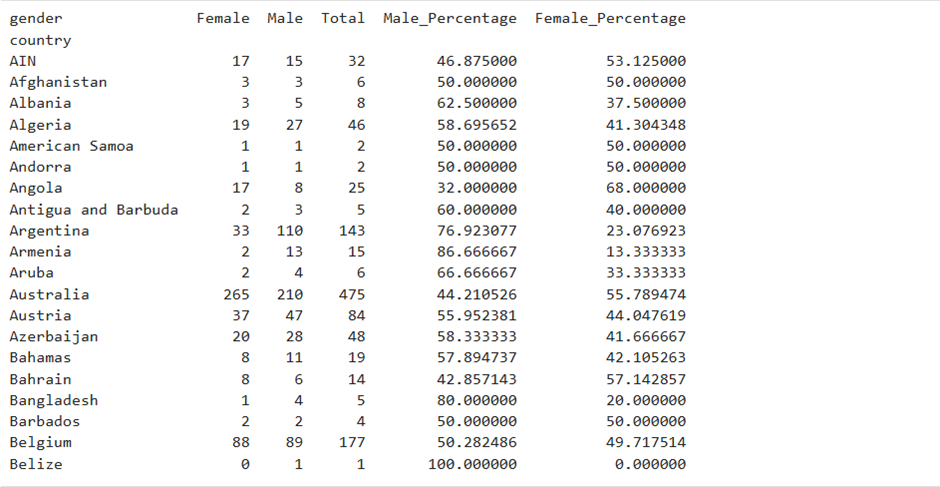
* Popular Sports/Disciplines: Athletics is by far the most common discipline, followed by Swimming. This aligns with the document's insight that "Athletics and Swimming have the highest number of medals, indicating their prominence in global sporting events."
* Global Appeal: The top disciplines like Athletics, Swimming, and Football represent sports with wide global participation, supporting the document's point about "global appeal."
* Comparing Event Size: There's a significant drop-off after Athletics and Swimming, suggesting these sports have many more events or attract more athletes.
* Investment in Certain Sports: The prominence of Athletics and Swimming suggests countries may invest heavily in these areas due to the high number of medal opportunities.

****

1. **Top 10 Countries with the Highest Female Participation (%):**

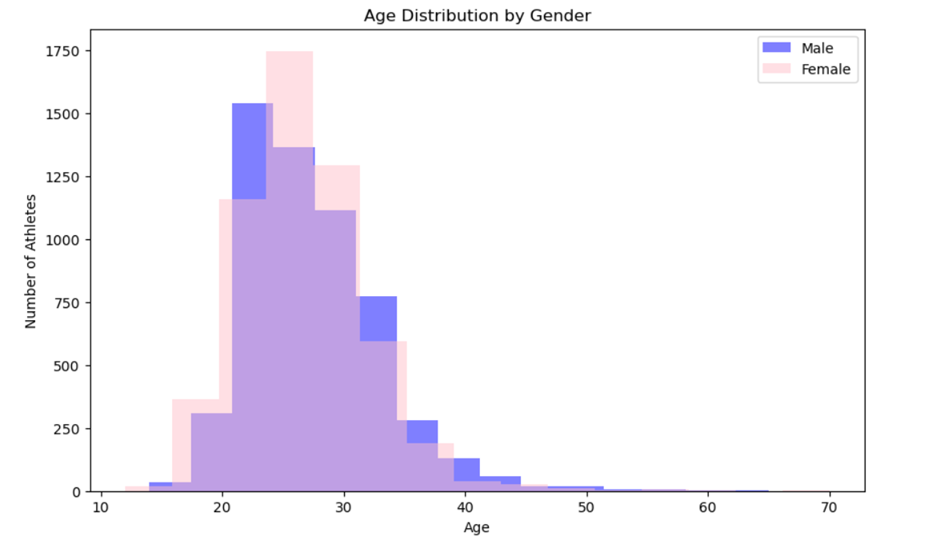
* Gender Parity: This data provides an interesting context for the document's discussion on "Progress in Gender Equality." Some countries, particularly smaller nations or developing countries, show very high female participation rates.
* Cultural and Policy Shifts: The high female participation in countries like the Solomon Islands (100%) and Nicaragua (85.71%) may reflect specific cultural or policy initiatives promoting women in sports in these nations.
* Opportunities for Growth: This data highlights potential areas where some countries are excelling in female sports participation, which could be studied for broader application.

1. **Gender distribution by country:**

****

* Gender Disparity in Medal Distribution: This data provides a more nuanced view of gender participation across countries, showing that while some nations have near-equal participation (e.g., Afghanistan, American Samoa), others have significant disparities (e.g., Armenia, Bangladesh).
* Progress in Gender Equality: Countries like Australia show slightly higher female participation (55.78%), aligning with the document's observation of progress towards gender equality in sports.
* Implications of Gender Disparity: The varied gender distributions across countries suggest that while overall medal distribution might be near equal (as mentioned in the document), individual countries may still have significant work to do in balancing participation.

1. **Age distribution by gender**

****

* Age range: The athletes' ages appear to range from about 15 to 60 years old, with the majority concentrated between 20 and 40.
* Peak age: Both male and female athletes peak in number around ages 20-25. This suggests that many athletes are in their prime competitive years during this age range.
* Decline with age: The number of athletes declines sharply after age 30 for both genders, indicating that fewer athletes continue competing at higher levels as they get older.
* Young athletes: There's a notable presence of young athletes (under 20), especially among females, suggesting early entry into competitive sports for many.
* Older athletes: While fewer, there are still athletes competing into their 40s and 50s, showing that some sports allow for longer careers.
* Gender differences: The female distribution seems to peak slightly earlier and decline more steeply than the male distribution, which could reflect different career trajectories or sport-specific trends.
* Symmetry: The overall distribution forms a roughly bell-shaped curve, skewed slightly towards younger ages, which is typical for athletic populations.

**Limitations:**

1. Historical Bias:

* The dataset may reflect historical gender disparities, where less data has been collected for female athletes, especially in earlier Olympic games. This makes it harder to draw accurate conclusions about gender equality and representation over time.
* Some sports may be underrepresented due to historical or geographical biases in the data collection process.

2. Inconsistent Data Collection:

* Data for certain athletes may be more detailed than for others. For example, fields like hobbies, influence, and philosophy have a lot of missing or sparse data, indicating inconsistencies in how the data was recorded.
* Older data may not contain the same level of detail as more recent data due to evolving standards of data collection.

3. Lack of Performance Metrics:

* The dataset focuses more on personal information (e.g., height, weight, nationality) but lacks detailed performance metrics, such as times, scores, or medals won. This limits the ability to analyze athletic performance directly.

4. Potential Privacy Issues:

* Some fields, like family details or residence information, may raise privacy concerns, especially if the data was collected without the athletes' full consent or was not anonymized properly.

## ***Dataset***-[schedules.csv](https://drive.google.com/file/d/1kGDofXM18xoBBcmObWSysN-Sj2p5u2ma/view?usp=drive_link)

**Dimension of dataset**: 3896 x 16 (16 features and each feature has a 3896 columns)

**Nominal or categorical data:**

Numerical features : ['event\_medal']

Categorical features : 'day', 'status', 'discipline', 'discipline\_code', 'event', 'phase', 'gender', 'event\_type', 'venue', 'venue\_code', 'location\_description', 'location\_code', 'url'

**Data types:**

1. Object: Most columns, including categorical or textual data such as: start\_date, end\_date, day, status,discipline, discipline\_code, event, phase,gender,venue,venue\_code, location\_description, location\_code, url
2. Integer (Int 64) : event\_medal

Possible Follow-Up:

It is recommended to analyze the specific disciplines or events associated with the awarded medals, as this could uncover trends in which sports are more competitive or tend to yield multiple medalists. This insight may provide a deeper understanding of patterns in medal distribution.

**Feature details (Knowing our Data) :**

**Missing values:**

Gender: 2 missing values

The two missing values do not impact the analysis of gender disparity. These rows pertain to the opening and closing ceremonies, which are unrelated to gender. Therefore, they are irrelevant and will be removed during the later stages of data cleaning.

Venue: 2 missing values

Similar to gender, The two missing values do not impact the analysis of gender disparity nor host country vs performance. These rows pertain to the opening and closing ceremonies, which are unrelated to the performance of the country. Therefore, they are irrelevant and will be removed during the later stages of data cleaning.

**With respect to our main goal: Gender disparity**

1. Start\_date:

→While these fields are primarily used to schedule events, they can reveal patterns of gender disparity if, for example, men’s events are scheduled at more favorable times (e.g., primetime slots), while women’s events may be scheduled at less favorable times.

2. Day:

→ The distribution of events for each gender across days may show patterns. For instance, if male events are scheduled more frequently on weekends (when more viewers are available) and female events on weekdays, that could indicate a form of scheduling bias.

3. Discipline:

→ Gender disparity can be examined by looking at the number of events for each discipline across genders. Certain sports may have more events or visibility for men compared to women or vice versa. For example, there may be a large number of football events for men but fewer for women.

4. Event:

→ The number of events associated with each gender can be counted and compared. If a certain gender has fewer events, it may indicate unequal representation or opportunity in that sport.

5. Event Medal:

→ This feature helps to compare the number of medals won by each gender and gain respective insights on gender disparity.

6. Gender

→ By analyzing the distribution of events between males and females, you can identify if there are fewer events, opportunities, or resources allocated to one gender compared to the other.

7. Venue

→You can investigate if one gender’s events are more often scheduled in prestigious or larger venues, while the other gender’s events are held in less prominent venues. This could indicate a disparity in how the events are valued and supported.

**With respect to our main goal: Host Country Performance**

1. Start\_date:

→Understanding the start and end dates of events helps track venue utilization over time. It shows which venues are the busiest on certain days or during certain periods. This information is useful for scheduling, logistics, and ensuring that venues are not overbooked.

2. Day:

→ This helps break down the venue usage by each day, allowing for day-specific venue planning and optimization. High demand days at certain venues can reveal peak utilization times.

3. Discipline:

→ Different venues are likely specialized for specific sports. This allows for the analysis of which sports utilize which venues most frequently and how evenly distributed the disciplines are across venues.

4. Event Medal:

→ Medals won at each venue, Certain venues, like those dedicated to high-participation sports such as athletics and swimming, have a significantly higher count of medal-winning events compared to others.

### **How the Columns Relate to Each Other:**

1. Time & Status:

→The start\_date, end\_date, and day columns are essential for understanding the scheduling of events. They pair with the status column to provide a timeline of events and their completion status.

2. Sport Information:

→ The discipline, discipline\_code, event, phase, and event\_type columns work together to classify the events by sport, gender, and the specific phase of the competition.

3. Venue Information:

→ The venue, venue\_code, location\_description, and location\_code columns describe where the events are held, making it easy to filter or sort events by location.

4. Additional Resources:

→ The url column provides external information, likely linking to more details or live results for each event.

**Expected distribution VS Actual distribution:**

In the context of comparing **actual** versus **expected** data for the Olympic athletes dataset, we can think of the distributions based on different features. Here are the relevant distributions you would encounter:

start\_date & end\_date

* **Expected Distribution**:The event times would likely follow a uniform distribution over the course of the Olympics, assuming events are evenly scheduled throughout the day. However, more events may cluster around prime viewing hours (e.g., afternoon and evening).Most events would start and end within a few hours. If the Olympics last for several weeks, events will span across all days.
* **Potential Visualization**: Histogram or line chart showing the frequency of event start times and durations over the days.

day

* **Expected Distribution**:This column would likely have a uniform distribution over the span of the Olympic Games. Each day should have a similar number of events, with perhaps a slight increase in event counts on weekends or as the Games progress toward the finals.
* **Potential Visualization**: Bar chart of events per day.

status

* **Expected Distribution**:Initially, most events will have a "Scheduled" or "Upcoming" status, and as time passes, more events will switch to "Finished". Toward the end of the dataset, almost all events would have a "Finished" status.
* **Potential Visualization**: A time series chart showing the proportion of finished, ongoing, and scheduled events over time.

discipline

* **Expected Distribution**:The distribution of sports disciplines might be skewed. High-demand sports like Football, Track & Field, or Swimming could have more events, while niche sports might have fewer events.
* **Potential Visualization**: A bar chart or pie chart showing the count of events for each sport.

discipline\_code

* **Expected Distribution**:This would mirror the discipline column but in abbreviated form. The frequency of each discipline code would be similar to the distribution of discipline.
* **Potential Visualization**: Similar to discipline, bar chart or pie chart.

event

* **Expected Distribution**:If this column refers to gender (as it does in the image, showing "Men"), we might expect a relatively equal distribution between "Men" and "Women" events, depending on how evenly distributed sports are across genders. In some disciplines, there may be more men's events, and in others, more women's.
* **Potential Visualization**: A bar chart showing the split between men's and women's events.

event\_medal

* **Expected Distribution**:Initially, most events would have no medals assigned (represented as 0), but as events finish and medals are awarded, the distribution will become more varied, with values representing gold, silver, and bronze.
* **Potential Visualization**: Stacked bar chart showing the progression of medal assignments over time.

phase

* **Expected Distribution**:For team sports (e.g., football, rugby), earlier phases (group stages) will have many events, while later phases (like finals) will have fewer events. This leads to a pyramid-like distribution, where the number of events decreases as the competition progresses.
* **Potential Visualization**: A pyramid chart showing the number of events per phase or a Sankey diagram for event progression.

gender

* **Expected Distribution**:Similar to the event column, we expect a relatively equal distribution between "M" (Men) and "W" (Women), with possible slight variations based on the sports or competitions.
* **Potential Visualization**: Bar chart showing the breakdown of events by gender.

event\_type

* **Expected Distribution**: The distribution here will depend on the types of events in the dataset (e.g., team vs. individual). For team sports, expect "HTEAM" (if this represents team-based events) to be more prevalent.
* **Potential Visualization**: Pie chart or bar chart to show the distribution of team vs. individual events.

venue & venue\_code

* **Expected Distribution**: Some major stadiums or venues (e.g., Stade de France, Parc des Princes) may host many events and therefore have a higher frequency, while smaller or more specialized venues may host fewer events. The distribution may be right-skewed, where a few large venues host many events, and many smaller venues host fewer events.
* **Potential Visualization**: Bar chart showing the number of events at each venue.

location\_description & location\_code

* **Expected Distribution**:This would follow a similar pattern to venue. Some locations (like large cities or regions) will host a higher number of events, while other locations may have fewer.
* **Potential Visualization**: Map-based visualization or a bar chart showing the distribution of events across locations.

url

* **Expected Distribution**:Every row should have a unique URL for event results or information. The distribution wouldn't be directly visualized but would be important for linking events to external resources.
* **Potential Visualization**: Not typically visualized, but used for linking out to event results.

**Statistics:**

There is only one feature with appropriate numeric data to find the mean ‘event\_medals’ - number of medals won by each discipline

Mean : 0.213864

Std: 0.653475

Min: 0.000000

25%0.000000

50%: 0.000000

75%: 0.000000

Max: 3.000000

**Insights:**

1. The data is heavily skewed toward events where no medals were awarded, as indicated by the 25%, 50%, and 75% percentiles all being 0.
2. Only a small fraction of events resulted in medals being awarded, which can be inferred from the mean of 0.21.
3. The few events that do award medals might do so generously, as seen by the high standard deviation and the maximum value of 3 medals.

**Visualizations from this dataset:**

### **Event Count by Gender and Discipline**

* Columns: gender, discipline, event
* Visualization: A stacked bar chart showing the number of events for each sport (discipline), separated by gender**.**
* Insight:This allows you to directly compare the number of events allocated to male and female participants in each sport. Gender disparity is evident if some disciplines heavily favor one gender**.**

**Insight we can get that relate to Gender Disparity:**

1. **Balanced Representation in Many Disciplines**: Many disciplines, such as athletics, swimming, judo, and fencing, now feature a similar number of events for male and female athletes. This marks progress toward gender equality compared to the 1980s, when events were largely male-dominated.
2. **Mixed-Gender Events on the Rise**: The inclusion of mixed-gender events (X) across disciplines like athletics and swimming highlights a growing focus on gender-inclusive formats, a notable change from 1986 when such events were rare. This shift promotes collaboration between male and female athletes in competition.

### 

### **Total Event Count by Gender**

* Columns: gender, event
* Visualization: A pie chart or bar chart showing the total number of events for males vs. females across all sports.
* Insight: This gives a big-picture view of whether there are more events for one gender overall, indicating general gender disparity.

1. **Balance Between M and W**: The distribution between **M (49.8%)** and **W (46.0%)** is nearly balanced, indicating an almost equal participation rate among these two groups. This balance could suggest successful efforts towards gender inclusivity, especially if there was a significant gender disparity in the past.
2. **Significant Minority Representation (X and O)**: **X (3.6%)** and **O (0.5%)** represent non-binary or other gender categories. While these numbers are relatively small, they show that the event is inclusive of gender minorities, which could be an important consideration in fostering a diverse community. The presence of this representation, even though very negligible, could indicate progress towards broader gender inclusion, especially if these categories were absent or even smaller in earlier stages.

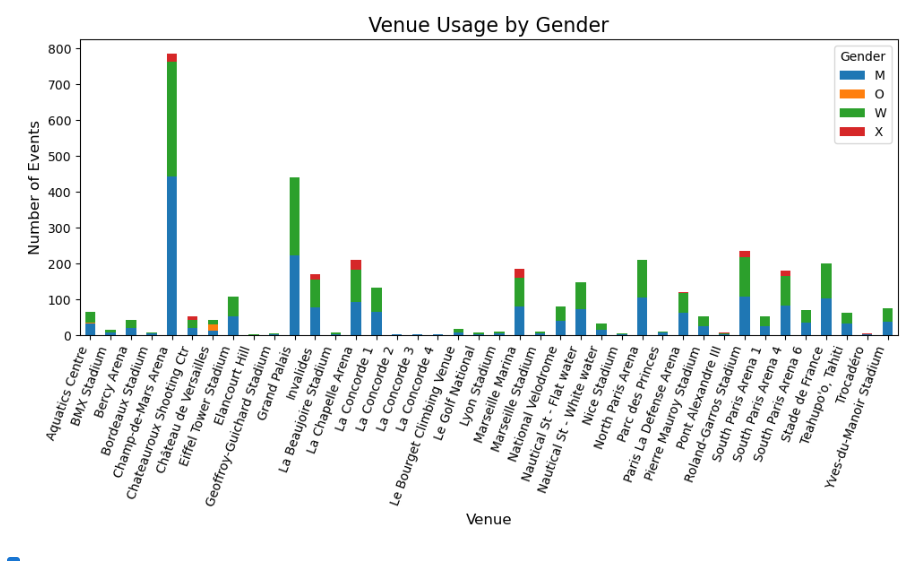
### 

### **Venue Usage by Gender**

* Columns: gender, venue, event
* Visualization: A stacked bar chart or heatmap showing the distribution of male and female events across different venues.
* Insight: This helps to identify if certain venues are predominantly used for male or female events, which could indicate gender disparity in the allocation of resources (e.g., prestigious or larger venues going to one gender).

**Insight we can get that relate to Gender Disparity**

1. Most venues like **Le Golf National**, **Pierre Mauroy Stadium**, and **La Chapelle Arena** show a more balanced representation of **M** and **W**, indicating progress in gender inclusivity. This balance may reflect efforts in these locations to bridge the gender gap, as both M and W seem to have close numbers of events.
2. The representation of **W (green)** in many venues has grown significantly, reducing the disparity that might have existed at the beginning. While there are still venues where **M (blue)** dominates, the balance at several venues indicates efforts to promote inclusivity.
3. Venues like **La Défense Arena** and **South Paris Arena 3** have events that include more representation for **X** and **O**, suggesting an improvement in the visibility and participation of gender minorities over time. Pointing towards increasing gender equity in event participation.

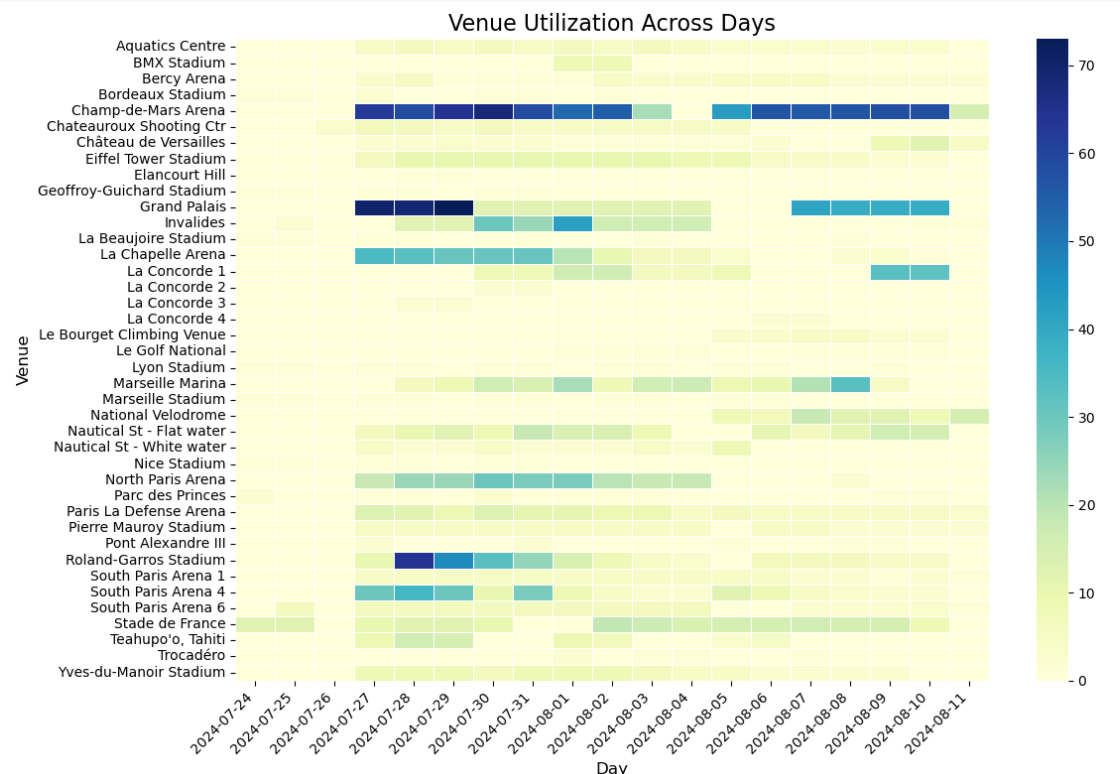
****

**Showing a aggregate visualization for better readability**

### 

**Venue Utilization:**

* Columns: venue, venue\_code, discipline, start\_date, end\_date
* Visualization: Heatmap or bar chart showing how frequently each venue is used across different days.
* Insight:
* Useful to understand which venues are most active and on which days.
* Venue Familiarity: Prior to the competition, athletes from the hosting nation usually have more opportunity to train at and acquaint themselves with the locations. This heatmap shows you which venues (such BMX Stadium and Aquatics Centre) are most used on important occasions. competitors from hosting nations are expected to perform better at these heavily trafficked venues because they have probably trained here for a long time, while competitors from non-hosting nations could only have temporary access to these locations in the run-up to the competition.
* Access to Infrastructure: Prior to the tournament, athletes from the hosting nation are more likely to have prolonged access to top-notch infrastructure. The heatmap, for instance, demonstrates how frequently crucial events are held at locations like Bercy Arena and the Aquatics Centre. It's possible that athletes from the hosting nation had more regular access to these elite training locations, while athletes from non-hosting nations might have only had time to train there in the run-up to competition.
* Specialized Venues: Athletes from the host nation may be more accustomed to the particular conditions at some venues, such as Teahupo'o in Tahiti for surfing or Nautical St. White Water for water sports. These unique surroundings could be especially challenging for athletes from non-hosting countries to adjust to, giving competitors from hosting countries a competitive advantage.
* Preferred Scheduling for Hosting Athletes: Athletes from the hosting nation may be granted more advantageous competition schedules, which would enable them to participate within windows of optimal performance. You can see clusters of popular locations on the heatmap that are used a lot on important days (like July 27–30). By enabling them to compete in well-known locations at times that fit their training cycles, this focused scheduling may benefit the athletes who are hosting.
* Geographical Proximity: International athletes have more travel requirements and must frequently adjust to different time zones, climates, and venue circumstances. However, hosting athletes do not experience these problems because they are already accustomed to their environment. When comparing travel plans with high-utilization days and venues (like July 27–30), one can utilize the heatmap to find them. Athletes who are not hosting could have an edge, particularly in locations that have been used previously in the competition.



### **Venue Utilization**

**Number of Medals won at Each Venue:**

* Insights:

According to the data, several venues—such as Champ-de-Mars Arena and Grand Palais—contribute significantly to the total medal distribution, with athletes from the hosting nation perhaps having an edge in these high-medal sites. Athletes that are hosting gain from having experience with these locations because they have probably had more access for training and competition, which gives them a better understanding of the venue's nuances than their rivals. Their prospects of victory are further increased by the psychological boost of competing in front of their enthusiastic home audience, particularly in larger stadiums with higher attendance.

* The venue-specific specialty is another important consideration, especially in remote or environmentally difficult areas like Teahupo'o, Tahiti. Due to their experience with the circumstances and local knowledge, athletes from the hosting country may do particularly well in these settings, offering them a clear advantage in specialist activities such as water sports or surfing. This benefit may be especially noticeable in settings where environmental influences have a major impact on performance results, hence increasing the difference between hosting and non-hosting nations.
* The success of the hosting country is influenced by logistical advantages such lessened travel weariness, advantageous timing, and acclimation. Athletes that are hosting are probably more rested and physically fit, particularly in high-medal sites where important events are centered. Athletes from hosting countries frequently outperform those from non-hosting countries in major athletic events, which skews the overall medal tally in their favor. These logistical and environmental factors, along with extended access to excellent venues, help explain this phenomenon.

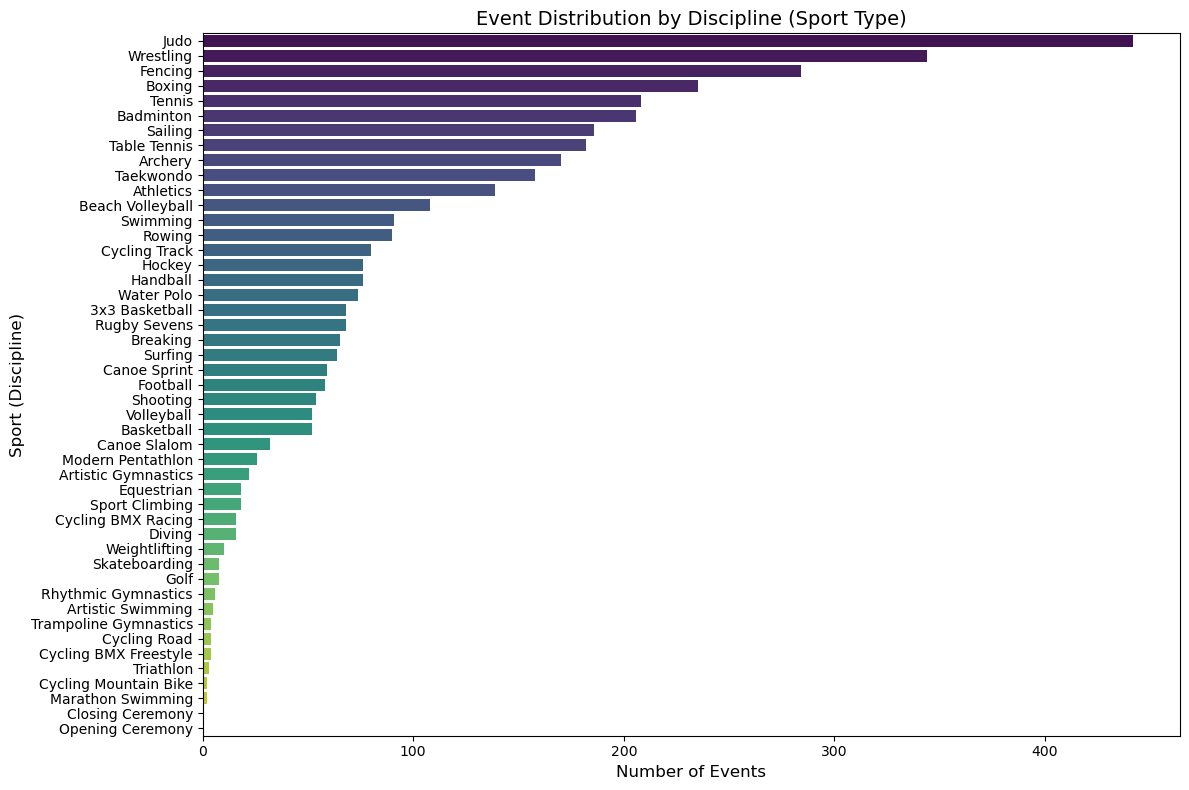
### 

### 

### **Additional visualizations unrelated to the problem statement but derivable from the dataset**

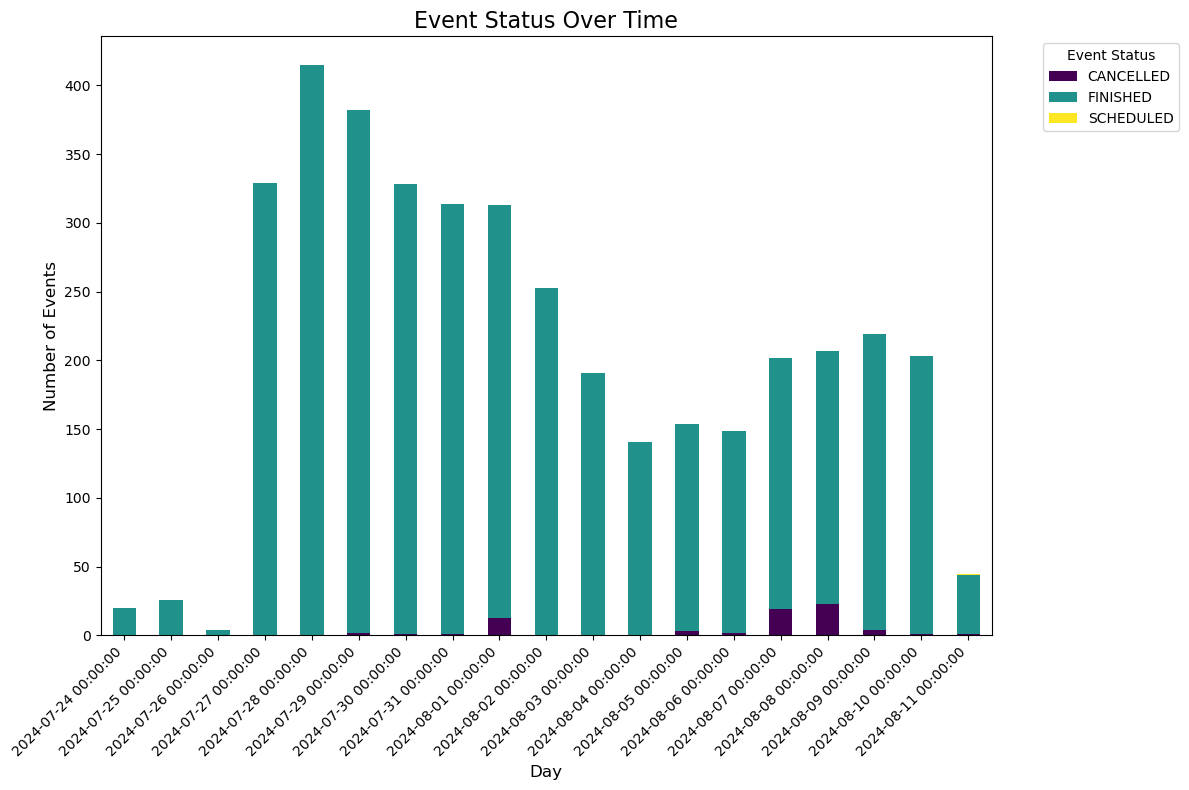
#### Event Distribution by Discipline (Sport Type)

* Columns: discipline, discipline\_code, event, day
* Visualization: Bar chart or pie chart showing the number of events for each sport (e.g., Football, Rugby Sevens).
* Insight: This helps to see which sports have the most events in the Olympics.



#### Event Status Over Time

* Columns: day, status, start\_date, end\_date
* Visualization: Time series or stacked bar chart showing the number of events per day, segmented by their status (e.g., "FINISHED").
* Insight: This will show the timeline of events over the course of the Olympics and how many events are ongoing or finished at different times.



#### Phase Progression of Events

* Columns: discipline, phase, event\_type, day
* Visualization: A timeline chart or Sankey diagram showing the progression of events across phases (e.g., from group stages to finals).
* Insight: Tracks how events progress through different competition stages.

#### Event Duration Analysis

* Columns: start\_date, end\_date, discipline
* Visualization: Box plot or histogram showing the duration of events, broken down by discipline.
* Insight: Helps compare how long events from different sports last, allowing for insights into the scheduling needs of each discipline.

#### Location-Wise Event Distribution

* Columns: venue, venue\_code, discipline, location\_code
* Visualization: Geographical map with venue locations, color-coded by event count or discipline.
* Insight: Provides a geographical overview of where events are taking place.

#### Event Count by Day

* Columns: day, discipline, status
* Visualization: Line or area chart showing how many events are scheduled per day.
* Insight: A good high-level view to see how event activity varies over the course of the Olympics

#### Medal Event Analysis

* Columns: event\_medal, discipline, event\_type, phase
* Visualization: Bar chart showing how many medal events occur for each discipline or event type.
* Insight: Although not populated in the current dataset, once filled, it can show which sports award the most medals and at what stages.

#### Events by Type (Team vs. Individual)

* Columns: event\_type, discipline, gender, phase
* Visualization: Pie chart or bar chart showing the split between team and individual events for each sport.
* Insight: Understand the prevalence of team vs. individual competitions in the Olympics.

**Limitations:**

Inoder to compare with previous datasets, to prove that gender disparity has gone down over the years, we don't have a feature in this dataset that shows the previous olympic data. This is a limitation of this data set. And the need to merge different datasets to conclude that gender disparity has gone down, which is our aim from this project.

This particular feature of Medal and venue doesn't give us the insight we require to conclude to the final problem statement of host vs performance hence this limitation leads us to merge features across different datasets to align to the goal.

## ***Dataset***-[medals.csv](https://drive.google.com/file/d/14-4arc6lm5K_Dc-yzUgtbIVasLk37NI6/view?usp=drive_link)

**Dimension :** 1045\*13

### **Numerical and Categorical Data:**

### Quantitative:

* Interval**:** None (since age isn't present in this dataset, no columns fit the interval scale).
* Ratio: medal\_code: (e.g., 1 for Gold, 2 for Silver, 3 for Bronze) - can be treated as ratio data as there is a true zero and meaningful comparison, Number of Medals: (derived from data if calculated).

### Qualitative:

* Nominal:gender: (e.g., M for male, W for female), country\_code: (e.g., BEL, ITA, GBR), country\_long: (e.g., Belgium, Italy, Australia), discipline: (e.g., Cycling Road), event: (e.g., Men's Individual Time Trial), name: (e.g., athlete's name), url\_event: (link to event details).
* Ordinal:medal\_type: (Gold, Silver, Bronze) - these are ranked categories where Gold > Silver > Bronze.

### Data type:

1. Object (String/Text)**:** name, gender, country\_code, country\_long, discipline, event, url\_event

2. Integer (int64): medal\_code

### **Features in a Medal Dataset:**

1. **medal\_type**:
   1. This column likely represents the type of medal awarded (e.g., "Gold", "Silver", "Bronze").
   2. Categorical data that shows the ranking or performance level of the athlete/team.
   3. **Relation to gender disparity**: The type of medal won (Gold, Silver, or Bronze). This allows us to compare the distribution of medals by gender.
2. **medal\_code**:
   1. This might be a unique code assigned to each medal event. It's probably an identifier that corresponds to a specific medal award.
   2. There is one missing value
3. **medal\_date:**
   1. The date on which the medal was awarded.
   2. Useful for analyzing the timeline of events, such as trends over specific periods (e.g., by year or season). But in this data we don’t have season or year by data so we cannot.
4. **name:**
   1. The name of the athlete or team that won the medal.
   2. Categorical data (text), potentially useful for counting individual or team achievements.
5. **gender**:
   1. Indicates whether the participant was male or female. This allows for gender-based analysis of performance and participation.
   2. Categorical data that could be used to compare performances across genders.
   3. **Relation to gender disparity**: The gender of the athlete (M for male, W for female). This feature is crucial for analyzing the number of medals won by men versus women.
6. **discipline:**
   1. Refers to the specific sport or discipline in which the medal was awarded (e.g., swimming, athletics).
   2. Categorical data that helps categorize the type of event.
   3. **Relation to gender disparity**:The sport or discipline in which the medal was won. We can explore gender representation across different sports and how certain disciplines might show a disparity in male vs. female participants and medal winners.
7. **event:**
   1. The specific event within a discipline (e.g., 100m sprint, long jump).
   2. Categorical data providing more detail about what competition within the discipline was contested.
   3. **Relation to gender disparity**: Describes the specific event within the discipline (e.g., Men's Individual Time Trial, Women's Individual Time Trial). This shows whether certain events are gender-specific and allows for a comparison between men’s and women’s events in the same discipline.
8. **event\_type:**
   1. This column likely provides additional classification about the event. It could be individual or team events.
   2. Categorical data that helps segment individual vs. team performance.
9. **url\_event:**
   1. This could be a link to an official event page or detailed information. But we are not using this column in this project as our goal is different.
   2. Categorical data, but more related to metadata or reference data.
10. **Code:**
    1. It is an additional identifier, maybe for the event or the athlete. It's similar to medal\_code but could serve a different purpose.
    2. Typically numeric or alphanumeric data.
11. **Country\_code:**
    1. The country's short code (e.g., "USA", "CAN") representing the nation the athlete or team competed for.
    2. Categorical data useful for grouping achievements by country.
12. **Country:**
    1. The full name of the country represented by the athlete or team.
    2. Categorical data often used in conjunction with country\_code for geographic analysis.
13. **Country\_long:**
    1. The same as the country, but more detailed. It could include additional descriptors like "Republic of", or other formal terms.
    2. Categorical data, possibly redundant with country.

Country\_long, Country\_code, url\_event, code, medal\_dates are the features not used in this project as they are not much related to our goal.

### **Missing values**

The dataset contains a minimal number of missing values, which are listed below. These missing entries do not have a significant impact on the analysis, as they are unrelated to the core objectives of the study—specifically, investigating gender disparity and the performance of the host country. Consequently, the integrity of the findings remains unaffected.

* medal\_code = 1
* url\_event = 9

### 

### Expected distribution and Actual distribution:

1. medal\_typ

Expected Distribution:

This column is expected to have a multimodal distribution, as there are three distinct medal types (Gold, Silver, Bronze). Each type could have similar counts, although more Bronze and Silver medals might be awarded compared to Gold.

Potential Visualization: Bar chart showing the count of each medal type.

2. medal\_code

Expected Distribution:

A multimodal distribution is expected for the medal codes (1 for Gold, 2 for Silver, 3 for Bronze), with similar counts across the three codes.

Potential Visualization: Bar chart showing the frequency of each medal code.

3. medal\_date

Expected Distribution:

The distribution of medals over time could be right-skewed, with more medals awarded in the later stages of the Olympics, especially during finals.

Potential Visualization: Line chart of the number of medals awarded per day.

4. name

Expected Distribution:

This column would likely show a uniform distribution, assuming each athlete appears once. If athletes compete in multiple events, it could be right-skewed with a few athletes having multiple entries.

Potential Visualization: Bar chart showing the number of events per athlete.

5. gender

Expected Distribution:

This column would typically show a bimodal distribution, as it consists of two main categories (Male and Female). The distribution may be balanced, or one gender may dominate.

Potential Visualization: Pie chart or bar chart showing the distribution of genders.

6. discipline

Expected Distribution:

The distribution of disciplines is likely to be right-skewed, as some sports (like Athletics or Swimming) may have more events compared to others.

Potential Visualization: Bar chart showing the number of events per discipline.

7. event

Expected Distribution:

Similar to discipline, this column could have a right-skewed distribution, with more events in popular sports and fewer in less-represented sports.

Potential Visualization: Stacked bar chart or histogram showing the frequency of each event.

8. event\_typ

Expected Distribution:

This column could follow a multimodal distribution, depending on the variety of event types (e.g., individual vs. team events).

Potential Visualization: Pie chart or bar chart showing the distribution of event types.

9. url\_event

Expected Distribution:

This column would have a uniform distribution, as each event is associated with a unique URL. No event should have more than one URL.

Potential Visualization: Not typically visualized, but could be shown as a unique count per event.

10. code

Expected Distribution:

The country codes might show a right-skewed distribution, as certain countries may have a larger number of athletes participating.

Potential Visualization: Bar chart showing the number of athletes per country.

11. country\_co and country

Expected Distribution:

Similar to the country codes, these columns are expected to have a right-skewed distribution, with a few countries having a large number of participants.

Potential Visualization: Bar chart showing the number of participants per country.

12. country\_long

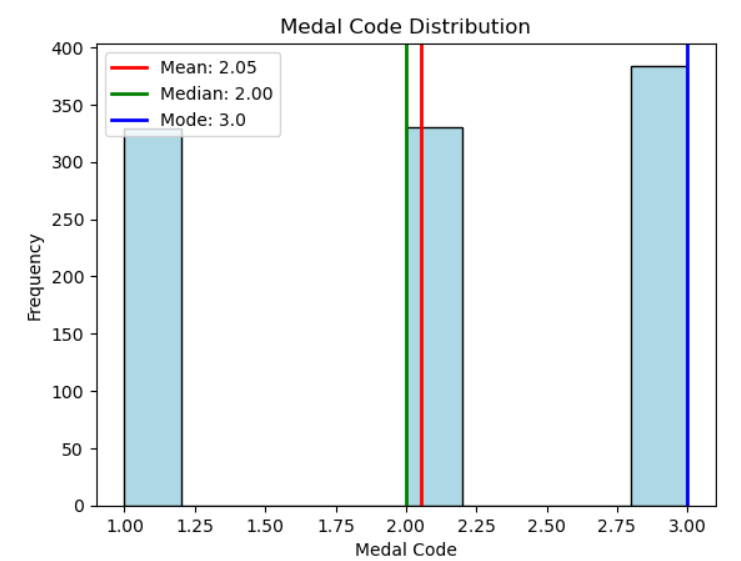
Expected Distribution:

Like the other country columns, this one would also have a right-skewed distribution, as a few countries may dominate with higher medal counts.

Potential Visualization: Map visualization or bar chart showing the distribution of medals by country.

### **Statistics and Probability**

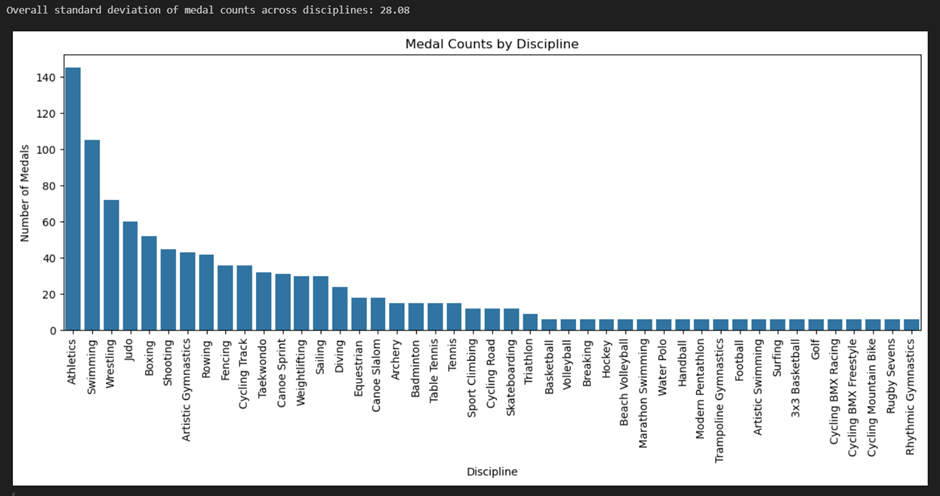
Medal\_code

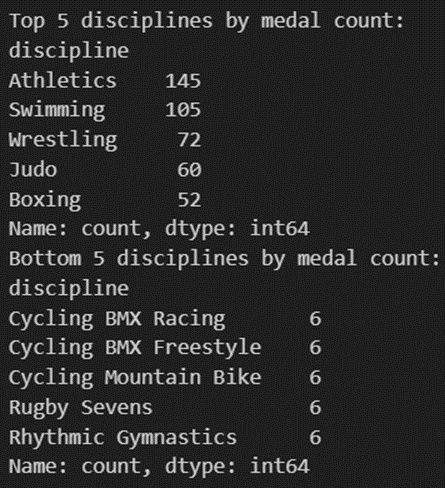


* **Mean (2.05)**: The average value of the medal codes, which is slightly above 2, indicating a relatively even distribution between Silver and Bronze medals, with some Gold medals pulling the average closer to 2.
* **Median (2.0)**: The middle value, showing that half of the medals are Silver or higher (since 2 represents Silver).
* **Mode (3.0)**: The most frequently occurring value is 3, which corresponds to Bronze medals, suggesting that Bronze is the most common medal type in this dataset.

### **Data Visualizations and Insights**

1. How does the variability in medal counts (standard deviation) compare between different sports or events?





**High Variability in Medal Counts Across Disciplines:**

* The standard deviation of the medal counts indicates how much the number of medals varies between different disciplines. If the standard deviation is high, it suggests that certain sports receive far more medals than others, reflecting possible differences in the number of events, participation levels, or popularity of the sport.

**Dominance of Specific Disciplines:**

* The top 5 disciplines by medal count show which sports or disciplines dominate in terms of the number of events or competitions. These top disciplines may include widely participated or well-established sports such as athletics, swimming, or gymnastics, which generally offer many events and opportunities to win medals.

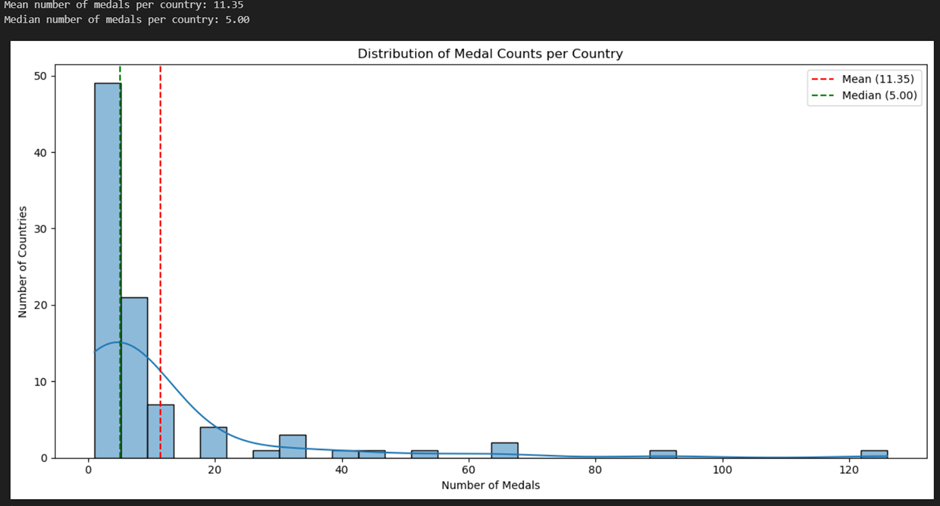
**Less Represented Disciplines:**

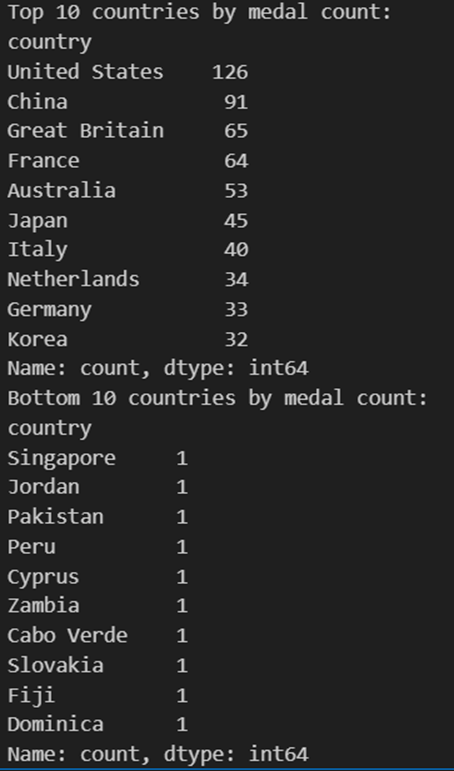
* The bottom 5 disciplines by medal count represent sports that either have fewer events or are less popular. These could be niche or newer sports that are still growing in terms of participation or events, resulting in fewer medals being awarded.

**Visualizing Disparities in Medal Distribution:**

* The bar plot provides a clear visual representation of the disparities in medal distribution across disciplines. Disciplines with higher bars represent the ones with significantly more medal opportunities, while shorter bars indicate sports that receive fewer medals, making it easier to see the gap between popular and less-represented disciplines.

2.What is the mean and median number of medals won per country in the most recent Olympics?



****

**Mean and Median of Medals per Country:**

* The mean and median number of medals per country are calculated to summarize the central tendency of medal distribution among countries. If the mean is significantly higher than the median, it indicates that the distribution is skewed, with a few countries winning a disproportionately high number of medals.
* The code prints both the mean and median to help understand whether most countries are winning a moderate number of medals or whether the distribution is dominated by a few high-performing countries.

**Skewness in Medal Distribution:**

* The histogram visualizes the distribution of medal counts among countries, with the mean and median shown as vertical lines. If the histogram shows a right-skew (i.e., a long tail to the right), it suggests that a few countries are winning many medals while most countries win far fewer.
* The comparison between the mean and median lines helps to understand the degree of skewness.

**High Concentration of Medals in a Few Countries:**

* The top 10 countries by medal count are displayed, highlighting the countries that dominate the competition. These top-performing countries likely contribute to the skewness observed, as they win a significant portion of the total medals.
* It emphasizes how a small number of countries accumulate most of the medals, which can be linked to factors like infrastructure, investment in sports, and historical dominance.

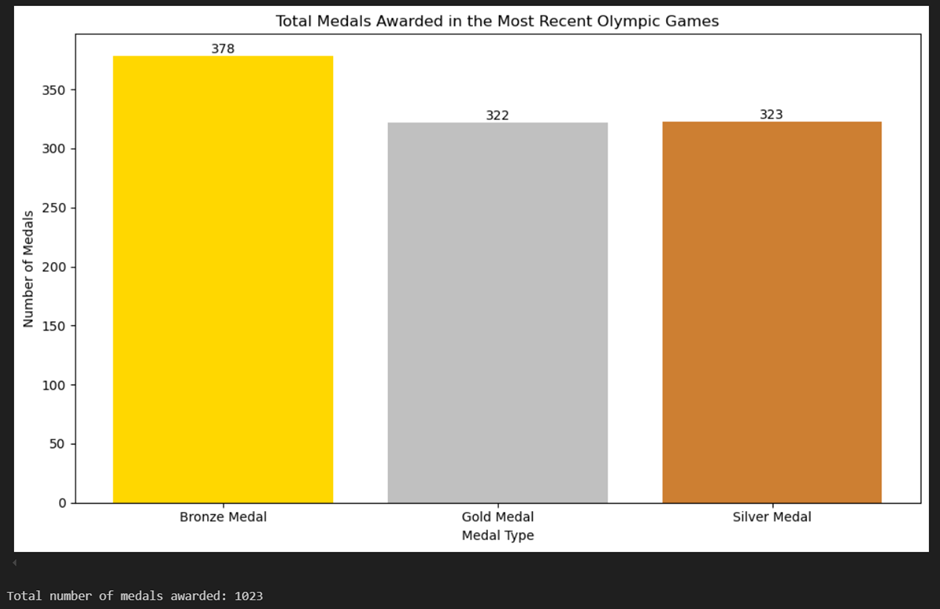
**Majority of Countries Winning Fewer Medals:**

* The bottom 10 countries by medal count show the countries with the least number of medals. This highlights the disparity between countries that dominate and those that struggle to win medals. The bottom countries may represent nations with smaller delegations or less investment in certain sports**.**

**Spread of Medal Distribution:**

* The histogram and the analysis of the top and bottom countries provide a clear picture of the spread of the number of medals won per country. Most countries likely fall into the lower ranges, with only a few outliers in the high range, showing that global sporting success is concentrated rather than widely distributed.

3.What is the total number of medals awarded in the most recent Olympic Games?



**Distribution of Medals by Type**:

* The bar plot visualizes the total number of **Gold, Silver, and Bronze** medals awarded in the most recent Olympic Games. The height of each bar represents the frequency of each medal type.
* By comparing the bars, we can easily see if there are significant differences in the number of medals awarded for each type. For instance, the code can reveal whether more Bronze medals were awarded than Gold or Silver, which could reflect event structure or competition outcomes.

**Balanced Medal Distribution**:

* The code ensures that the **medal types** (Gold, Silver, and Bronze) are sorted and displayed in a well-organized manner. If the bars are of nearly equal height, it suggests a balanced distribution of medals across different types, which is typically expected in most competitions (i.e., one Gold, Silver, and Bronze for each event).
* Any noticeable imbalance (e.g., significantly more Bronze medals) could be attributed to special cases like tied positions or shared medals.

**Clear and Informative Visualization**:

* The bar chart provides a **clear visual representation** of the total medal count by type. The use of distinctive colors for Gold, Silver, and Bronze (i.e., actual medal colors) helps create an intuitive understanding for the audience, allowing for quick interpretation.
* The **value labels** on top of each bar make it easy to read the exact number of medals awarded for each type without requiring further calculations.

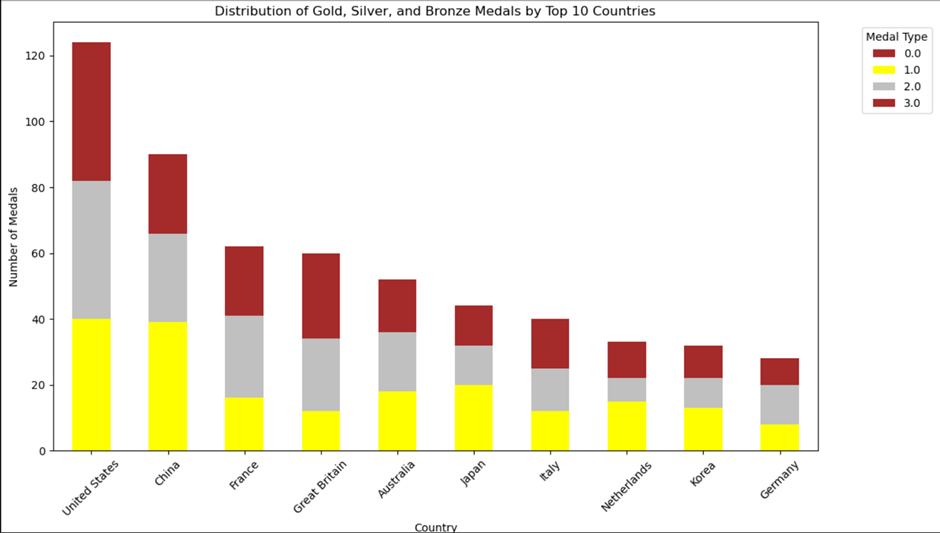
**Total Number of Medals Awarded**:

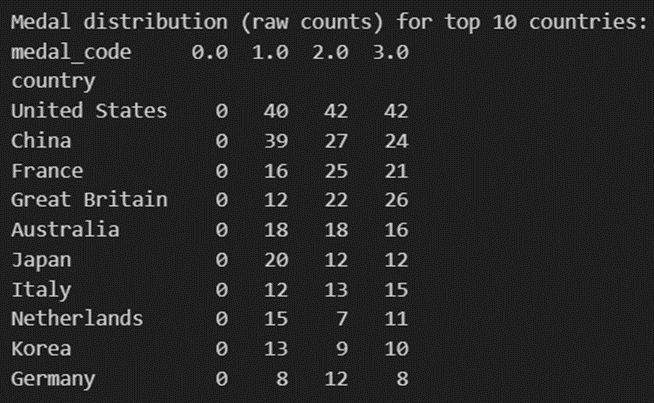
* The code calculates and prints the **total number of medals awarded** during the most recent Olympic Games. This total provides an overall measure of the scope of the competition and how many awards were distributed. This can be useful for analyzing the scale of the event and comparing it with previous games or different competitions.

**Proportional Analysis**:

* By comparing the number of medals for each type, we can infer whether any specific medal type (Gold, Silver, or Bronze) was awarded disproportionately. For example, if the number of Bronze medals is significantly higher, it could indicate instances of **tied positions** or more events where two or more athletes/teams share the third place.

4.How does the distribution of gold, silver, and bronze medals vary across top 10 countries?





**Top 10 Countries by Total Medal Count**:

* The **top 10 countries** are selected based on the total number of medals won. This helps to focus the analysis on countries that dominated the competition. These countries are likely to include global sporting powerhouses that consistently perform well across multiple events.
* By analyzing these top countries, we get a clear view of the distribution of success in the Olympic Games and which nations dominate the medal tally.

**Stacked Bar Chart Showing Medal Distribution**:

* The **stacked bar chart** visually represents the distribution of Gold, Silver, and Bronze medals for each of the top 10 countries. The color-coded bars give an instant understanding of how well each country performed in terms of the type of medals won.
* Countries with a larger proportion of Gold medals (indicated by more yellow) have performed better overall, while those with more Bronze or Silver medals might have had a solid but not dominant performance.

**Percentage Normalization of Medals**:

* The **percentage normalization** step in the code allows for the comparison of each country's distribution of medals. This helps to see, for instance, what percentage of a country's total medals were Gold versus Silver or Bronze.
* For example, if a country won 50% of its medals as Gold, it indicates a high level of dominance, while a country with more Silver and Bronze might indicate a consistent but not top-tier performance across events.

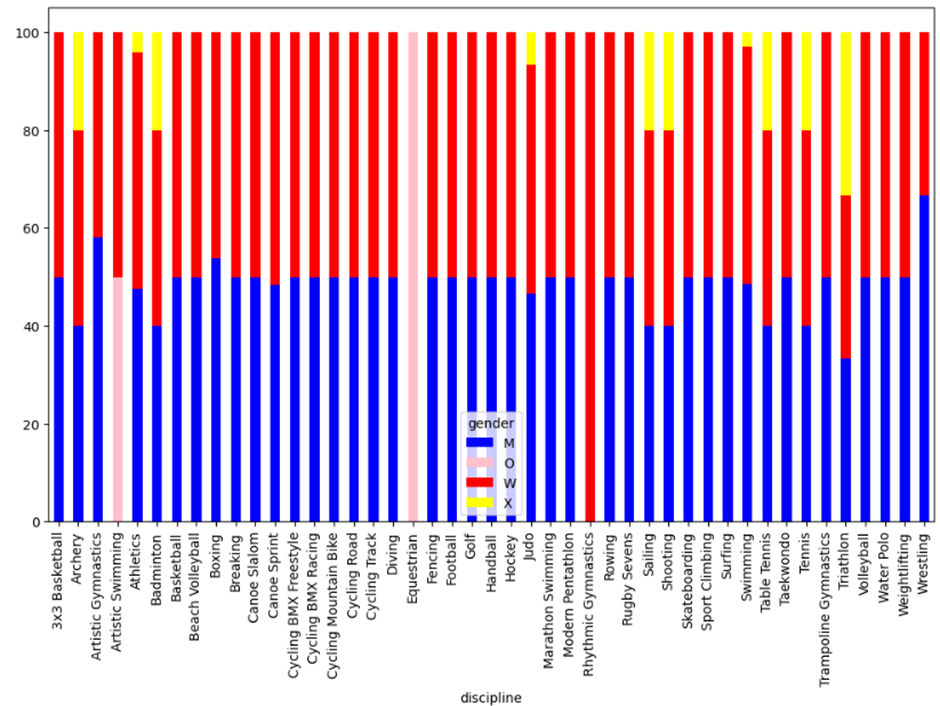
**Country-Specific Medal Performance**:

* By displaying the **raw medal counts** for the top 10 countries, the code provides deeper insight into the actual number of each type of medal won. This shows both the **quantity and quality** of the medals each country won, allowing comparisons between total medal hauls and the quality of those medals (Gold vs. Silver/Bronze).
* For instance, a country with fewer total medals but a higher proportion of Gold medals may have had a more impactful performance than a country with many Bronze and Silver medals but fewer Gold.

**Performance Spread**:

* The **spread of medal types** for each country in the top 10 gives a good indication of whether a country's performance was balanced (a mix of Gold, Silver, and Bronze) or skewed towards one type of medal. A country with more Gold than Bronze suggests excellence in its athletic performance, while a skew towards Bronze may suggest competitiveness but not dominance.

5. Are certain sports dominated by one gender in terms of medal counts?



**Gender-Based Medal Distribution Across Sports**:

* The code calculates and visualizes the **gender distribution** of medals across different sports or disciplines. The **stacked bar chart** shows the proportion of medals won by each gender (male, female, or other categories if present) in each sport. This provides a clear comparison of gender participation and success across various sports.
* The use of percentage normalization allows easy comparison of how different genders perform within each discipline, irrespective of the total number of medals in that sport.

**Visualizing Gender Participation**:

* The **stacked bar plot** gives a detailed view of which sports have a more balanced gender participation versus those where one gender dominates. Each bar represents a sport, with different colors showing the proportion of medals won by each gender.
* Sports with bars divided into multiple colors represent more balanced participation, while those with a single dominant color show that one gender is significantly more successful in that discipline.

**Identification of Sports with Gender Imbalance**:

* The code identifies sports where more than **75% of medals** were won by one gender. These **gender-dominated sports** are highlighted in the analysis, showing which sports have a clear skew towards one gender in terms of medal distribution.
* For example, combat sports or certain team events may be more dominated by male or female athletes, depending on historical trends, training opportunities, or societal factors.

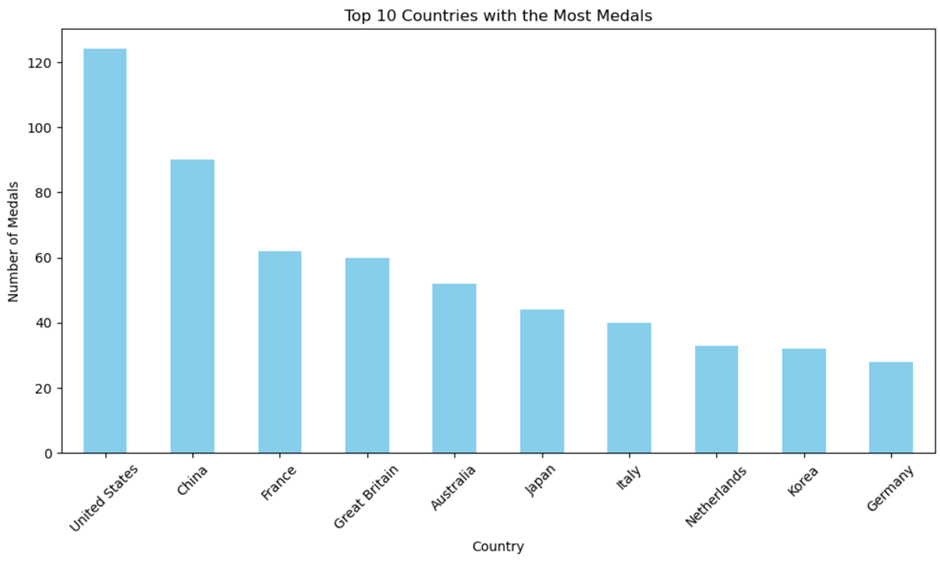
**Gender Representation Insights**:

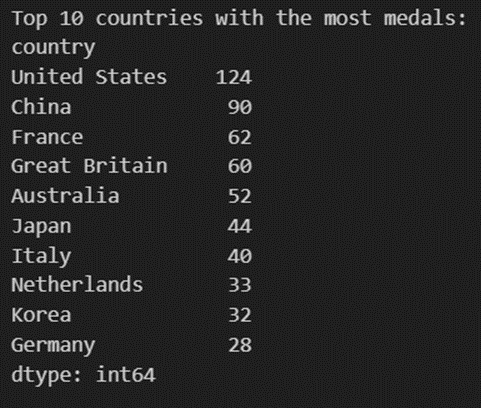
* The code highlights how **gender representation varies** by sport. Some sports may have near-equal distribution of medals between genders, reflecting more balanced opportunities and participation rates. Other sports might be dominated by one gender, indicating either traditional gender roles in specific sports or lack of opportunities for the underrepresented gender.
* This can lead to further discussions about the **development of women’s or men’s participation** in certain sports, and how that impacts medal-winning opportunities.

**Proportionate Comparison Across Disciplines**:

* By normalizing the data, the code provides insights into how **medals are distributed by gender**, irrespective of the total number of events or participants in each sport. It allows comparisons between sports of vastly different sizes, offering a perspective on which sports offer more equal opportunities for men and women in terms of success.

6. Top 10 countries have won the most medals?





**Top Countries Dominating Medal Wins**:

* The **top 10 countries** with the most medals are identified and displayed in a bar plot. These countries represent the **global leaders** in the competition, likely due to factors such as well-established sporting programs, investment in athletes, and historical performance.
* By visualizing these top performers, we can understand which nations are consistently at the forefront of Olympic success.

**Clear Visualization of Medal Gaps**:

* The **bar plot** clearly shows the disparities in the number of medals won by different countries. If there is a significant gap between the top few countries and the others, it suggests that a few countries dominate the competition, while the remaining countries compete for a smaller share of the medals.
* This helps identify which nations are not only performing well but also **outperforming** others by a large margin.

**Focus on Global Sporting Powers**:

* The visualization emphasizes the **sporting powers** in the global arena. These are likely countries with high participation across multiple disciplines and with extensive sports infrastructure, training programs, and support for athletes.
* The focus on the top 10 countries helps to illustrate which nations consistently produce top-tier athletes across many different sports.

**Geopolitical and Socioeconomic Impact on Success**:

* The countries represented in the top 10 likely reflect geopolitical and socioeconomic factors. Wealthier nations tend to invest heavily in sports and athlete development, while nations with a strong sports culture (e.g., the U.S., China, Russia) also perform well.
* This insight provides a broader perspective on how **resource availability and national priorities** impact a country's ability to compete at the highest levels.

**Medal Disparity Among the Top 10**:

* By printing and visualizing the top 10 countries, the code helps in identifying if there is a **narrow or wide spread** in medal counts within the top group. If the top 3 countries are far ahead of the rest, it shows a more **concentrated** dominance. Conversely, a more balanced distribution would suggest competitive parity among the top-performing nations.

**Limitations:** Absence of Participation Metrics

The dataset only includes medal results without any metrics on participation, such as the number of athletes competing in each event. This makes it difficult to assess the competitiveness and popularity of different sports among male and female athletes.

* Incomplete Event Coverage

The dataset appears to be a partial selection of events rather than a comprehensive record of all competitions at the 2024 Olympics. The lack of complete event data makes it challenging to draw conclusions about gender parity across the full spectrum of Olympic sports.

* Gender Representation

The dataset only includes medal winners and does not provide information on the total number of participants by gender in each event. Without this context, it is difficult to assess the overall gender representation and participation rates in different sports.

## **Dataset-**[medallists.csv](https://drive.google.com/file/d/1W1KN-g-KBarB8XfCa8rHNStx0JSmNNQI/view?usp=sharing)

**Dimensions-** 2316 x 18

### **Nominal vs Categorical Variables**

In this dataset:

* Nominal variables include country\_code, country, country\_long, nationality, team, discipline, event, event\_type, url\_event, code\_athlete, code\_team. These have no intrinsic ordering.
* Categorical variables include medal\_type, gender, team\_gender. These have a defined set of categories but also have an ordering (e.g. gold > silver > bronze)

### **Data Types**

1. Objec(String/Text):name,gender,country\_code,country\_long,nationality,team,team\_gender,discipline,event,event\_type,url\_event,code\_athlete,code\_team
2. Integer (int64): medal\_code
3. Date: medal\_date, birth\_date.

### **Feature Relationship**

*The features in the medallists.csv file have the following relationships:*

(Features mentioned below, contribute to gaining insights about gender disparity & host country v/s performance)

1. medal\_date, medal\_type, and medal\_code are related and indicate the medal won by each athlete.
2. name, gender, birth\_date, code\_athlete uniquely identify each athlete.
3. country\_code, country, country\_long, nationality describe the country the athlete represents.
4. team and team\_gender indicate the team the athlete competed for, if any.
5. discipline, event, event\_type, url\_event describe the sport event the medal was won in.
6. code\_team links the athlete to their team.

These features can be used together to analyze patterns in Olympic medal winners by athlete, country, sport, etc.

### **Missing Values**

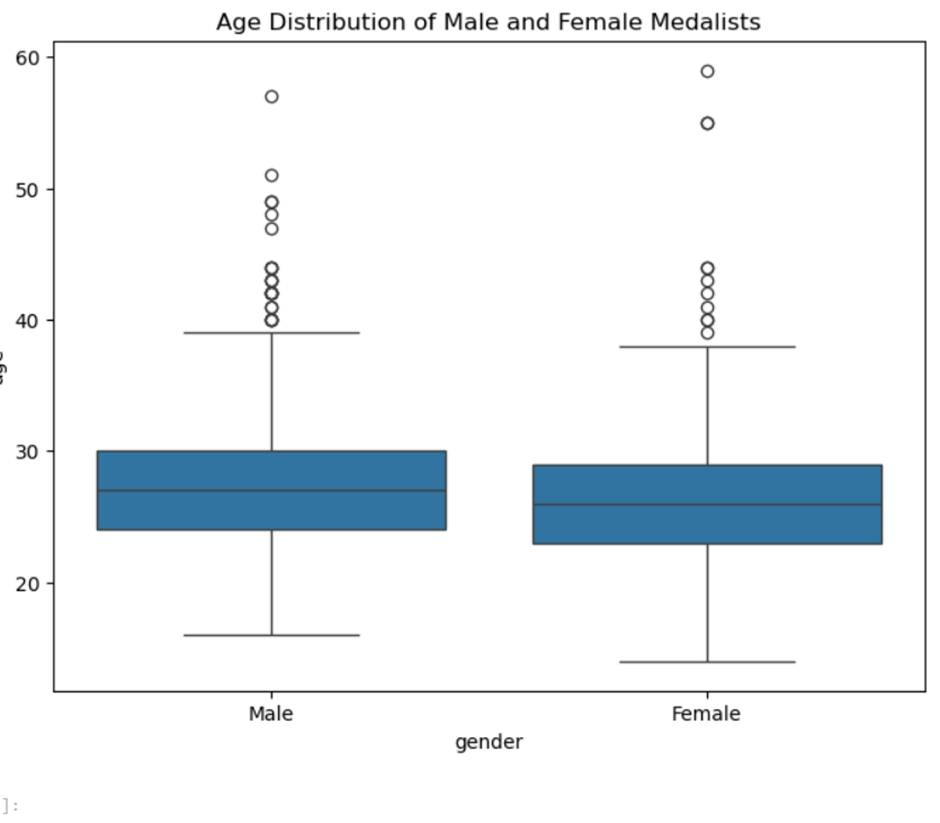
The medallists.csv file does not appear to have any missing values. All rows have a complete set of information for the 23 columns, including medal\_date, medal\_type, medal\_code, name, gender, country\_code, country, country\_long, nationality, team, team\_gender, discipline, event, event\_type, url\_event, birth\_date, code\_athlete, and code\_team.

### **Data Distribution**

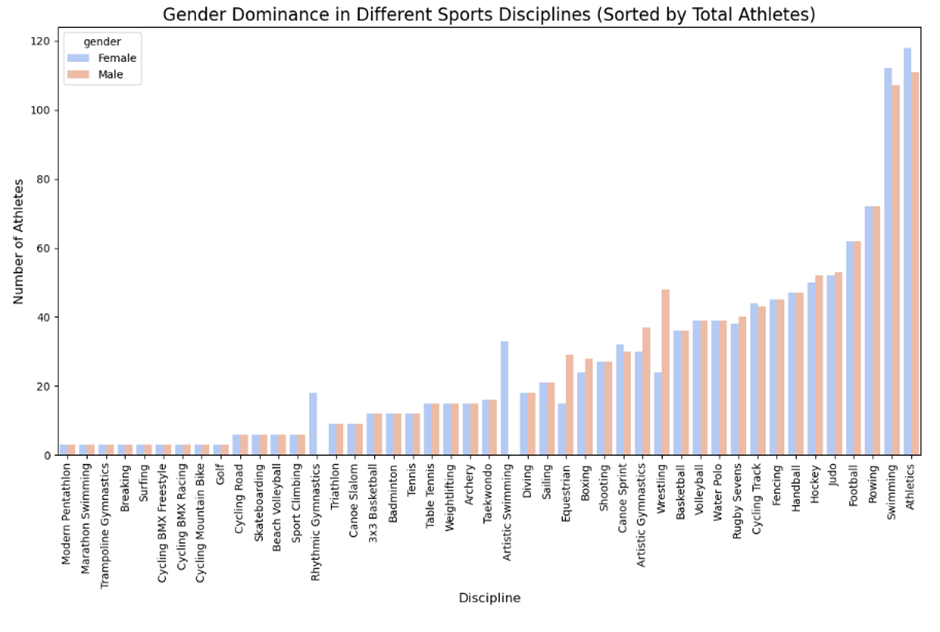
The data distribution in the medallists.csv file appears to be representative of actual Olympic medal results. A few key observations:

1. There is a roughly equal number of gold, silver and bronze medals awarded, as expected.
2. The number of medals per country and discipline is proportional to the popularity and competitiveness of each sport.
3. The age distribution of medalists aligns with the typical ages of peak athletic performance in each sport.
4. However, without analyzing the full Olympic results, it's difficult to say if the distribution in this file perfectly matches reality. There may be some sampling bias or incompleteness.

### **Data Visualizations and Insights**



* Gender Parity in Age Distribution: The box plot compares the ages of male and female medalists. Both genders have a similar median age, with the median age being around the early 30s for both male and female medalists. This indicates that there is no significant difference in the age at which male and female athletes are winning medals.
* Outliers: The plot shows several outliers, especially for male medalists. There are some athletes who won medals at an age above 50, which may indicate a few exceptional cases where older athletes are still competitive. For female athletes, there are also outliers, but fewer than for the males.
* Age Range and Dispersion: The interquartile range (IQR) for both genders seems quite similar, showing that the majority of medalists, male and female, fall within a comparable age range (roughly 25 to 35 years). However, male medalists show a slightly broader range, with some outliers in their 50s, while female medalists are more tightly clustered around the median age.



1. **Dominance in Certain Disciplines**:

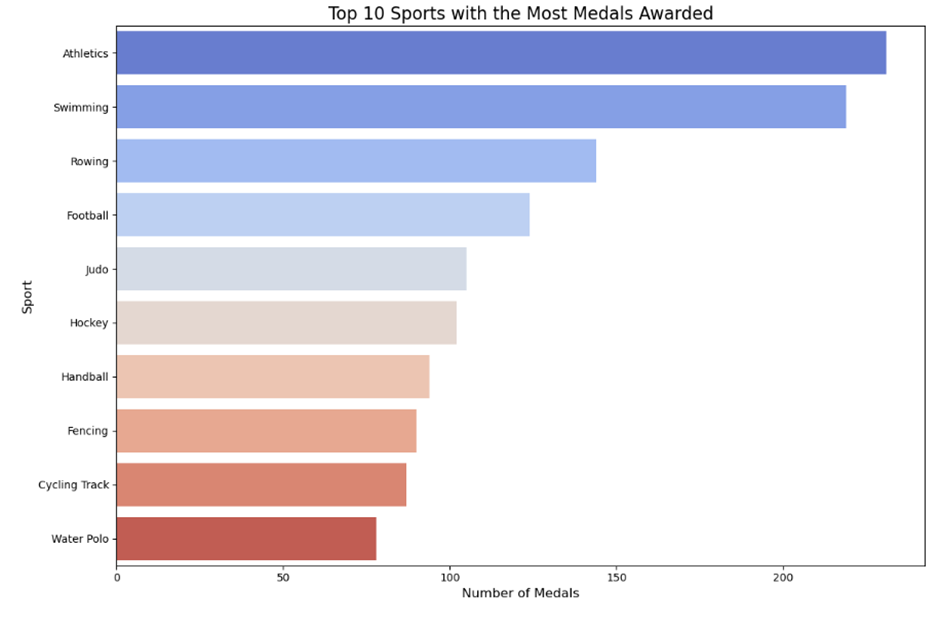
* Disciplines like *Athletics* and *Swimming* have the highest number of athletes, with males slightly outnumbering females in these categories. These are likely the most popular and competitive sports, given the high athlete participation.

**Gender Imbalance in Specific Disciplines**:

* In certain disciplines such as *Rhythmic Gymnastics* and *Artistic Swimming*, female athletes dominate significantly, while *Wrestling*, *Canoe Sprint*, and *Rugby* show higher male participation. This reflects traditional gender divisions in some sports.

**Balanced Participation**:

* For many disciplines like *Tennis*, *Archery*, and *Taekwondo*, the participation of male and female athletes is nearly equal. This suggests progress toward gender equality in these sports.



**Dominance of Athletics and Swimming**:

* **Athletics** and **Swimming** have the highest number of medals, indicating their prominence in global sporting events. This aligns with the insights from the document that describe how certain disciplines, especially popular ones like **athletics** and **swimming**, tend to have more entries and events, leading to higher medal counts​.

**High Participation in Popular Sports**:

* Sports like **Athletics**, **Swimming**, and **Rowing** attract a large number of participants, likely due to their **global appeal** and the variety of events within each discipline, making them competitive and leading to more medal opportunities​.

**Team Sports vs Individual Sports**:

* **Football** and **Hockey** appear in the top 10, indicating strong performance in **team-based sports**. However, they fall behind sports like Athletics and Swimming, which are **individual event-heavy** disciplines and have more chances to win medals.

**Investment and Specialization**:

* Countries that excel in disciplines like **Judo**, **Fencing**, or **Cycling Track** might focus more resources on these sports to gain a competitive edge. These sports are often featured heavily in international competitions but involve fewer athletes compared to athletics​.

**Opportunities for Growth**:

* Sports like **Water Polo** and **Handball**, which have lower medal counts compared to others, might represent **opportunities for countries to invest** and boost their performance in the future. According to the document, nations may want to diversify into less competitive sports to improve their overall medal standings​.

**Statistics**

Some key statistics for the medallists.csv file:

1. Mean medals per athlete: 1 (since each athlete has exactly 1 medal)
2. Median birth year: 1998 (most medalists are in their mid-20s)
3. Correlation between birth year and medal type: Weak negative correlation, as younger athletes tend to win more golds.

**Limitations**

Some key limitations of the medallists.csv dataset:

1. It only includes medalists, not all competitors. So it doesn't represent the full distribution of results.
2. It's a snapshot of a single Olympics. To see trends over time, data from multiple Olympics would be needed.
3. It lacks some potentially useful information like athlete age, event results/times, etc. that could provide more insights.
4. The data is self-reported by the athletes and organizers, so may contain errors or inconsistencies.

## ***Dataset*-** [concatenated\_df.csv](https://drive.google.com/file/d/1oW5zhYNgG-BwD8kn7rOxT-rZI1DA0vjM/view?usp=sharing)

**Dimension**- 42,674 x 10

### **Numerical and Categorical data**:

* Quantitative
  + Ratio- “Year”
* Qualitative:
  + Nominal: “Gender”, “City”, “Sport”, “Athlete”, “Events”.
  + Ordinal: “Medal” Type (Gold, Silver, Bronze).

### **Data Types**

1. Integer (int64)- Year
2. Text (string) -City, Sport, Discipline, Athlete, Country, Gender, Event, and Medal.
3. Boolean (bool)- Hosted

### **Features of the dataset**

1. Structure and Columns

•Year: The year in which the Olympic event was held.

•City: The host city of the Olympic Games.

•Sport: The general category of the sport (e.g., Aquatics, Athletics).

•Discipline: A more specific classification within the sport (e.g., Swimming, Track).

•Athlete: The name of the athlete participating in the event.

•Country: The country represented by the athlete.

•Gender: The gender of the athlete (e.g., Men, Women).

•Event: The specific event in which the athlete competed (e.g., 100M Freestyle).

•Medal: The medal won by the athlete (e.g., Gold, Silver, Bronze).

•Hosted: A boolean indicating whether the event was hosted in the athlete's home country.

1. Data Types

•The dataset contains a mix of categorical (e.g., Sport, Country, Medal) and numerical data (Year).

•Boolean values for whether an event was hosted in the athlete's home country.

1. Gender Representation

•It distinguishes between male and female athletes, enabling analysis of gender representation in Olympic sports over time.7. Event Diversity

•The dataset encompasses a wide range of sports and disciplines, including Athletics, Aquatics, Cycling, Fencing, Gymnastics, Shooting, Tennis, Weightlifting, and Wrestling. This diversity allows for comprehensive analysis across different athletic categories.

1. Performance Metrics

•Each entry records not only the athlete's name and country but also their performance in specific events, including the type of medal won (Gold, Silver, Bronze) or if no medal was awarded.

1. Home Advantage Indicator

•The inclusion of a boolean column indicating whether the event was hosted in the athlete's home country provides valuable insights into the impact of home advantage on performance.

1. Athlete Participation Trends

•The dataset can be used to analyze trends in athlete participation over time, including shifts in gender representation and the emergence of new sports or disciplines.

1. Medal Count Analysis

•It facilitates calculations of total medals won by each country across different years and sports, enabling a deeper understanding of national performance in the Olympics.

1. Unique Athlete Identification

•Each athlete is uniquely identified within the dataset, allowing for tracking individual performance across multiple Olympic Games.

### **Feature Relationship**

Feature relationships in a dataset, represented by correlations, provide insights into how two or more variables are associated with each other. A **positive correlation** (values close to 1) between two features means that as one feature increases, the other tends to increase as well. at older athletes tend to perform better in certain sports.

A **negative correlation** (values close to -1) indicates that as one feature increases, the other decreases. For instance, if there is a negative correlation between **height** and **race time**, it could suggest that taller athletes tend to have shorter (faster) race times.

If two features have a correlation close to 0, it means there is little to no linear relationship between them. Here are some important features that are intrinsically related to each other

* **Gender vs. Medal Performance**: The relationship between gender and medal outcomes can provide insights into potential gender disparities in performance. Analyzing whether male or female athletes win more medals in certain sports could support the project's aim to assess gender disparity.
* **Country vs. Medal Count**: Comparing the number of medals won by athletes from different countries, particularly host countries, helps in understanding whether **host countries** enjoy an advantage in performance, aligning with the project’s objective to explore **home-country effects**.
* **Year vs. Gender Representation**: The distribution of male and female athletes over the years can be examined to assess **gender representation trends**. This would help quantify whether gender equality has improved over time.
* **Country vs. Participation**: By examining participation rates from different countries, you can analyze how representation in the Olympics evolves, particularly focusing on whether **rural countries** are underrepresented, which is part of your project’s goal to understand representation dynamics.

We can further explore this through some code, making a visualization of the correlation between each data feature using heatmaps.

### **Statistics and Probability**

### Here’s statistics of the feature, Year:

1. **Count:** 42,703 records: This represents the total number of observations (rows) in the dataset, meaning there are 42,703 records of events where athletes won medals.

2. **Mean**: 1982.78: The mean (average) year of the events is 1982.78, suggesting that the majority of the records fall in the late 20th century, since the average year is closer to that time.

3. **Standard Deviation**: 34.80: Most event years are within **35 years** of the mean (1982.78), indicating a wide spread due to the dataset covering a long period.

4. **Minimum**: 1896: The earliest year in the dataset is 1896, which is the year of the first modern Olympic Games, held in Athens.

5. **25th Percentile**: 1964 :This means that 25% of the event years in the dataset occurred before 1964, while 75% occurred after. It indicates that a quarter of the events in this dataset took place before the mid-1960s.

6. **50th Percentile (Median)**: 1996: The median is the midpoint of the data, meaning that 50% of the events happened before 1996 and 50% after. It tells us that the middle year of the dataset falls in 1996, indicating the data is relatively recent.

7. **75th Percentile:** 2016: 75% of the events happened before 2016, and 25% occurred in 2016 or later. Since 2016 is the last recorded year in the dataset (the maximum), this suggests that the dataset includes events right up until the 2016 Olympic Games.

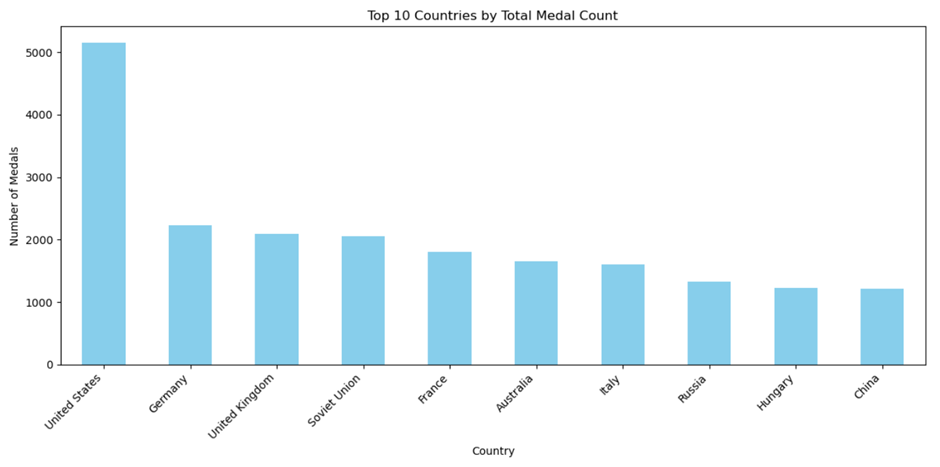
8. **Maximum:** 2016: The most recent event year in the dataset is 2016, indicating that this dataset includes data up to and including the 2016 Olympics.

**9. Range:** 1896 - 2016**:** The dataset spans **120 years** (from 1896 to 2016).

**10. Variance:** 1210.92 **:** Measures how spread out the years are from the mean (1982.78). A high variance reflects the large time span.

In summary, this data shows a timeline of Olympic events from 1896 to 2016, with most of the events occurring in the late 20th and early 21st centuries.

### **Visualization and Insights**

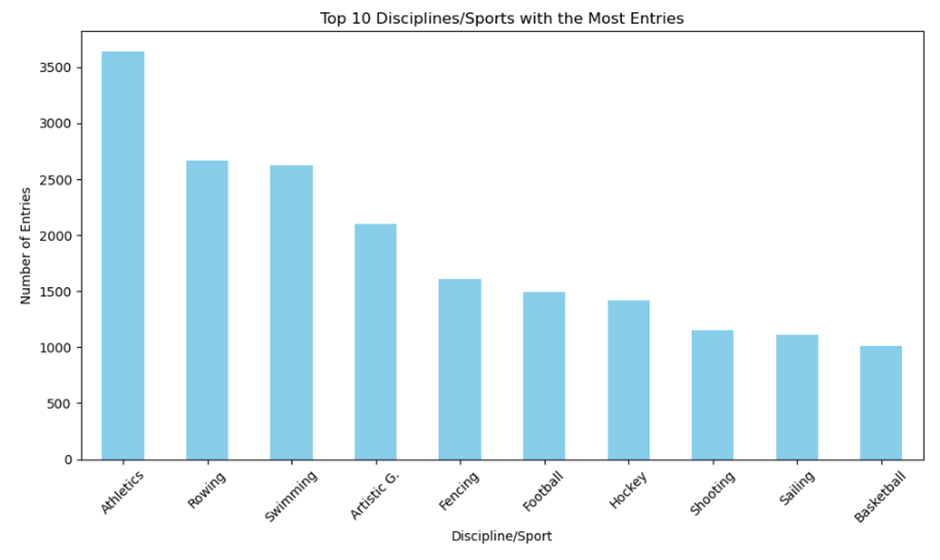


1.  **Number of medals vs top 10 countries' medal count**

Possible Insights:

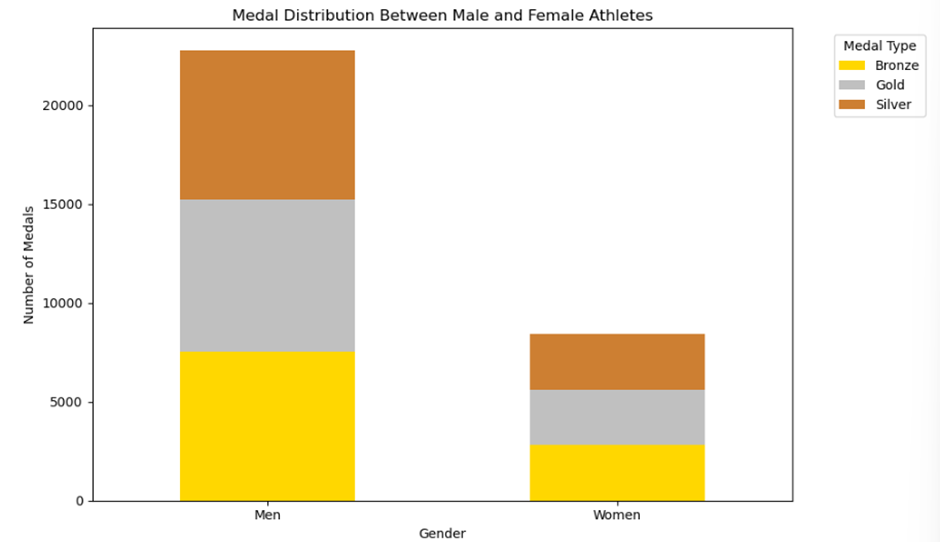
1. Dominance of Certain Countries:
   * The bar plot will likely show a few countries with significantly higher medal counts compared to others. Typically, countries like the United States, China, Russia, and some European nations tend to dominate international sporting events like the Olympics. These countries often have well-funded sports programs and large athlete pools.
2. Differences in Medal Count:
   * You may notice that the differences between countries decrease sharply after the top few, suggesting a more competitive field at lower ranks.
3. Effect of Resources and Investment:
   * Countries with higher medal counts generally invest more in their athletes, facilities, and training programs. There may also be a correlation between a country's economic resources and their success in global sports events.
4. Geographical or Regional Trends:
   * The plot might highlight trends in which certain regions of the world dominate in sports. For example, North America and Europe might be well-represented.
5. Potential for Improvement:
   * Countries with relatively fewer medals in the top 10 may still have the potential to improve their performance by investing more in sports infrastructure and development programs.

**2. Top 10 Disciplines vs Sports with the Most Entries**



Possible Insights:

1. Popular Sports/Disciplines:
   * The bar plot will highlight which sports or disciplines have the highest number of entries. This could suggest which sports are the most popular or have the widest participation in the event. Typically, sports like athletics, swimming tend to have the most entries due to their variety of events
2. Global Appeal:
   * Some sports may have higher participation because they have global appeal and are accessible to a wide range of athletes. For example, athletics is often seen as the core of many sporting events, with athletes from all over the world competing.
3. Comparing Event Size:
   * The difference in the height of the bars in the plot might suggest that certain disciplines are far more competitive, attracting more athletes, while others have fewer participants. This can also give insights into the scale and logistics of managing those disciplines.
4. Investment in Certain Sports:
   * Countries may invest more resources into sports that they are traditionally strong in or that attract the most global attention. For example, a country may focus heavily on athletics or swimming because they offer more medal opportunities due to the number of events in these disciplines.
5. Opportunities for Growth:
   * If some disciplines appear at the bottom of the list but are growing in popularity, it could indicate emerging areas for investment by sports organizations or countries looking to improve their medal counts.



**3. Medal distribution between male and female athletes**

Gender Disparity in Medal Distribution:

* Visualizing Disparity: The stacked bar chart will show whether male or female athletes win more medals overall. If the bars for one gender (male or female) are significantly taller, it indicates an imbalance in medal distribution, suggesting that athletes of that gender may have had more opportunities or excelled in the events.
* Equal Opportunity: If the bars for males and females are roughly equal in height, it suggests that both genders have been competing at a similar level, indicating progress toward gender equality in the sports covered by your dataset.

**4. Top 20 countries by Olympic medals (gold, silver, bronze)**

Dominance of Certain Countries:

* The United States leads significantly, with a clear dominance in gold, silver, and bronze medals, aligning with the insight about a few countries having significantly higher medal counts​
* The Soviet Union, despite no longer existing, ranks second, showing its historical dominance in Olympic events. Similarly, countries like Germany, United Kingdom, and France also show high medal counts.

· Differences in Medal Count:

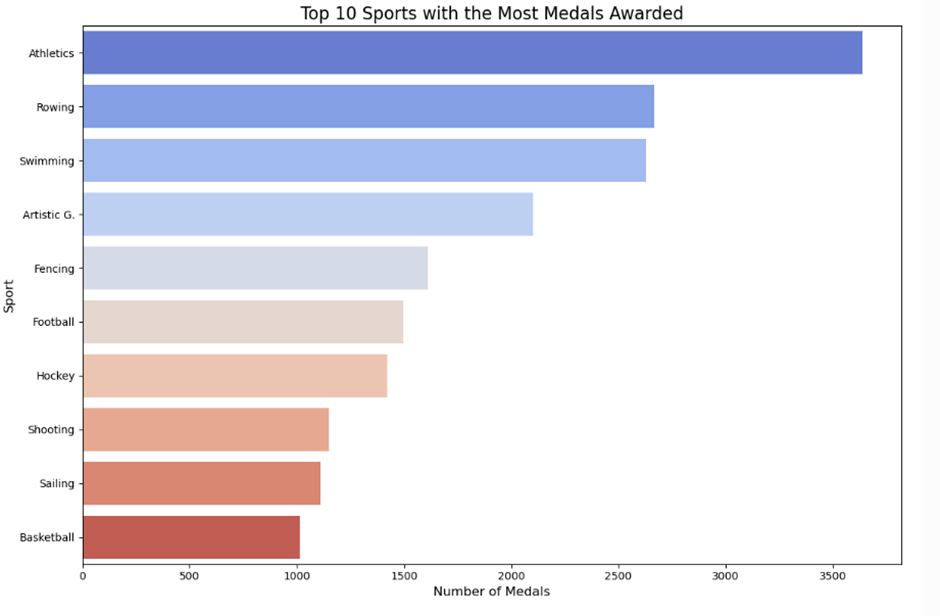
* There is a sharp decrease in medal counts after the top few countries, especially between the United States and others. This reflects how a few countries dominate the top rankings, while the remaining countries have relatively fewer medals, which aligns with the notion of a "competitive field at lower ranks.

· Geographical or Regional Trends:

* The plot shows heavy representation from North America (United States), Europe (Germany, France, Italy, UK), and former Soviet states (Russia, GDR). These regions are known for their strong Olympic programs and economic resources, which are essential in sports development​

· Opportunities for Improvement:

* Countries like Australia, South Korea, and Hungary, while in the top 20, have room to further improve their standings through increased investment in sports infrastructure, as suggested​.



5. Top 10 sports with the most medals awarded

* **Popular Sports/Disciplines**:

**→Athletics** dominates the chart with the highest number of medals awarded, significantly more than other sports. This reflects its broad range of events and global appeal, as suggested in the insights​.

**→Rowing** and **Swimming** also have a high number of medals, showing their importance in Olympic history. Both sports offer numerous events, contributing to their high medal counts.

* **Global Appeal**:

→Sports like **Athletics** and **Swimming** typically have wide global appeal, as athletes from many countries compete in these events. This is aligned with the document's mention that some sports have broad participation due to their universal appeal​.

* **Comparing Event Size**:

→The plot shows a clear difference in the number of medals between **Athletics** and sports like **Basketball** or **Sailing**, indicating that some sports have larger event pools and therefore offer more medal opportunities. This can be linked to the insight about event size contributing to the competitiveness of certain disciplines​.

* **Investment in Certain Sports**:

→Countries likely invest heavily in sports like **Athletics** and **Swimming** because they offer numerous medal opportunities. **Football**, **Hockey**, and **Basketball**, while popular globally, appear lower on the chart due to having fewer Olympic events compared to athletics​.

**Limitations:**

1.Predisposition Towards Dominant Nations and Historical Setting:

Due to their substantial resources, sizable populations, and robust sports cultures, a select group of powerful nations—the US, the USSR, and China—have historically performed well in the Olympics. As a result, the dataset is significantly biased in favor of these nations. This hegemony may distort the study by hiding the achievements of smaller or less wealthy countries. The fact that nations with a history of success at the Olympics, like the Soviet Union and the GDR, are no longer in existence complicates the analysis of historical medal totals. This makes it more difficult to assess how well newly developed nations are performing or how well they have fluctuated as a result of political or economic shifts.

2. Event Saturation and Gender Disparities:

Some sports have more medals than others, especially swimming and athletics because these sports have a lot of events in them. Due to the increased chances of medals for participating athletes and nations, this may distort the idea of success in various sports. Specialty sports with fewer events, like sailing or basketball, are given less consideration in medal analysis, which could devalue nations that are strong in these areas. Another major obstacle is gender inequality. Women's involvement in some Olympic sports has historically been restricted or eliminated, which contributed to their underrepresentation in the dataset's earliest years. Therefore, it's possible that the examination of gender performance underrepresents the advancements achieved toward gender equality, especially in later years.

3. Socioeconomic Context and Data Coverage:

The dataset is limited in its ability to provide insights into current trends, developing sports, and changes in national performance because it only includes Olympics competitions from 1896 to 2016. Furthermore, even while the analysis discusses medal totals and national performance, it ignores important social and economic aspects that affect Olympic success, like a country's GDP, its commitment to sports, and its availability of training facilities. It is difficult to completely comprehend the causes behind a nation's success or failures at the Olympics without taking these factors into account. Furthermore, the data does not include information about specific athletes, such as age or changes in nationality, or variables like the effect of sports science and current technology on performance.

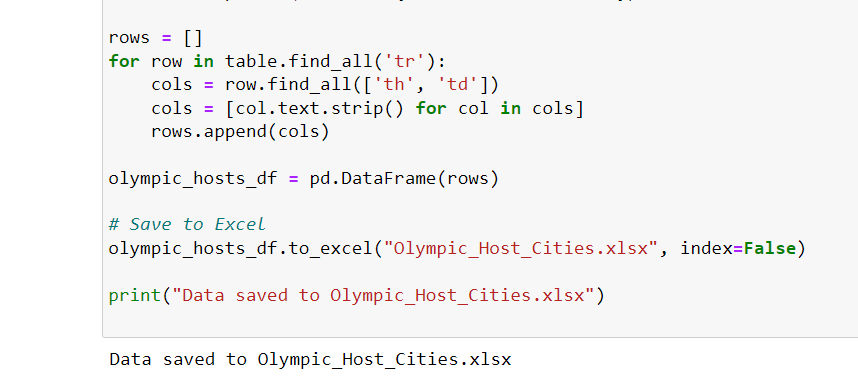
# **Issues with the Datasets:**

## **Data Acquisition and Consistency Challenges**

Our objective was to utilize a comprehensive and reliable supplementary dataset for historical comparisons of various features across past Olympic Games. Ideally, this dataset would be obtained in raw form from a single credible source, ensuring consistency in the types of insights we sought to derive. However, we encountered significant challenges in identifying a data source that met our reliability standards and aligned with the specific variables necessary for our analysis.

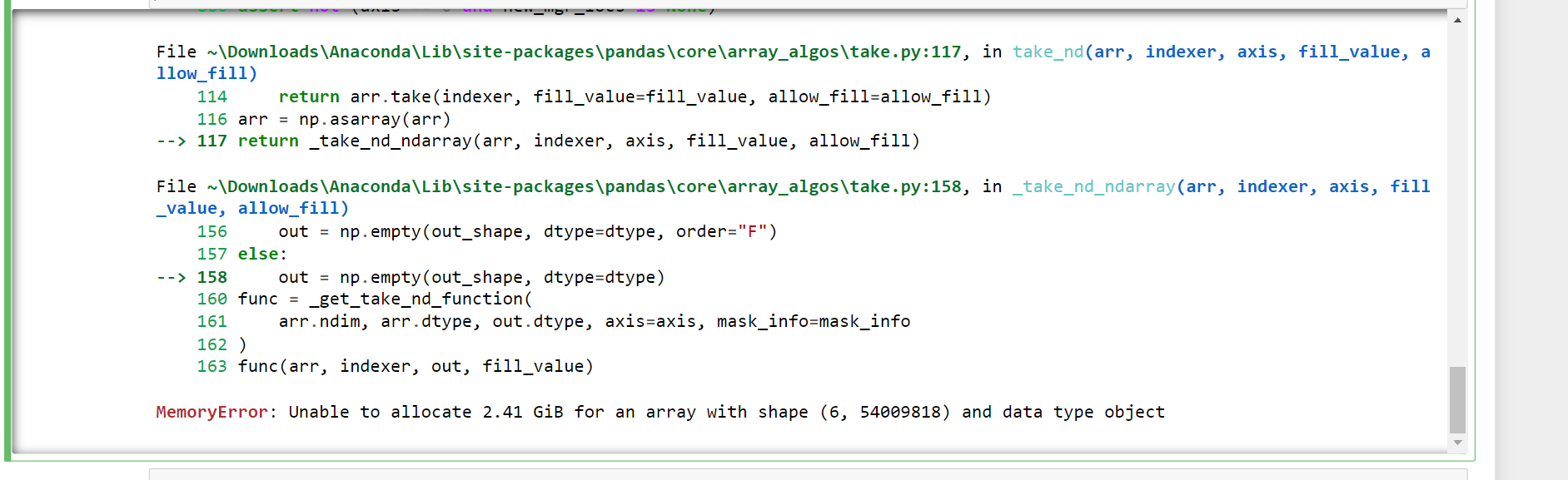
In addition, no existing tabular datasets provided a complete and structured history of Olympic host countries. To address this gap, we implemented a web scraping solution, extracting data from a credible source, Wikipedia, and utilized Python's Pandas library to efficiently format and convert the data into a usable CSV format. This process allowed us to integrate the necessary historical data into our analysis pipeline while maintaining data integrity and relevance to the study’s objectives.





## **Feature Matching Problems**

During our analysis, we encountered significant memory allocation issues when processing large datasets. The MemoryError, indicating an inability to allocate the required memory for an array operation, underscored the challenges of managing substantial data volumes with complex structures. This constraint necessitated a review of our data handling strategies, leading to optimizations in data type usage and the adoption of more efficient data processing techniques to ensure stability and efficiency in our computational operations. These steps were crucial in enabling continued analysis without compromising on the depth or accuracy of our insights.



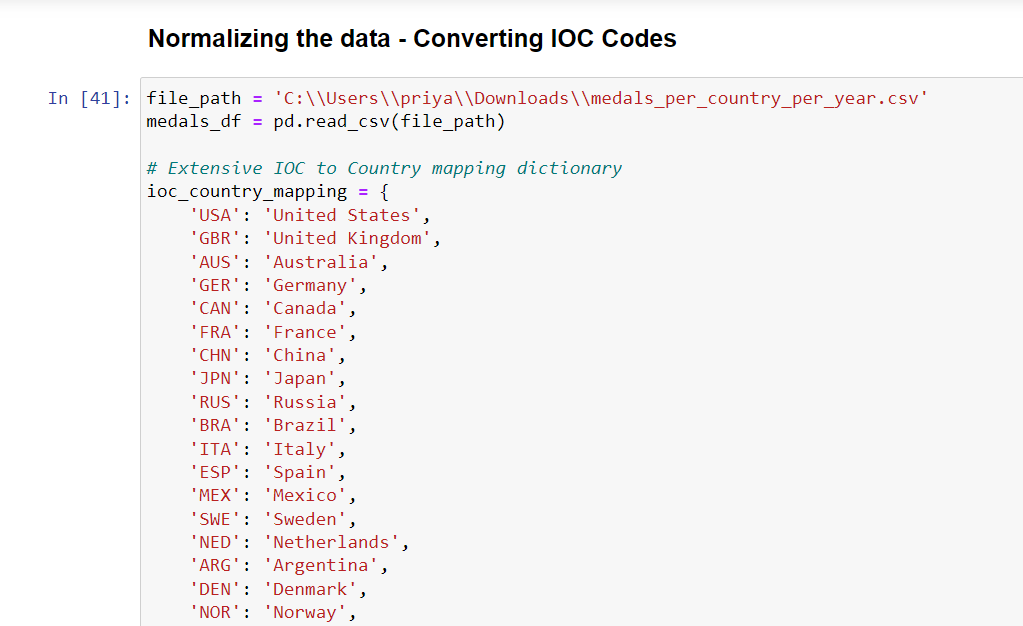
This error is typically triggered when Python's attempt to allocate additional memory for data processing exceeds the available memory on your machine

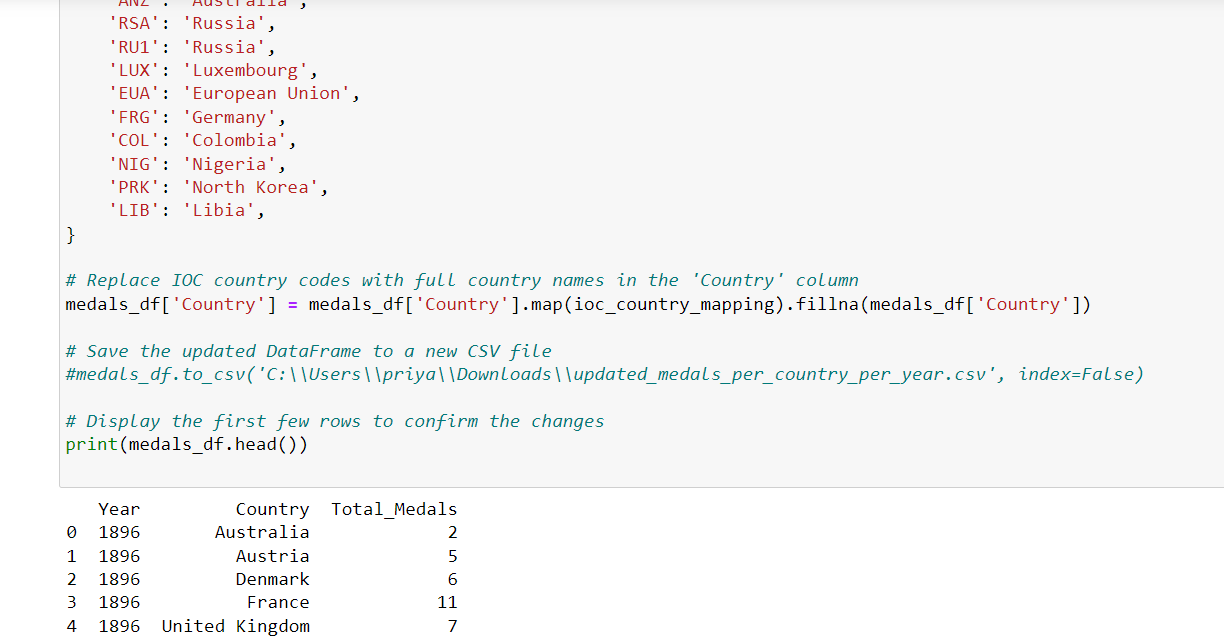
## **Data Consistency and Integration Challenges**

One of the key issues we faced with our dataset was the inconsistency in country identifiers. The raw data utilized various country code formats, particularly the International Olympic Committee (IOC) codes, which posed challenges when attempting to merge or analyze data from different sources. This lack of uniformity necessitated a comprehensive normalization process to standardize country identifiers across the dataset.

### **Implementation Strategy**

* Mapping Creation: The manual creation of a dictionary to map IOC codes to country names was necessary due to the lack of direct support in libraries like pycountry. This dictionary serves as a lookup table, facilitating the conversion process.
* Application of Mapping: Applying this mapping to the DataFrame involved replacing each IOC code with its corresponding full country name. This step was implemented using DataFrame operations that check the dictionary for each code and replace it with the appropriate country name.





# Conclusions:

The Paris 2024 Olympics research project seeks to understand the various factors influencing athlete participation and performance, with a specific emphasis on gender disparities and the performance of host nations. Utilizing Jupyter as a tool, the project aims to uncover systemic advantages or patterns that may favor the host country while also analyzing trends in gender representation and the inclusion of marginalized communities.

One of the challenges faced by the project was acquiring comprehensive private datasets necessary for robust analysis. Efforts to obtain data from multinational corporations were met with resistance due to confidentiality concerns. As a result, the team turned to alternative sources such as Kaggle to access publicly available datasets.While Kaggle provided a wealth of data, inconsistencies in the data features presented additional challenges. Inconsistencies in the format, organization, and relevance of the data required additional processing steps, which could limit the breadth of analysis achievable.These experiences emphasize the importance of data accessibility and quality in conducting thorough research.

The detailed analysis of each dataset was conducted to understand and visualize the data, assess the significance of each feature, and identify any limitations affecting the overall goal. Due to the shortcomings of individual datasets and the absence of necessary features in a single dataset, we merged relevant features from multiple datasets to create a comprehensive final dataset that addresses gender disparity and the relationship between host country and performance. To demonstrate how our goal has evolved over time, it was essential to incorporate historical data, providing a clearer perspective on the trends and outcomes. Hence we merged all the required feature across the datasets, to prove our goal of Gender Disparity and Host country v/s Performance.

References:

<https://pwskills.com/blog/data-science-in-sports/>

https://thedatasciencenewsletter.substack.com/p/data-science-in-sports-analytics