Zip Code Wilmington Data5.1
Tim Linkous

Olympic Success

Introduction

The Summer Olympics are back in full force in Paris, and we have been watching these athletes compete for their respective countries for the past few days. But what factors have led to success in the Olympics over time? The purpose of this analysis is to explore various aspects of Olympic data, focusing on athlete characteristics, medal distribution, and historical trends to discover just that.

Data

The dataset utilized was selected from a zip file obtained from the following GitHub repository (https://github.com/ZCW-J101D51/DataAnalysis) and contains 70,000 entries of Olympic athletes between the 1896 to 2016 Olympic games. Data preparation included cleaning the data to remove entries with missing values in critical categories used within the analysis. Given the range of time between the earliest and the most recent Olympics

included within the dataset, an effort was made in several instances to include data from more recent years (this will be defined in the analysis).

Some of the resources used throughout this analysis include Python and Pandas were utilized to read, sort, clean, and manipulate the data. Jupyter Notebooks was the interactive environment in which analysis and visualization was performed. Matplotlib, Seaborn, and Plotly libraries were used for plotting and visualization.

Methods

This section describes the analytical methods and tools used to explore the dataset. The objective was to uncover trends, correlations, and insights related to athlete performance, medal distribution, and historical patterns.

1. Data cleaning with pandas
    - pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language. It is helpful in providing data in a DataFrame, a data structure constructed with rows and columns, similar to a database or Excel spreadsheet.
    - It was used in this analysis to remove missing values in critical categories used, including *Age, Height, Weight,* and *Medal*. This helped prepare the data for interpretation and visualization.
2. Exploratory Analysis: visualization with Matplotlib, Seaborn, and Plotly.
    - These are popular Python Libraries. Seaborn is great for quickly creating visually appealing plots with minimal code, while Matplotlib offers more customization options and fine-grained control over every aspect of a plot. Plotly allows users to create beautiful interactive web-based visualizations that can be displayed in Jupyter notebooks,

These methods and libraries were helpful in sorting and providing visualizations to the dataset, and served as powerful tools for understanding data patterns, identifying trends, and communicating results.

Analysis

This section will contain the various analyses that we ran to try to identify trends. First, just how difficult is it to place and receive a medal as an Olympic athlete? Based on our dataset of 70,000 entries, the following is the medal results:

```
Total athlete entries: 70000
Total medal winners: 9690
Percentage of medal winners: 13.84%
```

So less than 14 % of all entries received a medal. We then pursued different avenues to try to compare medal winners.

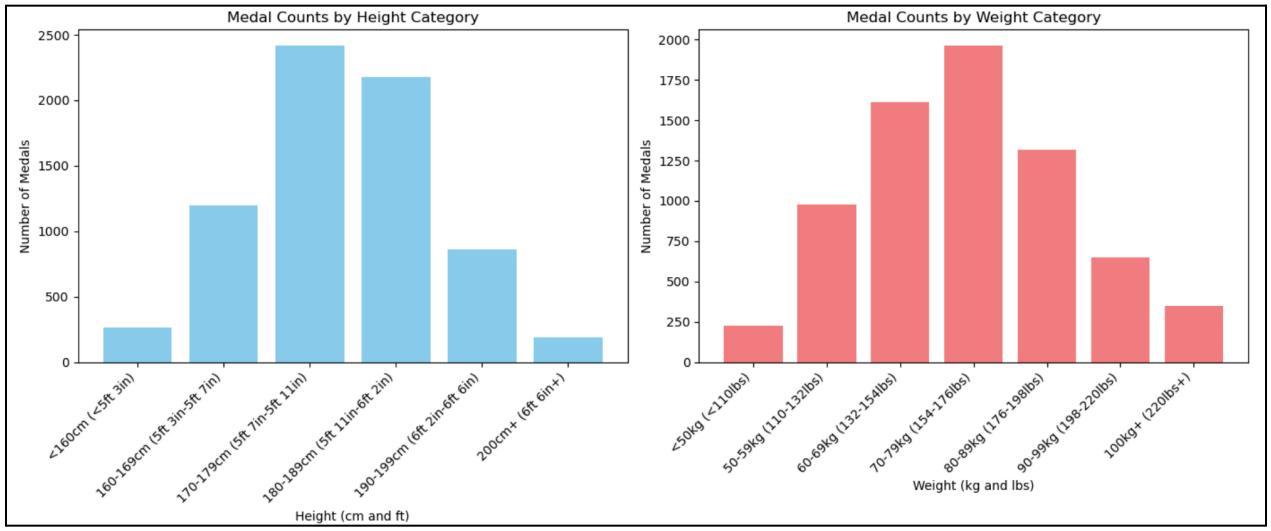
- Male vs: Female:

	Total Participants	Medalists	Medalist Ratio (%)
Male	26,108	3,128	11.98
Female	14,024	2,065	14.72

Note — we have filtered this comparison from 1980 on. This was an effort to get more recent data and avoid skewed data, as it is likely that significantly less women participated in the earlier years of the data. (https://www.topendsports.com/events/summer/women.htm)

The data shows that since 1980, although there are more male participants and overall medals, women have medaled at a slightly higher ratio than men.

- Height and weight data:

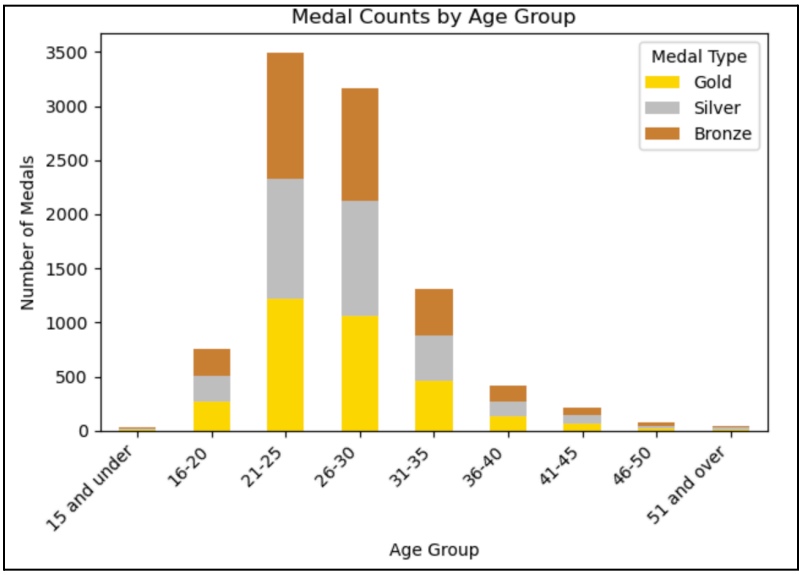


This data was worth exploring, but appears to follow a natural distribution of human height and weight. Given the large sample size, this is expected. There are certainly sub-sections of this data that would likely make sense (basketball medalists will be taller, equestrian medalists shorter and lighter, etc.) and it would be insightful to analyze height and weight distributions within individual sports to uncover sport-specific characteristics and requirements.

- Medals by age group:

The data was separated to find medalists in the following age groups: '15 and under', '16-20', '21-25', '26-30', '31-35', '36-40', '41-45', '46-50', '51 and over'.

```
#Medals by age group

df_cleaned = df.dropna(subset=['Age', 'Medal']).copy()

bins = [0, 15, 20, 25, 30, 35, 40, 45, 50, 100]
labels = ['15 and under', '16-20', '21-25', '26-30', '31-35', '36-40', '41-45', '46-50', '51 and over']
df_cleaned['Age Group'] = pd.cut(df_cleaned['Age'], bins=bins, labels=labels, right=False)
                                                    #bins are 'left-includive': includes 15--> exludes 20, etc.
```

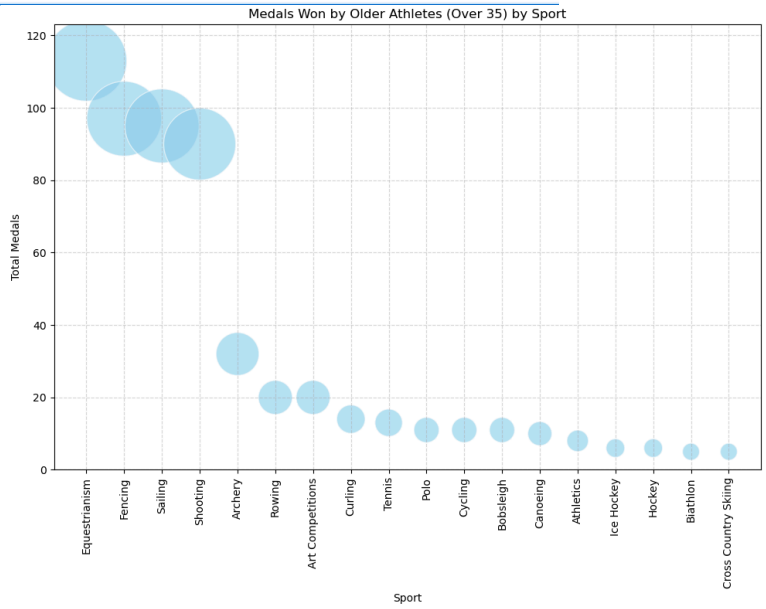


This analysis confirms that a vast majority of medal winners are between the ages of 21 and 30, as expected for the prime of an athlete's career. While some aspects were expected, others were interesting and resulted in further tests. While relatively small amounts compared to the prime athletic ages, participants of older ages were able to win

medals, but in what sports/competitions? We sorted the data including medalists over 35 years of age and presented our findings below:

```
older_threshold = 35

older_medalists = df_cleaned[(df_cleaned['Age'] > older_threshold)] #filter for older athletes

#count medals by sport
medals_by_sport = older_medalists['Sport'].value_counts().reset_index()
medals_by_sport.columns = ['Sport', 'Total Medals']
print(medals_by_sport)
```

	Sport	Total Medals
0	Equestrianism	113
1	Fencing	97
2	Sailing	95
3	Shooting	90
4	Archery	32
5	Rowing	20
6	Art Competitions	20
7	Curling	14
8	Tennis	13
9	Polo	11
10	Cycling	11
11	Bobsleigh	11
12	Canoeing	10
13	Athletics	8
14	Ice Hockey	6
15	Hockey	6
16	Biathlon	5
17	Cross Country Skiing	5
18	Wrestling	4
19	Volleyball	4
20	Tug-Of-War	4
21	Basketball	4
22	Water Polo	4
23	Alpinism	3
24	Handball	3
25	Gymnastics	3



This data revealed that medals won by older competitors were heavily skewed towards four events in particular: *Equestrian, Fencing, Sailing,* and *Shooting.* Older athletes are proven to be more successful in events that require skill, precision, and experience rather than pure physical strength or endurance. These sports also allow for a prolonged career of the competitors.

*Fun Note: Interesting discovery in this section.. If you notice, the 7th ranking event in the chart above is Art Competitions, which were held as Olympic events from 1912 through the 1948 Olympic games, Medals were awarded in five categories: architecture, literature, music, painting and sculpture. Jean Jacoby, of the Netherlands, was the most successful Olympic Artists, winning his second gold medal for this drawing, titled "Rugby":

Sources:https://www.olympic-museum.de/art/artcompetition.php

https://olympics.com/en/news/look-to-the-past-when-olympic-medals-were-awarded-for-architecture-music-and-lit

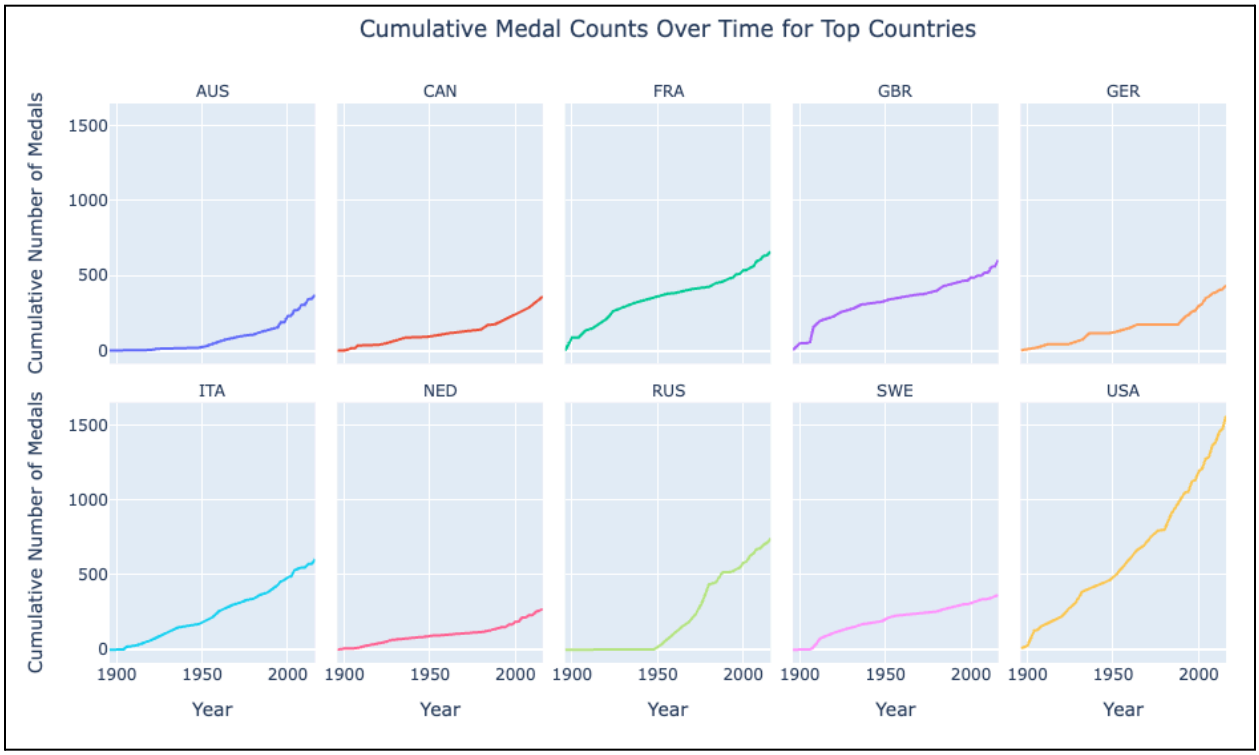
- Country comparison: cumulative medals over time:

In this method, we focused on the period from 1960 onward in an effort to acknowledge the increased global reach of the Olympic games and introduction of new countries, which leads to increased competition for the events. For example, the Stockholm Olympics in 1912, featured athletes from 28 nations (the most up until that point), compared to 87 in 2016 at Rio (https://www.history.com/topics/sports/modern-olympic-games-timeline). Also of note, we have combined the medal counts of the USSR (URS) and Russia (RUS) and the data was cleaned to remove rows with missing values in critical columns such as Year, NOC (National Olympic Committee), and Medal. Below is a visual showing our code to sort the data and display the top cumulative medal earning countries. Although the top 50 countries were sorted in this analysis, the top 20 are displayed for brevity.

```
# Data Cleaning: Drop rows with missing Year, NOC, or Medal
df_cleaned = df.dropna(subset=['Year', 'NOC', 'Medal']).copy()
df_cleaned = df_cleaned[df_cleaned['Year'] >= 1960]   #Filter for Post-1960 data

#combine USSR (URS) and Russia (RUS)
df_cleaned['NOC'] = df_cleaned['NOC'].replace('URS', 'RUS')

#total number of medals by country
medals_by_country = df_cleaned['NOC'].value_counts().reset_index()
medals_by_country.columns = ['Country', 'Total Medals']

# Sort the countries by total medals and select the top 50
top_countries = medals_by_country.sort_values(by='Total Medals', ascending=False).head(50)
print("Countries with the most medals since 1960: ")
print(top_countries)
print("Note: USR and RUS medals have been combined in this dataset.")
```

```
Countries with the most medals since 1960:
   Country  Total Medals
0      USA          1007
1      RUS           668
2      ITA           387
3      AUS           327
4      GER           297
5      FRA           280
6      GBR           255
7      CAN           255
8      GDR           203
9      ROU           192
10     NED           176
11     HUN           169
12     ESP           165
13     BRA           161
14     NOR           149
15     SWE           138
16     CHN           122
17     FRG           119
18     CUB           107
19     SUI            85
20     POL            75
```

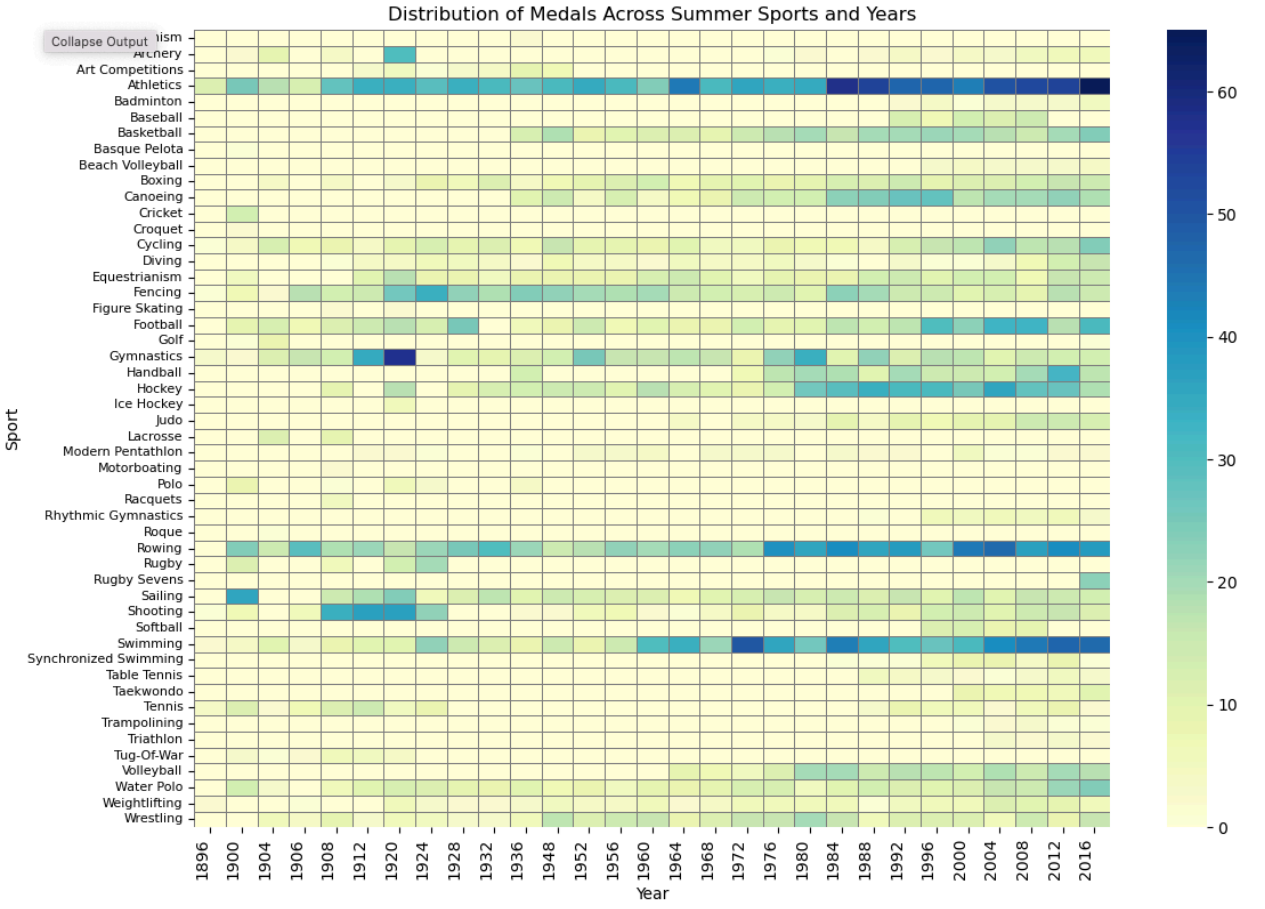
The medal counts have been dominated by the United States since 1960, with Russia emerging as a distant second. The group shows countries from relatively diverse regions and highlights the increasing global nature and competition of the Olympics. To show the total cumulation of medals awarded to the top performing countries throughout the entirety of the dataset, we have provided a line graph visual display using the Plotly library.

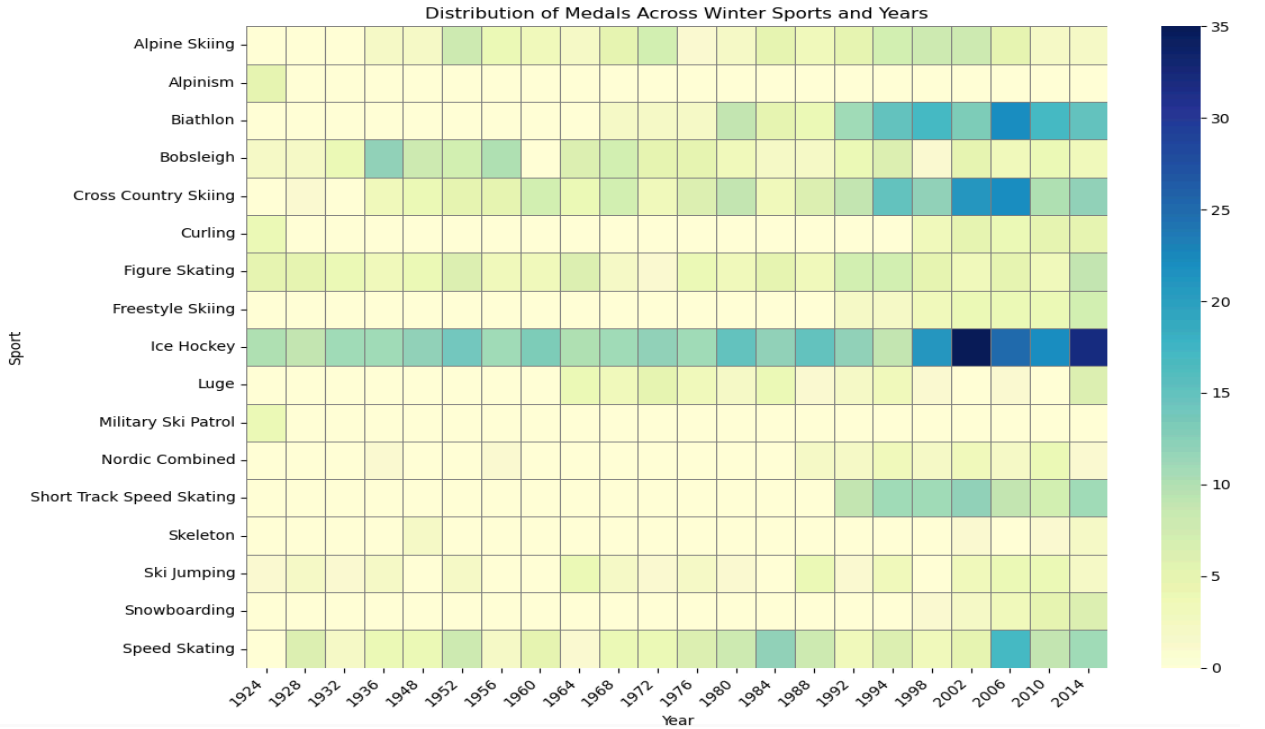


Of note is the USA's steep and consistent dominance throughout the dataset, Russia's rapid cumulative medal growth starting from the mid-20th century, and Germany with a noticeable acceleration post-1990 following reunification and fall of the Berlin Wall.

- Success by Sport:

FInally, we have focused a portion of the analysis on the distribution of medals across individual sports over time for both the Summer and Winter Olympics. This allows us to visualize which sports have gained or lost prominence over time. By analyzing these trends, the analysis can reveal shifts in popularity, the impact of rule changes, or the introduction of new sports and new athletes to the Olympics. The below visual is a heatmap (using Matplotlib), which provides an intuitive way to identify trends, with the intensity of color representing the number of medals awarded in a particular sport for a given year.


Distribution of Medals Across Summer Sports and Years
Collapse Output
Archery
Art Competitions
Athletics
Badminton
Baseball
Basketball
Basque Pelota
Beach Volleyball
Boxing
Canoeing
Cricket
Croquet
Cycling
Diving
Equestrianism
Fencing
Figure Skating
Football
Golf
Gymnastics
Handball
Hockey
Ice Hockey
Judo
Lacrosse
Modern Pentathlon
Motorboating
Polo
Racquets
Rhythmic Gymnastics
Roque
Rowing
Rugby
Rugby Sevens
Sailing
Shooting
Softball
Swimming
Synchronized Swimming
Table Tennis
Taekwondo
Tennis
Trampolining
Triathlon
Tug-Of-War
Volleyball
Water Polo
Weightlifting
Wrestling
Sport
1896 1900 1904 1906 1908 1912 1920 1924 1928 1932 1936 1948 1952 1956 1960 1964 1968 1972 1976 1980 1984 1988 1992 1996 2000 2004 2008 2012 2016
Year
0 10 20 30 40 50 60


Distribution of Medals Across Winter Sports and Years
Alpine Skiing
Alpinism
Biathlon
Bobsleigh
Cross Country Skiing
Curling
Figure Skating
Freestyle Skiing
Ice Hockey
Luge
Military Ski Patrol
Nordic Combined
Short Track Speed Skating
Skeleton
Ski Jumping
Snowboarding
Speed Skating
Sport
1924 1928 1932 1936 1948 1952 1956 1960 1964 1968 1972 1976 1980 1984 1988 1992 1994 1998 2002 2006 2010 2014
Year
0 5 10 15 20 25 30 35

Based on the summer sport heatmap, Swimming, Rowing, and Athletics (Track & Field) show consistent and high medal counts across many years, while there is an emergence in popularity in Rugby Sevens and Football (Soccer). For the winter games, Ice Hockey has predominantly featured the most medals in recent years.

Results

This section will cover the key findings within the analysis, which will help to provide insights on performance and trends.

Key Findings:

1. **Cumulative Medal Counts**:
    - The United States leads with over 1,000 medals since 1960, demonstrating its dominance in the Olympics.
    - Russia and Germany also rank among the top-performing countries.
2. **Age and Sports Success**:
    - Older athletes have found success in less physically demanding sports such as Equestrian, Shooting, and Archery. These sports emphasize skill and experience over physical strength, which allows athletes to compete beyond their athletic primes
3. **Gender Participation**:
    - Female participation in the Olympics has increased significantly since 1980, with female athletes achieving higher medalist ratios compared to their male counterparts (14.7% compared to 12.0%).
4. **Sports Evolution over time:**
    - Traditional sports like Gymnastics, Swimming, and Athletics (Track) continue to secure a large number of medals, while there are a number of emerging sports capturing the attention of the athletes and audience.

This analysis highlights the changing nature of the Olympics over time by showcasing historical trends provided by the data. The most successful athletes come from a few select countries: USA, RUS, ITA, AUS, GER, FRA, GBR, CAN; and are predominantly in the prime ages for peak athletic performance (ages 21-30). However, older athletes have found

success in less physically demanding sports that require more experience and precision, such as Equestrian, Fencing, Sailing, Shooting, and Archery. Many of the medalists come from a select range of popular sports, including Gymnastics, Swimming, Athletics (Track), and Ice Hockey; however, there are new emerging sports that will look to gain popularity in the coming events.

Conclusion

What factors have led to success in the Olympics over time? Based on this dataset, there are several factors. Age and experience, physical attributes, country for which you are participating, and the sports/events you are participating in. Which sports/events will emerge in the coming years? Will there be success by more diverse countries? We can't wait to see the data. Enjoy the Olympics!!

Further Thoughts/Exploration:

- Why are some countries more successful than others? More success in wealthier nations due to more access to training?
- Is there geopolitical relevance to success? Individual case studies could be performed on certain countries: Russia's success during the Cold War and competition with the USA, Germany's rapid success post- 1990.
- Which countries offer significant incentives for earning medals? (or hopefully not, but possible punishments for poor performance?) If so, how does that impact success?
- Which sports, if any, will take over the popularity in the Olympic games? Have they already, given that the dataset ends at the 2016 games?

*Note: this analysis was performed as an assignment by a student of Zip Code Wilmington in the Data 5.1 cohort in August of 2024.. Thank you for reading.