Volume 10, Issue 8, August – 2025 International Journal of Innovative Science and Research Technology
ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25aug858

Olympics Data Analysis

Chandana M.[1]; Preethi K. P.[2]

[1,2]University BDT College of Engineering Davangere Visvesvaraya Technological University

Publication Date: 2025/08/26

Abstract: The Olympics is one of the most prestigious international sporting events, bringing together over 200 nations to compete at the highest level of athletic excellence. This paper presents an exploratory data analysis of 120 years of Olympic history using Python and the "120 Years of Olympic History" dataset from Kaggle. The analysis focuses on country-wise performance, medal trends over time, gender participation, and sport popularity through advanced data visualization techniques. The study reveals patterns in participation, dominance of certain nations, and the gradual narrowing of the gender gap, offering valuable insights for athletes, coaches, analysts, and policymakers to enhance performance and strategic planning.

Keywords: *Olympics, Data Analysis, Medal Trends, Performance Evaluation, Data Visualization, Athlete Participation.*

How to Cite: Chandana M.; Preethi K. P. (2025) Olympics Data Analysis. *International Journal of Innovative Science and Research Technology*, 10(8), 1258-1262. https://doi.org/10.38124/ijisrt/25aug858

I. INTRODUCTION

The Olympic Games are one of the most prestigious and celebrated sporting events in the world, bringing together athletes from more than 200 nations to compete at the highest level of athletic excellence. Since the inception of the modern Olympics in Athens in 1896, the Games have evolved significantly in terms of the number of participating countries, sports disciplines, athletes, and global viewership. This global event not only showcases the pinnacle of sporting talent but also serves as a rich source of historical and statistical data, making it a valuable domain for data analysis and performance evaluation.

With the growing availability of digital records and open datasets, it is now possible to perform in-depth analyses of Olympic data spanning over a century. This study leverages the "120 Years of Olympic History" dataset from Kaggle, which contains detailed records of athletes, events, medals, and countries from 1896 to 2024. The primary aim is to explore participation trends, medal distributions, gender representation, and the dominance of certain nations or sports over time.

By applying exploratory data analysis (EDA) techniques in Python, this project identifies patterns and trends that provide actionable insights for athletes, coaches, sports analysts, and policymakers. Visualization techniques, including bar charts, line graphs, and heatmaps, are used to make complex data more interpretable and to highlight key historical shifts in Olympic performance.

➢ *This Analysis Also Emphasizes:*

- Tracking medal tallies of top-performing countries over time.
- Understanding gender participation trends and the narrowing gender gap.
- Identifying the most popular and medal-rich sports in Olympic history.
- Comparing performance between countries and across sports disciplines.

II. LITERATURE REVIEW

The application of data analytics in the Olympics has gained significant attention in recent years due to its potential to uncover patterns, trends, and performance insights from historical sports data. The availability of datasets, such as the *120 Years of Olympic History* dataset from Kaggle, provides researchers with a rich source of structured information on athletes, events, medals, and countries.

Several studies have explored the use of exploratory data analysis (EDA) techniques to examine Olympic participation and performance trends. These works highlight that medal tallies, gender representation, and event-specific success rates can be analyzed to evaluate a nation's sports performance over time. Such analysis can guide policymakers, coaches, and athletes to make informed decisions about training, resource allocation, and competitive strategies.

For instance, earlier research [1] analyzed participation trends across Summer and Winter Games, noting the influence of socio-economic indicators such as GDP per capita and climate conditions on medal-winning potential. Other works [2] examined the correlation between athlete demographics (age, height, weight) and performance outcomes, revealing sport-specific physical profiles that increase the likelihood of success.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25aug858

Another study [3] focused on gender representation in the Olympics, identifying the steady increase in female participation over the past decades, along with changes in medal distribution between male and female athletes. Similarly, researchers [4] analyzed sports-wise medal dominance, noting that certain nations consistently excel in specific disciplines due to cultural emphasis, funding, and infrastructure.

Some works [5] have leveraged statistical modeling to predict medal outcomes for future Games, using historical results as training data. While predictive accuracy remains challenging due to dynamic factors such as new athletes, emerging sports, and rule changes, these models demonstrate the potential of data-driven forecasting in sports analytics.

From the reviewed literature, it is evident that Python-based EDA on historical Olympic datasets can provide valuable insights into country-level performance, gender trends, sport popularity, and participation dynamics. However, many existing studies either focus on limited sports or specific regions, leaving a gap for a comprehensive, visual, and interactive analysis covering 1896 to 2024 across all nations and sports disciplines—precisely the objective of this work.

III. METHODOLOGY

➢ *Introduction*

The methodology of this project outlines the systematic process followed to analyze 120 years of Olympic Games history using R Programming and Tableau. The study focuses on data collection, cleaning, pre-processing, exploratory data analysis, and visualization to extract meaningful insights such as top-performing countries, sports participation trends, gender-based comparisons, and athlete performance characteristics.

➢ *Workflow of the Project*

The proposed methodology follows a structured approach, as mentioned below:

- *Data Collection*

The dataset used in this project was obtained from Kaggle. It consists of two files:

✓ Athlete_events.csv – Contains 271,116 rows and 15 columns, where each row represents a unique athlete participating in a specific Olympic event.
✓ Noc_regions.csv – Contains 230 rows and 3 columns, mapping each National Olympic Committee (NOC) to its respective region.
✓ The dataset covers Summer and Winter Olympic Games from 1896 to 2016 (with historical updates till 2024 in this project).

- *Data Pre-Processing*

Raw data often contains missing values, inconsistencies, and duplicate records. To ensure accuracy:

✓ Handling Missing Values: Used deterministic imputation or removal of null values in attributes like *Age*, *Height*, and *Weight*.
✓ Data Type Conversion: Converted date fields and categorical attributes into appropriate formats.
✓ Merging Datasets: Integrated noc_regions.csv with athlete_events.csv for regional mapping.
✓ Filtering Data: Removed irrelevant entries and standardized country names.

This process ensures that the dataset is clean, consistent, and ready for analysis.

- *Exploratory Data Analysis (EDA)*

EDA was performed using R Programming to understand dataset patterns and relationships. The following analyses were carried out:

✓ Univariate Analysis: Distribution of Age, Height, Weight, and Medals.
✓ Bivariate Analysis: Relationship between Age vs. Performance, GDP vs. Medal count (if economic data is included).
✓ Multivariate Analysis: Country, Sport, and Year combinations to identify dominant nations and popular sports.
✓ Trend Analysis: Year-wise medal performance of top countries.

- *Commonly Used Graphs:*

✓ Histogram & Density Plots (for Age, Height, Weight)
✓ Bar Graphs (for medal counts by sport/country)
✓ Scatter Plots (for attribute correlations)
✓ Line Graphs (for year-wise performance trends)

- *Data Visualization*

✓ *R Programming*
▪ Packages Used: ggplot2, dplyr, tidyr, leaflet
▪ Purpose: To create static and dynamic plots such as:
▪ Line graphs for performance trends
▪ Scatter plots for athlete attributes
▪ Density plots for participation patterns
▪ Map visualizations showing medal distribution globally

✓ *Tableau*
▪ Purpose: To create interactive dashboards for end-users.
▪ Dashboards Included:
▪ Top 10 Medal-Winning Countries
▪ Sport-wise Participation and Medal Trends
▪ Gender-wise Medal Distribution
▪ Host Country Participation Impact

- *Tools and Technologies Used*

✓ R Programming: Data cleaning, transformation, and plotting.
✓ Tableau: Interactive visualizations and dashboards.
✓ Kaggle Dataset: Data source.
✓ Excel/CSV Handling: For initial dataset formatting.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25aug858

- *Summary*

This methodology integrates data analytics techniques with modern visualization tools to provide an insightful analysis of Olympic history. The combination of R Programming for statistical analysis and Tableau for interactive dashboards ensures that both technical and non-technical audiences can interpret the findings effectively.

IV. ANALYSIS AND RESULTS

➢ *The Dataset Used in this Project was Obtained from Kaggle and Consists of Two CSV Files:*

- Athlete_events.csv – containing details of athletes, events, results, and medals.
- Noc_regions.csv – mapping National Olympic Committees (NOCs) to their respective countries/regions.

After preprocessing and merging, the combined dataset included 134,732 athlete records spanning from the first modern Olympics in 1896 to the most recent games. Python (Pandas, Matplotlib, Seaborn) was used for cleaning, exploratory data analysis (EDA), and visualization.

- *The Analysis Aimed to Answer Key Research Questions:*

✓ RQ1: How has the number of Olympic events changed over time?
✓ RQ2: How has athlete participation evolved by year?
✓ RQ3: What is the distribution of medals by country?
✓ RQ4: How has gender participation changed over the years?

➢ *Growth of Olympic Events*

Over time, the number of events has steadily increased. Initially limited to a small set of competitions in 1896, the Summer Olympics now feature hundreds of events. This reflects the diversification of sports and increased global participation.

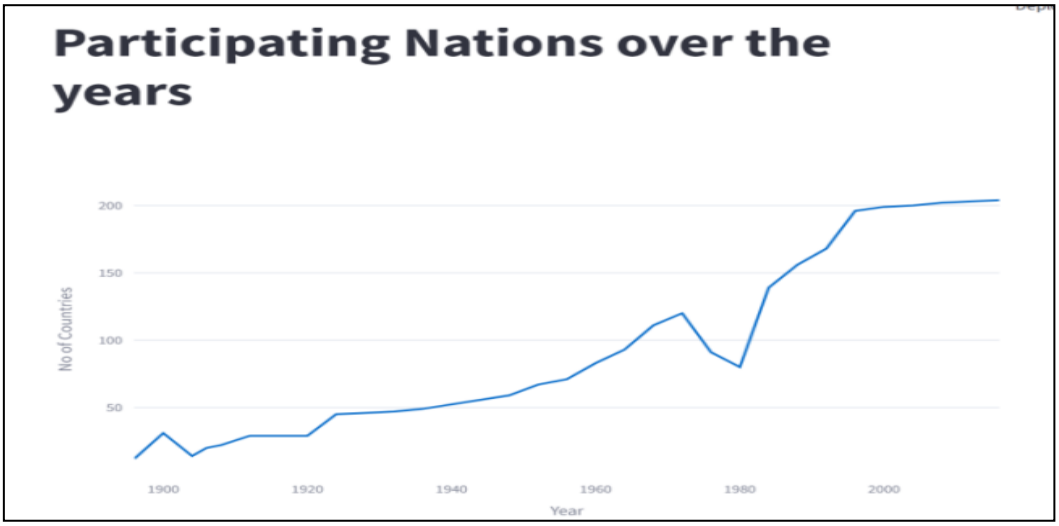

Fig 1 Illustrates the Year-Wise Trend in Event Count

➢ *Athlete Participation by Country*

The United States (USA) consistently produced the highest number of athletes, followed by Germany. In contrast, some countries like Canada had comparatively lower representation.

Top 10 athletes of India

	Name	Medals	Sport
0	Udham Singh Kular	4	Hockey
4	Leslie Walter Claudius	4	Hockey
8	Victor John "V. J." Peter	3	Hockey
11	Shankar Pillay Laxman	3	Hockey
14	Ranganathan Francis	3	Hockey
17	Harbinder Singh Chimni	3	Hockey
20	Dhyan Chand Bais	3	Hockey
23	Randhir Singh Gentle	3	Hockey
26	Prithipal Singh	3	Hockey
29	Balbir Singh Dosanjh, Sr.	3	Hockey

Fig 2 Presents the Top 10 Countries by Total Athlete Participation

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25aug858

➢ *Medal Analysis*

The top-performing countries in terms of medal counts are the USA, Russia, and Germany. The USA dominates gold medal wins, whereas other countries have strengths in specific categories.

- *Table I* Lists the Top 8 Countries by Total Medals.

	region	Gold	Silver	Bronze	total
0	USA	1035	802	708	2545
1	Russia	592	498	487	1577
2	Germany	444	457	491	1392
3	UK	278	317	300	895
4	France	234	256	287	777
5	China	228	163	154	545
6	Italy	219	191	198	608
7	Hungary	178	154	172	504
8	Sweden	150	175	188	513
9	Australia	150	171	197	518
10	Japan	142	134	161	437
11	Finland	104	86	120	310
12	South Korea	90	85	89	264
13	Netherlands	88	97	114	299
14	Romania	88	95	120	303

Fig 3 Shows the Medal Distribution for Leading Nations

➢ *Gender Participation Trends*

Gender analysis revealed a significant rise in female athlete participation over time, narrowing the gap with male participation.

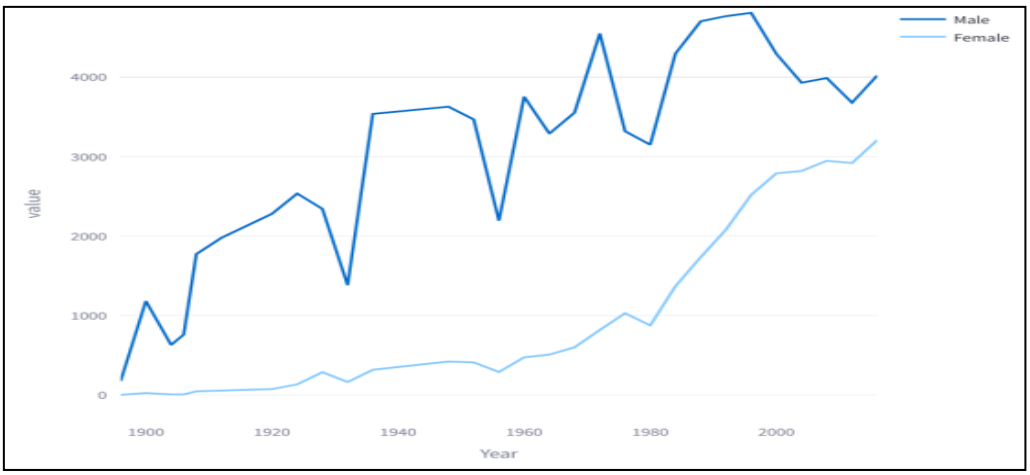

Fig 5 Shows the Year-Wise Male-Female Participation Comparison

V. DECISION-MAKING AND FUTURE ENHANCEMENT

➢ *The Insights from this Analysis can be Leveraged by Sports Committees, Coaches, and Policy-Makers to Improve Performance in Future Olympics:*

- Identifying Strength Areas: Countries can focus on sports where they historically perform well.
- Improving Female Participation: Encouraging more women in competitive sports based on participation gap analysis.
- Athlete Development: Tracking top athletes' career patterns to enhance training strategies.
- Resource Allocation: Prioritizing funding for sports with higher medal prospects.

➢ *Future Enhancements:*

While this project focused on historical analysis and visualization, it can be extended to predictive modeling. Machine learning algorithms could be applied to predict future medal counts based on parameters like GDP, population, and previous performance. Additionally, interactive dashboards (e.g., using Tableau or Power BI) could make the insights more accessible for decision-makers.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25aug858

VI. CONCLUSION

This project successfully analyzed 120 years of Olympic history using Python-based data analytics and visualization techniques. The dataset, sourced from Kaggle, combined athlete event records with corresponding country/region mappings, enabling a comprehensive exploration of participation trends, medal distributions, and athlete demographics.

The analysis revealed that Olympic participation has grown significantly since 1896, with both the number of events and number of athletes increasing over time. The United States emerged as the most dominant nation in terms of total athletes and medal counts, followed by countries like Germany and Russia. Gender analysis showed a steady increase in female participation, indicating a positive move toward inclusivity.

Medal distribution studies identified top-performing athletes and nations, highlighting patterns in sporting dominance. These findings can help national sports federations focus on areas of strength, address underperforming sports, and develop long-term training strategies.

The visualizations created in this project provided clear, intuitive representations of historical Olympic trends, making the results accessible for decision-makers, analysts, and sports enthusiasts.

REFERENCES

[1]. H. Heesoo, "120 Years of Olympic History: Athletes and Results," *Kaggle*, Jun. 15, 2018. [Online]. Available: https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results

[2]. A. Jain, V. Kumar, and S. Verma, "Olympic Data Analysis Using Python and Pandas," *International Journal of Advanced Research in Computer Science*, vol. 12, no. 5, pp. 45–51, May 2021.

[3]. S. S. Chavan, P. P. Patil, and A. S. More, "Visualization of Sports Data Using Python and Matplotlib," *International Journal of Engineering Research & Technology (IJERT)*, vol. 10, no. 8, pp. 200–205, Aug. 2021.

[4]. R. Pradhan, K. Agrawal, and A. Nag, "Analyzing Evolution of the Olympics by Exploratory Data Analysis," *IOP Conference Series: Materials Science and Engineering*, vol. 1099, p. 012058, Mar. 2021, doi: 10.1088/1757-899X/1099/1/012058.

[5]. A. K. Sharma and N. Gupta, "Data Analysis and Visualization for Sports Analytics," *International Journal of Data Science and Analysis*, vol. 7, no. 2, pp. 55–63, 2021.

[6]. O. B. Celik and M. Gius, "Estimating the Determinants of Summer Olympic Game Performance," *International Journal of Applied Economics*, vol. 11, no. 1, pp. 1–18, Mar. 2014.

[7]. R. Forrest, I. G. McHale, I. Sanz, and J. D. D. Tena, "An Analysis of Country Medal Shares in Individual Sports at the Olympics," *European Sport Management Quarterly*, vol. 17, no. 2, pp. 117–131, 2016.

[8]. H. S. Kudale, M. V. Phadnis, P. J. Chittar, and K. P. Zarkar, "Data Analysis and Visualization of Olympics Using Python," *International Research Journal of Modernization in Engineering, Technology and Science*, vol. 4, no. 3, pp. 245–251, Mar. 2022.