McGill

MGSC 661

FINAL PROJECT

JOSHUA POOZHIKALA

261184574

December 12th 2023

Introduction

The Olympics are a globally significant event that transcends cultural boundaries, uniting athletes worldwide in a celebration of diversity and unity. Beyond its cultural importance, the Games provide a platform for athletic expression and is a place where body types and compositions can be seen in all shapes and sizes. While both are exceptional athletes, a powerlifter's height and weight differ significantly from those of a swimmer. The Olympics serve as a unique platform where all forms of physical excellence are celebrated equally, a distinction that few places in the world can claim. The objective of this predictive analysis task is to identify different clusters of athletes based on body composition specifically focusing on weight and height. If effective, the output of this model can have wide-ranging applications in the sports industry, offering practical insights for optimizing athletic performance, health management, and strategic decision-making. These insights could be instrumental in enhancing training regimes, designing specialized equipment and facilities, and informing policy decisions for sports federations and committees. After identifying distinct clustering patterns in the dataset, a logistic regression classification model will be developed to predict an athlete's success based solely on their weight and height. This classification task aims to determine if an athlete's success can be predicted irrespective of their chosen sport, focusing exclusively on their body composition.

Data Description

About the Dataset

The dataset used within the analysis highlights historical information on all the Olympic Games from Athens 1896 to Rio 2016. The dataset contains a total of 271116 rows and 15 columns. Each row corresponds to an individual athlete that competed in an Olympic Event. The target variable within the dataset is Medal which refers to the placing of the athlete being either Gold, Silver Bronze or NA referring to no placing at all. Refer to Appendix 1 for a more complete understanding of the variables held within the dataset.

Exploratory Data Analysis (EDA) and Data Preprocessing

To further understand the dataset, exploratory data analysis and data preprocessing were undertaken as foundational steps for developing both clustering and predictive classification models. The process began with an examination of the dataset's dimensions and structure. Subsequently, data transformation steps were implemented: weight and height were converted to integers, and the sex column was factorized to ensure its treatment as a categorical variable in the models. A significant aspect of preprocessing involved handling missing values. For the 'Age' variable, missing entries were replaced with the dataset's mean age. Additionally, rows with missing 'Height' or 'Weight' data were removed to maintain data integrity. A crucial step was the redefinition of the target variable 'Medal'. It was transformed into a categorical variable, with the introduction of a new category named 'No Medal'. This allowed for the replacement of NA entries in the dataset with 'No Medal', thereby preparing the data for more effective modeling.

EDA steps were taken to understand the relationships between variables and the overall behavior of data within the dataset. An understanding of the distribution of gender within the dataset was key. By looking at the number of unique male and female participants Males almost double females in terms of participation in the Olympics (See Appendix 2). This can be attributed to female participation in the Olympics only starting in 1900 (Wikipedia contributors,2023). as well as having women being the lower percentage of participants in Olympic games every year with 2020 having the largest percentage of women with 48% (International Olympic Committee.n.d.). Another exploratory step undertaken involved plotting the heights and weights of athletes from instances in the dataset (see Appendix 4). Creating these plots facilitated the identification of the median weight and height of the athletes and provided insights into the behavior of outliers at both the lower and upper ends. For height the main bulk of athletes were around 160 -190 cm. While for weight, athletes were typically found within the range of 50kg to 100kg. This wide range of data can be explained by the various sports held within the Olympics that require a variety of body types.

Model Selection and Methodology

For an effective data analysis of the dataset, two modeling techniques were employed. The first technique involved the use of clustering to group athletes based on their height and weight. The second technique was the development of a classification model, designed to predict whether an athlete would receive a medal, based solely on their body composition.

Clustering Model

Initially the creation of the clustering model started with the separation of data into male and female subsets. The elbow method was applied to each subset to determine the optimal K value for each clustering model. The Elbow Method is a heuristic used in determining the number of clusters in a dataset. This method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. When looking at a plot the optimal number of clusters is chosen at the point where the rate of decrease sharply changes. This indicates a balance between maximizing the number of clusters and minimizing the total within-cluster variance. Appendix 5 highlights the elbow graphs for each subset of data both elbow graphs identify a range between 3 and 4 clusters to be the optimal amount for the clustering task. A cluster task was performed on both sets of data however the large dimensionality of the dataset made interpretation of the results difficult (See Appendix 6 & 7). As a result of this it was determined that a better option would be to create subsets of male and female athletes in both summer and winter Olympics and performing clustering on all 4 of the subsets of data. Again, the Elbow method was used to determine the optimal amount of clustering to be used within the model for each respective subset of data (See Appendix 8)

Logistic Regression Classification Model

To develop a logistic regression model that classifies an athlete of winning a medal or not based on factors such as sex weight height and age a few steps were taken. In addition to the data preprocessing steps described earlier, the construction of the classification model also involved dividing the dataset into two distinct subsets: a training set, which constituted 70% of the data, and a testing set, which accounted for the remaining 30%. This split was crucial for training the model and subsequently evaluating its performance. This partitioning was essential for validating the model's performance on unseen data. A logistic regression model was trained using the training set with 'Age', 'Sex', 'Height', and 'Weight' as predictor variables, and the binary outcome of medal status was modeled using a binomial family in the glm function. Predictive accuracy was subsequently evaluated on the test set, with a threshold of 0.5 used to classify predictions. A confusion matrix was generated to provide a clear visualization of the model's performance, and overall accuracy was calculated to quantify the model's effectiveness in predicting an athlete's success in winning a medal.

Results

Clustering Model

The updated clustering models pertaining to each gender and season can be seen in Appendix 9. The results of the clustering analysis can be seen in Appendix 10. Although the dimensionality of the dataset still makes the clusters difficult to interpret the split of the entire dataset into 4 different categories allows for some interpretations to be made. The K-means clustering analysis of Olympic athletes by height and weight across summer and winter sports reveals distinct physical profiles that correlate with the athletes' sports disciplines. For summer athletes, Cluster 1 predominantly consists of gymnasts and track athletes, suggesting a leaner physique conducive to agility and endurance. Cluster 2 includes a more varied set of body types with sports like rowing and basketball, indicating a combination of height and muscular build. Cluster 3 is characterized by taller athletes in sports such as swimming and athletics, where both speed and endurance are key. In the winter sports category, Cluster 1 is dominated by endurance sports such as cross-country skiing, where a lighter build is beneficial. Cluster 2 shows a prevalence of sports like ice hockey and bobsleigh, which favor more muscular and heavier builds, likely due to the strength and stability required. Finally, Cluster 3 for winter athletes also tends towards taller individuals with a balance of strength and agility, suitable for sports like alpine skiing. The scatter plots visually confirm these groupings, demonstrating how K-means clustering can effectively distinguish between the various physical demands of different Olympic sports. An interesting thing that should be noted is that the clusters for both genders are similar highlighting that the weight and height plays a similar role in each sport for both genders.

Logistic Regression Classification Model

The logistic regression model developed to predict whether an athlete would receive a medal at the Olympic Games, based on their age, sex, height, and weight, was evaluated for its performance using several metrics.

The model achieved an overall accuracy of 85.39%. This accuracy level was consistent with the No Information Rate, and the p-value indicated that the accuracy was not significantly better than what would be expected by random chance (p-value > 0.05). The Kappa statistic was essentially zero, suggesting that the agreement between the predictions and the actual values was no better than random.

The confusion matrix revealed that the model predicted 'no medal' (class 0) correctly for 52,989 cases, but it failed to accurately predict a single case of 'medal won' (class 1), as indicated by the zero count for true positives. Conversely, the model had only two instances where it incorrectly predicted that an athlete would win a medal when they did not. The high number of false negatives (9,065) resulted in a recall (sensitivity) of 100% but a specificity of 0%, meaning that while the model captured all the true 'no medal' cases, it failed to capture any 'medal won' cases correctly.

The Positive Predictive Value (PPV) or precision was 85.39%, meaning that when a 'no medal' prediction was made, it was correct 85.39% of the time. However, the Negative Predictive Value (NPV) was 0%, indicating the model failed to correctly predict any true 'medal won' instances.

The prevalence of 'no medal' was 85.39%, which was equal to the detection rate due to the model predicting 'no medal' for all cases. The Detection Prevalence was 100%, reflecting the fact that all predictions were for 'no medal'.

The Balanced Accuracy, which is the average of sensitivity and specificity, was 50%. This, combined with the other metrics, highlighted significant issues with the model's predictive performance. Despite high accuracy, the model's inability to correctly predict 'medal won' cases meant that it was essentially functioning as a naive classifier, predicting 'no medal' in almost all cases.

The statistically significant McNemar's Test (p-value < 2e-16) indicated that there was a significant difference between the number of false negatives and false positives, further emphasizing the model's bias towards predicting 'no medal'.

In conclusion, while the model appeared to have high accuracy and precision, it performed no better than random chance when it came to distinguishing between medal winners and non-winners. It was evident that the model's predictions were heavily skewed towards one class, which raises concerns about its practical utility.

Classification/Predictions and Conclusions

Overall while clustering tasks and logistic regression classification tasks were performed their performance and capabilities were limited. The K-means clustering model, was able to cluster instances based on weight and height requirements for some sports. For example, sports such as basketball and volleyball were grouped together in the same cluster which makes sense as both sports require taller and leaner individuals. The clusters effectively split the data into groups that shared similar body types and compositions.

The logistic regression model on the other hand was less successful. While the model was able to effectively predict when no medal would be won by an athlete it was no better than random chance at predicting when a medal would be won. In the future there are a few methods that can be done to improve the model's ability to distinguish between medal winners and non-winners based on body composition.

Addressing class imbalance through resampling methods, such as oversampling the minority class or under sampling the majority class, could provide a more balanced dataset for the model to learn from. Alternatively, applying advanced algorithms like SMOTE (Synthetic Minority Over-sampling Technique) can generate new, synthetic examples of the minority class to enhance the training process.

Feature engineering could also play a critical role in enhancing model performance. By incorporating additional variables that highlight body composition factors that may influence an athlete's likelihood of winning a medal —such as VO2 Max Score, Body fat Percentage, muscle density, or psychological factors—the model would have access to more nuanced information that might improve its predictive accuracy.

Switching to more complex models that can capture non-linear patterns and interactions between variables might also be beneficial. Techniques such as decision trees, random forests, or gradient boosting machines can often model complex relationships more effectively than logistic regression. However, these methods would require data that is more nuanced and specific.

Regularization methods like ridge or lasso regression can help prevent overfitting by penalizing the model for excessive complexity. This approach can encourage the model to focus on the most informative features, potentially improving its generalization to new data.

Finally, Cross-validation techniques could be more effectively employed to robustly assess the model's performance, and a careful re-examination of the data preprocessing steps could be undertaken to ensure that they do not introduce unintended biases. By implementing these changes, the predictive power of the model stands a better chance of improving, providing a more accurate tool for forecasting an athlete's potential to secure a medal based on there body composition.

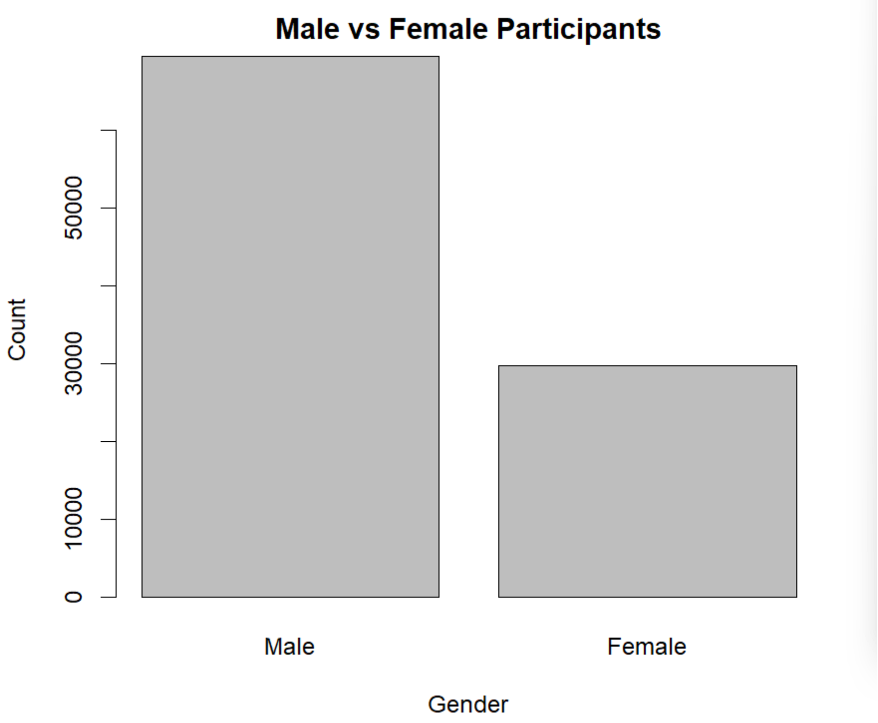
Appendix

Appendix 1: Dataset Variable Description

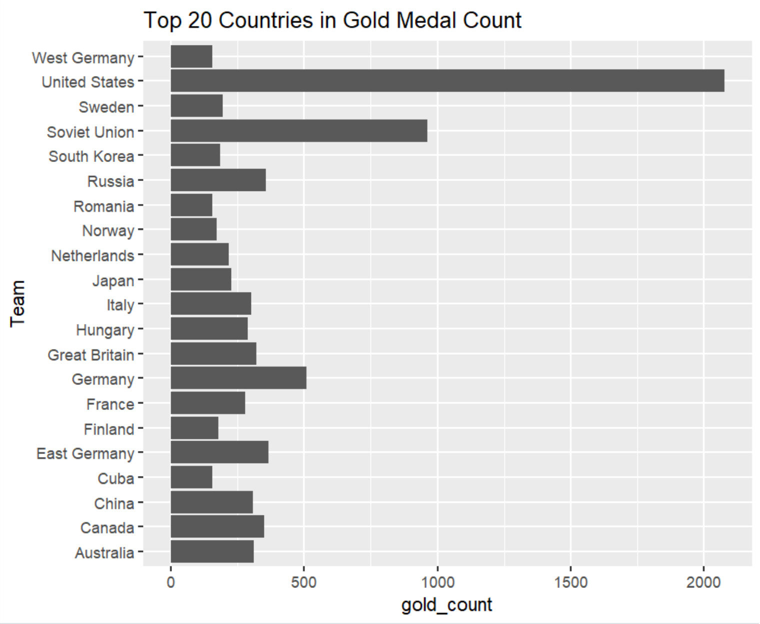
Variable	Description

ID	Unique number for each athlete
Name	Athlete’s First and Last Name
Sex	Male or Female (M or F)
Age	Integer
Height	In Centimetres
Weight	In Kilograms
NOC	National Olympic Committee 3- letter code (National Olympic Committee Athlete is Competing For)
Team	Country the Athlete is Competing for
Games	Year and Season that athlete competed in
Year	Integer
Season	Summer or Winter
City	Host City
Sport	
Event	
Medal	Gold, Silver Bronze, or NA

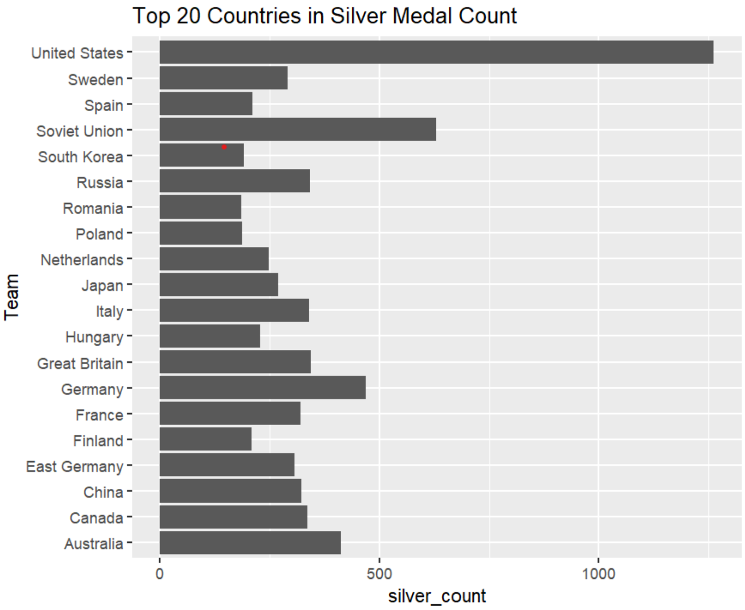
Appendix 2: Male vs Female Participation in Olympics

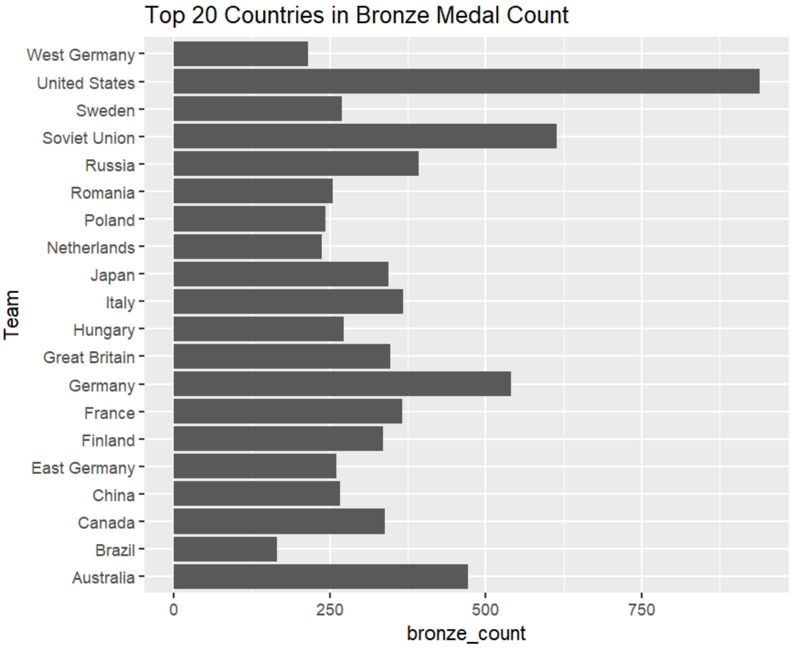


Appendix 3: Top 20 Countries in Medal Count- Gold, Silver and Bronze

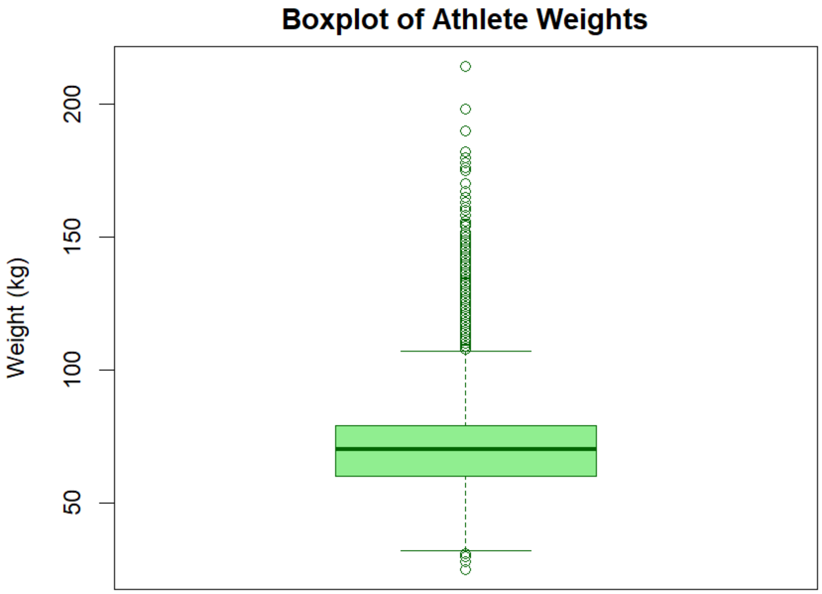


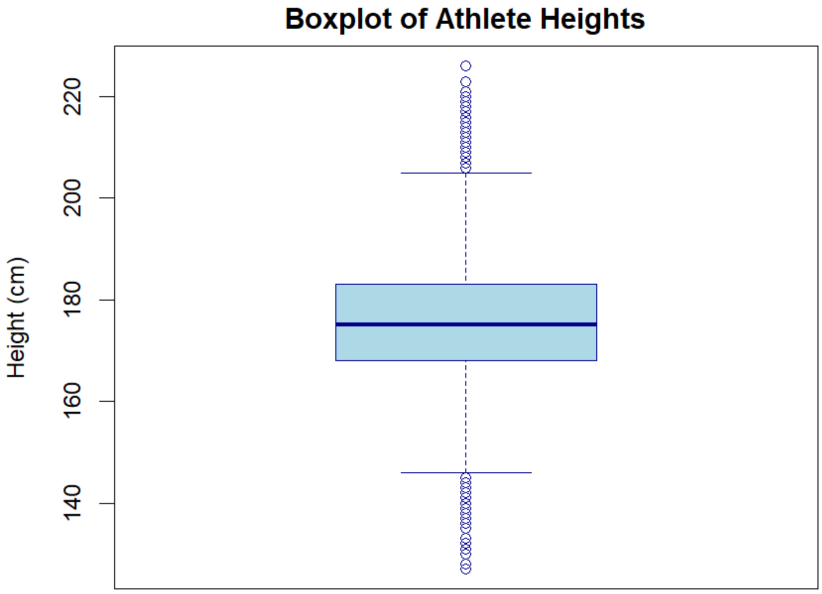
Appendix 3 Cont.: Top 20 Countries in Medal Count- Gold, Silver and Bronze



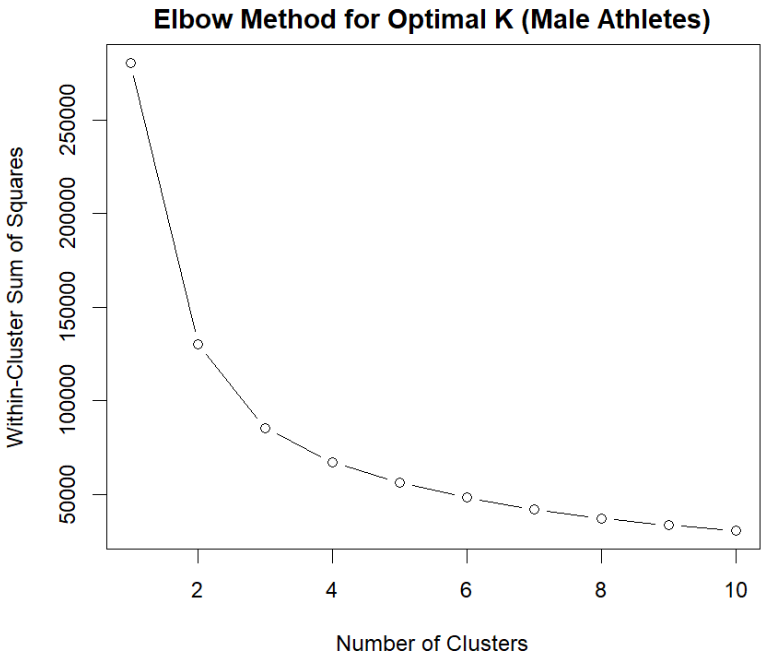


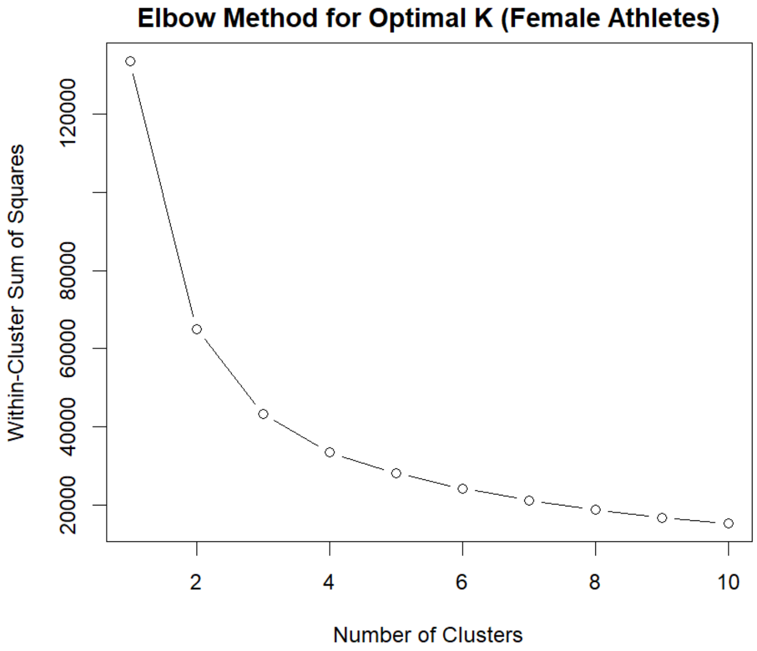
Appendix 4: Boxplot of Athlete Weights and Heights



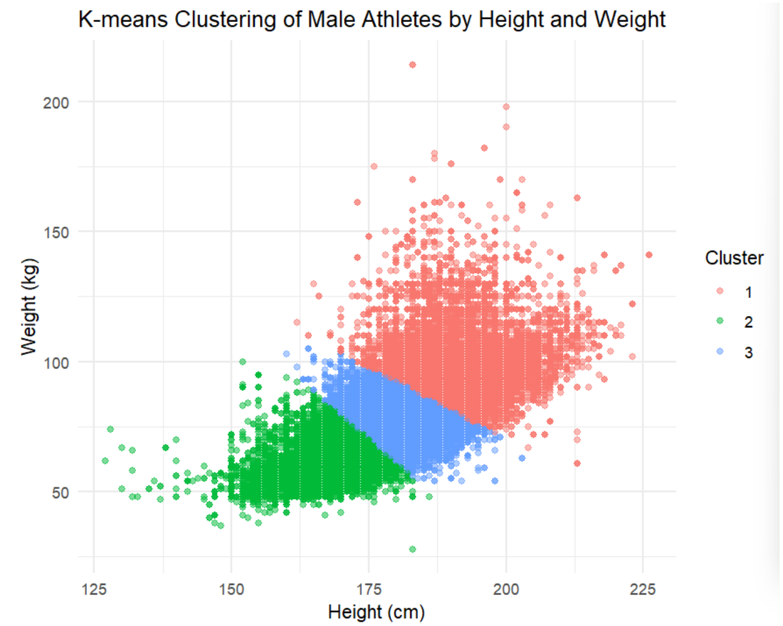


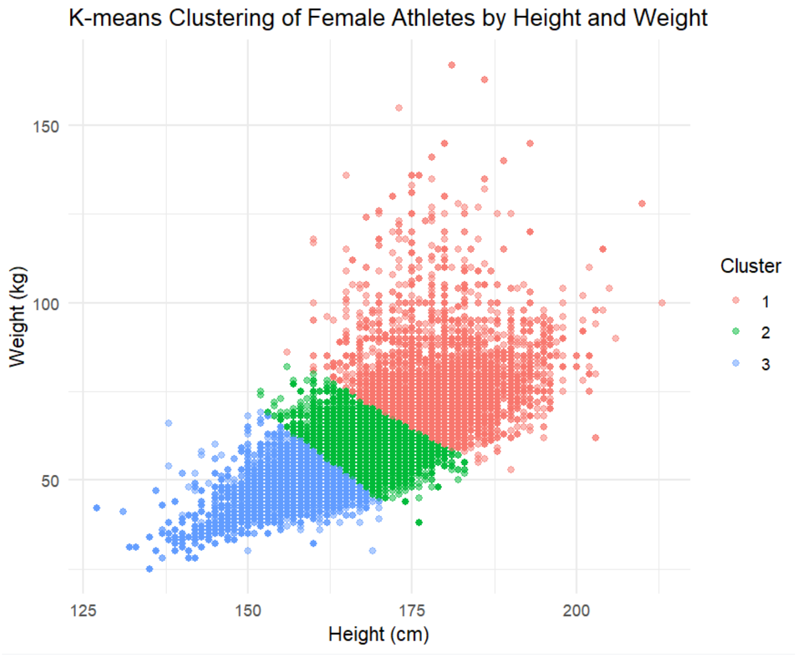
Appendix 5: Elbow Method for Optimal K Male and Female- Version 1


Elbow Method for Optimal K (Male Athletes)
Within-Cluster Sum of Squares
50000
100000
150000
200000
250000
2
4
6
8
10
Number of Clusters

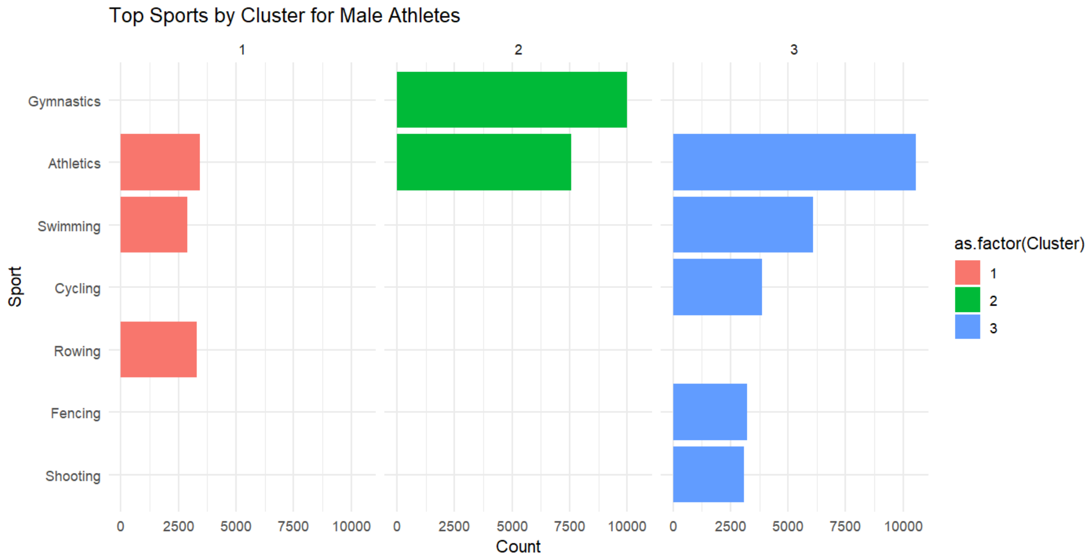

Elbow Method for Optimal K (Female Athletes)
Within-Cluster Sum of Squares
20000
40000
60000
80000
120000
2
4
6
8
10
Number of Clusters

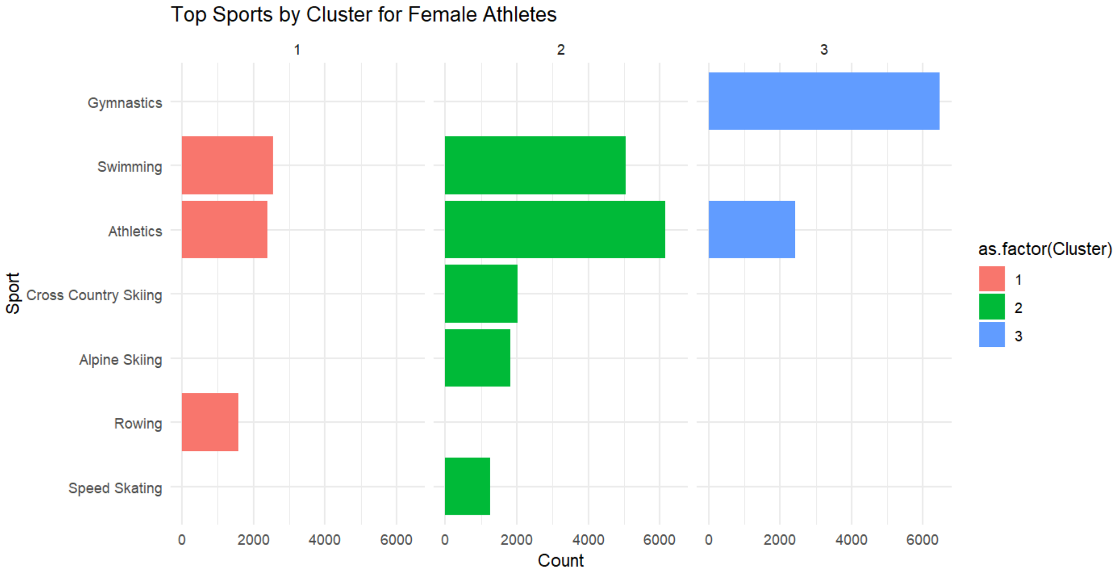
Appendix 6: K-Means Clustering Results for Female and Male Athletes by Height and Weight- Version 1


K-means Clustering of Male Athletes by Height and Weight

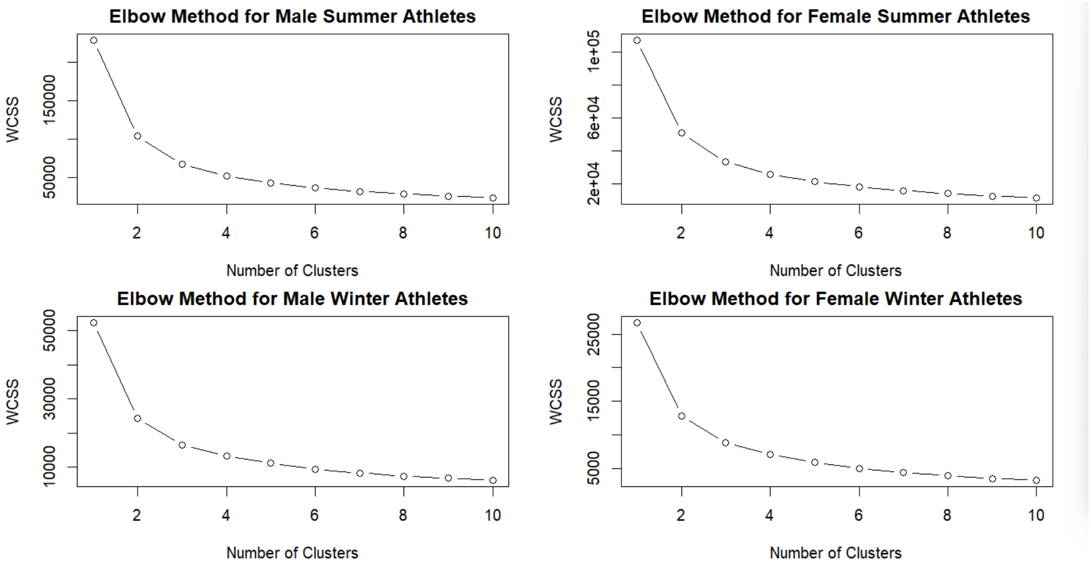

K-means Clustering of Female Athletes by Height and Weight

Appendix 7: Interpretation of Clustering Results-Version 1

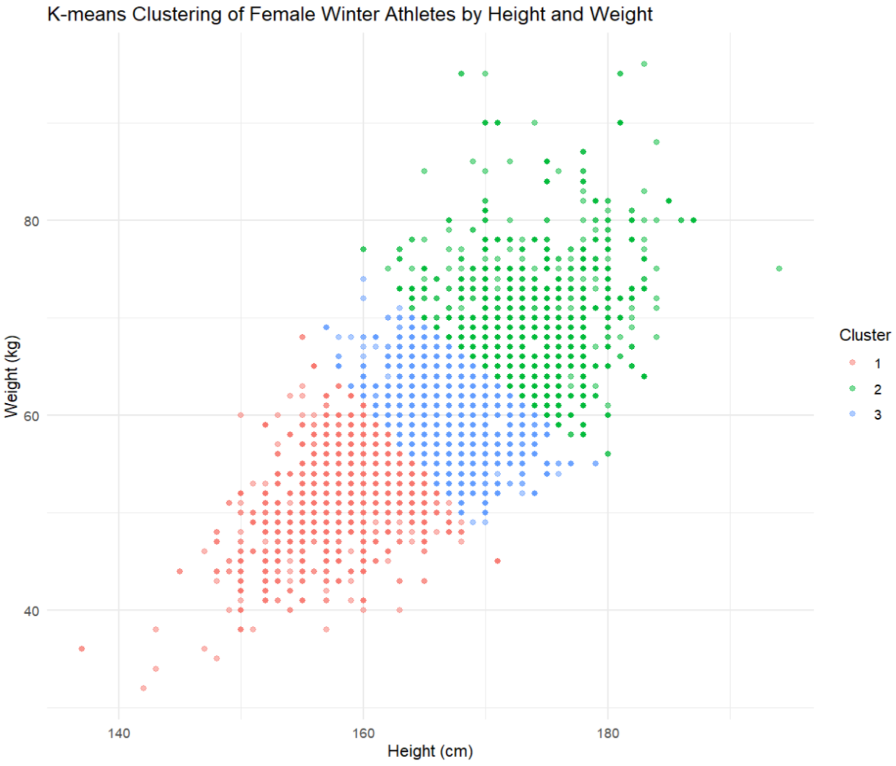




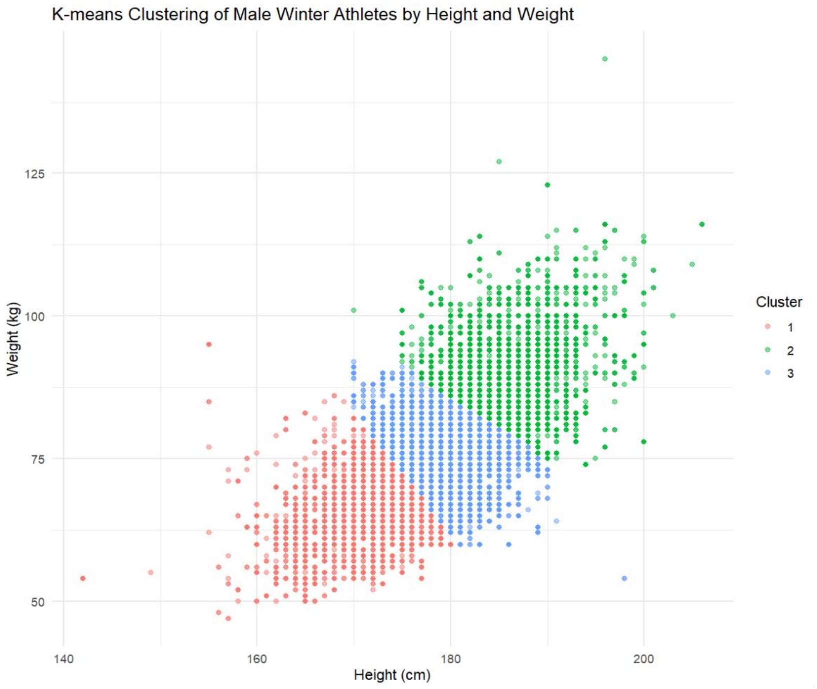
Appendix 8: Elbow Method for Male and Female Summer and Winter Athletes- Version 2

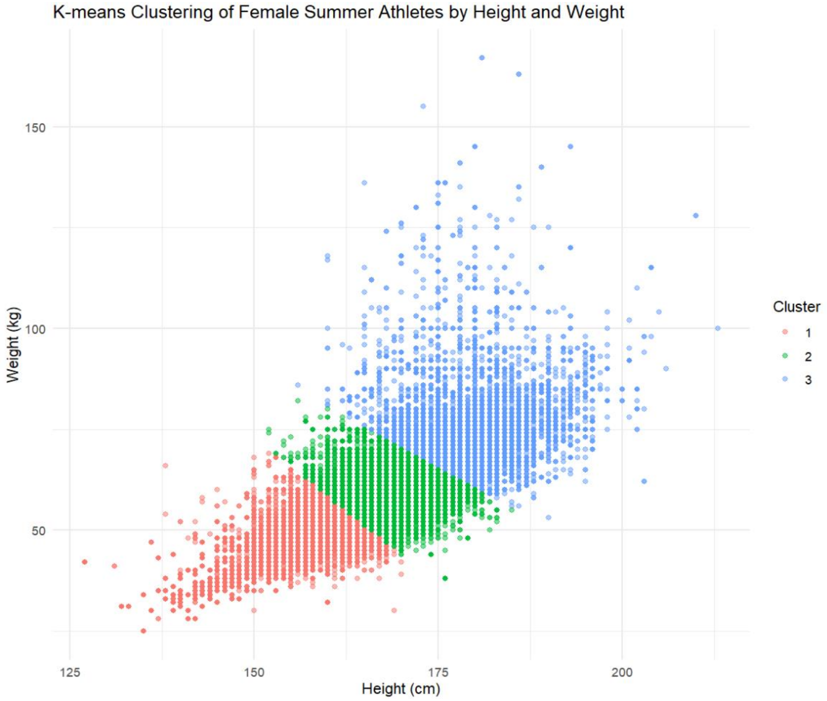


Appendix 9: K-Means Clustering Results- Version 2

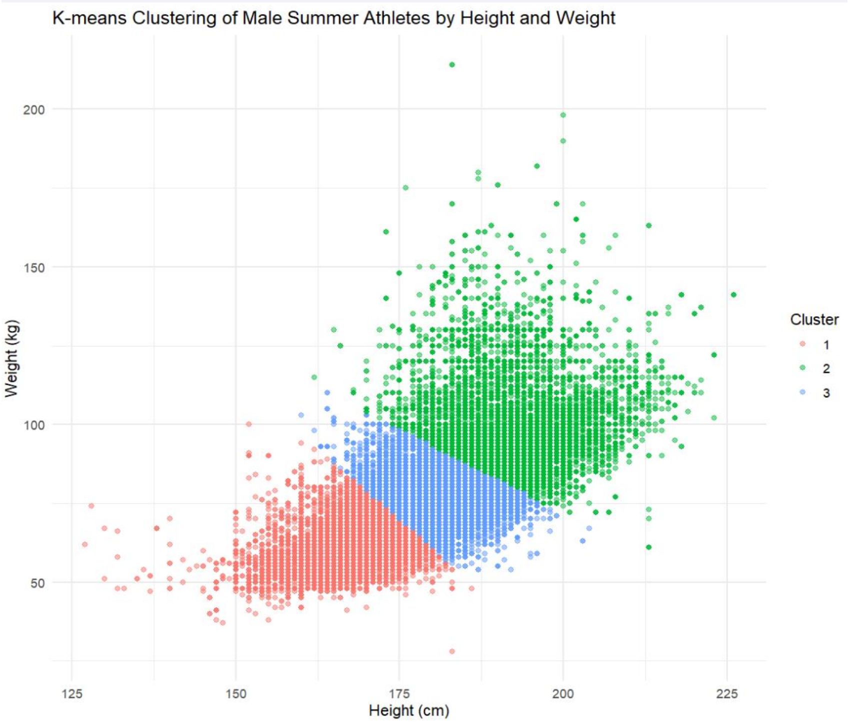


Appendix 9 Cont.: K-Means Clustering Results- Version 2





Appendix 9 Cont.: K-Means Clustering Results- Version 2



Appendix 10: Top Sports for Each Clustering Task -Version 2

K Means Clustering of Male for Summer Athletes by Height and Weight

Cluster 1

Sport	Count
Gymnastics	9921
Athletics	7286
Boxing	2825
Wrestling	2581
Cycling	2153

Cluster 2

Sport	Count
Athletics	3208
Rowing	3156
Swimming	2664
Basketball	1769
Volleyball	1239

Cluster 3

Sport	Count
Athletics	11082
Swimming	5432
Cycling	4014
Shooting	3217
Canoeing	2813

K Means Clustering of Female for Summer Athletes by Height and Weight

Cluster 1

Sport	Count
Gymnastics	6390
Athletics	2297
Swimming	825
Diving	557
Judo	320

Cluster 2

Sport	Count
Athletics	6470
Swimming	5346
Fencing	1181
Gymnastics	1175
Hockey	971

Cluster 3

Sport	Count
Swimming	2289
Athletics	2225
Rowing	1541
Volleyball	1085
Basketball	963

K Means Clustering of Male for Winter Athletes by Height and Weight

Cluster 1

Sport	Count
Cross Country Skiing	1782
Ski Jumping	1130
Alpine Skiing	977
Biathlon	895
Speed Skating	627

Cluster 2

Sport	Count
Ice Hockey	1852
Bobsleigh	1463
Alpine Skiing	996
Speed Skating	604
Cross Country Skiing	415

Cluster 3

Sport	Count
Cross Country Skiing	2252
Alpine Skiing	1763
Biathlon	1598
Ice Hockey	1502
Speed Skating	1233

K Means Clustering of Female for Winter Athletes by Height and Weight

Cluster 1

Sport	Count
Cross Country Skiing	1022
Biathlon	1763
Figure Skating	599
Alpine Skating	429
Speed Skating	425

Cluster 2

Sport	Count
Alpine Skiing	798
Speed Skating	633
Cross Country Skiing	471
Ice Hockey	365
Biathlon	263

Cluster 3

Sport	Count
Cross Country Skiing	1593
Alpine Skiing	1377
Biathlon	908
Speed Skating	883
Short Track Speed Skating	389

Citations

International Olympic Committee. (n.d.). Women in the Olympic Movement. Retrieved December 10, 2023, from https://stillmed.olympics.com/media/Documents/Olympic-Movement/Factsheets/Women-in-the-Olympic-Movement.pdf

Wikipedia contributors. (2023, April 4). Participation of women in the Olympics. In Wikipedia, The Free Encyclopedia. Retrieved December 10, 2023, from https://en.wikipedia.org/wiki/Participation_of_women_in_the_Olympics