Adnane El Bouhali, Saba Shahsavari, William Liaw

Visualizing Olympic Migration and Medal Trends: A Geopolitical Perspective

Report submitted for the Data Visualization course (CSC_51052_EP) as part of the Master Data AI at Institut Polytechnique de Paris.

Palaiseau

2024

Adnane El Bouhali, Saba Shahsavari, William Liaw

Visualizing Olympic Migration and Medal Trends: A Geopolitical Perspective

Report submitted for the Data Visualization course (CSC_51052_EP) as part of the Master Data AI at Institut Polytechnique de Paris.

Advisor: Emmanuel Pietriga

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.

RÉSUMÉ

Ce rapport présente un projet de visualisation de données axé sur l'analyse des migrations d'athlètes et de la performance des médailles aux Jeux Olympiques de 1986 à 2022. En utilisant le jeu de données Olympic Games Medals 1986-2022, nous explorons les principales tendances telles que les migrations d'athlètes et l'accumulation temporelle de médailles par pays. Les visualisations sont conçues pour mettre en lumière les dynamiques de représentation athlétique et les distributions des médailles au fil du temps. Des visualisations interactives, incluant des flux migratoires, des cartes choroplèthes et des graphiques linéaires, ont été implémentées à l'aide de D3.js afin d'offrir une analyse complète des données. Ce projet souligne l'importance de la visualisation de jeux de données complexes et multidimensionnels pour extraire des informations significatives sur les tendances historiques influençant la performance olympique.

Mots-clés: Jeux Olympiques, migration des athlètes, visualisation de données, D3.js, tendances temporelles, cartes choroplèthes, visualisations interactives.

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. This report emphasizes the role of data visualization in interpreting complex datasets and uncovering meaningful historical and competitive trends.

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.

Once the data was cleaned, various visualizations were created to explore patterns and trends in the dataset. Python, with libraries such as `pandas` and `matplotlib`, was used to generate preliminary exploratory plots. These static visualizations provided insights into data distributions, temporal trends, and relationships between variables. Examples include boxplots for identifying invalid birth years, line charts for visualizing medal accumulation over time, and scatterplots for exploring correlations.

For more advanced and interactive visualizations, JavaScript and the D3.js library were employed. These tools facilitated the creation of dynamic visualizations, such as choropleth maps for visualizing medal distributions across countries and flow diagrams for illustrating athlete migration patterns. The interactive visualizations incorporated features such as tooltips, zooming, and filtering, allowing users to explore the data in depth and tailor their analysis to specific interests.

This combination of Python for initial exploratory analysis and JavaScript with D3.js for dynamic and interactive visualizations ensured a comprehensive approach to analyzing and presenting the data. Together, these tools provided both high-level summaries and detailed explorations, enhancing the overall analytical process and user experience.

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 extreme ages (e.g., over 150 years), were identified in the dataset. These anomalies introduced inaccuracies in calculating athletes' ages at their first Olympic appearance and other related analyses.

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]. Using these references, red shaded regions in the visualizations highlight implausible age values falling outside this range.

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.

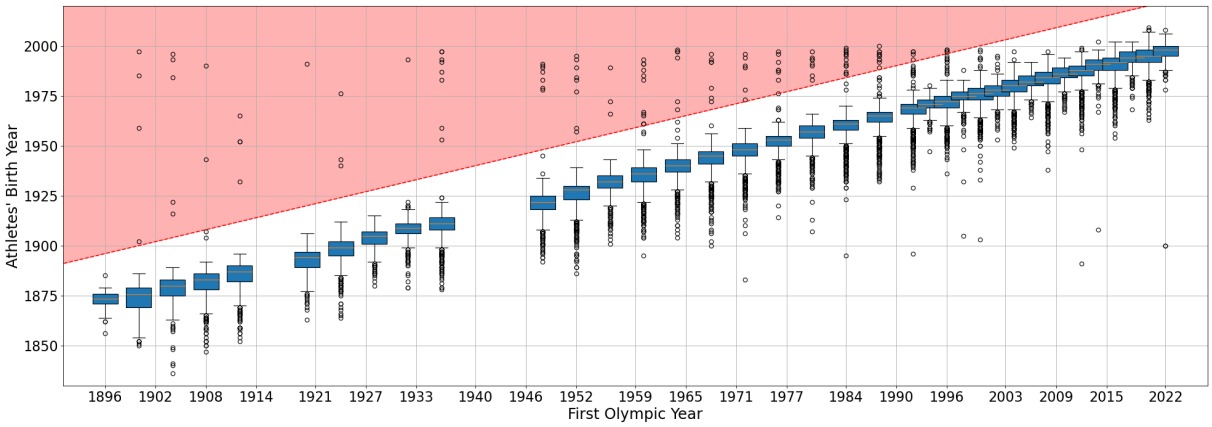

Figure 1: Distribution of Athletes' Birth Years by First Olympic Appearance

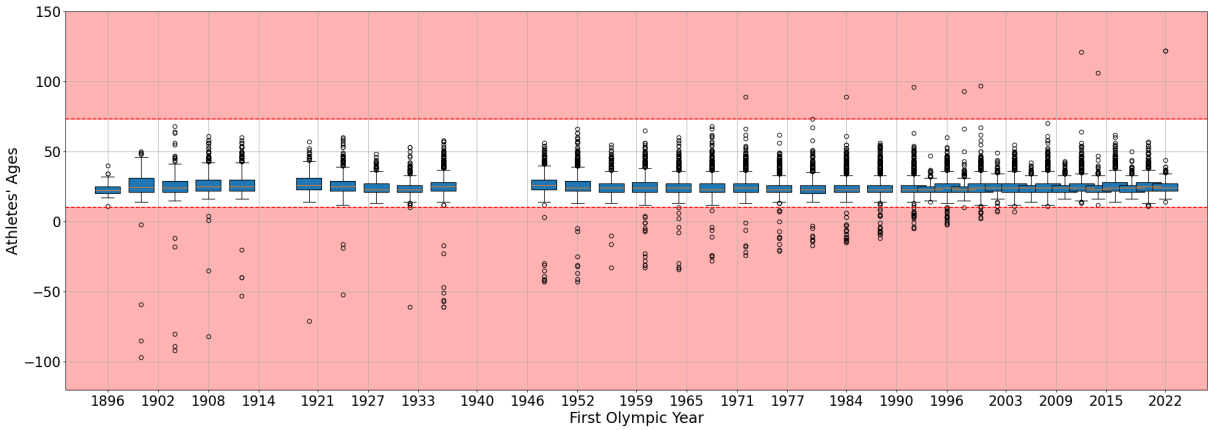

Figure 2: Distribution of Athletes' Ages at Their First Olympic Appearance

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.

By standardizing geographical data, we ensured consistency and improved the

dataset's usability for visualizations involving country-specific analyses.

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.

These gaps were addressed where possible, and records that could not be resolved were excluded from further analysis.

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 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.

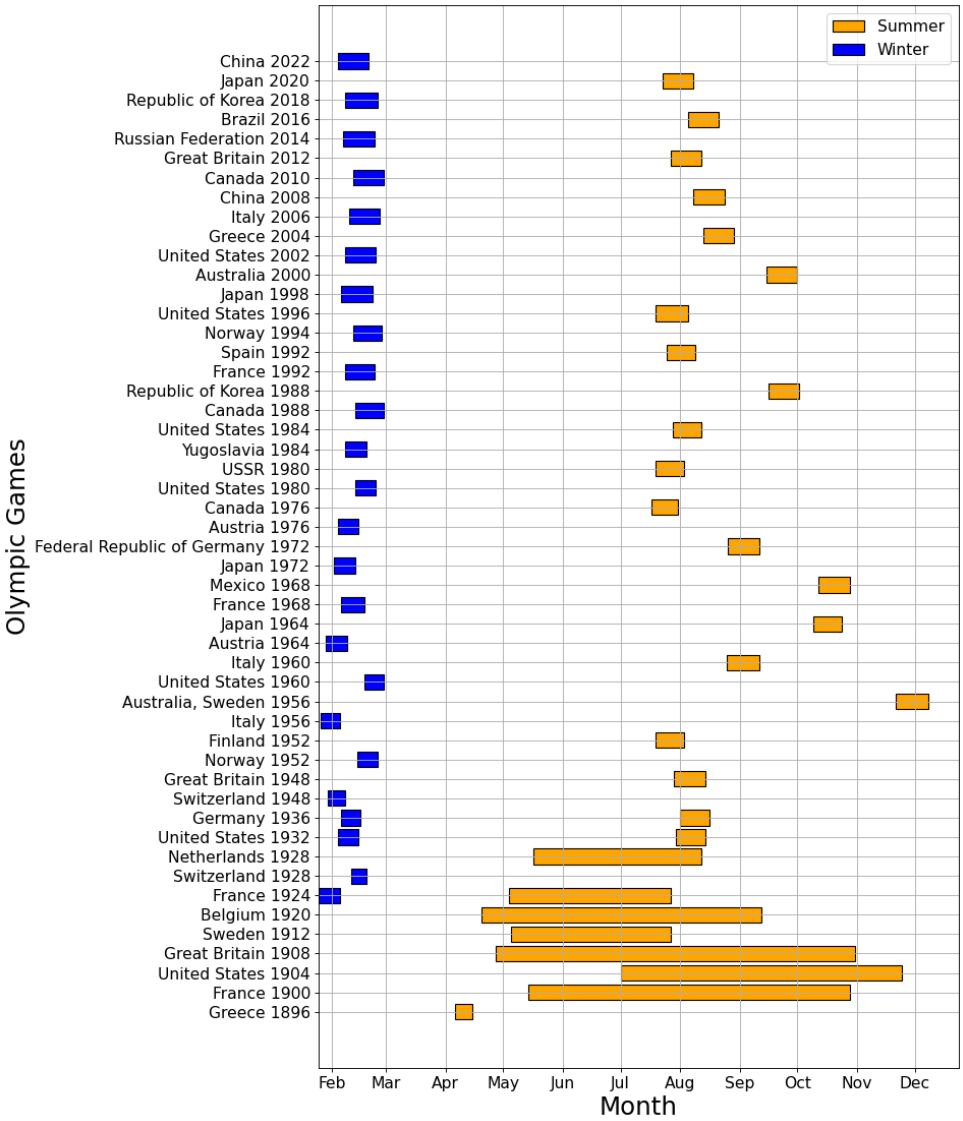

Figure 3: Gantt Chart of Olympic Games Duration

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

PLACEHOLDER ...

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.

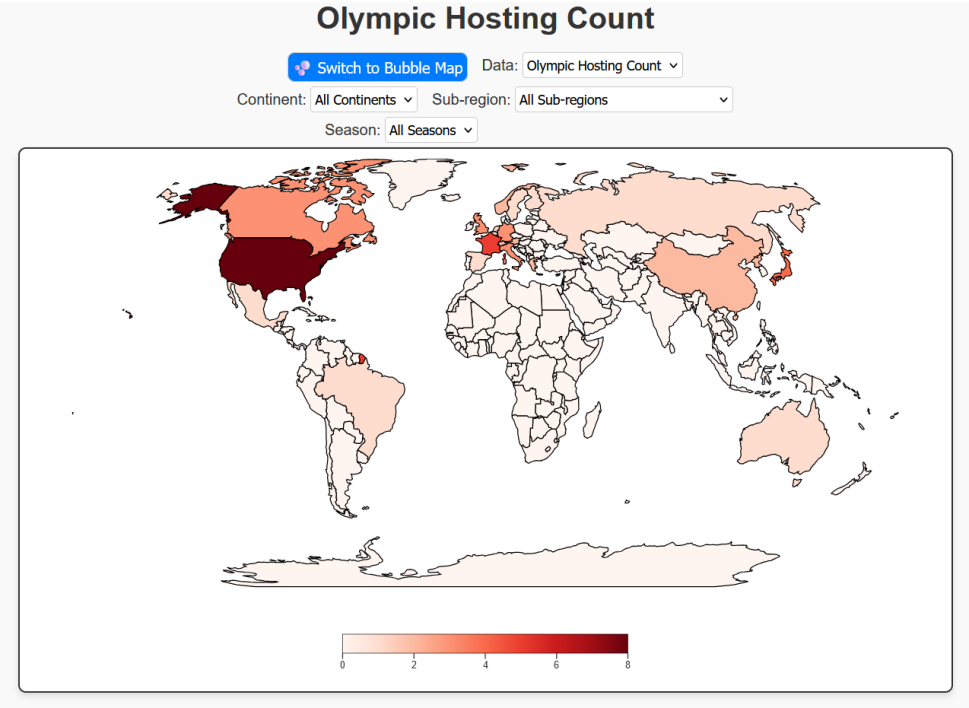

Figure 4: Dynamic Choropleth Map: Visualizing Olympic Hosting Trends

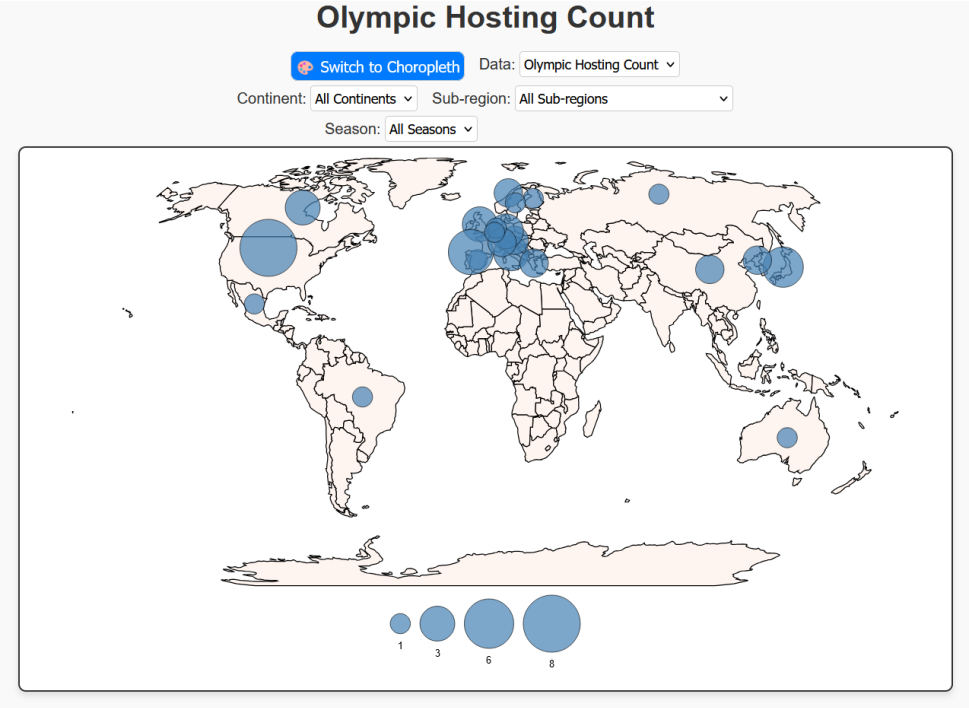

Figure 5: Dynamic Bubble Map: Visualizing Olympic Hosting Trends

4.3.1 Features of the Maps

The maps provide the following functionalities:

- **Visualization Modes:** Users can switch between a choropleth map (Figure 4) and a bubble map (Figure 5) 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.
- **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 num-

ber 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.

4.4 ADNANE PLOTS

PLACEHOLDER

5 CONCLUSION

In this work we have considered ...

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.