🏆 Olympics: Exploratory Data Analysis 🏆

Table of Contents :

1. Introduction

```
Welcome to a deep dive into the history and intricacies of the Summer Olympics (1896-2024)!
```

The Olympic Games is a prestigious international sporting event held every four years. The event brings together athletes from various nations to compete in a wide range of sports and disciplines.

The Olympics transcend borders, bringing athletes from diverse nations to showcase their talents. This notebook explores the evolution, milestones, and trends of the Summer Olympics.

2. Importing Libraries

In [3]:

```python
import numpy as np
import pandas as pd
import seaborn as sns
custom_params = {"axes.spines.right": False, "axes.spines.top": False}
sns.set_theme(style="ticks", rc=custom_params)
import matplotlib.pyplot as plt
%matplotlib inline
from wordcloud import WordCloud
from PIL import Image
import warnings
warnings.filterwarnings("ignore")
print("✅Libraries Imported Successfully")
```

```
✅Libraries Imported Successfully
```

3. Reading Data

In [5]:
```
# we will be combining the following datasets for our analysis
df = pd.read_csv("olympics_project/olympics_dataset.csv")                          # main dataset
noc = pd.read_csv("olympics_project/NOC_regions.csv")                            # for accurate 'region'
```

In [6]:
```
# random 3 rows
df.sample(3)
```

Out[6]:

	player_id	Name	Sex	Team	NOC	Year	Season	City	Sport	Event	Medal
89826	108776	Jo Yun-jeong	F	South Korea	KOR	2000	Summer	Sydney	Tennis	Tennis Women's Doubles	No medal
48878	58169	Joanne Dow	F	United States	USA	2008	Summer	Beijing	Athletics	Athletics Women's 20 kilometres Walk	No medal
205388	250022	Rupeni Varea	M	Fiji	FIJ	1996	Summer	Atlanta	Weightlifting	Weightlifting Men's Light-Heavyweight	No medal

In [7]:
```
# detailed information of each column - dtype, non-null values etc.
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 252565 entries, 0 to 252564
Data columns (total 11 columns):
 #   Column     Non-Null Count   Dtype 
---  ------     --------------   ----- 
 0   player_id  252565 non-null  int64 
 1   Name       252565 non-null  object
 2   Sex        252565 non-null  object
 3   Team       252565 non-null  object
 4   NOC        252565 non-null  object
 5   Year       252565 non-null  int64 
 6   Season     252565 non-null  object
 7   City       252565 non-null  object
 8   Sport      252565 non-null  object
 9   Event      252565 non-null  object
 10  Medal      252565 non-null  object
dtypes: int64(2), object(9)
memory usage: 21.2+ MB
```

In [8]:

```
# descriptive statistics of numeric columns
df.describe()
```

Out[8]:

	player_id	Year
count	2.525650e+05	252565.000000
mean	2.305499e+05	1981.743908
std	4.289330e+05	32.596548
min	0.000000e+00	1896.000000
25%	5.713700e+04	1960.000000
50%	1.356110e+05	1988.000000
75%	2.118590e+05	2008.000000
max	9.460001e+06	2024.000000

In [9]:
```python
# rows and columns size
df.shape
```

Out[9]:
```
(252565, 11)
```

In [11]:
```python
# random 3 rows of NOC (region data)
noc.sample(3)
```

Out[11]:

	NOC	region	notes
107	KGZ	Kyrgyzstan	NaN
52	CUB	Cuba	NaN
24	BIZ	Belize	NaN

4. Processing Data

In [14]:
```python
# joining the two datasets on 'NOC'
df = df.merge(noc, on="NOC", how="left")
```

In [15]:
```python
# filtering data since we will only be doing analysis on summer olympics
df = df[df['Season']=="Summer"]
```

In [16]:
```python
df.columns
```

Out[16]:
```
Index(['player_id', 'Name', 'Sex', 'Team', 'NOC', 'Year', 'Season', 'City',
       'Sport', 'Event', 'Medal', 'region', 'notes'],
      dtype='object')
```

In [17]:
```python
# dropping unnecessary column
df.drop(columns='notes', inplace=True)
```

In [42]:
```python
# renaming column 'region' to 'Country'
df.rename(columns={'region': 'Country'}, inplace=True)
```

```
df.rename(columns={'player_id': 'ID'}, inplace=True)
```

In [19]:
```
# dropping duplicate values
df.duplicated().sum()
```

Out[19]: 0

In [21]:
```
# checking for null values
missing_data_df = pd.DataFrame({
    'Count': df.isnull().sum(),
    'Percentage': round((df.isnull().sum() * 100 / len(df)), 2).values
 })
# Displaying the table
print(missing_data_df)
```

```
           Count  Percentage
player_id      0        0.00
Name           0        0.00
Sex            0        0.00
Team           0        0.00
NOC            0        0.00
Year           0        0.00
Season         0        0.00
City           0        0.00
Sport          0        0.00
Event          0        0.00
Medal          0        0.00
Country     1139        0.45
```

In [31]:
```
# Find all athletes in the Refugee Olympic Team
refugee_athletes = df[df['NOC'] == 'ROT']
print(refugee_athletes)
```

	player_id	Name	Sex	Team	NOC	Year	\
5523	6267	Paulo Lokoro	M	Refugee Olympic Athletes	ROT	2016	
6828	7908	Rami Anis	M	Refugee Olympic Athletes	ROT	2016	
6829	7909	Rami Anis	M	Refugee Olympic Athletes	ROT	2016	
18062	21529	Yiech Biel	M	Refugee Olympic Athletes	ROT	2016	
26367	31708	Mabika Bukasa	F	Refugee Olympic Athletes	ROT	2016	
33708	40238	James Chiengjiek	M	Refugee Olympic Athletes	ROT	2016	
98467	119392	Yonas Kinde	M	Refugee Olympic Athletes	ROT	2016	
116168	141669	Anjelina Lohalith	F	Refugee Olympic Athletes	ROT	2016	
116229	141753	Rose Lokonyen	F	Refugee Olympic Athletes	ROT	2016	
122530	149306	Yusra Mardini	F	Refugee Olympic Athletes	ROT	2016	
122531	149307	Yusra Mardini	F	Refugee Olympic Athletes	ROT	2016	
131684	160069	Popole Misenga	M	Refugee Olympic Athletes	ROT	2016	

	Season	City	Sport	\
5523	Summer	Rio de Janeiro	Athletics	
6828	Summer	Rio de Janeiro	Swimming	
6829	Summer	Rio de Janeiro	Swimming	
18062	Summer	Rio de Janeiro	Athletics	
26367	Summer	Rio de Janeiro	Judo	
33708	Summer	Rio de Janeiro	Athletics	
98467	Summer	Rio de Janeiro	Athletics	
116168	Summer	Rio de Janeiro	Athletics	
116229	Summer	Rio de Janeiro	Athletics	
122530	Summer	Rio de Janeiro	Swimming	
122531	Summer	Rio de Janeiro	Swimming	
131684	Summer	Rio de Janeiro	Judo	

	Event	Medal	Country
5523	Athletics Men's 1,500 metres	No medal	NaN
6828	Swimming Men's 100 metres Freestyle	No medal	NaN
6829	Swimming Men's 100 metres Butterfly	No medal	NaN
18062	Athletics Men's 800 metres	No medal	NaN
26367	Judo Women's Middleweight	No medal	NaN
33708	Athletics Men's 400 metres	No medal	NaN
98467	Athletics Men's Marathon	No medal	NaN
116168	Athletics Women's 1,500 metres	No medal	NaN
116229	Athletics Women's 800 metres	No medal	NaN
122530	Swimming Women's 100 metres Freestyle	No medal	NaN
122531	Swimming Women's 100 metres Butterfly	No medal	NaN
131684	Judo Men's Middleweight	No medal	NaN

In [32]:
```
# Identify possible Mixed Teams by checking missing region before 1920
mixed_teams = df[(df['Country'].isna()) & (df['Year'] < 1920)]
print(mixed_teams)
```

```
        player_id          Name Sex     Team  NOC  Year  Season       City  \
51268       61080  Fritz Eccard   M  Unknown  UNK  1912  Summer  Stockholm
107183     130721     A. Laffen   M  Unknown  UNK  1912  Summer  Stockholm

                   Sport                               Event     Medal  \
51268   Art Competitions  Art Competitions Mixed Architecture  No medal
107183  Art Competitions  Art Competitions Mixed Architecture  No medal

       Country
51268      NaN
107183     NaN
```

- The data is easy to understand.
- Most of the columns will be helpful in analyzing the data.
- The data type of each column is appropriate.
- Only **"Country"** field has null values. Because some athletes have missing country values because they competed as Refugee Athletes (ROT) under the Olympic flag or were part of Mixed Teams (ZZX) with members from multiple countries.

In [35]:
```
# adding 4 new columns from 'Medal'- Gold, Silver, Bronze, No medal
pd.get_dummies(df['Medal'])
```

Out[35]:

	Bronze	Gold	No medal	Silver
0	False	False	True	False
1	False	False	True	False
2	False	False	True	False
3	False	True	False	False
4	False	False	True	False
...	...	...	...	...
252560	False	False	True	False
252561	False	False	True	False
252562	False	True	False	False
252563	True	False	False	False
252564	True	False	False	False

252565 rows × 4 columns

In [36]:
```
df = pd.concat([df, pd.get_dummies(df['Medal'])], axis=1)
```

In [37]:
```
# new df shape
df.shape
```

Out[37]:
```
(252565, 16)
```

5. Analysis and Inerences

5.1 Primary Analysis

In [38]:
```
# years when summer olympics were held
years = df['Year'].unique()
years.sort()
print(f"Olympics were introduced in {years[0]}; since then, {len(years)} summer olympics have been held.")
```

```
Olympics were introduced in 1896; since then, 31 summer olympics have been held.
```

In [39]:
```
df['City'].nunique()                            # Number of host cities
```

Out[39]:
```
23
```

In [40]:
```
df['Country'].nunique()                         # Number of countries participated so far
```

Out[40]:
```
205
```

In [43]:
```
df['ID'].nunique()                              # Numer of participants so far
```

Out[43]:
```
235903
```

5.2 Medal Statistics

In [44]:
```
df.groupby('NOC').sum()[['Gold', 'Silver', 'Bronze']].sort_values('Gold', ascending=False)
```

Out[44]:

	Gold	Silver	Bronze
NOC			
USA	2716	1539	1366
URS	832	635	596
GBR	716	813	753
GER	634	613	721
FRA	583	712	660
...	...	...	...
LBR	0	0	0
LES	0	0	0
LIB	0	2	2
LIE	0	0	0
LBA	0	0	0

234 rows × 3 columns

- While rechecking values online, it was noticed that the count of medals was not matching.
- It seems like each medal earned in a team event was counted separately.

  -----> For example, if India won a gold medal in hockey (a team of 11 plus extras), all the medals would be counted separately instead of one.
- To overcome this issue, we'll focus only on a few selected columns and try to remove the duplicated rows to get an accurate medal count.

In [45]:

```
df['Year'].min()
```

Out[45]: 1896

In [46]:
```
# demonstrating the issue
df[(df['NOC']== 'IND') & (df['Medal']=='Gold')].head()
```

Out[46]:

	ID	Name	Sex	Team	NOC	Year	Season	City	Sport	Event	Medal	Country	Bronze	Gold	No medal	Silver
4186	4732	Shaukat Ali	M	India	IND	1928	Summer	Amsterdam	Hockey	Hockey Men's Hockey	Gold	India	False	True	False	False
4190	4736	Syed Ali	M	India	IND	1964	Summer	Tokyo	Hockey	Hockey Men's Hockey	Gold	India	False	True	False	False
4460	5032	Richard Allen	M	India	IND	1928	Summer	Amsterdam	Hockey	Hockey Men's Hockey	Gold	India	False	True	False	False
4461	5033	Richard Allen	M	India	IND	1932	Summer	Los Angeles	Hockey	Hockey Men's Hockey	Gold	India	False	True	False	False
4462	5034	Richard Allen	M	India	IND	1936	Summer	Berlin	Hockey	Hockey Men's Hockey	Gold	India	False	True	False	False

- Focusing on the top 5 rows, **ID 4732 and ID 5032** won gold medals in **Hockey** at the **1928** Summer Olympics. This should be counted as one medal and not individually. Hence, the medal count is to be corrected.

In [49]:
```
# Creating a dataframe for the correct medal count
# droping duplicates
medals = df.drop_duplicates(subset=['Team', 'NOC', 'Year', 'City', 'Sport', 'Event', 'Medal'])
```

In [99]:
```
medal_tally = medals.groupby('Country').sum()[['Gold', 'Silver', 'Bronze']].sort_values('Gold', ascending=False).reset_index()
```

In [100…
```
medal_tally['Total']= medal_tally['Gold'] + medal_tally['Silver'] + medal_tally['Bronze']
```

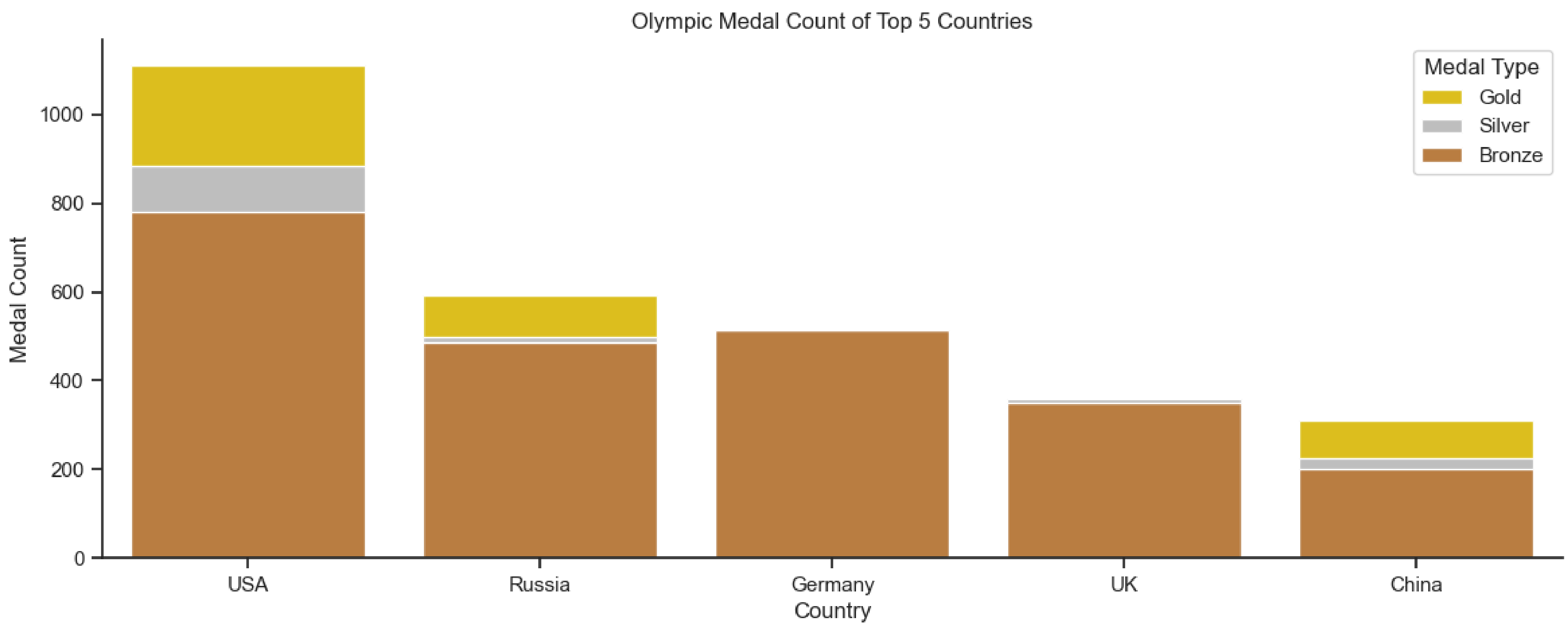
5.2.1 Top 5 Countries of All Time

In [101…
```
top5 = medal_tally.head()
top5
```

Out[101…

	Country	Gold	Silver	Bronze	Total
0	USA	1113	885	782	2780
1	Russia	592	498	487	1577
2	Germany	465	480	515	1460
3	UK	313	360	350	1023
4	China	309	224	201	734

In [53]:
```
plt.figure(figsize=(14, 5))
sns.barplot(x='Country', y='Gold', data=top5, color='gold', label='Gold')
sns.barplot(x='Country', y='Silver', data=top5, color='silver', label='Silver')
sns.barplot(x='Country', y='Bronze', data=top5, color='#cd7f32', label='Bronze')
plt.title('Olympic Medal Count of Top 5 Countries')
plt.xlabel('Country')
plt.ylabel('Medal Count')
plt.legend(title='Medal Type')
plt.show()
```



5.3 Participation trend over the years

In [54]:
```
# participation of countries wrt year
nations_over_time = df.drop_duplicates(['Year', 'Country'])['Year'].value_counts().reset_index(name='Nations Participated').sc
nations_over_time.head(3)
```

Out[54]:

	Year	Nations Participated
30	1896	12
24	1900	31
29	1904	14

In [55]:
```
plt.figure(figsize=(14, 5))
sns.lineplot(x='Year', y='Nations Participated', data=nations_over_time, marker='o', color='#8A2BE2')
plt.title('Participating Nations over the Years')
```

```
plt.xlabel('Year')
plt.ylabel('Nations Participated')
# Show the plot
plt.show()
```

- With a humble beginning of only 12 teams participating in 1896, the Olympics has evolved into a global phenomenon, attracting the participation of more than 200 teams in the 2016 edition. Over the years, the Games have gained immense fame, becoming a symbol of international unity, athletic excellence, and cultural diversity.
- There is a dip in the participation trend line because in the year 1980, when the Olympics were held in Moscow, many countries boycotted the games due to Russia's attack on Afghanistan.

5.4 Top performance in each Edition

In [56]:
```
won_df = df[(df['Gold']==1) | (df['Silver']==1) | (df['Bronze']==1)]
```

In [57]:
```
won_df['Total'] = won_df['Gold']+won_df['Silver']+won_df['Bronze']
won_df.head()
```

Out[57]:

	ID	Name	Sex	Team	NOC	Year	Season	City	Sport	Event	Medal	Country	Bronze	Gold	No medal
3	3	Edgar Aabye	M	Denmark/Sweden	DEN	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold	Denmark	False	True	False
12	37	Arvo Aaltonen	M	Finland	FIN	1920	Summer	Antwerpen	Swimming	Swimming Men's 200 metres Breaststroke	Bronze	Finland	True	False	False
13	38	Arvo Aaltonen	M	Finland	FIN	1920	Summer	Antwerpen	Swimming	Swimming Men's 400 metres Breaststroke	Bronze	Finland	True	False	False
15	41	Paavo Aaltonen	M	Finland	FIN	1948	Summer	London	Gymnastics	Gymnastics Men's Individual All-Around	Bronze	Finland	True	False	False
16	42	Paavo Aaltonen	M	Finland	FIN	1948	Summer	London	Gymnastics	Gymnastics Men's Team All-Around	Gold	Finland	False	True	False

In [58]:
```
top_player_by_edition = won_df.groupby(['Name','Year'])['Total'].sum().sort_values(ascending=False).reset_index()
```

In [59]:
```
# top player by edition
top = top_player_by_edition.drop_duplicates(subset='Year').sort_values(by='Year')
```

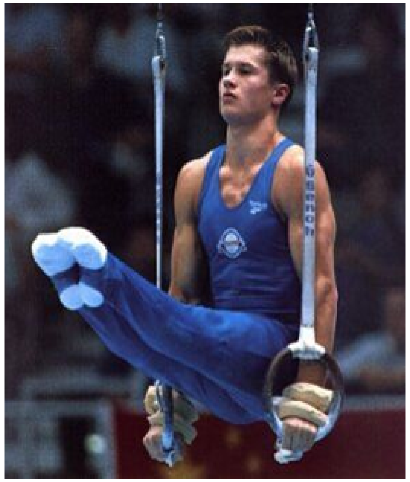
In [60]:
```
top.groupby('Name')['Year'].count().sort_values(ascending=False).reset_index(name='Count').head()
```

Out[60]:

	Name	Count
0	Michael Ii	4
1	Larysa (diriy-)	2
2	Aleksey Nemov	2
3	Aleksandr Dityatin	1
4	Matthew Biondi	1

. Larysa Semenivna Latynina, Aleksey Yuryevich Nemov, and Michael Fred Phelps have all set outstanding performances in multiple editions.

Larysa Semenivna Latynina



Aleksey Yuryevich Nemov

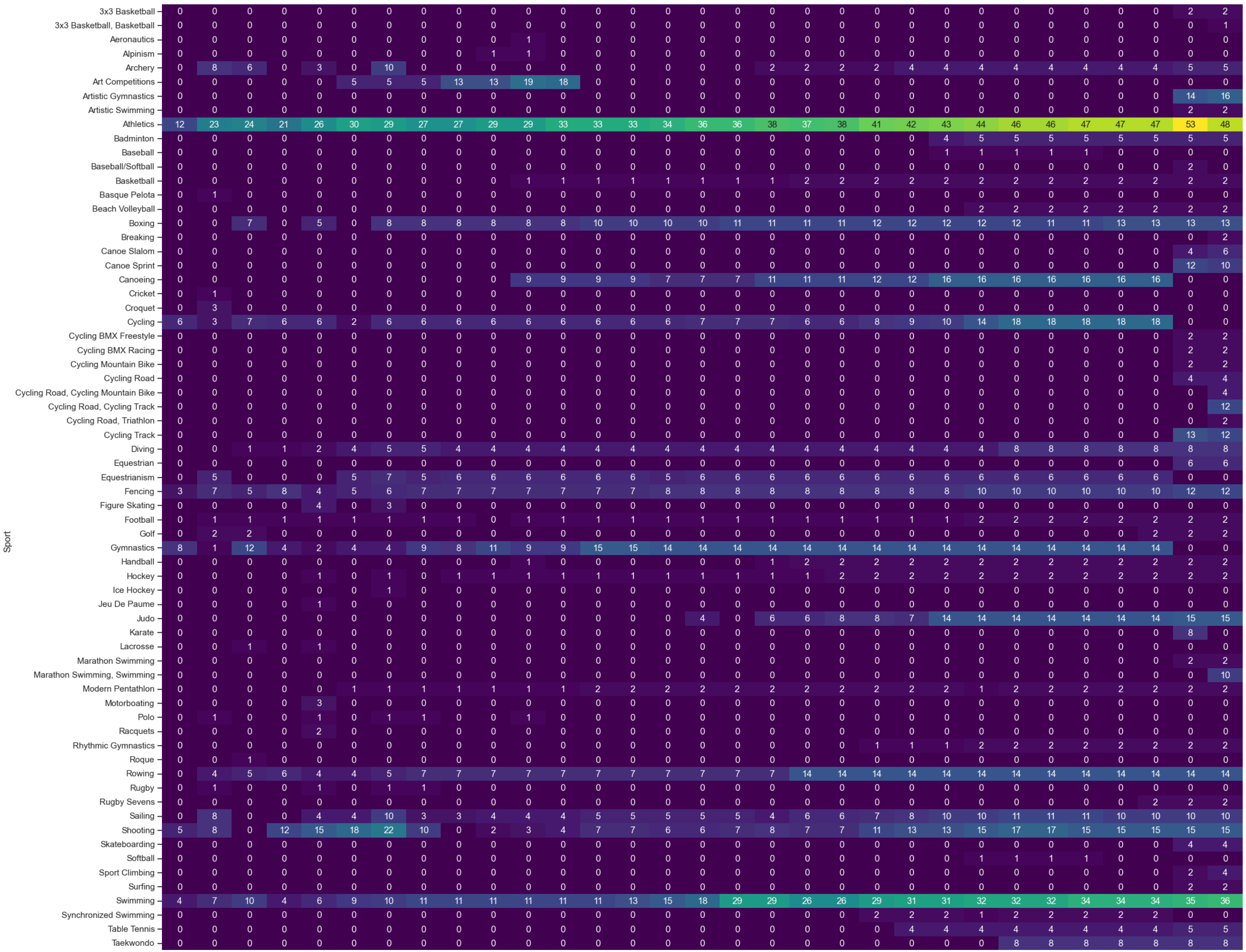
Michael Fred Phelps

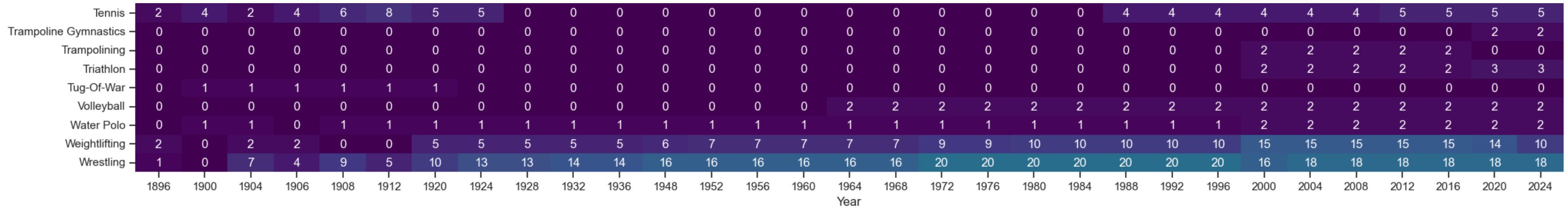
5.5 Trend analysis of events held

In [62]:
```
trend = df.drop_duplicates(['Year', 'Sport', 'Event'])
```

In [63]:
```
# Creating a heatmap illustrating the distribution of sports events over the years
event = trend.pivot_table(index='Sport', columns="Year", values='Event', aggfunc='count').fillna(0).astype(int)
plt.figure(figsize=(25,25))
```

```
sns.heatmap(data=event, annot=True, cmap='viridis', cbar=False)
plt.show()
```


Sport
3x3 Basketball
3x3 Basketball, Basketball
Aeronautics
Alpinism
Archery
Art Competitions
Artistic Gymnastics
Artistic Swimming
Athletics
Badminton
Baseball
Baseball/Softball
Basketball
Basque Pelota
Beach Volleyball
Boxing
Breaking
Canoe Slalom
Canoe Sprint
Canoeing
Cricket
Croquet
Cycling
Cycling BMX Freestyle
Cycling BMX Racing
Cycling Mountain Bike
Cycling Road
Cycling Road, Cycling Mountain Bike
Cycling Road, Cycling Track
Cycling Road, Triathlon
Cycling Track
Diving
Equestrian
Equestrianism
Fencing
Figure Skating
Football
Golf
Gymnastics
Handball
Hockey
Ice Hockey
Jeu De Paume
Judo
Karate
Lacrosse
Marathon Swimming
Marathon Swimming, Swimming
Modern Pentathlon
Motorboating
Polo
Racquets
Rhythmic Gymnastics
Roque
Rowing
Rugby
Rugby Sevens
Sailing
Shooting
Skateboarding
Softball
Sport Climbing
Surfing
Swimming
Synchronized Swimming
Table Tennis
Taekwondo



- Athletics and Swimming have shown a consistent upward trend over the years, establishing themselves as the most popular and widely participated sports to date. - Wrestling, weightlifting, shooting, rowing, judo, gymnastics, cycling, canoeing, and boxing are emerging as the next trending sports, gaining increased attention and participation.

5.6 Popularity of Olympics

- The Olympics has grown exponentially in popularity, captivating audiences worldwide with its celebration of sport, unity, and the pursuit of athletic excellence. - Below we'll see Olympics trend wrt country and gender.

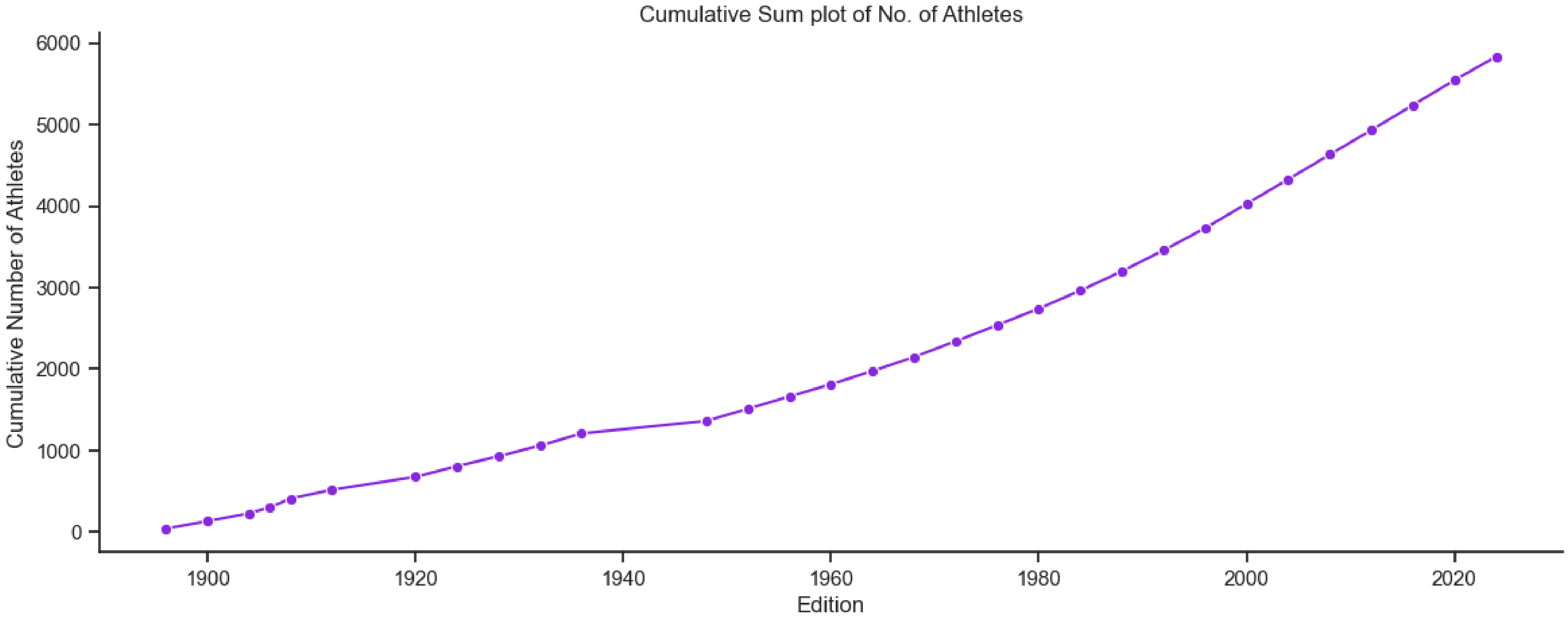
In [65]:
```
Athletes_over_time =  df.drop_duplicates(['Year', 'Event'])['Year'].value_counts().reset_index().sort_values('Year')
Athletes_over_time.head(3)
```

Out[65]:

	Year	count
30	1896	43
28	1900	90
27	1904	95

In [66]:
```
cumulative_sum = Athletes_over_time['count'].cumsum()
plt.figure(figsize=(14, 5))
sns.lineplot(x=Athletes_over_time['Year'], y=cumulative_sum, marker='o', color='#8A2BE2')
```

```
plt.title('Cumulative Sum plot of No. of Athletes')
plt.xlabel('Edition')
plt.ylabel('Cumulative Number of Athletes')
plt.show()
```



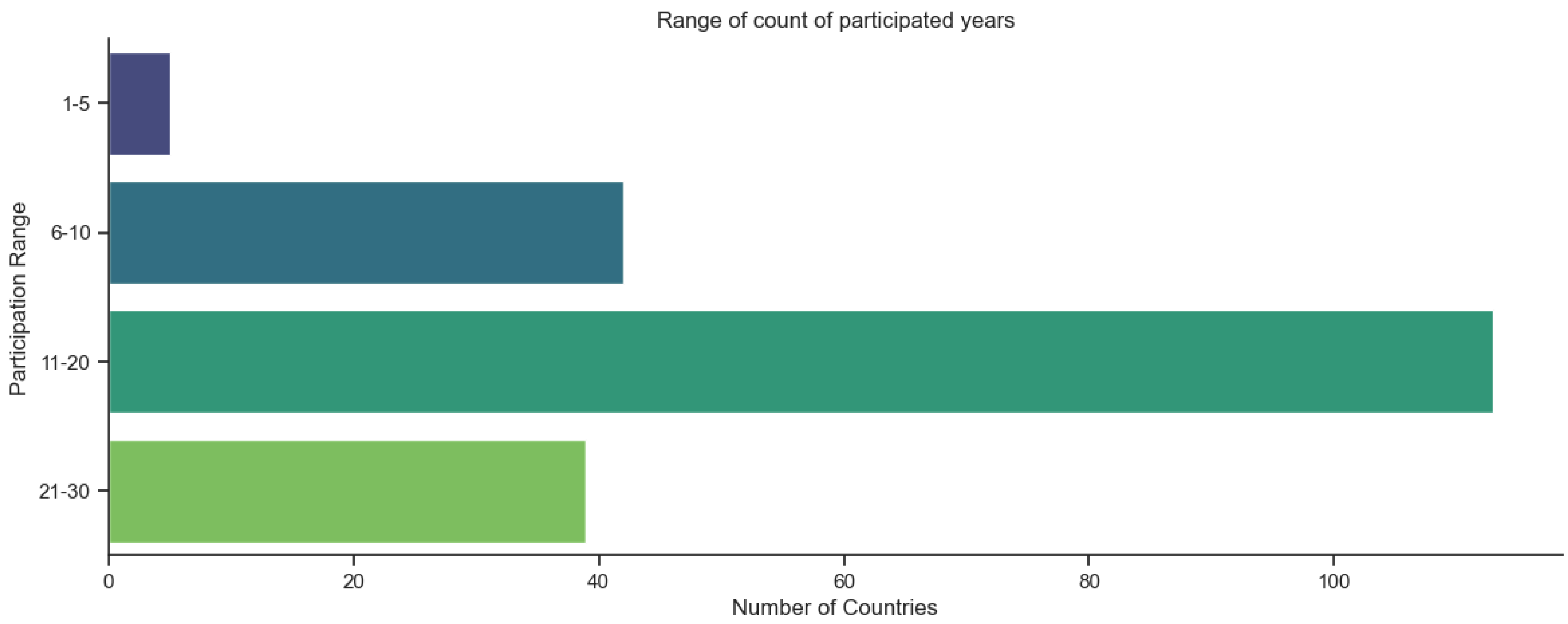
- The journey began with a modest count of 43 athletes and has now expanded to over 6000 athletes. Olympics has provided deserving athletes with the support and recognition they truly deserve.

5.6.1 Country wise Trend

In [68]:
```
country_df = df.drop_duplicates(subset=['Year', 'Country'])
```

In [69]:
```
country_trend = country_df['Country'].value_counts().reset_index(name='Count')
```

In [70]:
```
country_trend_sorted = country_trend.sort_values(by='Count', ascending=False)
clubbed_countries = pd.cut(country_trend_sorted['Count'], bins=[1, 5, 10, 20, 30], labels=['1-5', '6-10', '11-20', '21-30'])
country_trend_sorted['Clubbed'] = clubbed_countries
plt.figure(figsize=(14, 5))
sns.countplot(y='Clubbed', data=country_trend_sorted, palette='viridis')
plt.xlabel('Number of Countries')
plt.ylabel('Participation Range')
plt.title('Range of count of participated years')
plt.show()
```



In [78]:
```
# Countries which has participated in all the Editions held so far
top_participation = country_trend[country_trend['Count']==31]
top_participation
```

Out[78]:

	Country	Count
0	Switzerland	31
1	Australia	31
2	Greece	31
3	UK	31
4	France	31
5	Italy	31

- Notably, Switzerland, Australia, Greece, the United Kingdom, France, and Italy have demonstrated unwavering commitment to the Summer Olympics by participating in every edition of the event.

5.6.2 Gender wise Trend

In [79]:
```
athlete_df = df.drop_duplicates(['Name'])
```

In [80]:
```
male_df = athlete_df[athlete_df['Sex']=='M'].groupby('Year')['Name'].count().reset_index(name='No. of Males')
female_df = athlete_df[athlete_df['Sex']=='F'].groupby('Year')['Name'].count().reset_index(name='No. of Females')
```

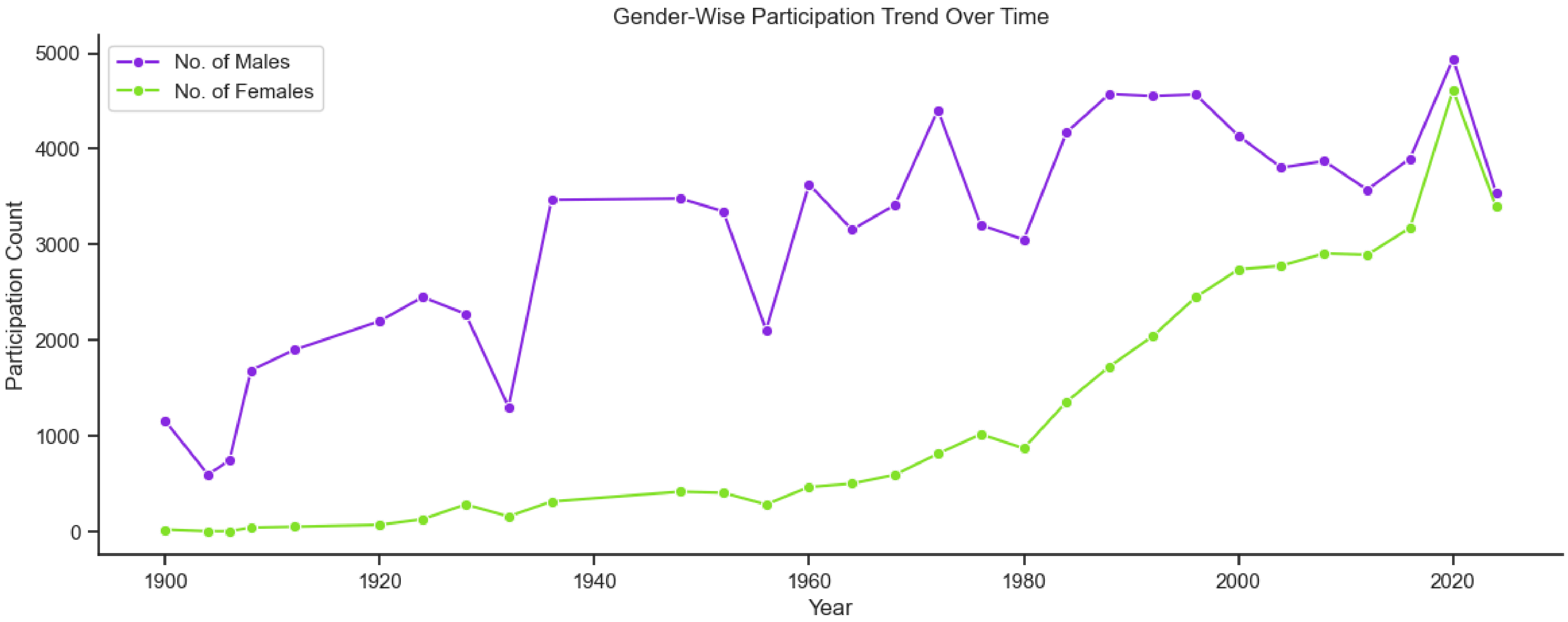
In [81]:
```
gender_df = pd.merge(male_df, female_df).sort_values(by='Year', ascending=True)
gender_df.head(3)
```

Out[81]:

	Year	No. of Males	No. of Females
0	1900	1160	23
1	1904	598	6
2	1906	746	6

In [82]:

```
plt.figure(figsize=(14, 5))
sns.lineplot(data=gender_df, x='Year', y='No. of Males', label='No. of Males', marker='o', color='#8A2BE2')
sns.lineplot(data=gender_df, x='Year', y='No. of Females', label='No. of Females', marker='o', color='#83E22B')
plt.xlabel('Year')
plt.ylabel('Participation Count')
plt.title('Gender-Wise Participation Trend Over Time')
plt.show()
```



- In the early years (1900-1920), there are significant imbalances in gender participation, with a notably higher number of male athletes compared to females. World War II (1939-1945) appears to have influenced a dip in overall participation, with a subsequent rebound in the post-war years. - The years 2012 and 2016 stand out as notable for achieving high levels of female participation, suggesting a continued focus on gender inclusivity.

5.7 India at Olympics

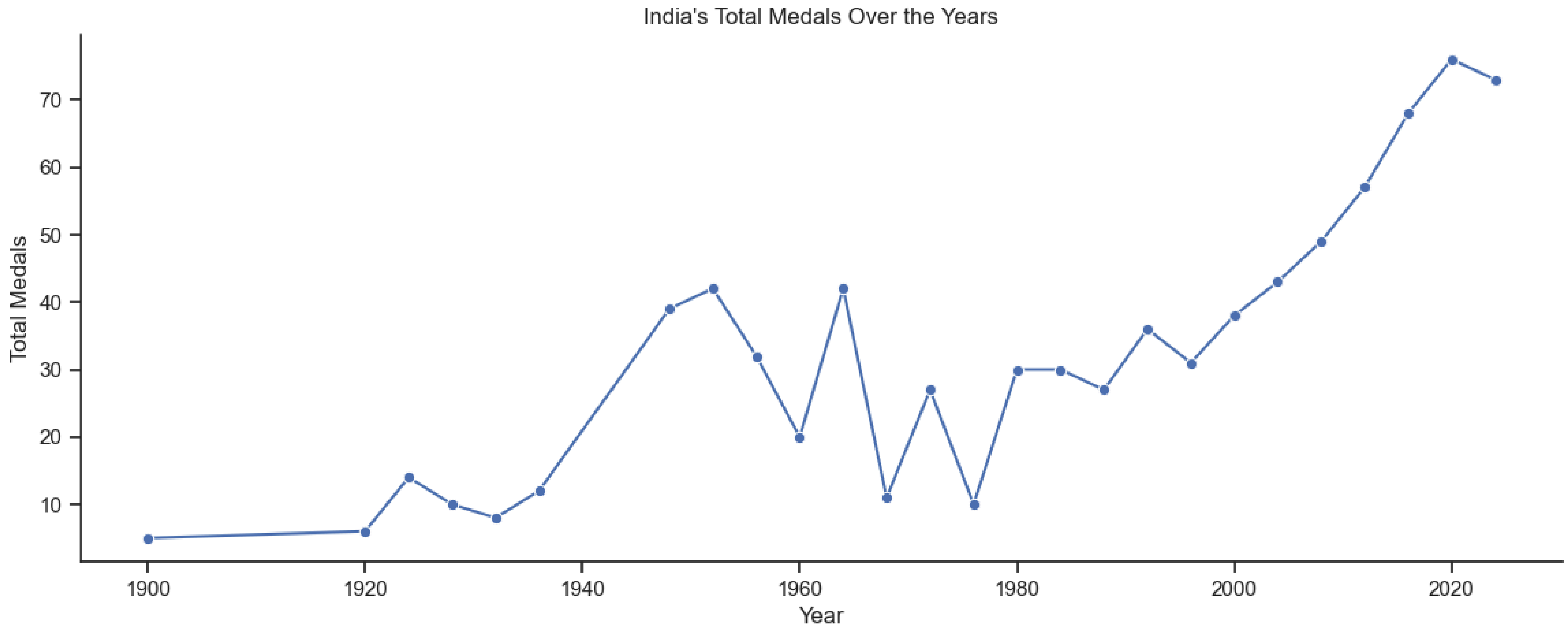

TYKA

In [84]:
```
# Records of all the medals won by India
India = medals[medals['Country']=='India']
```

5.7.1 Medals Analysis

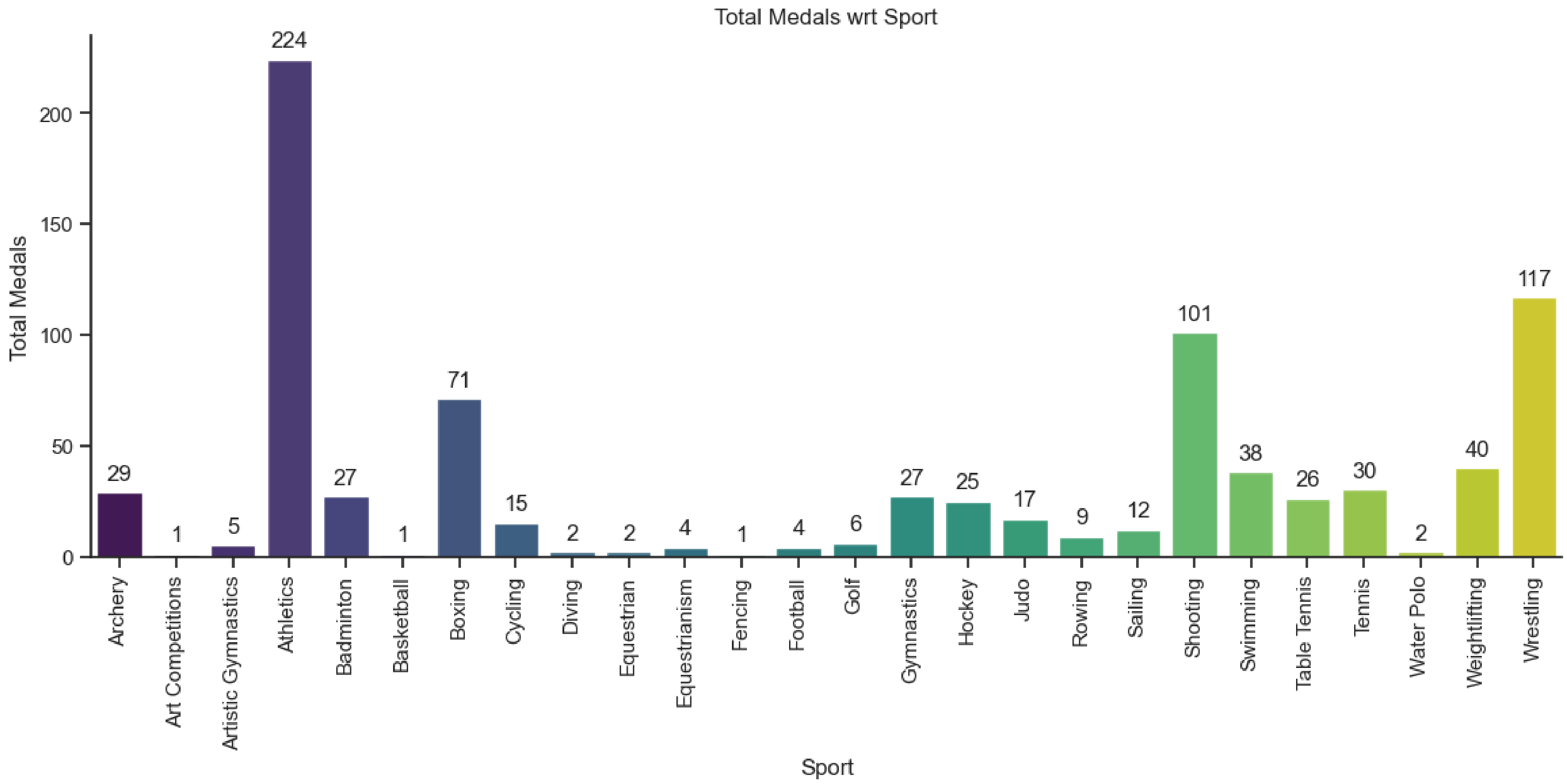
In [85]:
```
Ind_medals = India.groupby('Year')['Medal'].count().reset_index()
```

In [86]:
```
plt.figure(figsize=(14, 5))
sns.lineplot(data=Ind_medals, x='Year', y='Medal', marker='o', palette='viridis')
plt.xlabel('Year')
plt.ylabel('Total Medals')
plt.title("India's Total Medals Over the Years")
plt.show()
```



In [87]:
```
medal_ind = medals[(medals['Medal'].notna()) & (medals['Country']=='India')]
medal_ind = medal_ind.groupby(['Sport'])['Medal'].count().reset_index()
```

In [90]:
```
plt.figure(figsize=(14, 5))
sns.barplot(data=medal_ind, x='Sport', y='Medal', palette='viridis')
for p in plt.gca().patches:
   plt.gca().annotate(f"{int(p.get_height())}",
                      (p.get_x() + p.get_width() / 2., p.get_height()),
                       ha='center', va='center', xytext=(0, 10), textcoords='offset points')
plt.xticks(rotation=90)
plt.xlabel('Sport')
plt.ylabel('Total Medals')
plt.title('Total Medals wrt Sport')
plt.show()
```



- India's golden era in Olympic medals began in 1928, with an impressive haul of 25 medals in field hockey, setting the tone for dominance in subsequent years.
- India's total medal count has seen a significant upward trend, especially from the 2000s onward.
- After some fluctuations in earlier decades, recent years show a steady increase, indicating improved performance, better training facilities, and rising global competitiveness.
- India has won the most medals in Athletics, Wrestling, and Shooting, showcasing dominance in these sports. Other key contributors include Boxing, Weightlifting, and Hockey, reflecting a mix of strength, endurance, and traditional excellence.

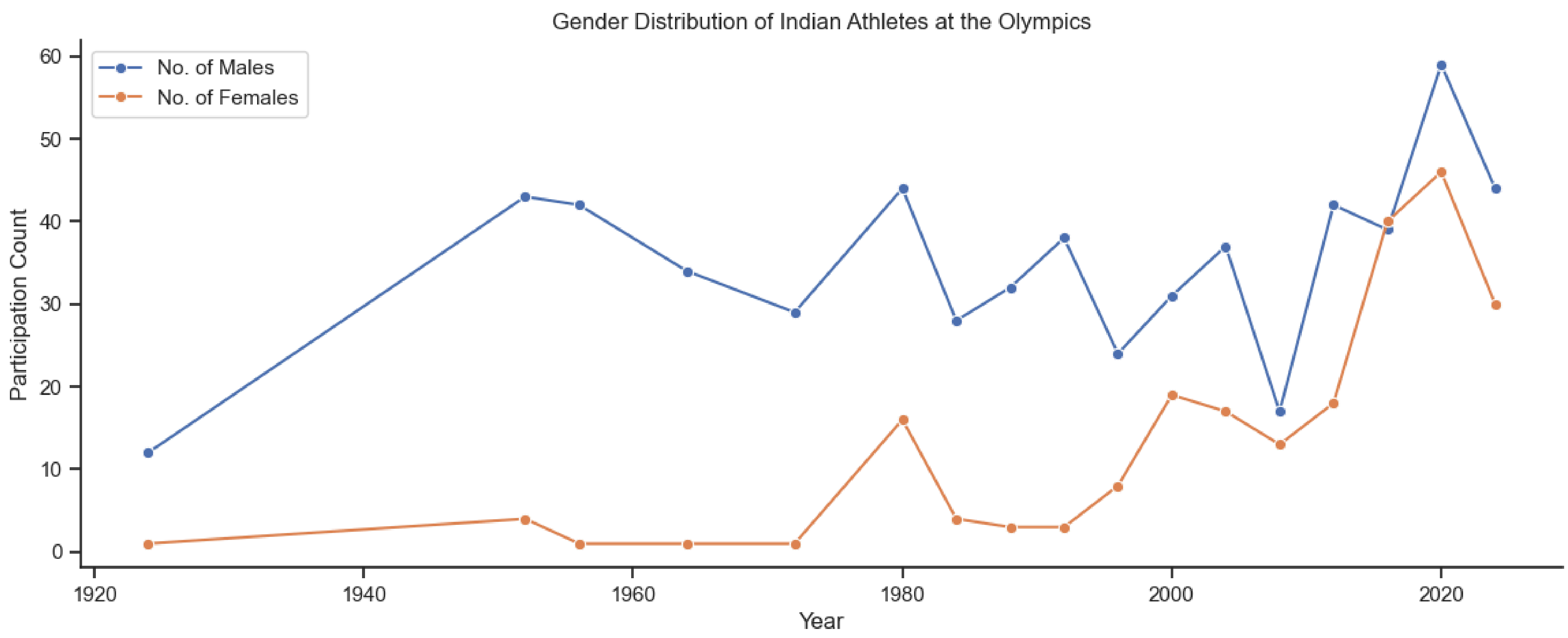
- India’s Olympic journey reflects a progressive rise in performance, with certain sports dominating medal wins. The increasing trend suggests strong future potential with continued investment in sports infrastructure and athlete development.

5.7.2 Gender wise Participation

In [92]:
```
male_ind = athlete_df[(athlete_df['Sex']=='M') & (athlete_df['Country']=='India')].groupby('Year')['Name'].count().reset_index
female_ind = athlete_df[(athlete_df['Sex']=='F') & (athlete_df['Country']=='India')].groupby('Year')['Name'].count().reset_ind
```

In [93]:
```
gender_ind = pd.merge(male_ind, female_ind)
```

In [94]:
```
plt.figure(figsize=(14, 5))
sns.lineplot(data=gender_ind, x='Year', y='No. of Males', label='No. of Males', marker='o', palette='viridis')
sns.lineplot(data=gender_ind, x='Year', y='No. of Females', label='No. of Females', marker='o', palette='viridis')
plt.title('Gender Distribution of Indian Athletes at the Olympics')
plt.xlabel('Year')
plt.ylabel('Participation Count')
plt.legend()
plt.show()
```



- **Male Dominance Historically** – The participation of Indian male athletes has always been higher, especially in the early years of the Olympics.
- **Rising Female Participation** – Female athlete participation remained low for decades but has significantly increased since the 2000s, narrowing the gender gap.
- **Recent Growth in Representation** – Both male and female athlete participation peaked in recent Olympics, reflecting India's growing investment in gender-inclusive sports.

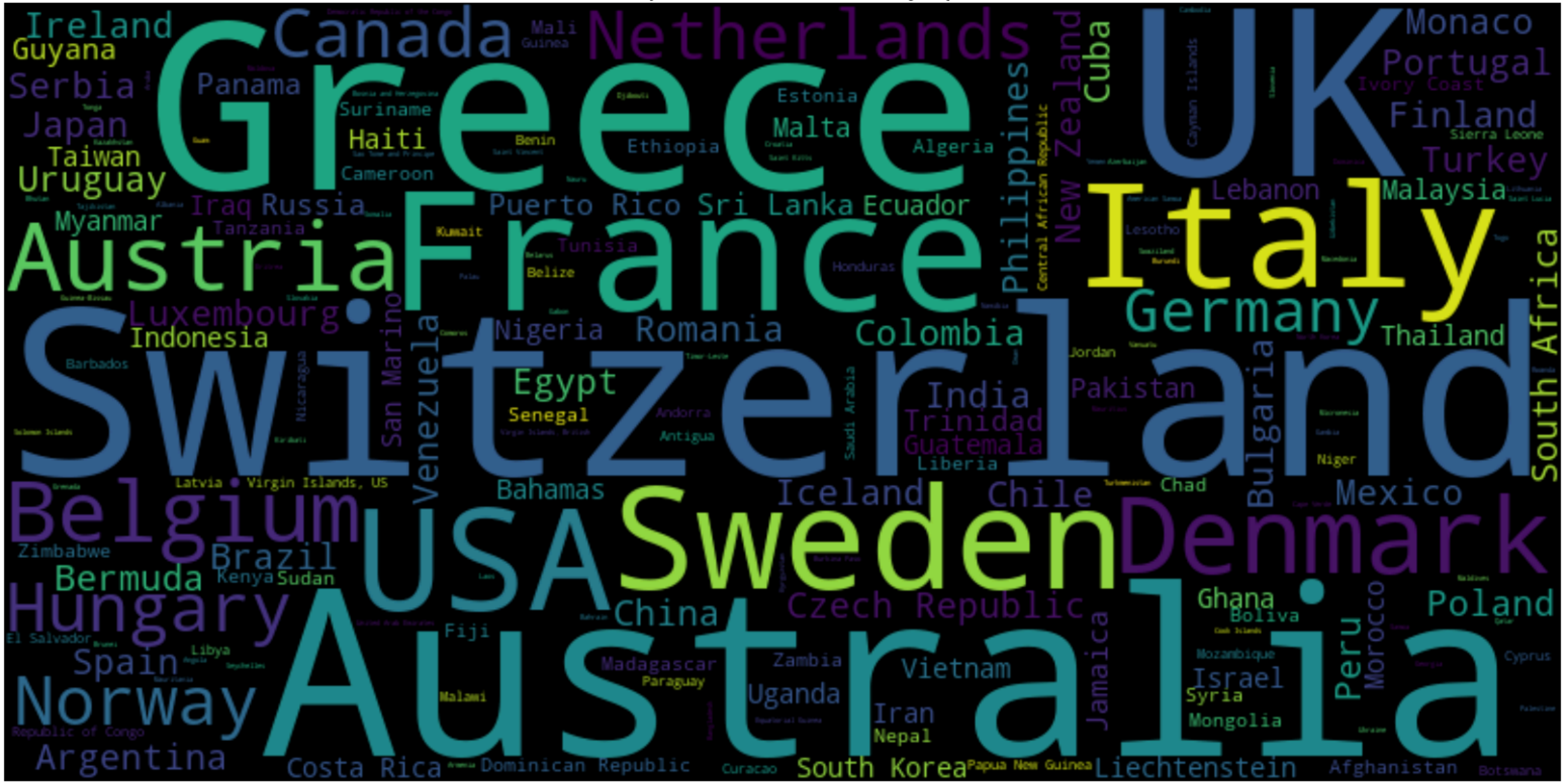
5.8 Word Cloud

In [97]:

```
wordcloud = WordCloud(width=800, height=400).generate_from_frequencies(country_trend.set_index('Country').to_dict()['Count'])
plt.figure(figsize=(20,10))
plt.imshow(wordcloud, interpolation='bilinear')
```

```python
plt.title('Participated countries of Olympics',fontsize = 20)
plt.axis("off")
plt.show()
```

Participated countries of Olympics



6. Conclusion

Historical Evolution:

- The Summer Olympics began with 12 participating nations in 1896 and has grown to engage over 200 nations in recent editions.
- The dip in participation during the 1980 Moscow Olympics boycott impacted 65 nations, highlighting geopolitical influences.

Medal Statistics:

- The refinement of our analysis identified and corrected discrepancies in medal counts, offering accurate insights.
- Top-performing nations, such as the United States, the Soviet Union, and China, have consistently dominated the medal tally with significant counts, e.g., the USA's 2780 total medals.

Participation Trends:

- The visual representation illustrated steady growth, with 43 participating nations in 1896 expanding to more than 200 in recent editions.
- Switzerland, Australia, Greece, the United Kingdom, France, and Italy participated in all 31 editions, showcasing enduring commitment.

Top Performers:

- Athletes like Larysa Semenivna Latynina, Aleksey Yuryevich Nemov, and Michael Fred Phelps set records with multiple wins, e.g., Phelps' 28 total medals.

Event Trends:

- Athletics and Swimming maintained consistent popularity, with Wrestling, Weightlifting, and Judo gaining traction over the years.
- The heatmap highlighted the distribution of sports events, offering insights into the dynamics of Olympic disciplines.

Popularity and Inclusivity:

- The growth in popularity was quantified by the increasing number of athletes, reaching over 5000 in recent editions.
- The rise in female participation, with notable spikes in 2012 and 2016, reflects ongoing efforts for gender inclusivity.

India's Olympic Journey:

- India's initial dominance in field hockey, evident with 25 medals in 1928, transitioned into diversified wins in shooting and badminton.
- India's total medal count showed irregular trends before 2000, with periods of rise and decline.
- A sharp and consistent increase in medals after 2000 indicates improved sports infrastructure, training, and government support.

To conclude:

The numerical analysis reaffirms the monumental growth of the Summer Olympics, shaping it into a global spectacle.
As nations anticipate future editions, the data supports the expectation of increased participation, diversity, and continued excellence, fostering the Olympic spirit.
The combination of historical context and numerical insights offers a comprehensive understanding of the Summer Olympics' enduring significance and its impact on the world of sports.
The numbers validate the trends, achievements, and the universal appeal of this iconic sporting event. 🏆 🌍

Thank you!