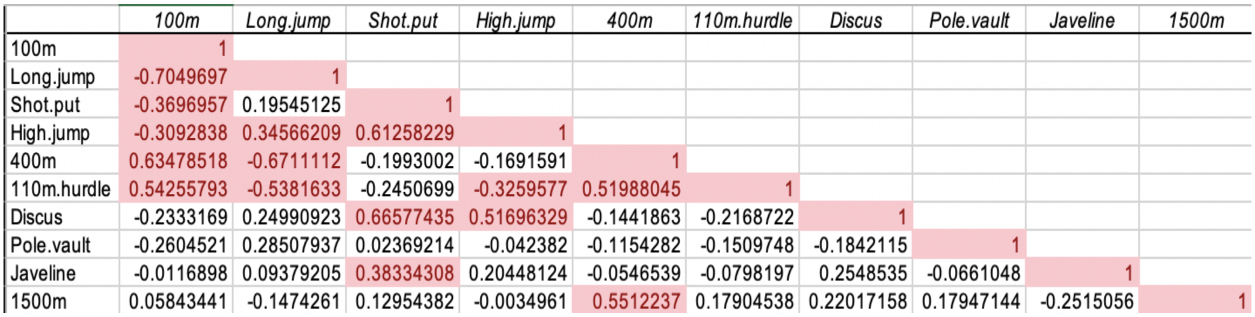
Principal Component Analysis Project
Of Decathlon Athletes

Hanna Song
EMSE 6765

Introduction

This report is conducted by 28 of the Decathlon Athletes Dataset using Minitab and Excel. From Principal Component Analysis, the project is to re-expressing the multivariate data to define the patterns from reducing the variables. The correlation matrix is provided in Table 1 with Threshold of 0.3

Table 1. Correlation Matrix



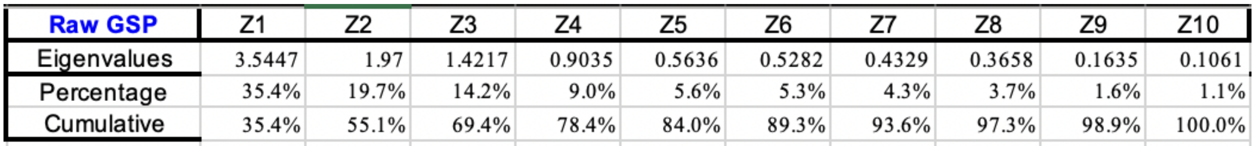
	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m
100m	1									
Long.jump	-0.7049697	1								
Shot.put	-0.3696957	0.19545125	1							
High.jump	-0.3092838	0.34566209	0.61258229	1						
400m	0.63478518	-0.6711112	-0.1993002	-0.1691591	1					
110m.hurdle	0.54255793	-0.5381633	-0.2450699	-0.3259577	0.51988045	1				
Discus	-0.2333169	0.24990923	0.66577435	0.51696329	-0.1441863	-0.2168722	1			
Pole.vault	-0.2604521	0.28507937	0.02369214	-0.042382	-0.1154282	-0.1509748	-0.1842115	1		
Javeline	-0.0116898	0.09379205	0.38334308	0.20448124	-0.0546539	-0.0798197	0.2548535	-0.0661048	1	
1500m	0.05843441	-0.1474261	0.12954382	-0.0034961	0.5512237	0.17904538	0.22017158	0.17947144	-0.2515056	1

Table 1 shows how 100m running has about half of correlation between other sport events while long jump and shot put lead to the second. This report will discuss in detail about principal component analysis performed to assess these decathlon sports.

Reduced Model Development

To start PCA, we need to compute Eigenvectors and Eigenvalues by Minitab. The eigenvalues explain the percentage of variance, and the cumulative contribution of each component are listed in Table 2.

Table 2. PCA Eigenvalues and their contributions



Raw GSP	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
Eigenvalues	3.5447	1.97	1.4217	0.9035	0.5636	0.5282	0.4329	0.3658	0.1635	0.1061
Percentage	35.4%	19.7%	14.2%	9.0%	5.6%	5.3%	4.3%	3.7%	1.6%	1.1%
Cumulative	35.4%	55.1%	69.4%	78.4%	84.0%	89.3%	93.6%	97.3%	98.9%	100.0%

Figure 1: Scree Plot

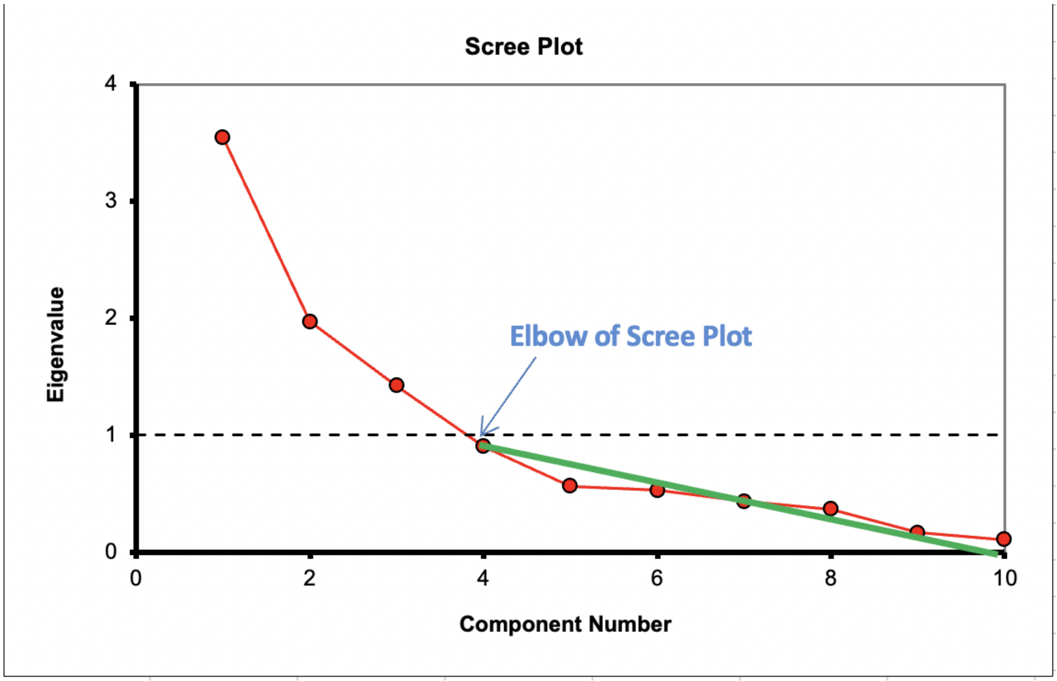
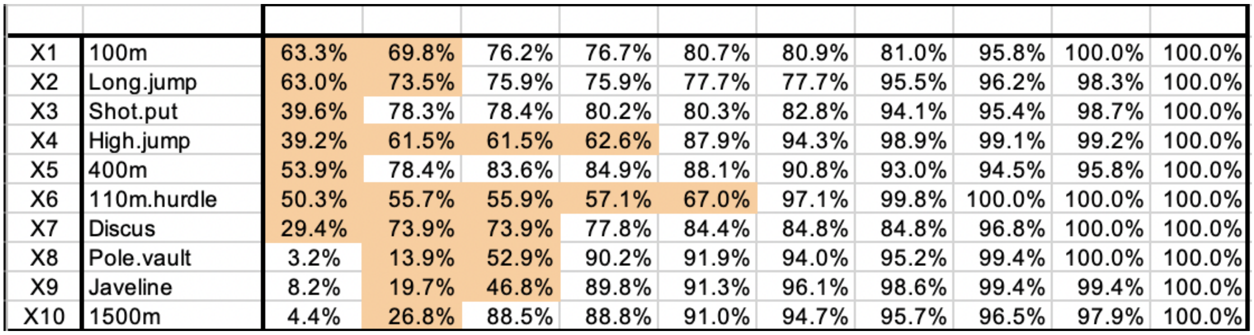


Table 3: Percentage of each Athletes



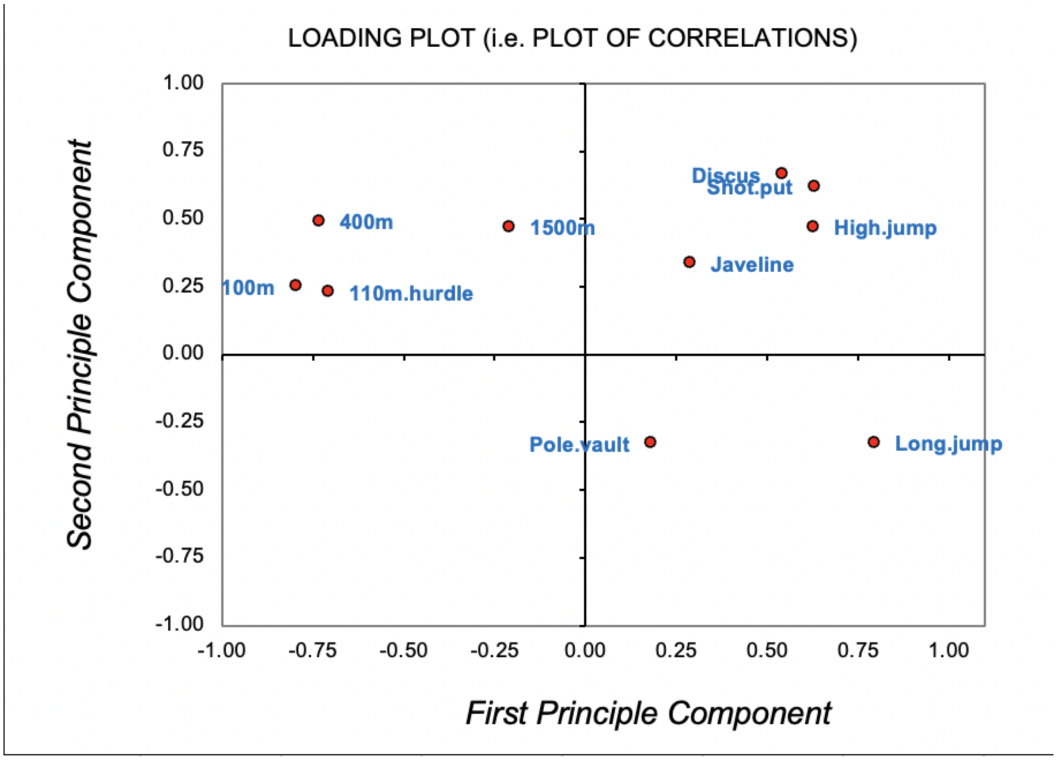
X1	100m	63.3%	69.8%	76.2%	76.7%	80.7%	80.9%	81.0%	95.8%	100.0%	100.0%
X2	Long.jump	63.0%	73.5%	75.9%	75.9%	77.7%	77.7%	95.5%	96.2%	98.3%	100.0%
X3	Shot.put	39.6%	78.3%	78.4%	80.2%	80.3%	82.8%	94.1%	95.4%	98.7%	100.0%
X4	High.jump	39.2%	61.5%	61.5%	62.6%	87.9%	94.3%	98.9%	99.1%	99.2%	100.0%
X5	400m	53.9%	78.4%	83.6%	84.9%	88.1%	90.8%	93.0%	94.5%	95.8%	100.0%
X6	110m.hurdle	50.3%	55.7%	55.9%	57.1%	67.0%	97.1%	99.8%	100.0%	100.0%	100.0%
X7	Discus	29.4%	73.9%	73.9%	77.8%	84.4%	84.8%	84.8%	96.8%	100.0%	100.0%
X8	Pole.vault	3.2%	13.9%	52.9%	90.2%	91.9%	94.0%	95.2%	99.4%	100.0%	100.0%
X9	Javeline	8.2%	19.7%	46.8%	89.8%	91.3%	96.1%	98.6%	99.4%	99.4%	100.0%
X10	1500m	4.4%	26.8%	88.5%	88.8%	91.0%	94.7%	95.7%	96.5%	97.9%	100.0%

The first step of PCA is to determine how many components to be included in a reduced model. I will use Kaiser's rule (Table 2) and scree plot (Figure 1) for the report. From Kaiser's rule, the first three components are retained since these are all greater than one. Second approach is to examine the scree plot from figure 1 where it recommends retaining the first three components as well. From the scree plot and table, the first three components account for 69.4% of the total variance in the dataset. Moreover, Table 3 shows that using the third component is to increase minimum variance of 13.9% to 52.9% which seems valid for using the third one as well. Therefore, my final approach is to select a reduced model using three components.

Data Assessment

This analysis is performed using the loading plots of the first and second component provided in Figure 2. These first two components account for 55% of the total variation which have a good description of the dataset. The figure shows striking characteristics of both components.

Figure 2: Loading Plot of 1st and 2nd Component



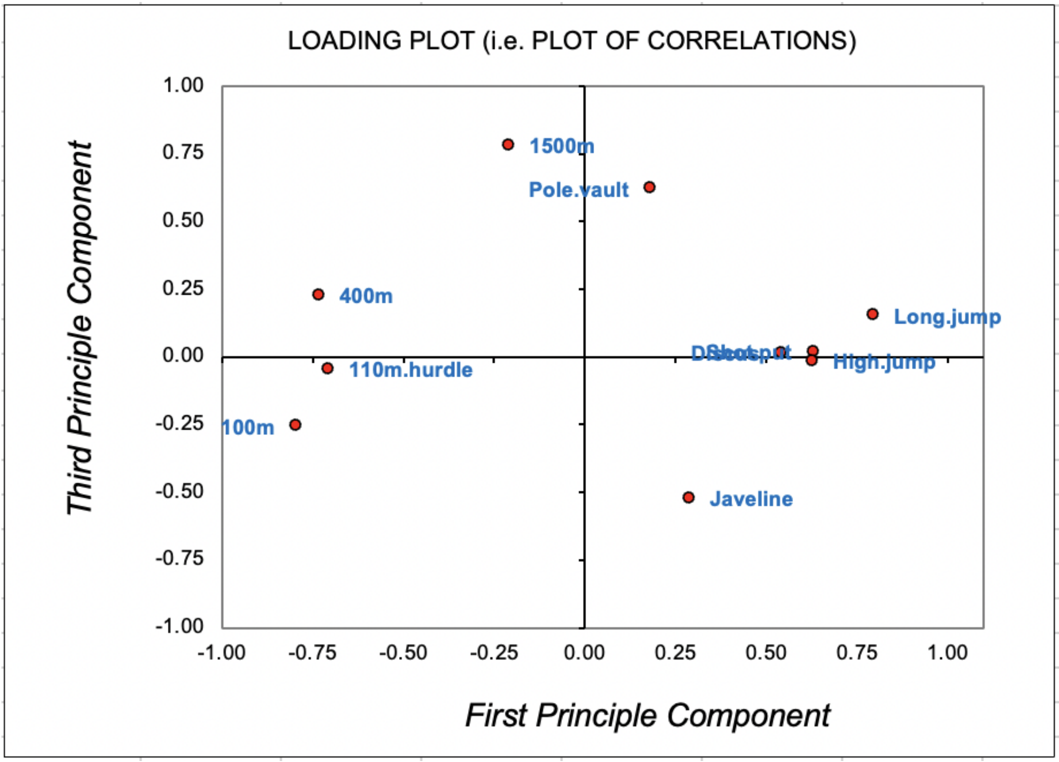
Overall, the loading plot (figure 3) shows correlation of each different type of sports together. For instance, the running sports of 110m, 400m, 110m hurdle, and 1500m correlated on the top left, while throwing sports of discus, shot put, and javelin correlated on the top right. This suggests that sports with similar characteristics are strongly correlated together, moreover, indicates athletes performance is relatively good in similar sports.

However, Pole vault and long jump also have a low correlation with other sports. They are somewhat different from other sports in that pole vault is also using equipment to assist an athlete to pass over the bar, and long jump is a jumping sport but rather it attempts to leap as far as possible. This indicates athletes might need other skills to have a good performance.

First Principle Component shows is the component of common variation. It shows sports on the positive axis have good results from athletes while sports on the negative axis represent a lack of results from the athletes. We can conclude that athletes have good abilities in high jump, long jump and shot put following up with discus, javelin and pole vault. However, athletes have poor abilities in running sports where all of the running sports are on the negative axis.

Second Principle Component is to classify the sport from insecurity that pole vault and long jump are only in the negative axis. It shows the classification of the sports that pole vault and long jump have different characteristics with other sports. Both of them are considered to be easily injured in which athletes would have significant injuries from the sport events. While sports in the positive axis seem to be softer sports than these two.

Figure 3: Loading Plot of 1st and 3rd Component



Further looking into strong correlation between throwing sports, I added figure 4 (loading plot of 1st and 3rd component). This plot indicated throwing sports (Discus and Shot put) are strongly positively correlated together, however, Javelin seems to be slightly off from strong correlation in which Javelin is using a spear while discus and shot put is using rounded equipment (disc and ball). This also indicates athletes performing in either discus or shot put will definitely have a good performance on both.

Figure 5: Score Plot of 1st and 2nd Component

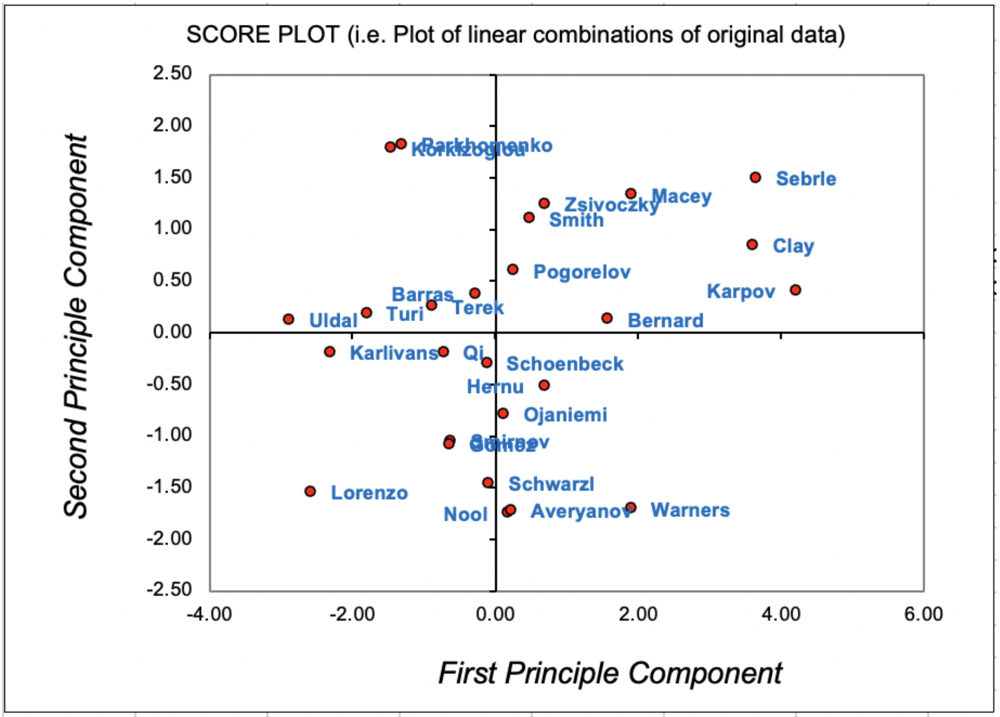


Figure 5 shows the score plot of first and second principle components which allows us to determine the performance of athletes during Decathlon. From the plot, we can determine how many athletes are clustered near zero which also means normally distributed. Moreover, three athletes seem to have significant results: Sebrie, Clay and Karpov. They are located in the strongly positive axis on the first and second principle that indicates their performance is excessively good. While one athlete, Lorenzo, seems to have a poor performance located in the strongly negative axis of both components.

In conclusion, the analysis from this report indicates that athletes' performance varied from their capabilities of what kind of sports they are good at. Moreover, the normally distributed score plot suggests that Decathlon athletes are well-rounded athletes even if they are varied with athletic capabilities.