Case Study: Performance Analysis throughout the History in Summer and Winter Olympics

Prabodha Abeynayake
Department of Computer Engineering
University of Sri Jayewardenepura
Rathmalana, Sri Lanka
pyabeynayake@gmail.com

Udaya Wijenayake
Department of Computer Engineering
University of Sri Jayewardenepura
Rathmalana, Sri Lanka
udayaw@sjp.ac.lk

***Abstract*—The Olympics is the world's foremost multisport international event that occurs every four years in 2 seasonal editions which are the Summer and Winter Olympics respectively. With more than 200 nations participating in hundreds of events, it is a matter of prestige for any nation and the dream of every sportsperson to win a medal in the Olympics. Despite the massive hard work of the sportspersons, many nations fail to win medals whereas some nations are able to grab most of the medals under the names of their nations. Therefore, an analysis of the history of the Olympics is needed to understand the performances of nations, athletes and events to observe certain patterns occurring throughout the history of the Olympics, in order to help the nations to improve themselves. The primary objective of this paper is to analyze the historical data of both the Winter and Summer Olympics under 14 different factors using Exploratory Data Analysis (EDA) techniques and statistical methods to evaluate the performance of nations and athletes as well as the historical evolution of Olympics and correlation of medal win of sports persons with external variables such as GDP and population. The analyses provide accurate insights into the performance of nations in Olympics throughout the history and help sportspersons to analyze their own and the competitor's performances. In this paper, the data visualization aspect of EDA is used to provide a statistical view of factors that led to the evolution of the Olympic games as well as the performance of nations and sportspersons and helps countries with average and poor performances to improve themselves in upcoming Olympic editions.**

***Index Terms*—Data Analysis, Data Visualization, Olympics, Performance Analysis, Statistical Correlation**

I. Introduction

The Olympics is considered the world's only truly global, multisport celebratory international event. It provides a common global platform for all sportspersons and nations to fairly compete with each other and showcase their skills and talents. Winning a medal in the Olympics is a lifelong dream of most sportspersons as well as a matter of prestige for any nation. Modern Olympic games are inspired by the ancient Greek Olympic games held in Olympia, Greece from 776 BC to 393 AD. [1] Olympics occur every four years and have two different editions differed by the season, which are the Summer and Winter Olympics respectively.

The 1896 Summer Olympics in Athens, Greece marked the inception of the modern international Olympic Games, featuring 241 male athletes from 14 nations in 43 events [2]. The Winter Olympics, which began in 1924 in Chamonix, France, had 260 male athletes from 16 countries competing in 16 events [3]. Over the years, both editions have evolved significantly, enhancing diversity, increasing events, nations, and introducing female participation. The 2020 Tokyo Summer Olympics was graced with 11,420 athletes from 206 nations in 339 events [4], while the 2018 Pyeongchang Winter Olympics featured 2,833 athletes from 92 nations in 102 events [5]. These figures highlight the continuous evolution of the Olympics.

When observing the Olympic Games, discernible patterns emerge, including recurring countries in the medal tally leaderboards, consistent underperformance of third-world nations, and sustained excellence by specific countries and athletes in particular sports. Analyzing these patterns offers insights into the Olympics' evolution and the performance of nations and athletes. Such analyses serve as performance indicators, aiding nations and athletes in their quest for improvement in future Olympic editions. The attainment of Olympic medals reflects the dedication and hard work of athletes. However, external factors, such as a nation's investment in sports, represented by its GDP, can influence medal outcomes. Statistical correlation analysis of the Olympics can gauge the strength of the relationship between external variables and medal achievements.

The main objective of this study is to analyze the historical data of both the Winter and Summer Olympics using Exploratory Data Analysis (EDA) techniques and statistical methods to evaluate the performance of nations and athletes as well as the historical evolution of the Olympics and correlation of medal win of sports persons with external variables such as GDP and population. The analysis includes the visual representation of changes in trends in certain aspects of the Olympics to provide accurate insights for the nations and sportspersons to improve their future performances.

Section II contains the related works in the domain and section III discusses the overall methodology of the study. Section IV discusses the results and section V includes a concise conclusion of the study.

II. Related Work

Interpreting and analyzing data which highlights the essence of data analysis, constitutes a critical aspect of big data

analysis. Within the domain, extensive research has been conducted on the analysis of Olympic games such as statistical analysis, performance analysis, predictive analysis as well as EDA. Predictive analytics is an analytical approach that leverages current and historical data to forecast future events. [6] Predictive analysis has been a popular approach among researchers in the Olympic domain.

A nation's Olympic performance could be predicted using past performance data. One such study utilized logistic regression and the country's highest score from previous Olympic participations to predict their chances of winning gold in 2016. [7] Predictive analytics has been utilized to forecast the probability of a medal-winning athlete repeating their success in the upcoming Olympics. The study involved interpreting the probability of a medal win by an athlete in an upcoming Olympics, given the fact that they have won medal/medals in previous editions of the Olympics. [8]

Machine learning techniques, specifically heuristic prediction, have been employed to predict a nation in view of the Olympic awards using data from the 2012 London Summer Olympics [9]. In addition to predictive analysis, a study explored statistical differences in age and swimming styles during the 2016 Summer Olympics [10]. Another study delved into the underlying legacies and factors influencing the hosting of the 2016 Rio Olympics, aiming to provide insights into successful Olympic hosting [11].

EDA serves as a crucial and widely employed method for examining the evolution of the Olympics. This analytical approach involves the visual representation of extensive data through graphs, charts, and visual formats, facilitating a deeper understanding of complex datasets. EDA enables the identification of patterns, trends, and outliers, fostering the generation of hypotheses for further analysis [12].

Parveen et al. [13] utilized EDA techniques on data spanning the 1896-2012 Summer Olympics, effectively comparing the overall performance and contributions of participating countries. The focus was on assessing the growth in each country's Olympic performance over time, analyzing factors such as total number of medals won, individual country performance, and comparisons between participants .

Pradhan et al. [14] analyzed the evolution of the Olympics by EDA using R on 1896-2016 Summer Olympics data. This analysis focused on factors that lead to the evolution of the Olympic Games and the improvement of countries/players over time in a visual format such as the density of players in the Olympic Games based on Age, the total number of medals won by various countries in Olympic games, male and female participants in Olympic games in different years and heights and weights of the players who have won the maximum number of medals.

Apart from these approaches in EDA, Rathke et al. [15] adopted a distinct approach to assess a country's success in the Olympics through a combination of efficiency analysis and the significance of sports in its society using stochastic frontier analysis.

III. Methodology

This paper aims to analyze the historical data of both the Winter and Summer Olympics using EDA techniques and statistical methods to evaluate the performance of nations and athletes as well as observe the historical evolution of the Olympics and the correlation of medal wins of sports persons with external variables. To identify these factors and conduct a comparative study among them, a step-by-step approach has been employed.

A. Selection of Dataset

We have used data from various datasets to facilitate the requirements of the different analyses. We have majorly used 2 datasets that contained a satisfactory amount of volume and variety in data.

1st dataset "120 years of Olympic history: athletes and results" [16] contained a 271,116 rows and 15 columns of data on individual sportspersons that participated in any Olympic event in the Summer/Winter Olympics from 1896-2016. It consisted of details such as age, height, weight, gender, sportsperson's nation, National Olympic Committee (NOC) region code and medal-winning status (No Medal/ Gold/Silver/Bronze). The data were employed to analyze the player performance, and historical evolution of the Olympics as well as facilitate comparative studies between nations' performances.

2nd dataset, "Olympic Sports and Medals, 1896-2014" [17] consisted of a dataset of GDP and population estimates of nations with their respective IOC country codes. This dataset was employed to assess the correlation of medal wins of sports persons with external variables.

B. Data Preprocess

The next step involves data preprocessing, which entails converting raw data obtained from data sources into useful data by meticulously examining for inaccuracies, incompleteness, and null values. The datasets often contained null values in fields such as height and weight of sportspersons. Some other fields also occasionally employed null values. The null values needed to be eliminated by replacing the valid data as their presence led to erroneous results and greatly affected visualized graphical outputs.

Three methods were employed for data pre-processing. First, we attempted to fill in the null values using manual methods by searching for the correct values in available sources related to the Olympics. The data fields subjected to this method were NOC region, GDP, population, country and heights weights of famous sportspersons.

Following manual preprocessing, the remaining null values in numerical fields were subjected to deterministic imputation. This method involves replacing missing values based on other observations within the same column. For numerical fields such as heights and weights of sportspersons, basic numeric imputation was applied, replacing null values with the mean for a fair assumption. Additionally, for categorical data which

was the team name, Hot deck imputation was utilized, replacing null values with similar values from other records within the same column.

C. Exploratory Data Analysis

Following the data pre-processing phase, our study delved into comprehensive data analysis using EDA across 14 distinct factors covering a broad spectrum of dimensions and exploring facets that were not previously addressed in other relevant studies. This analytical approach aimed to assess the performance of nations and athletes while providing insights into the historical evolution of the Olympics. The 14 factors explored encompassed Men vs. Women participation over the years, Overall performance of all the nations throughout the history of the Olympics, Countries' performances over the years, Best-performed fields of sports for a particular country, Comparison of performance between the countries throughout history, Overall statistics of Olympics (Number of Nations, Hosts, Sports, Events), Nations' participation over the years throughout the history, Events of the games over the years, Most Successful sportspersons of overall Olympics, Most Successful sportspersons per sport, Sportspersons' optimal weight and height per sport for winning Gold, Silver and Bronze Medals, Sportspersons' age distribution with respect to sport, Evolution of each event of a particular sport throughout the year and Top athletes of a country.

At the end of the analyses, diverse graphical representations, including bar graphs, box plots, line graphs, heat maps, and scatter plots, were generated. This visual exploration facilitated flexible analysis and comparison of data, enhancing the intuitive understanding of information. The examination of various plots and visualizations allowed for deeper insights and meaningful conclusions to be drawn from the analysis.

Python, along with its libraries such as pandas, numpy, and scipy were employed for data analysis. The graphical representations of the analysis results were created using Matplotlib, Seaborn, and Plotly libraries.

D. Correlation Analysis

Olympic medal achievements result from years of athletes' dedicated efforts, yet the impact of external factors on these victories cannot be ignored. Factors such as a nation's investment in sports, represented by its GDP, play a significant role. Statistical correlation analyses provide insights into the relationship between external variables and medal wins.

In this study, we employed two statistical correlation methods, Spearman and Pearson, to assess the correlation between medal wins and a country's GDP and population. The Pearson method, denoted by Equation 1, measures linear correlation between continuous variables. Spearman, a non-parametric measure, evaluates monotonic relationships. We selected these methods due to their appropriateness for analyzing the specific nature of Olympic data, capturing both linear and non-linear correlations effectively in a rigorous scientific manner.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{1}$$

where:

$\bar{x}$: mean of x

$\bar{y}$: mean of y

x_i : individual data points in x

y_i : individual data points in y

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{2}$$

where:

d_i : the difference between the ranks of corresponding pairs

n : number of pairs of observations

The two methods have been effectively used to evaluate the relationship between external variables, GDP and population to the medal win in the Olympics.

E. Result Analysis

The outcomes of EDA and visualizations underwent rigorous analysis to extract meaningful insights and draw significant conclusions. This entailed a meticulous examination of results, pattern identification through graphical visualizations, and interpretation of numerical findings, providing a nuanced understanding of underlying factors.

IV. Results

We have conducted a thorough analysis of several factors encompassed in the datasets and have generated multiple graphs that vividly illustrate the shifts in Olympic Games trends across the years. Below are some of the key findings resulting from our research.

A. Nations' participation over the years throughout the history

Figure 1 illustrates a consistent rise in the number of participating nations in both the summer and winter Olympic games, with occasional fluctuations. The data suggests a near saturation point for summer Olympic participation, exemplified by 204 nations in the 2016 edition, encompassing a significant majority of the 206 Olympic nations globally by the 2020 Tokyo Olympics. In contrast, the Winter Olympics exhibit a continuous growth trend, hinting at the potential inclusion of more nations in forthcoming editions.

The graph also highlights significant historical events, notably the sharp decline in participating nations during the

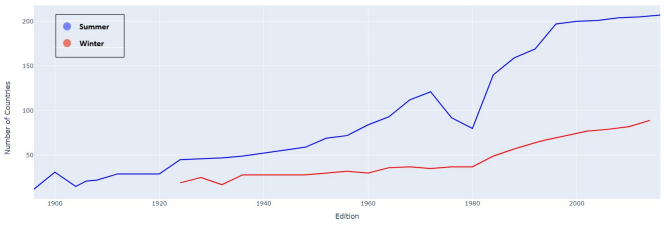

Fig. 1. Nations' Participation throughout the History of Summer and Winter Olympics

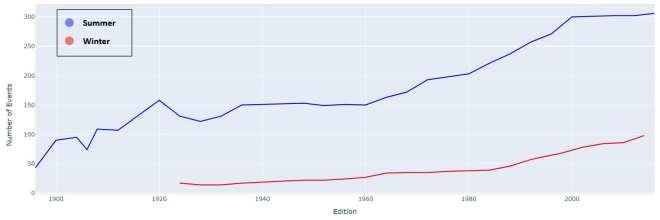

Fig. 2. Number of events throughout the history of summer and winter olympics

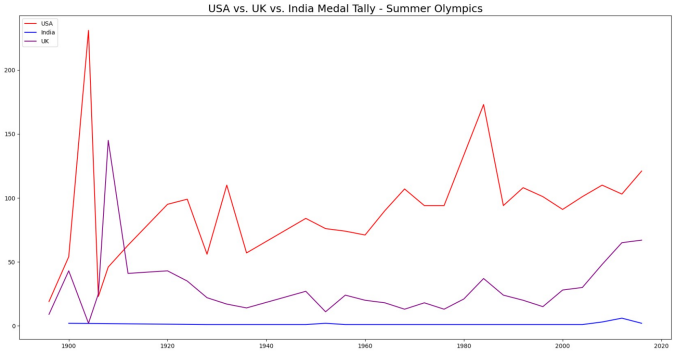

Fig. 3. USA vs. UK vs. India comparison based on total medals won per each edition of Summer Olympics

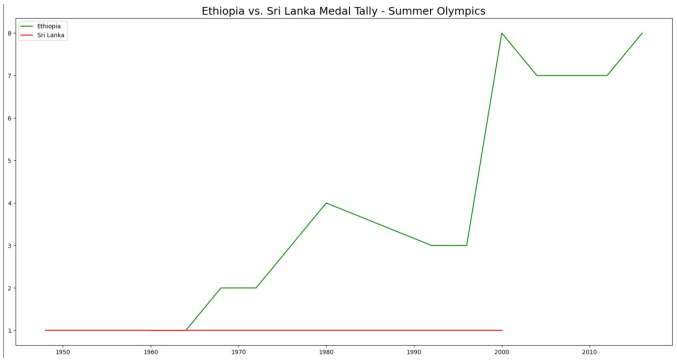

Fig. 4. Ethiopia vs. Sri Lanka comparison based on total medals won per each edition of Summer Olympics

TABLE I
PERFORMANCE OF EACH CLASSIFICATION METHOD BASED ON URL CLASS

Variable	Correlation Coefficient	
	Pearson Method	*Spearman Method*
Population	0.21	0.42
GDP	0.44	0.46

1980 Moscow Olympics. This downturn resulted from a widespread boycott triggered by the Soviet Union's invasion of Afghanistan in 1979, underscoring the Olympics' vulnerability to geopolitical events.

B. Events throughout the History

Based on Figure 2, which illustrates the total number of events in every sport for both the summer and winter editions of the Olympic Games, our analysis reveals that the number of events has gradually increased over time in both seasons of the Olympics. However, the number of events in the Summer Olympics has reached a saturation point in recent consecutive editions, while the number of events in the Winter Olympics has continued to grow gradually over the past few editions. As a result, we anticipate that the overall number of events in the Winter Olympics to grow more compared to the Summer Olympics in upcoming editions.

C. Sport Evolution throughout the History

We generated distinct heatmaps illustrating the evolution of sports in the Summer and Winter Olympics. The y-axis represents various sports, and the x-axis denotes the years of Olympic editions. Each heatmap cell signifies the number of events for a specific sport in a given edition. These heatmaps provide insights into sport evolution, event counts, and sport popularity (indicated by color). Our analysis highlighted athletics as the most popular sport in the Summer Olympics (47 events in 2016) and Speed Skating and Cross-Country Skiing leading in the Winter Olympics (12 events each in 2014). Trends show discontinued sports like Art competitions and recent additions like Table Tennis and Taekwondo. Certain sports, such as Archery, exhibit volatile evolution with discontinuous phases.

D. Best Performing Nations in Olympics

Our analysis has identified the top 15 nations with the best performance in the Winter and Summer Olympics, based on the total number of gold medals they have won as of 2016. Notably, 9 countries (United States, Russia, Germany, France, Italy, Sweden, Finland, and South Korea) have secured positions in the top 15 best-performing nations in both the Winter and Summer Olympics. Furthermore, 11 of these top-performing nations are also members of the G20, a group consisting of the world's most advanced and emerging economies. These countries together account for over 85% of global GDP, 75% of global trade, and around two-thirds of the world's total population [18]. Even the remaining countries among the top 15 positions are considered economically well-developed based on GDP data. Therefore, it could be argued that the economically well-developed nations are also performing well in the Olympics.

E. Countriwise Comparison

Through our study, we compared the performance of nations with similar economic standings. Figure 3 illustrates a comparison between the United States of America (USA), the United Kingdom (UK), and India based on their overall medal tallies in each Olympic edition throughout history. The graph reveals consistent superiority of the USA over the UK except for a single instance, with India showing comparatively poor performance. Similarly, comparing two developing nations, Ethiopia has significantly outperformed Sri Lanka in Olympic history (Figure 4). This suggests that even nations of similar

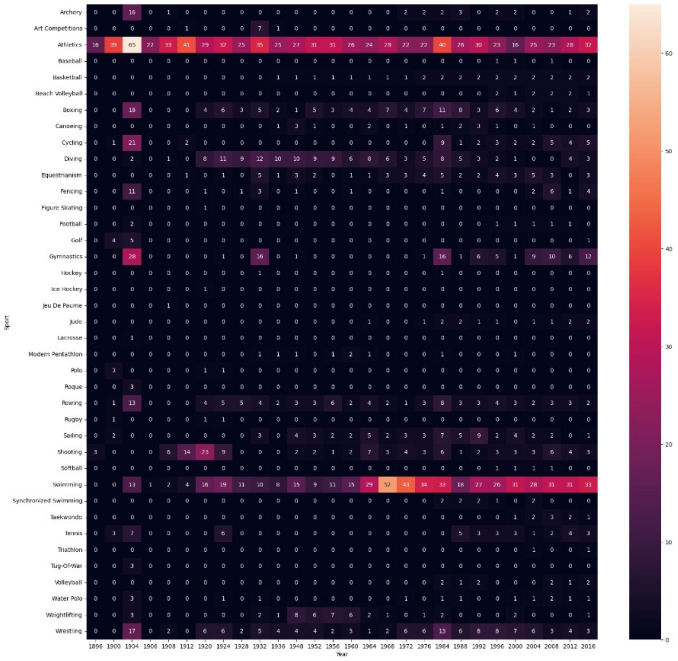

Fig. 5. USA Event Heatmap throughout the history of Summer Olympics

financial caliber could exhibit substantial differences in their Olympic performances.

The main objective of this study is to assist nations and athletes in enhancing their Olympic performance. To achieve this objective, individual heatmaps have been created for each Olympic nation, illustrating their strengths and weaknesses. Each element in these heatmaps represents the number of medals a nation has earned in each sporting event across all Olympic editions. By analyzing these heatmaps, nations and athletes could identify their areas of excellence and areas that need improvement. This information serves as a valuable guide for optimizing performance in future Olympic editions.

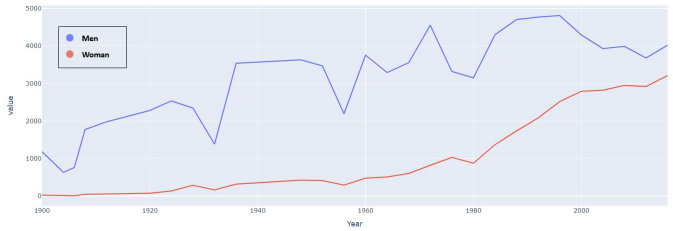

Fig. 6. Male vs. Female Sportspersons' participation in Summer Olympics

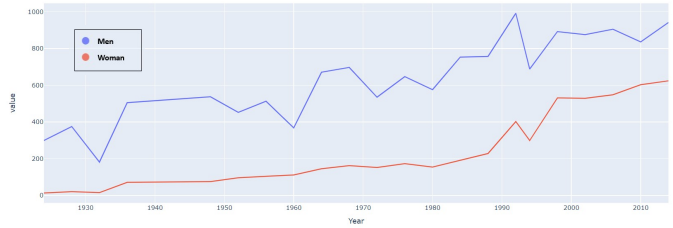

Fig. 7. Male vs. Female Sportspersons' participation in Winter Olympics

Based on the analysis of Figure 5, which depicts the performance of the USA in the Summer Olympics, it is apparent that athletics and swimming are the sports where the nation has shown considerable strength, thereby directly contributing to its top position in the leaderboards. However, despite displaying excellent performance in sports such as diving during the mid of Olympic history, the nation has lost all traces of victory in recent years. Additionally, it highlights the volatile nature of the USA's performance in some sports. Gymnastics displays an unstable history, with sporadic successes and declines. While recent consecutive editions have seen a rise in the USA's gymnastics performance, it has experienced a consistent decline in sports such as wrestling. In light of these observations, the USA could benefit from paying attention to these details to improve its performance in future Olympic events.

Similarly, any nation could leverage its heatmaps to discern strengths and weaknesses, facilitating targeted strategies and resource allocation to enhance performance in upcoming Olympic editions.

F. Male Vs. Female Sportspersons' Participation

Figures 6 and 7 reveal the evolving trend of gender participation in the Summer and Winter Olympics. Initially exclusive to male athletes, both events have gradually integrated female participants over the years. The graphs depict a continuous rise in overall participation, with a notable and consistent increase in female representation compared to the dramatic ups and downs observed in male participation across Olympic history.

The observed trends in the graphs suggest a continued rise in women's participation in future editions of both the Summer and Winter Olympics, reflecting efforts to promote gender equality in sports. However, external factors, including shifts in global politics and economic conditions, may influence sportsperson participation levels for both genders.

G. Height and Weight Vs. Medal Win

Height and weight are major aspects that a sportsperson keeps a keen eye on, as they directly affect the performance of sportspersons. Based on our analysis, we have created scatter plots for each sport, depicting the relationship between the height and weight of sportspersons and their corresponding medal wins. These scatter plots reveal a connection between height, weight, and medal wins in many sports.

Based on Figure 8, illustrating the Height and Weight scatter plot of female gymnastics, the analysis indicates that the optimal height range for winning gold medals in this sport is 180 - 190 meters. Most medal wins fall within the height range of 170 - 190 meters and weight range of 60 - 100 kg, with outliers present. Winning a medal outside of this range becomes more challenging, suggesting that athletes maintaining heights and weights within these ranges may increase their chances of success in upcoming editions. This information serves as valuable guidance for sportspersons aiming to excel in future Olympic competitions.

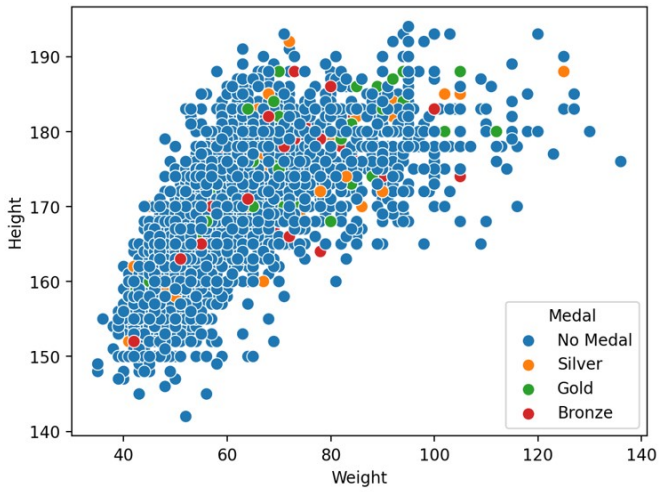

Fig. 8. Height and Weight vs. Medal Win for Female Gymnastic in Summer Olympics

H. Correlation Analysis

To observe the strength of the relationship between the external variables and the medal win in the Olympics, we have conducted a statistical correlation analysis using Pearson and Spearman Correlation methods. The results of the statistical analysis using Pearson and Spearman correlation methods (Table I) showed that the Pearson method yielded a lower correlation coefficient value compared to the Spearman method for the population. This suggests that there may not be a direct relationship between the population of a nation and the medal wins. However, for the GDP, both methods obtained values around 0.45. While this may not be considered an extremely high coefficient value, it is still a higher positive value, indicating a moderate positive relationship between a nation's GDP and the medal wins. Therefore, it can be concluded that there is a relationship, although not the strongest, between a nation's GDP and winning medals in the Olympics.

V. Conclusion

The study aimed to analyze the historical data of both the Winter and Summer Olympics to evaluate the performance of nations and athletes and observe the evolution of the Olympics in order to help the nations to improve themselves for a better performance in upcoming editions of Olympics. Employing EDA under 14 factors and statistical correlation analysis, we examined trends such as the increasing participation of nations over Olympic history, the growth in the number of events, varying trends in sports events, the performance of nations, and the relation between factors such as age, height, weight and medal wins. The results revealed the evolution of the Olympics in terms of participation, events, and gender representation. There is also an age, height and weight range to win a medal per particular sport.

Through the correlation analysis, we could conclude that the GDP of the country depicted by the amount the country could invest in the Sports sector, contributes positively to the medal win while the population does not affect much stronger. Nations also could improve themselves in particular sports and events through the insights given by heatmaps while identifying their and opponents' strengths and weaknesses. The use of various graphical formats such as line graphs, scatter plots, bar graphs has been employed to visually represent and validate the analysis of these factors. These visualizations provide a clear and concise way to present complex data, making it easier to understand and interpret the results of the analysis.

Acknowledgment

I wish to extend my gratitude to Dr. Udaya Wijenayake, my esteemed supervisor, whose role has been paramount in the successful completion of this paper. His multifaceted role as both my supervisor and co-author highlights the profound nature of his contribution. His unwavering commitment to guiding, mentoring, and actively collaborating throughout this study has been instrumental in shaping its outcome.

References

[1] International Olympic Committee. (2023, Mar) History of the IOC. [Accessed: 18-Apr-2023]. [Online]. [Online]. Available: https://olympics.com/ioc/ancient-olympic-games
[2] ——. (2022, Jun) Athens 1896 summer olympics - athletes, medals & results. [Accessed: 18-Apr-2023]. [Online]. [Online]. Available: https://olympics.com/en/olympic-games/athens-1896
[3] ——. (2022, Jun) Chamonix 1924 winter olympics - athletes, medals & results. [Accessed: 18-Apr-2023]. Olympics.com. [Online]. [Online]. Available: https://olympics.com/en/olympic-games/chamonix-1924
[4] ——. (2023, Feb) Tokyo 2020 summer olympics - athletes, medals & results. [Accessed: 18-Apr-2023]. Olympics.com. [Online]. [Online]. Available: https://olympics.com/en/olympic-games/tokyo-2020
[5] ——. (2023, Feb) Pyeongchang 2018 winter olympics - athletes, medals & results. [Accessed: 18-Apr-2023]. Olympics.com. [Online]. [Online]. Available: https://olympics.com/en/olympic-games/pyeongchang-2018
[6] C. Elkan, "A prediction model for which country will win the highest number of "gold" in 2016," 2016.
[7] A. Sen and G. Margaj, "Predictive analytics and data mining," 05 2010.
[8] D. M. Leonard, "Data mining of sports performance data," 2011.
[9] C. S. Thirumalai, M. Sankar, and A. Vijayalakshmi, "Heuristics prediction of olympic medals using machine learning," 04 2017.
[10] J. M. Gonzalez Rave, I. Yustres Amores, and D. Juárez, "Swimming performance analysis in 2016 summer olympic games," *Retos*, pp. 256–259, 01 2017.
[11] N. Mel'nikova and A. Nikiforova, "Olympic legacy: Comparative analysis of performance of national teams of host countries of olympic winter games (1988-2014)," Ph.D. dissertation, Russian State University of Physical Culture, 2015.
[12] M. Komorowski, D. C. Marshall, J. D. Salciccioli, and Y. Crutain, "Exploratory data analysis," in *Secondary Analysis of Electronic Health Records*, 2016, pp. 185–203.
[13] D. Yamunathangam, G. Kirthicka, and S. Parveen, "Performance analysis in olympic games using exploratory data analysis techniques," *International Journal of Recent Technology and Engineering*, vol. no. 201974, pp. 251–253, 2019.
[14] R. Pradhan, K. Agrawal, and A. Nag, "Analyzing evolution of the olympics by exploratory data analysis using r," *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, 2021.
[15] A. Rathke and U. Woitek, "Economics and olympics: An efficiency analysis," *SSRN Electronic Journal*, 2007.
[16] R. Griffin, "120 years of olympic history: athletes and results," Dataset, Massachusetts, United States, Cambridge, May 2018.
[17] T. Guardian, "Olympic sports and medals, 1896-2014," Dataset, Manchester, United Kingdom, London, 2017.
[18] G20. About g20. Accessed: 19-Apr-2023. [Online]. Available: https://www.g20.org/en/about-g20/