The Olympic Dream: A Definitive Database of Every Olympian

Key Words: Definite Olympic Athletes Dataset | Sports Data Analytics | Data Integration | Data Standardisation | Data Enrichment

The Olympics are the world's biggest sports event, and yet no comprehensive dataset of all Olympic athletes has existed until now. Researcher Prof. Dr. Gijsbert Oonk started an initiative to compile a comprehensive and accurate database of Olympic athletes. Before this project, existing Olympic athlete data was often inaccurate and fragmented, halting research on this topic. Advancement in sports-related research requires a reliable and comprehensive dataset to analyse trends and correlations related to for example athletes' backgrounds, performance, and the socio-economic contexts of their home countries.

Erasmus Data Collaboratory | House of AI support

Professor Oonk sought the assistance of the EDC data lab team to meticulously consolidate, clean, and enrich Olympic athletes' data from multiple sources to create a single, authoritative registry. The data analysts from EDC's data lab used a multi-faceted approach for the necessary data integration and cleaning. The first step involved matching and merging the data. Athletes and their events were identified and matched across different sources using unique IDs, names, and event details. Unmatched records were manually reviewed to ensure accuracy. This was followed by data correction and standardisation, which was needed for mapping the country abbreviations and cross-referencing them against the official IOC (International Olympic Committee) list. Lastly, to achieve an extensive dataset, the data was significantly enriched with information from various specialised sources. Athlete-specific details, event data, geographical coordinates, and economic data were sourced and added to the set. EDC provided the support and the framework to work through the data and overcome the limitations of disparate datasets, which suffered from inaccuracies, missing information, and a lack of standardisation.

Impact

The methodologies and the resulting large-scale dataset created for this project by EDC's data lab team have wide-reaching applications across multiple disciplines. Comprising over 300,000 rows of athlete-event combinations, this definitive dataset serves as a unique and authoritative resource for researchers, journalists, and sports enthusiasts alike. The comprehensive nature of the data allows for in-depth, "big data" analysis across various disciplines, from sports science and history to sociology and economics. Potential avenues for future research include exploring long-term trends in athlete demographics, examining the relationship between economic development and Olympic performance, and mapping the global distribution of athletic talent. By building on EDC's advanced data management capabilities, this project establishes a strong foundation for future socio-economic and historical research related to the Olympic movement.

Stakeholders: Gijsbert Oonk | Erasmus School of History, Culture, and Communication (ESHCC) | Erasmus University Rotterdam, International Olympic Committee (IOC)
Specific EDC expertise used: Matching and Merging Data | Data Integration | Data Standardisation | Data Enrichment |
Tech/Tools Used: Python, Pandas, Excel

Testimonial by researcher: *"Working with the EDC team was an outstanding experience. They were friendly, professional, and far beyond what I expected. Their support was instrumental in helping us validate, clean, and enrich Olympic athlete data drawn from multiple, often inconsistent sources. The analysts at EDC's data lab approached the challenge with impressive precision and care. From intelligently matching and merging records across databases to meticulously correcting and standardising country codes using official IOC references, every step was handled with expertise.*

What truly stood out was their ability to enrich the dataset and meaningfully integrating athlete profiles with geographic and economic data from specialised sources. Thanks to their multi-faceted and collaborative approach, we now have a single, authoritative dataset that we can confidently rely on. EDC not only provided the technical framework to navigate the complexity of disparate datasets but did so with a level of support and friendliness that made the entire process smooth and enjoyable.

With this updated dataset we are ready for the next steps!"