

120 Years of Olympic History: Athletes and Results

Shijie Cheng



Intro



Motivation

Spanning over a century, the Olympic Games have witnessed the triumphs and challenges of generations, making Olympic history a rich data that mirrors global cultural and social developments.

Data Source

<https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results/data>

athlete_events.csv: Contains records of individual athletes' performances across different Olympic Games, including details like the athlete's name, sex, age, team, NOC code, the game year, sport, event, and medals won.

noc_regions.csv: Provides a mapping between the NOC codes and the corresponding countries or regions, facilitating an understanding of the athletes' national affiliations.

Objective

Utilizing two comprehensive datasets from Kaggle, I plan to examine the development history of the Olympics, focusing on:

1. Medal distribution
2. Athlete demographics
3. National performances (China vs. US)

Extraction, Transform, and Load (ETL)



athlete_events.csv

```
In [3]: athlete_data.head()
```

```
Out[3]:
```

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

noc_regions.csv

```
In [4]: noc_region.head()
```

```
Out[4]:
```

	NOC	region	notes
0	AFG	Afghanistan	NaN
1	AHO	Curacao	Netherlands Antilles
2	ALB	Albania	NaN
3	ALG	Algeria	NaN
4	AND	Andorra	NaN

1. Merge these two datasets

2. Removed some redundant columns

1. "Team", "NOC", and "region" all mean the country of players

2. The "notes" column is irrelevant to the topic of my project.

Final DataFrame

	ID	Name	Sex	Age	Height	Weight	Games	Year	Season	City	Sport	Event	Medal	region
0	1	A Dijiang	M	24.0	180.0	80.0	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	China
1	2	A Lamusi	M	23.0	170.0	60.0	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN	China
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN	Denmark
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold	Denmark
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN	Netherlands

Exploratory Data Analysis (EDA)

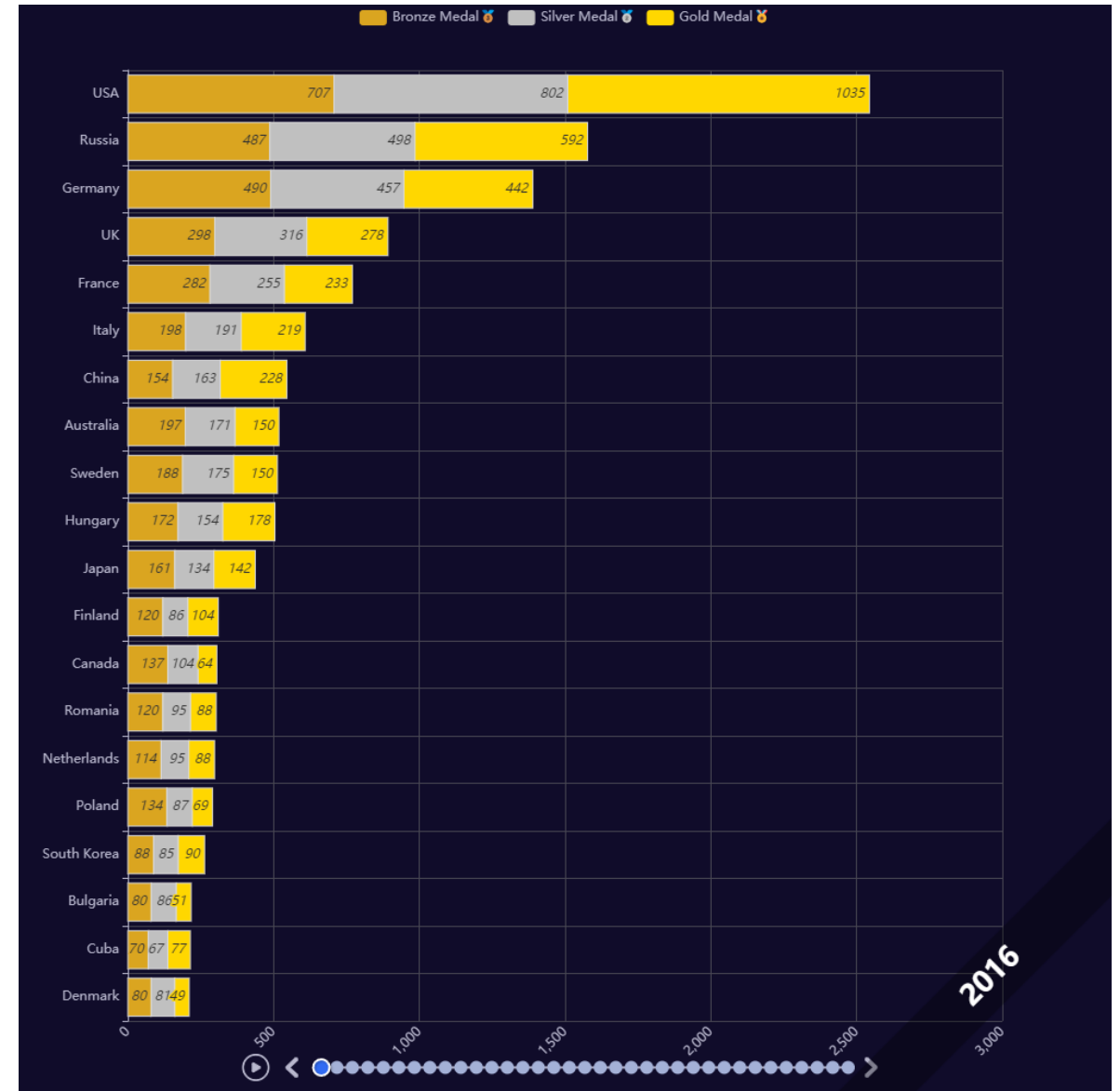


Medal distribution

- As of the 2016 Summer Olympics, **the United States** took **first place** by a wide margin, winning **2,544 medals and 1035 gold medals**, which are both nearly **twice as many as the second-place Russia**.
- China**, a well-known sports powerhouse, is only in **the seventh place**, behind the United Kingdom, France and Italy.
(The reason for this will be found in the later section China vs. US)

Conclusion:

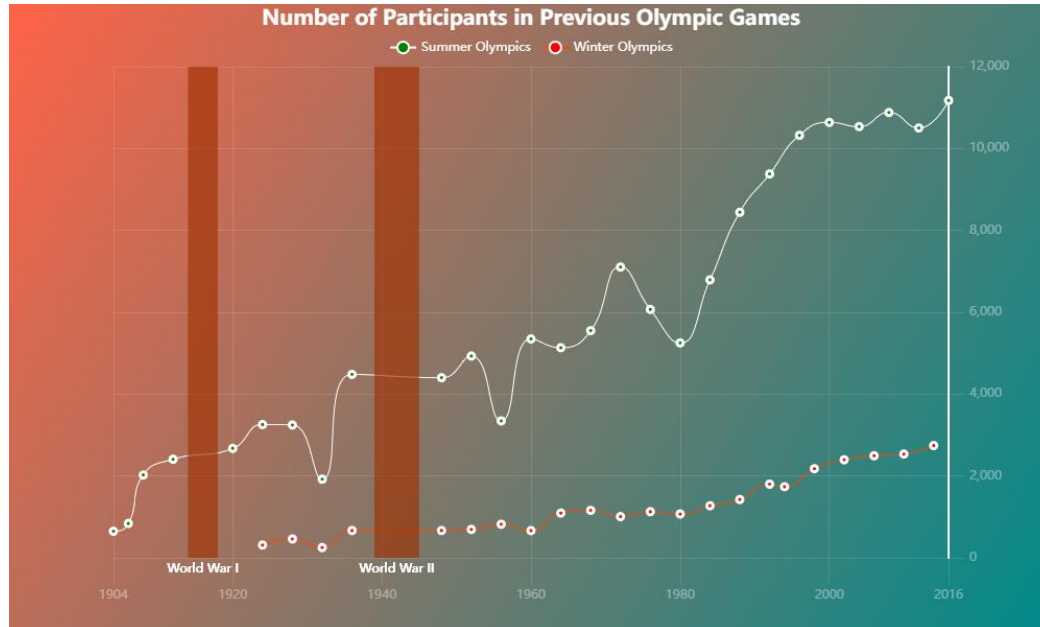
In 120 years of Olympic history, **the United States** has **dominated the gold and medal standings by a breakneck margin**.



Athlete Demographics



Numbers of Athletes



- The number of participants in each **Summer Olympics is 4-8 times higher than in the Winter Olympics.**
- The overall number of participants is **trending upwards.**

Trends of Women Participating

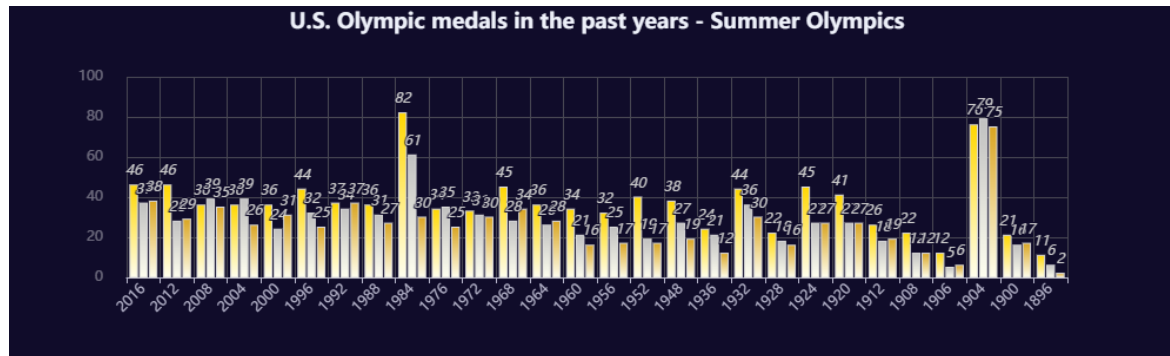
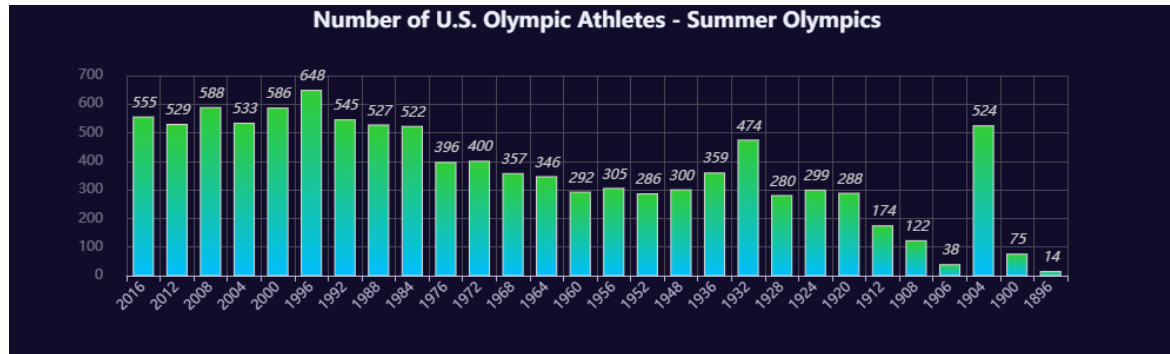


- **In the beginning**, the Olympics was basically a “**man’s sport**” with only **single digit female athletes.**
- However, **in recent Olympics**, the number of male and female participants **has basically equalized.**

National performances (China vs. US)

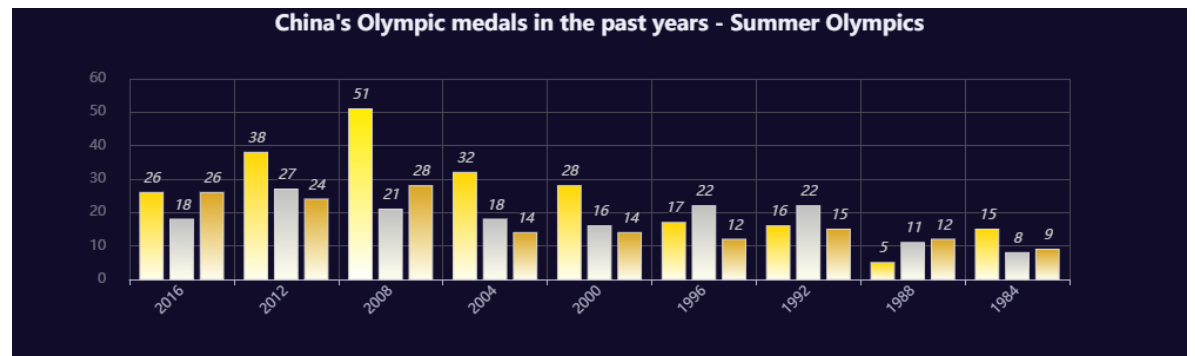
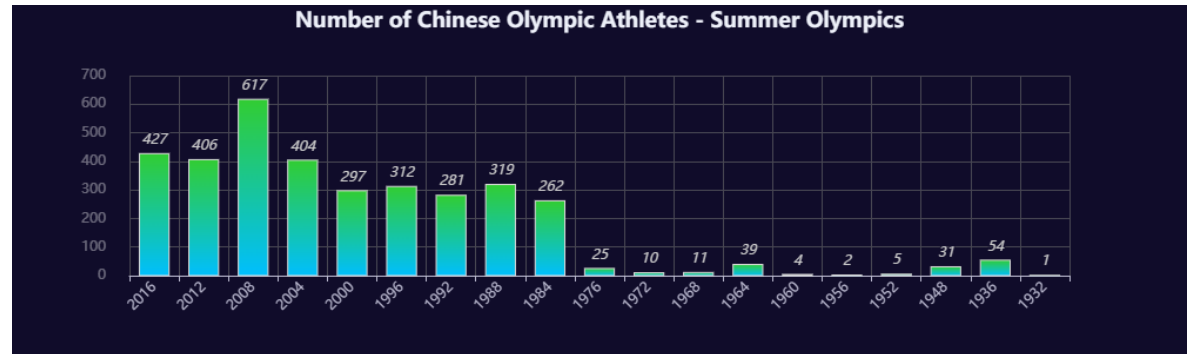


US



- **United States** has maintained a relatively high and stable number of participants and medals **throughout its 120-year Olympic history**.

China



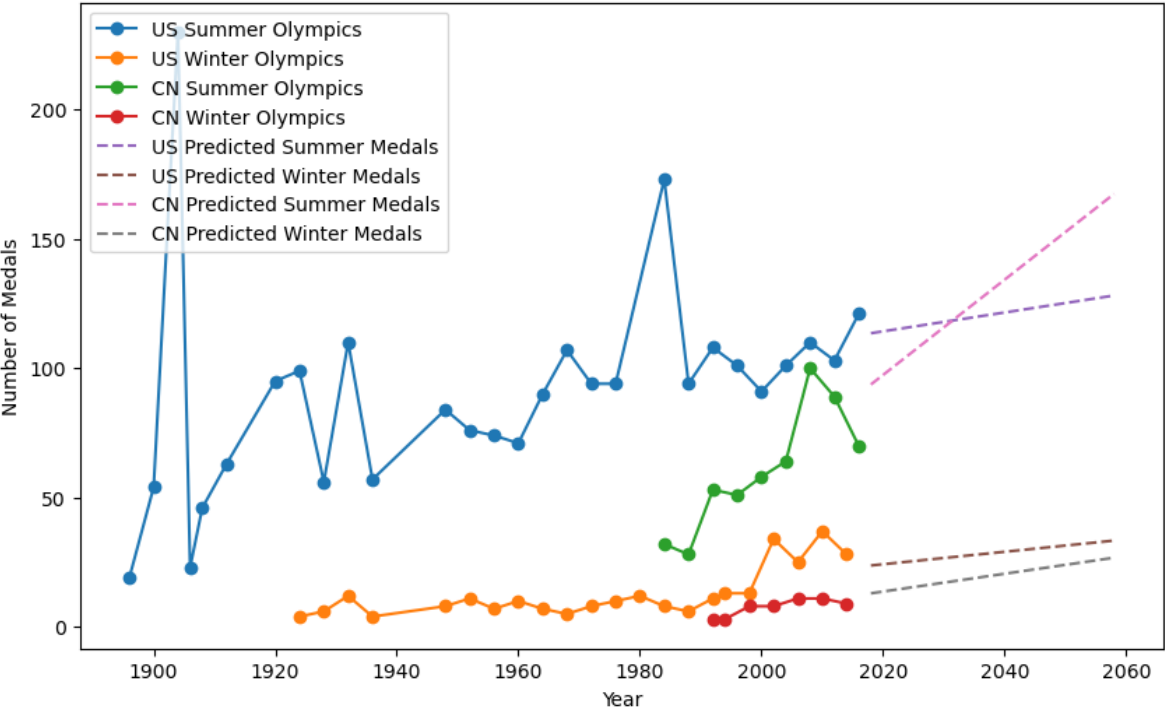
- **China** only started to win medals and have a relatively high and stable number of participants **until 1984**.

Model Prediction (China vs. US)



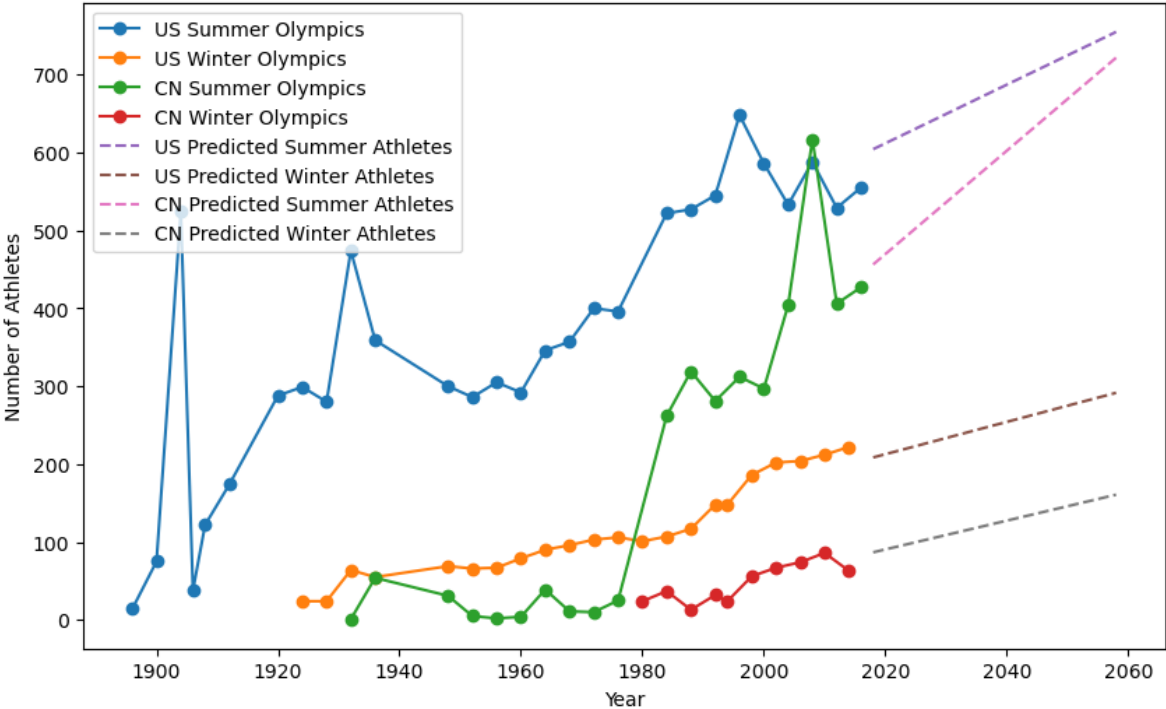
Medal Number Prediction

Predicted Medal Number for US and CN in Olympics until 2060



Athletes Number Prediction

Predicted Number of Athletes for US and CN in Olympics until 2060



Although United States is **ahead of** China so far **in terms of both number of medals and participants**, in the future China seems **to have a tendency to overtake** the US in both parts.