Kéli Davis, Maybel Herrera, Sandy Leung
Final Project Report
09 September 2024

Olympics Analysis

1. Executive Summary

This report examines the evolving dynamics of the Summer Olympics from the 1800s to 2024, with a focus on medal distributions, gender participation, top-performing sports, and the dominance of various countries. Through this analysis, we reveal key trends and shifts that have shaped the Olympic Games over more than a century.

2. Introduction

The Olympic Games stand as the pinnacle of international sports, drawing the best athletes from across the globe every four years. This project delves into the historical and current performance trends in the Olympic Games, with a particular emphasis on the 2020/24 Tokyo Olympics, comparisons to previous summer and winter games, and any preliminary data from the 2024 Paris Olympics. Leveraging datasets from Kaggle, this analysis will explore athlete demographics, medal distributions, and evolving performance patterns over time.

3. Data Source and Cleaning

These are the following dataset sources that were leveraged to explore the various aspects of the Olympic games:

- https://www.kaggle.com/datasets/stefanydeoliveira/summer-olympics-medals-1896-2024

Overall, four datasets were being explored, including 'olympics_dataset.csv', 'Medals_2020.xlsx', 'medals_2024.csv', and 'Athletes_2024.csv', ‘Tokyo_2020_tweets’.

One of the initial data cleaning steps involved ensuring that all four datasets were free of missing values (NAs). With the project's focus on the Tokyo 2020 and 2024 Olympics, filtering was applied to isolate and create a new dataframe specifically containing data from the summer Tokyo 2020/24 Games.

4. Special Processes

Decision tree model is used to predict how likely countries are to attain Gold, Bronze, Silver or no medals. This model analyzes various factors, such as age, gender and team, to classify each country’s probability of attaining the different types of medals.

5. Method of Data Analysis

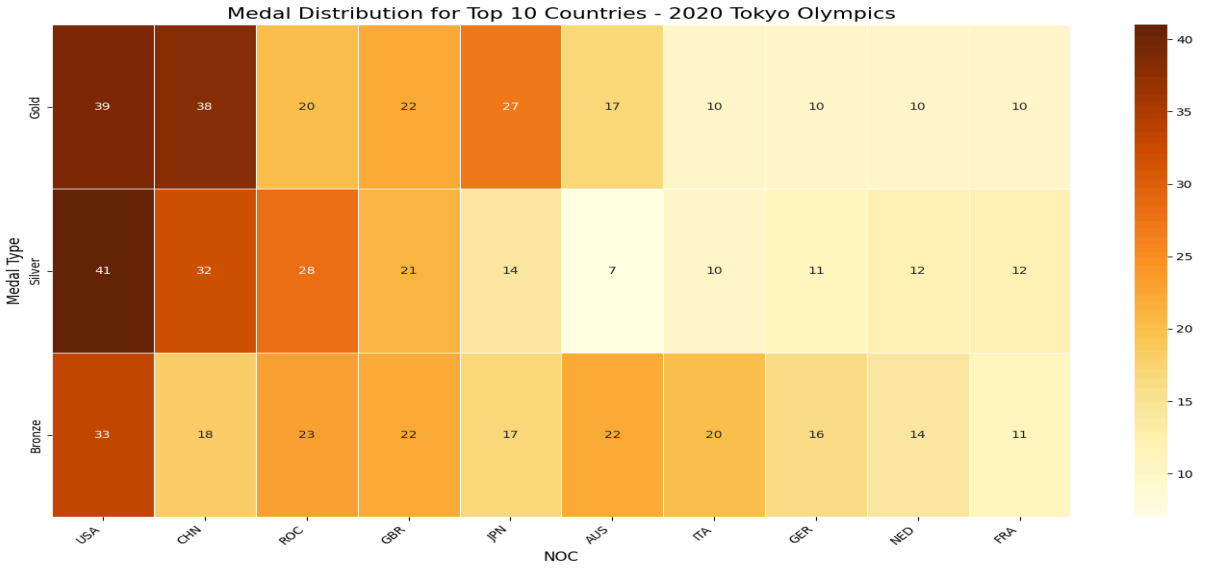
Various methods are being utilized for analysis, with a focus on visualizations such as bar graphs, line graphs, and heatmaps.

5.1 2020 Olympics: Analysis for 2020 Tokyo Olympics

Questions:
1. What are the top 10 highest-performing countries based on medal counts?
2. Top sports based on medal count?
3. What country dominated what sport?
4. What was the gender participation in each sport? Was it evenly distributed? Are there sports more male-dominated or female-dominated?
5. Athletes per country that participated. Which country had more athletes? Did that correlate with the high medal count?

Top Performing Countries:
What are the top 10 highest-performing countries based on medal counts?

	NOC	Gold	Silver	Bronze	Total
0	United States of America	39	41	33	113
1	People's Republic of China	38	32	18	88
2	ROC	20	28	23	71
3	Great Britain	22	21	22	65
4	Japan	27	14	17	58
5	Australia	17	7	22	46
6	Italy	10	10	20	40
7	Germany	10	11	16	37
8	Netherlands	10	12	14	36
9	France	10	12	11	33

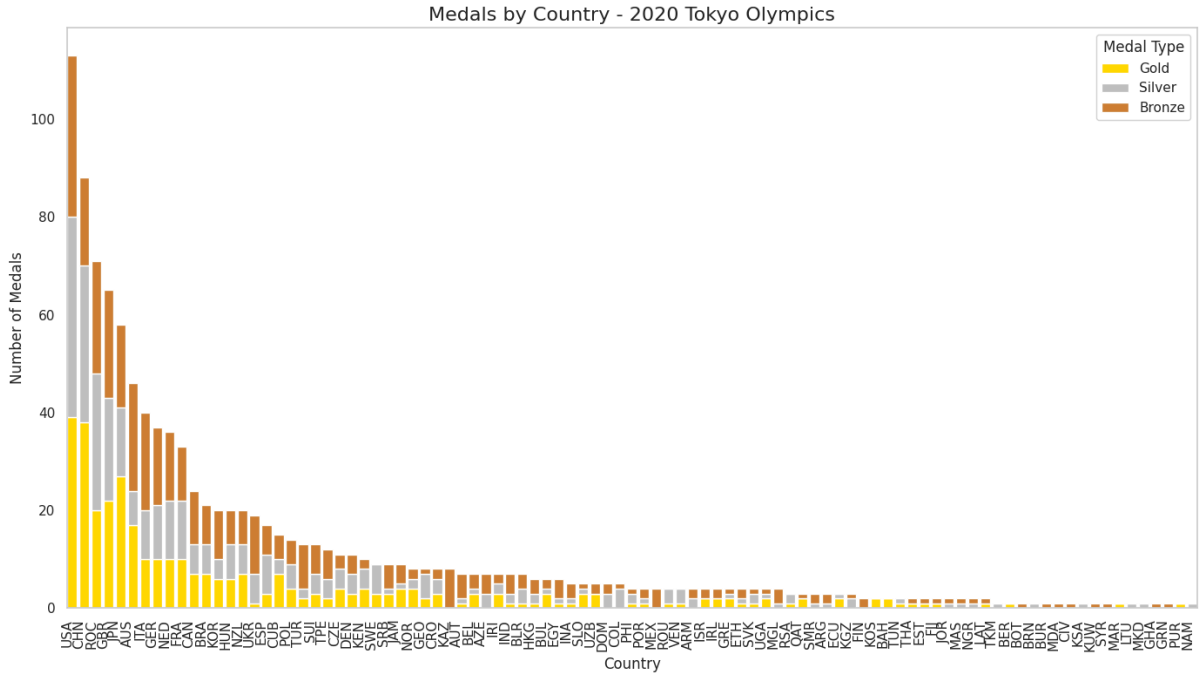


- **United States of America (USA)**: The USA stands out as the top-performing country, securing the highest number of medals across the Gold, Silver, and Bronze categories. This demonstrates the country's consistent excellence across a wide range of sports.
- **People's Republic of China (CHN)**: China follows closely behind the USA, also exhibiting a strong presence in the Gold and Silver medal categories. The dominance

of these two countries is apparent in the significant height of their bars compared to other nations.

- **Great Britain (GBR) and ROC (Russian Olympic Committee)**: Both Great Britain and ROC also performed strongly, each securing numerous medals. Their performance is particularly notable in the Gold and Silver categories.
- **Japan (the host nation)** also performed well, with a respectable number of medals, making it to the top 10 on the NOC list.
- Countries like **AUS (Australia)**, **ITA (Italy)**, **GER (Germany)**, **NED (Netherlands)**, and **FRA (France)** fall into the mid-tier category. These countries show a balanced distribution of medals but do not reach the same heights as the top-performing nations

Analysis of overall medal-winning distribution?

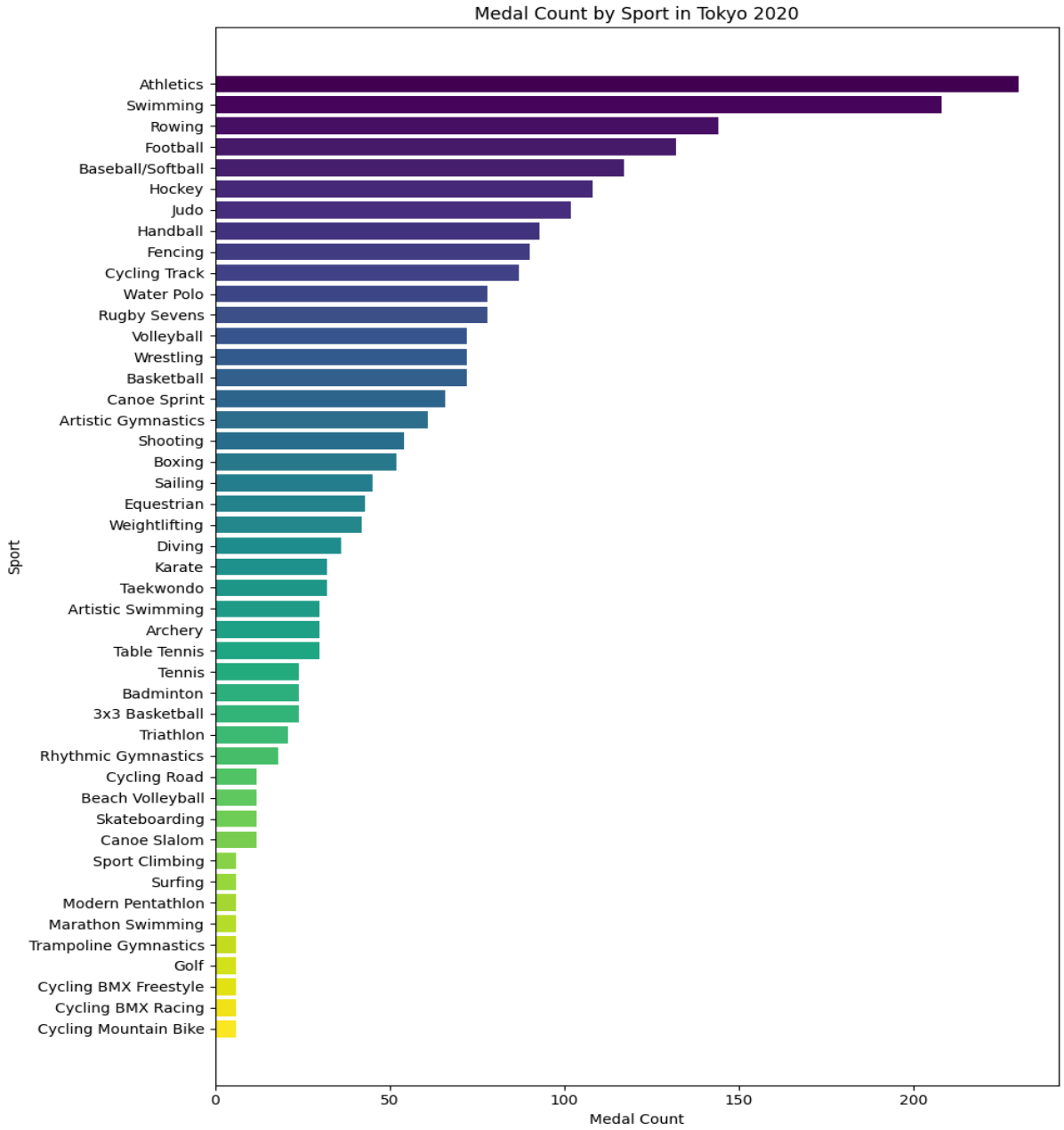


- The distribution shows a steep drop-off after the top-performing nations, with many countries earning only a handful of medals. This demonstrates the high level of competition at the top and the challenge for many nations to reach the podium.
- As the chart progresses to the right, the total number of medals diminishes. Countries like CAN (Canada), BRA (Brazil), KOR (South Korea), and others have fewer medals, indicating less dominance in the games but still achieving notable success
- The distribution is highly skewed, with a small group of countries earning a large proportion of the medals, while the majority earned far fewer. This pattern is typical of the Olympics, where historically, a select few nations with robust sports programs and greater resources tend to dominate the overall medal count. Despite this, the achievement of any medal is a testament to the hard work, dedication, and skill of the athletes and their support teams.

Analysis of Top Sports at the 2020 Tokyo Olympics Based on Total Medals:

Top sports based on the medal count?

The chart highlights the broad range of sports where athletes compete for medals, from team-based sports (Football, Basketball) to individual disciplines (Shooting, Weightlifting)**.**



Athletics and Swimming lead by a large margin, with Athletics having the most medals (over 200), followed closely by Swimming.

Other sports such as Rowing, Football, and Baseball/Softball also contribute significantly to the medal count.

To answer our question the top sports based on the medal count in the Tokyo 2020 Olympics, as shown in the bar chart, are:

1. **Athletics** – The sport with the highest number of medals, exceeding 200.

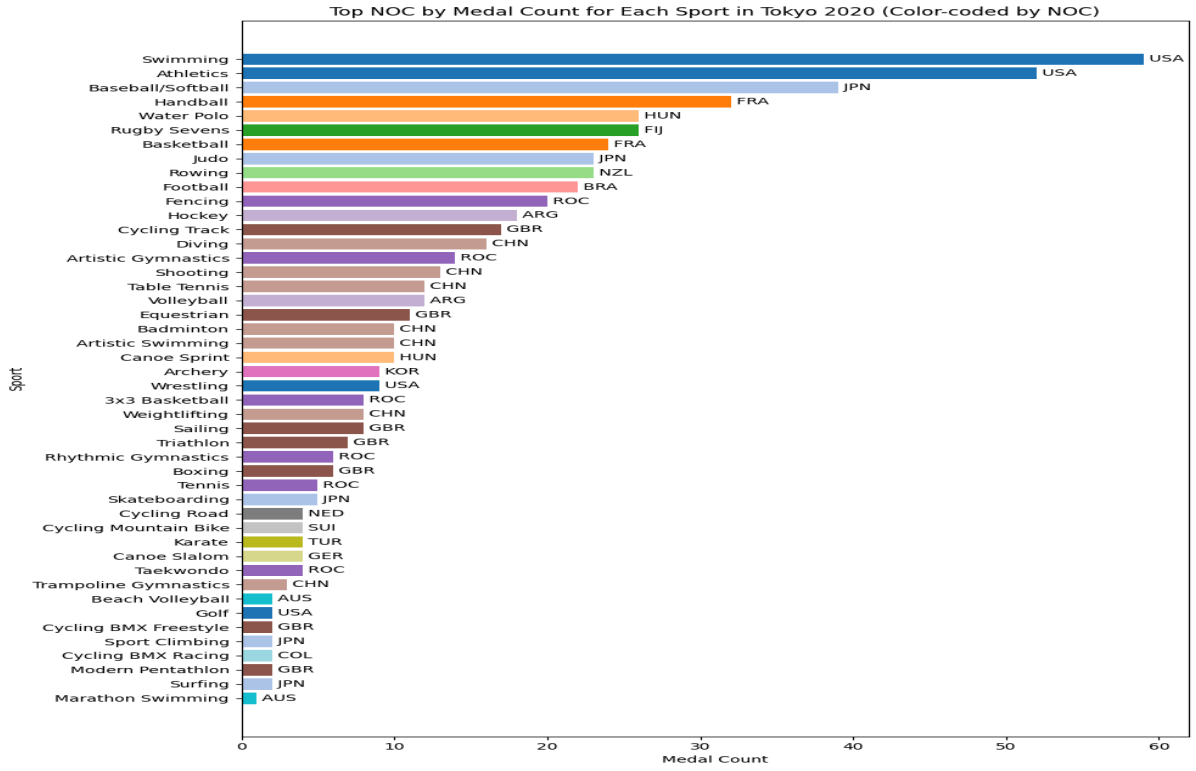
2. **Swimming** – Close behind, Swimming ranks second with a substantial number of medals, also over 150.
3. **Rowing** – Takes the third spot with a strong medal count.
4. **Football** – This team sport holds a significant position in fourth place.
5. **Baseball/Softball** – Rounding out the top five with a considerable number of medals.

These top sports show a large gap in medal counts compared to others, indicating their prominence in the Olympics in terms of the number of events and competitive opportunities for athletes.

Analysis of top National Olympic Committees for each sport.

What country dominated what sport?

This visualization highlights the dominance of specific National Olympic Committees (NOCs) in various sports during the 2020 Tokyo Olympics.



- **USA** dominated two of the most competitive sports, **Swimming** and **Athletics**, with a medal count exceeding 50 in both categories, indicating a strong performance in these disciplines.
- **Japan (JPN)** led in **Baseball/Softball** and **Surfing**, showcasing their expertise in both a traditional sport and a newer Olympic event.
- **France (FRA)** dominated **Handball** and **Basketball**, while **Hungary (HUN)** led in **Water Polo**, reinforcing their strength in team sports.

Team Sports:

- Several NOCs dominated traditional team sports: **Fiji** in **Rugby Sevens**, **Brazil (BRA)** in **Football**, and **New Zealand (NZL)** in **Rowing**.

Smaller Disciplines:

- For many of the smaller, less medal-heavy sports, a variety of NOCs are dominant. For example, **Germany (GER)** led in **Canoe Slalom**, and **ROC** (Russian Olympic Committee) in **Fencing** and **Artistic Gymnastics**.

Broad Representation:

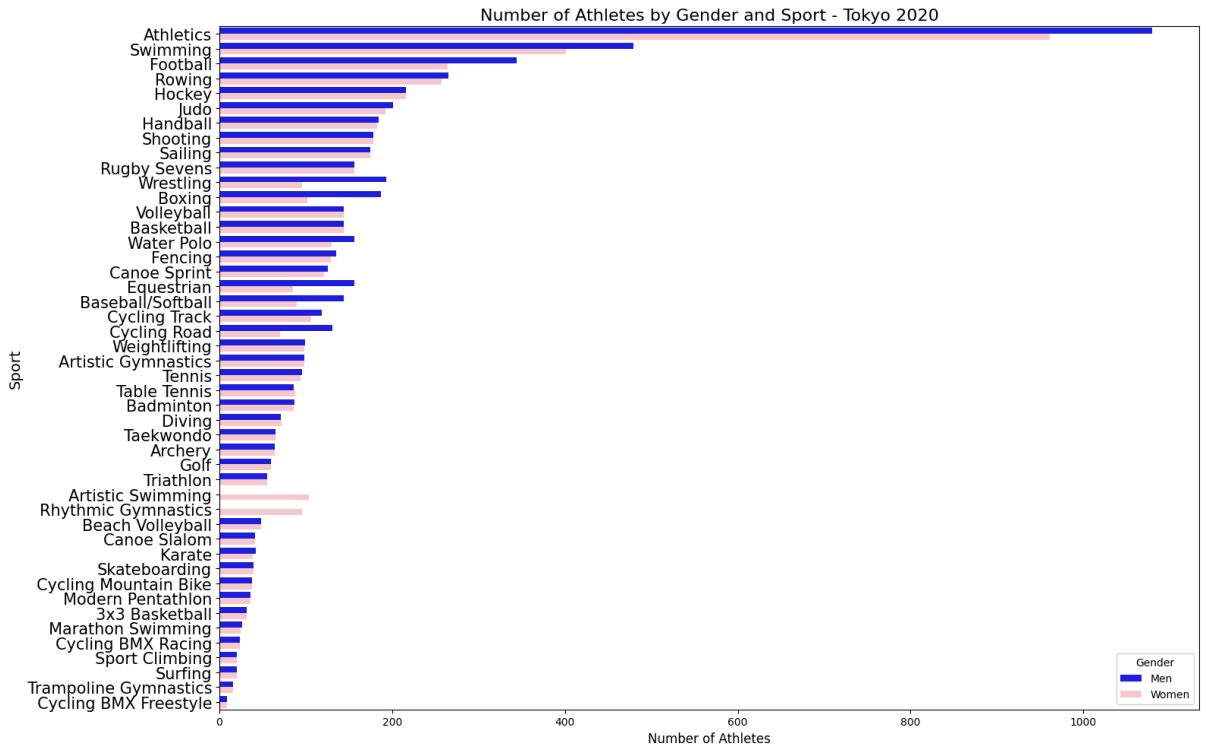
- The chart shows a broad representation of NOCs across different sports, highlighting that no single NOC dominated across the board. For instance, **China (CHN)** shows strength in precision and technical sports like **Shooting**, **Table Tennis**, and **Weightlifting**.

  In summary, the USA, Japan, and China emerged as key leaders in terms of dominating most sports based on total medal counts.

Analysis of Event Distribution by Gender and Sport - Tokyo 2020 Olympics

What was the gender participation in each sport? Was it evenly distributed? Are there sports more male-dominated or female-dominated?

Was Gender Participation Even?



Even Participation:

- Athletics and Swimming continue to show relatively even participation between men and women.
- Hockey, Judo, and Rowing show balanced gender participation

Uneven Participation:

- **Boxing** and **Wrestling** remain male-dominated sports.
- **Artistic Swimming** continues to have a strong skew towards female athletes, as originally noted.

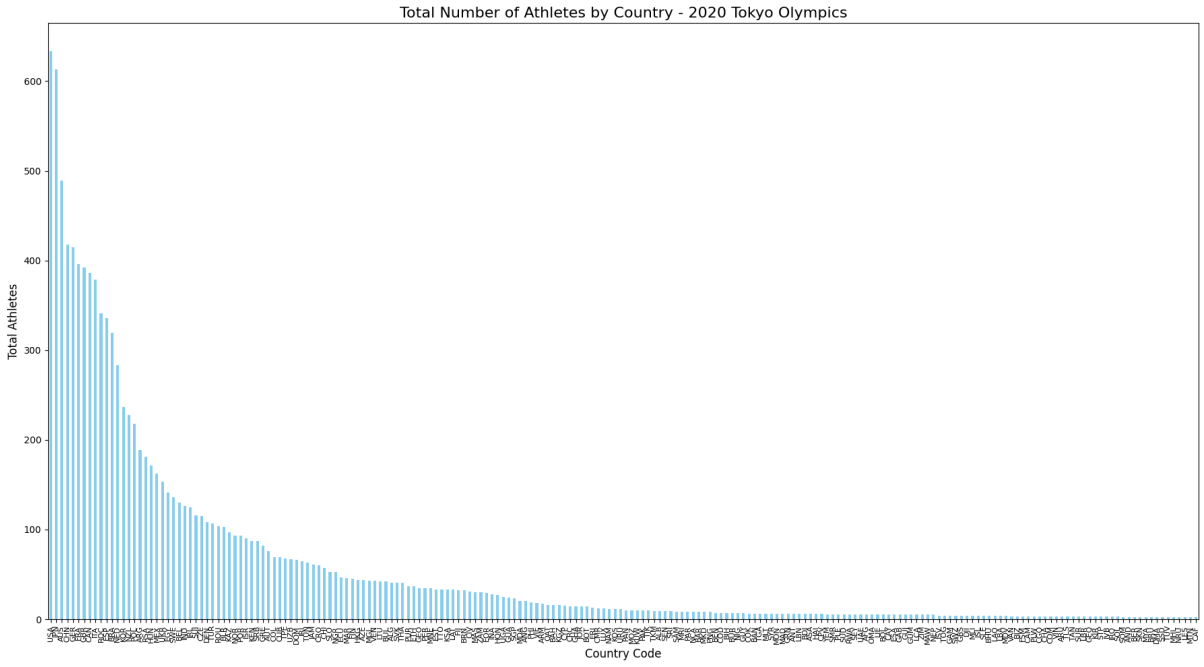
Summary

The Tokyo 2020 Olympics made strides in balancing gender representation across various sports, with many sports showing equal or nearly equal numbers of events for men and women. However, there are still areas with noticeable gender disparities, particularly in sports traditionally associated with one gender, such as Artistic Swimming for women and certain combat sports for men. Sports with a good balance in participation, reflect successful efforts towards gender parity. The balanced numbers in these sports indicate that both men and women are equally encouraged and supported in these disciplines.

Analysis of Athlete Distribution Per Country - Tokyo 2020 Olympics

Athletes per country that participated. Which country had more athletes? Did that correlate with the high medal count?

The bar chart represents the total number of athletes participating from each country (NOC) in the 2020 Tokyo Olympics.



General Distribution:

- The chart shows a steep decline after the top 2 countries, with many nations sending fewer athletes. This reflects the varying levels of resources and focus on the Olympics across different countries.
- Countries like the Marshall Islands (MHL), Nauru (NRU), Lesotho (LES), Mauritania (MTN), and Central African Republic (CAF) had the smallest delegations, each sending only 2 athletes. These smaller teams typically focus on one or two sports, often reflecting the limited sports infrastructure and resources available in these nations.

Top 10 Countries by Athlete Representation:

NOC	
USA	633
JPN	613
AUS	489
CHN	418
GER	415
FRA	396
GBR	392
CAN	386
ITA	379
ROC	341

Let's discuss the top 5 NOC:

- **USA**: The United States had the highest number of athletes with a total of 633. This strong presence is typical for the USA, which is known for its large and diverse participation across numerous sports.
- **Japan (JPN)**: Japan, as the host nation, had 613 athletes participating, reflecting the advantage of host nations in fielding a larger team.
- **Australia (AUS)**: Australia had a significant representation with 489 athletes, reflecting its strong sporting tradition, particularly in swimming and other water sports.
- **China (CHN)**: China sent 418 athletes, consistent with its position as a global sporting powerhouse, especially in sports like gymnastics, table tennis, and diving.
- **Germany (GER)**: Germany's delegation included 415 athletes, showcasing its broad participation in both traditional and modern Olympic sports.

Diversity of Participation:

- The diversity in the number of athletes per country highlights the global nature of the Olympics, with 206 countries represented. While some countries have a large number of athletes participating across many sports, others are represented by just a few athletes, often in specific events.

Host Nation Advantage:

- Japan's large athlete count can be attributed to its role as the host nation. Host countries often benefit from increased quotas and automatic qualification in many sports, allowing them to field more athletes than they might otherwise.

Resource Disparities:

- The disparity between the number of athletes from different countries underscores the varying levels of investment in sports development. Wealthier nations with more developed sports infrastructures can support larger teams across a wider range of sports.

If more athletes equal more medals, how did ROC and Great Britain come in 3rd and 4th in medal counts but they had fewer athletes than Japan who came in 5th?

Medal Counts and Athlete Participation:

ROC (Russian Olympic Committee): Won a total of 71 medals.

Great Britain (GBR): Won a total of 65 medals.

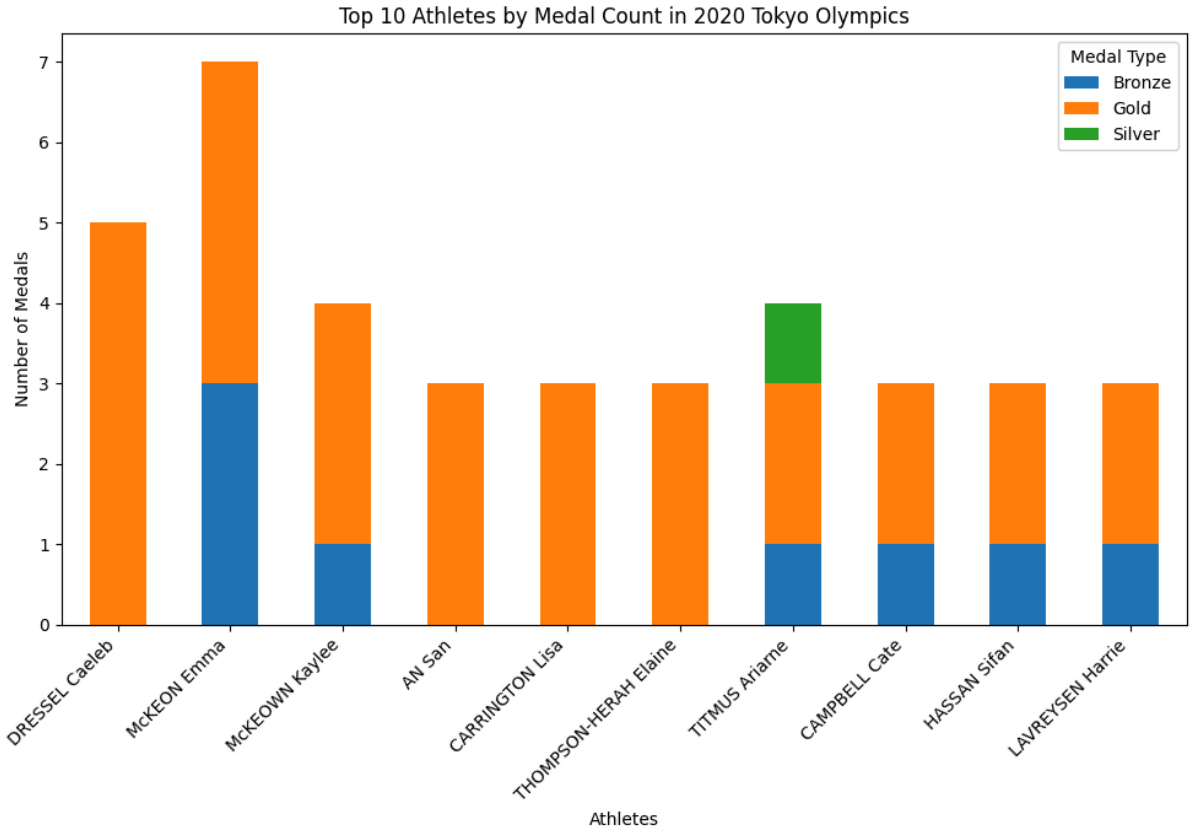
Japan (JPN): Won a total of 58 medals.

Despite Japan having significantly more athletes (613), both ROC (341 athletes) and Great Britain (392 athletes) managed to outperform Japan in terms of total medals won. This suggests that ROC and Great Britain were more efficient in converting their athlete participation into medal-winning performances.

Top contenders in the 2020 Tokyo Olympics

The bar chart displays the top 10 athletes based on the total number of medals won during the 2020 Tokyo Olympics. The athletes are represented on the y-axis, while the total number of medals is shown on the x-axis. The bars are color-coded to distinguish between two teams: the United States and China.

Medal	Bronze	Gold	Silver
Name			
DRESSEL Caeleb	0	5	0
McKEON Emma	3	4	0
McKEOWN Kaylee	1	3	0
AN San	0	3	0
CARRINGTON Lisa	0	3	0
THOMPSON-HERAH Elaine	0	3	0
TITMUS Ariarne	1	2	1
CAMPBELL Cate	1	2	0
HASSAN Sifan	1	2	0
LAVREYSEN Harrie	1	2	0



Caeleb Dressel and **Emma McKeon** are the standout athletes, each with 7 medals. Dressel won primarily gold medals, while McKeon had a more balanced distribution of gold and bronze medals.

Kaylee McKeown followed with a total of 4 medals, consisting mostly of gold medals.

The rest of the athletes, including **An San**, **Lisa Carrington**, and **Elaine Thompson-Herah**, have slightly fewer medals, with most athletes in this group earning 3 or 4 medals each.

Ariarne Titmus is unique in this group for having a prominent number of silver medals, standing out as the only athlete with a significant contribution from silver medals in this top 10 list.

Summary

This analysis provides a snapshot of global participation in the 2020 Tokyo Olympics, reflecting the competitive nature of the games and the broad appeal and reach of the Olympic movement worldwide. The COVID-19 pandemic likely caused a decrease in the number of athletes participating, particularly from smaller nations or in sports where qualification was heavily disrupted. The postponement, health concerns, travel restrictions, and financial impacts were major factors influencing athlete participation.

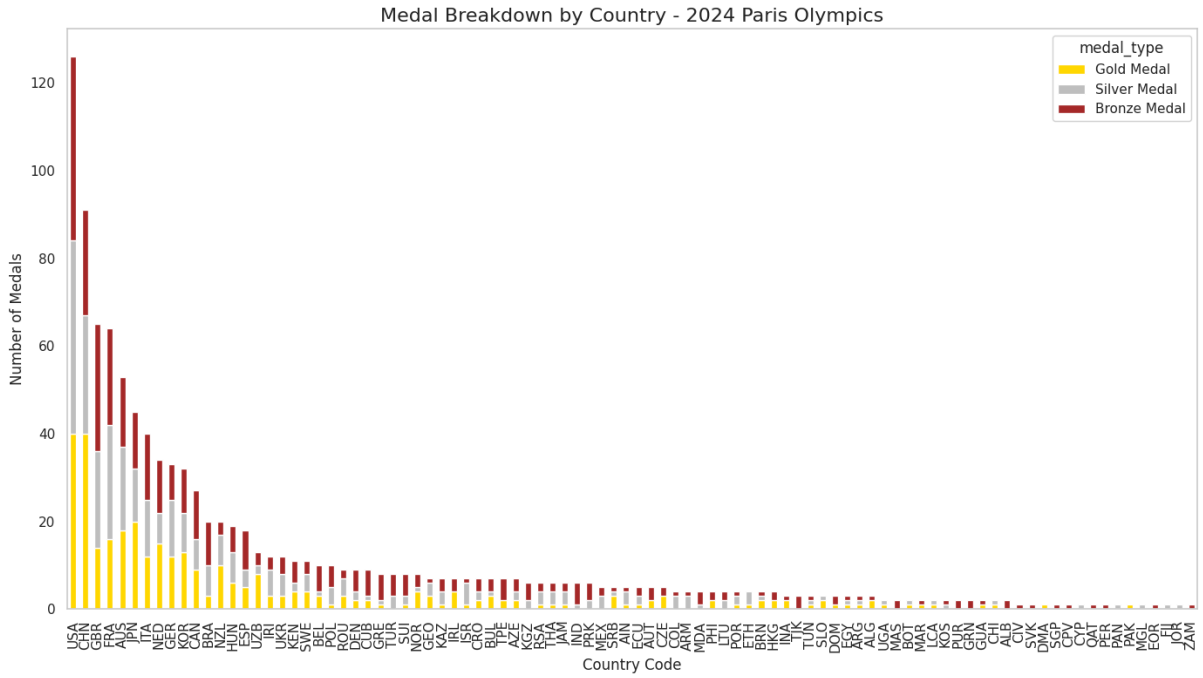
Notably, the countries with the highest medal counts were generally those that sent the largest number of athletes, suggesting that a larger pool of athletes provided these nations with more opportunities to compete across various events, thereby increasing their chances of winning medals. The substantial representation of athletes from countries like the USA, Japan, and China played a significant role in their top performance, demonstrating the importance of both depth and breadth in an Olympic team.

However, the 2020 Tokyo Olympics also presented notable exceptions to this trend. ROC and Great Britain, despite having significantly fewer athletes than Japan, outperformed the host nation in total medals won. This achievement underscores that strategic focus, athlete preparation, and efficiency in converting participation into podium finishes can sometimes outweigh sheer numbers. The ability of ROC and Great Britain to excel with a smaller delegation highlights the nuanced dynamics of Olympic success, particularly in the context of the unique challenges posed by the global pandemic.

5.3 2024 Olympics

Comparison Analysis from 2020 Summer Olympics to 2024

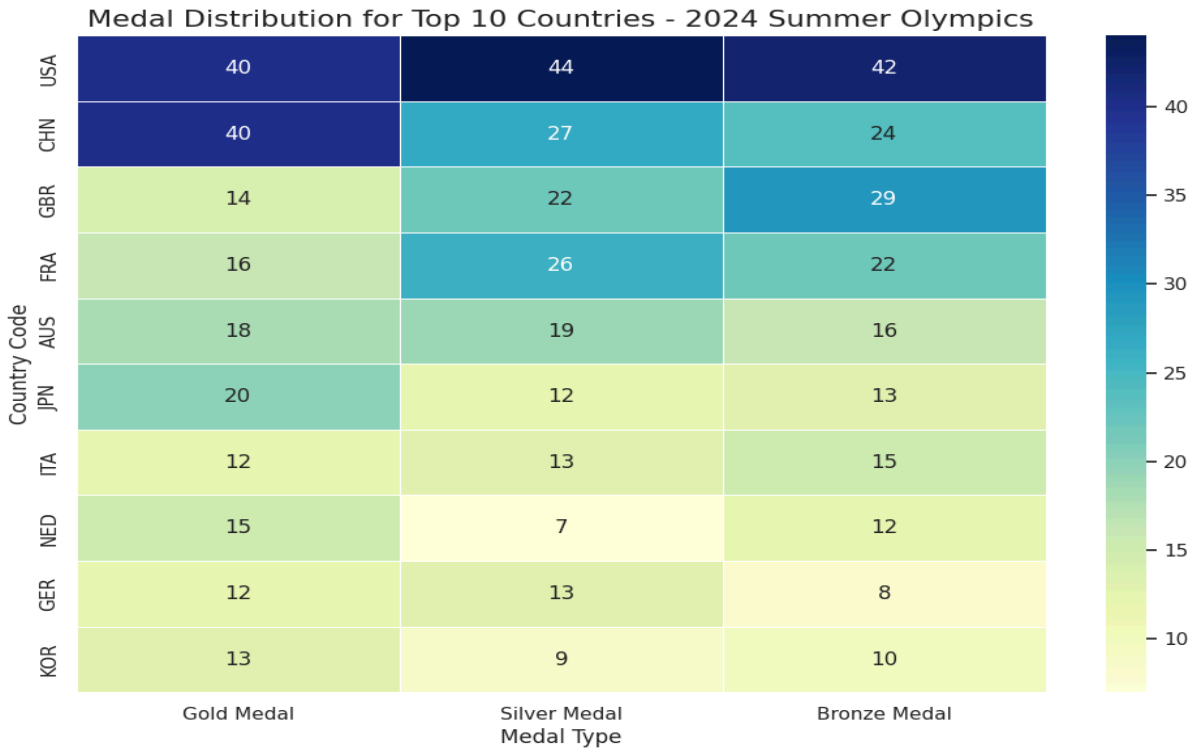
Analysis of 2024 top 10 NOC:
Questions :
Total medal counts



- In 2024, the steep decline in medal counts after the top-performing countries is more pronounced compared to the 2020 Olympics, indicating a concentration of medals among a few top nations.
- Russia, who was not allowed to compete, significantly impacted the total medal count distribution.
- France experienced a substantial increase in medals, likely benefiting from home advantage.
- The USA maintained its position as the top medal-winning country, with a marked increase in total medals from 2020 to 2024.

What are the top 10 NOC based on medal counts?

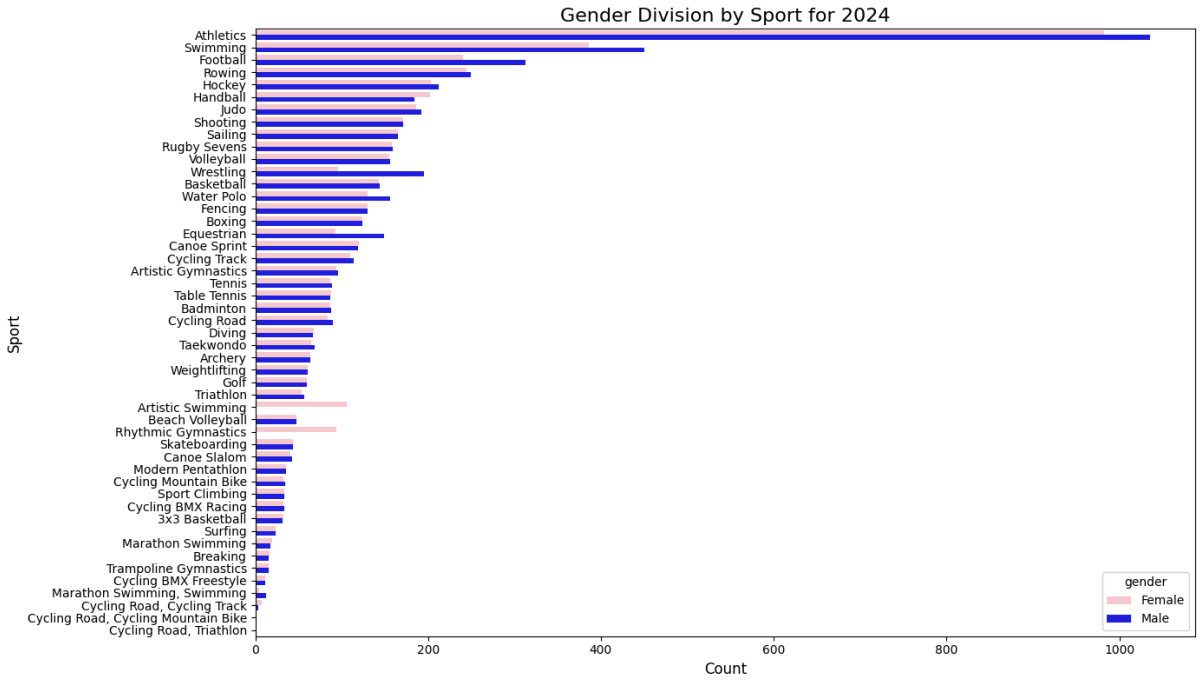
medal_type	Bronze Medal	Gold Medal	Silver Medal	Total
country_code				
USA	42	40	44	126
CHN	24	40	27	91
GBR	29	14	22	65
FRA	22	16	26	64
AUS	16	18	19	53
JPN	13	20	12	45
ITA	15	12	13	40
NED	12	15	7	34
GER	8	12	13	33
KOR	10	13	9	32



- The United States and China have maintained dominance, with slight increases in their overall medal counts from the 2020 to 2024 games.
- France went from number 10th place in 2020 to 4th place, they Increased total medals by 31 (from 33 in 2020 to 64 in 2024) this could be due to the games being hosted in Paris, France.
- Italy remained the same with 40 total medals won.
- Korea emerged strongly in 2024, entering the top 10 with 32 total medals.
- Russia did not make it to the top 10 as they were banned from the games due to political reasons(war in Ukraine)

Gender analysis comparison:

What was the gender participation in each sport? Are there sports more male-dominated or female-dominated?

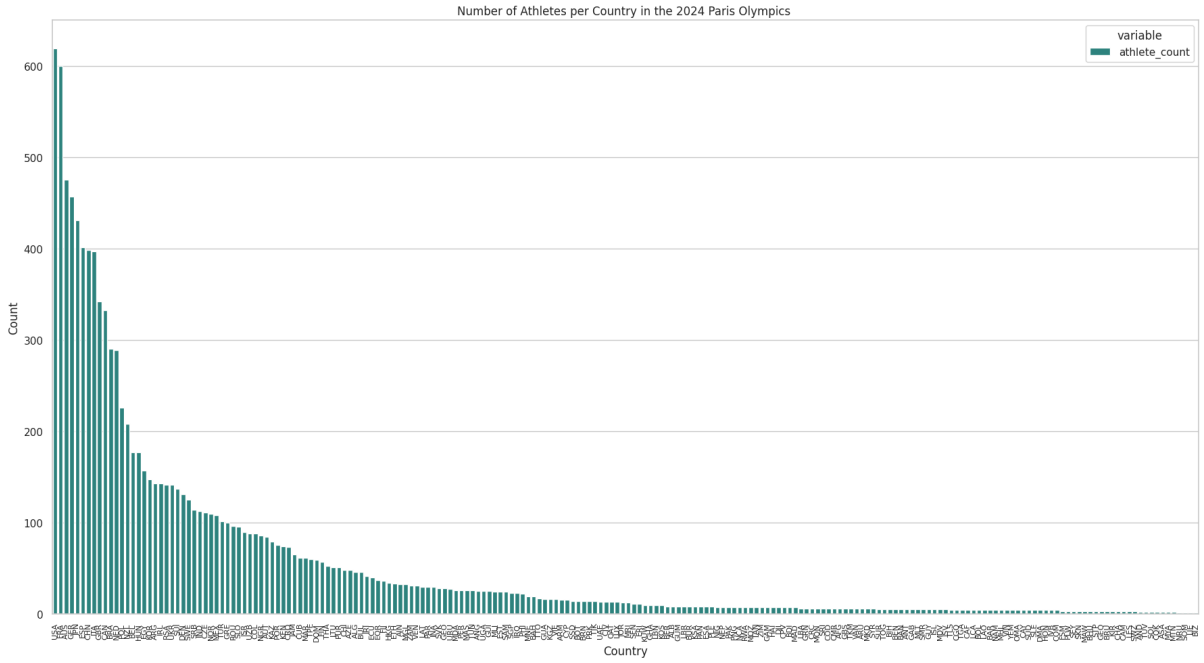


- The highest athlete count is still in Athletics, Swimming rowing, and football.
- Rhythmic Gymnastics gender participation has not changed since the last summer Olympics. We again see a strong female dominance, consistent with traditional trends in this sport. Artistic Swimming also continues to be female-dominated.
- Wrestling and Football remain a male-dominated sport, with the data showing a substantial difference between male and female participation**.** However, the gap between males and females seems to have shortened showing a slight change towards more females participating.
- Boxing has shown a huge change as this year it had equal female-male participation, unlike the 2020 Tokyo Olympics where it was more male-dominated.

Question: Which country had more athletes?

Top 10 Countries by Athlete Representation in 2024:

	country_code	athlete_count
197	USA	619
65	FRA	600
11	AUS	475
73	GER	457
97	JPN	431
60	ESP	401
38	CHN	398
93	ITA	397
69	GBR	342
33	CAN	332



- In both the 2020 Tokyo Olympics and the 2024 Paris Olympics, the United States (USA) has the highest number of athletes (even higher than the hosting nation)
- Like the summer 2020 Tokyo Olympics, the hosting nation had the second-highest number of athletes.
- Again, we see a steep decline after the top 10-15 countries skewed left. The consistency in the distribution patterns suggests the same countries might be participating with similar-sized teams across both Olympics.

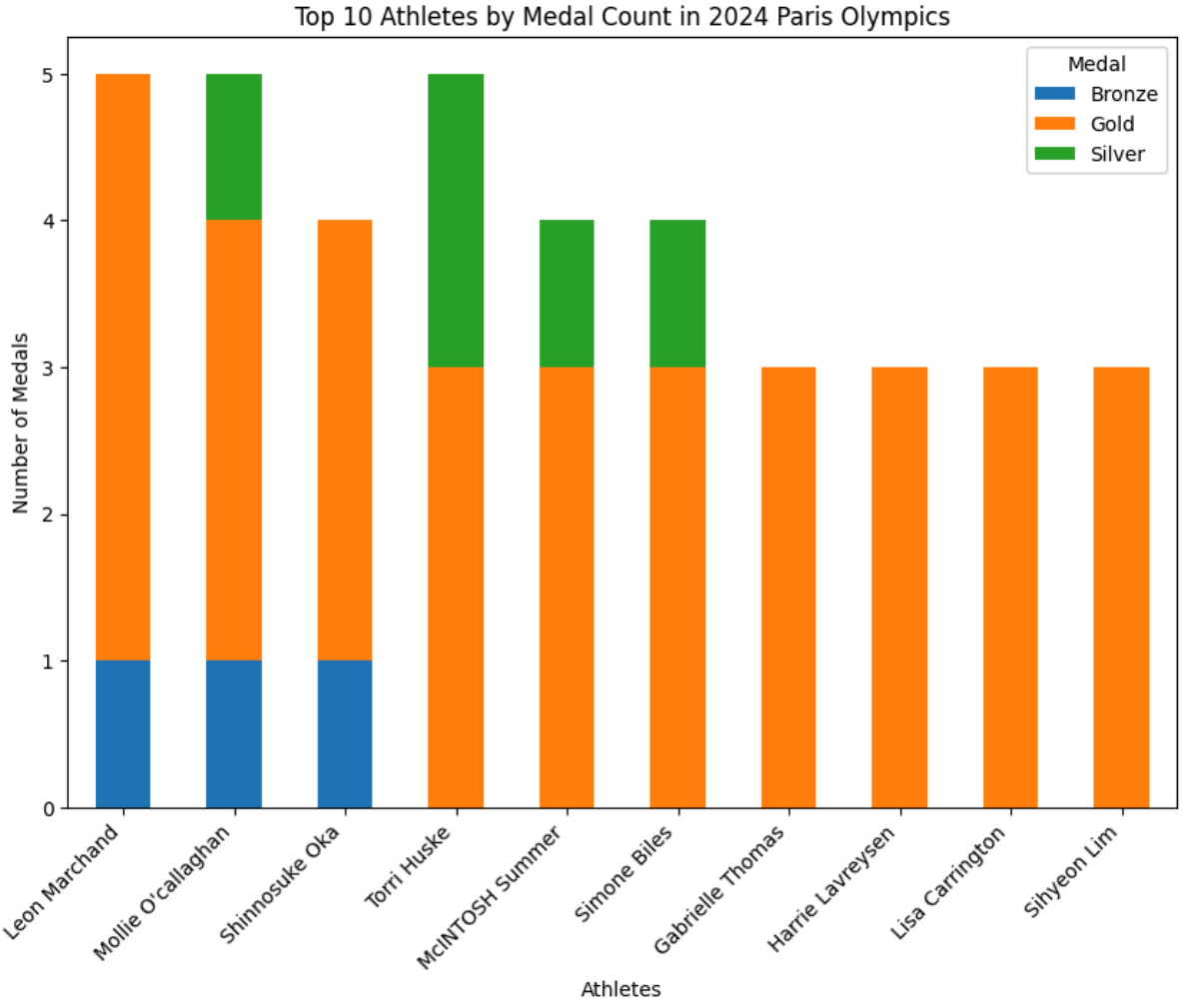
Did that correlate with the high medal count?

From the analysis, there is a clear indication that a higher number of athletes tends to correlate with a higher medal count. The USA and China, which consistently fielded the largest teams, dominated the medal tally. The steep decline in both the number of athletes and medal counts after the top 10 countries further reinforces this correlation.

However, it's important to note that other factors, such as the quality of athletes, funding, and home advantage (as seen with France), also play significant roles. While the number of athletes strongly indicates potential success, it is not the sole determinant.

Top 10 athlete contenders

Name	Medal	Bronze	Gold	Silver
Leon Marchand		1	4	0
Mollie O'callaghan		1	3	1
Shinnosuke Oka		1	3	0
Torri Huske		0	3	2
McINTOSH Summer		0	3	1
Simone Biles		0	3	1
Gabrielle Thomas		0	3	0
Harrie Lavreysen		0	3	0
Lisa Carrington		0	3	0
Sihyeon Lim		0	3	0



This chart shows the top 10 athletes by the number of medals won in the 2024 Paris Olympics, categorized by the type of medal (gold, silver, and bronze). Key insights from the data include:

Top performers are:

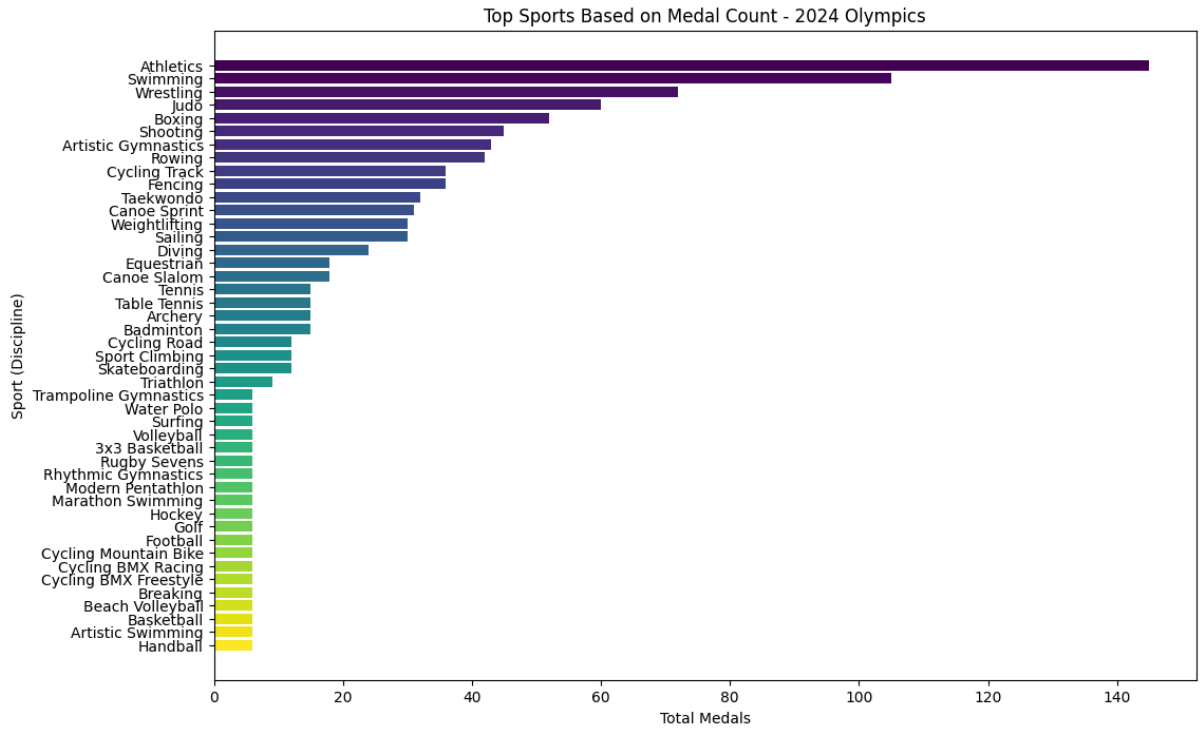
Leon Marchand leads the group with 5 medals, 4 gold, and 1 bronze. His strong performance places him at the top.

Mollie O'Callaghan follows closely with 5 medals, 3 gold, 1 silver, and 1 bronze, indicating versatility across multiple events.

Gold dominance:

Several athletes, including Gabrielle Thomas, Harrie Lavreysen, Lisa Carrington, and Sihyeon Lim, have only won gold medals, reinforcing their consistent top-tier performances without placing lower in their events.

Top sports based on medal count?



The top sport by medal count in the 2024 Paris Olympics is still athletics, by swimming. Athletics leads significantly, which might indicate a larger number of events or more competitive participation in this sport in 2024.

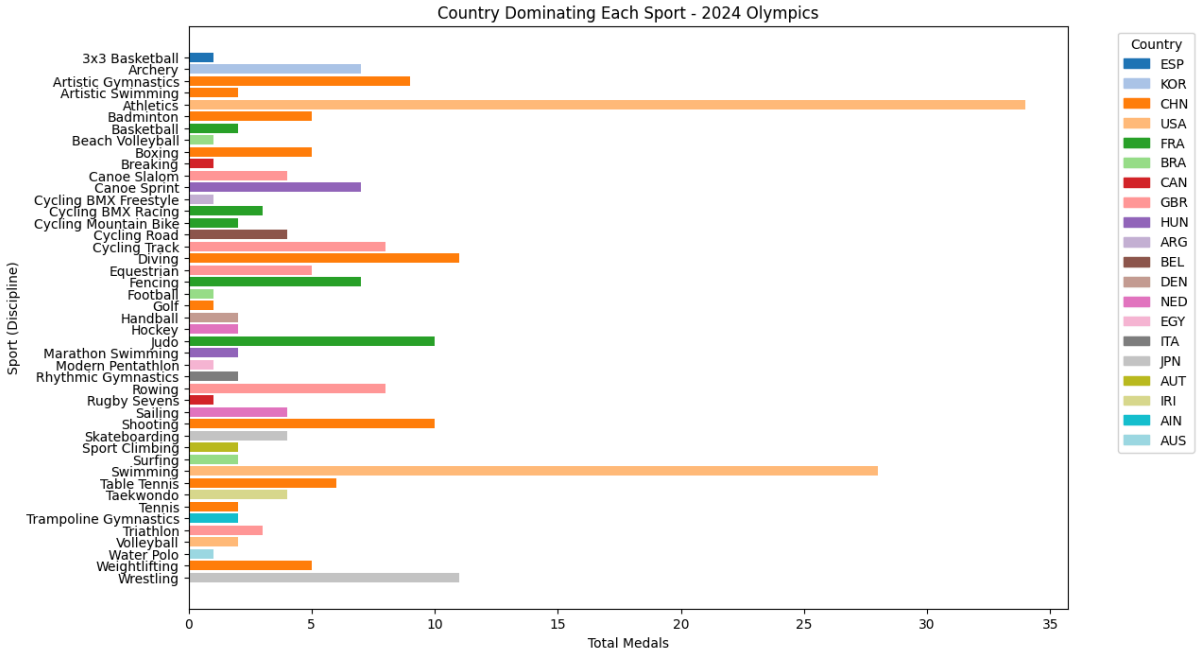
The dominance of swimming might reflect the high number of events and strong athlete participation. However, it is notable that there were concerns over the cleanliness of the Seine River, where some of the open-water swimming events took place. It would seem that this would negatively affect athletes due to health concerns which might discourage athletes from participating, potentially leading to a different sport taking the lead in medal counts. Yet,

Swimming still managed to dominate, which underscores the resilience and determination of the athletes despite the challenging conditions.

Similarly, in the 2020 Tokyo Olympics, athletics was the top sport, with swimming coming in second. The Swimming events in Tokyo took place in a controlled indoor environment (the Tokyo Aquatics Centre), which ensured optimal conditions for athletes. This difference in setting could suggest that the more challenging outdoor environment in Paris influenced the dynamics of the competition.

The consistent presence of Artistic Gymnastics in the top ranks highlights this sport's continued significance and competitiveness across both Olympics.

3. What country dominated what sport?



- The USA maintained its dominance, particularly in Athletics, and Swimming.
- Japan was a stronger competitor for wrestling and China took the lead in Diving and shooting.
- France, as the host country, made notable strides in sports like Judo, Fencing Cycling, surfing, and Triathlon, which might reflect a home advantage.

Summary

In summary, the 2024 Paris Olympics showcased a more pronounced concentration of medals among the top-performing nations, with the USA and China maintaining their dominance from the 2020 Tokyo Olympics. The absence of Russia significantly impacted the distribution, while France, as the host nation, saw a remarkable rise, likely benefiting from home advantage.

Notably, Korea emerged as a new contender in the top 10, while Italy remained consistent in its performance. Sports like Athletics and Swimming continued to lead in terms of medal counts, reinforcing their prominence in the Olympic landscape. Interestingly, there were shifts in gender participation, with sports like Boxing seeing equal male-female participation for the first time, signaling a positive change toward gender equality.

The data also reaffirmed the correlation between a larger athlete delegation and a higher medal count, but it's clear that factors such as athlete quality, funding, and home advantage play crucial roles in a nation's success. The resilience of athletes, particularly in Swimming, despite challenges like the Seine River conditions, and the continued excellence in Artistic Gymnastics, highlights the evolving nature of Olympic competition.

Ultimately, while familiar powerhouses continue to lead, the 2024 Paris Olympics also saw shifting dynamics, with emerging nations and evolving sports contributing to a diverse and exciting Olympic Games.

5.4 Summer Olympic Trends Over Time

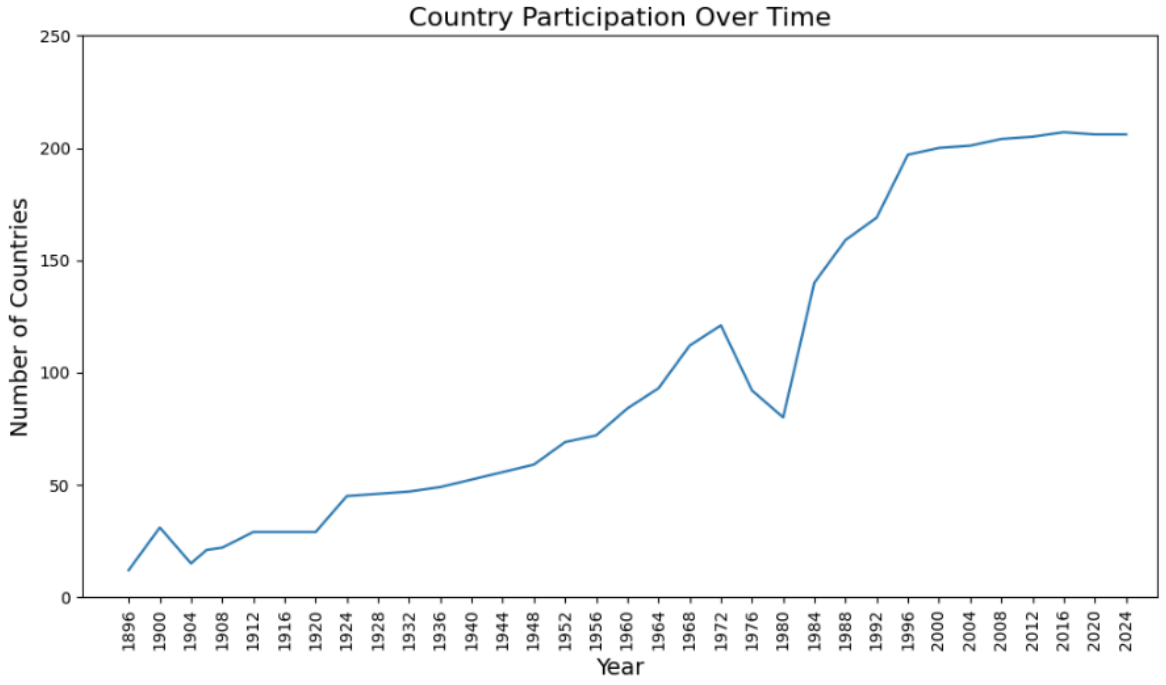
Questions:

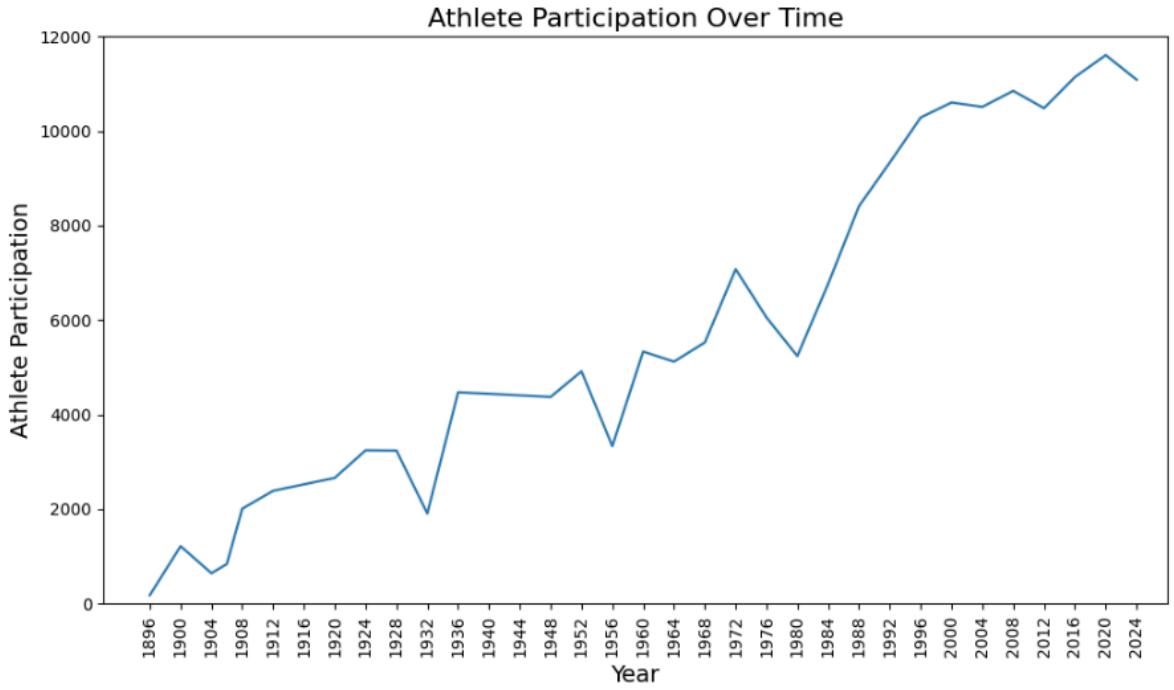
1. How has country participation in the olympic summer games changed over time? How have major world events affected participation?
2. How has athlete participation in the olympic summer games changed over time?
3. Compare trends in male and female participation over the last 124 years
4. How has sport participation and types of sports included changed over time?

The first occurrence of the modern Olympic games was held in 1896 in Greece. The first Games brought together over 12 nations and more than 240 athletes. The event featured 9 sports with about 50 events. Only men were allowed to compete. The modern summer games have occurred every 4 years since 1896 with a few exceptions. The most notable exceptions occurred in the early 1900s with no games between 1912 and 1920 then again between 1936 and 1948. These periods line up with the occurrences of World War 1 and 2.

The number of countries and athletes that have participated in the Olympic Games since 1896 has exploded over the past 124 years. The figure below shows the number of countries participating in the summer games from 1896 to 2020. The number of countries participating increased from 12 in 1986 to over 200 in 2000. After 2000, the countries participating leveled off around 205.

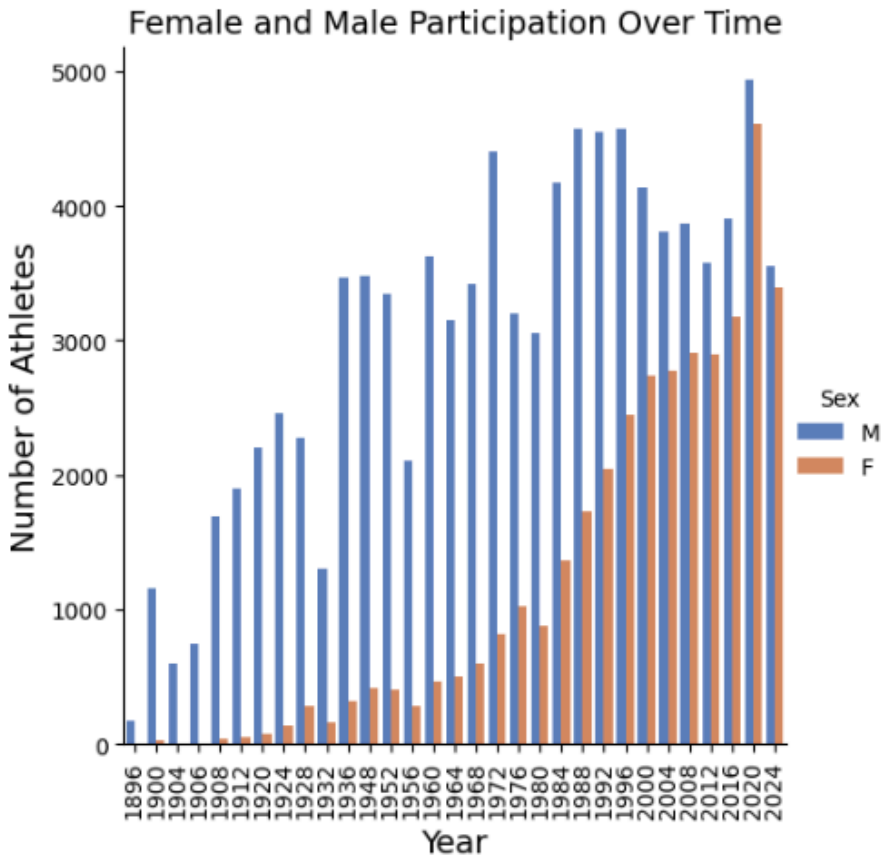
Country participation increased steadily from 1900 to 1976. Participation in the 1976 games decreased 30% from the 1972 games. Many nations boycotted New Zealand's participation in the games due to the country's apartheid system. The downward trend continued in the 1980. Participation decreased another 15% (or 51% from 1972) as many countries boycotted the Soviet Union's participation due to their invasion of Afghanistan.



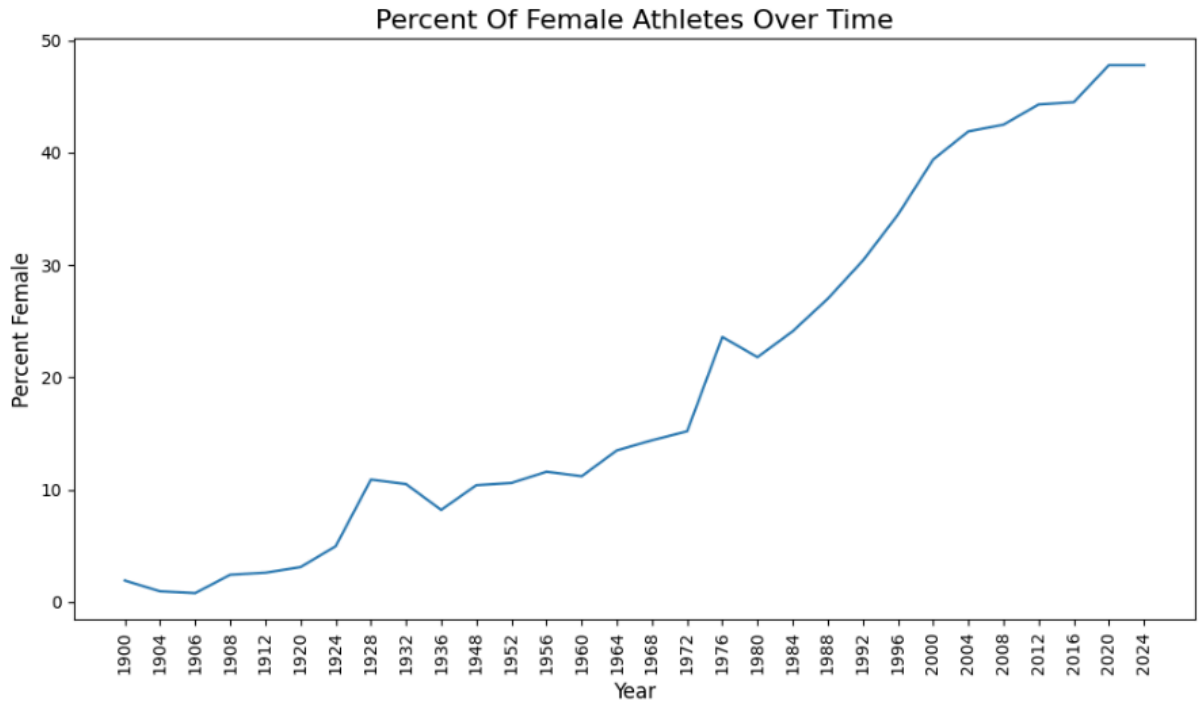


The number of athletes participating follows a similar trend as country participation. There is a steady increase from 1904 to 1932 where participation decreases sharply by 68%. This is likely due to the worldwide effect of the Great Depression. In addition, the games were held in Los Angeles. This could have contributed to the low participation as travel costs for many athletes may have been too high. Participation decreased again in 1956 due to various boycotts including the Netherlands, Spain, and Switzerland boycott of the Soviet Union's invasion of Hungary. Participation decreased 1976-1980 due to boycotts. After 1980, participation increased steadily at a 5-10% increase per event then levelled off around 2000, with over 10,000 athletes participating at each event.

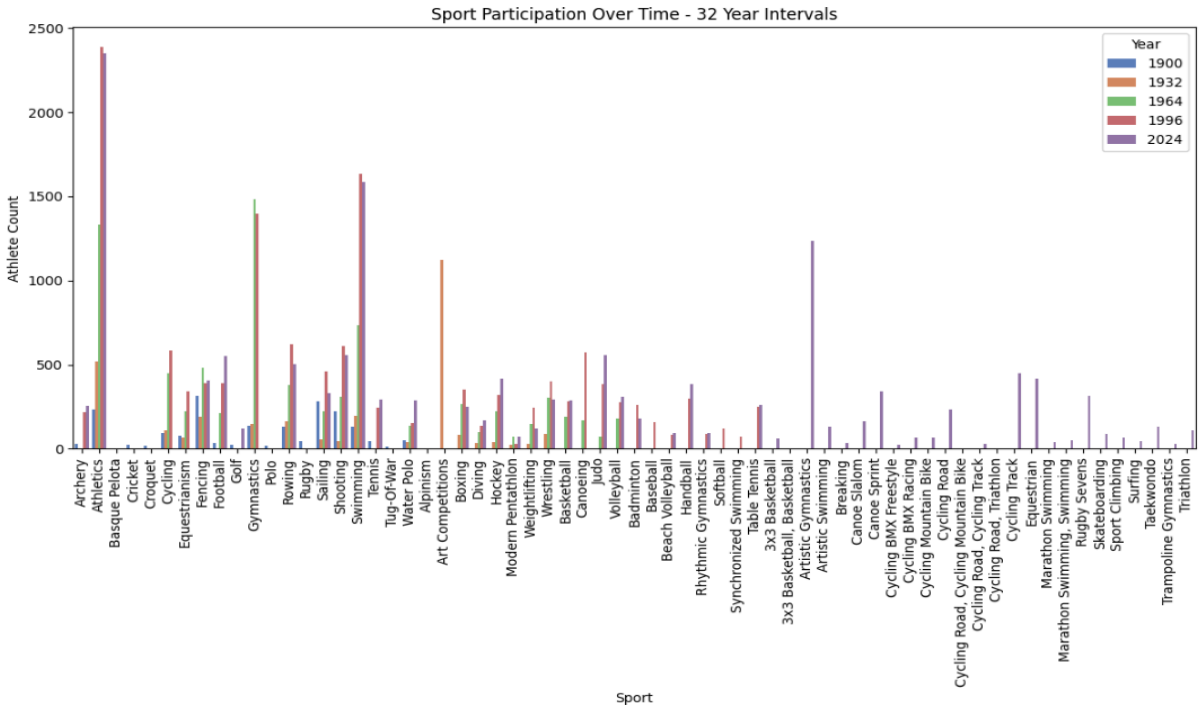
Historically, the Olympic games have been dominated by men. Women were not allowed to compete in the first-ever modern Olympics in 1896. However, women were allowed to compete in 1900. The graph below compares male and female athlete participation over time. Male participation in the games has always been greater than female participation. However, female participation has increased tremendously since the early years of the games.



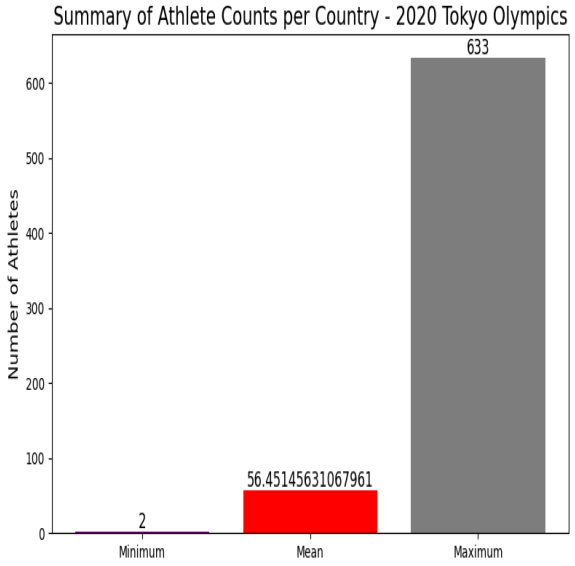
In 1900, only 23 women competed which was less than 2% of the total participation. Over the next 124 years, the percent of women competing increased from less than 2% to over 47%.

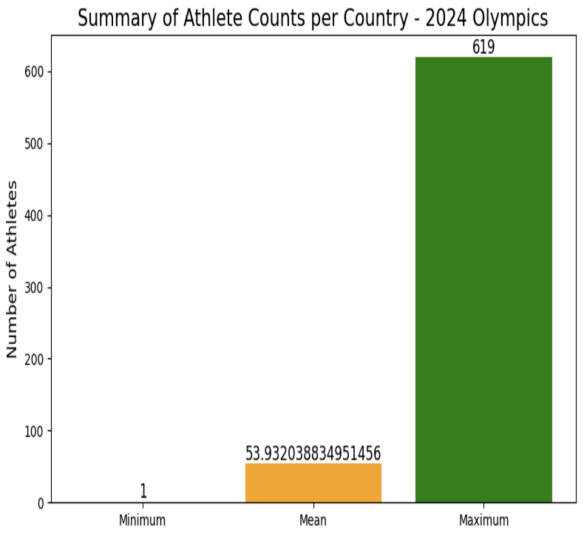


In the early years of the games, the number of sports was limited to only 9 types of sports. As the game's popularity grew, the types of sports also grew. In 2024, there were 50 types of sports included in the games. The figure below shows how participation in the type of sport has changed and how sport categories have expanded over time. Throughout time, athletics, swimming, and gymnastics has had the highest athlete participation. Other sports like Tug-O-War, art competitions, cricket, and croquet were only available in the early 1900's. Many sports were added between 1964 and 1996. Sports continued to be added between 1996 and 2024.



Descriptive statistics:





In both Summer Olympics, the maximum and minimum values for the number of athletes per country were in a similar range. The minimum was 1 or 2 athletes, and the maximum was in the 600 range with 2020 having a max count of 633 and 2024 showing a lower max count of 619. The mean value has a minor difference .For the 2020 Tokyo Olympics, the average was 56 athletes per country, while in 2024, it dropped to an average of 54 athletes per country.

Classification Model:

Can we predict future Countries' medal wins?

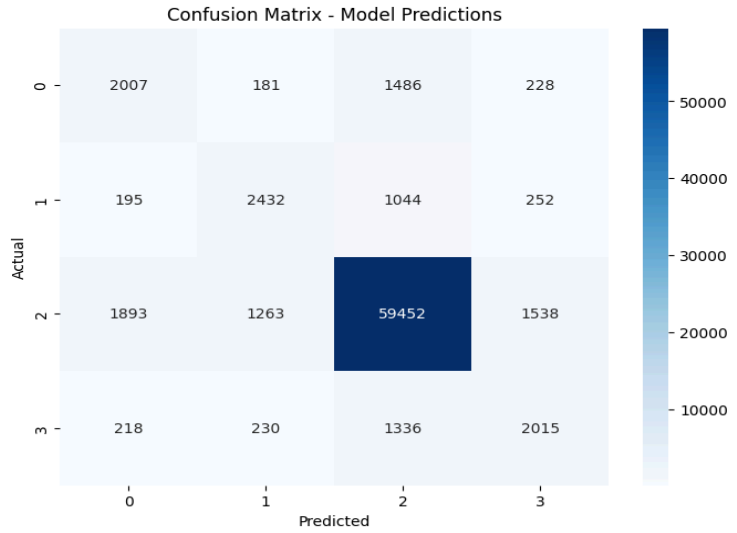
With a wealth of data extending back to the 1800s, this historical dataset provides a rich foundation for training classification models. In this case, a decision tree was developed to assess the potential of machine learning techniques in predicting future medal outcomes for countries, classifying them of no medals, Gold, Silver, or Bronze.

```
Confusion Matrix:
[[ 2007   181  1486   228]
 [  195  2432  1044   252]
 [ 1893  1263 59452  1538]
 [  218   230  1336  2015]]

Classification Report:
              precision    recall  f1-score   support

           0       0.47      0.51      0.49      3902
           1       0.59      0.62      0.61      3923
           2       0.94      0.93      0.93     64146
           3       0.50      0.53      0.51      3799

    accuracy                           0.87     75770
   macro avg       0.62      0.65      0.64     75770
weighted avg       0.87      0.87      0.87     75770
```



Analysis of the Decision Tree Model Performance
Class 0: No Medal
Class 1: Bronze
Class 2: Silver
Class 3: Gold

Confusion Matrix:

The confusion matrix shows how well the model classified each class (0: No Medal, 1: Bronze, 2: Silver, 3: Gold). The largest block in the matrix corresponds to class **2 (Silver)**, which has the highest number of correctly classified instances (59,452). However, there are notable misclassifications for other classes, especially in class **0 (No Medal)** and class **3 (Gold)**:

- Class **0 (No Medal)**: Out of 3,902 actual instances, 2,007 were correctly predicted, with significant confusion with class **2 (Silver)** (1,486 misclassified).
- Class **1 (Bronze)**: Out of 3,923 actual instances, 2,432 were correctly predicted, with 1,044 misclassified as class 2 (Silver).
- Class **3 (Gold)**: Out of 3,799 actual instances, 2,015 were correctly predicted, with 1,336 misclassified as class 2 (Silver).

Classification Report:

The overall accuracy of the model is **87%**, but there is a clear imbalance in the model's performance across different classes:

- **Class 2 (Silver)** has the highest precision, recall, and F1-score (all around 0.93). This shows that the model performs exceptionally well in identifying Silver medalists, likely due to the large number of samples in this class (64,146).

- **Class 0 (No Medal)** and **Class 3 (Gold)** have lower precision and recall values (~0.50). The model struggles more to distinguish these classes, possibly due to fewer samples or higher class overlap with Silver medalists.
- **Class 1 (Bronze)** has moderate performance with a precision and recall around 0.60, indicating some level of confusion between Bronze and Silver medalists.

Decision Tree

To reduce the complexity of the decision tree and prevent overfitting, We implemented **pre-pruning** by setting the maximum depth to 3 (max_depth=3) and requiring at least 50 samples to split a node (min_samples_split=50). These parameters ensure that the model captures meaningful patterns without focusing on noise or outliers.

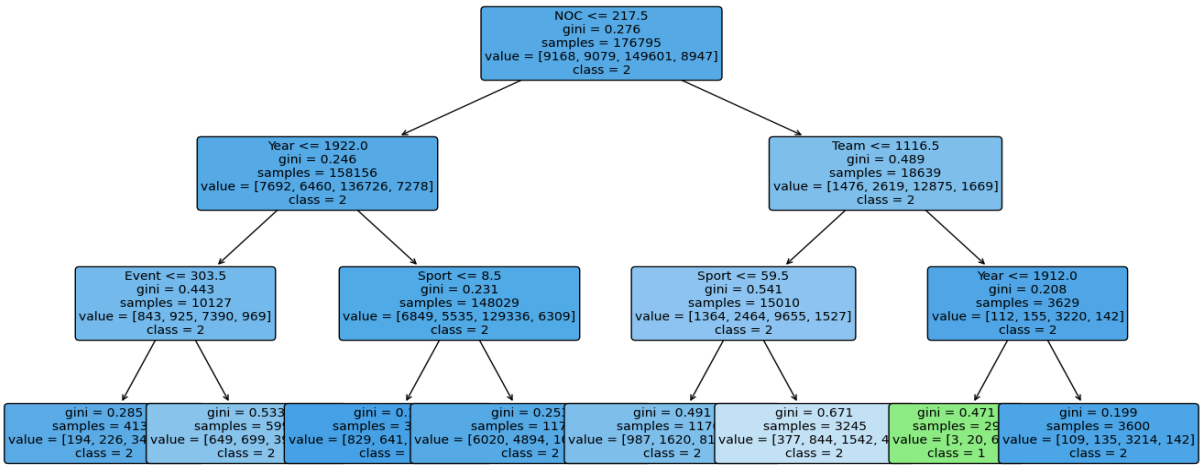
After retraining the pruned decision tree, I visualized the resulting structure using Scikit-learn's plot_tree function, with the nodes filled and rounded for clarity. The tree's reduced depth made it easier to interpret, showing clear decision paths for the classification task. This approach resulted in a simpler and more generalizable model, which performed better on unseen data compared to an unpruned tree

The decision tree is largely dominated by **class 2 (Silver)**, which reflects a strong class imbalance in the dataset.

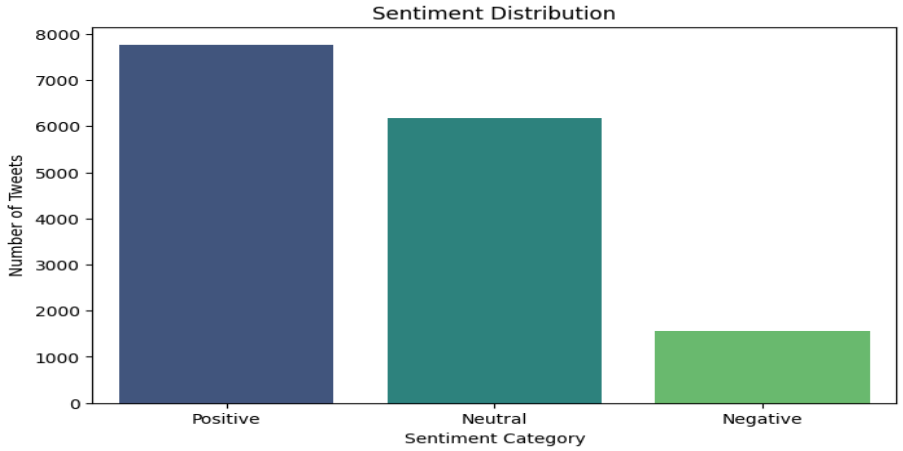
The tree uses NOC, Year, Event, Sport, and Team features to split the data and reduce Gini impurity.

The splits in the left subtree (based on Year) achieve lower Gini impurities, resulting in purer nodes. In contrast, the right subtree (based on Team) has a higher Gini impurity, indicating more class mixing.

Class 2 (Silver) is the predicted class in most nodes, except for one node with a class 1 (Bronze)



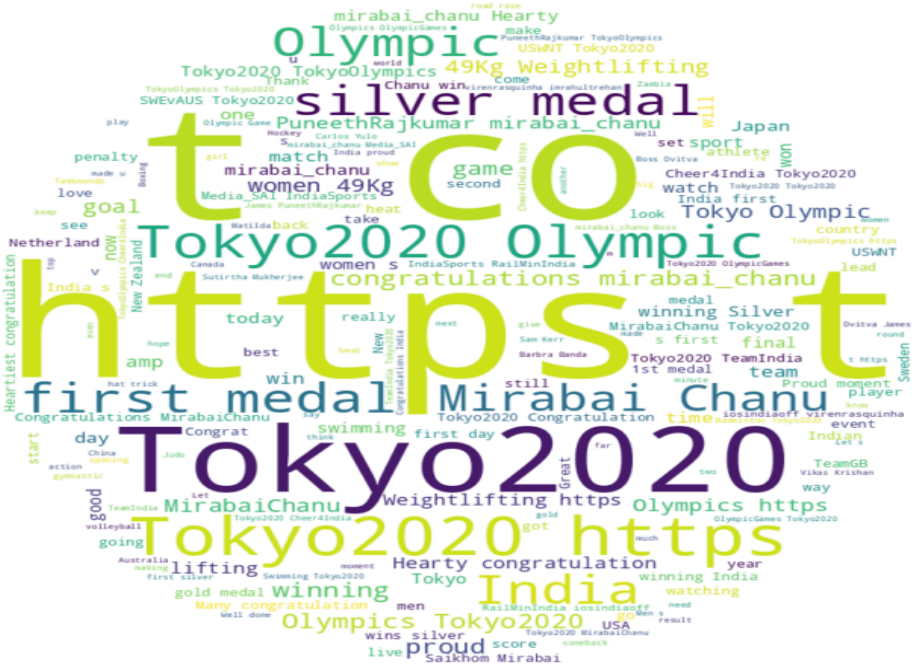
Sentiment Distribution Analysis:



- The bar chart shows the **sentiment distribution** for the tweets related to the **Tokyo 2020 Olympics**.
- Most tweets express **positive sentiment**, with around 7,800 tweets categorized as positive.
- **Neutral sentiment** comes next, with approximately 6,800 tweets, indicating a significant number of tweets with neither strong positive nor negative sentiments.
- **Negative sentiment** is the least frequent, with just under 2,000 tweets.

This suggests that the overall conversation around the Tokyo 2020 Olympics is largely positive or neutral, with relatively few negative remarks.

Word Cloud Analysis:



- The word cloud highlights the most frequently used words in the tweets.
- The most prominent words include "Tokyo2020," "Olympics," "medal," "Mirabai Chanu," and "India", indicating that these terms are central to the discussions.
- Specific references to athletes like Mirabai Chanu, as well as phrases such as "Silver medal" and "first medal," suggest that these events and achievements were popular topics.
- The presence of terms like "t.co" and "https" indicates that many tweets included links, possibly sharing news articles, results, or media related to the event.

Overall, the positive sentiment and frequent mentions of achievements and key figures point to a celebratory tone surrounding the 2020 Olympics on social media.