202510-Spring 2025-ITCS-3162-001-Introduction to Data Mining-Project 1

Minthra Khounsavath

February 7, 2025

1 Kaggle Data Set

https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results/data?select=athlete$_e vents.csv$

2 Introduction

Historical records from the Olympic Games from 1896 up to 2016 [1] form the basis of the analyzed dataset. The dataset contains extensive records about 134732 athletes who competed in different sports events and received medals and nationalities and physical information. The vast dataset gives researchers a special opportunity to observe Olympic Games development through population data analysis of athlete achievements and sport movement patterns.

3 Personal Insights into Olympic Sports Dynamics

For the analysis of Olympic data my main interest focused on understanding how individual versus team sports changed over different Olympic competitions. My objective was to monitor sport participation patterns because they reveal changing societal sports preferences and identify the most successful sports between individual and team categories based on medals received. The exploration of these trends included careful implementation of bar charts, line graphs, scatter plots and pie charts which served as visualization tools in my research. The illustrations made complex information accessible to wider audiences by quantifying the data while simultaneously making the complicated information more understandable thus meeting my objective of clarifying Olympic competition trends.

The investigative process revealed important advantages for uncovering specific trends that would aid in making strategic Olympic training decisions while benefiting from participation. My investigation encountered major difficulties because the data set did not fully complete itself. Some athlete records failed to provide complete details about their medal awards thus creating analytic uncertainties. When interpreting distribution findings of medals this limitation required researchers to proceed with caution since the available dataset probably did not accurately show all Olympic athletic achievements. My analysis pushed through obstacles because I wanted to increase knowledge of athletic achievements in the Olympics.

Data Preprocessing Overview

Data preprocessing is a critical phase in any data analysis project, especially when dealing with complex datasets such as those from the Olympic Games. The objective of this phase was to accurately categorize the sports into individual and team events, and to handle any ambiguities in event classification to ensure the reliability of subsequent analyses.

Classification of Sports

The classification process involved segregating sports into three distinct categories:

- **Individual Sports:** These are sports where athletes compete solo, such as Judo, Alpine Skiing, and Boxing. The classification was straightforward for these events based on the nature of the competition.
- **Team Sports:** Sports requiring team collaboration, like Basketball and Football, were grouped under team sports, reflecting their inherent dependency on team dynamics.
- **Ambiguous Sports:** Sports like Swimming and Athletics, which could involve both individual and team competitions (e.g., relays), were initially flagged as ambiguous.

Handling Ambiguities

For sports categorized as ambiguous:

1. Events were further scrutinized to determine whether they should be classified under individual or team sports by examining keywords in the event descriptions such as "Individual" or "Team".
2. A heuristic was applied to manage entries not clearly defined by initial filters, particularly for those entries missing comprehensive event descriptions. For instance, entries with athletes' names containing nested quotes typically indicative of nicknames, a common feature in team sports, were classified as team events.

Final Aggregation

After rigorous filtering and classification:

- All entries were consolidated into comprehensive datasets representing individual and team sports. This step ensured that the final datasets—individual and team—were accurate and inclusive of all relevant entries.
- This process guaranteed that the datasets used for further analysis were reflective of the actual competition types, thereby enhancing the reliability of findings from the analysis.

4 Data Visualization

Using different visualizations in my analysis of the Olympic dataset through a Kaggle notebook enabled me to discover and show relationships and patterns in the data. I used bar charts and line graphs and scatter plots to present data that showed the evolution of athletes taking part in Olympics through time and how medals distributed between individual and team sports. The graphics displayed time-based patterns alongside complete representations of how Olympic sports modify over time. These visual graphs improved research capabilities to comprehend Olympic changes in sports participation and success patterns which revealed common trends as well as unexpected results. The selected approach enabled me to develop a strong narrative integrating numeric information with actual athletic competition events and sports business strategy.

4.1 Visualization 1: Decadal Trends in Olympic Sports Participation

Figure 1 presents detailed information using bar charts to display athlete participation changes between individual sports and team sports throughout numerous decades. The chart displays the exact numbers which demonstrate that individual participation began with 358 people while team members made up 22 in the 1890s but exploded in the 2010s to reach 28,222 individuals and 7,679 teams. Individual sports witnessed a substantial upswing after the 1960s because the participant count expanded from 11,258 in the 1920s to 38,752 during the 2000s. The numbers of team sports participants have increased gradually whereas the

numbers of individual participants have grown substantially during the years before the 2000s. Using this visual representation enables readers to grasp both the numerical growth in participant numbers along with shifts in Olympic event preferences and structural changes that happened over time.

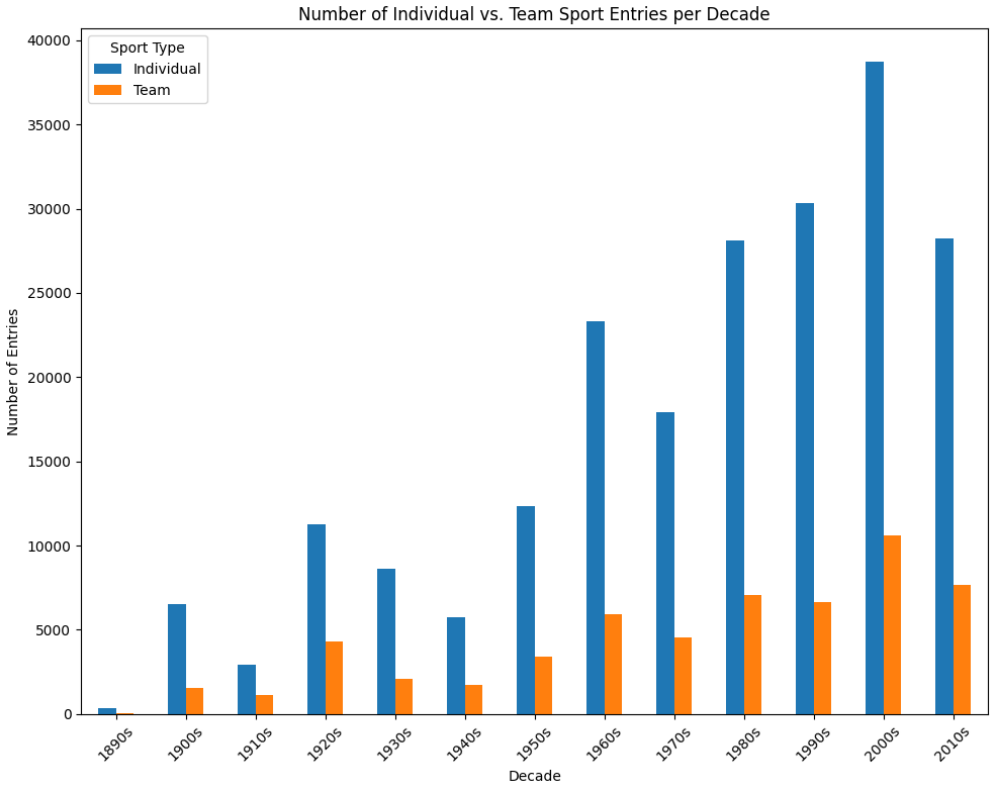

Figure 1: Decadal Trends in Olympic Sports Participation

4.2 Visualization 2: Stacked Bar Chart of Olympic Participation by Sport Type and Decade

A stacked bar chart in Figure 2 depicts athlete participation levels in individual sports and team sports during different decades which matplotlib generated from a Python programming environment. The visualization demonstrates the total athlete number growth from 380 participants in 1890s to 48,357 in 2000s through distinct accumulative data presentation. This bar chart shows how participation numbers in individual sports maintained their elevated rate during a growing trend of team sports throughout different decades. Individual sports maintained a higher rate of engagement than team sports because the 2000s featured 38,752 participants while team sports drew only 10,605 participants. This visual presentation shows how Olympic sports have become progressively popular while growing in number between 1890 and 2000 leading to a complete overview of athlete involvement development within official Olympic competition.

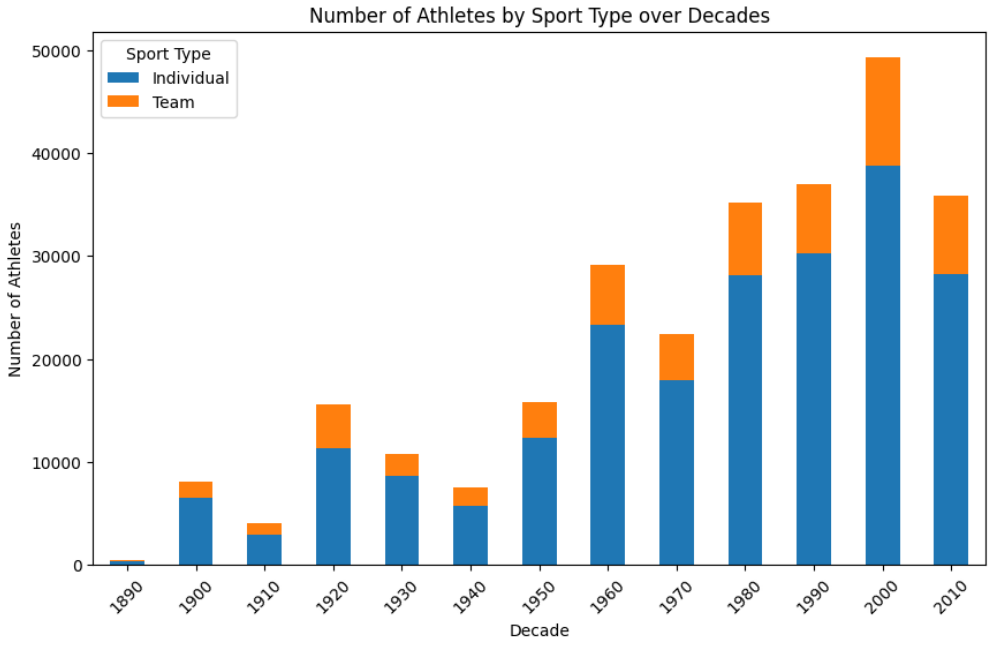

Figure 2: Stacked Bar Chart of Olympic Participation by Sport Type and Decade

4.3 Visualization 3: Line Graph of Olympic Participation Trends by Sport Type Over Time

The participation trends of individual and team sports throughout different decades appear in Figure 3 which was created using matplotlib and seaborn within a Python workspace. The visual representation in the graph illustrates Olympic sports athlete involvement by tracing their participation rates through time from the 1890s up to the 2010s. The 1960s initiated a sudden upswing in individual sports attendance which peaked at 38,752 participants in the 2000s before numbers decreased slightly to 28,222 in the 2010s according to the research data. Participation numbers among team sports rose concurrently with individual sports throughout different decades yet maintained lower growth rates over time.

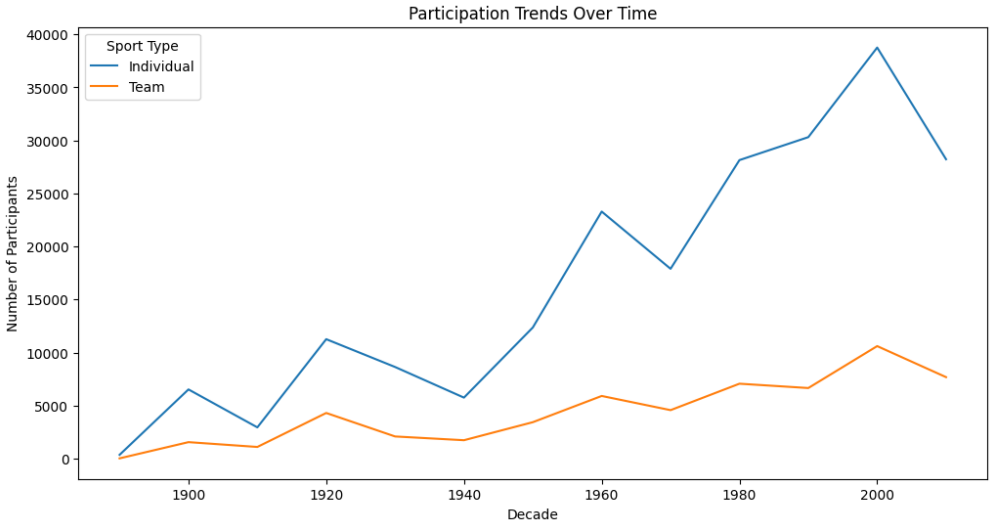

Figure 3: Line Graph of Olympic Participation Trends by Sport Type Over Time

4.4 Visualization 4: Scatter Plot of Medal Count Versus Number of Athletes by Olympic City and Decade

Figure 4 uses scatter plot analysis to examine the connection between participating athletes and awarded medals concerning different Olympic decades and host cities. The visual presentation uses event sizes (athlete numbers) to evaluate success (medal achievement) thus demonstrating the relationship between athlete participation and medal totals. The strong connection between athlete numbers and medal counts is demonstrated by Beijing during the 2000 games and London during 2010 with 10,899 participants and 10,517 participants receiving 2,048 and 1,941 medals respectively. Smaller Olympic competitions like the Lake Placid 1930 event demonstrated different competitive patterns because they had 252 participants who won 92 medals.

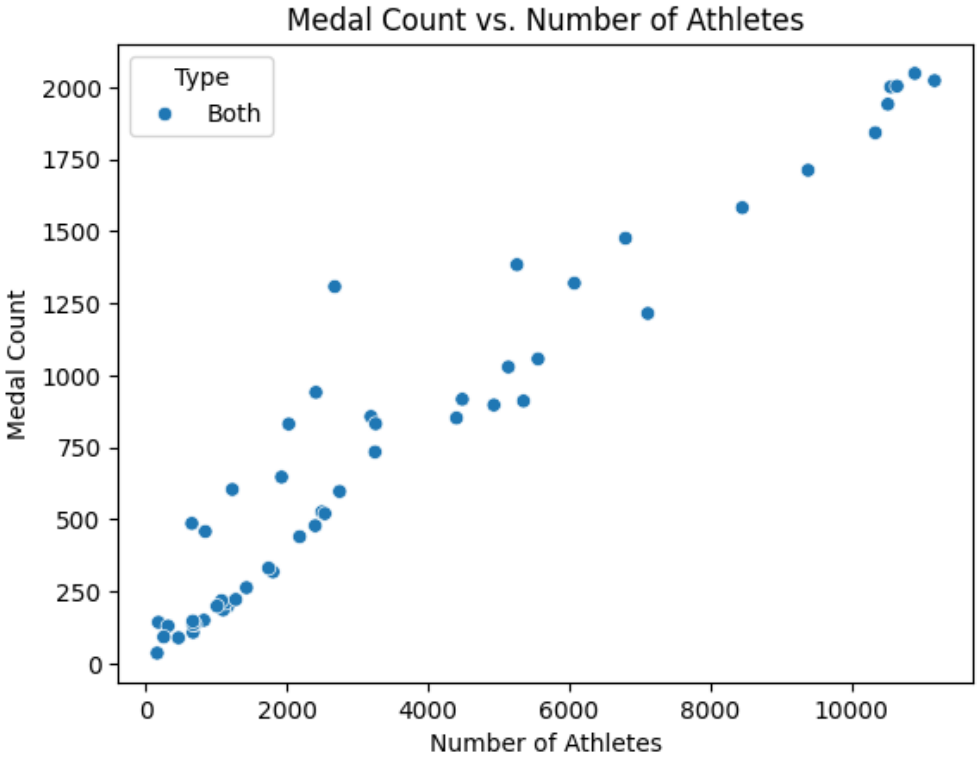

Figure 4: Scatter Plot of Medal Count Versus Number of Athletes by Olympic City and Decade

4.5 Visualization 5: Longitudinal Trends in Participation for Top 15 Olympic Sports

Over the years Figure 5 displays data about the historical development of top 15 Olympic sports through participation rates presented using line graphs. The chart features lines which depict the different sports with the highest participation levels comprising Athletics and Gymnastics among others including Wrestling. This visualization shows how Swimming participation kept climbing annually up to 942 participants during 2016 while Athletics remained a constant sport since 1896 when the modern Olympics began.
The graphical representation demonstrates how different sporting events have gained and lost popularity along with significant transformations that occurred in Olympic competition since 1896. The first year of Cycling competition in 1896 began with 19 participants who gradually resulted in substantial growth throughout subsequent years. The visual presentation of data through storytelling reveals the movement of worldwide sports popularity as well as its connection to cultural developments along with social and technological advancements.

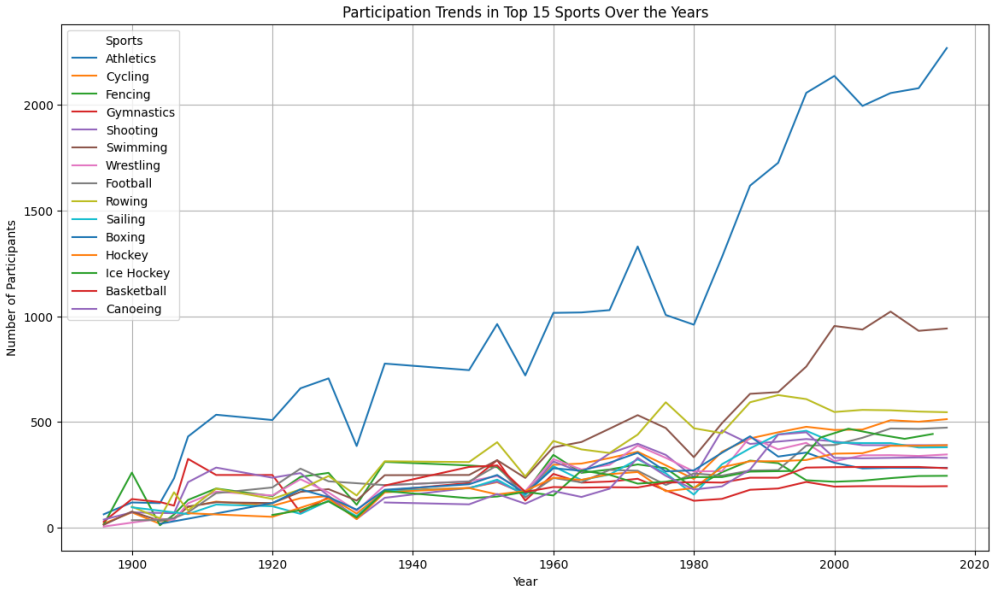

Figure 5: Longitudinal Trends in Participation for Top 15 Olympic Sports

4.6 Visualization 6: Comparative Analysis of Olympic Medal Distribution: Individual vs. Team Sports and Gender Disparities

1. **Figure 6: Medal Comparison in Individual vs. Team Sports**
   The bar chart shows that individual athletes won more medals than teams because participants secured 8,240 bronze medals along with 8,101 gold medals and 7,901 silver medals while teams received 5,055 bronze medals together with 5,271 gold medals and 5,215 silver medals. The picture demonstrates how individual sports yield higher success rates although team sports might have better opportunities.

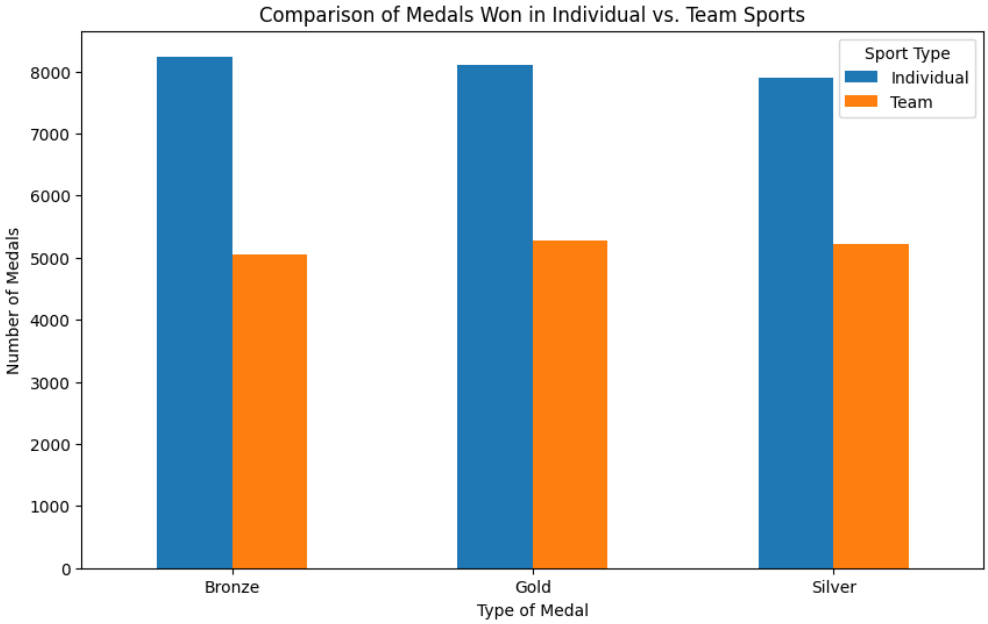

Figure 6: Medal Comparison in Individual vs. Team Sports

2. **Figure 7: Gender-Specific Medal Wins**
   The bar chart depicts gender variations of Olympic medals by displaying male athletes obtained 9,524 bronze along with 9,625 gold pieces and 9,381 silver whereas female athletes earned 3,771 bronze medals and 3,747 gold medals alongside 3,735 silver medals. This chart illustrates the substantial gender differences in medal awards while emphasizing the ongoing debate about gender equality in athletic competition.

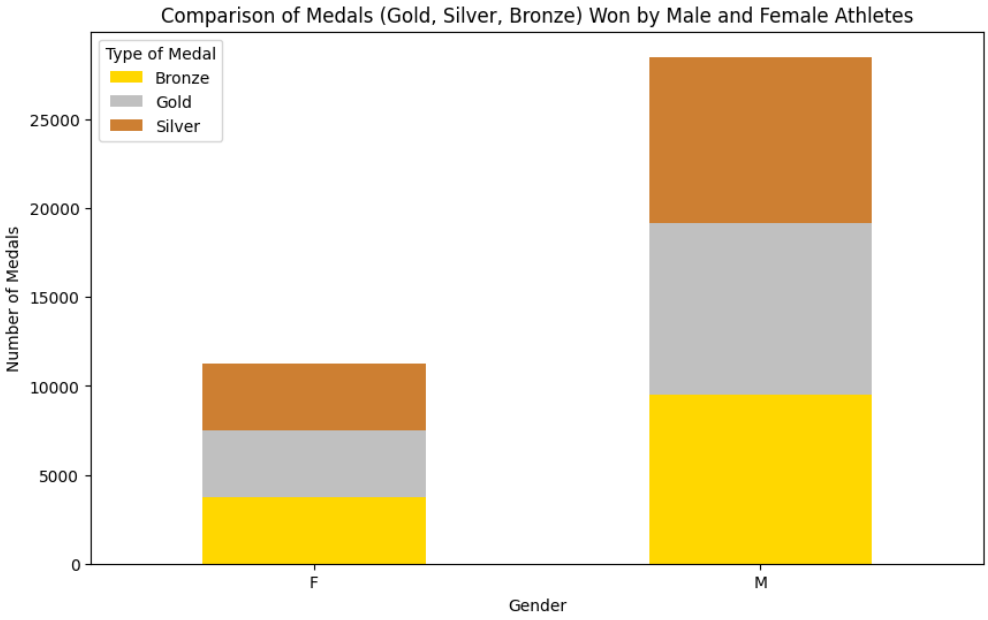

Figure 7: Gender-Specific Medal Wins

3. **Figure 8: Proportional Medal Distribution by Gender**
   A double pie chart shows how male and female athletes obtained their Olympic medals through separate proportional segments that share uniform color indicators for each medal type. The proportion of male and female athletes at the Olympics shows minimal disparity even though males have higher absolute participation numbers because gender equality benefits female athletes equally in performance abilities.

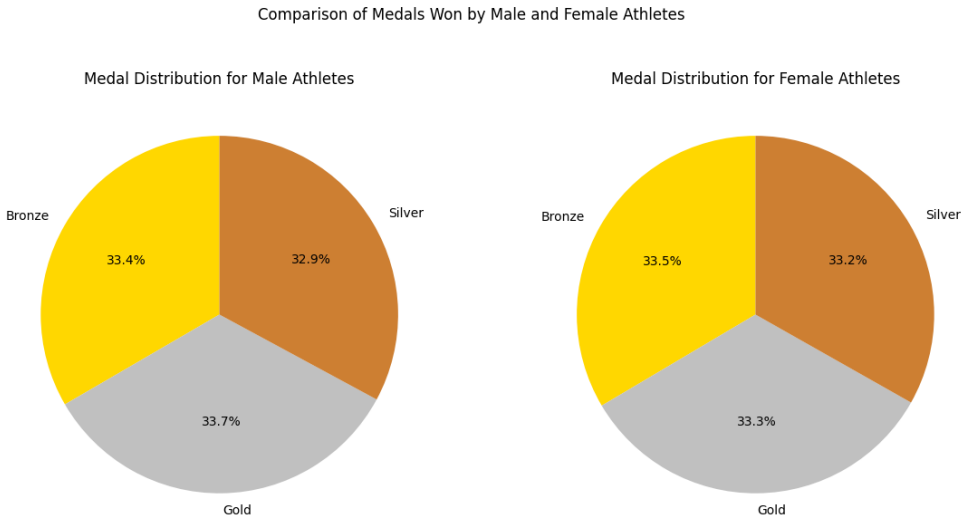

Figure 8: Proportional Medal Distribution by Gender

4.7 Visualization 7: Gender Participation in Olympic Sports: Individual, Team, and Overall Analysis

The collection of pie diagrams demonstrates how women and men are distributed across sports both as individuals and teams and as total Olympic competitors. The stats indicate that individual sports see males lead with 153,721 participants who make up 71.7% while the remaining 28.3% are made up of 60,708 female participants. Team sports feature more men than women according to the second chart since men account for 42,873 participants (75.6%) and women total 13,814 participants (24.4%). A breakdown of Olympic participants through gender statistics shows that males made up 196,594 participants (72.5%) compared to females who totaled 74,522 participants (27.5%). This demonstrates ongoing gender inequality in Olympic sports.

5 Impact of Visualized Trends in Olympic Sports Data

A set of visuals based on Olympic sports metrics displays findings that deliver educational but complex comprehension points. The visual representations track distinct patterns and inequalities that create comprehensive knowledge about sporting events throughout major decades and female versus male involvement and between solo and team-based competitions. The stacked bar charts alongside line graphs demonstrate participation and success rate changes so we can understand possible trends in sports policies together with training methods and global sports attitudes since the 1960s.
The visual displays reveal major gender inequalities and sports type predominance that can guide efforts to distribute resources adequately and establish gender equitable sports environments and advance all-inclusive sports facilities. The explicit data provides evidence of male dominance in sporting representation across individual and team sports since 1960 which calls for research about fundamental causes together with possible solutions to address this gap.

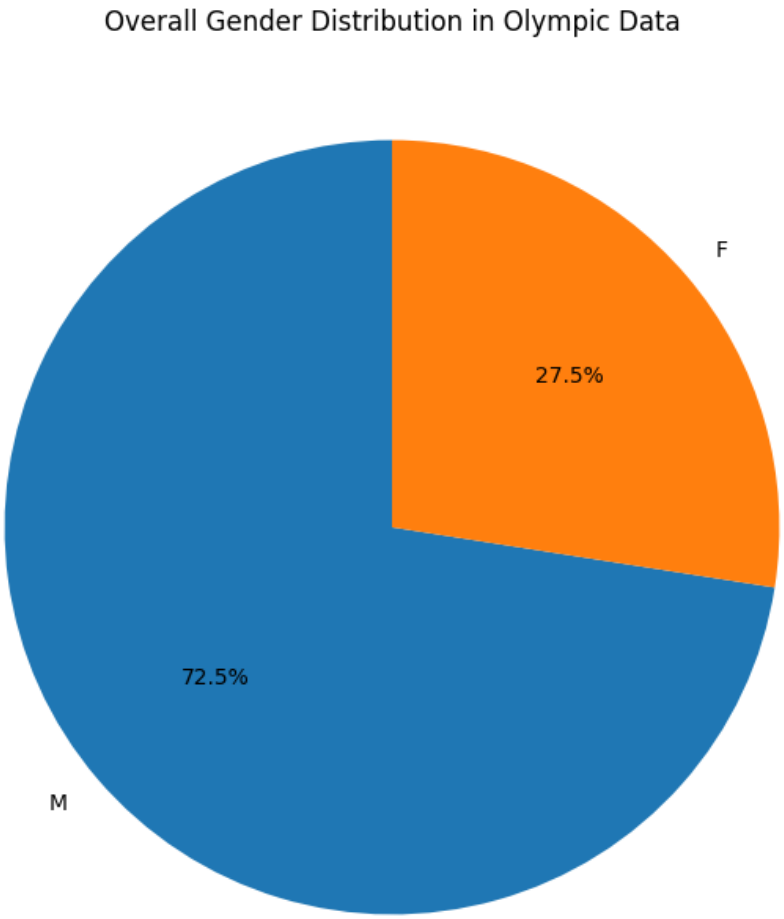

Figure 9: Overall Gender Distribution

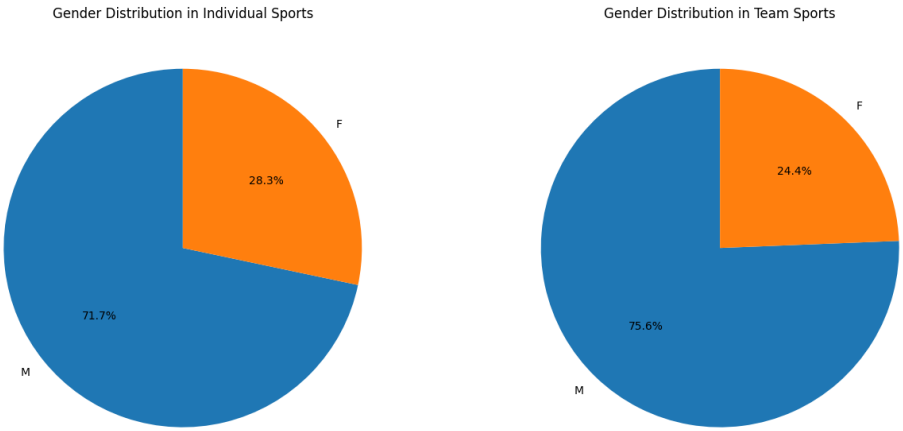

Figure 10: Team vs Individual Gender Distribution

The analysis restrictions stem from incomplete and imprecise data regarding medal awards because some entries lacked this information making the results potentially unreliable. The study could experience limitations in generalizing Olympic findings due to specific constraints regarding time period and location coverage of its data sources. The visualizations deliver solid results yet show only a single view which neglects several aspects including economic situations, political influences as well as technical sports advancements and sensitive situational elements.

References

[1] S. R. LLC, "120 years of olympic history: Athletes and results." Historical data on modern Olympic Games from Athens 1896 to Rio 2016, 2018.