Final Project

The Shape of Champions: Unraveling the Truth Behind Athletic Body Stereotypes in Various Sports

Presented to
Professor Juan Serpa

By
Delisle, Audrey - 261142504

MGSC 661 Multivariate Statistics - Section 076

McGill University - Desautels Faculty of Management
Sunday December 10th 2023

Table of Contents

1. Introduction

1. 1 Project Summary

Have you ever dreamed of swooshing through a basketball court, only to hear someone say you're not quite tall enough for the hoop's height? Or maybe you've imagined swinging from the uneven bars in gymnastics, but whispers about being too tall or too slender followed you? Perhaps you've wrestled with the idea of stepping onto the mat, yet you were discouraged by a leaner silhouette that didn't seem to fit the robust image of a wrestler. Stereotypes in sports are as old as the games themselves, each echoing the belief that certain physical traits predestine success. In this project, we'll dribble through the world of basketball, balance on the beam of gymnastics, grapple with wrestling, dive into swimming, and take strides in running to investigate if there's truth behind these body stereotypes. By examining medal-winning Olympic athletes across these five sports known for their physical typecasts, we'll explore the correlation between an athlete's age, height, weight, BMI, and nationality, and their likelihood participating in a certain sport.

1.2 Project Goals

The overarching objective of this project is to analyze and understand the factors contributing to an athlete's probability of participating in five specific Olympic sports - Basketball, Gymnastics, Wrestling, Athletics, and Swimming. Throughout the project, we will navigate through different data analysis steps, each obtaining various goals. These include:

1) Data Exploration and Preprocessing: We want to begin by meticulously processing the Olympic dataset by filtering, categorizing, and transforming variables to ensure relevancy and accuracy in the analysis. We will do some feature engineering, check collinearity, outliers, and specific filtering.

2) Predictive Analysis: Next, we wish to employ logistic regression, linear discriminant analysis (LDA), and principal component analysis (PCA) to identify key predictors and their influence on the likelihood of an athlete's success in these sports.We will analyze the results of this data analysis to see if stereotypes really do stand out in this olympic dataset.

3) Practical Application: Lastly, is is important to translate the analytical insights into practical recommendations, particularly for sports recruitment and strategy development, enhancing the understanding of sports dynamics based on physical, geographical, and demographic factors.

All in all, this project aims to offer a data-driven perspective on athlete selection and training, contributing valuable insights for sports professionals and enthusiasts.

2. Data Description

2.1 Introduction to the Dataset & Preprocessing

The dataset I selected is comprised of Olympic Games records, initially containing 271,116 rows and 15 columns. It includes athlete demographics and performance metrics across multiple events. The preprocessing of this dataset was an important step for the needs of the project.

To begin with, there was the addition of BMI, calculated from athletes' height and weight. This provided an additional variable to assess the physical attributes of athletes. The dataset was further filtered to include only data from 1980 onwards, ensuring the analysis remained relevant to the modern context of the Olympics. This was particularly important because earlier editions of the Games included non-athletic events, like art competitions, which are not part of the contemporary Olympic events. Focusing on medal winners was another deliberate choice, narrowing the dataset to successful athletes (all athletes with a medal) and thereby aiming to understand the attributes of those who have achieved a certain level of recognition in their fields. I then decided to include athletes' nationalities, mapped to their respective continents, which introduced a geographical dimension to the analysis.

2.2 Dependent Variable (Y)

For this analysis, the dependent variable (Y) is the probability of an athlete participating in one of five selected sports: Basketball, Gymnastics, Wrestling, Athletics, and Swimming. I chose these 5 sports because I felt like they had a lot of various physical body stereotypes.[1] They also captured a diverse range of athletic disciplines encompassing team sports, individual performance, and different physical demands. The probability was calculated as the proportion of athletes participating in each of these sports out of the total number of athletes, as seen in Appendix A for reference. Using probability as the dependent variable also aligns well with the logistic regression models employed later in the analysis. This choice makes the interpretation of results more intuitive, as the outcomes directly correlate to the likelihood of an athlete participating in a sport.

2.3 Independent Variables x_i Overview

With a focus on five sports, the dataset encompasses a range of independent variables that offer insights into the characteristics influencing an athlete's likelihood of participating in a particular sport. These variables include age, height, weight, gender, BMI, and the athletes' continental representation. I ran summary statistics and plots to get a good overview of these variables that can be found in Apendix B, C and D.

[1] Burtka, J. (Jan 16, 2020). Too tall, too short, too big: Athletes breaking body stereotypes. Global Sports Matters. https://globalsportmatters.com/science/2020/01/16/too-tall-too-short-too-big-athletes-breaking-body-stereotypes/

Age is a crucial factor in sports performance, often indicating the phase of peak physical capability. In our dataset, athletes' ages range widely, suggesting varied career spans across different sports. The median age typically falls in the mid-20s, a period often associated with optimal athletic performance. However, the age distribution also reflects the unique demands and accessibility of each sport, with some sports like Gymnastics skewing towards younger participants (youngest at 13 years old).

Height and weight are significant in sports, as they can provide advantages or indicate suitability for certain types of physical activity. For example, basketball players in the dataset are generally taller and heavier, which aligns with the sport's physical demands. On the other hand, gymnasts, who require agility and flexibility, tend to be shorter and lighter. BMI also has similar behaviour because it is derived directly from height and weight. However, it gives a nuanced outlook to see what athletes lean more towards a normal weight or over/underweight.

Gender distribution varied across the selected sports, reflecting the natural segregation in Olympic competitions. While some sports showed a balanced representation of both genders, others were dominated by one, aligning with historical trends and cultural inclinations towards certain sports. This variable was later excluded from the model, given the gender-specific nature of Olympic events.

The continent variable offered insights into the geographical distribution and cultural preferences in sports participation. Certain continents showed a stronger presence in specific sports, indicative of regional popularity or infrastructure supporting those sports. For example, most basketball players came from Northern America and Southern Europe, whereas the athletics athletes came from Sub Saharan Africa, Northern America and Latin America. This variable adds a cultural and geographical dimension to the analysis, highlighting the diversity in sports participation on a global scale.

2.4 Correlations and Multicollinearity

Regarding outliers, these were observed across various variables, as indicated by the box plots in the appendices B, C and D. However, the decision was made to retain these outliers in the dataset. Even though these vould affect my model, they are still athletes, they are not typos, and hence I decided that all athletes should be included. To address multicollinearity, a correlation matrix heatmap was created to visualize the relationships between all numeric variables (view Appendix E). This analysis revealed a high degree of correlation between BMI, height, and weight. Given that BMI is a derived measure based on height and weight, its strong correlation with these two variables was expected. However, in predictive modeling, such high collinearity can distort the significance and interpretation of these predictors. Based on these findings, the decision was made to remove BMI from the analysis. Similarly, the variable 'Sex_M' was also removed from the analysis. In the context of the Olympics, where men and women compete in separate categories, using gender as a predictor could introduce bias and unfair comparisons.

3. Model Selection

3.1 Methodology

The methodology adopted for this project centered around a combination of classification techniques and unsupervised learning to analyze the Olympic dataset. The objective was to explore the probability of an athlete participating in one of five selected sports (Basketball, Gymnastics, Wrestling, Athletics, Swimming) based on prior physical attributes and ethnicity. Considering the nature of the dependent variable — a categorical outcome representing sports participation — classification methods were deemed most appropriate. To begin, logistic regression was employed and it provided insights into the influence of my predictors on the probability of an athlete's participation in a specific sport. Linear regression could not be used, because as seen in class, it has no boundaries and probabilities need to be between 0 and 1. Hence, logistic regression with the glm() function was used. Next, linear discriminant analysis (LDA) was chosen as a complementary approach to logistic regression. It helped in identifying the variables that best seperate athletes into the different sports. Lastly, PCA was used to gain more insight on how many predictors should be included in the model. It assisted in identifying the structure of the data and the key components that explain the most variance in athlete characteristics.

3.2 Classification: Logistic Regression and LDA

In the logistic regression models, each sport was treated as a separate binary outcome, and logistic regression models were developed for each sport. The models included predictors such as age, height, weight, and continent, chosen based on their relevance to athletic performance. Since there were over 10 continent sub-regions, I decided to chose the 5 most popular/un-popular ones based on the box plot in Appendix B. This consisted of Northern America, Eastern Europe, Eastern Asia, Sub-Saharan Africa, and Southern Europe. Thus, I had 8 predictors in total.

To compliment my findings from logistic regression, I looked at linear discriminant analysis. LDA helped in visualizing the data in a reduced number of dimensions while maintaining the separation between the different sports categories. This method provided a clear visual representation of how athletes from different sports cluster based on their physical and geographical attributes.

3.3 Unsupervised Learning: PCA

Lastly, principal component analysis was conducted as a part of unsupervised learning to further dive into the dataset's structure. PCA helped in identifying the principal components that capture the maximum variance in the dataset. This approach was particularly useful in understanding the relationships and patterns among the variables without the influence of the predefined categorization of sports. It also gave me a better idea of how many predictors to include in a model like classification.

4. Results

4.1 Logistic Regression Model

To begin, In the logistic regression analysis for each of the five sports (view Appendix F), the models revealed patterns that align with common stereotypes about these sports. For instance, in basketball, positive coefficients were observed for variables like age and height, particularly a coefficient of 0.17 for height, indicating that an increase in height by one unit (cm) boosts the probability of being a successful basketball athlete in the Olympics by 0.17. Regions like Northern America, Eastern Europe, Eastern Asia, and Southern Europe also showed positive associations, while weight and Sub-Saharan Africa had negative coefficients. Gymnastics exhibited a negative relationship with age, height, and weight, implying a decrease in the probability of success with increasing values in these factors. Swimming showed positive coefficients for height and Northern America, but negative for other variables. Wrestling and athletics highlighted the positive impact of age and weight, with athletics also showing a significant positive coefficient (2.88) for Sub-Saharan Africa. The p-scores indicate that most variables are significant, but a couple results had p-scores above 0.1, suggesting cautious interpretation.

4.2 Linear Discriminant Analysis Model

To continue, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) were employed to explore the relationships between athletes' physical attributes and their sports. The LDA, focusing on weight and height, produced a partition plot with an error rate of 0.515 (view Appendix G). This plot clustered gymnasts in the low weight and height range, basketball players in the high height segment, while athletes in wrestling and athletics were positioned in the high weight-average height zone, and swimmers in the middle. This was pretty in tune with prior stereotypes, but my error rate was pretty high and swimmers were not clearly classified. Due to these reasons, I decided to run a quadratic discriminant analysis to see if I could decrease my error rate and better classify the sports.

For QDA, I also included age to see if that would make a difference. I got 3 partition plots, but the one with the lowest error rate was age against height, with a rate of 0.435 (view Appendix H). This means that my model correctly classified 56.5% of my data points. This is not great, but is a lot better than the linear discriminant model. This improved model more distinctly separated swimmers by average height but younger age. Gymnasts remained characterized by lower height and age, basketball players by average age but greater height, and wrestlers and athletes in athletics by higher age and average height, showcasing the nuanced physical profiles prevalent in these sports.

4.3 PCA

In the PCA analysis of the Olympic dataset, principal components were visualized to discern relationships between variables and sports categories. The visualization, generated through PCA, revealed distinct directional arrows for each variable, indicating their influence on different sports (view Appendix I). However, it is tricky to specifically analyze a visual graph. Due to this, I took a look at the numerical output (view Appendix I). The insights I got from this were that most of the variability found across sports can be explained by four variables: height, weight, Sub-Saharan African ethnicity, and age. I can see that these absolute values are significantly higher than the others in PCA1.

I then conducted an analysis of principal components for variance explanation (view Appendix J), which revealed that with five components, around 40% of the dataset's variance is explained. This percentage increases with the number of components, reaching about 80% with twelve components. However, considering the dataset's limited number of variables (eight in total, with five being continents), it was crucial for me to utilize all predictors. Each predictor provided unique insights into different sports, underscoring the significance of their inclusion in the analysis. This approach was particularly fitting for this dataset, where the specific sports categories were distinct and required comprehensive variable inclusion for accurate analysis. In larger datasets with more variables, PCA would be more important in reducing dimensionality and focusing on the most impactful predictors.

5. Classification/predictions and conclusions

5.1 Findings Summary

Each sport showcased distinct characteristics influencing athlete success. Basketball highlighted the significance of height and regions like Northern America and Eastern Europe. Gymnastics showed a negative correlation with age, height, and weight, indicating a preference for younger, lighter athletes. Swimming's success was closely tied to height, with a strong representation from Northern America. Wrestling and Athletics emphasized age and weight, with a notable impact from Eastern European and Sub-Saharan African backgrounds. The LDA revealed moderate differentiation among sports based on height and weight, with a 0.515 error rate. QDA further refined this by incorporating age, resulting in a lower error rate of 0.435. This indicated a more nuanced approach to athlete classification. The PCA analysis highlighted the primary variables driving variance across sports. Height, weight, and regional factors like Sub-Saharan African ethnicity were key determinants. The percentage of variance explained by the principal components indicated that around 40% of sport classification could be attributed to the first five components, increasing to approximately 62% with eight components.

These findings offer an overall view of the factors influencing athlete success across various Olympic sports, providing a data-driven foundation for future sports analysis and recruitment strategies.

5.2 Managerial conclusions

This project went through a lot of analysis, but all in all, I can conclude that it is clear that there are physical body stereotypes present in different athletes competing in a variety of sports. The information gained from my report can have multiple impacts on the real world, as seen below;

Recruitment Strategy: The study's results are invaluable for sports recruiters. For instance, basketball recruiters should focus on taller athletes, as height significantly impacts performance in this sport. Similarly, swimming recruiters could prioritize individuals with an average age but taller stature, aligning with the sport's physical demands.

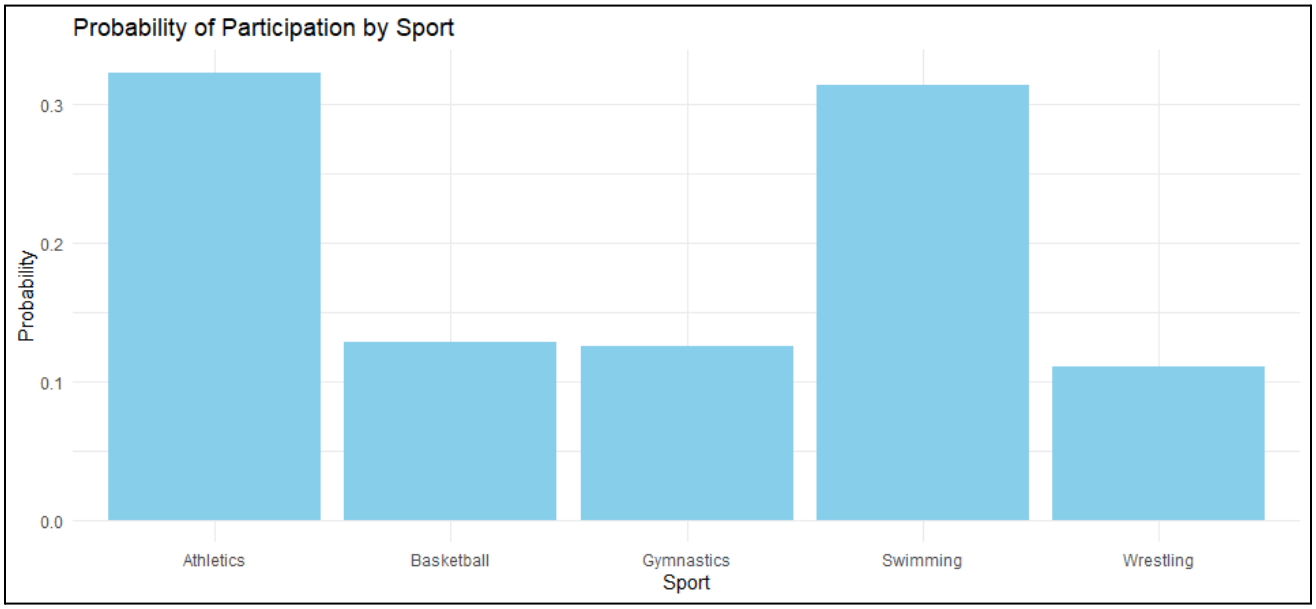
Training Focus: Coaches can use these insights for targeted training. Understanding that weight and height influence success in wrestling and athletics means training regimens can be tailored to enhance these aspects, maximizing an athlete's potential in their chosen sport.

Demographic Considerations: Regional factors also play a crucial role. Athletes from Northern America and Eastern Europe show a higher propensity for success in certain sports. This could guide international recruitment strategies, focusing efforts on regions with a higher likelihood of producing successful athletes in specific disciplines.

6. Appendices

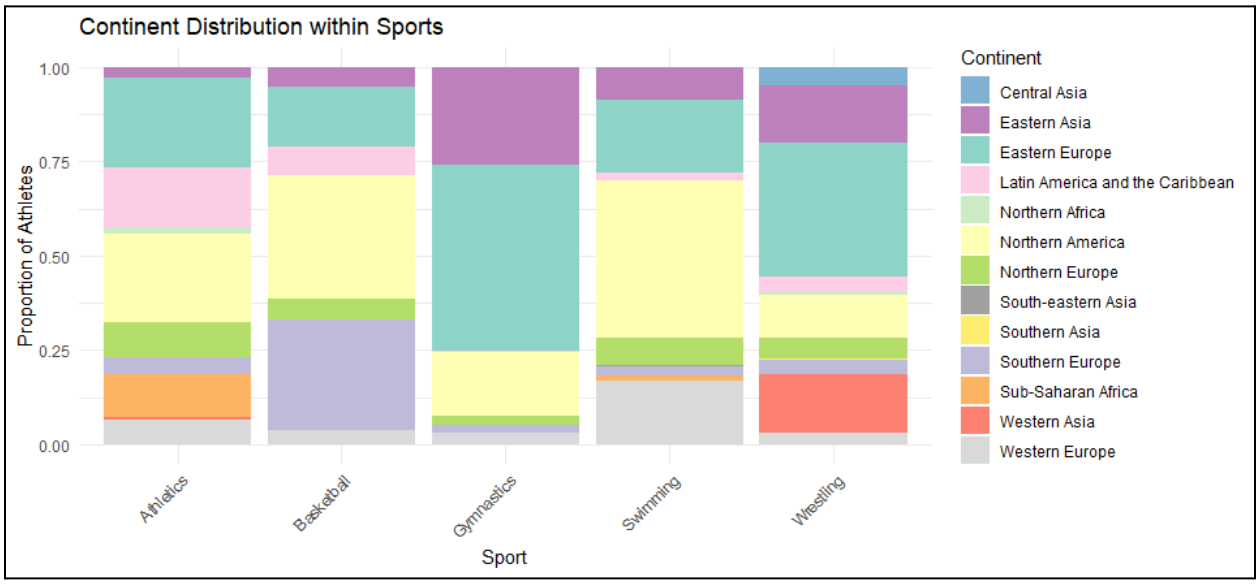
Appendix A : Graph of Probability of Participation

Graph 1: Probability of Participation in 5 Olympics Sports

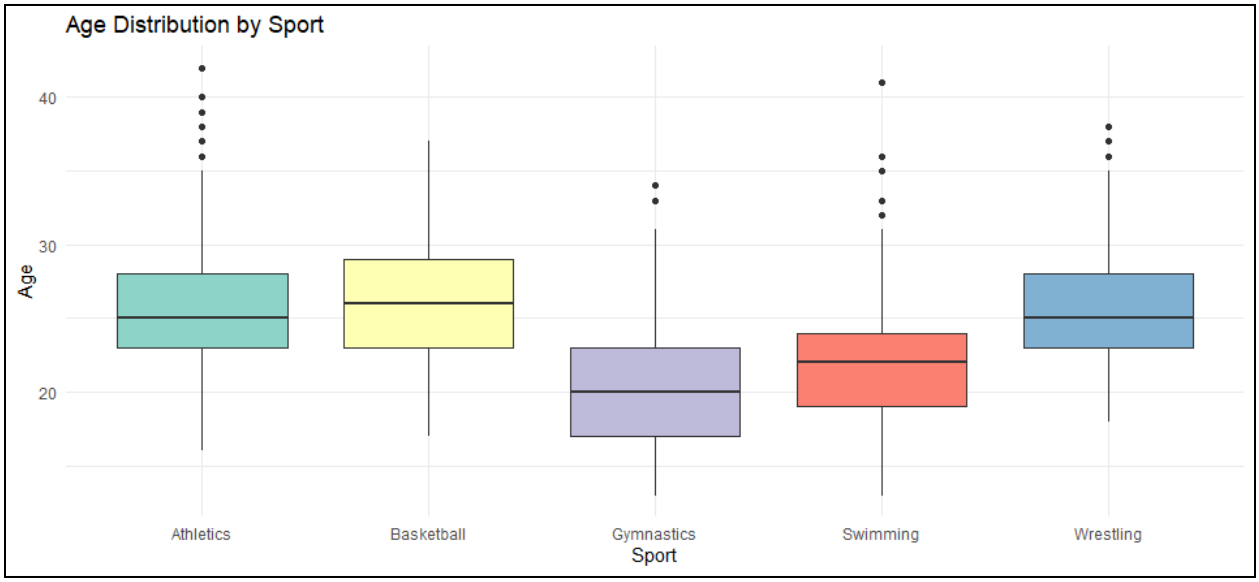


Appendix B : Variable Analysis Plots I

Graph 2: Barchart of Continent Distribution within Sports

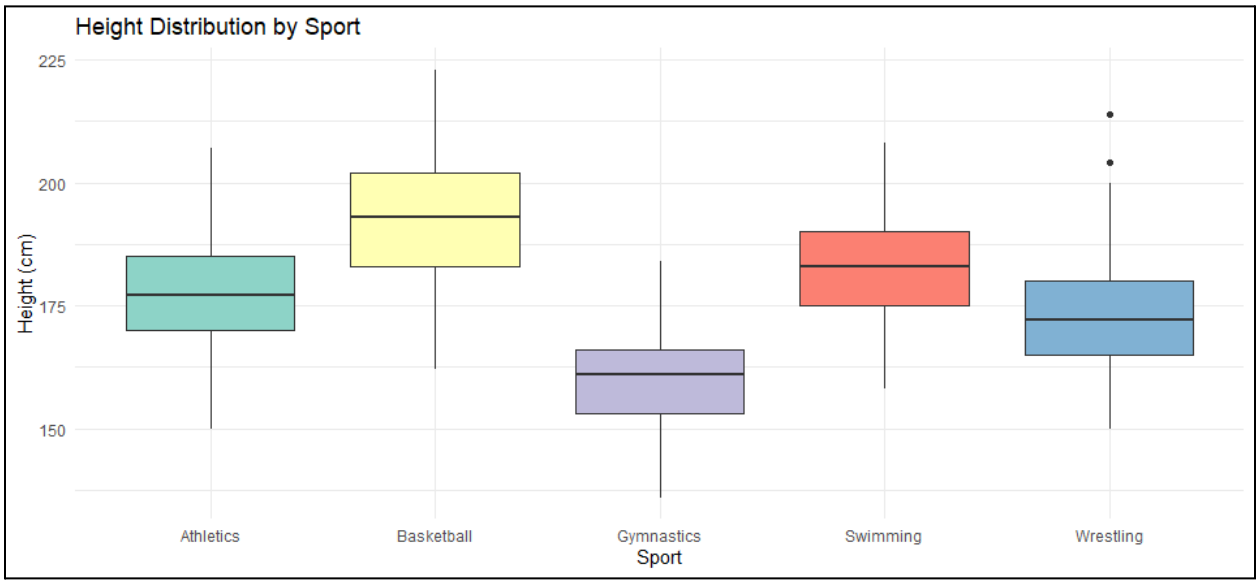


Graph 3: Boxplot of Age Distribution by Sport

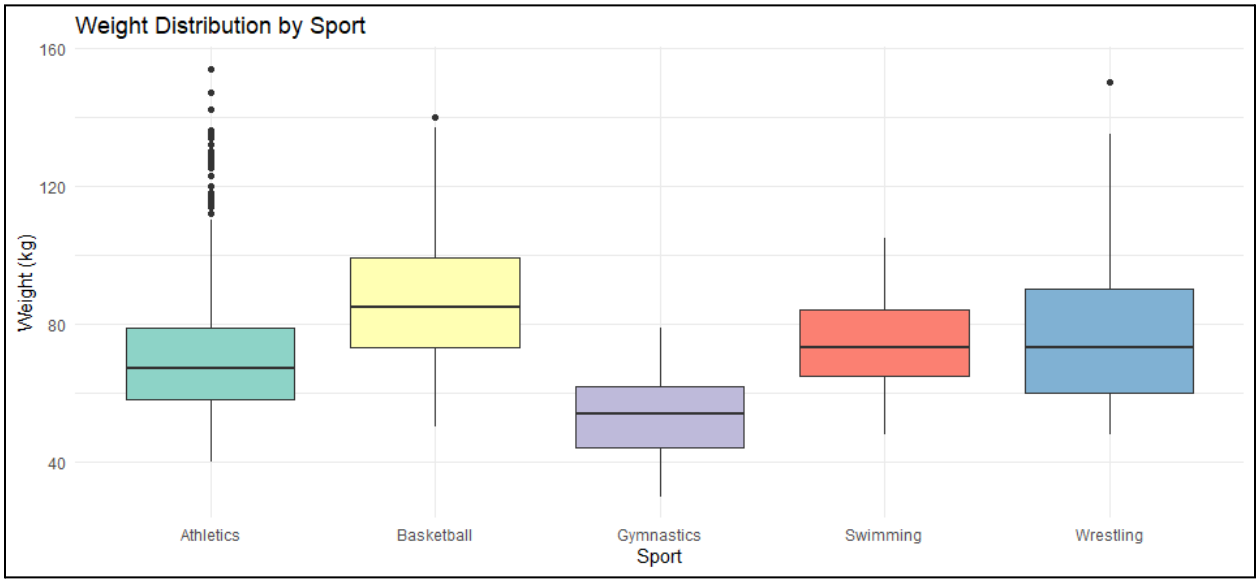


Appendix C : Variable Analysis Plots II

Graph 4: Boxplot of Height Distribution by Sport

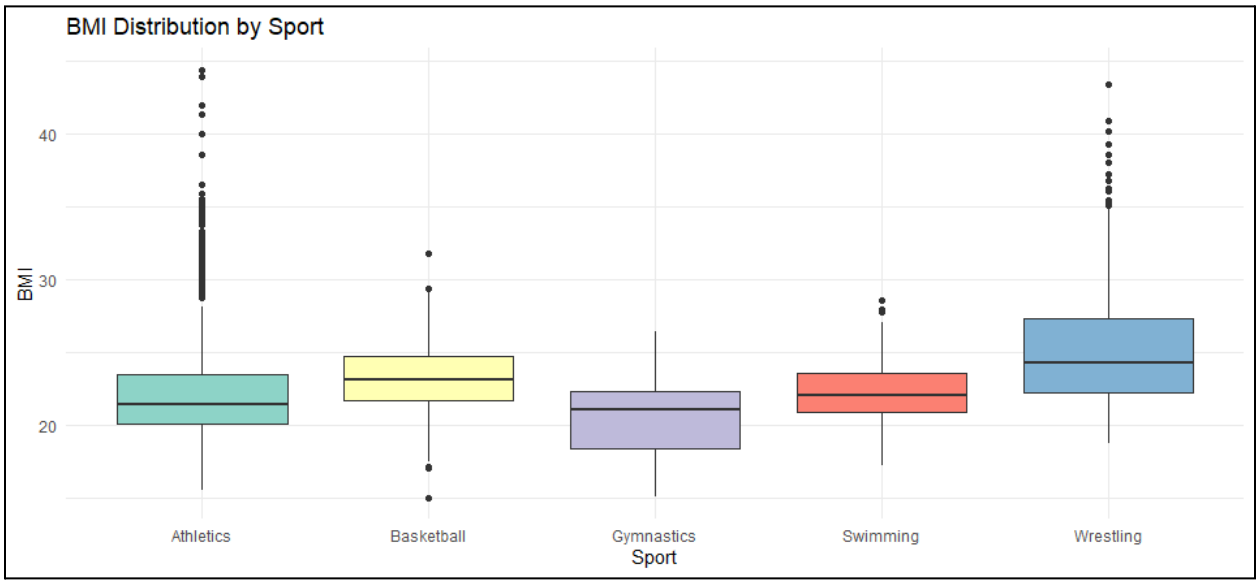


Graph 5: Boxplot of Weight Distribution by Sport

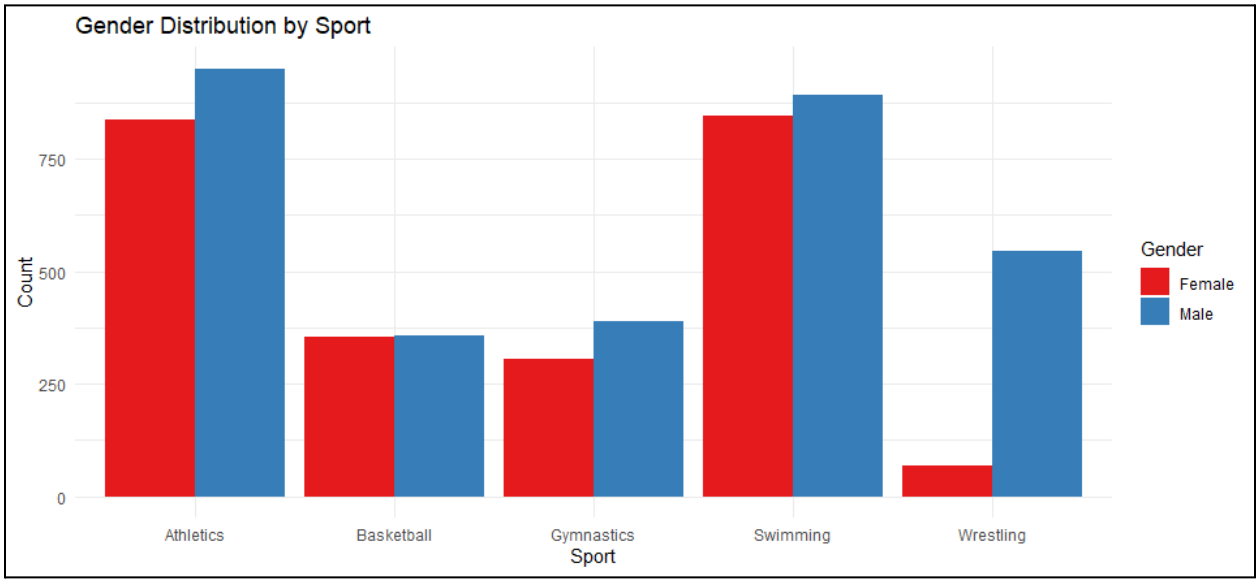


Appendix D : Variable Analysis Plots III

Graph 6: Boxplot of BMI Distribution by Sport

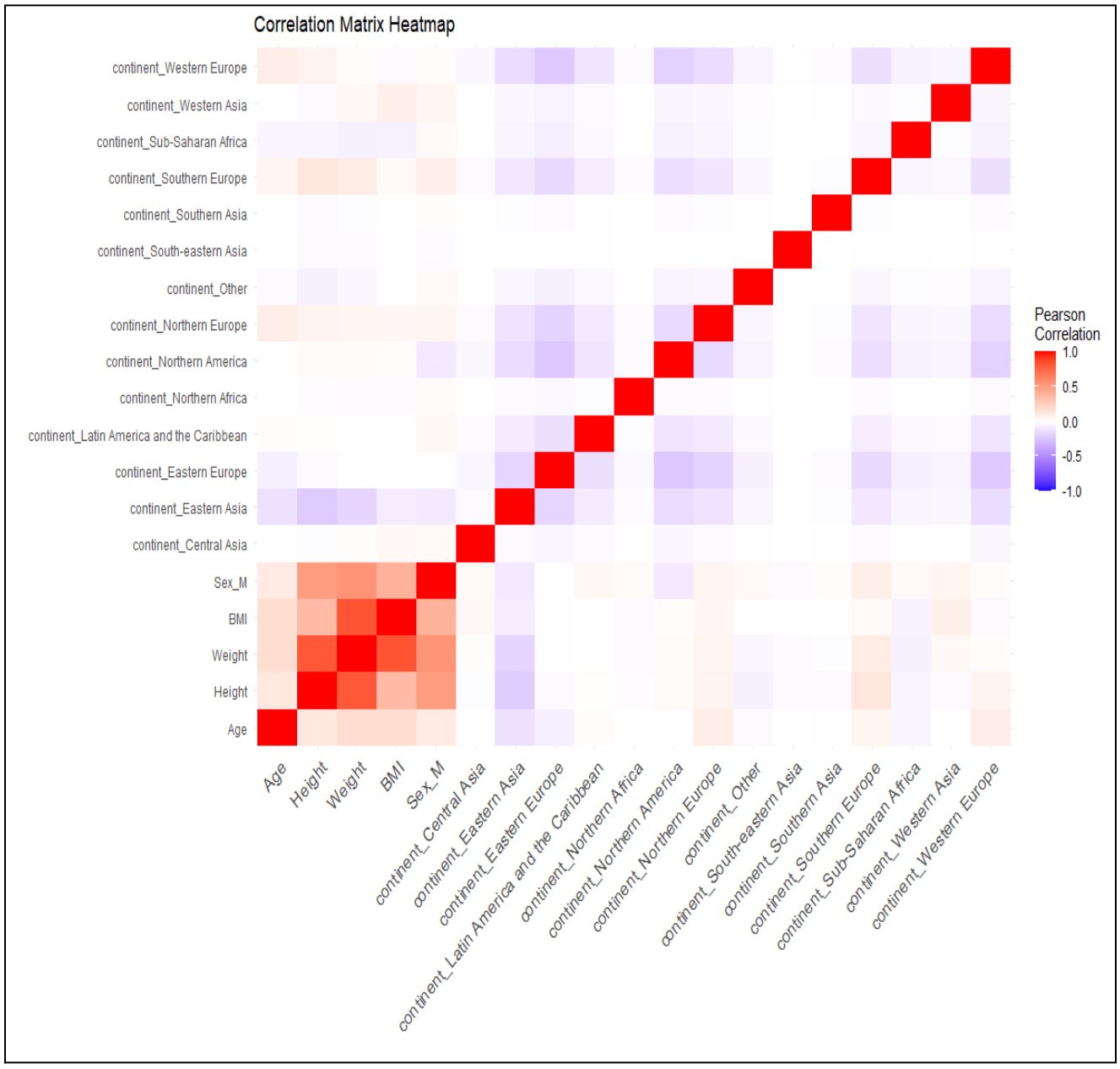


Graph 7: Double Barchart of Gender Distribution by Sport



Appendix E : Correlation Matrix Heatmap

Graph 8: Correlation Matrix Heatmap



Appendix F : Logistic Regression Results

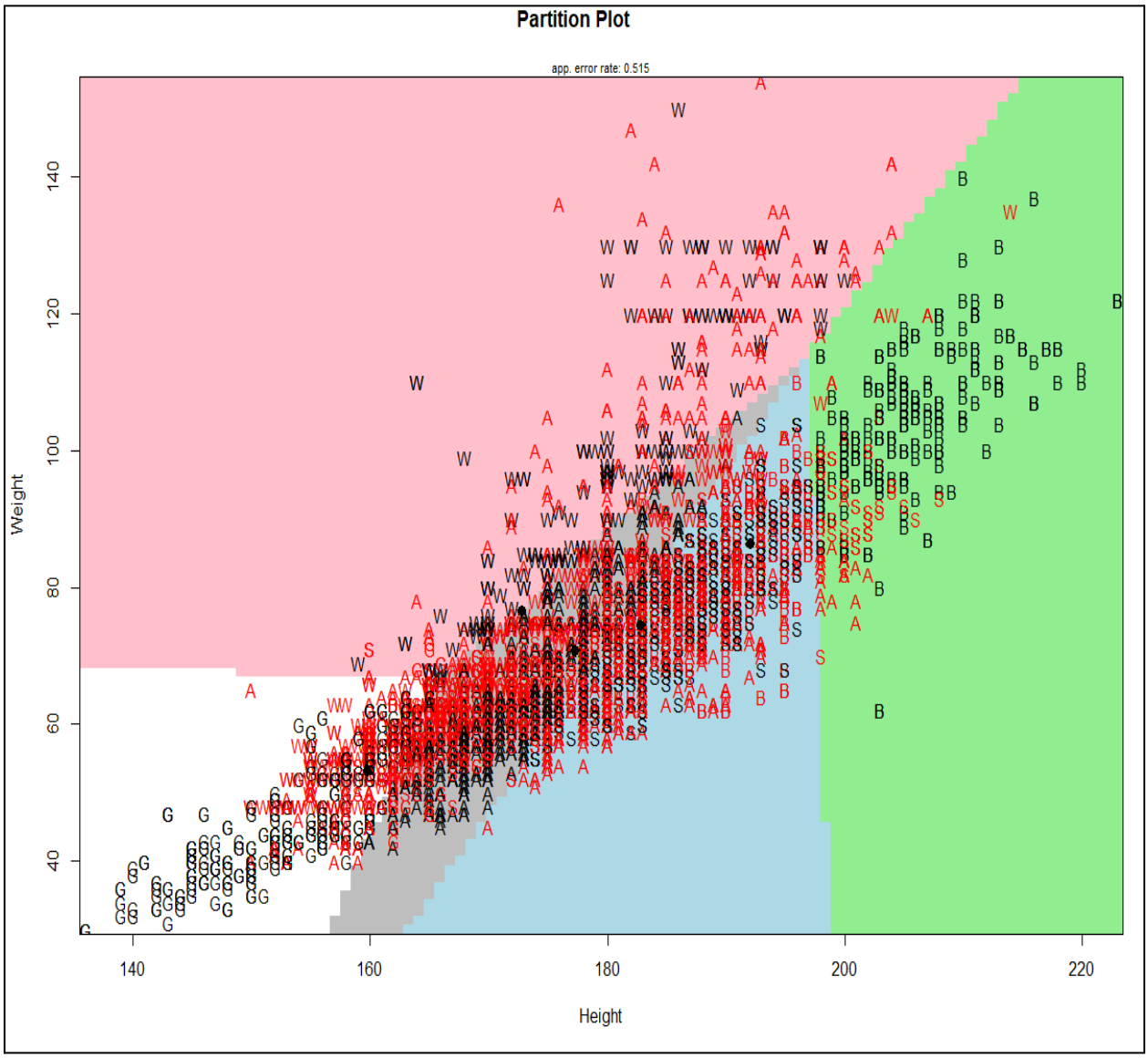
Table 1: Logistic Regression Results for Dependant variable (5 sports) based on predictors

Regression Results

	Dependent variable:				
	Basketball	Gymnastics	Swimming	Wrestling	Athletics
	(1)	(2)	(3)	(4)	(5)
Age	0.013	-0.185***	-0.289***	0.009	0.020***
	(0.009)	(0.013)	(0.009)	(0.009)	(0.005)
Height	0.170***	-0.157***	0.136***	-0.121***	0.034***
	(0.007)	(0.009)	(0.005)	(0.006)	(0.004)
Weight	-0.048***	-0.004	-0.066***	0.067***	-0.038***
	(0.005)	(0.007)	(0.004)	(0.003)	(0.003)
Northern America	1.207***	1.356***	0.946***	-0.289**	0.505***
	(0.111)	(0.193)	(0.067)	(0.143)	(0.067)
Eastern Europe	0.251*	2.331***	-0.492***	0.605***	0.240***
	(0.132)	(0.172)	(0.080)	(0.102)	(0.066)
Eastern Asia	0.876***	1.664***	-0.272**	0.275**	-1.299***
	(0.200)	(0.181)	(0.109)	(0.135)	(0.158)
Sub-Saharan Africa	-12.999	-14.027	-0.417*	-13.881	2.880***
	(198.390)	(311.834)	(0.234)	(204.682)	(0.121)
Southern Europe	1.432***	0.639*	-1.747***	-0.417*	-0.551***
	(0.118)	(0.333)	(0.180)	(0.223)	(0.125)
Constant	-31.928***	26.367***	-15.187***	12.335***	-6.402***
	(1.075)	(1.201)	(0.680)	(0.851)	(0.592)
Observations	21,439	21,439	21,439	21,439	21,439
Log Likelihood	-2,373.056	-1,704.471	-4,537.375	-2,471.176	-5,674.120
Akaike Inf. Crit.	4,764.113	3,426.943	9,092.749	4,960.351	11,366.240
Note:			*p<0.1; **p<0.05; ***p<0.01		

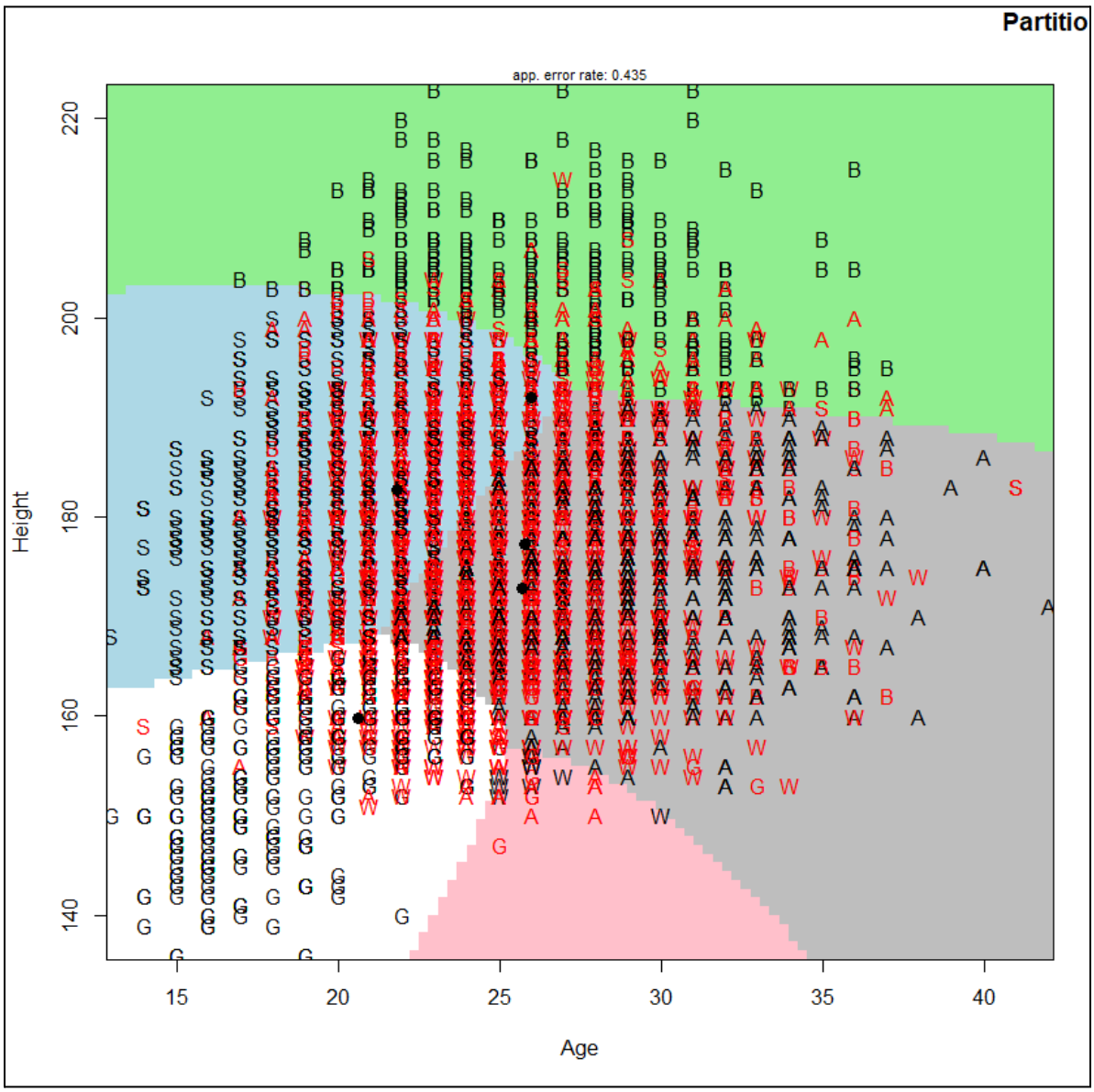
Appendix G : LDA Partition Plot

Graph 9: LDA Partition Plot for Weight against Height



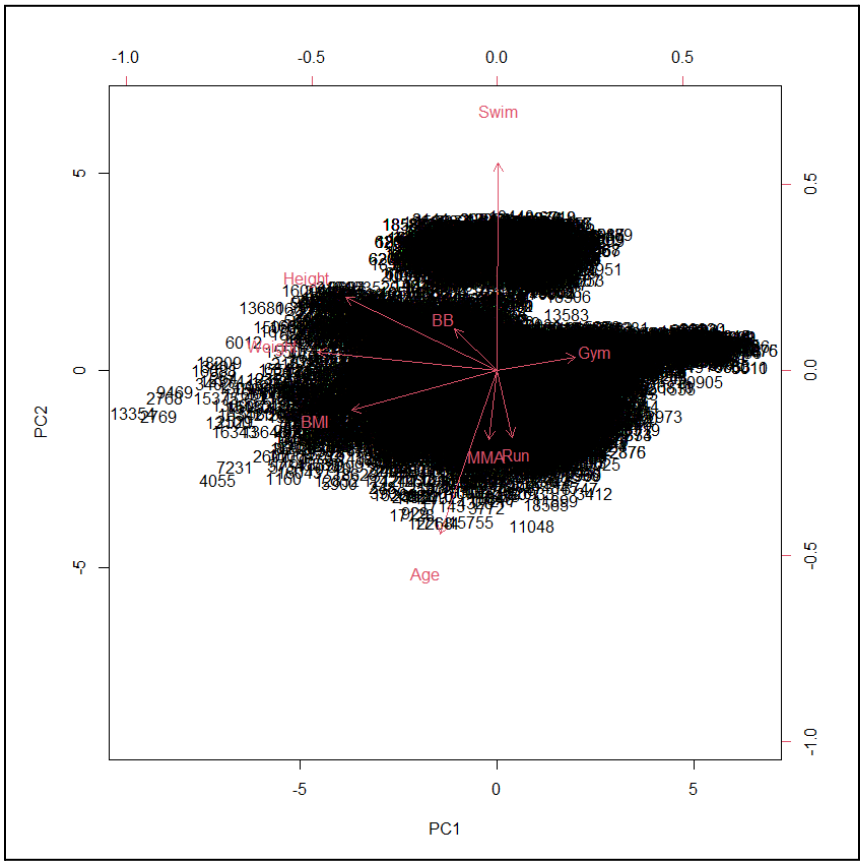
Appendix H : QDA Partition Plot

Graph 10: QDA Partition Plot for Height against Age



Appendix I : PCA Results

Graph 11: Visulazation of PCA Results for PC1 and PC2



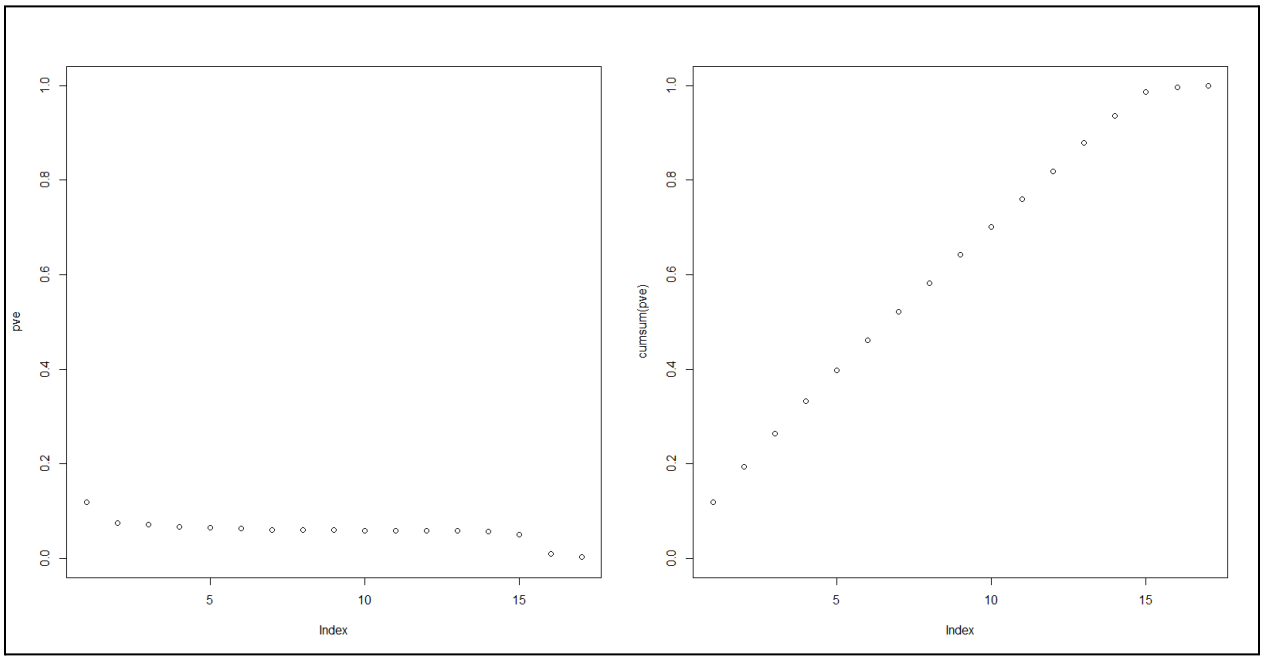
Tabe 2: Numerical Output of PCA Results for PC1 to PC8

```
Standard deviations (1, .., p=8):
[1] 1.4144907 1.1166804 1.0680543 1.0139302 1.0017588 0.9602391 0.6884382 0.4288584

Rotation (n x k) = (8 x 8):
                                 PC1         PC2         PC3         PC4         PC5         PC6
Age                         0.23634484 -0.09410896  0.04392985 -0.10050808 -0.57096740 -0.75489174
Height                      0.63764913  0.02893334  0.02434374  0.09096601  0.28932639  0.01722139
Weight                      0.63493490  0.03418695  0.01925447  0.13116001  0.27610941 -0.07019861
continent_Northern America  0.05585013 -0.58921166 -0.59127081  0.03057356 -0.05480872  0.19935848
continent_Eastern Europe   -0.01855546  0.77890249 -0.27672292  0.03575116 -0.08413857  0.05579468
continent_Eastern Asia     -0.30354541 -0.15326497  0.40717779  0.46844498  0.41020089 -0.32350930
continent_Sub-Saharan Africa -0.07028050 -0.03013081  0.04873383 -0.84007030  0.44963851 -0.20108093
continent_Southern Europe   0.18400996 -0.10430788  0.63463480 -0.19248030 -0.36541374  0.48679090
                                 PC7         PC8
Age                         0.15505888  0.057830305
Height                     -0.01664441  0.706693454
Weight                      0.06912531 -0.701550198
continent_Northern America  0.50600407  0.019660836
continent_Eastern Europe    0.55103122  0.035848904
continent_Eastern Asia      0.47220168  0.056904436
continent_Sub-Saharan Africa  0.20842574 -0.003173518
continent_Southern Europe   0.38053930 -0.012143262
```

Appendix J : PCA Percentage of Variance Explained

Graph 12: Cummulative Percentage of Variance Explained Plot



7. Code

Please view the attached R file with this submission in order to view the code.