Demographic and Performance Trends in Olympic Athletes (1896–2016)

Keon Augustine 003006978

Objective

This project explores Olympic athletes' demographic composition, physical characteristics, and performance trends over time. By analyzing variables such as age, gender, height, weight, and medal distribution, we aim to better understand patterns in athlete selection, participation, and achievement. The final goal is to highlight notable disparities or evolutions within the data and draw insights for future sports research.

Introduction

The Olympic Games, one of the most significant global sporting events, have evolved considerably since their modern inception in 1896. This dataset, containing records of over 270,000 athletes, provides a unique opportunity to explore historical trends and patterns in elite athleticism. With increasing attention to gender equity, athletic performance, and sports science, such an analysis is timely and insightful.

Method: Data Wrangling and Handling Missing Values

The raw dataset presented notable challenges due to missing or incomplete data, particularly in the Age, Height, Weight, and Medal columns. These attributes are critical for the study and thus were cleaned as follows:

- Records missing all three core physical features (Age, Height, Weight) were removed.

- Medals were treated as categorical, with nulls indicating no medal won.

- The cleaned dataset was saved as athlete_events_cleaned.csv for subsequent analysis.

Descriptive statistics were derived using NumPy, and visualizations were produced using Matplotlib and Pandas to ensure clarity and interpretability.

Storytelling Through Data:

The analysis sought to uncover stories in the data that would reveal the evolution of athletic participation, representation, and achievement. Some guiding questions included:

- How has gender representation changed over time?

- What is the typical age or physical profile of an Olympic athlete?

- Are there observable trends in physical characteristics?

- What role does gender play in medal outcomes?

Data Visualization & Interpretation:

1. **Age Distribution of Athletes:** A histogram revealed that most athletes compete between the ages of 20 and 30. This trend corresponds to peak physical performance windows across most sports. Outliers existed, particularly in sports with older average competitors like equestrian and shooting.
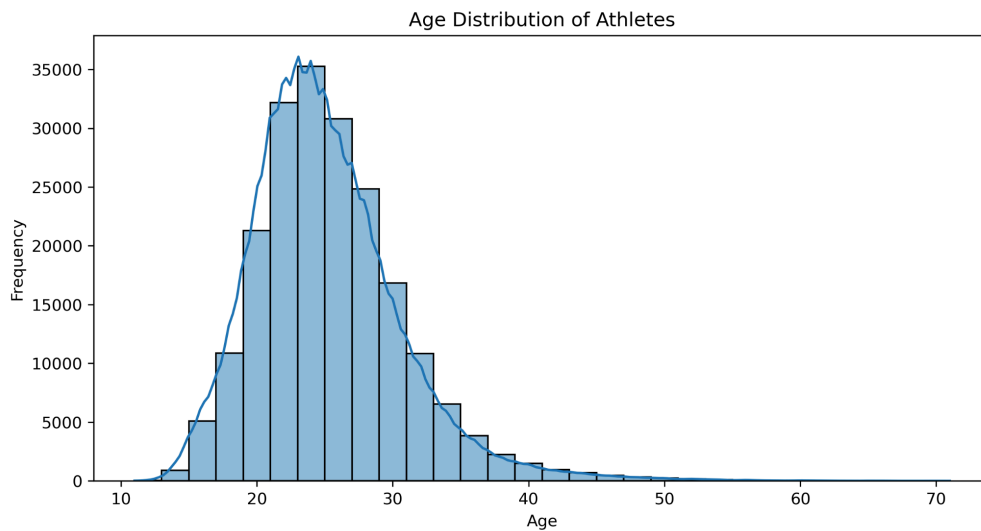

2. **Height Distribution by Gender:** A boxplot comparison showed male athletes are taller on average than female athletes. The spread within each gender group suggests that height is not uniformly advantageous but rather sport-dependent.
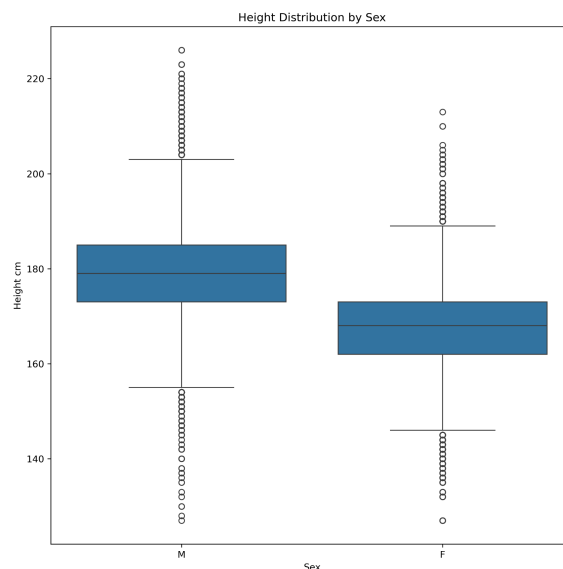
3. **Gender Count Over Time:** A line plot tracking gender count by Olympic year demonstrated significant progress in female representation. From near invisibility in early years to parity with male athletes in recent decades, the Olympics have increasingly supported gender inclusivity.
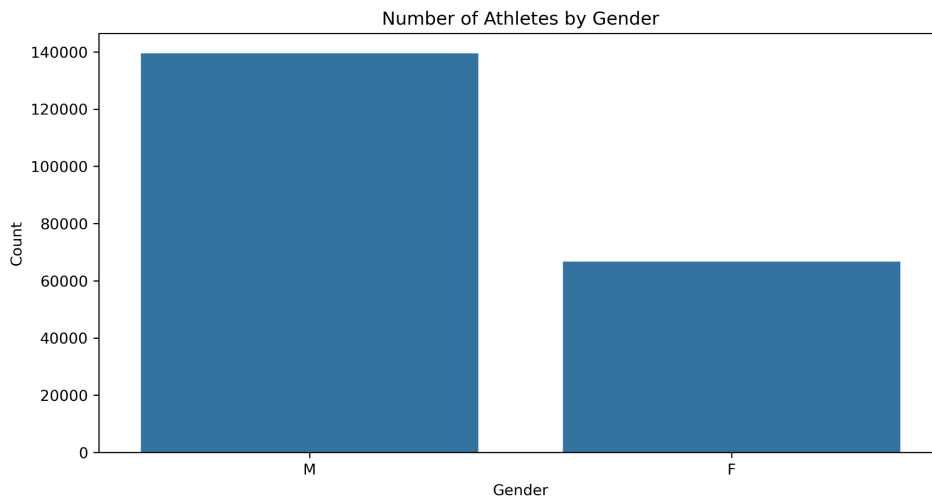

4. **Average Height Over Time:** Tracking average height revealed a gradual increase, possibly attributed to improved nutrition, selection bias toward taller individuals, or shifts in sports prioritizing height.
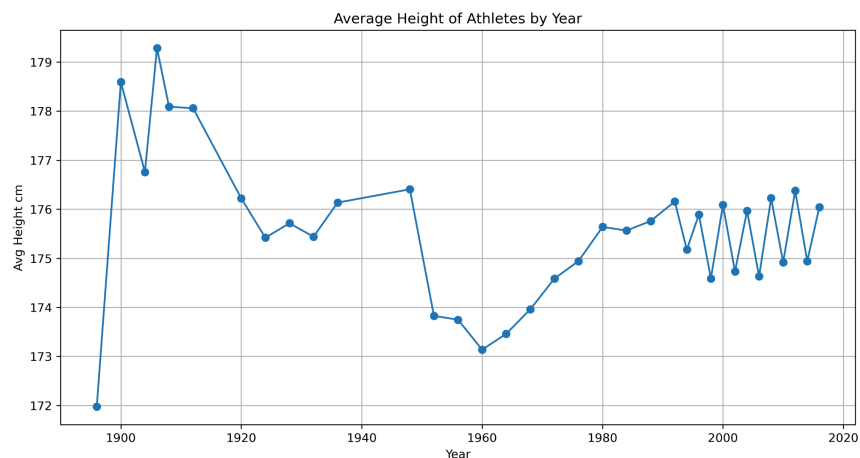
5. **Bubble Plot: Age vs. Weight:** This plot captured the cluster of most athletes within the 20–30 year and 60–90 kg range. Taller athletes tend to weigh more, with bubble size (height) reinforcing this pattern.


Age vs Weight(bubble is height)
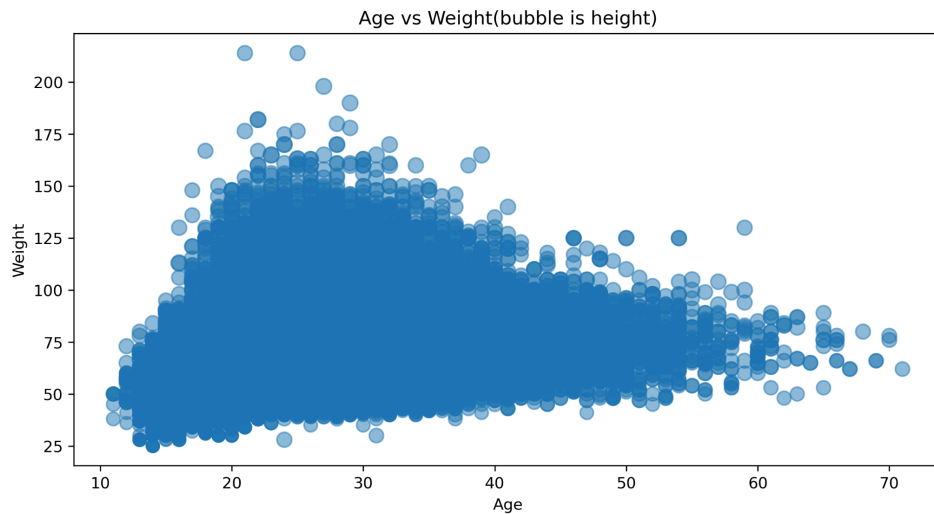
6. **Medal Distribution by Type:** A pie chart illustrates that bronze medals are most commonly awarded, followed by silver and gold. This reflects the higher number of bronze awards, especially in events with two third-place recipients (e.g., boxing, judo).


Medal Distribution
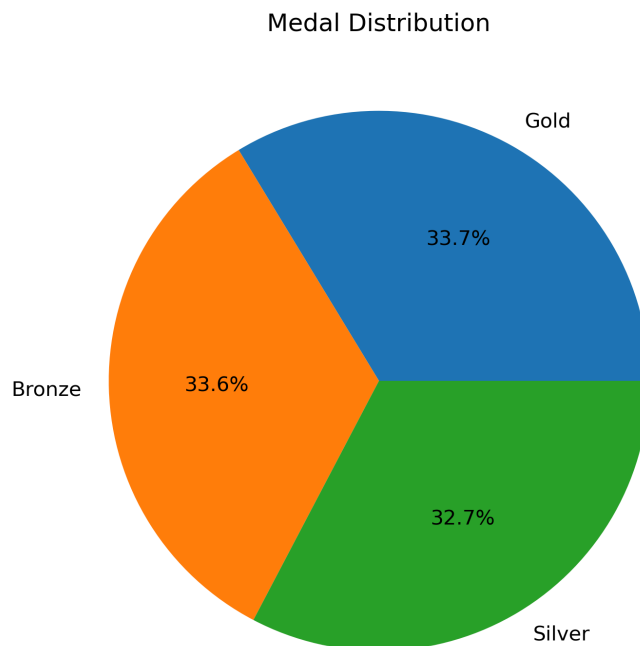
7. **Average Weight by Gender:** A bar chart comparing average weight reaffirmed biological distinctions, with male athletes generally heavier. The variation also reflects different sport requirements (e.g., heavyweights vs. gymnasts).
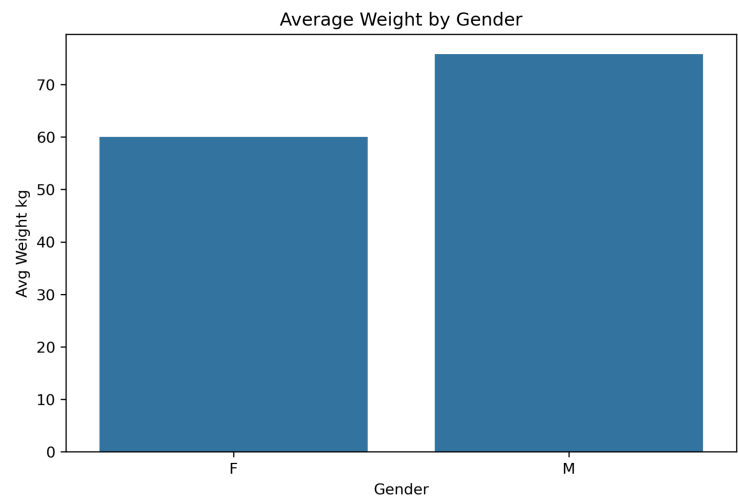

8. **Correlation Heatmap:** This matrix confirmed a strong positive correlation between height and weight, a weaker link between age and weight, and little correlation between age and height. It underscores the diversity of athletic profiles.
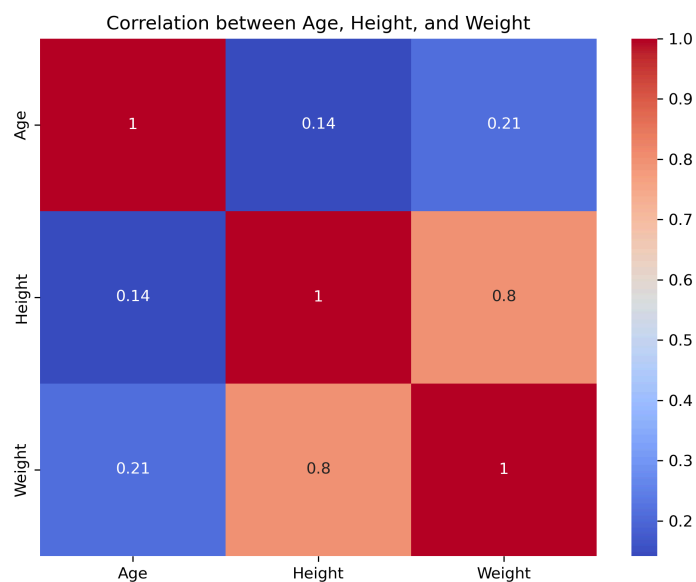
9. **Height vs. Weight Scatter by Gender:** This plot emphasized gender-based clustering, with a near-linear relationship between height and weight. It suggests consistent body proportions within genders despite sport variation.
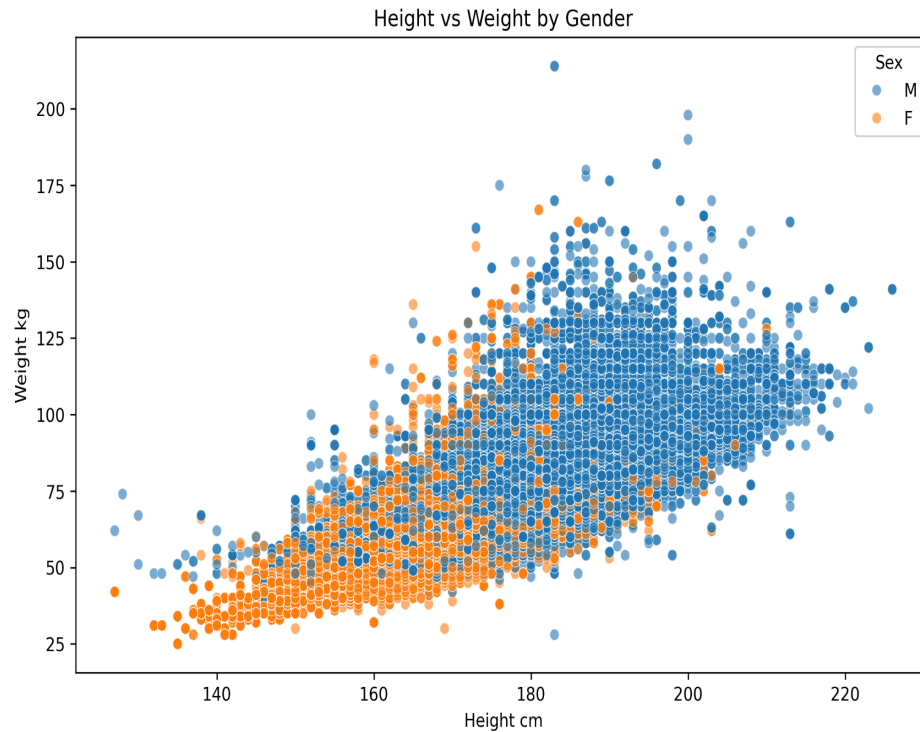
10. **Medals by Gender and Type:** A grouped bar chart showcased that male athletes historically won more medals across all categories, though female medals have increased significantly since the 1980s.
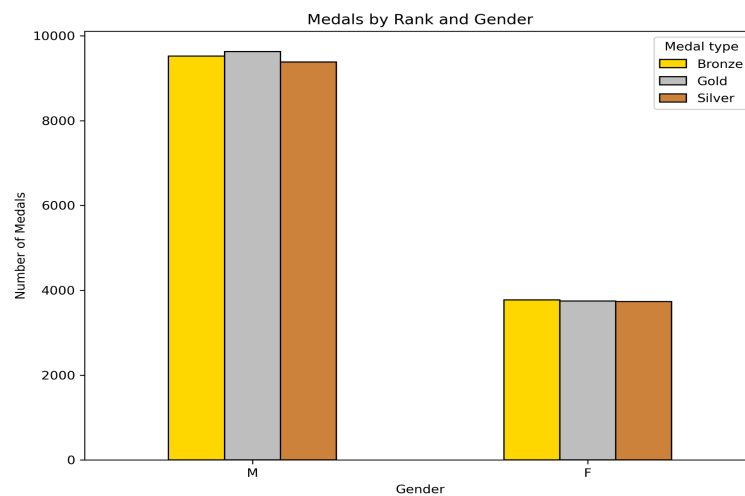
Conclusion

The Olympic athlete dataset reveals a rich history of changing demographics and performance patterns. The most profound insight is the rapid increase in female participation and the diversification of athlete profiles across sports. Strong correlations between height and weight suggest consistent training regimes and physical expectations, though outliers challenge strict profiling. Issues such as missing data were handled conservatively to preserve analytical integrity. Further studies could explore nationality-based trends, multi-sport athletes, or post-career impacts.

References:

- Kaggle: Olympic History Dataset (athlete_events.csv)

- NumPy, Matplotlib, and Pandas documentation