The Evolution of the Olympic Games: A Visual Analytics Approach

Sajjaad Jurawon

Abstract— In this project, we use visualisation techniques and visualisation tools such as Python libraries (matplotlib, geopandas) and Tableau, to analyse how different aspects of the Olympic Games have changed over time. 120 years of data has been used to perform this analysis. We looked at diversity in the Olympic Games and found out that in the early years, most partcipating countries were westerners compared to nowadays where there is almost a complete global participation. In terms of gender, an equal proportion of men and women has been seen nowadays compared to the early years where the games were mainly for male athletes. We also look at the number of sports played and learn that addition and removal of sports are strategic moves to appeal to viewers and attract younger generations. Top performing countries were mainly westerners such as USA in the first few years of the Olympics however, nowadays there is more competitiveness. Using Kmeans clustering, we have grouped athletes using physical attributes, sports played and success in the Olympics to identify changes in the characteristics of participants. The results showed that the average BMI of athletes has remained relatively stable over the decades, the diversity in BMI has increased, reflecting the inclusion of a wider variety of sports requiring different physical profiles. Moreover, the steady growth in athlete participation across all clusters highlights the expansion and increased inclusivity of the Olympic Games over time.

1 Problem Statement

The Olympic Games, a major sporting event every four years that brings millions of people worldwide to celebrate the best of humanity. It covers a wide range of sports where athletes from different countries compete; it is an important symbol of sporting events and cultural exchange [1]. Since the first "modern" Olympic games in Athens, 1896, where only thirteen countries were competing to today where more than 200 hundred nations competing in a variety of sports. It has a rich evolutionary history where many changes can be visualized using data of over a century.

In this study, we aim to use visual analytics techniques to investigate the evolution of the Olympic games. Using historical data of the Olympic, we attempt to answer the following questions:

- How diversity of participants in terms of gender, nationality, and representation across regions have changed since the first "modern" Olympic Games?
- How has the number of sports in the Olympic Games changes?
- Which countries have been leading the rankings and how their results evolved over time?
- Have the physical profiles of athletes evolved?

Through this study, we may identify valuable insights that could help sports organizations and policymakers seeking to foster inclusivity. Understanding the evolution of countries' performances and athlete profiles can help with strategic decisions around funding, training, and event planning. Ultimately, this work highlights how data-driven insights can help ensure the Olympic Games continue to serve as a global stage for cultural exchange, fair competition, and more importantly, the celebration of human potential.

2 State of the Art

The article, "Performance Analysis of Olympic Games using Data Analytics" by Asha et al. (2023), explores the use of data analytics to identify patterns and trends in the Olympic Games [2]. The study uncovers insights into the historical and contemporary patterns of the Games by leveraging data analytics to analyze athlete performance, medal distributions, and demographic trends. Methodologies such as descriptive statistics, data cleaning, and inferential analysis are used alongside predictive models, such as regression and decision trees, to forecast future performance trends. Line trends, bar charts and histograms are used through the article to illustrate findings. A key focus of the article is the analysis of medal trends revealing long-term patterns in the dominance of specific countries and variations in performance over decades. The article also emphasizes gender participation, showing a steady increase in female representation over time. The study shows how data-driven methodologies can help provide valuable insights into training strategies, decisions made by stakeholders and improvement in structure.

While the article provides valuable insights, it also identifies challenges, such as the limitations posed by inconsistent data. The complexity of Olympic events and the diversity of sports further complicate the analytical processes. This underscores the need for more robust datasets and splitting the analysis into different sports to better capture the dynamics of the Games.

Another article, by Chowdary et al. (2024), "From Athens to Rio: A Comprehensive Data Analysis and Visualization of 120 Years of Olympic History", applied data visualization and analytical techniques on the same dataset used in our project to uncover important trends in different areas of the Olympic Games [3]. Using Tableau and R, they created a dynamic dashboard. Multiple diagrams were provided in the article. The authors performed a trend analysis, visualizing long-term trends, such as increase in countries participation, gender diversity and changes in athletes. Performance analysis showcased dominant countries and showed how emerging countries have improved over time.

This article successfully uses diagrams to explore key aspects of Olympic history, such as dominant countries, popular sports, and athlete demographics. The visualizations

provide compelling evidence of how the Olympic Games have evolved in terms of global participation, gender representation, and athlete profiles.

The article "Olympic Data Analysis using Machine Learning" by Bhosale et al. (2024) provides an in-depth exploration of Olympic data over time using machine learning and highlights the importance of explorative data analysis before implementing machine learning algorithms [4]. In this article, the author applied various EDA methods such as histograms, bar charts, and scatter plots, ensuring that the data is clean, reliable, and ready for advanced analysis. The article provided some prior insights which were proven to be helpful in the analysis of our project and further reinforce the results we subsequently observed.

3 Properties of the Data

The main dataset used to perform the analyses can be obtained from Kaggle. The dataset contains information about Olympic athletes, events and results for 120 years all the way from the first "modern" 1896 Athens Olympic Games to 2016 Rio [5]. The dataset consists of two CSV files, and we are using the dataset on athletes and events. The second dataset covers regions names and information about NOC. According to the Kaggle dataset description, the data was compiled from the official International Olympic Committee (IOC) website and Sports-Reference.com.

The dataset contains 217117 rows and 15 variables covering athlete details, event details and results. There is data on 35 sports and over 400 events across Summer and Winter Olympic Games. Table 1 below shows the list of variables and their description.

Table 1: Variables in the dataset

Variables	Description
ID	Unique number for each athlete
Name	Athlete's name
Sex	Male or Female
Age	Athlete's age
Height	Athlete's height (cm)
Weight	Athlete's weight (kg)
Team	Team name
NOC	National Olympic Committee
Games	Year and Season
Year	Year Games were held
Season	Summer or Winter
City	Host city
Sport	Sport name
Event	Event name
Medal	Gold, Silver, Bronze or Nan

The dataset is well-structured with a mixture of categorical, numerical and temporal data types. However, missing values are observed in the variables: age', height' and 'weight'. Figure 1 below shows a heatmap of missing data.

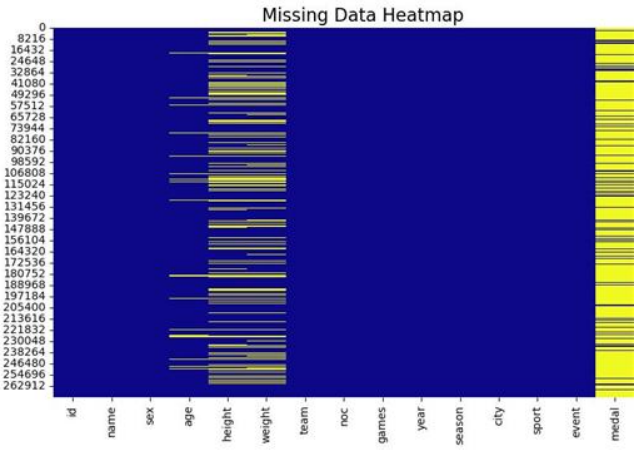


Figure 1: Heatmap of missing values

To preserve the temporal aspect of the data and prevent biasing over different time periods, missing values in the 'age' column have been filled using the mean for the specific year of competition. For the missing values in the 'height' and 'weight' columns, we have used the average values per country which preserve the natural variations that exist across different populations. The missing values in 'medal', as indicated in figure 1, suggests that those athletes have not received any medal; these are not missing observations. We have visualized the distribution of medals in figure 2 below. The distribution of gold, bronze and silver medals are similar with athletes with no medal being the majority.

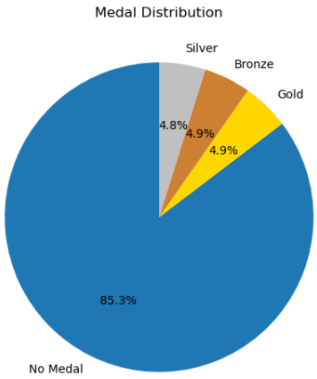


Figure 2: Distribution of medals among athletes

4 Analysis

4.1 Approach

Figure 3 shows the approach chosen to analyse several aspects of the Olympic Games that have changed over time. Below we discuss further all the computational methods that have been considered for each steps and explain how we have attempted to convert raw data into visuals to help with human reasoning

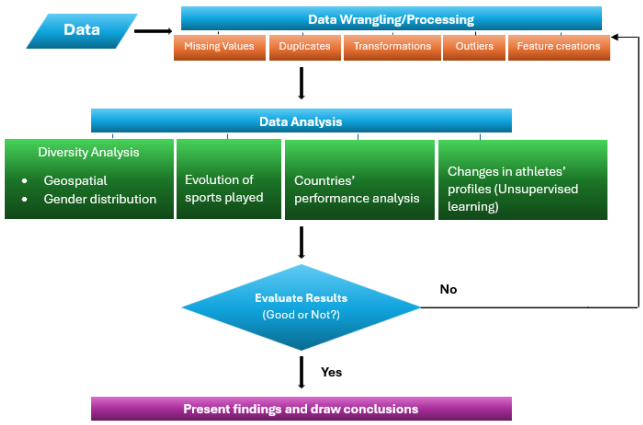


Figure 3: Flow of work map

DATA

The first step is to obtain all the required data for each section of the study. As previously mentioned, we are using Olympics Data from Kaggle and we will also be using vector data containing geographical features for countries, also known as a shapefile. The shapefile is found online on World Bank Data Group [5].

DATA WRANGLING/PROCESSING

This step is to ensure that the dataset is ready for the analysis. We have already identified missing values and replaced them earlier. We also check for duplicates, grammatical error, perform data tranformations where skewness is observed and remove outliers. These would ensure data quality and accuracy, statisticla integrity and consistency in our results. We also create new features tha. t are believed to be useful in the analyses.

DATA ANALYSIS

In this step, we implement a variety of visualisation techiniques paired with analytical knowledge to create diagrams that would help understand the evolution of the Olympic Games.

- Diversity Analysis

We perform diversity analysis to indentify by how much the Olympic Games are more inculsive towards underepresented regions and between genders. Geospatial visualizations will provide a clear and intuitive understanding of how the Olympic evolved in terms of geographic inclusivity. This transform complex data into useful insights ehich help understand the magnitude and pace of inclusiveness in events. Similarly, for gender, we can identify trends from when women's games were introduced. This makes it easier to understand the dynamics of gender equity in sports.

- Evolution of sports played

We also analyse how the number of sports played have changed over the years. A line trend chart will provide temporal perspective. We cannot see temporal changes in raw data.

- Countries' performance Analysis

Visualization will help identify patterns and trends than raw data. We will be able to compare the performance of top countries in the Olympics.

- Changes in athletes' profiles

Thorugh machine learning techniques we analyse how characteristics of athletes changes over different time periods. Clustering techniques groupong athletes based on physical attributes can be difficult to interpret without visualising the results. By transforming the raw clustering results into temporal graphs or period-specific charts, we can goin additional insights on the evolution of athletes' profiles.

EVALUATION OF RESULTS

In this section, we check if the results generated are accurate and help answer our research questions. We attempt to make changes to the data if results are biased or inaccurate.

RESULTS/FINDINGS

We present the findings found from the analyses once that we have confirmed that the results are accurate and unbiased. We draw conclusions from them and reflect on the work done.

4.2 Process

Data Wrangling/Processing

- Outliers Analysis

Since we have already replaced missing values in the previous section, we will focus on outliers as the first step in this section.

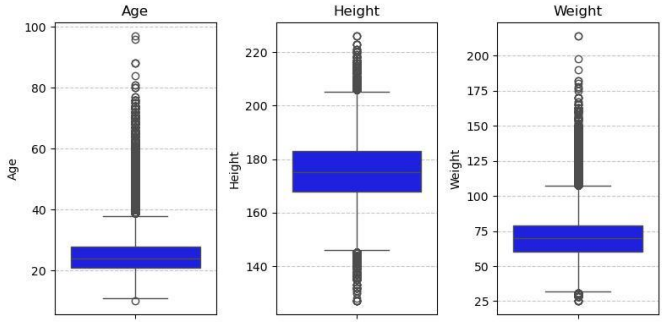


Figure 4: Boxplots of numerical variables (Outliers)

From figure 3 below, extreme outliers can be observed for age, but this information is accurate; transformation or their removal will be considered if distortions/biasness is observed. For height and weight, winsorization is implemented since the outliers are meaningful. Winsorization is a technique to limit extreme values in data by using percentiles [6].

- Feature creations

The first variable created is one where the names of countries have been changed to match the names identified in the shapefile. This has been done to ensure that all countries are included when the two datasets are merged.

The second variable created is the athlete BMI (Body Mass Index), which is a measure to assess athlete's body size by calculating the ratio of their bodyweight to their height. The following formula is used:

$$BMI\ (Body\ Mass\ Index) = \frac{Weight\ (kg)}{(Height\ (m))^2}$$

Weight is already in kg in the dataset, height however had to be converted from cm to m.

Another variable created is a 'decade' feature which is created based on the year the Olympic Games were held. For instance, the Olympic Games held in 1992 and 1996 were in the decade 1990. This proves useful when trying to group the data into specific periods.

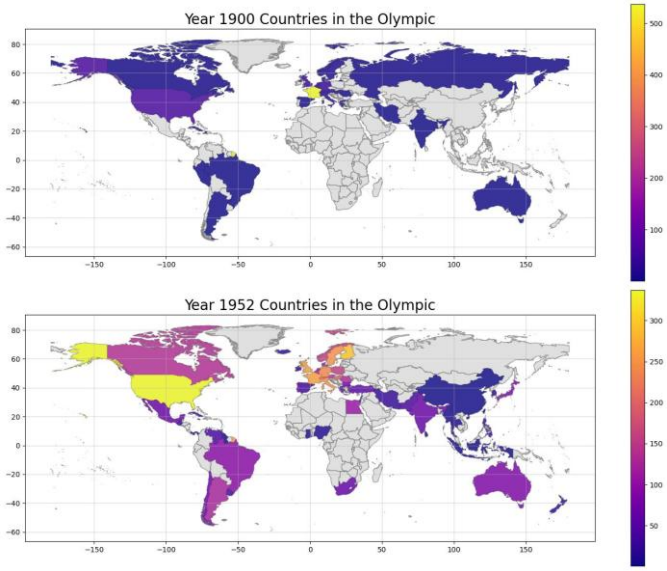
Since there are a lot of sports in the Olympic Games, we have created a 'sport category' variable where the different sports have been grouped into five types: Team Sports, Strength and Power, Precision and Skill, Endurance and Stamina, Mixed and Other. This has been done for easier interpretation of results.

Data Analysis

Diversity Analysis

- Geospatial Analysis

The first part of the diversity analysis consists of the choropleth world map using geopandas where countries participating in the Olympic Games with the number of athletes are displayed on a world map. We have used data for three different years to see how diversity evolved in the Olympics. We opted for the year 1900 as this captures the initial Olympic Games where participants were mostly westerners. The year 1952 was post-World War and during the Cold War. We expect to see some interesting changes in that year. The final year we considered is 2016; this consists of the Olympic Games as we know it where globalization and the internet have made the event more accessible.



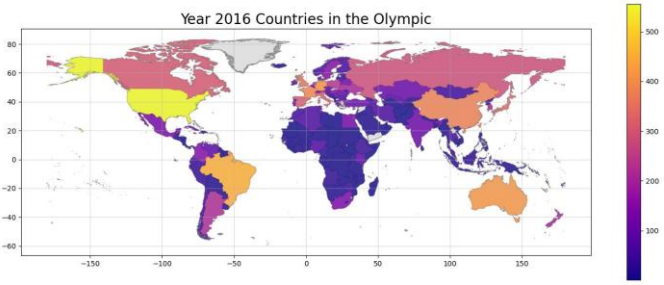


Figure 5: Geographical representation of the evolution of diversity in the Olympic Games

From the maps in figure 5, we can see that the Olympic Games have expanded significantly over the years. In 1900, only a few countries, primarily in Europe and North America participated, with very low representation from other countries. By 1952, more countries from South America, Oceania and Asia had joined, with Africa remaining underrepresented. The increase in 1952, post-World War II events, signified the end of colonialism and the beginning of the Cold War, more countries were seeking international recognition by participating if global events. In 2016, there is widespread global participation. Most countries are represented with USA, Europe, Brazil, Australia, Canada and China bringing than 300 athletes to compete.

- Gender Distribution Analysis

To analyze how the distribution of the gender of athletes has changed since the beginning of the "modern" Olympics, we have converted the ratio of male participants to female participants into percentage. A stacked area chart is used to see the changes.

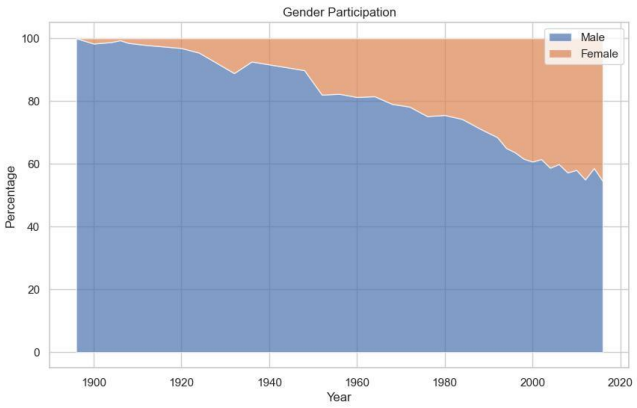


Figure 6: Percentage of Male vs Female athletes' participation

From figure 6, we can a clear consistent increase in female participation. In the early years, female athletes consisted of a very small proportion of the participants. In the first year of the Olympic, female participation was zero. By mid-20th century, female participation had grown but was still very low. A sharp upward trend with the gap between the genders narrowing is observed from 1970 and onwards. In latest years, the participation of female athletes is near 50%.

Evolution of Sports played

To evaluate the sports played during the Olympics, we opted for a line trend to see when new sports have been added or removed. We have split the data into Summer Olympics and Winter Olympics to better visualize how the number of sports changed over time for the specific Olympic Games.

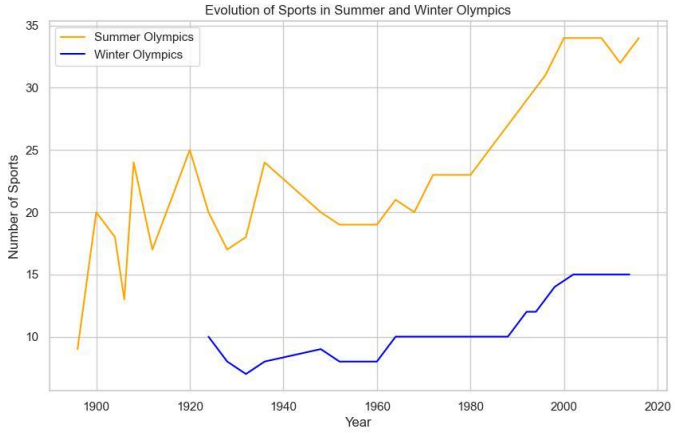


Figure 7: Sports at Summer and Winter Olympics

The first Olympic Games in 1896 featured nine sports. A general increase in the number of sports is observed. However, fluctuations are seen. In the early 20th century, some sports such as tug-of-war or club swinging were included in the games between 1900 to 1920, but these were later removed [7]. A decrease in the mid-20th century can be seen, which was due to the adjustments made by the IOC. Certain sports were removed since those had inconsistently been included in earlier games [8]. In recent years, new disciplines have been included aiming at younger audiences. For instance, breakdancing has seen a debut in the 2024 Paris Olympics [9].

Countries' performance Analysis

To analyze how the performance of countries changed over the years, we have selected the top 10 performing countries for the Summer Olympics and Winter Olympics. Combining both events in one diagram leads to unrealistic fluctuations since there are lower medal counts in the Winter Olympics.

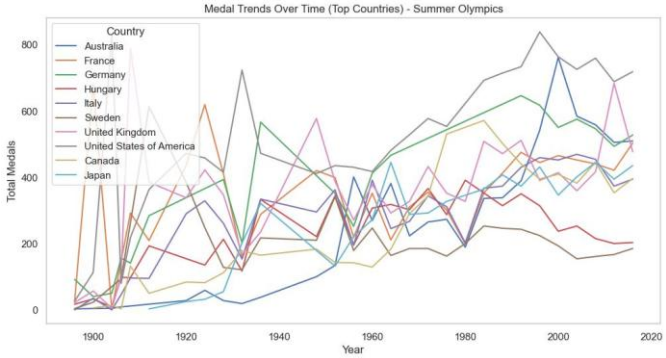


Figure 8: Top performing countries in Summer Olympics

Top performing countries differ between summer and winter Olympic games. We focus on the Summer Olympics for this section of the analysis. The graph for the Winter Olympics is provided in the Appendix section below. The USA consistently maintained a strong position in the Summer Olympics. A higher dominance is observed past the mid-20th century, peaking in the late 20th century and early 2000. Countries like France, Italy and Germany exhibits fluctuations. Australia, Japan and Canada show increasing trends. The dips around the World War periods: 1914 to 1918 and 1939 to 1945, correspond to Olympic cancellations during those years. A noticeable gap is observed.

Changes in athletes' characteristics

To visualize the changes in the physical attributes of athletes, we initially opted with simpler line trend which can be seen in Appendix 2. However, a flat trend was observed, and this does not consider the different characteristics of athletes. Hence, we opted to unsupervised learning to create clusters of athletes for each decade and see how this has shifted over time.
K-means clustering is the algorithm of choice and to find the right number of clusters, an elbow plot is used (Appendix 3). BMI, decade, sport category, sex and medal are the features used. BMI is the main characteristic to track changes in athletes' physical attributes. Decade help us capture temporal trends while medal helps us differentiate top athletes from non-medalists. Sport category is necessary since different physical traits are needed for different sports. Sex accounts for gender-related differences in performance.

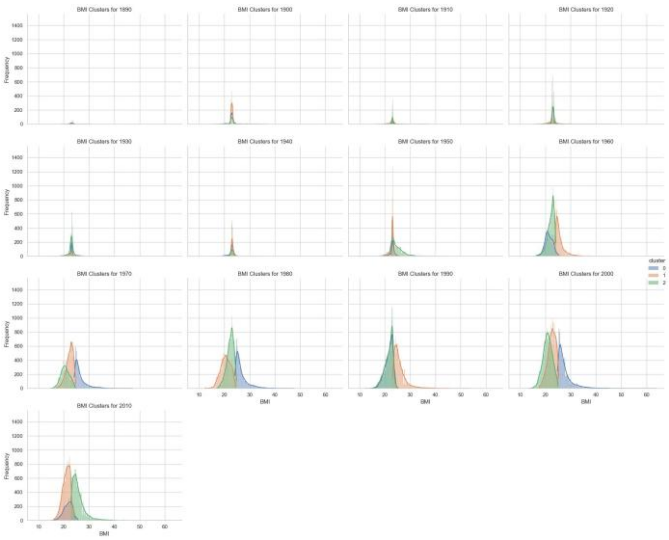


Figure 9: Faceted Plot of Cluster of athletes' BMIs for each decade

Figure 9 shows the different clusters obtained for each decade. Three clusters are identified: lean athletes (low BMI), balanced athletes (with moderate BMI) and power athletes (high BMI) but there are overlapping among them. In the early decades (1900-1940) clusters are narrowly spread with most athletes concentrated in the narrow range of 20-26 BMI. For most decades, athletes with low BMI (lean physique) dominates. In the mid and late 20th century (1950-1980), more variation is observed. In recent decades, (1990-2010), a broader range of BMI values is seen, reflecting, increased participation and diversity in body types.

4.3 Results

We have seen that the Olympic games have become more inclusive nowadays towards underrepresented people compared to the first few years. The maps reflect the impact of globalization, decolonization, and international cooperation in increasing global engagement in sports. Although 2016 show almost complete global participation, the number of athletes representing each country in not the same. Some imbalances can still be seen where many countries are unrepresented; they provide less athlete and have less support.

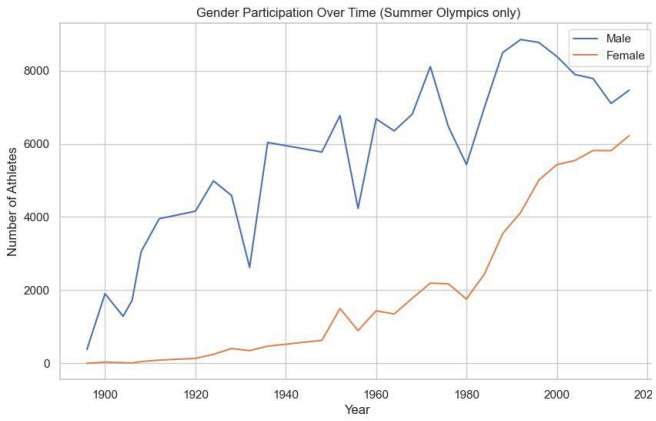


Figure 10: Number of Male and Female athletes

In terms of gender, an equal percentage was observed as of 2016. However, from figure 10, we can see that along with a substantial increase in women participation, there has also been a decrease in male athletes.

We found out that the sports played evolve to match global interests and are strategically removed or added. We have also seen that in the early years western countries dominated the games but post-1950 more competitiveness is seen.

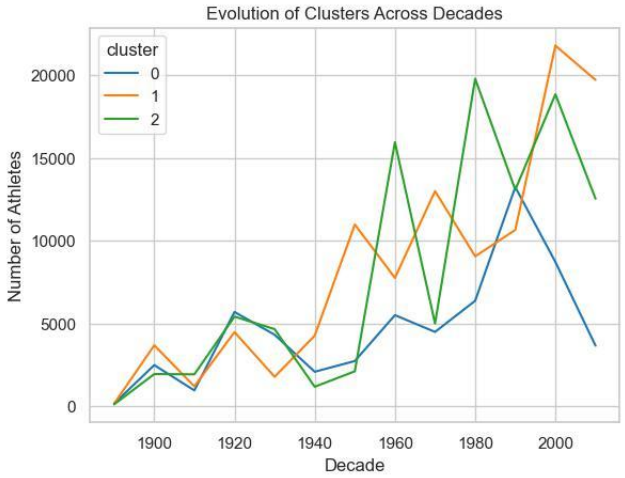


Figure 11: Evolution of athletes' characteristics across decades

By using unsupervised machine learning, we identified that most athletes fall into a central BMI range, with fewer athletes in higher and lower ranges across all decades. From figure 11, we can see that the number of athletes has grown consistently across all clusters, with a significant rise after the 1950s. Cluster 1, which consists of athletes with low BMI, appears to become dominant after 1960, indicating a shift in athlete characteristics with the addition of more sports and greater international participation. Clusters representing higher BMI ranges have grown over time, which may reflect the increasing prominence of sports requiring greater strength and power (e.g., weightlifting).

5 Critical Reflection

The Olympic Games have changed considerably throughout the years. We have implemented the temporal information in the dataset to visualize those changes and find underlying insights into what might have been the cataclysm for those changes.

While we have successfully conveyed how participation of countries has evolved at different intervals and identified the number of athletes from participating countries, the choice of years (1900, 1952, and 2016) covers significant gaps. By including more intermediate years (e.g., every two or three decades) we could have seen a clearer picture of when key changes occurred.

As for the analysis on gender, we have seen that there has been changes in both men and women leading to a near equal participation rate. However, a more in-depth analysis could be conducted to find out the reasons for these changes, for instance the addition or removal of sport s might be one of the reasons.

Similarly, the analysis of the number of sports played was very brief. We have overlooked expansion within sports. A more detailed analysis could be done for specific time periods to investigate what sports have been removed or expanded upon.

When analysing performance of countries, we were able to identify all time top performing countries and how periods of war lead to a decrease in participation where the dips in medal count during those periods are clearly visible. The graph however has a lot of overlapping lines, and we have not been able to include new top performing countries, like China. Using an interactive dashboard may help solve some of those issues or by splitting into smaller groups, for instance analyse top performing countries for a specific region.

Through clustering, we have been able to identify different categories of athletes and how these categories have changed over decades. We have seen that there are more variations as the Olympic Games become more inclusive and introduce different types of sports. Through clustering we were hoping to find a pattern that we could possibly be able to link to some changes in the games but despite a constant rise due to increased participation, fluctuations were observed. This section of the project could further be improved be looking at the clusters within each sports category and investigate whether training and nutrition led to changes in the characteristics of athletes. For better visualisation, we could use variables like 'weight' and 'height', which are more suitable for Kmean clustering. We have not been able to include how the other variables like 'medal', 'sex' and 'sport category' affect the clusters in our diagram. A more in depth-analysis in the

attributes of athletes could yield more significant insights that may help trainers and athletes improve their programmes.

Throughout this project, we have explored various aspects of the Olympic Games and how they have changed since the 1896 Athens Games to the 2016 Rio Games. We believe that the findings from each section provides key insights that could be used as a foundation for more in-depth analysis on specific areas of the Olympics for better generalization of results. Some future work, we suggest getting more data on athlete capabilities such as endurance, strength and stamina and see the impact using an interdisciplinary approach over the years.

Table of word counts

Problem statement	250
State of the art	454
Properties of the data	301
Analysis: Approach	428
Analysis: Process	1201
Analysis: Results	247
Critical reflection	540

REFERENCES

The list below provides examples of formatting references.

[1] Savić, Z. (2007). The Olympic Games as a cultural event. Acta Universitatis Palackianae Olomucensis. Gymnica, 37(3), 7–11

[2] V. Asha, S. P. Sreeja, B. Saju, C. S. Nisarga, P. N. Gowda, and A. Prasad, "Performance analysis of Olympic Games using data analytics," in *Proceedings of the Second International Conference on Electronics and Renewable Systems (ICEARS-2023)*, 2023, [ISBN: 979-8-3503-4664-0].

[3] P. H. Chowdary, V. Kaur, A. Kaur, K. Krishan, and T. Nandeesh, "From Athens to Rio: A Comprehensive Data Analysis and Visualization of 120 Years of Olympic History," in *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, Mar. 2024.

[4] V. R. Bhosale, T. Bhadrike, G. Khare, I. Darvesh, and R. Gajakos, "Olympic data analysis using machine learning," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 6, no. 4, Apr. 2024. [e-ISSN: 2582-5208].

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

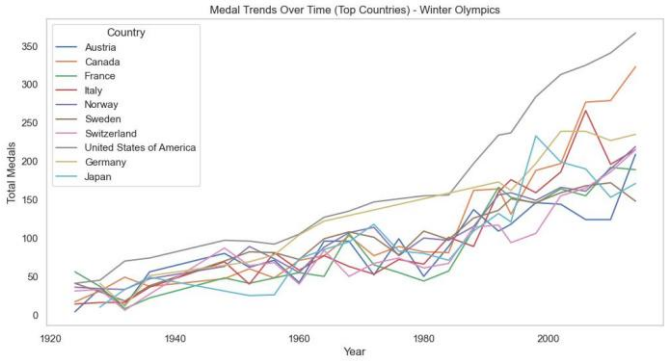
[6] J J. De Lange and J. M. Van Ree, "Endorphinergic and serotonergic mechanisms of tolerance to the behavioural effects of ketamine in rats," *European Journal of Pharmacology*, vol. 228, no. 2, pp. 265–270, 1992. [Online]. Available: https://doi.org/10.1007/BF02922904. [Accessed 30 Dec 2024]

[7] "Which Sports Have Been Dropped from the Olympics?," *World Atlas*, [Online]. Available: https://www.worldatlas.com/articles/which-sports-have-been-dropped-from-the-olympics.html. [Accessed: 30 Dec 2024].

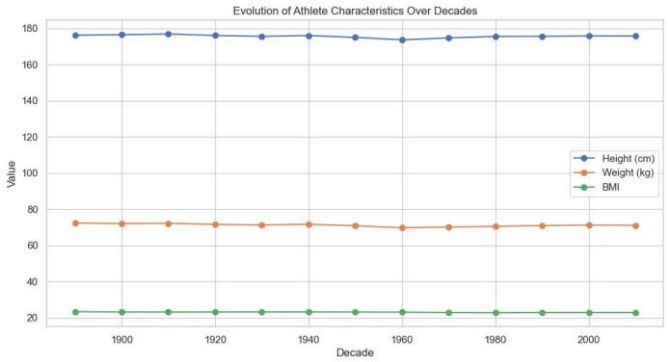
[8] Wikipedia contributors, "Olympic sports," *Wikipedia*, The Free Encyclopedia, [Online]. Available: https://en.wikipedia.org/wiki/Olympic_sports. [Accessed: 30 Dec 2024].

[9] R. Hickok, "Sports of the Olympic Games," *Topend Sports*, [Online]. Available: https://www.topendsports.com/events/summer/sports/index.htm . [Accessed: 31 Dec 2024].

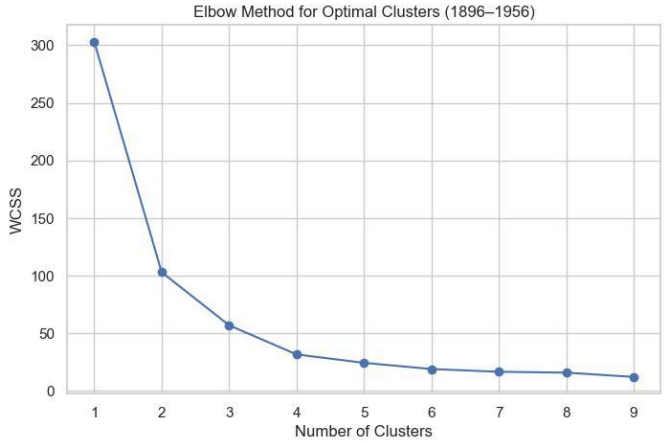
APPENDIX



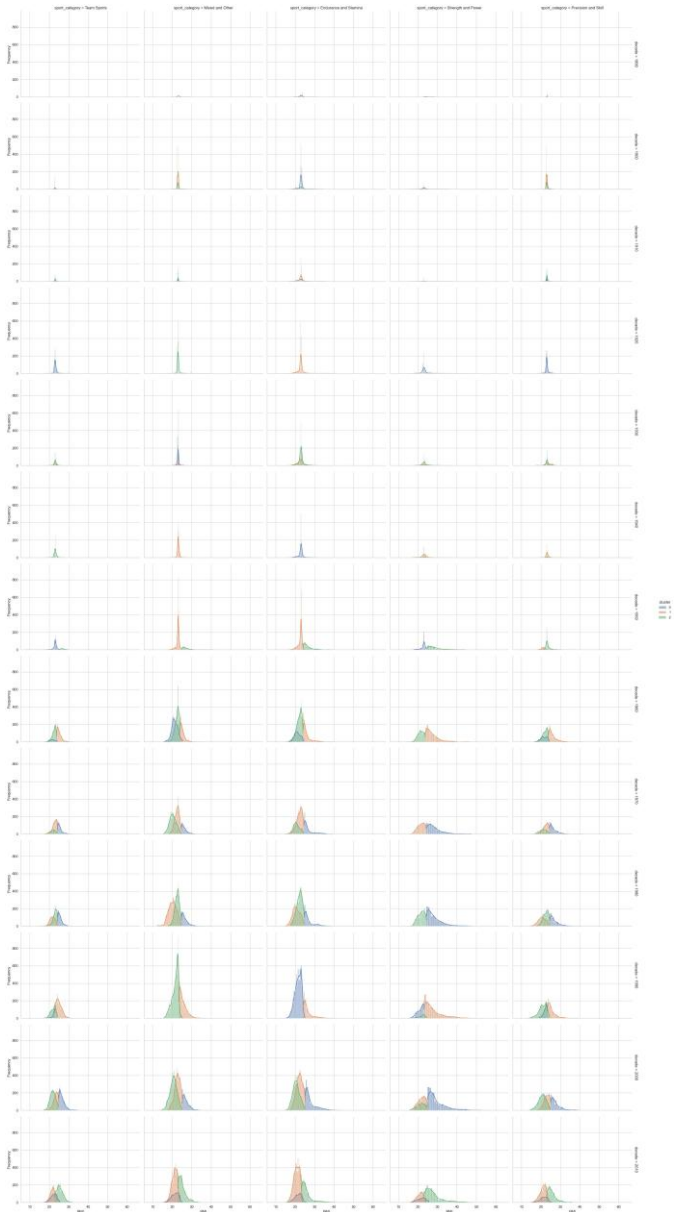
Appendix 1: Top performing countries in Winter Olympics



Appendix 2: Average height, weight and BMI of athletes for each decade



Appendix 3: Elbow plot



Appendix 4: Clustering results for each sports category and decade