WHAT MAKES A CHAMPION WIN?

A Study to Analyse Key Attributes of
Olympic Athletics and Gymnastics Champions

Final Project on MGSC 661-MultiVariate Statistics

1 Introduction

The Olympic Games stand as a symbol of global sportsmanship and athletic prowess, offering a rich compendium of data that chronicles the evolution of human physicality and performance across more than a century. This paper ventures into a thorough examination of a historical dataset that spans all the Olympic Games from Athens 1896 to Rio 2016. The goal of this study is to understand what sets Olympic winners apart and how these factors have evolved. Central to the investigation is the analysis of winners in Athletics and Gymnastics, the top two sports, by participation. By applying unsupervised learning techniques, the study seeks to validate observed insights and uncover the underlying factors contributing to an athlete's triumph at the Olympics. The adopted approach was multi-faceted: First, hierarchical clustering helps to identify athlete profiles in Athletics and Gymnastics, highlighting unique traits that match various sports events. This detailed analysis helps us understand how athletes' physical attributes and strategies have changed over time. Second, Principal Component Analysis (PCA) simplifies the complex data, revealing the key factors that contribute to Olympic success and how the characteristics of winning athletes have shifted throughout history. However, the implications of this study extend far beyond the academic realm. The insights derived offer invaluable guidance for coaches, trainers, and sports strategists within the competitive sports arena. It offers a clear, data-driven guide for developing Olympic-level athletes. By understanding and following the key physical and performance traits of successful Olympians over time, sports professionals can improve their training methods to match the qualities of top athletes.

2 Data Description

2.1 Data Source and Variable Descriptions

The foundation of the analysis is a comprehensive historical dataset that encompasses the entirety of the modern Olympic Games, ranging from Athens 1896 to Rio 2016. Sourced from Kaggle, this extensive dataset comprises over 271,000 records, with 15 diverse attributes, providing a comprehensive view of each athlete's Olympic journey. The columns are as follows:

Data Type	**Columns**	**Meaning**
Numerical	ID (identifier)	Athlete's ID
	Age	Age of Athlete
	Height	Height of Athlete – in centimeters
	Weight	Weight of Athlete - in kilograms
	Year	Year of the Olympic entry
Categorical	Name (identifier)	Athlete Name
	Sex	Sex of Athlete- M or F
	Team	Team the Athlete is Representing
	NOC	National Olympic Committee - 3 letter code
	Games	What Olympics the athlete partook in - Year & Season
	Season	Winter or Summer Olympics
	City	Host city
	Sport	What sport athlete partook in
	Event	What event athlete partook in
	Medal	What kind of Medal the athlete won - Gold, Silver, Bronze, or Null

Table 1: Olympic Games Data Dictionary

2.2 Exploratory Data Analysis

2.2.1 Rising Trends: The Growth of Participant Numbers Over Time

It was observed that there was a gradual increase in the number of participants over time. Two long periods without any Games, between 1912-1920 and 1936-1948, correspond to WWI and WWII. After the year 1994, the Summer and Winter Olympic Games were split and held during separate years, hence the graph shows different points.

Figure 1: Participants (winner and non-winner) over time

2.2.2 The Youth and Agility Trend: Younger, Lighter Participants Lead the Way

Winners of Olympics are typically young, with a concentration in a specific range of height and weight. The age distribution is right-skewed, suggesting winners are predominantly in their 20s, while height and weight distributions are roughly normal, indicating common physical profiles among champions whereas for winners of Gymnastics, there is a peak at a shorter height.

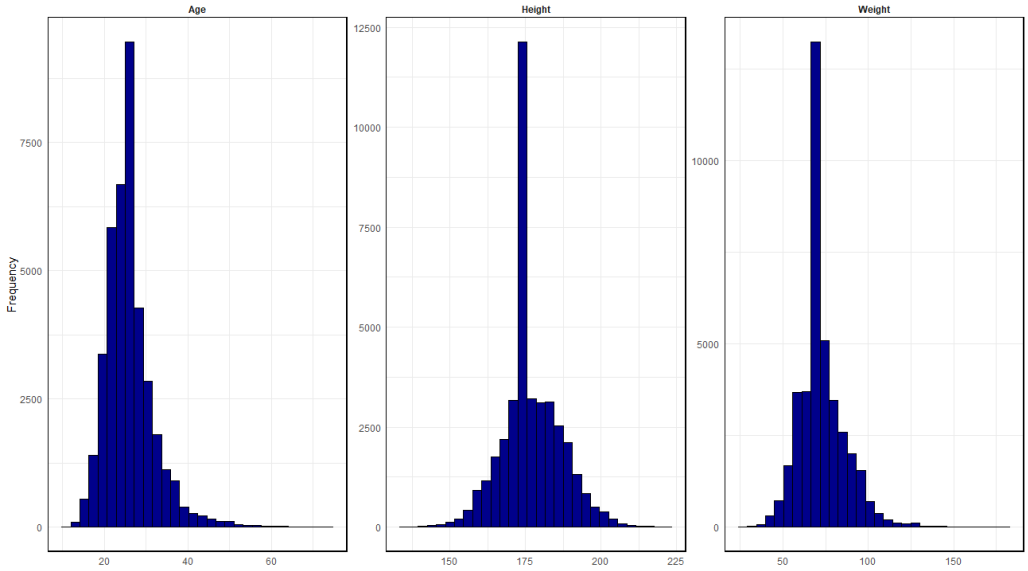

Figure 2: Distribution of numerical variables - Age, Height, and Weight.

2.2.3 Recent Gymnastics Dominance: The Ascent of Romania and China

Countries like the USA, Germany, and the UK have had high dominance in this sport. Russia experienced notable variations due to the dissolution of the Soviet Union, leading to fluctuations in its athletic performance, especially in the early 1990s. China has shown a marked increase in gymnastic medals, reflecting its growing focus on international sports. Romania's notable peak in gymnastics is attributed to factors like a successful cohort of gymnasts and enhancements in coaching and training.

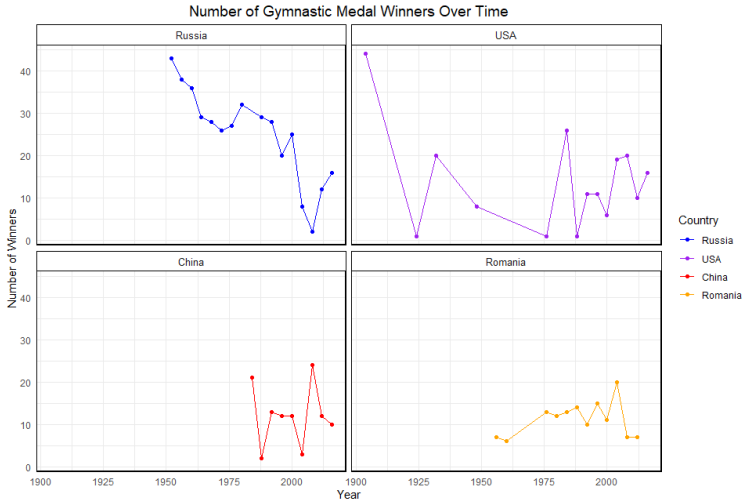

Figure 3: Winners of Gymnastics from Russia, USA, China, and Romania over time

2.2.4 Evolution of Physical Traits in Athletics Champions Through the Years

The evolution in athletes' physical profiles over the last century shows early preferences for taller champions, shifting recently to shorter athletes, possibly due to changing sports types or training focusing on agility. There's also a move from heavier to lighter athletes, reflecting a shift in importance from strength to speed and flexibility. Additionally, the rising average age of athletes suggests a growing emphasis on experience and maturity, likely aided by advances in sports medicine and training. This study will further explore these trends.

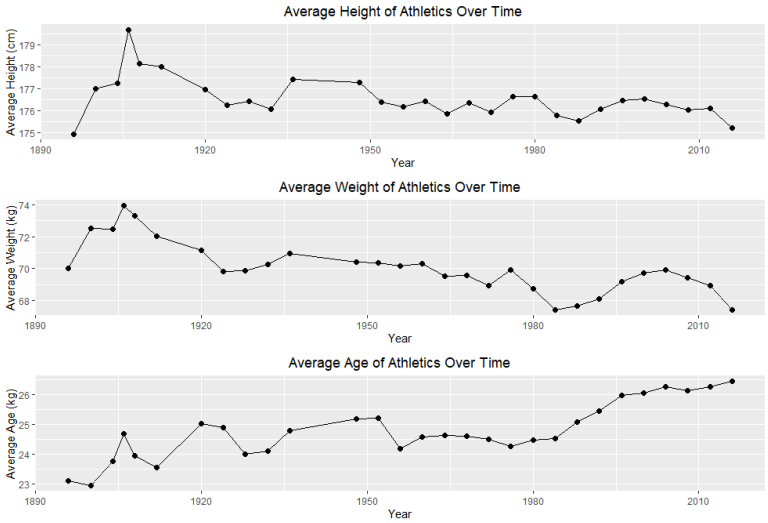

Figure 4: Change in physical attributes among winners of Athletics over time

2.3 Data Preprocessing

The Olympic dataset's preprocessing involved multiple steps for analysis readiness. Identifiers were removed, and rows lacking all key attributes (height, weight, medal status) were discarded. Missing values in age, height, and weight were filled with mean values. The 'Medal' column was encoded to differentiate between medal winners ("Winner" variable set to 1) and non-winners (0), and then dropped. Duplicate entries were removed for data uniqueness. Age, height, and weight were standardized to integer format. The 'Season' variable was numerically encoded, and a new 'Country' column was created from an external dataset mapping regions to NOCs. Age, height, and weight were normalized to ensure equal feature weighting in the analysis. The dataset was then filtered specifically for athletics (Athletics and Gymnastics) for detailed analysis.

3 Model Selection & Methodology

3.1 Hierarchical Clustering

3.1.1 Data Sampling

The study commenced by focusing on a subset of data specifically related to Athletics from the preprocessed dataset, while the cluster analysis was conducted on the entire dataset for Gymnastics. This approach allowed for a manageable yet representative dataset for each sport.

3.2 Clustering Method

Hierarchical clustering was chosen for this study primarily due to its effectiveness in handling mixed data types, which included both categorical and numerical variables. This method does not require the number of clusters to be pre-specified, allowing for a more natural and intuitive grouping based on the data. The dendrogram produced by hierarchical clustering makes it an ideal choice for exploratory analysis in sports analytic.

The Gower distance, a measure ideal for mixed data types, was calculated for both subsets using the `daisy` function from the `cluster` package in R. This distance metric considers the different types of variables and scales them appropriately, providing a robust choice for datasets with varied attributes. For the clustering process, the complete linkage method was employed, implemented via the `hclust` function in R. The dendrogram tree was cut to form two clusters for each sport using the `cutree` function. This division facilitated the classification of athletes into two distinct groups, as it was based on their characteristics at the point where the tree's length was the longest compared to subsequent branches (Dendogram in Appendix 6.3). The number of clusters, two in this case, was chosen to simplify the analysis and to focus on the most significant differences within each sport.

3.3 Principal Component Analysis (PCA)

PCA was employed to reduce the dimensionality of the athletics data while retaining the most informative features contributing to variance within the dataset. The aim was to uncover underlying variables that explain the success of athletes in the sports of athletics and gymnastics. A subset of the preprocessed dataset was created specifically for PCA, comprising only numerical variables.

4 Results and Classification

4.1 Hierarchical Clustering

4.1.1 Athletics

Hierarchical clustering resulted in two distinct clusters that have helped to derive some interesting insights from Athletics winners. Cluster 1 is distinguished by athletes who are generally taller and heavier. The physical attributes of these athletes suggest a propensity towards sports where height and weight are advantageous, such as field events (like javelin, shot put, discus throw) or sprinting events where power and speed are crucial. This cluster also shows a strong historical presence across a wide range of Olympic years, indicating a consistent representation of this athlete profile over time. The standard deviations for age, height, and weight are relatively moderate, suggesting some diversity within the group, but still within a range that highlights their physical prowess. The country representation in this cluster is noteworthy, with a significant presence of athletes from countries with long-standing and well-established athletics programs like the USA, Germany, and Russia, which have historically invested heavily in these sports.

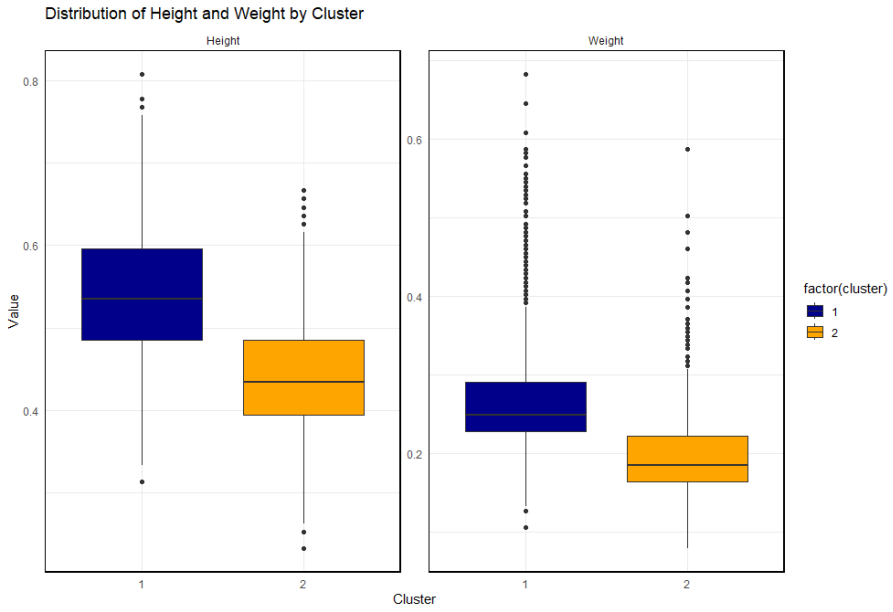

Figure 5: Variation of physical features of Athletics winners in 2 clusters

Cluster 2, on the other hand, comprises athletes who are younger on average, shorter, and lighter. These characteristics are indicative of athletes who may excel in endurance-based events like long-distance running or events that require agility and precision, rather than brute strength. The standard deviations for their physical attributes are slightly lower, pointing to a more homogeneous group in terms of physicality. This cluster's prominence in more recent Olympic years (notably in 2012, 2008, and 2016) reflects evolving training techniques and a shift in the athletic events that gained prominence. The country distribution in this cluster is also diverse but with notable mentions of countries like Kenya and Jamaica, which are renowned for their excellence in endurance running and sprints, respectively. This cluster represents a different aspect of athletics, where speed endurance, and agility are more significant than physical size.

4.1.2 Gymnastics

Hierarchical clustering also resulted in two distinct clusters that have helped to derive some interesting insights from Gymnastic winners. In terms of physical attributes, in Cluster 1 the height and weight of gymnasts are higher than that of Cluster 2 with low variation indicating consistency. Athletes in this cluster exhibit an average age slightly higher than that of Cluster 2. Notably, the standard deviation for age in this cluster is relatively low, indicating a narrower age range among the athletes. One striking feature of Cluster 1 is the significant presence of athletes from various countries, with a diverse distribution across multiple nations. This diversity suggests that Cluster 1 comprises gymnasts from different parts of the world, each contributing to their unique gymnastic styles and techniques.

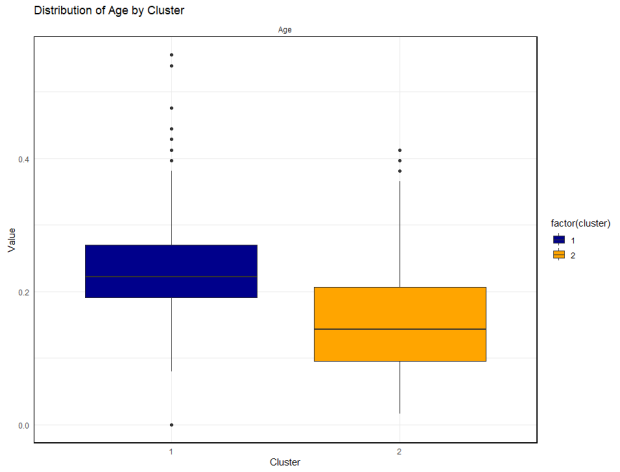

Figure 6: Variation of physical features of Gymnastics winners in 2 clusters

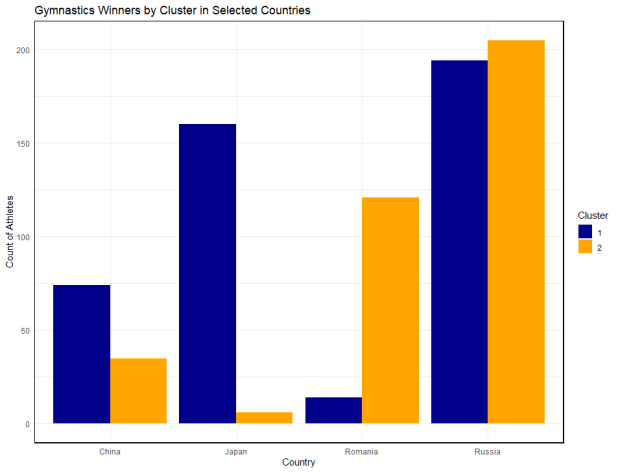

Figure 7: Variation in Nationalities of Gymnastics winners in 2 clusters

In contrast to Cluster 1, Cluster 2 Gymnastics winners show distinct characteristics. Gymnasts in this cluster are notably younger on average compared to Cluster 1. Despite their youth, Cluster 2 gymnasts demonstrate remarkable skills and achievements. Their average height and weight are lower than that of Cluster 1. Interestingly, Cluster 2 exhibits higher variance indicating a wider range of physical attributes among the athletes. One of the most significant differentiating factors of Cluster 2 is the concentration of athletes from specific countries, particularly with a notable presence from a single country or a small group of nations. This suggests a more focused and possibly specialized approach to gymnastics training within this cluster. The younger age and higher variability in physical attributes indicate that Cluster 2 represents a group of emerging gymnastic talents who have shown exceptional promise at a relatively early stage in their careers. Overall, Cluster 2 appears to be characterized by a unique combination of youth, potential, and specialization, which sets it apart from Cluster 1 in the realm of gymnastics winners.

4.2 Principal Component Analysis

4.2.1 Athletics

PCA has provided a valuable reduction of the dataset, highlighting the key factors that contribute to Olympic success in athletics. The analysis suggests that both temporal and physiological factors play a significant role in Olympic success in athletics. Weight and height show a justifiable correlation. The combination of 'Age', 'Sex', 'Height', and 'Weight' in PC2 suggests that an athlete's physical and demographic profile is crucial in winning medals. If taller and heavier athletes are to one side of PC2, implies that events where such physical characteristics are advantageous (like throwing events) are clustered there. Conversely, events, where a lighter build is advantageous (like long-distance running), might be on the opposite side. The influence of the 'Year' variable on the second principal component (PC2) shows a strong loading for 'Year', which indicates a temporal trend in the data. This suggests

that the profile of winning athletes has evolved over the years. Possible interpretations could be advances in training techniques, changes in the selection of athletes, or the introduction of new events that favor different athlete profiles. This is consistent with the findings from Hierarchical clustering.

Figure 8: PCA on Athletics

4.2.2 Gymnastics

Gymnastics at the Olympic level requires a unique blend of strength, flexibility, and technical skill. The positive correlation of 'Height' and 'Weight' along PC2 indicates that physical attributes are significant. In gymnastics, where different events require different physical advantages, this pattern could mean that winners in certain types of events (e.g., vault or rings, which might favor a certain body type) are influencing this principal component. The orientation of 'Year' along PC1 implies that the physical characteristics of gymnastics winners have evolved. A trend toward winners with different body compositions in more recent years reflects changes in the sport's competitive landscape, training methods, or the introduction of new disciplines within gymnastics. This validates the findings from exploratory data analysis. The patterns are indicative of the sport's development over time. They reflect how gymnastics training, selection, and performance might have adapted to the evolving demands of the sport.

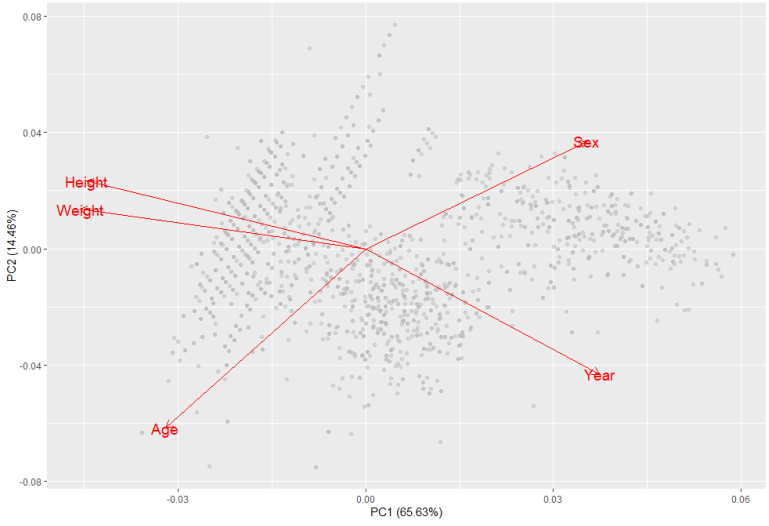

Figure 9: PCA on Gymnastics

5 Implication and Conclusion

The study's findings, using hierarchical clustering and PCA, reveal significant insights about the physical and performance characteristics of Olympic athletes in Athletics and Gymnastics which can be highly beneficial for coaches and managers of teams to strategically plan for enhancing performance.

For Athletics sport, the data enables the creation of tailored training programs. For athletes in the taller and heavier cluster, training can focus on power and speed, crucial for field events and sprints. Conversely, shorter and lighter athletes would benefit from training emphasizing endurance, agility, and precision. This information is also invaluable in talent identification and recruitment. Teams can focus on recruiting athletes with physical attributes suited to specific events. Adapting to temporal trends is crucial, as the 'Year' variable suggests that the profile of successful athletes evolves over time. Staying informed about these trends is essential for adapting training and recruitment strategies. Investment and resource allocation decisions can be informed by understanding which events align with the athletes' profiles. This guides the development of facilities, equipment, and training staff.

For Gymnastics, the data can guide the identification and development of young gymnasts. The younger age profile in Cluster 2 indicates a pool of emerging talents. Early identification of these athletes and nurturing their skills can be a strategic focus. This might include scouting for gymnasts with the potential to excel in specific events based on their physical attributes. Based on the clusters, teams can strategically decide which gymnastic events to focus on in competitions. Athletes from Cluster 1 might be steered towards events where their physical attributes give them an edge, while Cluster 2 athletes might be encouraged to participate in events that suit their agility and precision.

Additionally, the evolving nature of these sports, as revealed by the PCA, suggests that training methods and strategies should be dynamic, adapting to the changing landscape of Olympic competition. Coaches and managers should be aware of these trends and be prepared to evolve their approaches accordingly. This might include investing in new training techniques, focusing on different event types where their athletes have an advantage, and continuously monitoring changes in the sport to stay ahead. In conclusion, these findings provide a strategic roadmap for Olympic preparation and success, emphasizing the importance of aligning athlete development with the evolving profiles of successful Olympians.

6 Appendix

6.1 Main Data Source:

Kaggle Dataset - 120 Years of Olympic History: Athletes and Results

6.2 Data Source for Country:

Github Dataset - NOC with corresponding Regions/Country

6.3 Top sports are Athletics and Gymnastics

Since the top sports are Athletics and Gymnastics, these were chosen for the models in this paper

Sport	Count	%
Athletics	33238	15.1
Gymnastics	19714	8.93
Swimming	19640	8.90
Rowing	8755	3.97
Cycling	8372	3.79

Table 2: Distribution of Athletes in Top Sports

6.4 Welch Two Sample t-test Results on Athletics Hierarchical Clusters indicate that the differences in physical attributes are significant

Metric	Value
Data	Age by cluster
t-value	22.385
Degrees of Freedom (df)	1225.2
p-value	$< 2.2e-16$
Alternative Hypothesis	True difference in means is not equal to 0
95% Confidence Interval	[0.06753018, 0.08050461]

Metric	Value
Data	Height by cluster
t-value	24.695
Degrees of Freedom (df)	882.45
p-value	$< 2.2e-16$
Alternative Hypothesis	True difference in means is not equal to 0
95% Confidence Interval	[0.09882036, 0.11588385]

Metric	Value
Data	Weight by cluster
t-value	28.712
Degrees of Freedom (df)	833.24
p-value	$< 2.2e-16$
Alternative Hypothesis	True difference in means is not equal to 0
95% Confidence Interval	[0.06753111, 0.07744168]

6.5 Dendogram of Hierarchical Clustering

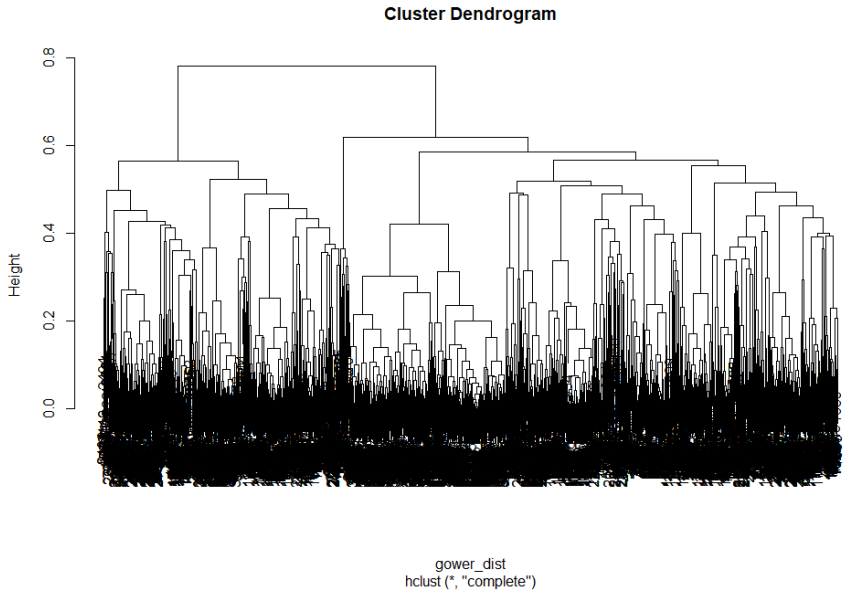

Figure 10: Dendogram of Athletics Winners

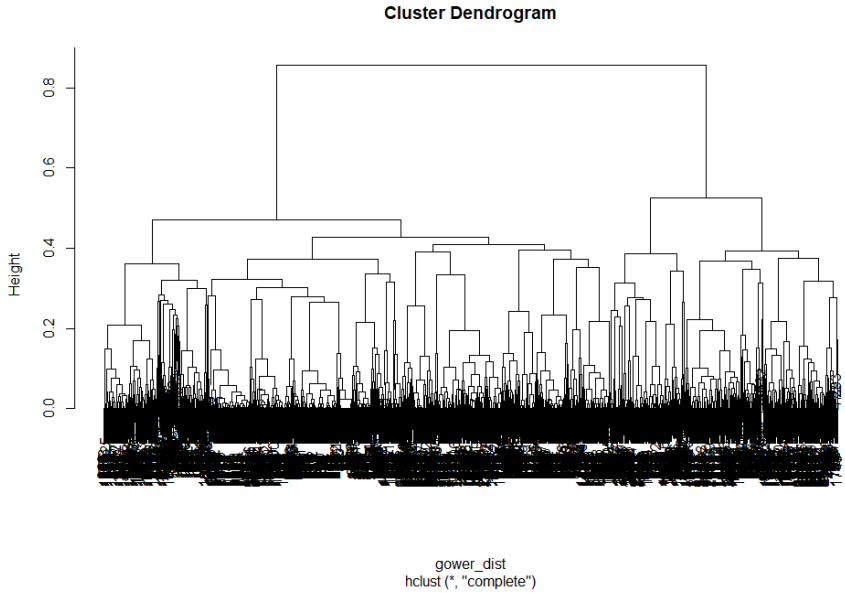

Figure 11: Dendogram of Gymnastics Winners