Adnane El Bouhali, Saba Shahsavari, William Liaw

Exploring Athlete Migration and Medal Trends in the Olympics: A Data-Driven Perspective

Palaiseau

2024

ABSTRACT

This report presents a data visualization project focused on analyzing athlete migration and Olympic medal performance during the Olympic Games from 1986 to 2022. Using the Olympic Games Medals 1986-2022 dataset, we explore key trends such as migration patterns of athletes and temporal medal accumulation across countries. The visualizations are designed to reveal patterns in athletic representation and medal distributions over time. Interactive visualizations, including flow maps, choropleth maps, and line charts, were implemented using D3.js to offer a comprehensive analysis of the data. The project underscores the value of visualizing complex, multi-dimensional datasets to derive meaningful insights into historical trends affecting Olympic performance.

Keywords: Olympic Games, athlete migration, data visualization, D3.js, temporal trends, choropleth maps, interactive visualizations.

LIST OF FIGURES

CONTENTS

1   INTRODUCTION

The Olympic Games serve as a remarkable global stage where athletes from various countries compete, showcasing their skills while embodying the cultural and historical narratives of their nations. Over more than a century, the Games have generated a rich repository of data that reflects the evolution of athletic participation, medal distributions, and international representation.

This project leverages advanced data visualization techniques to explore trends and patterns in the Olympic Games from 1986 to 2022. The focus is on uncovering insights into athlete migration and medal performance, using datasets sourced primarily from the Kaggle-hosted Olympic Games Medals 1986-2022 repository [Ivaniuk, 2022]. Through interactive visualizations created with tools like D3.js [Bostock, 2011], we aim to provide an intuitive understanding of the relationships and dynamics underpinning Olympic performance over time.

The following sections outline the dataset, the methodologies employed for data preparation and analysis, and the results derived from the visualizations.

2   DATASET OVERVIEW

This project relies on the *Olympic Games Medals 1986-2022* dataset, sourced from Kaggle [Ivaniuk, 2022]. This dataset was originally scraped from the official International Olympic Committee website [International Olympic Committee, 2024], ensuring its alignment with authoritative data sources. It consists of multiple files, each providing specific details about athletes, medal outcomes, event results, and hosting nations. Key files include:

- `olympic_athletes.csv`: Details about athletes, including demographics and participation history.

- `olympic_medals.csv`: Records of medals awarded across events and disciplines.

- `olympic_hosts.csv`: Information about hosting cities, countries, and the duration of each Olympic event.

- `olympic_results.csv`: Comprehensive details on event results, including participants and rankings.

A full description of the dataset can be found in Appendix A.

3   METHODOLOGY

The analysis of the Olympic dataset involved a combination of data preprocessing, cleaning, and visualization techniques to uncover meaningful insights. The raw dataset, despite its extensive coverage, required significant preprocessing and cleaning to address inconsistencies and ensure its reliability for analysis. Issues such as duplicate records, missing values, and invalid entries (e.g., unrealistic birth years and medal counts) were identified and carefully resolved. These preprocessing steps laid the foundation for the subsequent analysis, as detailed in the following subsections.

3.1   Inconsistent Data Entries

The raw dataset contained several inconsistencies in data entries that required attention during the preprocessing phase. Two significant issues encountered were duplicate records and invalid birth years. These problems had to be resolved to ensure the dataset's reliability and accuracy for analysis.

3.1.1   Duplicate Records

Duplicate entries for athletes and events were present in the dataset, leading to redundancies and inaccuracies. For instance, multiple records existed for the same athlete across different events or Olympic appearances, making it challenging to analyze unique participation trends. These duplicate entries were systematically identified using key attributes, such as athlete names, event details, and medal information, and subsequently removed to maintain data integrity.

3.1.2   Invalid Birth Years

Another major inconsistency was the presence of invalid or missing birth years for athletes. Implausible values, such as negative birth years or years indicating ex-

treme ages (e.g., over 150 years), were identified in the dataset.

To address this, reasonable age boundaries for Olympic participation were established based on historical records. According to sources, the youngest known Olympian was 10 years old [USA Today Sports, 2022], and the oldest recorded Olympian was 73 years old [Oldest.org, nd]. Figure 1 shows the distribution of athletes' birth years by their first Olympic appearance, with anomalies clearly visible as athletes cannot have their first Olympic appearance before the year their were born. The red shaded region highlights implausible values, including negative ages and values exceeding typical human lifespans. Similarly, Figure 2 illustrates the distribution of athletes' ages at their first Olympic appearance.
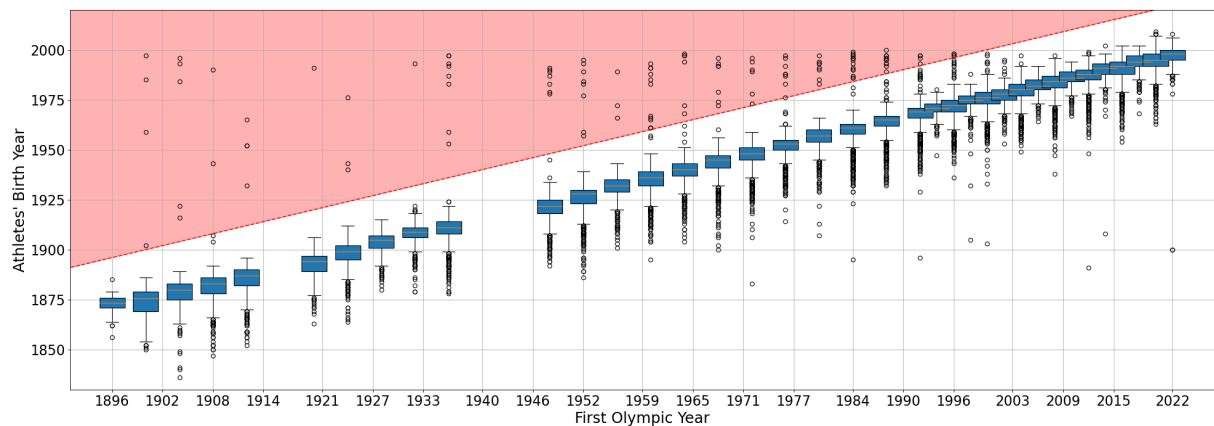

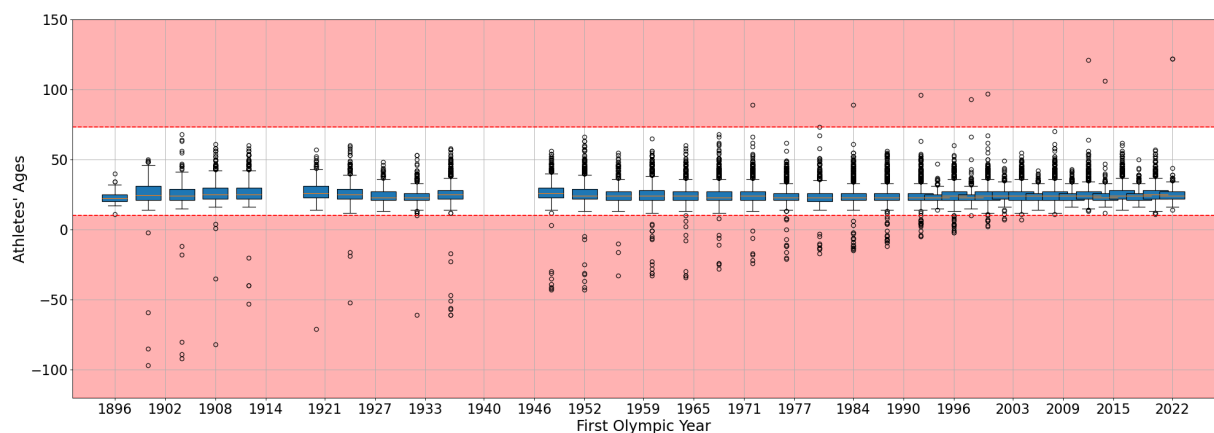Figure 1: Distribution of Athletes' Birth Years by First Olympic Appearance

Figure 2: Distribution of Athletes' Ages at Their First Olympic Appearance

3.2 Ambiguous Geographical Data

Geographical data in the raw dataset was often ambiguous or inconsistent. Key issues included:

- **Host City and Country Mapping:** Host cities were inconsistently labeled or lacked corresponding country information. To resolve this, city names were mapped to their respective countries using geocoding tools such as Nominatim [OpenStreetMap Contributors, nd].

- **Country Code Discrepancies:** Standardized two-letter (ISO 3166-1 alpha-2) and three-letter (ISO 3166-1 alpha-3) country codes were assigned to all entries to eliminate inconsistencies in naming conventions.

3.3 Missing Metadata

A significant portion of the dataset contained missing or incomplete metadata, particularly for athletes and events. Common issues included:

- **Incomplete Athlete Information:** Some athletes lacked URLs, full names, or demographic details. Such records were filtered out when critical information was unavailable.

- **Unresolved Medalist Metadata:** Certain medalists had incomplete associations with their events or disciplines, which limited their analytical use.

4 RESULTS

4.1 Analysis of Olympic Games Duration and Evolution

The Gantt chart in Figure 3 illustrates the scheduling and duration of both Summer and Winter Olympic Games from the inaugural Athens Games in 1896 to the events in 2022. It highlights key milestones and trends in the evolution of the Olympics.

The modern Olympics began in 1896 with exclusively male participants and a limited schedule of events lasting 10 days. By 1900, the Paris Games introduced
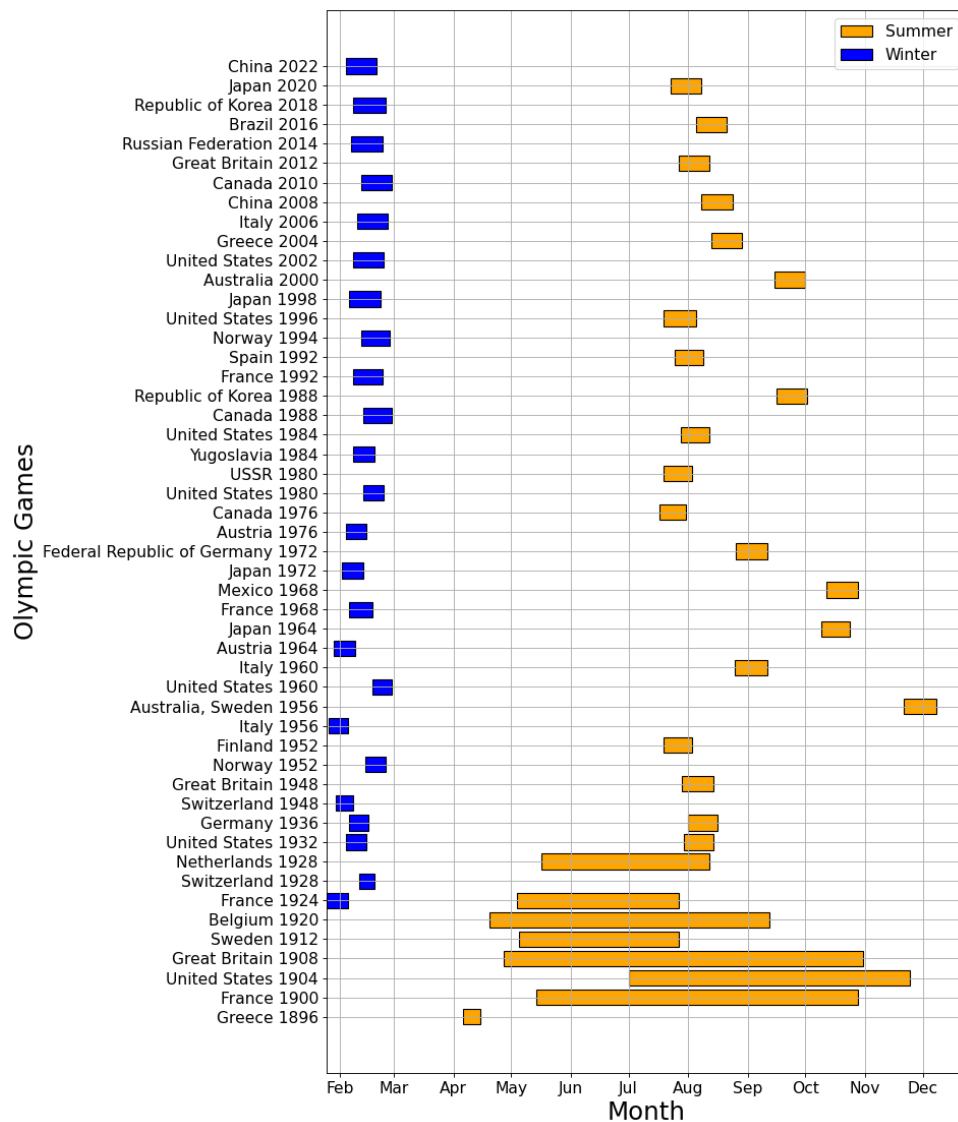
Figure 3: Gantt Chart of Olympic Games Duration

women's participation and expanded the schedule to over five months, integrating the Olympics into the World's Fair. This unusual duration reflected the scattered organization and the addition of new sports like golf, tennis, and rowing. Over time, the duration became more standardized, with events typically lasting around two weeks by the mid-20th century, reflecting the Games' growing scale and complexity.

The Winter Olympics were introduced in Chamonix, France, in 1924, marking the start of a separate seasonal competition for winter sports. Notably, France hosted both the Summer and Winter Games in the same year, solidifying its pivotal role in Olympic history.

Additionally, the chart captures regular scheduling patterns, with Summer Games

typically held between July and August and Winter Games in February. It also reveals disruptions caused by global conflicts, including cancellations during World War I and World War II.

4.2 Chord Diagram for Migration Flow

Given the increasing rate of immigration over the past few decades, we decided to analyze immigration in the context of the Olympic Games, specifically focusing on Olympic athletes who represent countries different from the ones in which they were born. A key component for this analysis was the inclusion of athletes' birth countries, which were not initially present in the dataset. To address this, we used web scraping to extract athletes' birth countries from their Wikipedia pages and added this information as a new feature to the dataset. Since there have been significant changes in the names and territories of countries over time, we decided to focus only on the last 10 Olympic Games to ensure consistency.

To visualize this analysis, we used a chord diagram. A chord diagram is a graphical tool designed to show relationships between entities, where the flow of migration is represented by the chords. The source and destination countries of the migration are indicated by the base and arrowhead of each chord, respectively. The width of each arc is proportional to the number of athletes migrating from one country to another.

In our chord diagram, the countries are grouped by sub-continent, with a total of 9 sub-continents, plus the Refugee Olympic Team, which is treated as a separate category. Each sub-continent is assigned a unique color theme to improve the clarity of the visualization. To make the analysis more manageable and focused, we only included countries with more than 10 immigrant athletes in the diagram. This filter ensures that the visualization highlights significant migration flows.

4.2.1 Features of the Diagram

An interactive feature of this diagram is that when hovering over a country's name, additional information is displayed, including the total number of incoming and outgoing immigrants for that country. Additionally, when hovering over a country, the opacity of arrows that do not originate from or end in that country is reduced. This
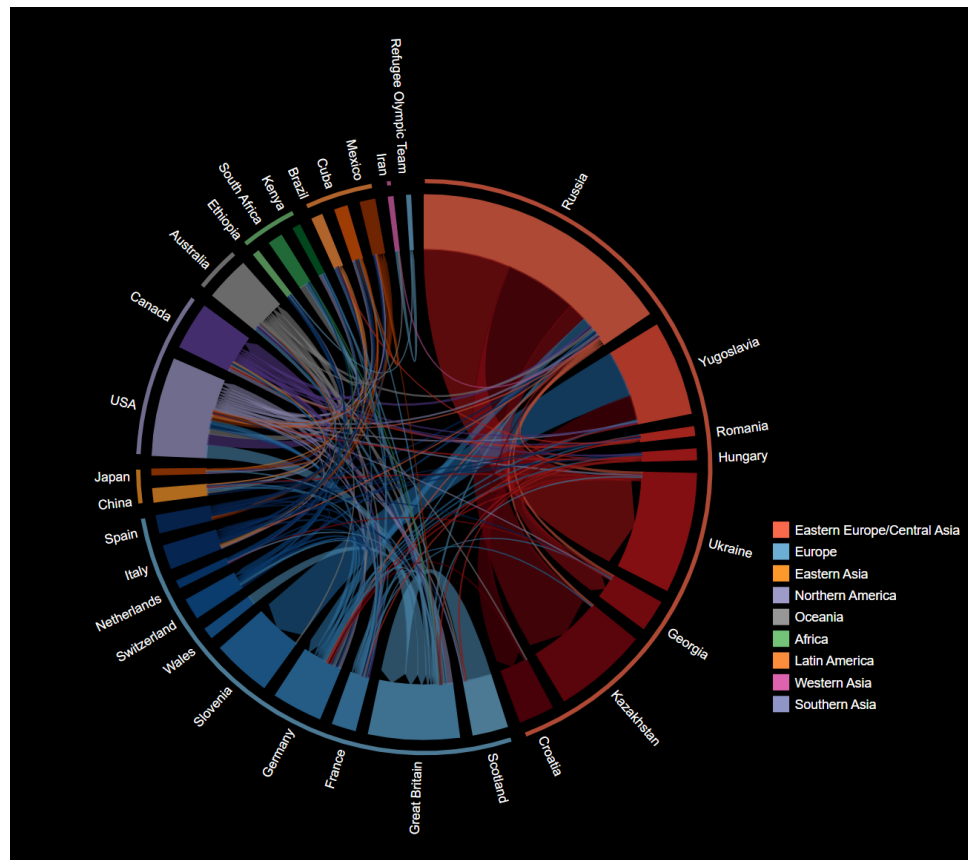
Figure 4: Chord Diagram: Visualizing Olympic Athletes' Migration Flows Since 2010

allows the user to focus on immigration flows specifically related to the selected country, highlighting it as the source or destination

4.2.2 Insights from the Diagram

The diagram provides valuable insights into immigration patterns. Most visible immigration flows align closely with political and geographical changes in the world. In particular, highly developed countries like the United States and Britain emerge as major destinations for immigrants, likely due to the better opportunities and higher quality of life they offer.

However, the diagram also reveals patterns specific to the unique nature of the dataset. For instance, Russia was banned from participating in the 2016 Rio and 2020 Tokyo Olympics due to doping violations. In such cases, many Russian athletes opted to compete under other countries' flags just to continue participating in the games.

Another intriguing trend is the movement of athletes between countries in pursuit

of better athletic opportunities. For example, some U.S. athletes appear to have moved to Latin American countries, possibly because it was easier to qualify for the Olympic team there.

4.3   Dynamic Choropleths and Bubble Maps

The dynamic choropleth and bubble maps were designed to provide an interactive exploration of Olympic trends worldwide. These maps enable users to visualize data such as the frequency of Olympic hosting by country, the number of athlete debuts, and medal counts. With dynamic filtering options and multiple visualization styles, they offer a versatile and engaging way to analyze key aspects of Olympic history.

The choropleth map excels at conveying spatial patterns and regional comparisons through color intensity, making it ideal for quickly identifying geographical trends. However, it may obscure information for smaller countries due to their limited map space. In contrast, the bubble map highlights individual data points with proportional circle sizes, ensuring that smaller countries remain visible and providing a more precise representation of numerical values. On the downside, bubble maps can become cluttered in regions with dense data points, potentially making it harder to interpret overlapping bubbles. Together, these visualization styles complement each other, balancing clarity and detail depending on the user's analytical needs.

4.3.1   Features of the Maps

The maps provide the following functionalities:

- **Visualization Modes:** Users can switch between a choropleth map (Figure 5) and a bubble map (Figure 6) for different visual representations of hosting frequency.

- **Filters and Options:** Filtering options include continents, sub-regions, seasons (Summer or Winter), medal types (Gold, Silver or Bronze), if applicable. These allow users to customize the view and focus on specific geographic or temporal trends.
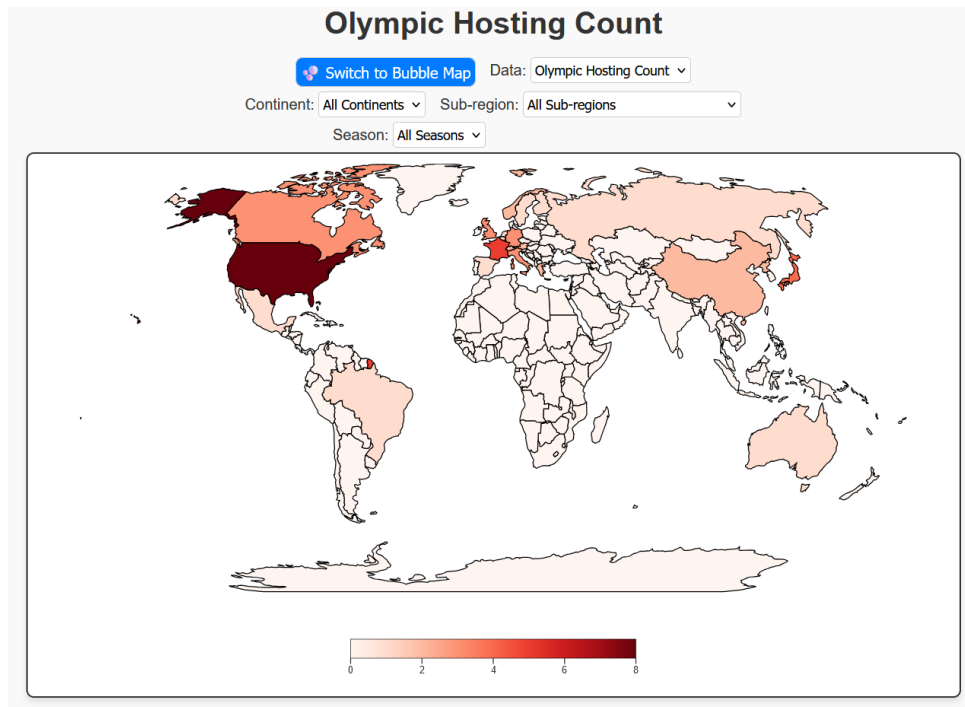
Figure 5: Dynamic Choropleth Map: Visualizing Olympic Hosting Trends

- **Interactivity:** Both maps feature hover-over tooltips displaying detailed information, such as the number of times a country has hosted the Olympics.

4.3.2 Insights from the Maps

These maps reveal several insights into Olympic hosting trends:

- Countries such as the United States, Great Britain, and France stand out in number of debuts and high hosting frequencies, reflecting their longstanding involvement in the Olympics.

- Hosting occurrences are concentrated in Europe and North America, particularly during the early years of the modern Olympics.

- The inclusion of filtering options enables the identification of hosting trends by region, season, and medal type, highlighting the geographical expansion of the Games over time.

The combination of choropleth and bubble maps exemplifies the power of interactive visualizations, enabling a flexible exploration of the data while revealing key trends and disparities in Olympic hosting.
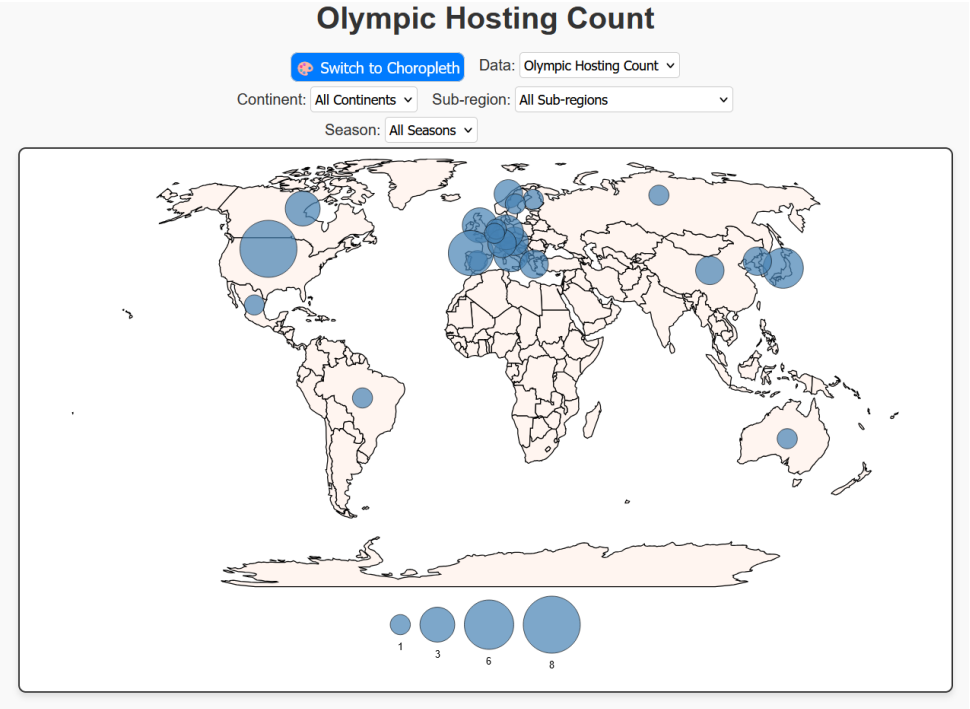
Figure 6: Dynamic Bubble Map: Visualizing Olympic Hosting Trends

4.4 Interactive Cumulative Medals Visualization

This section introduces an interactive visualization that dynamically displays the evolution of cumulative Olympic medals for the top 20 countries, based on total medal counts from 1896 to 2022. The visualization allows users to select specific countries, adjust the animation speed, and explore trends over time.

Figure 7: Interactive Line Chart of Cumulative Olympic Medals

4.4.1 Features of the Visualization

The interactive line chart provides the following functionalities:

- **Dynamic Country Selection:** A control panel allows users to select countries of interest, focusing the visualization on specific datasets.

- **Adjustable Speed:** Users can modify the animation speed through a slider, enabling both detailed and high-level explorations.

- **Dynamic Axes and Gridlines:** Axes and gridlines adjust dynamically during the animation, ensuring a clear representation of trends.
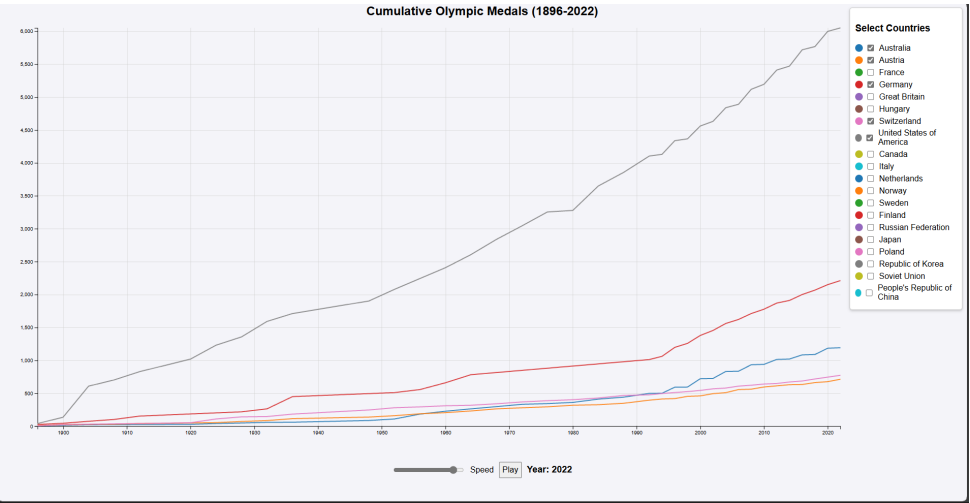
- **Year Tracking:** A label dynamically updates to display the current year during the animation, providing a temporal reference.

4.4.2 Insights from the Visualization

This visualization offers several insights into Olympic medal trends:

- Countries such as the United States, Soviet Union, and Great Britain demonstrate consistently high cumulative medal counts, reflecting their historical dominance in the Olympics.

- Newer Olympic nations exhibit rapid growth in cumulative medals, showcasing their emerging success over time.

- The interactive nature allows users to identify key turning points, such as the effects of global events (e.g., World Wars) on medal accumulation.

This visualization exemplifies the potential of dynamic and interactive tools in enhancing the exploration and understanding of historical data trends.

4.5 Olympic Medal Density Visualization

This section introduces an interactive map visualization of Olympic medal distributions over time. The map, shown in Figure 8, dynamically displays the density of medals won by countries in each Olympic Games from 1980 to 2022, allowing users to explore geographical trends in medal achievements.
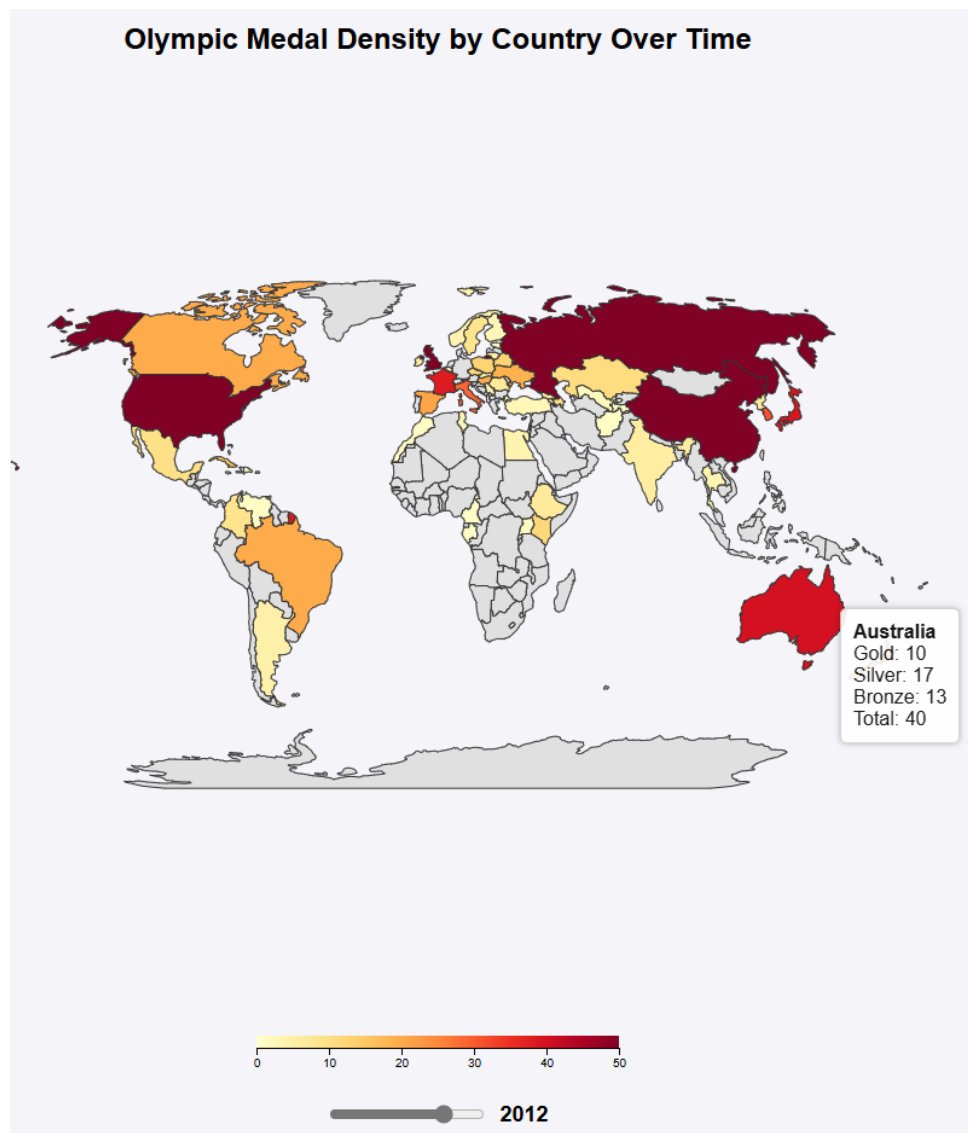
Figure 8: Olympic Medal Density Map: Medal Distributions by Country (1980-2022)

4.5.1 Features of the Visualization

This interactive map includes the following functionalities:

- **Dynamic Medal Visualization:** Countries are shaded based on the total number of medals won, with darker colors indicating higher medal counts.

- **Year Slider:** A slider allows users to select specific years, updating the map to reflect medal distributions for that year.

- **Tooltip Interactivity:** Hovering over a country displays detailed information about its medal achievements, including counts for Gold, Silver, Bronze, and Total

medals.

- **Color Legend:** A color bar legend provides a reference for interpreting the shading intensity, corresponding to medal density.

4.5.2 Insights from the Visualization

This visualization reveals significant patterns in Olympic medal distributions:

- Traditional powerhouses like the United States, Russia, and China consistently dominate the medal tallies, highlighted by darker shades on the map.

- Regional trends, such as strong performances by European nations and emerging contributions from African and Asian countries, become apparent over time.

- Disruptions in medal patterns during specific years, such as the Cold War-era boycotts of 1980 and 1984, are easily observable.

- The interactive slider enables a temporal analysis, allowing users to trace the evolution of Olympic success across decades.

This map provides an engaging and intuitive way to explore the geographical distribution of Olympic success, offering insights into historical trends and emerging nations in the Olympic arena.

4.6 Olympic Rivalries Network Visualization

This section presents an interactive network visualization that explores the complex relationships and rivalries between countries in the Olympic Games. The visualization, shown in Figure 9, highlights countries' total medal counts and their competitive interactions over time, with connections emphasizing shared participation in events.

4.6.1 Features of the Visualization

The interactive network provides the following functionalities:

- **Node Representation:** Each node represents a country, with its size proportional to the total number of medals won.
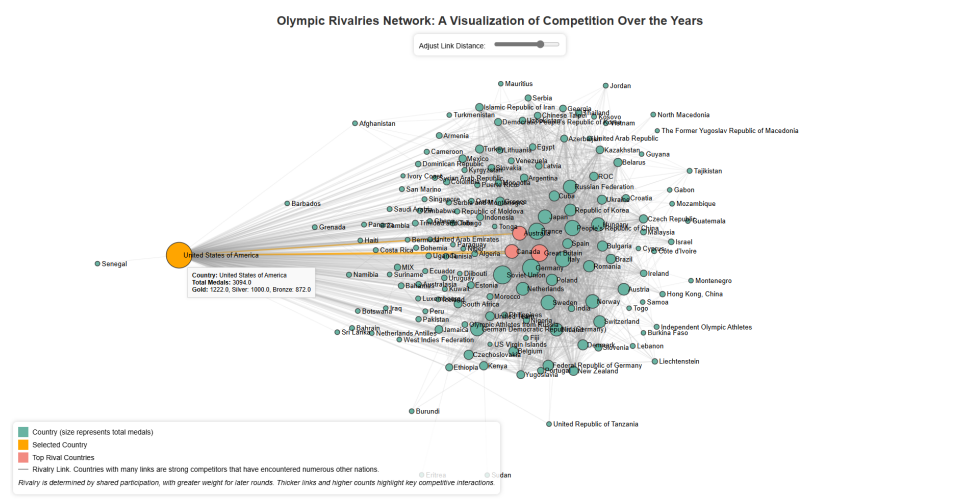
Figure 9: Interactive Olympic Rivalries Network: Countries and Medal Links

- **Link Representation:** Links between nodes represent competitive interactions, with thicker links indicating stronger rivalries based on shared participation, especially in later rounds.

- **Interactivity:**

  - Hovering over nodes displays a tooltip with the country name, total medals, and medal breakdown by type (Gold, Silver, Bronze).

  - Clicking a node highlights the top three rival countries and their competitive links.

- **Dynamic Adjustments:** A slider allows users to adjust the link distance, dynamically reorganizing the network for clarity.

- **Legend:** A detailed legend explains the meaning of node sizes, colors, and link thickness, ensuring user accessibility and understanding.

4.6.2 Insights from the Visualization

The Olympic Rivalries Network provides unique insights into the relationships and rivalries between countries:

- Countries with higher medal counts, such as the United States, Russia, and China, emerge as central nodes with numerous connections, highlighting their dominance in the Olympics.

- Strong rivalries are evident between geographically or politically significant nations, such as the USA and USSR during the Cold War era.

- Smaller nations with significant connections, such as Jamaica in athletics or Kenya in long-distance running, showcase specialized rivalries in specific sports.

- The interactive design allows users to explore the evolving competitive dynamics over time, emphasizing the global nature of the Olympics.

This visualization exemplifies the power of network analysis in uncovering patterns and relationships in complex datasets, offering a dynamic and engaging perspective on Olympic history.

4.7 Analysis of Olympic Participation Trends Over Time

This section provides an analysis of the trends in global participation in the Olympic Games from their inception in 1896 to the most recent Games in 2021. The visualization in Figure 10 captures the number of countries participating in each Olympic Games, highlighting key historical events and their impact on participation.

4.7.1 Features of the Visualization

This interactive line chart and marker-based visualization include the following elements:

- **Participation Trends:** A blue line representing the number of countries participating in each Olympic Games, excluding canceled years.

- **Event Markers:** Circles highlighting key historical events, such as World War-related cancellations and notable boycotts.

- **Shaded Regions:** Gray regions marking periods when the Olympics were canceled (e.g., during World War I and World War II).

- **Interactive Tooltips:** Hovering over points provides detailed information about the year, number of participating countries, and significant events.
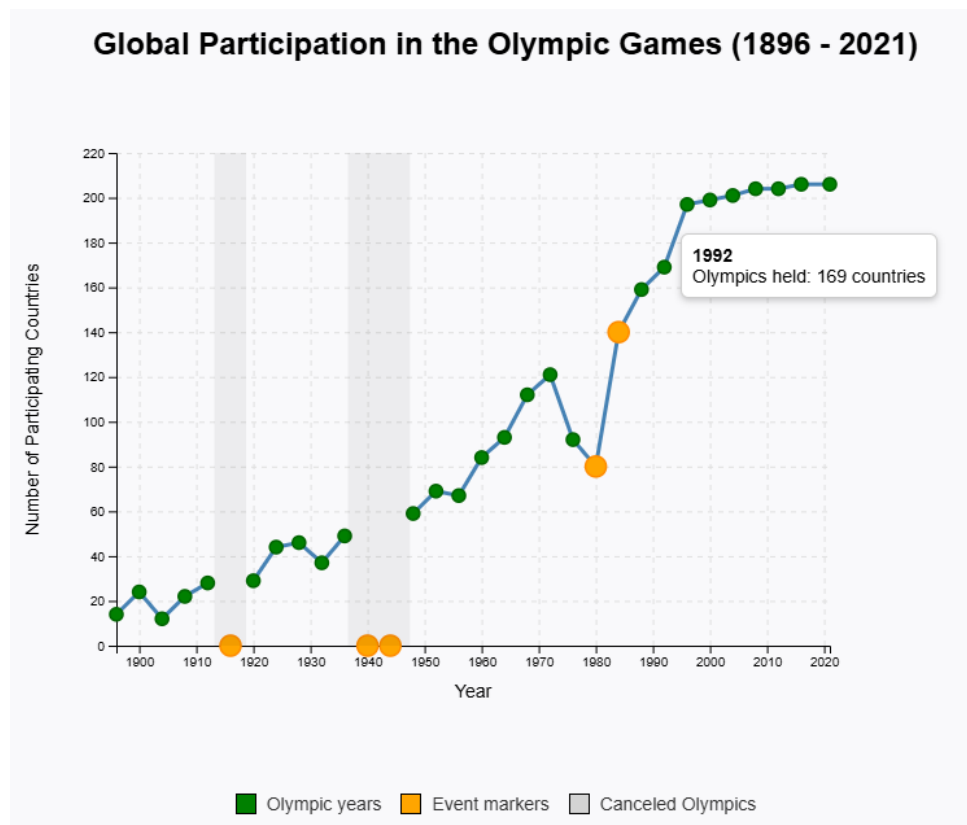
Figure 10: Olympic Participation Trends: Number of Participating Countries (1896-2021)

- **Legend:** A color-coded legend explains the meaning of the different visual elements, ensuring clarity and accessibility.

4.7.2 Insights from the Visualization

This visualization reveals several key insights into Olympic participation trends:

- Participation steadily increased from 14 countries in the inaugural Games to 206 countries in recent editions, reflecting the Olympics' growing global appeal.

- Significant dips in participation occurred during the 1980 and 1984 Games due to Cold War-era boycotts.

- Periods of cancellations during World War I and World War II (1916, 1940, and 1944) are clearly marked, emphasizing the impact of global conflicts on the Olympics.

- The introduction of new nations in the post-World War II era led to a sharp rise in

participation, with notable milestones in the 1992 Games (169 countries) and the post-Cold War 1996 Games (197 countries).

By combining historical context with interactive elements, this visualization provides a comprehensive overview of how global events and geopolitical shifts have shaped the Olympic Games over time.

4.8 Data Dashboard

This HTML page serves as a centralized hub for all visualizations, consolidating plots and charts to offer a clear and organized view of the data. Each plot acts as a clickable link, redirecting users to the corresponding full visualization page. A description overlay appears on hover, providing additional context for each plot.

The page leverages various HTML and CSS features, including iframe for embedding visualizations, zoom and transform for dynamic scaling, and hover effects for interactivity. Additionally, the responsive grid layout ensures the visualizations adapt seamlessly to different screen sizes, maintaining proper spacing and alignment for an optimized user experience.
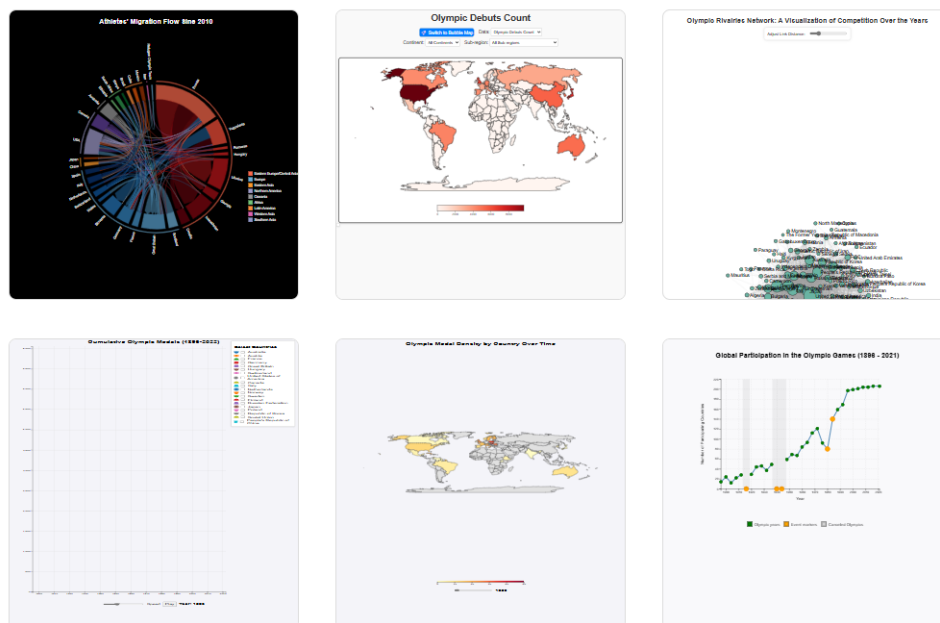

Figure 11: Data Dashboard: A centralized hub for displaying and accessing all visualizations

5   CONCLUSION AND FUTURE WORK

In this work, we have created multiple plots to analyze Olympic Games data and uncover deeply embedded patterns within it. These visualizations have allowed us to derive new insights about the Olympics from both geographical and social perspectives.

In future work, Olympic Games data can be analyzed from political and economic perspectives. By integrating this data with economic indicators such as the GINI index(a measure of income inequality within a country) and GDP(Gross Domestic Product), it would be possible to gain deeper insights into how economic conditions and political factors influence participation, performance, and trends in the Olympics.

REFERENCES

[Bostock, 2011] Bostock, M. (2011). D3.js - data-driven documents. `https://d3js.org/`. Accessed on 2024-12-11.

[International Olympic Committee, 2024] International Olympic Committee (2024). Olympics official website. `https://olympics.com/`. Accessed on 2024-12-11.

[Ivaniuk, 2022] Ivaniuk, P. (2022). Olympic games medals 1986-2022 dataset. `https://www.kaggle.com/datasets/piterfm/olympic-games-medals-19862018`. Accessed: 2024-12-11.

[Oldest.org, nd] Oldest.org (n.d.). 10 oldest olympians in history. `https://www.oldest.org/sports/olympians/`. Accessed: 2024-12-11.

[OpenStreetMap Contributors, nd] OpenStreetMap Contributors (n.d.). Nominatim geocoding tool. `https://nominatim.org/`. Accessed: 2024-12-11.

[USA Today Sports, 2022] USA Today Sports (2022). Who is the youngest olympian? `https://eu.usatoday.com/story/sports/olympics/2022/11/24/who-is-youngest-olympian/10380713002/`. Accessed: 2024-12-11.

APPENDIX A – COMPLETE DATASET DESCRIPTION

The *Olympic Games Medals 1986-2022* dataset, sourced from Kaggle [Ivaniuk, 2022], contains comprehensive information about the Olympic Games spanning from 1986 to 2022. The data, originally scraped from the official International Olympic Committee website [International Olympic Committee, 2024], includes details about athletes, medal results, hosting nations, and events. This section provides a complete overview of the dataset structure and its attributes.

The dataset consists of multiple files, each focusing on specific aspects of the Olympic Games. The key files and their contents are described below.

A.1 olympic_athletes.csv

The file `olympic_athletes.csv` includes the following attributes:

- `athlete_url`: The URL linking to the athlete's profile on the official Olympics website.

- `athlete_full_name`: The full name of the athlete.

- `games_participations`: The total number of Olympic Games in which the athlete participated.

- `first_game`: The first Olympic Games in which the athlete competed, specified by city and year.

- `athlete_year_birth`: The birth year of the athlete.

- `athlete_medals`: The total medals won by the athlete across all participations.

- `bio`: A brief biographical summary or description of the athlete, if available, which is often not the case.

A.2 olympic_medals.csv

The file `olympic_medals.csv` includes the following attributes:

- `discipline_title`: The title of the sport or discipline (e.g., `Swimming`, `Athletics`).

- `slug_game`: A unique identifier for the specific Olympic Games (e.g., `barcelona-1992`).

- `event_title`: The title of the specific event within the discipline (e.g., `5000m men`, `parallel bars men`).

- `event_gender`: The gender category of the event (`Men`, `Women`, or `Mixed` for mixed events).

- `medal_type`: The type of medal awarded (`GOLD`, `SILVER`, or `BRONZE`).

- `participant_type`: The type of participant (e.g., `Athlete` if individual, `GameTeam` otherwise).

- `participant_title`: The name or title of the participant (e.g., team name or athlete name).

- `athlete_url`: The URL linking to the athlete's profile (if applicable).

- `athlete_full_name`: The full name of the athlete (if applicable).

- `country_name`: The name of the country associated with the medal-winning participant.

- `country_code`: The two-letter country code (ISO 3166-1 alpha-2).

- `country_3_letter_code`: The three-letter country code (ISO 3166-1 alpha-3).

A.3 olympic_hosts.csv

The file `olympic_hosts.csv` includes the following attributes:

- `game_slug`: A unique identifier for the specific Olympic Games (e.g., `barcelona-1992`).

- `game_end_date`: The end date of the Olympic Games.

- `game_start_date`: The start date of the Olympic Games.

- `game_location`: The location of the Olympic Games (country).

- `game_name`: The official name of the Olympic Games (e.g., `Oslo 1952`).

- `game_season`: The season of the Games (`Summer` or `Winter`).

- `game_year`: The year in which the Olympic Games took place.

A.4   olympic_results.csv

The file `olympic_results.csv` includes the following attributes:

- `discipline_title`: The title of the sport or discipline (e.g., `Swimming`, `Athletics`).

- `event_title`: The title of the specific event within the discipline (e.g., `5000m men`, `tempest mixed`).

- `slug_game`: A unique identifier for the specific Olympic Games (e.g., `montreal-1976`).

- `participant_type`: The type of participant (e.g., `Athlete` if individual, `GameTeam` otherwise).

- `medal_type`: The type of medal awarded (`GOLD`, `SILVER`, `BRONZE`, or `None` for participants who did not win a medal).

- `athletes`: A list of athletes associated with the participant (useful for team events).

- `rank_equal`: Whether the rank is shared among multiple participants (`True` or `False`), if applicable.

- `rank_position`: The rank of the participant in the event.

- `country_name`: The name of the country associated with the participant.

- `country_code`: The two-letter country code (ISO 3166-1 alpha-2).

- `country_3_letter_code`: The three-letter country code (ISO 3166-1 alpha-3).

- `athlete_url`: The URL linking to the athlete's profile (if applicable).

- `athlete_full_name`: The full name of the athlete (if applicable).

- `value_unit`: The unit of measurement for performance values (e.g., seconds, meters).

- `value_type`: The type of performance value recorded (e.g., time, distance).

A.5 Dataset Summary

The dataset provides detailed information about:

- Over 120 years of Olympic Games history, spanning Summer and Winter Games.

- Thousands of athletes from different countries, disciplines, and events.

- Medal counts and distribution patterns across various sports and seasons.

- Host cities and countries, offering geographical insights into the Games.