

Olympic Data Analysis: Uncovering New Insights into Athletic Performance and Competition

Krishan Gopal Gupta¹ and Aditi Arora²

^{1,2}ABES Engineering College, Ghaziabad, India

Email: krishnagopalgupta94@gmail.com, aditi.arora@abes.ac.in

Abstract— The Olympic Games are known to be the world's leading sports event wherein thousands of athletes across the globe participate in numerous Games. More than two hundred countries take part in Olympics with an exponential increase in the number of participants over the years. Furthermore, this leads to an increase in the size of performance data records of various countries/athletes participating in a variety of different sports events in Olympic Games. In this work, our aim is to perform an exploratory data analysis on Olympic dataset that engage users with interactive dashboards to explore various data facets, including medal tallies, country-specific analyses, athlete-centric insights, and comprehensive overviews through specific python libraries to analyze, evaluate and retrieve useful insights about the performance comparison of various countries/individual athletes, growth rate of women participation so etc. Our dataset contains more than 2,71,116 records such as Athlete's name, gender, country, age, height, weight, Gold, Silver, Bronze Medals etc. We have used appropriate visualization technique such as heat maps, graphs, for data representation. Line Charts showed the real-time labeling in our user interface from the analyzed dataset. We have used one hot encoding technique to distribute the medals in gold, silver, and bronze categories respectively. The analysis done through visual tools provided useful insights about understanding various factors to work upon for the scope of improvement in future Olympics.

Index Terms— Olympics, Data Visualization, Exploratory Data Analysis, One hot encoding.

I. INTRODUCTION

The Olympic Games, also known as festival of athletes, is the most prestigious international sporting event which occurs once in every four years where potential athletes from more than 200 nations participate in various sports (Approx. 45) to showcase their brilliance and bring laurels for their countries. Pierre de Coubertin, a French educator and sports enthusiast had a vision to bring different countries and athletes on a common platform to compete and prove their sporting excellence so as to make their nations proud by fetching winning laurels for their country. He turned his vision into reality by initiating the first modern Olympic Games in Athens, Greece in 1896 [1]. Fig.1. shows the evolution of Olympics Games from 1896 onwards. In the first modern Olympic Games only 13 countries participated which was The United State, Sweden, The United Kingdom, Denmark, Bulgaria, France, Hungary, Italy, Australia, Austria, Germany, Greece, and Switzerland. The Above diagram show some highlights during the Olympics which was:

- 776 B.C.: The first ancient Olympic Games were held in Olympia.
- 1894: Baron Pierre de Coubertin was the founder of the first modern Olympic Games.

- 1896: First Modern Olympic Games in “Athens”, Greece.
- 1980: The United States boycotted the Olympic Games held in “Moscow”, Russia.
- 2000: The 2000 Olympics were marred by several doping scandals.
- 2016: Brazil, the host country won its first gold medal in football.

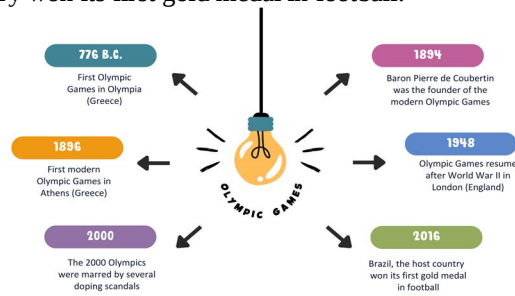


Fig.1. This graph shows the highlights from the beginning of the modern Olympic Games (1896) till 2016.

This research paper aims to analyse the following factors contributing to take necessary actions for future improvements among countries and players:

- Performance of various countries participated in Olympics over the period of 1896 to 2016.
- Participating ratio of women with respect to men over the duration of 1896 to 2016.
- Best performing athletes in the different Olympic Games.
- Motivated by the need to study such large *120-years-of-olympic-history-athletes-and-results* [2] dataset to get insights about the above- mentioned points, we aimed to perform exploratory data analysis (EDA) using python libraries and some powerful visualization tools to help countries and players analyze their status in Olympic Games and key aspects to work upon to perform better in future Olympics.

II. LITERATURE SURVEY

In this section, we present a comprehensive overview of the current research work done to analyze the evolution of Olympics over the period of time.

The data related to Olympic Games is collectively a huge dataset which is increasing exponentially over the years with the increase in the popularity of Olympic Games across the globe. Detailed analysis needs to be done to predict future trends on the basis of previous data through statistical visualization, country wise-medal; it's ranking throughout the Olympics, and number of increments of women participation. Not much relevant research has been done in this area so far. We compare the different factors of the 120 years of Olympic Games. We are using EDA technique [3] to analyse the vast array of information about the Olympic Games. Exploratory data analysis is a way of exploring datasets to find patterns, insights, and statistical visualization on the basis of graphs and charts. We are inspired by various research papers.

TABLE I: DESCRIPTION OF DIFFERENT RESEARCH PAPERS

Author[citation]	Paper	Description
Data, M. C., et al. (2016) [4]	Exploratory Data Analysis	Secondary analysis of electronic health records for insights
Shah, H., et al (2023) [5]	Olympics Data Analyzer with Prediction.	Analysing Olympic data with predictive modelling for future outcomes
Kulkarni, Nishant et al. (2022) [6]	Olympic Data analysis using data science	Applying data science techniques to analyse Olympic performance data.
Surya Sena Reddy, Suraj Kumar (2022) [7]	Data Analytics on Olympics	Utilizing data analytics to explore trends in Olympic data.
Pop, C. L. (2016) [8]	Athletics at the Summer Olympic Games. Marathon	Investigating athletics performance trends at the Summer Olympic Games.

These research papers have been studied so far in the Olympic Games but we want to elaborate more on this field. The main aim of our research papers is:

- Visualization of Participation of women over the years.

- Utilizing the idea of giving a broad overview, zooming in and filtering, and then giving more in-depth information when needed.
- Performance of an individual athlete.
- Performance of an individual country through visual aids.
- Implement multiple options in the Streamlit [9] web server UI to enable year wise analysis of an individual country.
- Providing the several aspects on the basis of Medal Tally (“Gold Medal”, “Silver Medal” and “Bronze Medal”), ‘Overall Analysis’(which involves the Participating of Nations over the years, Athletes over the years, No. of Events over-time, and Most successful Athletes), ‘Country wise analysis’ and ‘Athlete wise analyses.

III. PROPOSED APPROACH

In this work, we have considered a dataset of 120 years (From the beginning of Modern Olympics (1896) till 2016) which involves data collection from Kaggle dataset. We find out the insights of the Games on the basis of countries, athletes, medals and sports using exploratory data analysis. We have incorporated different visualization techniques such as bar charts, pie charts, comparison charts, and heat maps to show the pattern from the previous Olympic Games. We provide the user interface so that seekers can easily access the insights of the Olympic Games. The analytical findings of our work will help nations to understand the weaknesses and work on them to perform better in future Olympics. In the end, our aim is to supply countries with useful information, enhance their Olympic performance, observe further sporting events and make wise selections on the Games potential.

IV. METHODOLOGY

To study Olympic data, researchers employ a range of approaches, such as data cleansing and exploratory data analysis (EDA) [10]. These approaches can be used to examine athlete’s performance from the previous trends, women participation over the years, introduction of different Games in the Olympic Games. We will go through all the process i.e. Data Collection, Data transformation, Data Cleaning, Data Visualization, Data Analysis, User Interface, and finally we deploy our model on the Streamlit web server to keep our research online so that the seeker can easily retrieve the required data.

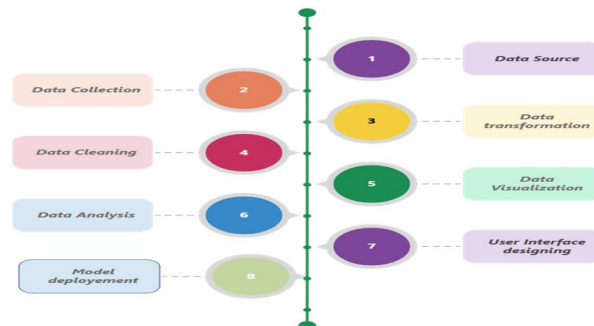


Fig. 2. The figures show the step-by-step process of our analysis

A. Data Collection

Firstly, we collected our data from 120-years-of-Olympic-history athletes-and-results [4]. We have considered two different datasets. The first dataset athlete_events.csv contains information about the athletes such as name, age, height, gender, event, weight, sex, medals and more, and the second dataset noc_regions.csv contains the region of the athletes and their respected National Olympic Committee (NOC) 3 letter code. The overall combined dataset contains more than 2, 00,000 rows and 20 columns.

B. Data Wrangling

This involves several steps to ensure the analysis is based on clean and reliable data:

1) Data Cleaning

This step involves handling missing values, removing duplicates, and dealing with outliers. We preprocess our data with the help of Deterministic imputation. Deterministic imputation is used to fill in missing information in a systematic and predetermined method rather than relying on randomness or statistical inference. This step helps

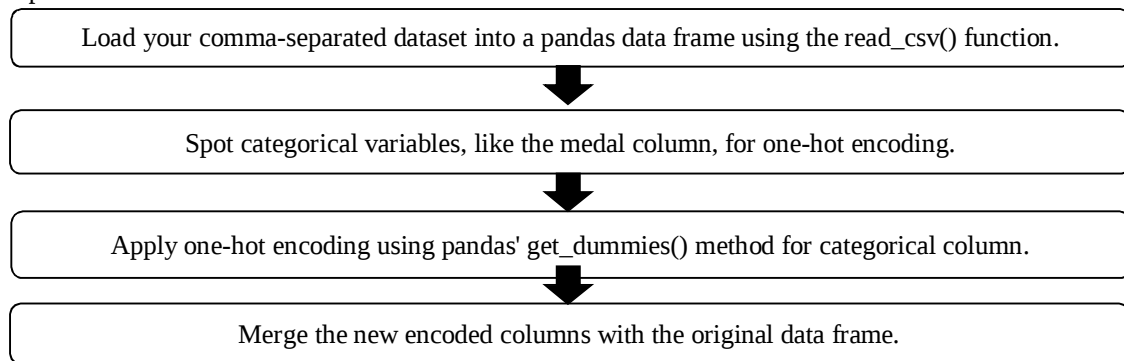
us to remove null values from our dataset, especially in fields like age, height, weight, etc. There are some methods that remove the null values from the given dataset which are:

TABLE II: METHODS FOR REMOVING NULL VALUES

Method	Approach
Naive Method	<code>df.dropna(how='any')</code>
Filtering Method	<code>df[df['col1'].notnull()]</code>
Deletion Method	<code>df.dropna(inplace=True)</code>
Imputation Method	<code>df.fillna(value=0)</code>

2) Data Transformation

The steps that we follow for data transformation:



3) Data Integration

We have two datasets i.e. athlete_event.csv and noc_region.csv from Kaggle, this steps involves combining them in a meaningful way. This could involve merging datasets on a common key, or aggregating data at a higher level.

4) Data Reduction

This involves simplifying your data, improving computation efficiency, and uncovering underlying its essence without losing its essence.

C. Exploratory Data Analysis

It is the most important part of our Olympic data analysis research. Exploratory data analysis is a way of analyzing the data and its attributes with the help of various graphs and plots. We can compare data in visual format using several plots i.e., Line Graph, Scatter Plot, and Histogram. EDA also gives us insights into our model such as:

- *Distribution of medals*: Exploratory data analysis can be used to visualize the distribution of medals across events, sports and countries. It also helps to identify different patterns and trends.
- *Relationship between variables*: EDA is also used to compare the relationships such as age and performance, or the relationship between height and weight.
- *Missing values*: We can also use them to find out the missing values from our dataset.

D. Analysis and Visualization

Olympic data is a rich and complex dataset that can be used to gain insights and comparative analysis with past Olympic performance of countries or athletes over different Olympic Games. This could highlight significant trends and changes. We are using different technique such as NumPy, matplotlib, seaborn, pandas etc. to analyse our dataset. We also categorize medals into three categories i.e. gold, silver, and bronze using one-hot encoding. It is a method by which we encode our variable into a binary vector and the value of each binary vector is 1 if the given value is present, otherwise 0 if it is null in the dataset. The advanced and interactive visualizations can provide deeper insights into the data. This could include:

- Time series plots to show the performance of a country or an athlete over time.
- An Interactive User Interface that allows users to experience different aspect of the data.
- Heat maps show the distribution of medals across different countries

V. RESULT

Our analysis reveals fascinating discoveries about Olympic performance. These various graphs, heatmaps, and charts depict the performance of individual countries and athletes across time. Below, we present our results showcasing these insights:

1) *Line graph for the Contribution of Men and Women from 1896–2016*: Previously Women were not allowed to participate in the Olympic Games, but in the late 19s and early 20s centuries women were allowed to participate in the Olympics, and the participation of women over the years also increased and the same have been shown in figure.

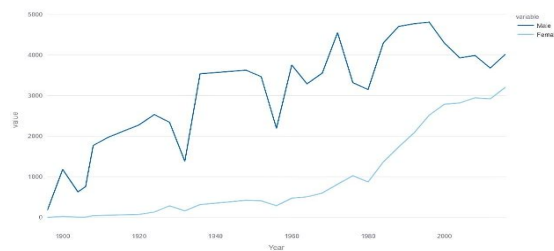


Fig. 3. The number of participation of women (w.r.t) to Men increased after 1980.

The Number of contributions of women with respect to men is always less because of many reasons such as family support, social support etc. Hélène de Pourtalès was the first female participant who played in the Olympics and won her first Olympic medal. In 1896, there no women Participation was shown. In the 1948 Olympics, there were approximately 4,104 athletes, of which around 390 were women. The ratio of women with respect to men in 1948 was roughly 1:10. In the 2016 Olympic, the total number of participations was 11180 in which 6146 were men and 5034 were women which shows the significant increment in women participation.

2) *Identify Number of Athletes over the years*: The line graph shown in Fig. 8. shows the increase in participation of athletes over the years. This graph is sometimes downward due to some issues such as, in 1916 Berlin was canceled the Olympics due to World War I. In 1980, the summer Olympics was boycotted by the USA in Moscow.

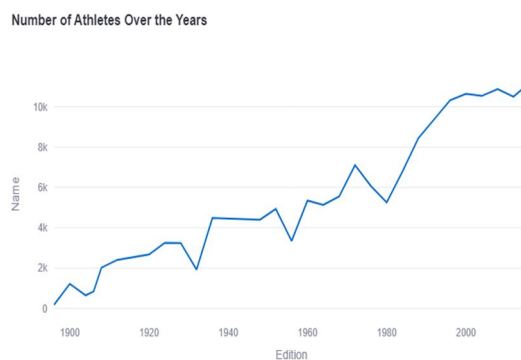


Fig. 4. The number of Participation of athletes was very low due to boycotts, and World Wars.

In 1984, Los Angeles was boycotted by the Soviet Union and its allies in the Soviet Union's invasion of Afghanistan, and there also many decades that were happened, but after 1988 we can see the number of increments of participation of athletes over the years respectively.

The increase in participation by athletes has both positive and negative implication, positive sides implies that the Olympic Games are more inclusive than ever, regardless of their background or financial situation, athlete have the equal opportunity to compete the with their highest potential, while negative sides in the Olympics increase the stress over the athletes, increasing in the number of competitors, lack of resources

3) *Line graph for the comparison between height and weight*: In the Olympics Games from 1896-2016, we find out the evolution of athlete' height and weight and its hidden pattern and trends over the years.

For example: After 1896, the height of the men's 100m Olympic champions was increased. The first winner, Tom Burke was 1.83m tall, while the 2016 champion has risen significantly as shown in Figure 4

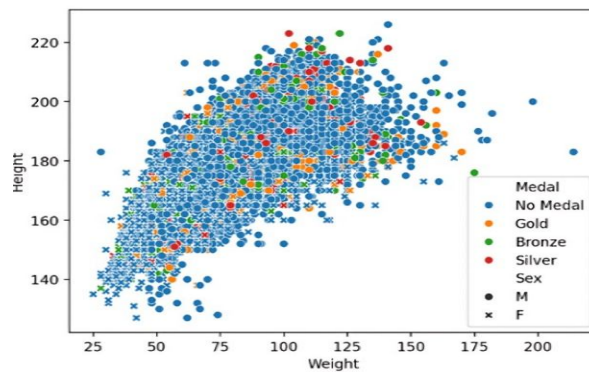


Fig. 5. scatter plot shows that there is a positive correlation between height and weight.

4) *Identify trends of Participation by countries over the years:* The graph shown in Figure 6 shows the rapid increase of participation of countries in modern Olympic Games from the beginning till 2016, with the exception of a few years during World War I and World War II.

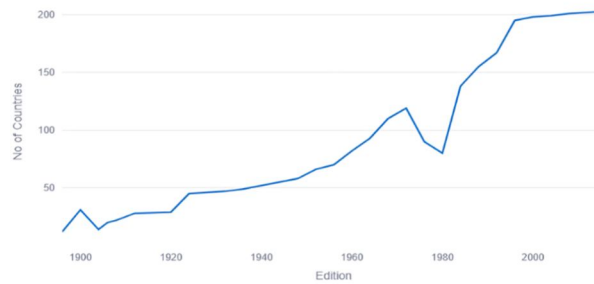


Fig. 6. The Line graph may sometimes show a downward trend due to external factors, such as wars, political conflicts, economic and problems.

After 1900, we can see the rising curve in the graph showing how the number of participants increased suddenly, displaying the increasing popularity of the Olympic Games.

5) *Distribution of Age with respect to sports (Gold Medallist):* The scatter plot in Figure 5 shows the average age of a Gold Medallist is between 30 to 40 years old, and for a silver medallist, the average age is between 40 and 50 years old and for Bronze Medallist have the average is between 40 and 50 years old respectively. The most important factor that can be considered during the World's largest Games events is how to adopt an individual game according to the distribution of their age. This impacts a lot on the athlete during the performance.

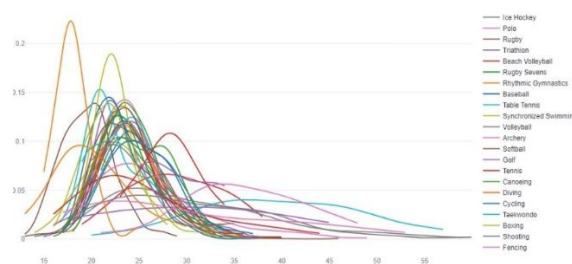


Fig. 7. The average age of the athlete with respect to the events.

6) *Heat Map of Number of Events over Time (Sports Event):* Heat maps are color-coded graphs that show how different data points relate to each other. They are often used to identify patterns and trends in data. It is a very powerful tool to visualize data because the user can quickly identify the hidden patterns.

The Heat Map shown in Figure 9 shows the following patterns:

- India won 11 medals from 1928 to 1980 in Hockey.

- India won 0 medals in athletics from 1900 to 2016.
- India won 2 medals in Badminton from 1900 to 2016.

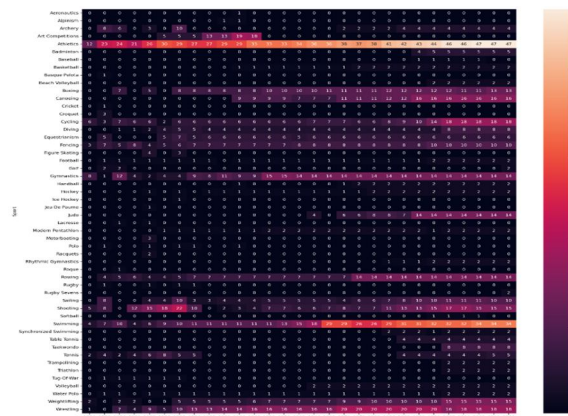


Fig. 8. This heat map shows the trends of various sports in the Olympic Games

E. User Interface

After Analysis of the dataset. We are set to develop a dynamic User Interface (UI) or the live Olympic data analysis application. Our UI acts as a portal to a comprehension understanding of Olympic data, recognize trends, and acquire valuable insights into the world of Olympic sports. Our UI provides a diverse range of functionalities, allowing users to toggle between various options such as Medal Tally, Overall Analysis, Country-wise Analysis and Gender-wise Analysis, to further empower user experience our UI will provide multiple options including the ability to filter data by year and country. This feature will allow users to conveniently assess the performance of a specific country in any given Olympic year, enhancing the overall usability and flexibility of our application.

F. Model Deployment

In the last stage of our research, we deploy our model on the Streamlit web server. Streamlit, an open- source Python library, to create powerful UI, custom web apps for machine learning and data science. The intuitive features and user-friendly design of the platform provides developers and researchers, to easily transform their application into dynamic accessible data apps. Streamlit also provides the pre-build feature of toggling between the dark and light mode without adding any functionality in your code. With minimal Python proficiency, one can easily construct a user interface through Streamlit.

VI. CONCLUSION

The Olympic Games is an international sports event where athletes from different countries participate and show their potential. This research work contained the 120 years of Olympic data from 1896 till 2016. We have two datasets on which we examine our data by combining both datasets which gives us more than 2,71,000 rows and 20 columns. We will provide a detailed analysis of the Olympic Games through graphs, charts, Scatter plots, and visual aids. We used Exploratory data analysis, one hot encoding method and Python libraries such as Pandas, NumPy, Seaborn, Streamlit, Matplotlib, etc. to analyze our dataset. The main purpose of our research work is to gain the insights of the participation of athletes and comparing the various factors of the Olympic Games such as participation of women over men, performance of an individual country and athlete, etc. Exploratory Data Analysis helps to gain insights into the datasets in the visual formats and interact it in the user interface with the help of Streamlit, after that we will deploy our model on Streamlit web server so anyone can interact with our application.

REFERENCES

- [1] https://en.wikipedia.org/wiki/Olympic_Games
- [2] <https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results>
- [3] D., Yamunathangam & Kirthicka, G. & Parveen, S.. (2019). Performance analysis in olympic games using exploratory data analysis techniques. International Journal of Recent Technology and Engineering. 7. 251-253.
- [4] M. C., Komorowski, M., Marshall, D. C., Saliccioli, J. D., & Crutain, Y. (2016). Exploratory data analysis. Secondary analysis of electronic health records, 185-203.
- [5] Shah, H., Sheth, J., Savla, H., Bansode, J., Patel, B., & Yewale, A. Olympics Data Analyzer with Prediction, 2023.

- [6] Kulkarni, Nishant & Patil, Pratik & Pande, Rugved & Patil, Dhiraj & Nair, Pranav & Prabhu, Parth & Doshi, Pratyush & Bhosale, Pranav. (2022). Olympic Data Analysis using Data Science. International Journal for Research in Applied Science and Engineering Technology. 10. 10.22214/ijraset.2022.48046.
- [7] Surya Sena Reddy, Suraj Kumar, Data Analytics on Olympics, International Journal of Research Publication and Reviews, vol. 3, no. 9, pp. 1405-1410, 2022. ISSN: 2582-7421.
- [8] Pop, C. L. (2016). Athletics at the Summer Olympic Games. Marathon, 8(2), 244-253.
- [9] <https://streamlit.io/>
- [10] Pradhan, Rahul & Agrawal, Kartik & Nag, Anubhav. (2021). Analyzing Evolution of the Olympics by Exploratory Data Analysis using R. IOP Conference Series: Materials Science and Engineering. 1099. 012058. 10.1088/1757-899X/1099/1/012058.