Final Individual Project:

How Physical Attributes Affect Olympics Sports Performance

Alisa Liu

MGSC 661: Multivariate Statistics - Fall 2023

Prof. Juan Camilo Serpa

Date: December 10th, 2023

1 Introduction

Analyzing athletes' physical attributes across diverse Olympic sports reveals critical insights into the link between physiology and performance. Each sport demands unique traits — gymnastics prioritizes flexibility and strength, while basketball emphasizes height and agility. This analysis informs talent development, training optimization, injury prevention, and athlete well-being. It intersects with sports science, medicine, talent management, and public interest in sports. This project focuses on how Age, Height, and Weight differ across sports and their impact on winning Olympic medals. Through dataset analysis, the goal is to provide actionable insights for coaches, fans, and aspiring athletes, bridging empirical findings with practical recommendations in the realm of sports.

2 Data Description

In the arena of Olympic sports analysis, understanding the diverse physical attributes of athletes takes center stage. The decision to scrutinize the distinctive features of athletes across various sports is pivotal in unraveling the intrinsic traits that contribute to success, offering a nuanced understanding of the physical dynamics that underpin excellence in each sporting discipline.

The dataset encompasses a total of 15 variables, encompassing ID, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, and Medal. Notably, ID and Name are served as labels and are excluded from the model's feature set. This collection contains vital information concerning Olympic events. The Team variable indicates the national representation in each event, offering light on the many countries taking part. The NOC code serves as a unique identifier for each National Olympic Committee, allowing for exact categorization. Within the Games variable, the combination of season and year uniquely specifies the edition of the Olympic Games, providing a simple temporal reference. The term Year refers to the calendar year of the Games, whilst Season distinguishes between Summer and Winter Olympics. City identifies the host city, giving the events a geographical context. The variable Sport classifies athletic disciplines ranging from basketball to judo and tug-of-war. Finally, Event provides precise information about each sport's specific tournament. This combined dataset provides a full view of the competing teams, temporal and geographical aspects, and the various sports disciplines featured throughout the Olympic events.

However, given the project's focus on physical attributes and ultimate performance, only Age, Height, Weight, and Medal are deemed pertinent. In examining the dataset's distribution, it is evident from the histogram (refer to Appendix Figure 2) that the majority of athletes cluster around the age of 25. The right-skewed distribution suggests a scarcity of athletes beyond the age of 30. Furthermore, the histograms (refer to Appendix Figures 3 and 4) for Height and Weight exhibit a broader spread, with notable outliers — an extreme value of 175 for Height and approximately 72 for Weight.

We propose a new variable, Body Mass Index (BMI), which is generated from the existing variables Height and Weight, to delve deeper into the examination of physical features.

$$BMI = \frac{\text{Weight (kg)}}{\text{Height (m)}^2}$$

In real-world circumstances, BMI is regarded as an important indicator for determining physical well-being. The BMI distribution (refer to Appendix Figures 5) is slightly skewed to the right. Notably, there is a strong frequency peak around a BMI of around 23. This observation implies a concentration of data points inside a range linked with health and normal weight, which provides useful insights regarding the dataset's overall physical state.

3 Model Selection and Methodology

The methods used is a multi-step procedure. Initially, clusters are established utilising selected physical qualities, exposing specific characteristics for each sport within these clusters. Following that, the insights gained from this clustering lead the construction of hypotheses regarding the probable impact of these traits on medal-winning circumstances. To validate these assumptions, a Random Forest model is used to determine feature importance. In addition, statistical tests, such as t-statistics and p-values, are used to carefully examine the validity of the specified assumptions. This comprehensive method combines clustering techniques and machine learning models, along with standard statistical studies, to provide a full examination of the relationship between physical qualities and medal outcomes in sports.

Prior to entering into the complexities of the clusters, we begin the investigation by creating box plots for each physical attribute to gain early insights. Examining the box plot for Height (refer to Appendix Figure 7), reveals striking trends; for example, athletes participating in Basketball and

Volleyball had significantly greater average heights than counterparts in other sports. Similarly, the box plot representing Weight (refer to Appendix Figure 8) highlights distinctions, such as athletes in Figure Skating having a lower average weight. With these preliminary findings, we anticipate the establishment of clusters that reflect these distinguishing qualities across various sports. These findings serve as a foundational guide for the future cluster analysis.

In our earlier exploratory data analysis, we uncovered correlations (refer to Figure 1), between Weight and BMI and between Height and Weight. To mitigate multicollinearity, there is a consideration to exclude Weight. However, given the nuanced nature of the relationship between height and weight, we opt to retain both variables. Our approach involves utilizing two distinct sets of features for subsequent analyses: Set 1 comprising Age, Height, and Weight, and Set 2 featuring Age, Height, and BMI. This strategic delineation aims to assess potential variations in clustering and Random Forest outcomes, offering a more nuanced perspective on the impact of these physical attributes on our analytical results.

	Age	Height	Weight	BMI
Age	1.0000000	0.1051740	0.1587619	0.1655990
Height	0.1051740	1.0000000	0.7866738	0.3121581
Weight	0.1587619	0.7866738	1.0000000	0.8249005
BMI	0.1655990	0.3121581	0.8249005	1.0000000

Figure 1: Correlation Matrix.

The decision to use K-Means clustering is motivated by its simplicity, computational efficiency, and interpretability. This approach is particularly useful when dealing with enormous datasets or scenarios requiring scalability. In our project, K-Means allows us to identify distinguishing characteristics linked with each physical attribute for each sport inside individual clusters. Furthermore, it allows for an easy count of athletes from each sport inside these clusters. This streamlined procedure improves our capacity to summarise and interpret the distinctive qualities of athletes across multiple sports, contributing to a more nuanced comprehension of the dataset and the ability to make additional assumptions. Nonetheless, choosing the appropriate value for k, the number of clusters, is a major difficulty. To remedy this, we use the Elbow approach, which is a methodology for determining the best k for our particular dataset. We can find the "elbow" point by analysing the distortion or inertia over multiple k values, showing a compromise between maximising the number of clusters and minimising intra-cluster variability. Following that, we offer a detailed cluster plot (refer to Appendix Figure 10) illustrating the distinctiveness of the four selected groups, providing a visual depiction of the best clustering solution for our data.

The use of Random Forest to assess feature importance in each sport allows for a full examination of how physical features contribute to predicting medal outcomes, exploiting the algorithm's capacity to capture complicated, non-linear interactions. However, it cannot tell us how those attributes affect the medal results, therefore, the inclusion of statistical tests, such as t-tests, serves as a critical validation step, rigorously evaluating whether detected variations in physical features between medalists and non-medalists are statistically significant. These tests not only give a typical method of hypothesis testing, but they also provide information on effect sizes, which improves our comprehension of the practical importance of observed differences.

4 Results and Conclusions

Two separate clustering conclusions emerge from the use of different sets of characteristics, but the elbow plots show a constant optimal k value of 4 across both findings. Despite some overlapping points in each plot, the main clustering patterns are clear and show sensible divisions.

The initial Weight-based grouping approach gives various insights. Cluster 1 athletes are distinguished by substantial Height and Weight, and they are primarily from Athletics, Swimming, Basketball, and Ice Hockey. Cluster 4, on the other hand, has athletes with the shortest Height and Weight, with Gymnastics having the highest athlete count of any sport. Meanwhile, the remaining two clusters represent athletes with intermediate attributes who lack obviously distinguishing characteristics.

Because there are far too many sports categories in this dataset, we only interpret a handful that have distinct properties. First, we look at the Gymnastics athletes in Cluster 4, because a considerable number of them have lower values in their height and weight. We assume that if they have lower values in their height and weight, they will have a better probability of winning a medal in this sport events. By looking into the result (refer to Appendix Figure 14), the Random Forest analysis in Gymnastics accentuates the significance of physical attributes, with Age identified as the most influential factor, followed by Height and Weight. The MeanDecreaseAccuracy values of 55.35366 for Age, 20.27574 for Height, and 22.45635 for Weight underscore their respective contributions to predicting medal outcomes. Aligning with these findings, the subsequent Welch Two Sample t-tests provide statistical validation. In particular, the t-test for Height yields a significant difference (t = -6.7469, p-value = 1.846e-11), as does the t-test for Weight (t = -6.1999, p-value = 6.526e-10), affirming that medalists tend to have higher mean heights and weights. This cohesive analysis suggests that, in Gymnastics, a combination of age, height, and weight plays a pivotal role in predicting and achieving medal success. This comprehensive review indicates that, whereas athletes in Gymnastics often have smaller heights and weights, these physical characteristics do not emerge as critical components in evaluating their performance. This result can emphasize the multidimensional nature of performance measures in the sport, which might driving a change in focus towards other factors like as skill mastery, flexibility, and strength for a more thorough assessment of gymnastic prowess.

Using the same approach, we focus on another set of parameters, including BMI (see Appendix Figure 12). Notably, four separate categories appear, with Cluster 3 standing out for having the greatest concentration of athletes with the greatest Height and Weight. It's worth noting that, despite Athletics' significant presence in other clusters, we illustrate our point in this context by utilising Rowing as an example. Based on the features in this cluster, we assume that an athlete with a bigger height and weight is more likely to win a medal. Then we apply the Random Forest analysis and find that, while Height does not show a statistically significant difference between medalists and non-medalists, BMI plays an important role in predicting success. Height and BMI had significant MeanDecreaseAccuracy values of 52.88618 and 51.25227, respectively. The subsequent t-tests confirm these findings, with the p-value for Height being non-significant (0.5462) and the p-value for BMI being somewhat more significant (0.01955). In practice, this means that a higher BMI is connected with medals in rowing, emphasising the importance of body composition and fitness levels over height alone. Upon closer examination of the mean BMI values, it becomes apparent that there is minimal disparity between athletes who secured a medal (23.35257) and those who did not (23.25433). This marginal difference challenges our initial assumption and aligns with the overarching conclusion drawn from the statistical results. Contrary to expectations, the similarity in BMI values suggests that, in the context of Rowing, these specific physical attributes may not serve as the primary determinants of athletes' performance outcomes. This nuanced understanding underscores the need for a more comprehensive exploration of factors beyond BMI in assessing and enhancing performance in Rowing.

Following a thorough examination, it is clear that, while certain physical characteristics such as height and weight may influence individuals' sports choices, they do not serve as direct indications of athletic ability. The primary predictors of success include elements such as intensive training regimens and skill mastery, in addition to natural features. Athletes should regard their innate abilities as advantages to be deliberately used, rather than restraints. This nuanced viewpoint allows for a shift in emphasis from inherent features to the development of skills and training strategies that genuinely contribute to peak performance. This understanding has enormous business value for coaches, athletes, and sports organisations. Training programmes and talent development tactics that are tailored to individual strengths and weaknesses, rather than fixed physical attributes, have the potential to maximise performance in sports competitions such as the Olympics. This adaptable strategy is consistent with the larger goal of cultivating excellence and resilience in the dynamic context of competitive sports.

5 Appendices

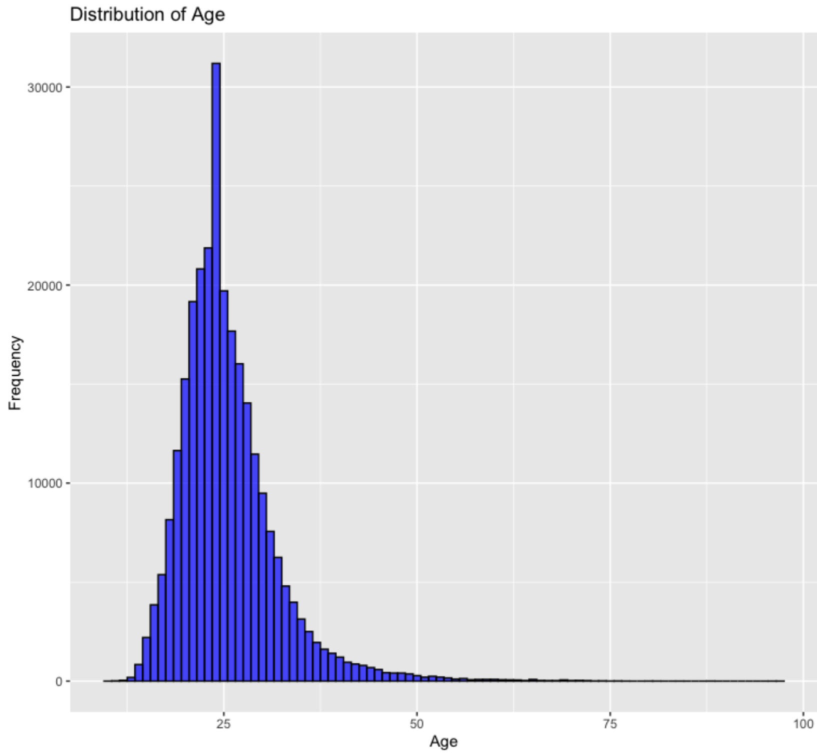

Figure 2: Histogram of variable ”Age”.

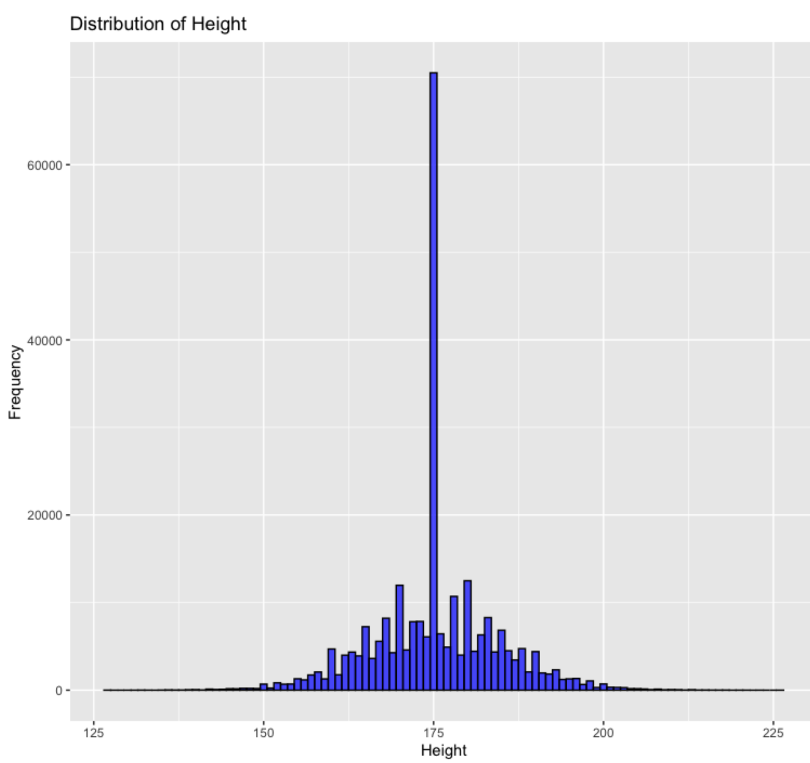

Figure 3: Histogram of variable ”Height”.

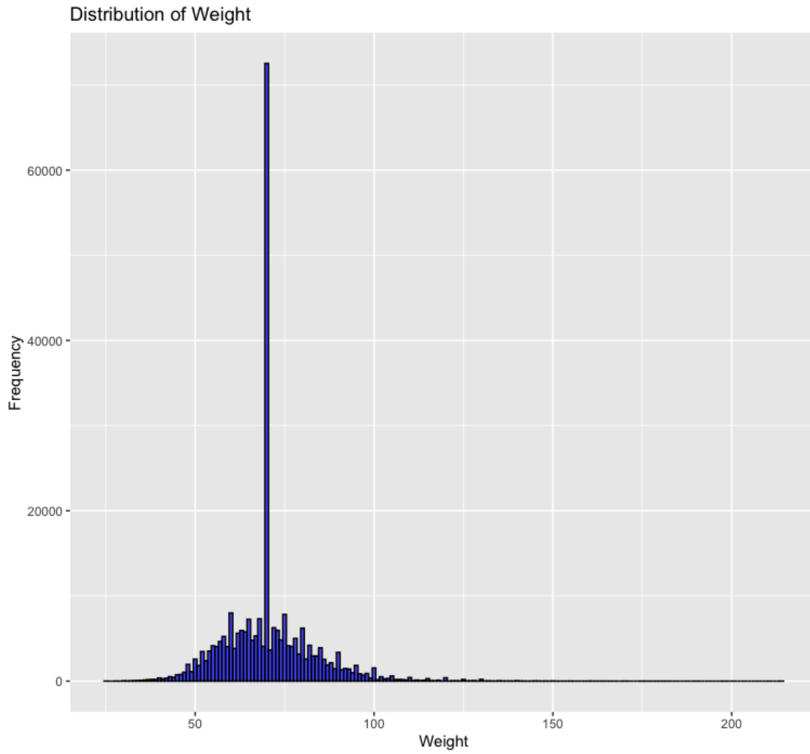

Figure 4: Histogram of variable "Weight".

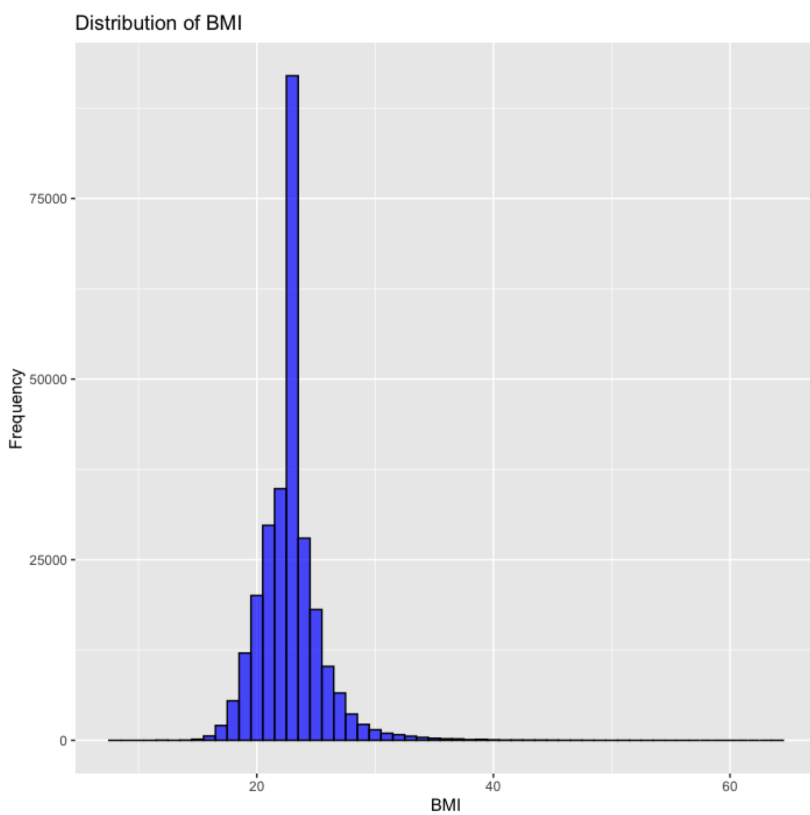

Figure 5: Histogram of variable "BMI".

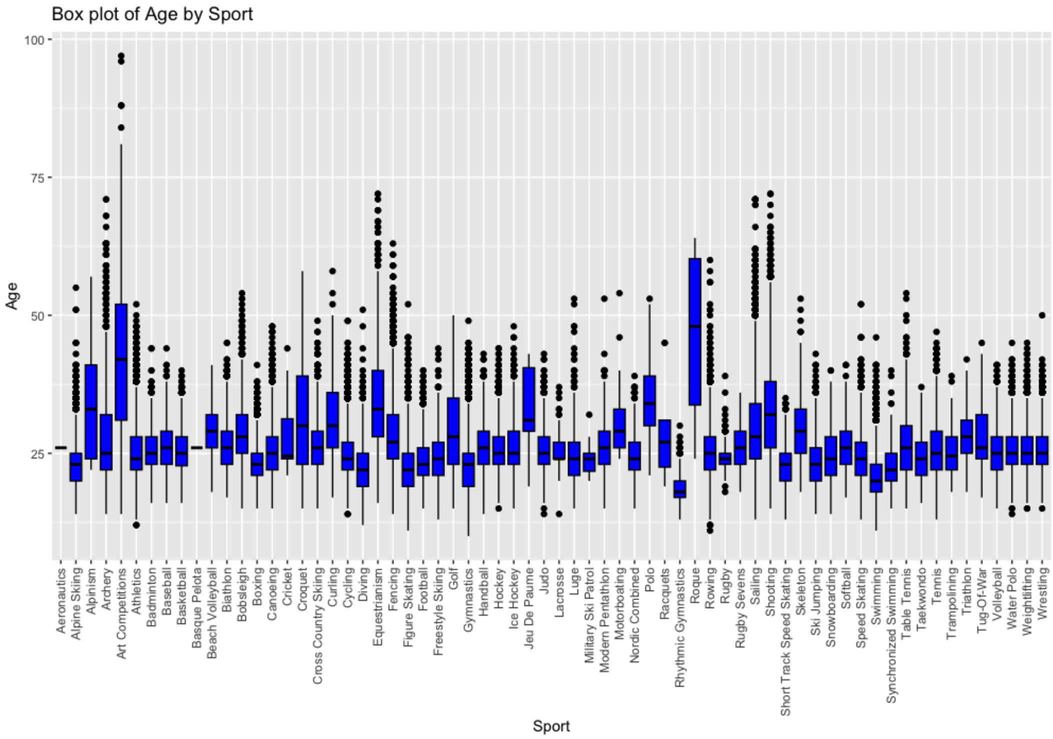

Figure 6: Box plot of variable "Age".

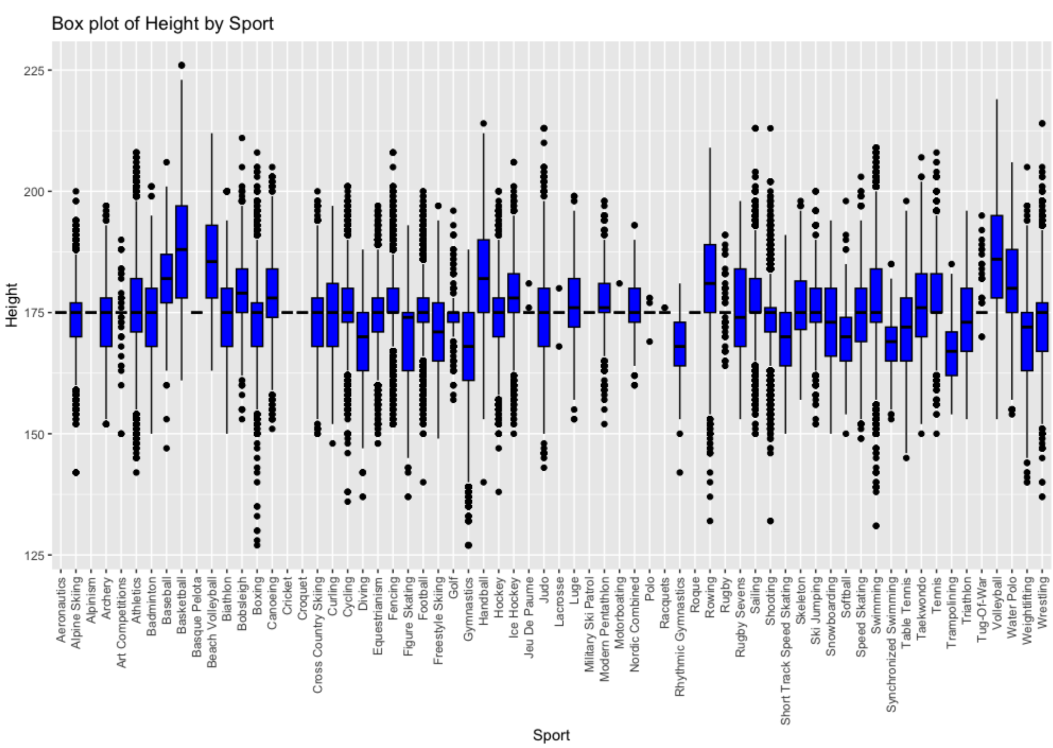

Figure 7: Box plot of variable "Height".

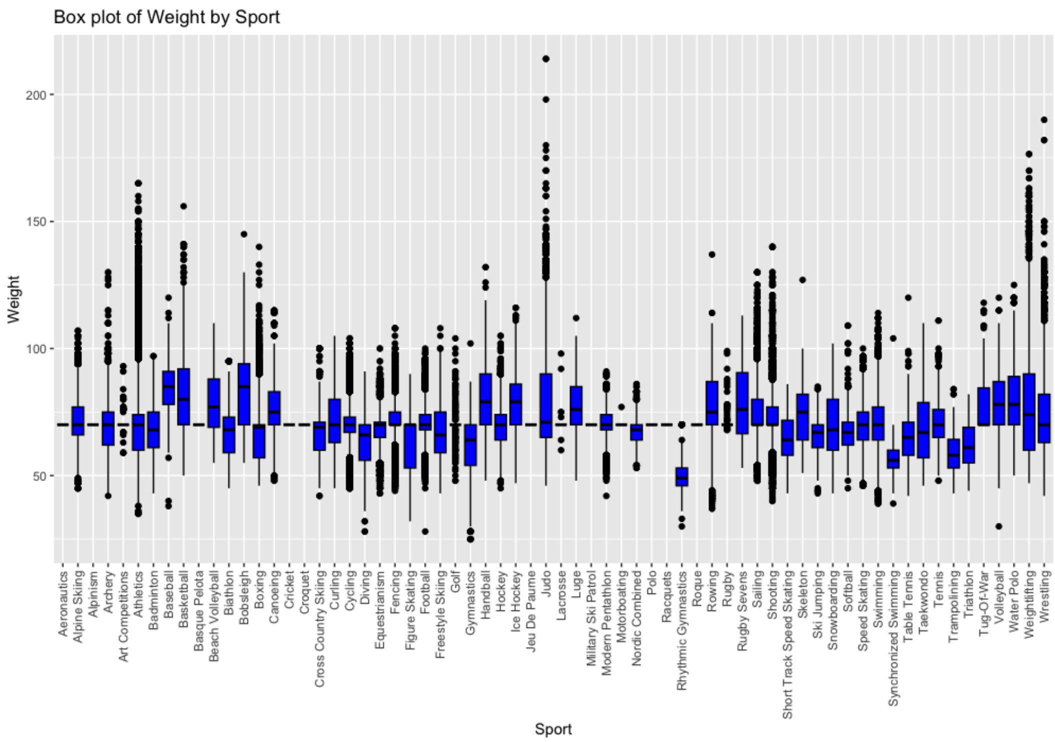

Figure 8: Box plot of variable "Weight".

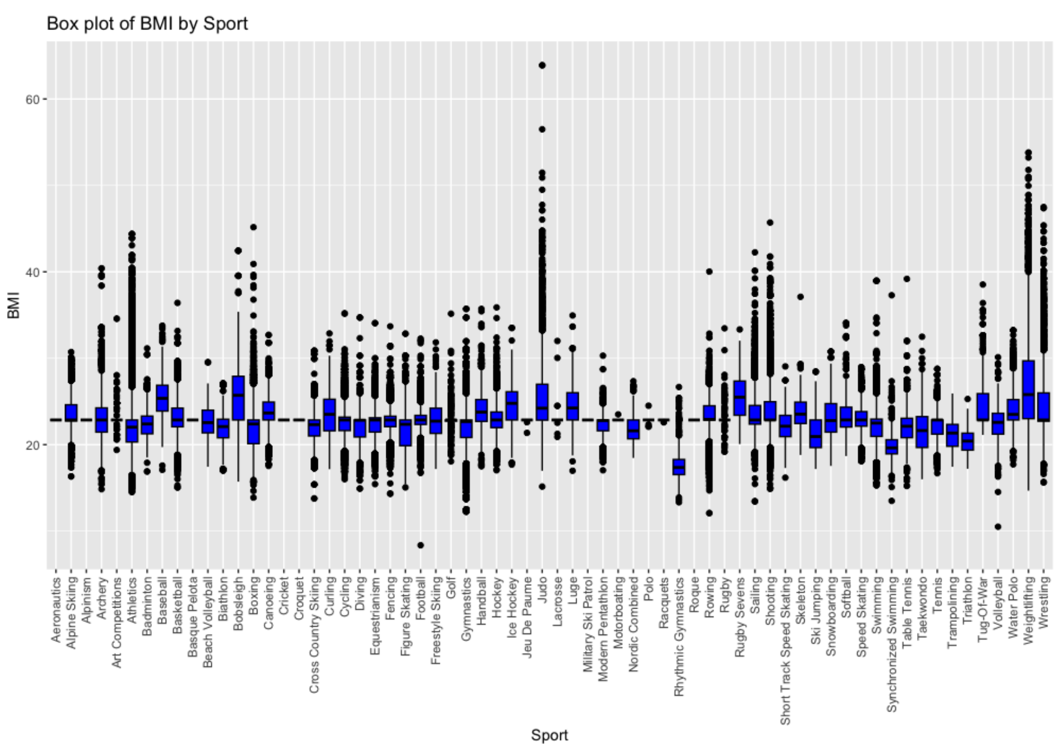

Figure 9: Box plot of variable "BMI".

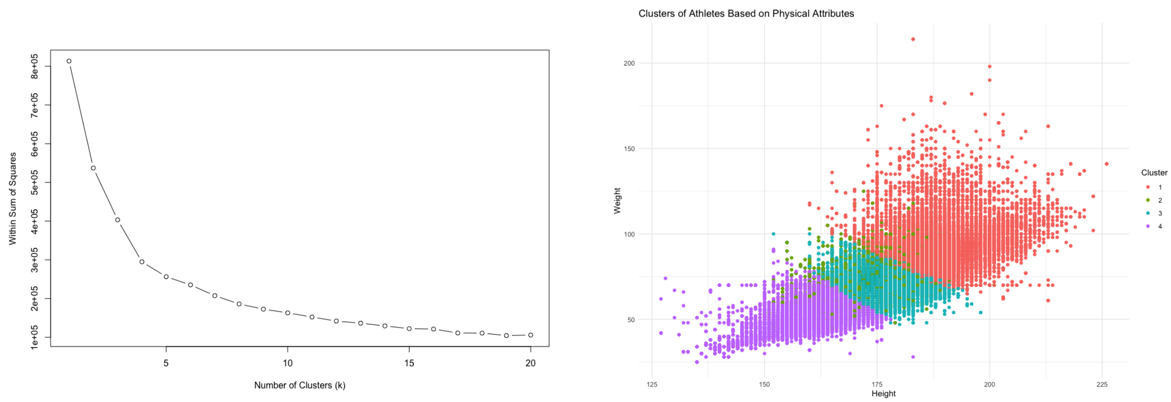

Figure 10: K-Means clusters using Height and Weight with the Elbow method.

Cluster 1 :

	Sport	Mean_Age	Mean_Height	Mean_Weight	Athlete_Count
1	Athletics	25.75113	188.0621	90.93000	6843
2	Swimming	22.43181	190.2442	84.33610	5250
3	Rowing	25.28597	190.6482	88.95722	4511
4	Basketball	25.47396	196.4587	91.76674	2688
5	Ice Hockey	26.55255	185.1747	89.33489	2141
6	Canoeing	25.45558	186.5385	86.22630	2015
7	Volleyball	25.70900	194.1779	86.89764	1866
8	Handball	26.91548	190.4632	91.27497	1822
9	Water Polo	26.13371	189.7796	91.26459	1765
10	Bobsleigh	28.04140	184.4457	93.96992	1546

Cluster 2 :

	Sport	Mean_Age	Mean_Height	Mean_Weight	Athlete_Count
1	Shooting	39.38831	174.0456	72.79813	5766
2	Equestrianism	39.47970	174.5631	68.98736	3719
3	Fencing	36.53620	175.4344	70.68016	2928
4	Athletics	33.96065	174.7033	65.93210	2872
5	Art Competitions	48.58259	175.0393	70.10481	2700
6	Sailing	38.76384	175.9943	72.94323	2096
7	Gymnastics	33.86355	173.3078	69.04680	1517
8	Cross Country Skiing	33.56125	174.3941	67.77160	1053
9	Cycling	33.69534	174.9617	67.68705	965
10	Rowing	34.42474	176.0686	69.67735	671

Cluster 3 :

	Sport	Mean_Age	Mean_Height	Mean_Weight	Athlete_Count
1	Athletics	23.95536	176.8457	68.75679	20252
2	Swimming	20.52282	177.1615	69.52754	13324
3	Gymnastics	23.79733	173.8090	69.06407	11605
4	Cycling	23.55761	176.1521	70.40494	7369
5	Alpine Skiing	22.66966	174.8158	71.52171	5343
6	Fencing	24.74650	176.2210	70.36614	5207
7	Football	23.25893	175.7147	70.80186	5009
8	Rowing	24.02621	176.8274	71.17844	4769
9	Cross Country Skiing	24.74058	176.2224	69.75500	4753
10	Shooting	25.14931	174.8616	71.41207	3389

Cluster 4 :

	Sport	Mean_Age	Mean_Height	Mean_Weight	Athlete_Count
1	Gymnastics	20.67547	159.5750	53.18060	13552
2	Athletics	24.44681	165.2309	54.88651	8657
3	Swimming	18.44317	165.8633	57.07184	4461
4	Cross Country Skiing	25.10949	164.5865	56.42190	2612
5	Boxing	22.72554	164.7401	54.75414	1993
6	Wrestling	24.63742	162.1749	57.23916	1892
7	Alpine Skiing	21.90112	164.4460	59.69165	1881
8	Biathlon	25.54533	164.3487	55.72908	1434
9	Cycling	24.40287	164.9403	58.03553	1323
10	Speed Skating	22.93427	163.7754	58.97535	1278

Figure 11: Characteristics of each Sport in Figure 10 clusters (partial results).

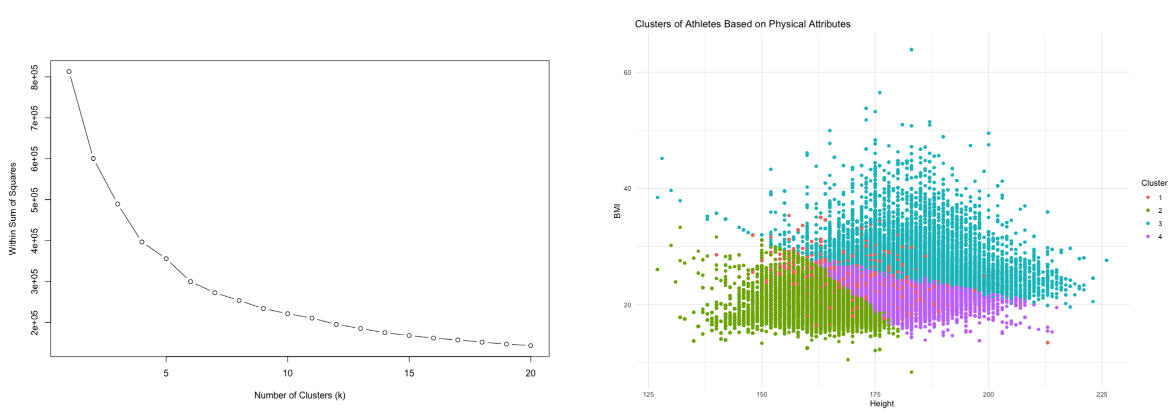

Figure 12: K-Means clusters using Height and BMI with the Elbow method.

Cluster 1 :

	Sport	Mean_Age	Mean_Height	Mean_BMI	Athlete_Count
1	Shooting	39.07466	173.8985	23.69176	5813
2	Equestrianism	38.88650	174.5310	22.63722	4000
3	Athletics	33.51590	174.3460	21.56456	3301
4	Fencing	35.95279	175.5014	22.96270	3283
5	Art Competitions	48.30463	175.0387	22.88181	2741
6	Sailing	38.21733	175.9843	23.35265	2227
7	Gymnastics	32.77042	172.7648	23.06135	2143
8	Cross Country Skiing	32.95839	174.0255	22.37917	1370
9	Cycling	33.20104	174.7626	22.16996	1154
10	Rowing	34.00125	176.1461	22.52594	801

Cluster 2 :

	Sport	Mean_Age	Mean_Height	Mean_BMI	Athlete_Count
1	Gymnastics	20.63252	159.8631	20.73934	13783
2	Athletics	24.25123	166.0856	20.04642	9517
3	Swimming	18.56071	166.7617	20.67749	5090
4	Cross Country Skiing	24.84021	165.0533	20.73983	2666
5	Boxing	22.62921	165.4995	20.08926	2136
6	Alpine Skiing	21.75483	164.9079	21.95606	1966
7	Wrestling	24.45170	162.3810	21.59691	1853
8	Biathlon	25.28552	164.8334	20.58774	1471
9	Cycling	24.15239	165.6713	21.14913	1378
10	Speed Skating	22.65471	164.3655	21.87865	1338

Cluster 3 :

	Sport	Mean_Age	Mean_Height	Mean_BMI	Athlete_Count
1	Athletics	26.57563	185.9369	28.33688	4390
2	Rowing	25.58846	190.9072	24.99394	3589
3	Ice Hockey	26.56707	183.0578	26.35210	2490
4	Swimming	23.57880	190.6476	24.65332	2094
5	Wrestling	26.45723	181.8118	29.30256	1929
6	Basketball	25.90889	198.7035	24.65611	1811
7	Canoeing	25.92180	184.5656	25.56038	1752
8	Bobsleigh	28.28176	182.9971	27.85206	1700
9	Weightlifting	25.74180	175.1795	31.89306	1677
10	Handball	27.14728	190.2851	25.65507	1582

Cluster 4 :

	Sport	Mean_Age	Mean_Height	Mean_BMI	Athlete_Count
1	Athletics	23.92734	178.7507	22.00780	21416
2	Swimming	20.75583	179.8860	22.29425	15788
3	Gymnastics	23.56542	173.9394	22.90710	10746
4	Cycling	23.47005	177.3018	22.60652	7595
5	Rowing	23.88234	178.9607	22.74389	5516
6	Fencing	24.55665	177.9326	22.59561	5472
7	Football	23.08332	176.7290	22.90098	5173
8	Alpine Skiing	22.47585	175.2740	23.21326	4949
9	Cross Country Skiing	24.60460	177.4788	22.50815	4866
10	Boxing	22.77608	176.7581	22.43566	3278

Figure 13: Characteristics of each Sport in Figure 12 clusters (partial results).

Sports	Age	Height	Weight
Gymnastics	55.35366	20.27574	22.45635
Athletics	16.85659	50.27649	49.75160
Swimming	69.18407	59.25329	65.11505
Rowing	11.62747	63.26647	60.77048
Basketball	0.6736021	24.7711265	28.9797045
Ice Hockey	-14.84683	43.26518	46.09104

Sports	t_{Height}	t_{Weight}	p_{Height}	p_{Weight}
Gymnastics	-6.7469	-6.1999	1.846e-11	6.526e-10
Athletics	-10.297	-8.9401	2.2e-16	2.2e-16
Swimming	-13.178	-12.355	2.2e-16	2.2e-16
Rowing	-0.60355	-1.756	0.5462	0.07915
Basketball	-7.8058	-6.5127	1.015e-14	9.74e-11
Ice Hockey	0.18851	-1.1793	0.8505	0.2384

Figure 14: RF Feature Importance and Statistical Testing (Weight).

Sports	Age	Height	BMI
Shooting	26.83545	23.73327	27.11112
Equestrianism	22.47770	34.53640	36.79829
Rowing	10.06576	52.88618	51.25227
Ice Hockey	-17.26967	28.43219	29.64652

Sports	t_{Height}	t_{BMI}	p_{Height}	p_{BMI}
Shooting	-2.8352	3.5275	0.004634	0.0004307
Equestrianism	-3.9018	1.0638	0.0001001	0.2876
Rowing	-0.60355	-2.3355	0.5462	0.01955
Ice Hockey	0.18851	-2.3752	0.8505	0.01761

Figure 15: RF Feature Importance and Statistical Testing (Weight).