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.

Despite its extensive coverage, the raw dataset required significant preprocessing and cleaning to address inconsistencies and prepare it for analysis.

2.1 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.

2.1.1 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.

2.1.2 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 [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.

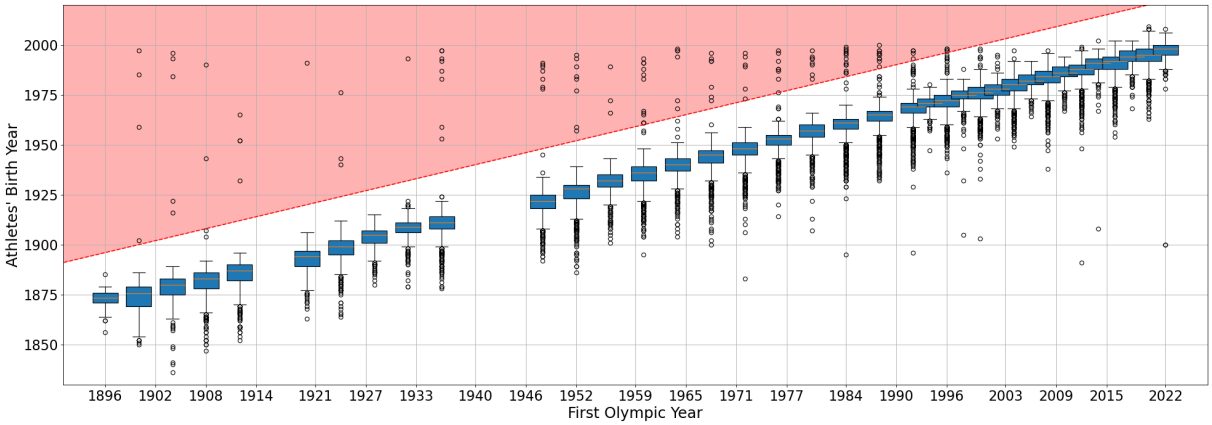

Figure 1: Distribution of Athletes' Birth Years by First Olympic Appearance

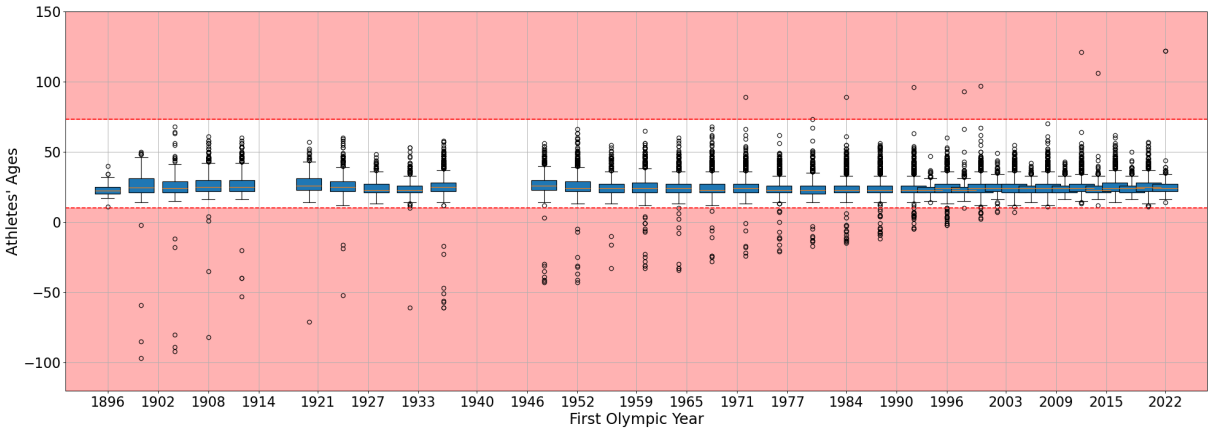

Figure 2: Distribution of Athletes' Ages at Their First Olympic Appearance

2.2 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 [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.

2.3 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.

3 METHODOLOGY

Introduction here...

4 RESULTS

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.

[Contributors, nd] Contributors, O. (n.d.). Nominatim geocoding tool. https://nominatim.org/. Accessed: 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.

[Sports, 2022] Sports, U. T. (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.