BREAKING THE OLYMPIC MOLD: A CENTURY OF ATHLETIC EVOLUTION THROUGH DATA

Ruhama Baruda

Department of Mathematics, Howard University

MATH014: Introduction to Data Science

April 24, 2025

OBJECTIVE

The purpose of this project is to explore how athlete participation, performance, and representation have evolved throughout the history of the modern Olympic Games. Using a dataset that spans from 1896 to 2016, this study applies data wrangling, visualization, and statistical interpretation to answer questions about medal distribution, gender disparities, seasonal representation, and sport-level trends.

DATASET OVERVIEW

This dataset, titled "120 Years of Olympic History: Athletes and Results," was compiled from sports-reference.com and sourced through Kaggle. The dataset includes athlete-level records from the modern Olympic Games spanning 1896 to 2016, encompassing both Summer and Winter Games. Each row in the dataset corresponds to a single athlete's participation in a specific Olympic event.

Link to dataset

- 271,116 rows
- 15 columns

Key Variables:

- athlete demographics (Age, Sex, Height, Weight)
- national affiliation (Team, NOC)
- event details (Sport, Event, Year, Season, City)
- Medal column

DATA CLEANING

- Missing values in Age, Height, and Weight were filled using the median value within each Sport and Sex grouping
- For athletes competing in rare or discontinued sports where no group median could be calculated, the overall column median was used as a fallback.
- The Medal column was also addressed by filling missing values with the string "No Medal" to distinguish non-medalists from incomplete data.

```
Count of missing values by column:

ID             0
Name           0
Sex            0
Age         9474
Height     60171
Weight     62875
Team           0
NOC            0
Games          0
Year           0
Season         0
City           0
Sport          0
Event          0
Medal     231333
```

DATA CLEANING

- 1,385 exact duplicates were identified and removed
- No anomalies in Height, Weight, or Age
- The Medal column was also addressed by filling missing values with the string "No Medal" to distinguish non-medalists from incomplete data.
- Team column was renamed to Country, and categorical text columns including Sport, City, and Event were standardized (title casing and whitespace)

New Columns:

- Is Medalist
- Decade
- BMI
- Medal Value

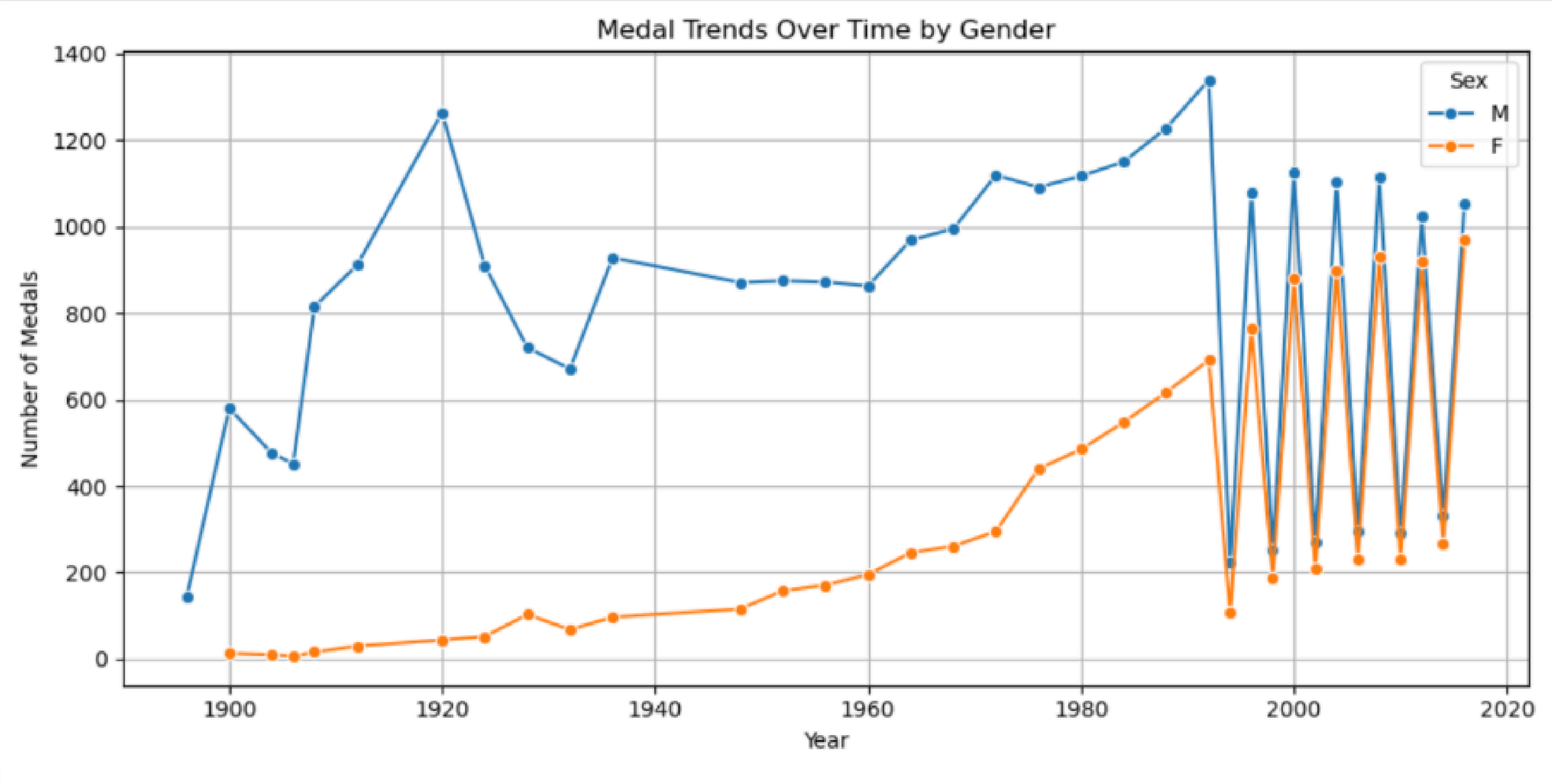
EXPLORATORY DATA

- Total unique athletes: 135571
- Top 5 Countries by Athlete Participation
- Top 10 Sports by Number of Events
- Athlete Reresentation by season:
  - Summer = 116776
  - Winter = 18958

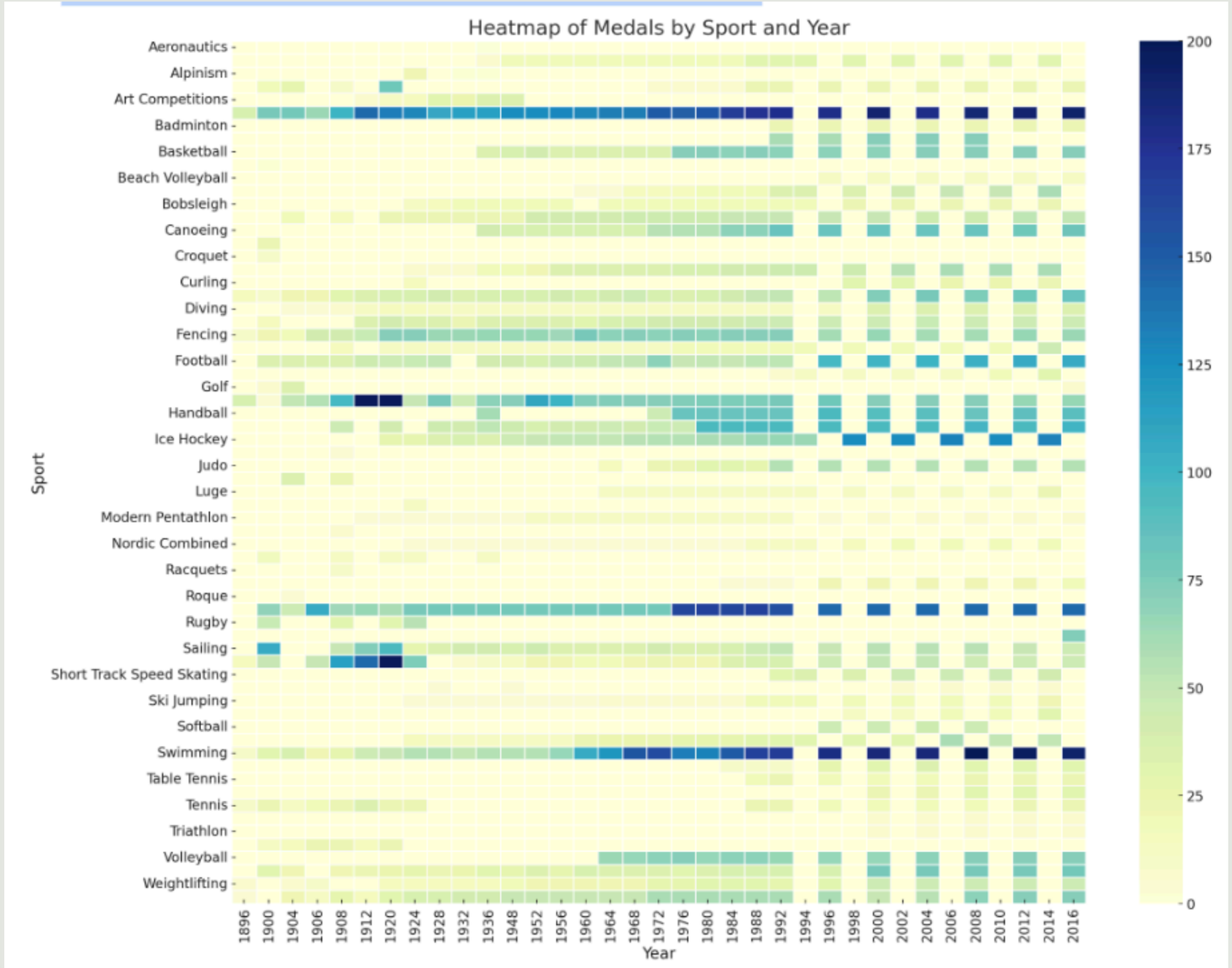
United States	17598
France	11817
Great Britain	11264
Italy	10213
Germany	9230

Shooting	83
Athletics	83
Swimming	55
Cycling	44
Sailing	38
Wrestling	30
Archery	29
Art Competitions	29
Canoeing	27
Gymnastics	27

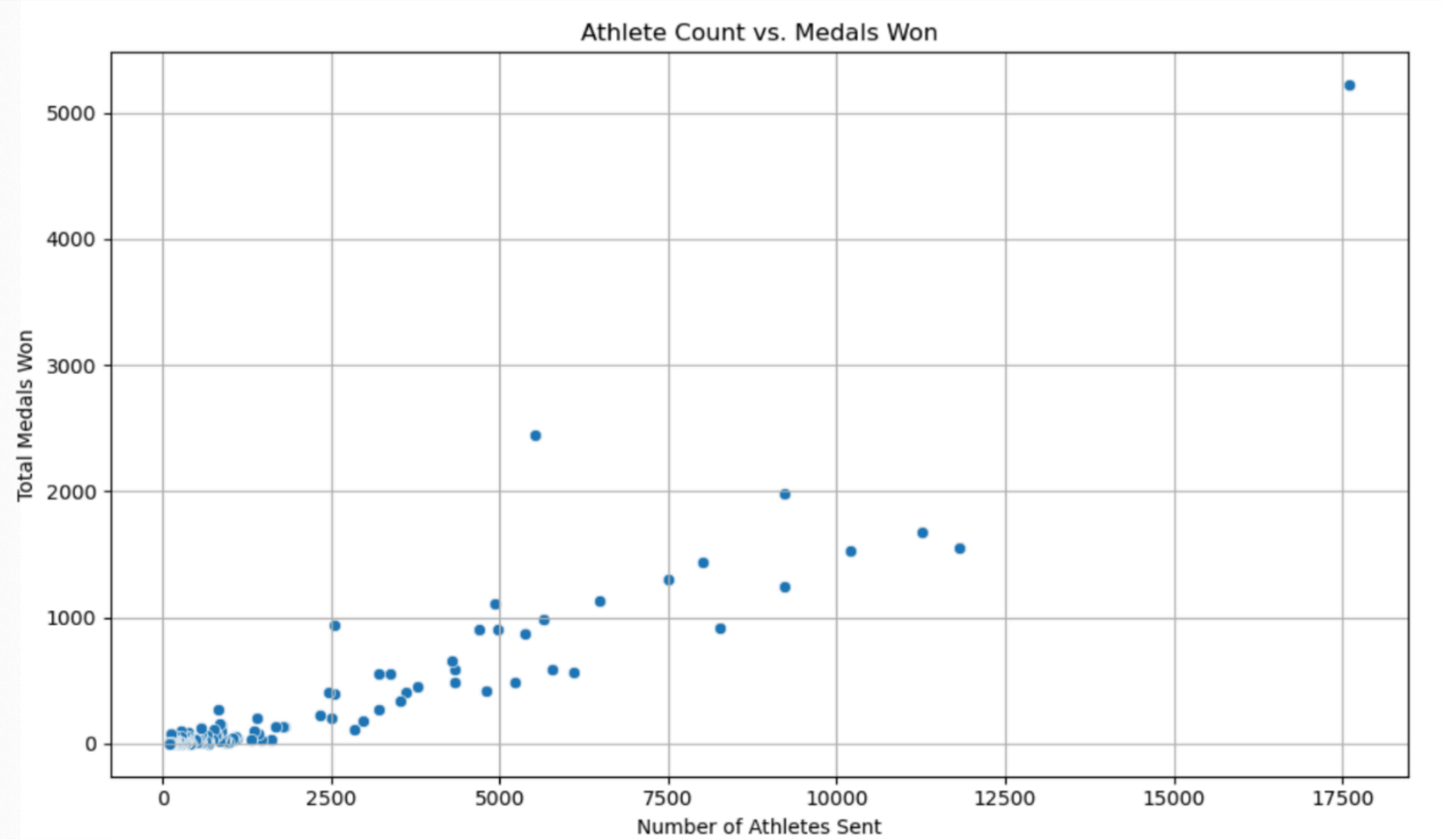
MEDAL TRENDS OVER TIME BY GENDER



HEATMAP OF MEDALS BY SPORT AND YEAR

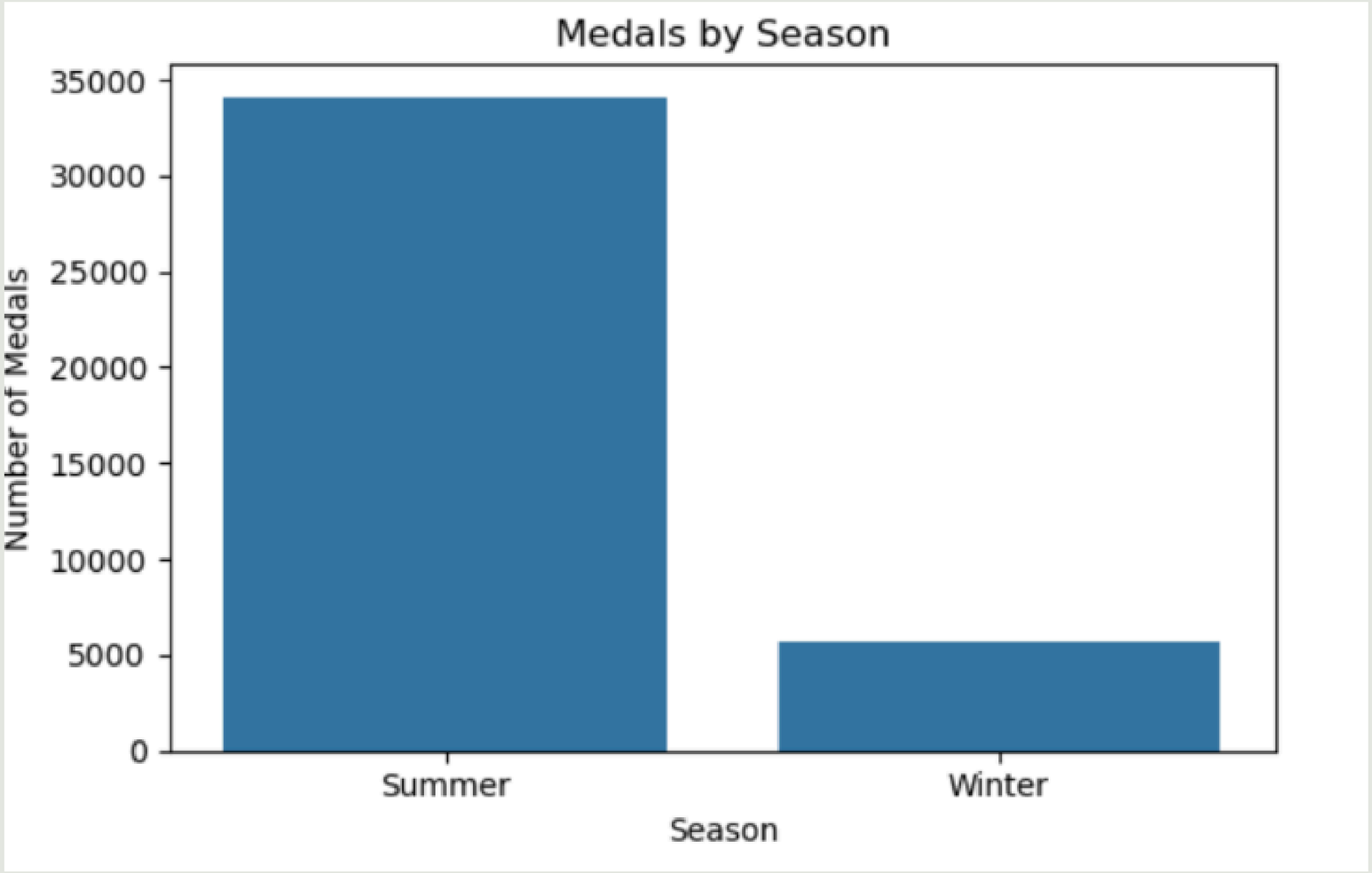


ATHLETE COUNT VS. MEDALS WON

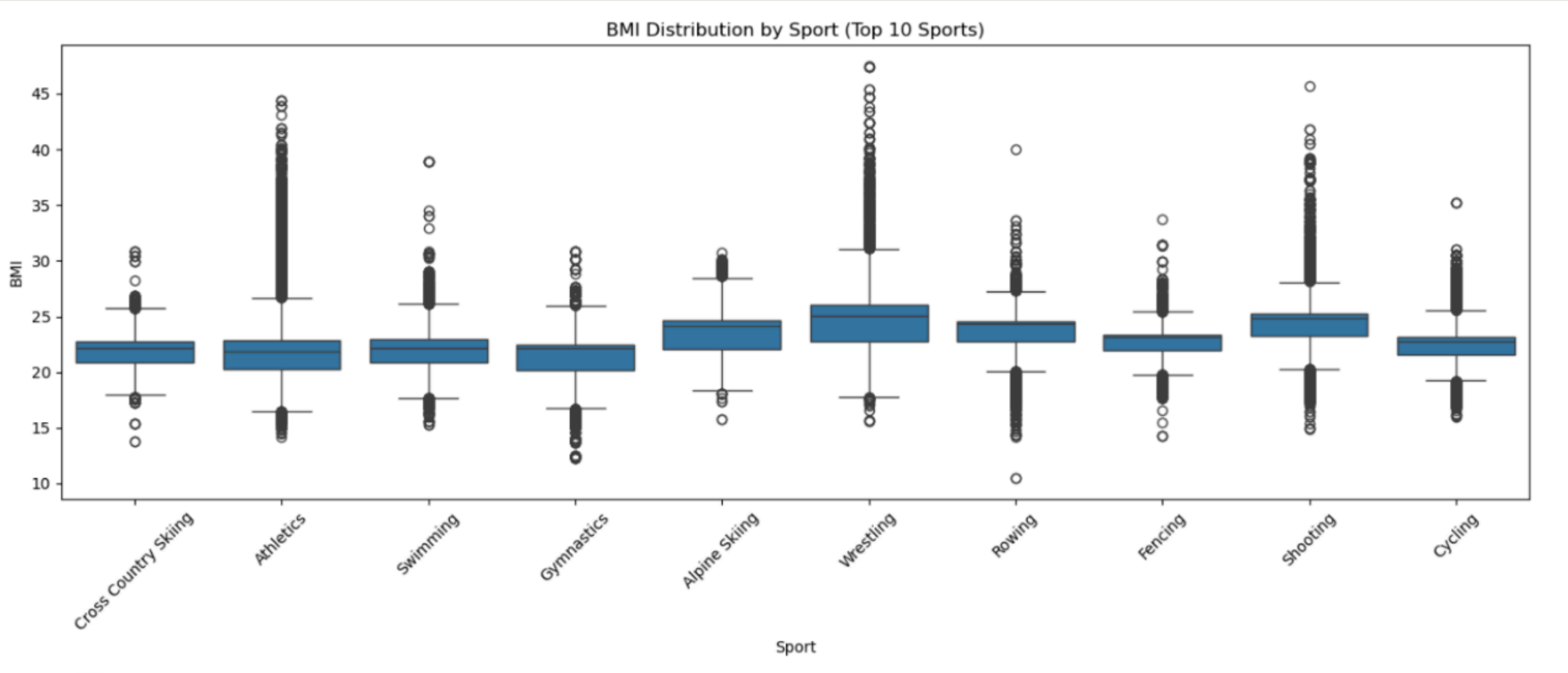


r = 0.90

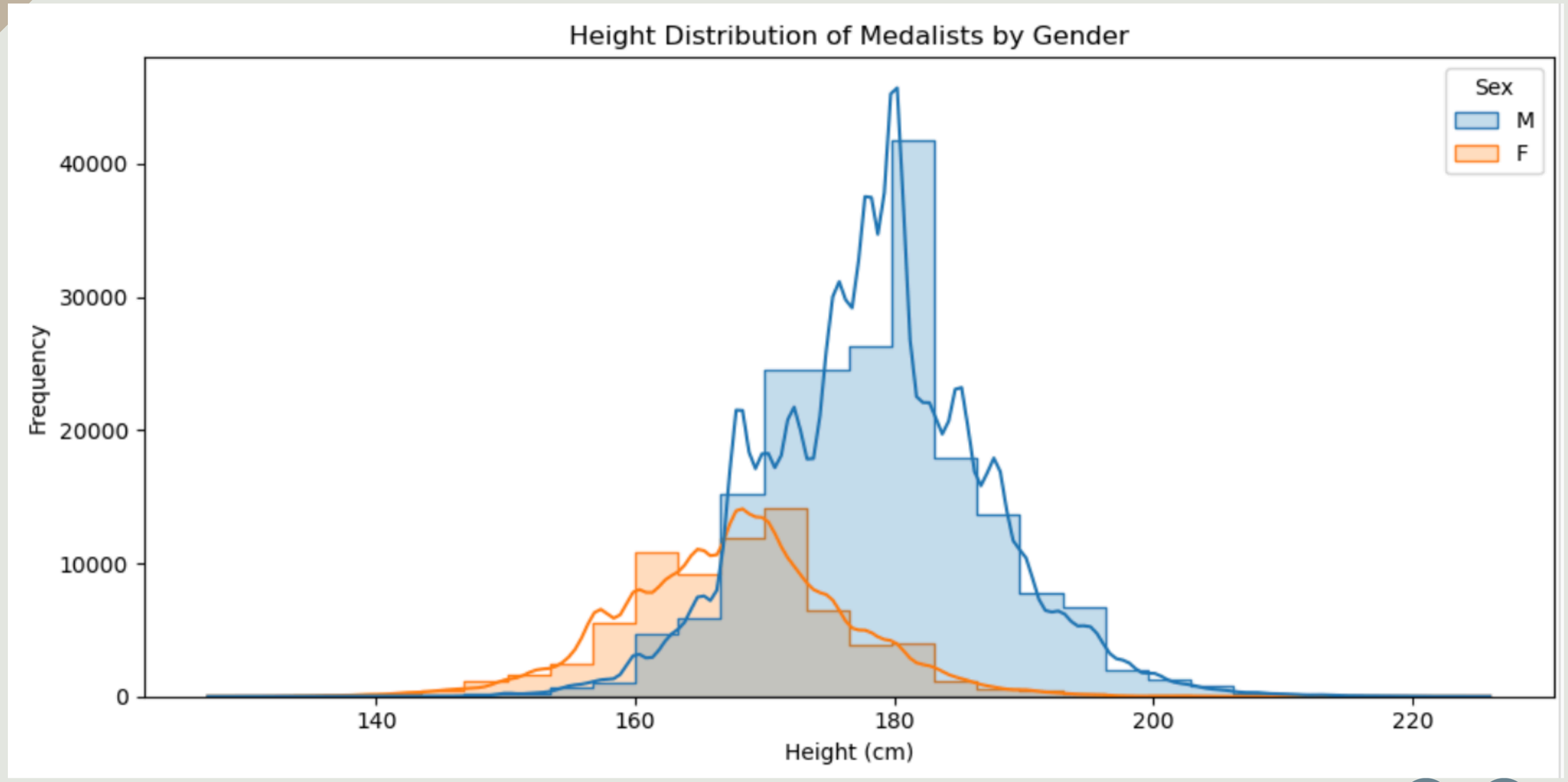
SEASON-SPECIFIC MEDAL TRENDS



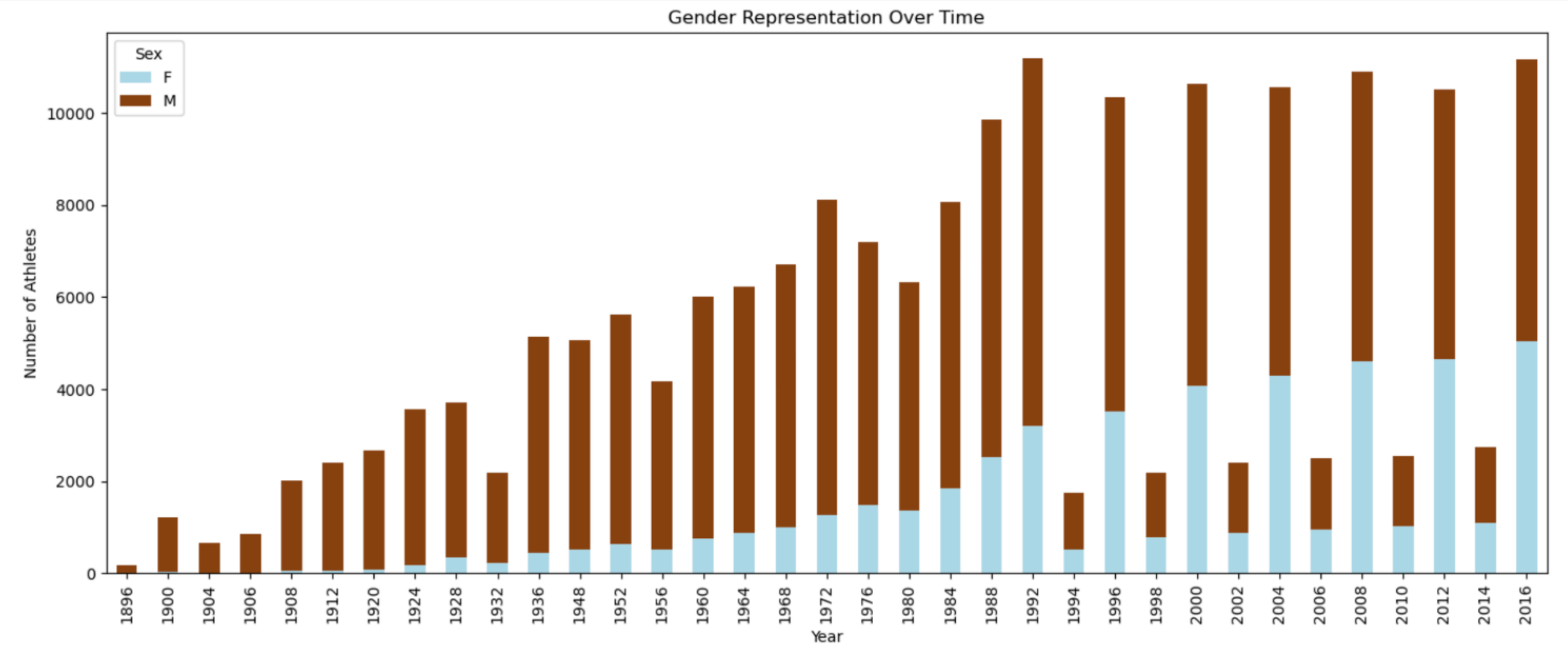
BMI DISTRIBUTION BY SPORT



HEIGHT DISTRIBUTION OF MEDALISTS BY GENDER



GENDER REPRESENTATION OVER TIME



INSIGHTS

1. **Representation Drives Results**
2. **The Summer Games Dominate**
3. **Physical Traits Vary Widely by Sport**
4. **The Olympics Reflect Broader Social Change**

REFERENCES

120 years of Olympic history: athletes and results

basic bio data on athletes and medal results from Athens 1896 to Rio 2016

kaggle.com