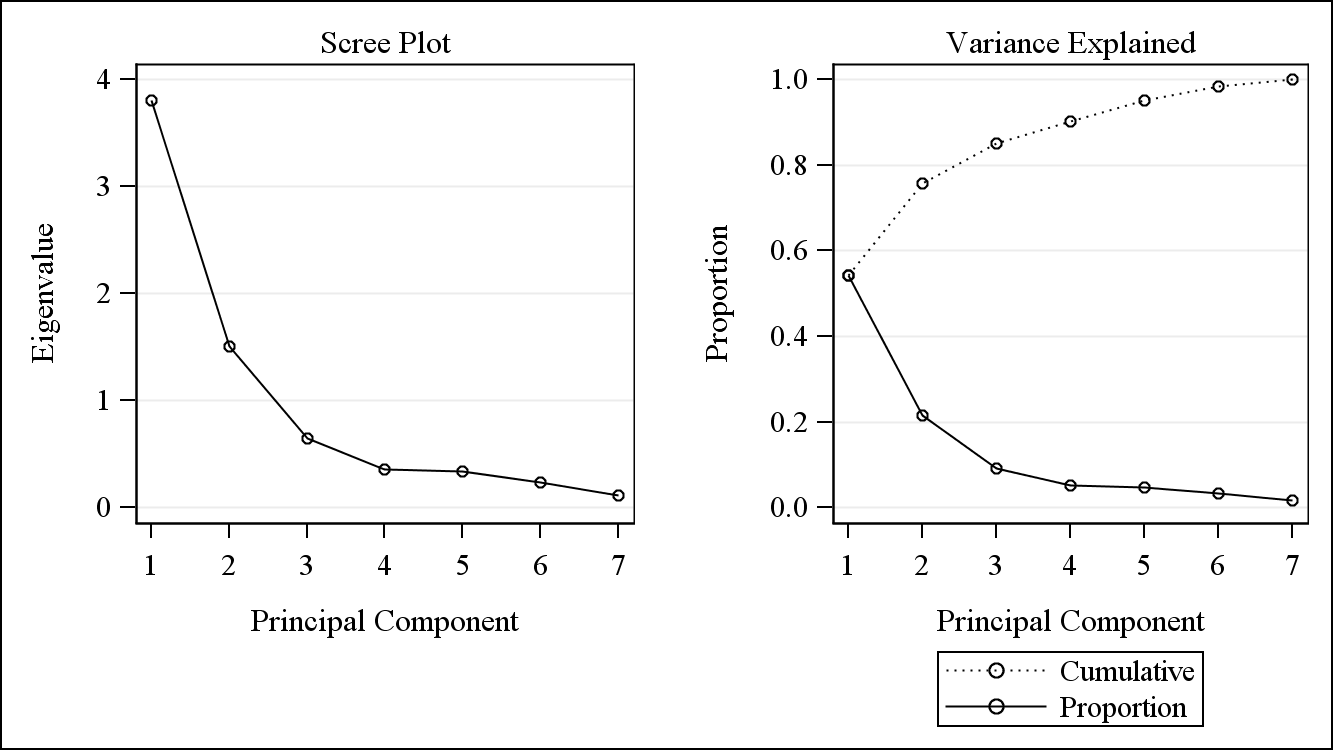
**STAT 448 Homework #5 Solution**

1)

Principal Component Analysis on Criminal Dataset

| **Eigenvalues of the Correlation Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 3.79947455 | 2.29714628 | 0.5428 | 0.5428 |
| **2** | 1.50232826 | 0.85252087 | 0.2146 | 0.7574 |
| **3** | 0.64980739 | 0.28975051 | 0.0928 | 0.8502 |
| **4** | 0.36005688 | 0.02089442 | 0.0514 | 0.9017 |
| **5** | 0.33916246 | 0.10390935 | 0.0485 | 0.9501 |
| **6** | 0.23525311 | 0.12133576 | 0.0336 | 0.9837 |
| **7** | 0.11391735 |  | 0.0163 | 1.0000 |

| **Eigenvectors** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** | **Prin6** | **Prin7** |
| **headlen** | 0.276304 | 0.364768 | 0.882275 | -.085739 | 0.067404 | 0.005385 | -.016387 |
| **headbr** | 0.211864 | 0.639204 | -.257528 | 0.687074 | -.081294 | 0.034956 | 0.017627 |
| **facebr** | 0.295145 | 0.512393 | -.381448 | -.698562 | 0.100718 | 0.033741 | -.074626 |
| **finger** | 0.437558 | -.234940 | -.069924 | 0.101600 | 0.619237 | 0.318242 | 0.503390 |
| **forearm** | 0.455705 | -.276667 | -.036669 | 0.113115 | 0.039077 | 0.290306 | -.784757 |
| **foot** | 0.450234 | -.178437 | -.059125 | 0.052999 | 0.034409 | -.870489 | 0.014451 |
| **height** | 0.435689 | -.179540 | -.006212 | -.081627 | -.769765 | 0.233030 | 0.352699 |



Part 1) Based on the average eigenvalue test, two principal components should be kept since only the first two have eigenvalues greater than 1 (the average eigenvalue). Based on the scree plot, it appears three principal components should be kept since after the third principal component the eigenvalues become relatively constant. The first four principal components explain 90% of the total variation in the data, so four principal components should be kept in order to explain a minimum total variance of 90%. Overall, these results indicate it may be appropriate to keep between two and four principal components.

Part 2) In many cases it is difficult to interpret more than the first couple components. In this case however, we can interpret most of them, though later components will be of little significance because they pick up relatively small amounts of the variation. Here are the interpretations:

* The first component is a general size feature. When any of the measurements increases, this component increases.
* The second contrasts head measurements to other body measurements. Individuals with really large second component would have large heads relative to overall body size. Really small second components would have small heads relative to overall body size.
* For the third component, the non-head measures are small compared to the head-related measurements. The head length measurement is positive and the head and face breadth measurements are negative, so this feature contrasts head height and width. Larger values indicate longer, thinner heads and smaller values indicate shorter, wider heads.
* In the fourth component, the head and face breadth measurement coefficients are several times larger than the coefficients for other measures. It’s not clear to me how the head and face breadth measurements are differentiated, but this component contrasts head width with face width. Larger values indicate wider heads relative to face width and smaller values indicate smaller heads relative to face width.
* The fifth is mostly contrasting finger width and height.
* The sixth contrasts finger, forearm and height length with left foot length.
* The last component contrasts finger length and height with forearm length.

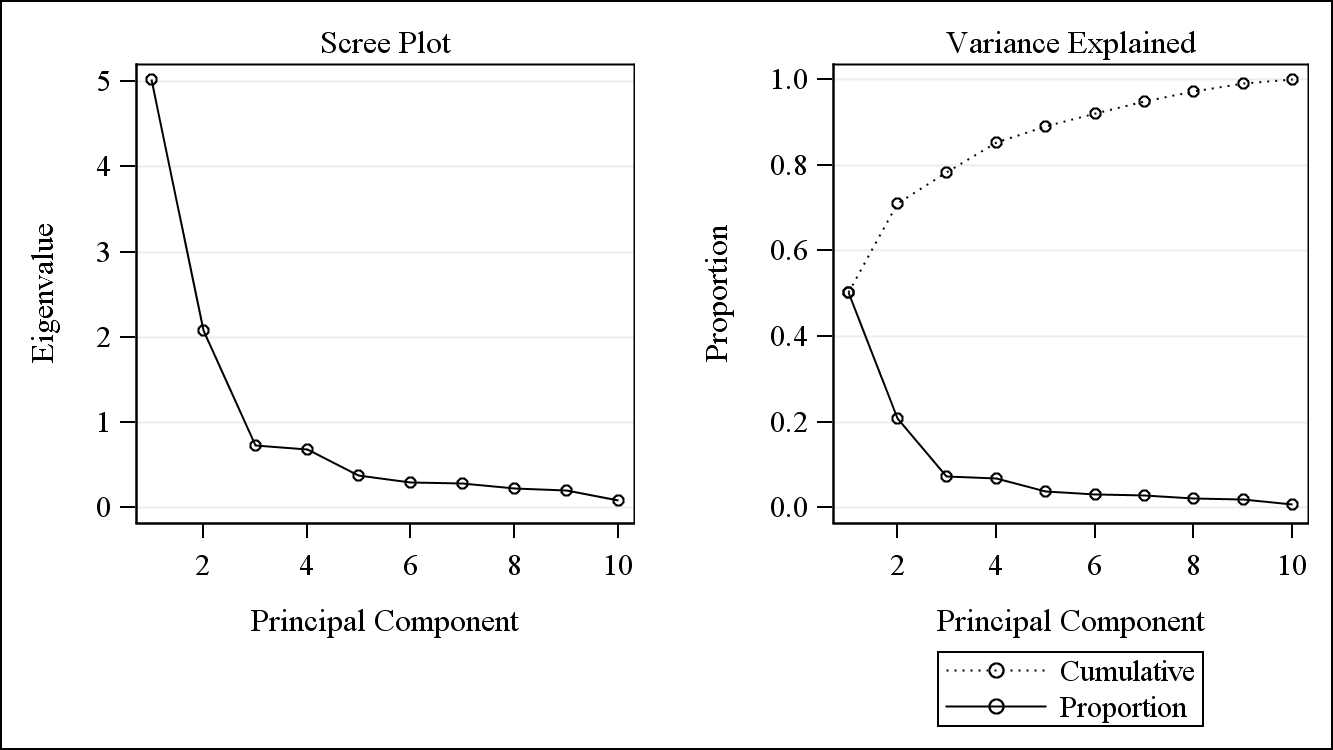
The last few components seem to be picking up on some symmetries or asymmetries that seem a bit atypical in general. The last components tell us relatively little about the general variation in the data. While the contrasts are interpretable, it should not be too surprising that the contrasts are a bit atypical—they are picking up on residual features in the original data.

2)

Principal Component Analysis on Decathlon Dataset

| **Eigenvalues of the Correlation Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 5.02351759 | 2.94361616 | 0.5024 | 0.5024 |
| **2** | 2.07990143 | 1.34443321 | 0.2080 | 0.7103 |
| **3** | 0.73546822 | 0.04972974 | 0.0735 | 0.7839 |
| **4** | 0.68573847 | 0.30947846 | 0.0686 | 0.8525 |
| **5** | 0.37626002 | 0.07415224 | 0.0376 | 0.8901 |
| **6** | 0.30210778 | 0.01660124 | 0.0302 | 0.9203 |
| **7** | 0.28550654 | 0.06172664 | 0.0286 | 0.9489 |
| **8** | 0.22377990 | 0.01904323 | 0.0224 | 0.9712 |
| **9** | 0.20473667 | 0.12175328 | 0.0205 | 0.9917 |
| **10** | 0.08298339 |  | 0.0083 | 1.0000 |

| **Eigenvectors** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** | **Prin6** | **Prin7** | **Prin8** | **Prin9** | **Prin10** |
| **run100** | 0.358634 | 0.204157 | -.300422 | -.175914 | -.272176 | -.107377 | 0.514409 | -.156995 | -.569470 | 0.096598 |
| **Ljump** | 0.361195 | 0.197895 | 0.076949 | 0.114807 | -.455732 | -.613165 | -.370971 | 0.253141 | 0.136399 | -.085474 |
| **shot** | 0.323991 | -.394275 | -.125214 | 0.173586 | 0.267976 | -.043078 | 0.285206 | 0.392126 | -.000555 | -.620479 |
| **Hjump** | 0.267477 | 0.007412 | 0.855859 | -.356066 | 0.065562 | 0.009049 | 0.146969 | -.148113 | -.052874 | -.135686 |
| **run400** | 0.294367 | 0.427198 | -.232214 | 0.016886 | 0.285836 | -.001911 | -.061246 | -.583619 | 0.367484 | -.339896 |
| **hurdle** | 0.373388 | 0.131247 | -.149139 | -.387308 | -.054891 | 0.372650 | 0.130146 | 0.419996 | 0.490603 | 0.308543 |
| **discus** | 0.306692 | -.416387 | -.046053 | 0.041081 | 0.445627 | -.426246 | 0.010495 | -.186134 | 0.077745 | 0.554693 |
| **polevlt** | 0.388947 | -.061929 | -.075492 | -.066216 | 0.155663 | 0.381392 | -.664091 | 0.029753 | -.472409 | -.006121 |
| **javelin** | 0.293275 | -.298380 | 0.135860 | 0.565922 | -.453519 | 0.374427 | 0.101264 | -.298874 | 0.169800 | 0.109295 |
| **run1500** | 0.083578 | 0.545602 | 0.230672 | 0.567738 | 0.359039 | 0.065941 | 0.150832 | 0.306116 | -.143247 | 0.222629 |



Based on the scree plot, it appears using either two or four principal components would be appropriate. This is identical to the result after the athlete who finished last was removed from the data. Based on the average eigenvalue, the first two principal components should be kept, which is also identical to the previous result. If we want the eigenvalues to explain 70% of the variation, two principal components should be kept, which differs from the previous result where four principal components should be kept based on the same criteria.

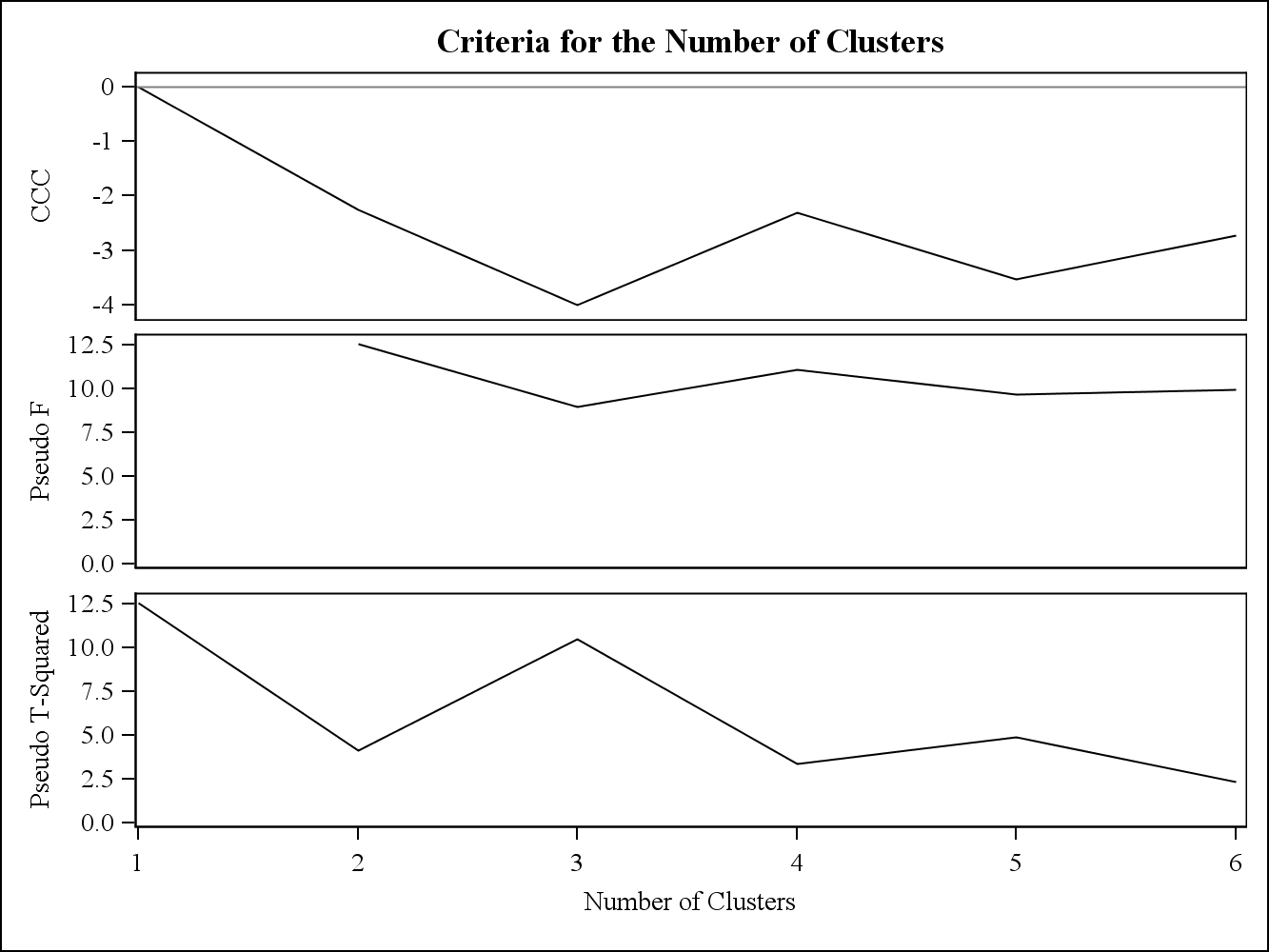
The first principal component has positive coefficients for every variable, indicating it represents the overall athleticism of the athlete. This was the same interpretation from the first principal component with the worst athlete removed from the dataset; however, the first principal component in this analysis explains about 16% more variation in the data compared to the first principal component in the previous analysis. Furthermore, the coefficients for shot put, high jump, discus and javelin (strength events) are larger in this principal component compared to the one with the worst athlete removed.

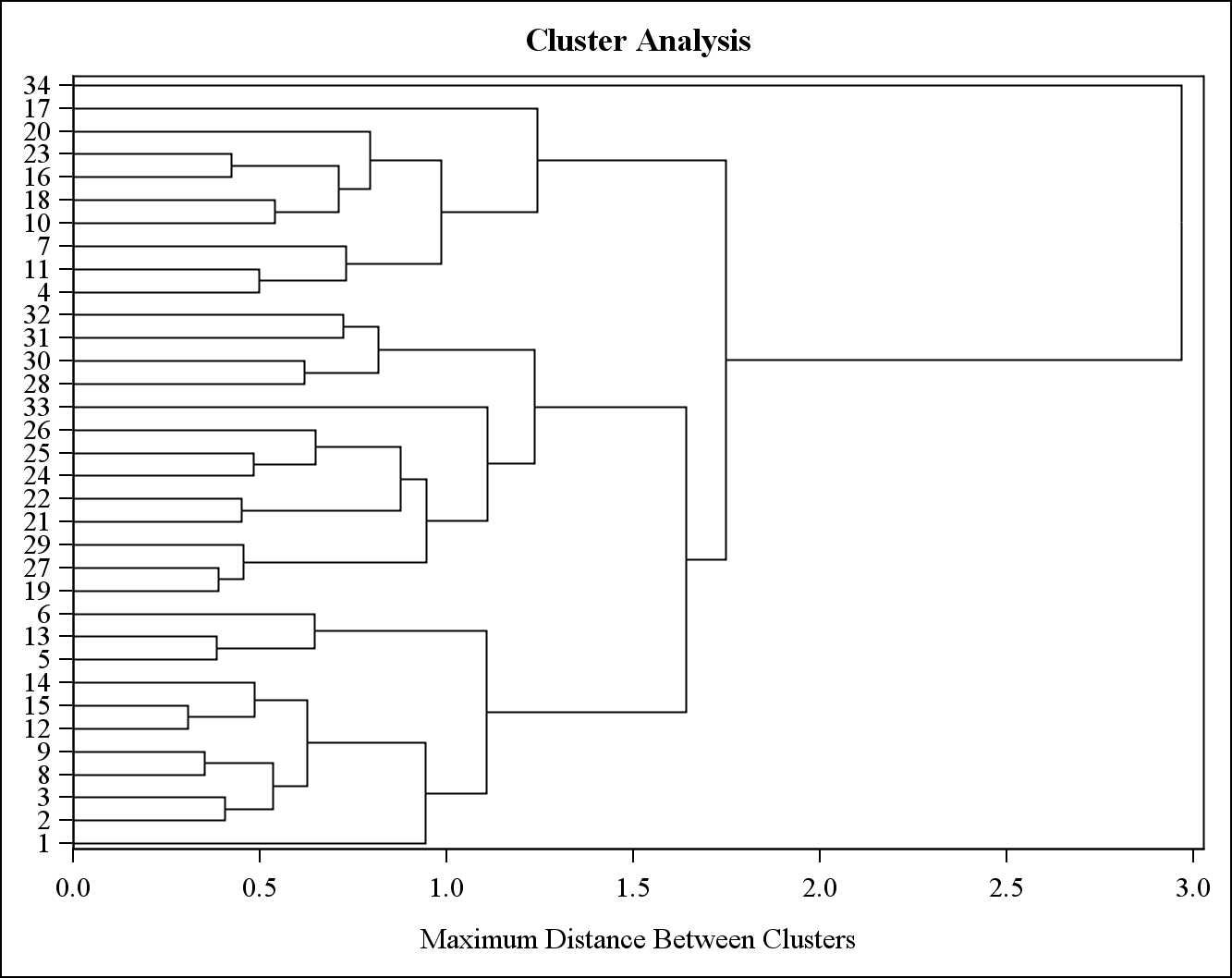
The second principal component has negative coefficients for sports that require strength such as shot put, javelin and discus and positive coefficients for sports that require speed and agility such as running events and long jump. This represents a contrast in athletes as some are better at strength events relative to speed events and others are better at speed events relative to strength events. This was the same interpretation from the second principal component with the worst athlete removed from the dataset; however, the amount of variation explained by this second principal component is about 5% lower than the second principal component with the worst athlete removed. Interestingly, although these principal components represent the same contrast, the signs of the coefficients are opposite, indicating athletes good at strength events compared to speed events will have a negative value for the second principal component in this analysis, while these same athletes would have a positive value for the second principal component in the analysis with the worst athlete removed from the data.

The third principal component is difficult to interpret. Overall, the results from this principal component analysis are similar to the results when the worst athlete was removed from the data.

3)

Complete Linkage Cluster Analysis on Decathlon Dataset





a) Based on the Cubic Clustering Criterion (CCC), a peak occurs at four clusters, which indicates four clusters should be chosen. Based on the Pseudo F statistic, three clusters should be chosen. Based on the Pseudo T-Squared statistic, it seems four clusters should be chosen; however, looking at the graph indicates that selecting two or four clusters would be reasonable. Since the linkage is complete, it is more appropriate to rely on the CCC, which indicates four clusters should be chosen.

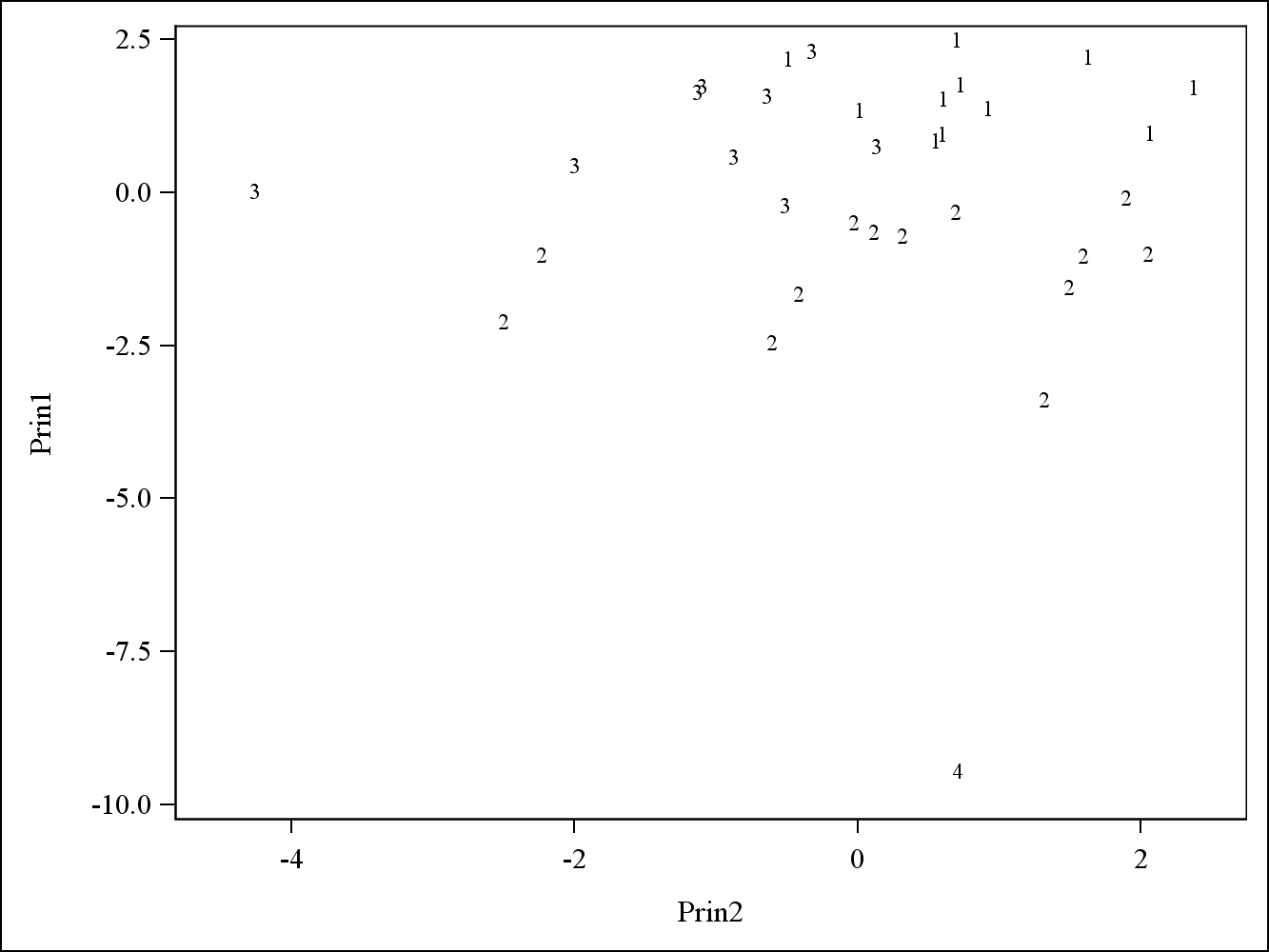
Looking at the dendrogram, four clusters should be chosen. If two clusters were chosen, one would only be one observation, which would not be helpful for analysis. Choosing four clusters would create three evenly size clusters and one cluster with only one observation. Overall, after inspecting the dendrogram and CCC, it’s clear that four clusters should be chosen.

Principal Component Analysis on Decathlon Dataset

| **Eigenvectors** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** | **Prin6** | **Prin7** | **Prin8** | **Prin9** | **Prin10** |
| **run100** | 0.358634 | 0.204157 | -.300422 | -.175914 | -.272176 | -.107377 | 0.514409 | -.156995 | -.569470 | 0.096598 |
| **Ljump** | 0.361195 | 0.197895 | 0.076949 | 0.114807 | -.455732 | -.613165 | -.370971 | 0.253141 | 0.136399 | -.085474 |
| **shot** | 0.323991 | -.394275 | -.125214 | 0.173586 | 0.267976 | -.043078 | 0.285206 | 0.392126 | -.000555 | -.620479 |
| **Hjump** | 0.267477 | 0.007412 | 0.855859 | -.356066 | 0.065562 | 0.009049 | 0.146969 | -.148113 | -.052874 | -.135686 |
| **run400** | 0.294367 | 0.427198 | -.232214 | 0.016886 | 0.285836 | -.001911 | -.061246 | -.583619 | 0.367484 | -.339896 |
| **hurdle** | 0.373388 | 0.131247 | -.149139 | -.387308 | -.054891 | 0.372650 | 0.130146 | 0.419996 | 0.490603 | 0.308543 |
| **discus** | 0.306692 | -.416387 | -.046053 | 0.041081 | 0.445627 | -.426246 | 0.010495 | -.186134 | 0.077745 | 0.554693 |
| **polevlt** | 0.388947 | -.061929 | -.075492 | -.066216 | 0.155663 | 0.381392 | -.664091 | 0.029753 | -.472409 | -.006121 |
| **javelin** | 0.293275 | -.298380 | 0.135860 | 0.565922 | -.453519 | 0.374427 | 0.101264 | -.298874 | 0.169800 | 0.109295 |
| **run1500** | 0.083578 | 0.545602 | 0.230672 | 0.567738 | 0.359039 | 0.065941 | 0.150832 | 0.306116 | -.143247 | 0.222629 |

| **Eigenvalues of the Correlation Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 5.02351759 | 2.94361616 | 0.5024 | 0.5024 |
| **2** | 2.07990143 | 1.34443321 | 0.2080 | 0.7103 |
| **3** | 0.73546822 | 0.04972974 | 0.0735 | 0.7839 |
| **4** | 0.68573847 | 0.30947846 | 0.0686 | 0.8525 |
| **5** | 0.37626002 | 0.07415224 | 0.0376 | 0.8901 |
| **6** | 0.30210778 | 0.01660124 | 0.0302 | 0.9203 |
| **7** | 0.28550654 | 0.06172664 | 0.0286 | 0.9489 |
| **8** | 0.22377990 | 0.01904323 | 0.0224 | 0.9712 |
| **9** | 0.20473667 | 0.12175328 | 0.0205 | 0.9917 |
| **10** | 0.08298339 |  | 0.0083 | 1.0000 |

Plot of 1st Principal Component vs 2nd Principal Component labeled by cluster



b) The first two principal components explain a minimum of 70% of variation in the data, indicating the first two principal components should be plotted against each other labeled by cluster. Based on the plot, it seems cluster 1 has positive values for principal component 1 and principal component 2. Cluster 2 appears to have negative values for principal component 1, while having a wide range of values with respect to principal component 2. Cluster 3 seems to have mostly positive values for principal component 1 and mostly negative values for principal component 2. Cluster 4 is a single observation and has the lowest value of principal component 1 by a wide margin.

c) Since cluster 1 has positive values for principal component 1 and 2, this indicates athletes in cluster 1 have better than average overall athletic ability and performed better at running and agility events compared to strength events such as shot put. Since cluster 2 has negative values for principal component 1, athletes in cluster 2 have lower than average overall athletic ability. Since cluster 3 has mostly positive values for principal component 1 and negative values for principal component 2, athletes in this cluster have better overall athletic ability than average and performed better at strength events compared to running and agility events. Finally, due to the extremely negative value cluster 4 has for principal component 1, the athlete in this cluster has the worst overall athletic ability compared to every other competitor, indicating why this athlete was grouped in his own cluster.